Statistica Sinica **7**(1997), 875-892

STATISTICAL APPLICATIONS OF THE POISSON-BINOMIAL AND CONDITIONAL BERNOULLI DISTRIBUTIONS

Sean X. Chen and Jun S. Liu

New York University and Stanford University

Abstract: The distribution of $Z_1 + \cdots + Z_N$ is called Poisson-Binomial if the Z_i are independent Bernoulli random variables with not-all-equal probabilities of success. It is noted that such a distribution and its computation play an important role in a number of seemingly unrelated research areas such as survey sampling, case-control studies, and survival analysis. In this article, we provide a general theory about the Poisson-Binomial distribution concerning its computation and applications, and as by-products, we propose new weighted sampling schemes for finite population, a new method for hypothesis testing in logistic regression, and a new algorithm for finding the maximum conditional likelihood estimate (MCLE) in case-control studies. Two of our weighted sampling schemes are direct generalizations of the "sequential" and "reservoir" methods of Fan, Muller and Rezucha (1962) for simple random sampling, which are of interest to computer scientists. Our new algorithm for finding the MCLE in case-control studies is an iterative weighted least squares method, which naturally bridges prospective and retrospective GLMs.

Key words and phrases: Case-control studies, conditional Bernoulli distribution, iterative weighted least squares, logistic regression, PPS sampling, Poisson-Binomial, survey sampling, weighted sampling.

1. Introduction

Suppose $Z_1, \ldots, Z_N$ are independently distributed Bernoulli random variables, each with probability of success p_i. We write $\mathbf{Z} = (Z_1, \ldots, Z_N)$ and $\mathbf{p} = (p_1, \ldots, p_N)$. Then $S_Z = Z_1 + \cdots + Z_N$ is called a *Poisson-Binomial* random variable with parameter $\mathbf{p}$. When all the p's are equal, this reduces to the Binomial distribution, which played an important role in the early history of probability theory. When N is large and all the p_i are small but not necessarily equal, the distribution of S_Z is well approximated by a Poisson distribution due to the well-known *Law of Small Numbers*. In this article, we are mainly concerned with exact computation of the distribution of S_Z. Closely related is the so-called *conditional Bernoulli* model defined as the conditional distribution of $\mathbf{Z}$ given that $S_Z = n$. As will be shown later, this model is very useful in a number of different areas.

It is easy to show that the exact formula for calculating the Poisson-Binomial distribution is

$$P(S_Z = n) = \left\{ \prod_{i=1}^{N} (1 - p_i) \right\} \sum_{i_1 < \cdots < i_n} w_{i_1} \cdots w_{i_n}, \tag{1}$$

where $w_i = p_i/(1 - p_i)$, for $i = 1, \ldots, N$, and the summation is over all possible combinations of distinct $i_1, \ldots, i_n$ from $\{1, \ldots, N\}$. A naive way of computing the summation on the right hand side of the equation needs to sum $N!/[n!(N - n)!]$ terms, which is impractical even when n and N are of moderate sizes. Recursive formulas that require only $O(nN)$ operations for computing this sum are analyzed in Section 2.

One of the main motivations for the investigation reported in this article is the problem of *weighted sampling* in survey studies. In many surveys, the individual units are not necessarily drawn with equal probabilities. A problem often considered in the literature, sometimes called "probability-proportional-to-size (PPS)" sampling, is to define a particular sampling scheme that achieves prespecified marginal probabilities π_i for the ith population unit to be included in the survey sample, where

$$0 < \pi_i < 1 \ \text{ for } \ i = 1, \ldots, N, \ \text{ and } \ \sum_{i=1}^{N} \pi_i = n. \tag{2}$$

Chen, Dempster and Liu (1994) proposed a maximum entropy distribution for the sampled units. Let $\mathbf{D} = (D_1, \ldots, D_N)$ be a random vector on space $\mathcal{D}^n$, where D_i takes the values 1 or 0 according to whether the ith unit is in or out of the sample, and

$$\mathcal{D}^n = \{\mathbf{d} = (d_1, \ldots, d_N) : \ d_i = 0 \text{ or } 1, \ \text{and } d_1 + \cdots + d_N = n\}.$$

Then the maximum entropy model for $\mathbf{D}$, on space $\mathcal{D}^n$, has the form

$$P(\mathbf{D} = \mathbf{d}) = \prod_{i=1}^{N} w_i^{d_i} \Bigg/ \sum_{\mathbf{z} \in \mathcal{D}^n} \left(\prod_{i=1}^{N} w_i^{z_i} \right), \tag{3}$$

where $(w_1, \ldots, w_N)$ is chosen to satisfy the constraints

$$\pi_i = E(D_i) = \sum_{\mathbf{d} \in \mathcal{D}^n} d_i P(\mathbf{D} = \mathbf{d}). \tag{4}$$

It can be easily shown that the maximum entropy model in (3) is just a conditional Bernoulli model with the w_i proportional to $p_i/(1-p_i)$. So from now on, we always use $\mathbf{D}$ to denote a conditional Bernoulli random vector on the space $\mathcal{D}^n$.

Methods for deriving the p_i or w_i from the π_i and for drawing samples from this model are provided in Chen, Dempster and Liu (1994). We discuss extensions of these methods in connection with a number of applications in GLM.

In analyzing survey data, it is often the case that only the marginal sample inclusion probabilities π_i, instead of the overall information about the sampling design, are available to analysts. A strong support for the use of the conditional Bernoulli model in PPS sampling comes from Sugden and Smith (1984), who show that only such a model can warrant the appropriateness of the π_i as a summary of the sampling design. Furthermore, the conditional Bernoulli model for PPS sampling enables an easy calculation of high-order joint inclusion probabilities and guarantees the nonnegativity of the Yates and Grundy (1953)'s variance estimator (Chen, Dempster and Liu (1994)).

The rest of the article is organized as follows. Section 2 illustrates and compares two methods for computing the Poisson-Binomial distribution, identifying circumstances under which one method is more efficient than the other. Section 3 displays two applications of the Poisson-Binomial distribution in generalized linear models (GLMs), one dealing with hypothesis testing in logistic regression model, and the other with case-control studies. Section 4 gives five efficient methods for sampling from the conditional Bernoulli model. Section 5 concludes with a brief summary.

2. Computation of Poisson-Binomial Probabilities

The distribution function of the Poisson-Binomial variable S_Z can be written as

$$P(S_Z = n) = \sum_{\mathbf{d} \in \mathcal{D}^n} \left(\prod_{i=1}^{N} w_i^{d_i} \right) \prod_{i=1}^{N} (1 + w_i)^{-1},\ n = 0, 1, \ldots, N, \tag{5}$$

where $w_i = p_i/(1 - p_i)$.

For operational purpose, we use the following notation: $S = \{1, \ldots, N\}$, capital letters such as A, B, or C for subsets of S, $A^c = S \backslash A$ for the complement of A in S, and $|A|$ for the number of elements of A. Also

$$R(k, C) \overset{def}{=} \sum_{B \subset C, |B| = k} \left(\prod_{i \in B} w_i \right) \tag{6}$$

for any non-empty set $C \subset S$ and $1 \leq k \leq |C|$. $R(0, C) = 1$, and $R(k, C) = 0$ for any $k > |C|$.

Using (6), we can rewrite (5) as

$$P(S_Z = n) = R(n, S) \prod_{i \in S} (1 + w_i)^{-1},\ n = 0, 1, \ldots, N. \tag{7}$$

We note from (7) that the function $R(n, S)$ differs from $P(S_Z = n)$ only by a normalizing constant, and thus completely characterizes the distribution of S_Z. Now we present two recursive methods for computing R economically.

Method 1. (Chen, Dempster and Liu (1994)) Define $T(i, C) = \sum_{j \in C} w_j^i$ for any $i \geq 1$ and $C \subset S$. Then for any $1 \leq k \leq |C|$,

$$R(k, C) = \frac{1}{k} \sum_{i=1}^{k} (-1)^{i+1} T(i, C) R(k - i, C). \tag{8}$$

Note that Method 1 is closely related to the Newton's identities for polynomials (Stein (1990)).

Method 2. (Gail, Lubin and Rubinstein (1981)) For any $C \subset S$ and $1 \leq k \leq |C|$,

$$R(k, C) = R(k, C \backslash \{k\}) + w_k R(k - 1, C \backslash \{k\}). \tag{9}$$

Method 2 was first proposed by Howard (1972) in her discussion of Cox (1972) for analyzing proportional hazard models with discrete survival times. Gail, Lubin and Rubinstein (1981) elaborate on the method and apply it further to retrospective studies. A new iterative weighted least squares method for the same application is proposed in Section 3.

By setting $w_i = 1$ for all i, we see that Method 1 is a natural generalization of

$$\binom{c}{k} = \frac{c}{k} \left[\binom{c}{k-1} - \binom{c}{k-2} + \cdots + (-1)^{k+1} \binom{c}{0} \right].$$

Similarly, Method 2 is a natural generalization of the formula

$$\binom{c}{k} = \binom{c-1}{k} + \binom{c-1}{k-1}. \tag{10}$$

The use of (9) for the recursive generation of $R(k, C)$ is illustrated in Table 1 for the case $n = 2, N = 4$. Starting from the upper-left corner of the table, i.e. $cell(0, 0)$, each cell in the table is generated recursively using (9) till $cell(n, N)$ at the lower-right corner is filled. Specifically, $cell(i, j)$ (where $1 \leq i \leq \min(n, j) \leq N$) is generated by $cell(i, j - 1) + w_j \times cell(i - 1, j - 1)$. Note that the desired quantity $R(2, \{1, 2, 3, 4\})$ is indeed given in $cell(2, 4)$.

We can compare the costs for computing $R(n, S)$ by the two methods above. For simplicity and convenience, we only consider arithmetic operations, such as additions and multiplications, used in each algorithm, but exclude non-arithmetic operations, such as managing the data arrays and checking whether $k > |C|$ for $R(k, C)$, from our discussion. This is because the efficiency of the non-arithmetic operations usually depend on the type of the computer, the programming language and the details in the implementation of the algorithm.

Table 1. Recursive generation of $R(n, S)$ for $n = 2$, $N = 4$

$$\begin{tikzcd}
 & & & & N & \\
 & & 0 & 1 & 2 & 3 & 4 \\
 & 0 & 1 \arrow[r] \arrow[dr, "\times w_1"] & 1 \arrow[r] \arrow[dr, "\times w_2"] & 1 \arrow[dr, "\times w_3"] & — & — \\
n & 1 & 0 & w_1 \arrow[r, "+"] \arrow[dr, "\times w_2"] & w_1 + w_2 \arrow[r, "+"] \arrow[dr, "\times w_3"] & w_1 + w_2 + w_3 \arrow[dr, "\times w_4"] & — \\
 & 2 & 0 & 0 & w_1w_2 \arrow[r, "+"] & w_1w_2+w_1w_3+w_2w_3 \arrow[r, "+"] & \begin{array}{c} w_1w_2 + w_1w_3 + w_1w_4+ \\ w_2w_3 + w_2w_4 + w_3w_4 \end{array}
\end{tikzcd}$$

For Method 1, it takes $N - 1$ additions to get $T(1, S)$, and $N - 1$ additions and N multiplications to get each $T(i, S)$ for $i = 2, \ldots, n$. Once all $T(i, S)$ $(i = 1, \ldots, n)$ are obtained, it takes $k - 1$ additions and $2k - 1$ multiplications to get $R(k, S)$ from $T(1, S), \ldots, T(k, S)$ and $R(1, S), \ldots, R(k, S)$ using the recursive formula (8), for each $k = 2, \ldots, n$. Thus it requires a total of $nN + (n^2/2) - (3n/2)$ additions and $nN - N + n^2$ multiplications to compute $R(n, S)$ from scratch. For Method 2, it can be seen from Table 1 that going from the $(k - 1)$th row to the kth row requires $N - n$ additions and $N - n + 1$ multiplications, for each $k = 1, \ldots, n$. Thus going through n rows to get the $cell(n, N)$ (i.e., $R(n, S)$) requires $nN - n^2$ additions and $nN - n^2 + n$ multiplications. The two methods have the same leading term, nN, in both the addition counts and multiplication counts, and thus require the same order of operations, $O(nN)$. There is virtually little difference in the computational cost between the two methods, especially when N and n are considerably large.

Nevertheless, note that the operations for Method 1 are spent mostly in computing $T(i, S)$ and the recursive formula (8) itself requires only $O(n^2)$ operations. Thus in the case where a large number of $R(k, C)$ with different k's and C's are to be computed, Method 1 requires considerably fewer operations than Method 2. We give three examples to illustrate this point.

Example 1. Chen, Dempster and Liu (1994) provide an iterative procedure for computing $\mathbf{w} = (w_1, \ldots, w_N)$ from the inclusion probabilities π_i (i.e., solving (4) for $\mathbf{w}$) which uses the following updating scheme:

$$w_j^{(t+1)} = \left. \frac{\pi_j R(n - 1, \{N\}^c)}{R(n - 1, \{j\}^c)} \right|_{\mathbf{w}=\mathbf{w}^{(t)}}, \ j = 1, \ldots, N-1; \ w_N^{(t+1)} = w_N^{(t)} = \pi_N. \quad (11)$$

At each iteration, we need to compute $R(n-1,\{j\}^c)$, $j=1,\ldots,N$, which requires $O(n^2N)$ operations if $T(i,\{j\}^c)$, $i=1,\ldots,n-1$, $j=1,\ldots,N$, are all available. On the other hand, it takes only $O(nN)$ operations to get all $T(i,\{j\}^c)$ using the formula $T(i,\{j\}^c)=T(i,S)-w_j^i$. In total, each iteration needs $O(n^2N)$ operations if Method 1 is used. However, no simplifications can be made on Method 2 and hence $O(nN^2)$ operations are required.

Example 2. (Chen (1993)) It is often desirable to get all the values of a Poisson-Binomial distribution. As shown in (7), this involves the computation of $R(n,S)$ for $n=1,\ldots,N$. For Method 1, this computation costs as much as the computation of $R(N,S)$ alone, which requires $3.5N^2-2.5N$ operations. For Method 2, each of $R(n,S)$, $n=1,\ldots,N$, has to be computed separately, which requires $N^3/3+N^2/2+N/6$ operations.

Example 3. (*Grouping*) In the case where there are less than N different weights in the population (e.g. within each of the case and the control group in a retrospective study, the units with the same covariate have the same weight), we can put the units with the same weight in the same group to form, say I, groups. Then the computation of $T(i,S)$ in Method 1 can be simplified by using $T(i,S)=\sum_{j\in I}n_jw_j^i$, where n_j is the number of units in the jth group with weight w_j and $\sum_{j\in I}n_j=N$. In such cases, the number of operations for computing $T(i,S)$ reduces to $(n-1)(2I-1)$. For instance, $I=2$ in the example of Section 3.2, thus Method 1 only requires $O(n^2)$ operations. However, the grouping does not help Method 2.

Besides computational cost, roundoff error is another concern that users of an algorithm generally have, especially when the algorithm is recursive and/or involves alternating series. Method 1 is both recursive and alternating, while Method 2 is recursive but not alternating. The recursive feature can cause problems for algorithms involving complex operations such as division and matrix inversion. Both Methods 1 and 2 use only additions and multiplications. Thus the recursive feature is not a problem for them. The alternating feature can cause problems for algorithms in which pairs of leading consecutive terms (e.g., $T(1,C)R(k-1,C)$ and $T(2,C)R(k-2,C)$ in (8)) are the same for the first several significant figures but have opposite signs. In such cases, special data type with extra long significant figures (e.g., "real*8" in FORTRAN and "double" in C) must be used to avoid roundoff error. To our best knowledge, no roundoff error problem is found in all numerical examples in which (8) is applied (e.g., case-control studies in Section 3.2, Chen (1993), and Chen, Dempster and Liu (1994)) and different sets of weights ranging from relatively homogeneous (i.e., $\max(w_i)/\min(w_i)$ is close to one) to relatively heterogeneous (i.e. $\max(w_i)/\min(w_i)$ is very large, say 6000) are used.

The following are some properties of the R function, which will be used to derive the weighted sampling schemes for the conditional Bernoulli model. They follow immediately from (6).

Proposition 1. *For any* $C \subset S$ *and* $1 \leq k \leq |C|$,
(a) $\sum_{j \in C} w_j R(k-1, C\backslash\{j\}) = kR(k, C)$;
(b) $\sum_{j \in C} R(k, C\backslash\{j\}) = (|C| - k)R(k, C)$;
(c) $\sum_{i=0}^{k} R(i, C)R(k-i, C^c) = R(k, S)$.

3. Applications in Generalized Linear Models

3.1. Hypothesis testing in logistic regression models

For binary response data, the linear logistic model is most popular for describing dependency of the success rate on some explanatory variables. Following the same notation as in Section 1, we let $Z_1, \ldots, Z_N$ denote the binary response variables, i.e., each Z_i is either zero or one; and let $\mathbf{x}_i$ be the $k \times 1$ covariate associated with Z_i. The *logistic regression model* for the Z_i is

$$\log\left\{\frac{P(Z_i = 1)}{P(Z_i = 0)}\right\} = \gamma + \mathbf{x}_i^T \boldsymbol{\beta}. \tag{12}$$

Hence

$$p_i = P(Z_i = 1) = \frac{\exp(\gamma + \mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\gamma + \mathbf{x}_i^T \boldsymbol{\beta})}. \tag{13}$$

Suppose we are interested in testing $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Since the complete sufficient statistics for γ and $\boldsymbol{\beta}$ are $\sum_{i=1}^N Z_i$ and $\sum_{i=1}^N Z_i \mathbf{x}_i$, respectively, the similar regions for the test are constructed from the conditional distribution of $T = \sum_{i=1}^N Z_i \mathbf{x}_i$ given $\sum_{i=1}^N Z_i = r$ (Cox and Hinkley (1974); pp 137), where r is the observed number of responses of one. Some Markov chain Monte Carlo methods (Kolassa and Tanner (1994); Diaconis and Sturmfels (1993)) can be applied to simulate this conditional distribution of T approximately. An exact method for computing the conditional distribution directly is provided by Hirji, Mehta and Patel (1987). Here we provide an exact way to simulate the Z_i, and therefore T, conditional on $\sum_{i=1}^N Z_i = r$.

If we let $w_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}_0)$, it is easy to see that the conditional distribution of $(Z_1, \ldots, Z_N)$ given $S_Z = r$ can be written as

$$P(Z_1 = z_1, \ldots, Z_N = z_N | S_Z = r) = \prod_{i=1}^N w_i^{z_i} \Big/ R(r, S),$$

which is a conditional Bernoulli distribution. Therefore, we can easily apply our schemes described in Section 4 to simulate the conditional distribution of $\mathbf{Z}$

given that the total number of ones is fixed at r. Thus the related conditional distribution of T can be simulated exactly.

In the special case of $\boldsymbol{\beta}_0 = \mathbf{0}$, the above simulation is reduced to simple random sampling. Another special case is when some of the coefficients of the covariates are known. Without loss of generality, suppose $\boldsymbol{\beta}^T = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)})^T$ and $\boldsymbol{\beta}^{(2)}$ is known. Then the construction of similar regions for $H_0 : \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}_0^{(1)}$ still requires weighted sampling, and our schemes can be applied.

3.2. Conditional inference in case-control studies

Let Z_i=0 or 1 be the disease status indicating for cases or controls, and let $\mathbf{x}_i$ be the observed exposure variable. In a case-control study, what one can observe is the information concerning the retrospective probability $p(\mathbf{x} \mid Z)$, i.e., the change of the distribution of exposure variable $\mathbf{x}$ caused by the change of the disease status. However, we are more used to prospective modeling, i.e., to thinking about $p(Z \mid \mathbf{x})$, the effect of changing $\mathbf{x}$ on the disease status.

Suppose prospectively $p(Z \mid \mathbf{x})$ can be modeled by a generalized linear model (GLM),

$$\log\left\{\frac{p(Z=1 \mid \mathbf{x})}{p(Z=0 \mid \mathbf{x})}\right\} = g(\mathbf{x}^T\boldsymbol{\beta})$$

in which g is a known link function, and $\boldsymbol{\beta}$ is a coefficient vector with the same dimension as $\mathbf{x}_i$. For example, in the logistic regression case, $g(x) = x$. When prospective data are available, techniques for fitting GLMs have been well documented (McCullagh and Nelder (1989)). In a retrospective study, however, such techniques can not be directly applied.

From basic probability formulas, we have for a retrospective study that

$$p(\mathbf{x} \mid Z) = \frac{p(Z \mid \mathbf{x})p(\mathbf{x})}{\int p(Z \mid \mathbf{x})p(\mathbf{x})d\mathbf{x}}.$$

Hence to estimate $\boldsymbol{\beta}$ directly, it is necessary to eliminate a possibly infinite dimensional parameter, $p(\mathbf{x})$. A convenient way of doing this is through a conditional inference argument, i.e., the conditional probability of observing the $\mathbf{x}$'s and the Z's given that $S_Z = n$ is

$$\exp(L) \overset{def}{=} p(Z_1, \ldots, Z_N, \mathbf{x}_1, \ldots, \mathbf{x}_N \mid S_Z = n) = \frac{\prod_{i=1}^N p(Z_i \mid \mathbf{x}_i)p(\mathbf{x}_i)}{\sum_\sigma \prod_{i=1}^N p(Z_{\sigma(i)} \mid \mathbf{x}_i)p(\mathbf{x}_i)}, \quad (14)$$

where σ denotes a permutation of $\{1, \ldots, N\}$, and the summation is over all such permutations.

The observed $\mathbf{Z}$ given $S_Z = n$ must be a point in $\mathcal{D}^n$. Let $\mathbf{d} = (d_1, \ldots, d_N)$ denote this point. Since the $p(\mathbf{x}_i)$'s cancel, the conditional likelihood in (14) becomes

$$\exp(L) = \frac{\prod_{i=1}^{N}[\exp\{g(\mathbf{x}_i^T\boldsymbol{\beta})\}]^{d_i}}{\sum_{i_1,\ldots,i_n}\prod_{j=1}^{n}\exp\{g(\mathbf{x}_{i_j}^T\boldsymbol{\beta})\}} = \exp\Big\{\sum_{i=1}^{N} d_i g(\mathbf{x}_i^T\boldsymbol{\beta}) - \log R(n,S)\Big\}, \quad (15)$$

where the right hand side is obtained by letting $w_i = \exp\{g(\mathbf{x}_i^T\boldsymbol{\beta})\}$ for all $i = 1, \ldots, N$ and $R(n,S) = \sum_{i_1,\ldots,i_n}(\prod_{j=1}^{n} w_{i_j})$, the same as in (6). (See Breslow and Day (1980) and Cox (1972) for related materials.) Let $\mathbf{D} = (D_1, \ldots, D_N)$ be a conditional Bernoulli vector whose distribution is the same as that of $\mathbf{Z}$ conditional on $S_Z = n$. We see that the likelihood in (15) is actually the probability of observing $\mathbf{D} = \mathbf{d}$.

Let $\partial L/\partial\boldsymbol{\beta}$ and $\partial^2 L/\partial\boldsymbol{\beta}^2$ denote the score function and the Hessian matrix respectively. Then

$$\frac{\partial L}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{N} g'(\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i d_i - \sum_{i=1}^{N} g'(\mathbf{x}_i^T\boldsymbol{\beta})\mathbf{x}_i E(D_i) = \mathbf{X}^T\mathbf{G}'\{\mathbf{d} - E(\mathbf{D})\},$$

$$\frac{\partial^2 L}{\partial\boldsymbol{\beta}^2} = -\mathbf{X}^T\mathbf{G}'\mathrm{Cov}(\mathbf{D}, \mathbf{D}^T)\mathbf{G}'^T\mathbf{X},$$

where $\mathbf{G}' = \mathrm{diag}\{g'(\mathbf{x}_1^T\boldsymbol{\beta}), \ldots, g'(\mathbf{x}_N^T\boldsymbol{\beta})\}$ is an $N \times N$ matrix. Therefore, finding $\boldsymbol{\beta}$ to maximize L is equivalent to finding $\boldsymbol{\beta}$ so that

$$\mathbf{X}^T\mathbf{G}'\mathbf{d} = \mathbf{X}^T\mathbf{G}'E(\mathbf{D}). \quad (16)$$

The Hessian matrix is just the covariance matrix $\mathrm{Cov}\{\mathbf{X}^T\mathbf{G}'\mathbf{D}, (\mathbf{X}^T\mathbf{G}'\mathbf{D})^T\}$. On the other hand, the MLE of $\boldsymbol{\beta}$ in the corresponding prospective model is given as the solution of $\mathbf{X}^T\mathbf{G}'\mathbf{z} = \mathbf{X}^T\mathbf{G}'E(\mathbf{Z})$, with the Hessian matrix $\mathrm{Cov}\{\mathbf{X}^T\mathbf{G}'\mathbf{Z}, (\mathbf{X}^T\mathbf{G}'\mathbf{Z})^T\}$. Therefore, the only difference between a retrospective model and its corresponding prospective model is that between $\mathbf{D}$ and $\mathbf{Z}$. Based on this observation, we provide an iterative weighted least squares method for finding the MCLE similar to that for finding the MLE in a prospective GLM.

Gail, Lubin and Rubinstein (1981) describe a quasi Newton-Raphson method to solve (16) for the case of a logistic regression model, i.e., $g(x) = x$, which requires numerical approximation to $\partial L/\partial\boldsymbol{\beta}$ and $\partial^2 L/\partial\boldsymbol{\beta}^2$. Howard (1972), however, uses recursive formulas to directly compute $\partial L/\partial\boldsymbol{\beta}$ and $\partial^2 L/\partial\boldsymbol{\beta}^2$. Compared with their methods, ours tends to be numerically more stable.

Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ be the mean vector of $\mathbf{D}$, and let W be its covariance matrix, which is of rank $N-1$ and is orthogonal to the constant vector. The Newton-Raphson's method in the form of iterative weighted least squares is

$$\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(0)} + \{\mathbf{X}^T(\mathbf{G}'^{(0)}W^{(0)}\mathbf{G}'^{(0)})\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{G}'^{(0)}(\mathbf{d} - \boldsymbol{\pi}^{(0)}), \quad (17)$$

where every unknown quantity on the right hand side is computed at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(0)}$.

Chen (1993) noted that the off-diagonal elements of W, i.e. $E(D_iD_j) - \pi_i\pi_j$, are usually much smaller than the diagonal elements $\pi_i - \pi_i^2$, especially when the number of cases n is small compared to the population size N or when the weights w_i have a narrow range. Thus, we can use the matrix $V = \text{diag}\{\pi_i - \pi_i^2\}$ in place of W to avoid the computation of off-diagonal elements. The new algorithm then becomes

$$\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(0)} + \{\mathbf{X}^T(\mathbf{G}'^{(0)}V^{(0)}\mathbf{G}'^{(0)})\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{G}'^{(0)}(\mathbf{d} - \boldsymbol{\pi}^{(0)}). \tag{18}$$

This substitution, as was pointed out by the Associate Editor, also helps make the algorithm more stable. It is easy to see that if the procedure converges, i.e. $\hat{\boldsymbol{\beta}}^{(1)} = \hat{\boldsymbol{\beta}}^{(0)}$, we must have $\mathbf{X}^T\mathbf{G}'\mathbf{d} = \mathbf{X}^T\mathbf{G}'\boldsymbol{\pi}$. Hastie and Tibshirani (1990) in Chapter 8 use a similar trick to deal with the Hessian matrix in fitting a generalized additive model for matched case-control data. The method illustrated here can be used to fit generalized additive models to more general retrospective studies.

To illustrate the use of our procedure, we consider a study of the effect of the drug sulphinpyrazone on cardiac death after myocardial infarction (Anturane Reinfarction Trial Research Group, 1978, 1980), which is cited in the first edition of McCullagh and Nelder (1989). The data are given in Table 2.

Table 2. A study of the effect of the drug suplhinpyrazone

X	cases (deaths from all causes)	control (survivors)	Total
1 (sulphinpyrazone)	41	692	733
0 (placebo)	60	682	742
Total	101	1374	1475

In this example, the population size $N = 1,475$ and the sample size $n = 101$. There is only one covariate, so the dimension of $\boldsymbol{\beta}$ is 1. As we mentioned earlier, $g(x)$ equals x in a logistic regression model, and therefore the corresponding $\mathbf{G}'$ is just the identity matrix. Our Newton-Raphson algorithm for this model reduces to $\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + (\mathbf{X}^TV^{(0)}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{d} - \boldsymbol{\pi}^{(0)})$.

We set a precision bound of six significant figures. Starting from $\hat{\beta}^{(0)} = 0$, the procedure converges to $\hat{\beta} = -0.395063$ with a standard error of 0.160734 after 17 iterations, which took 1.5 seconds on a Sun SparcStation 2 with a 40Mhz Weitek CPU chip.

A good choice for the starting point is the estimate of β based on the corresponding prospective linear model. Using GLIM to fit a prospective logistic regression model on the data, we obtain an estimate -0.395329 for β with a standard error of 0.209686. Starting from $\hat{\beta}^{(0)} = -0.395329$, the procedure converges to -0.395063 after 8 iterations, which took 0.8 second on the same machine.

In summary, three special "treatments" to the usual Newton-Raphson method (17) make the computation fast and more stable: (1) only the diagonal elements of W are used (note that in this example W is a $1,475 \times 1,475$ matrix and has in total 2,174,150 off-diagonal elements); (2) a reversed version of (11) in Section 2 is used for the computation of the π_i; (3) the trick for grouped populations as described in Example 3 of Section 2 is used to compute all R functions. Note that there are only two different weights in this example. It is easy to show that the grouping method requires about 15,000 operations for computing each R, whereas Method 2 in Section 2 (Gail, Lubin and Rubinstein (1981)) requires about 275,000 operations for the same calculation.

In the corresponding prospective GLM, the MLE for $\boldsymbol{\beta}$ imposes the same form of estimation equation as (16) except that $\mathbf{D}$ is substituted by the unconditional Bernoulli vector $\mathbf{Z}$. Thus the usual iterative least squares method using (17) has $W = \text{diag}\{p_i - p_i^2\}$, the covariance matrix for the unconditional Bernoulli random vector $\mathbf{Z}$ instead of the conditional one. Because of this similarity, we expect that numerical stability of our algorithm is comparable to that of a standard GLIM algorithm. In fact, in all the examples we have tested, our algorithm seems to be little affected by the starting point in terms of both numerical stability and the number of iterations. For example, when the starting point $\hat{\beta}^{(0)} = 4$ (a number with a different sign and magnitude from the true value), the procedure converges with the same precision in just 24 iterations within 2 seconds.

It is well-known that in logistic regression, treating retrospective data prospectively can produce valid inference for the regression coefficients, except for the interception term (McCullagh and Nelder (1989), pp 111-113). However, the estimate based on the conditional likelihood (14) often has smaller standard error than that based on the unconditional likelihood, as we have seen in the above example and in Lubin (1981). Moreover, GLMs other than the logistic model can be affected to a great extent by the choice between the prospective model and the retrospective model. Treating retrospective data as if they were drawn by prospective sampling would in general produce inferior estimates. Hence it is usually desirable to deal with the conditional likelihood (15) directly. In such cases, the iterative method in (17) is applicable.

4. Five Sample Selection Procedures

Now we discuss procedures for drawing a sample from the conditional Bernoulli model. Before we present the five sampling procedures, let us first consider a case study in order to understand the idea and the different use of each sampling procedure in response to various actual situations.

Case Study. Company XYZ is recruiting for a Research Associate position. By the time the company starts processing the applications, 1000 people have applied. The company wishes to interview as many candidates as possible but its budget is only enough for 10 interviews. To measure the qualification of the candidates, the company gives each candidate a rating based on his/her application materials (e.g. GPA, work experience, recommendation letters, etc.). Since there is a concern that application materials may not fully reflect the true "quality" of candidates, the company decides to select candidates randomly according to their ratings instead of taking the candidates with highest ratings. Suppose the candidates are "unconditionally independent" (i.e. they are independent without the constraint on the number of interviewees). Then a natural choice for the distribution of random sampling is a conditional Bernoulli model with the weights being ratings of the candidates.

With properly defined probabilities, the sampling procedure for the above situation can be carried out in any of the following ways:
(1) repeatedly select one candidate at a time from the unselected candidates until 10 people are obtained (Procedure 1, Drafting Sampling).
(2) check the candidates one at a time in any arbitrary order (e.g. by arrival time of the application) and decide for each individual whether or not he/she is selected with certain probability (Procedure 3, ID-Checking Sampling).
(3) line up all possible combinations of 10 people out of a pool of 1000 and decide which combination to be selected using only one random drawing (Procedure 5, Direct Sampling).

Case Study. (*continued*) After the company has selected 10 interviewees using any of the three procedures above, one or both of the following situations can occur:
Situation 1. Budget is increased to allow for more interviews.

In this case, it is not worthwhile to put the 10 already selected people back into the pool and select 11 (suppose one more interview is the case) out of 1000 — mainly because it will take a long time to get a sample of 11 which contains the 10 people previously selected. An easier way is to select one candidate out of the unselected 990 people with certain probability so that the joint probability of these 11 people still follows a conditional Bernoulli model (Procedure 2, Open-Market Sampling).
Situation 2. Some new applications come in.

Again, it is not worthwhile to put the 10 people already selected back into the pool of 1001 (suppose one new application is the case) and do it all over again. An easier way is to replace one of the 10 already selected people by the new candidate with certain probability, or otherwise keep the current sample (Procedure 4, Open-Pool Sampling).

We refer to the two situations above as having *non-fixed* n and *non-fixed* N, respectively. The two procedures above can also be used in the usual fixed n and N situations. For Situation 1, one can start off the procedure by pretending that there is only one interview allowed in the beginning and subsequently add one interview at a time until 10 interviews are filled. For Situation 2, one can pretend that in the beginning there are only 10 candidates in the pool and all of them are taken for interview. Then subsequently, other candidates are added one at a time and decision is made for each individual candidate as whether or not he/she should replace someone that is already in the sample.

Procedures 1 and 2 are given by Chen, Dempster and Liu (1994). Procedures 3 and 4 are natural generalizations of the "sequential selection sampling" and the "reservoir sampling" of Fan, Muller and Rezucha (1962), respectively. All five procedures are natural generalizations of the simple random sampling case as documented in Knuth (1968), Section 3.4.

We now present all five sampling procedures, of which Procedures 1 and 2 are given without any proof. Details about these two procedures can be found in Chen, Dempster and Liu (1994). As for the notation used in the procedures, we let A_k represent the set of indices of the selected units after step k and $S_k = \{1, \ldots, k\}$ for $k = 0, \ldots, N$ with $S_0 = \emptyset$. For example, the sizes of A_k in the first four procedures are k, k, $r \leq k$, and n, respectively.

Procedure 1. (*Drafting Sampling*) Start with $A_0 = \emptyset$. At step k ($k = 1, \ldots, n$), a unit $j \in A^c_{k-1}$ is selected into the sample (i.e. $A_k \leftarrow A_{k-1} \cup \{j\}$) with probability

$$P_1(j, A^c_{k-1}) := \frac{w_j R(n-k, A^c_{k-1} \backslash \{j\})}{(n-k+1) R(n-k+1, A^c_{k-1})}.$$

Procedure 2. (*Open-Market Sampling*) Start with $A_0 = \emptyset$. At step k ($k = 1, \ldots, n$), a unit $j \in A^c_{k-1}$ is selected into the sample (i.e. $A_k \leftarrow A_{k-1} \cup \{j\}$) with probability

$$P_2(j, A^c_{k-1}) := \sum_{i=0}^{k-1} \frac{w_j R(k-i-1, A^c_{k-1} \backslash \{j\}) R(i, A_{k-1})}{(k-i) R(k, S)}.$$

Procedure 3. (*ID-Checking Sampling*) Start with $A_0 = \emptyset$ and the first unit. By step k ($k = 1, \ldots, N$), suppose r out of the first $k-1$ units have been selected (i.e. $|A_{k-1}| = r$). Then the kth unit is selected into the sample (i.e. $A_k \leftarrow A_{k-1} \cup \{k\}$) with probability

$$P_3(k, r) := \frac{w_k R(n-r-1, S^c_k)}{R(n-r, S^c_{k-1})},$$

and is excluded from the sample (i.e. $A_k \leftarrow A_{k-1}$) with probability $1 - P_3(k, r)$.

Using the relation between Poisson-Binomial probability and R functions as in (7), P_3 can be interpreted as $P_3(k, r) = P(Z_k = 1|\sum_{i=k}^N Z_i = n - r)$. It is easy to see that Procedure 3 produces a sample from the conditional Bernoulli distribution by noting that for any $\mathbf{d} \in \mathcal{D}^n$,

$$\begin{aligned} P(\mathbf{D} = \mathbf{d}) &= \prod_{k=1}^{N} [P_3(k, \sigma_k)]^{d_k} [1 - P_3(k, \sigma_k)]^{1-d_k} \\ &= \prod_{k=1}^{N} \frac{w_k^{d_k} R(n - \sigma_k, S_k^c)}{R(n - \sigma_{k-1}, S_{k-1}^c)} = \prod_{k=1}^{N} w_k^{d_k} \Big/ R(n, S), \end{aligned}$$

where $\sigma_k = \sum_{j=1}^k y_j$ for $k = 1, \ldots, N$ and $\sigma_0 = 0$.

Another way to understand procedure 3 is through the "telescope law" in elementary probability theory. That is, the joint probability distribution of the conditional Bernoulli vector $\mathbf{D} = (D_1, \ldots, D_N)$ can be decomposed as, $P(D_1, \ldots, D_N) = P(D_1)P(D_2|D_1) \cdots P(D_N|D_1, \ldots, D_{N-1})$, which provides us the above sampling scheme. This idea can also be applied to sequentially classify N objects into k groups at random with sizes $n_1, \ldots, n_k$.

Procedure 4. (*Open-Pool Sampling*) Start with $A_n = \{1, \ldots, n\}$. Then at step k ($k = n + 1, \ldots, N$), a random number U is drawn uniformly from $[0, 1)$. If

$$U \geq \frac{w_k R(n-1, S_{k-1})}{R(n, S_k)} = \pi_{k,k},$$

then the old sample is kept (i.e. $A_k \leftarrow A_{k-1}$); Otherwise a unit $j \in A_{k-1}$ is chosen with probability

$$P_4(j, A_{k-1}) := \sum_{i=0}^{n-1} \frac{R(n-i-1, A_{k-1}^c \backslash \{j\}) R(i, A_{k-1})}{(n-i) R(n-1, S_{k-1})}$$

and is replaced by unit k (i.e. $A_k \leftarrow A_{k-1} \cup \{k\} \backslash \{j\}$).

The function $P_4(\cdot, A_{k-1})$ is in fact the selection probability used in the "backward" version of Procedure 2 and is shown in Chen, Dempster and Liu (1994) to be a probability density on A_{k-1}. The function $\pi_{k,k}$ is in fact the inclusion probability of the kth unit in a sample of size n when the population consists of only the first k units, i.e. $P(Z_k = 1|\sum_{i=1}^k Z_i = n)$.

Now we show by induction that a random sample selected by Procedure 4 is a sample from the conditional Bernoulli model. Let γ_k be the index of the unit chosen from A_{k-1} and $S_k = \{1, \ldots, k\}$. Assume $P(A_{k-1} = A) = R(n, A)/R(n, S_{k-1})$ for any $A \subset S_{k-1}$ with $|A| = n$ (which is true for $k = n+1$), we show that $P(A_k = B) = R(n, B)/R(n, S_k)$ for any $B \subset S_k$ with $|B| = n$. We first consider the case $k \in B$. By $a)$ and $c)$ of Proposition 1,

$$\begin{aligned} P(A_k = B) &= \sum_{j \in B^c} P[A_{k-1} = B \cup \{j\} \backslash \{k\}]\ P[\gamma_k = j \mid A_{k-1} = B \cup \{j\} \backslash \{k\}] \pi_{k,k} \\ &= \sum_{j \in B^c} \frac{R(n, B \cup \{j\} \backslash \{k\})}{R(n, S_{k-1})} \Big[\sum_{i=0}^{n-1} \frac{R(n-i-1, B^c \backslash \{j\}) R(i, B \backslash \{k\})}{(n-i) R(n-1, S_{k-1})} \Big] \\ &\quad \frac{w_k R(n-1, S_{k-1})}{R(n, S_k)} \\ &= \frac{R(n, B)}{R(n, S_k)} \sum_{i=0}^{n-1} \frac{R(i, B \backslash \{k\})}{(n-i) R(n, S_{k-1})} \Big[\sum_{j \in B^c} w_j R(n-i-1, B^c \backslash \{j\}) \Big] \\ &= \frac{R(n, B)}{R(n, S_k)} \sum_{i=0}^{n-1} \frac{R(i, B \backslash \{k\})}{(n-i) R(n, S_{k-1})} (n-i) R(n-i, B^c) \ = \ \frac{R(n, B)}{R(n, S_k)}. \end{aligned}$$

For the case $k \notin B$, we have $A_k = A_{k-1}$. Thus

$$\begin{aligned} P(A_k = B) &= \sum_{j \in B} P[A_{k-1} = B]\ P[\gamma_k = j \mid A_{k-1} = B](1 - \pi_{k,k}) \\ &= \frac{R(n, B)}{R(n, S_{k-1})} \Big[\sum_{j \in B} P_4(j, B) \Big] \frac{R(n, S_{k-1})}{R(n, S_k)} = \frac{R(n, B)}{R(n, S_k)}. \end{aligned}$$

Procedure 5. (*Direct Sampling*) Draw a random number U uniformly from $[0, R(n, S))$. Suppose by step k ($k = 1, \ldots, N$), r out of the last $k-1$ units have been selected. Then if $U \geq R(n-r, S_{N-k})$, the $N-k+1$st unit is selected and let $U \leftarrow [U - R(n-r, S_{N-k})]/w_{N-k+1}$; otherwise, the $N-k+1$st unit is not selected and the value of U does not change.

The use of Procedure 5 is in fact the *inverse* procedure for generating R. This is illustrated in Table 3, which is the same as Table 1 except that the movement is going backward from $cell(n, N)$ to $cell(0, 0)$. A sample drawn by Procedure 5 is represented by a path connecting $cell(n, N)$ and $cell(0, 0)$. Each path has exactly N edges of which n edges are diagonal and marked as "in" (included in the sample), and the other $N-n$ edges are horizontal and marked as "out" (excluded from the sample). Note that $cell(i, j)$ in the table gives the quantity $R(i, S_j)$. Thus it is easy to see from the definition of Procedure 5 that drawing a sample using Procedure 5 is equivalent to selecting a path according to the following rule: suppose we are currently in $cell(i, j)$, and if the current $U \geq cell(i, j-1)$, we then move to $cell(i-1, j-1)$ taking the diagonal edge; otherwise, we move to $cell(i, j-1)$ taking the horizontal edge.

Table 3. Direct sampling from the conditional Bernoulli model for $\overline{n} = 2$, $N = 4$

$$\begin{tikzcd}
 & & & & N & \\
 & 0 & 1 & 2 & 3 & 4 \\
0 & 1 & 1 \arrow[l, "\text{1 out}"'] & 1 \arrow[l, "\text{2 out}"'] & \text{—} & \text{—} \\
n \quad 1 & 0 & w_1 \arrow[ul, "\text{1 in}"] & w_1 + w_2 \arrow[l, "\text{2 out}"'] \arrow[ul, "\text{2 in}"] & w_1 + w_2 + w_3 \arrow[l, "\text{3 out}"'] \arrow[ul, "\text{3 in}"] & \text{—} \\
2 & 0 & 0 & w_1w_2 \arrow[ul, "\text{2 in}"] & w_1w_2 + w_1w_3 + w_2w_3 \arrow[l, "\text{3 out}"'] \arrow[ul, "\text{3 in}"] & w_1w_2 + w_1w_3 + w_1w_4 + w_2w_3 + w_2w_4 + w_3w_4 \arrow[l, "\text{4 out}"'] \arrow[ul, "\text{4 in}"]
\end{tikzcd}$$

We now compare the advantages and disadvantages of using the five procedures described above. A summary is given in Table 4, where three criteria are considered – number of operations, number of random draws and extra storage space needed in addition to N slots for the weights w_i. In terms of the computational cost (certain combination of the "number of operations" and the "number of random draws" depending on the type of the computer used), Procedure 2 (Open-Market) and Procedure 4 (Open-Pool) are the most expensive ones and the other three are about the same. In terms of extra storage space, Procedure 5 (Direct) requires considerably more space than the others.

Table 4. Comparisons of the five sampling procedures

	Operations	Random Draws	Extra Space
Drafting	$O(nN)$	n	$N + n$
Open-market	$O(n^2N)$	n	$2n$
ID-checking	$O(nN)$	$\in [n, N]$	$2n$
Open-pool	$O(n^2N)$	$2(N - n)$	$4n$
Direct	$O(nN)$	1	$n(N - n + 1)$

The relatively low computational efficiency of Procedures 2 and 4 is compensated by their special uses. Specifically, Procedures 2 and 4 can be used in the situations where the sample size n or the population size N is not known in advance, respectively. We can also combine them to deal with the situations where both N and n are unfixed. A classification of the five procedures is given in Table 5.

Table 5. Classification of the five sampling procedures

		N	
		fixed	non-fixed
n	fixed	Procedures $1 \sim 5$	Procedure 4
	non-fixed	Procedure 2	Procedure 2 + 4

5. Conclusions

In this article, we are able to connect several seemingly unrelated statistical problems by the Poisson-Binomial and conditional Bernoulli distributions. By doing this, we gain more understanding of and insights into these problems, both conceptually and computationally.

We have compared different methods for computing the exact distribution functions of Poisson-Binomial and conditional Bernoulli models and analyzed their computational complexities. As by-products, we found five efficient ways of sampling from the conditional Bernoulli distribution. These methods can be viewed as direct generalizations of methods for conducting simple random sampling, and can be useful to both applied statisticians and computer scientists (Knuth (1968)).

Sugden and Smith (1984) advocate the use of the Hájek model for PPS sampling in a survey. Although the Hájek model is essentially a conditional Bernoulli model, it can not guarantee that the marginal inclusion probability of each sample unit equals to its pre-specified value, violating a condition required by all PPS sampling. To correct it, an inversion scheme as illustrated in (11) needs to be employed in order to find a set of proper p_i's that give rise to the pre-specified marginal probability π_i. The method illustrated in this article is thus crucial for conducting a *proper* PPS sampling.

Lastly, we have found it conceptually more transparent and rewarding to think of the conditional likelihood of a retrospective GLM as a conditional Bernoulli probability distribution. This view helps us easily identify the connection and difference between a prospective and a retrospective GLM, and to design an iterative weighted least squares method for maximizing the conditional likelihood, similar to the one used in analyzing a prospective GLM. Because of this computational advance and the reasons discussed at the end of Section 3, we strongly advocate the use of MCLE for making inference in case-control studies.

Acknowledgements

S. X. Chen's research was partially supported by ARO Grant DAAL-0391-0089 and NSF Grant DMS-94-04396. J.S. Liu's research was partially supported by NSF Grants DMS-94-06044 and DMS-95-01570, and by the Terman Fellowship. The authors thank Professors I. M. Johnstone, D. B. Rubin, R. J. Tibshirani, W. H. Wong, the Associate Editor and three referees for insightful suggestions.

References

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, Volume 1 - The Analysis of Case-Control Studies.* IARC Scientific Publications 32, Lyon, France: IARC.

Chen, X. (1993). Poisson-Binomial distribution, conditional Bernoulli distribution and maximum entropy. Technical Report. Department of Statistics, Harvard University.

Chen, X. H., Dempster, A. P. and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457-469.

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

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Diaconis, P. and Sturmfels, B. (1993). Algebraic algorithms for sampling from conditional distributions. Technical Report, Department of Mathematics, Harvard University.

Fan, C. T., Muller, M. E. and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *J. Amer. Statist. Assoc.* **57**, 387-402.

Gail, M. H., Lubin, J. H. and Rubinstein, L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* **68**, 703-707.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hirji, K. F., Mehta, C. R. and Patel, N. R. (1987). Computing distributions for exact logistic regression. *J. Amer. Statist. Assoc.* **82**, 1110-1117.

Howard, S. (1972). Discussion on Professor Cox's paper. *J. Roy. Statist. Soc. Ser. B* **34**, 210-211.

Knuth, D. E. (1968). *The Art of Computer Programming*, vol. II. Reading, MA: Addison-Wesley.

Kolassa, J. E. and Tanner, M. A. (1994). Approximate conditional inference in exponential families via the Gibbs sampler. *J. Amer. Statist. Assoc.* **89**, 697-702.

Lubin, J. H. (1981). An empirical evaluation of the use of conditional and unconditional likelihoods for case-control data. *Biometrika* **68**, 567-571.

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

Stein, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson-Binomial distribution. Technical Report 354, Department of Statistics, Stanford University.

Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495-506.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc. Ser. B* **15**, 253-261.

Stern School of Business, New York University, New Youk, NY 10012, U.S.A.
E-mail: schen3@stern.nyu.edu

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
E-mail: jliu@playfair.stanford.edu

(Received October 1995; accepted November 1996)