Classical Multi-level and Bayesian Approaches to Population Size Estimation Using Multiple Lists*

Stephen E. Fienberg, Matthew S. Johnson, and Brian W. Junker
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890, U.S.A.

February 8, 1999

Abstract

One of the major objections to the standard multiple-recapture approach to population estimation is the assumption of homogeneity of individual "capture" probabilities. Modeling individual capture heterogeneity is complicated by the fact that it shows up as as a restricted form of interaction between lists in the contingency table cross-classifying list memberships for all individuals. Traditional log-linear modeling approaches to capture-recapture problems are well-suited to modeling interactions among lists, but ignore the special dependence structure that individual heterogeneity induces. A random-effects approach, based on the Rasch (1960) model from educational testing and introduced in this context by Darroch, et al. (1993) and Agresti (1994), provides one way to introduce the dependence resulting from heterogeneity into the log-linear model; however, previous efforts to combine the Rasch-like heterogeneity terms additively with the usual log-linear interaction terms suggest that a more flexible approach is required. In this paper we consider both classical multi-level approaches and fully Bayesian hierarchical approaches to modeling individual heterogeneity and list interactions. Our framework encompasses both the traditional log-linear approach and various elements from the full Rasch model. We compare these approaches on two examples, the first arising out of an epidemiological study of a population of diabetics in Italy, and the second a study intended to assess the "size" of the World Wide Web. We also explore extensions allowing for interactions between the Rasch and log-linear portions of the models in both the classical and Bayesian contexts.

Keywords: Log-linear models; Markov chain Monte Carlo methods; Multiple-recapture census; Quasi-symmetry; Rasch model.

*Supported in part by Grants REC-9720374 and DMS-9705032 from the National Science Foundation to Carnegie Mellon University. We are indebted to Steve Lawrence and Lee Giles for providing us with unpublished data for analysis. Alan Agresti, S. Basu, Annibale Biggeri, Richard Cormack, Jon Forster, Ed George, David R. Jones, and Steve Lawrence all provided helpful input of various sorts as we were working on different parts of the paper.

Contents

1 Introduction

Our goal in this paper is to re-examine the problem of estimating the size of a closed population using multiple lists or sources, often referred to as the multiple-recapture population estimation problem (e.g., see Bishop, et al., 1975) because of its origins for estimating wildlife and fish populations (e.g., see Petersen, 1896; Schnabel, 1938). In effect, we treat our lists as having been generated by sampling multiple times from the population and we identify individuals or objects according to the lists in which they were included.

We wish to estimate N, the unknown size of the population of individuals or objects of interest (e.g., people, fish, software errors, etc.), and we do so using the information gleaned from which objects were included in each of the J lists drawn from the population. We let $i = 1, \ldots, N$ index the objects, and $j = 1, \ldots, J$ index the lists. Our basic model has $N \times J$ random variables, X_{ij}, such that

$$X_{ij} = \begin{cases} 1, \text{ if object } i \text{ appears on list } j; \\ 0, \text{ otherwise.} \end{cases}$$

We let $p_{ij} = P[X_{ij} = 1]$, and n be the number of objects that appear on at least one list. Our goal is to estimate the number of unobserved objects, $N - n$ or, equivalently, to estimate N. To do this, we need a model which specifies:

- The probabilities of appearing in the various lists, i.e., capture probabilities;
- How the lists relate to one another, i.e., list dependencies; and
- The ways in which these capture probabilities and list dependencies vary across individuals.

The literature on capture-recapture methods is extensive and goes back many years to at least Petersen (1896). The earliest models for multiple-recapture methods (i.e., more than 2 lists) assumed that the various captures or lists were independent (e.g., Geiger and Werner, 1924; Schnabel, 1938) and that there were constant capture probabilities across individuals, although not necessarily across captures or lists. Although many authors expressed concern about the assumption of independence among lists, a general way to cope with this problem awaited other developments in statistics. Thus it was not until the 1970's that Fienberg (1972) introduced the role of log-linear models to provide for dependencies among the lists, and Sanathanan (1972) introduced the Rasch model to provide for the dependence induced by heterogeneity across individuals, but for independent lists. Fienberg (1992) and the International Working Group on Disease Monitoring and Forecasting (1995b) provide bibliographies of special relevance to the use of these methods in human populations.

The simplest model that allows for differences in capture probabilities and heterogeneity among individuals is due to Georg Rasch (1960), who derived it for scoring examination items in educational testing:

$$\log\left\{\frac{p_{ij}}{1 - p_{ij}}\right\} = \theta_i + \beta_j, \quad i = 1, \ldots, N; \quad j = 1, \ldots, J. \tag{1}$$

Note that when we set the θ_i in equation (1) equal to zero the log-odds of inclusion of object i on list j depends only on the list, and thus the model reduces to the traditional multiple-recapture model with independent lists. When the $\theta_i \neq 0$ and we treat them as random effects, this model is intrinsically multi-level, with lists at one level and individuals at another. Additional multi-level structure may be readily incorporated into this model, through either θ or β, depending on relevant object and list covariates. For example, see Johnson, et al. (1998) for a Bayesian version of this extension, and Wu, et al. (1997) for a classical/missing data formulation. The Rasch model and some of its natural generalizations play central roles in our work.

In this paper, we attempt to draw on the lessons from what was, until recently, three seemingly separate literatures on (1) log-linear models for multiple-recapture census problems, (2) Rasch models for individual heterogeneity, and (3) Bayesian hierarchical model approaches. For us these three approaches are intimately linked and this paper explores their relationships. In the remainder of this section we review aspects of the literature on heterogeneity, and Bayesian approaches. In Section 2, we introduce three examples to which we later apply our methodology: data simulated directly from the Rasch model; data from an epidemiological study of a population of diabetics in Italy; and data from six "web search engines" intended to assess the "size" of the World Wide Web. Then, in Section 3, we outline the elements of the Rasch model and its relationship with the usual log-linear models for population size estimation, and in Section 4 we present a fully-Bayesian hierarchical approach to the Rasch model which relaxes a seemingly necessary linear constraint in the log-linear formulation and takes into account previously ignored moment-inequality constraints. In Section 5, we apply both log-linear models with Rasch-like heterogeneity terms and our Bayesian hierarchical approach to the examples. We will see that these approaches, applied separately, work reasonably well but seem lacking: the dependence structure in multiple-recapture census data is often similar to the Rasch model, with departures that reflect non-symmetric dependence between lists, or partial symmetry features that represent "clumpy" heterogeneity of the objects being counted. "Generalized Rasch" models that allow interactions between parameters that express heterogeneous catchability of objects, or non-symmetric dependencies between lists offer some hope of a more parsimonious representation, and hence smaller standard errors of estimation for the unobserved count. In Section 6, we explore the relationship between generalized Rasch log-linear models and our hierarchical Bayes formulation of the Rasch model.

1.1 Heterogeneity Among Individuals

As we mentioned above, Sanathanan (1972, 1973) provided one of the early attempts to look at heterogeneity in the context of capture probabilities. She was interested in scanning experiments in particle physics and focused upon a Rasch model in which either the individual or the list parameters are viewed as independent draws from a parametrically-specified common distribution. Subsequently, Burnham and Overton (1978) described a similar model in which the source of variation in their capture probabilities comes solely from the heterogeneity among individuals, i.e., with capture probabilities $p_{ij} = p_i$. They took $p_1, \ldots, p_N$ to be a random sample from a probability distribution F and assumed that the random variables X_{ij} $(i = 1, \ldots, N; j = 1, \ldots, J)$ are mutually independent for given $(p_1, \ldots, p_N)$; this leads to a generalized Beta-binomial structure. To estimate the unknown population size N, they empirically estimated the probability distribution F and used a generalized jackknife approach in which the number of observed objects n serves as a naive estimate of N. Later efforts, especially in the wildlife literature, e.g., Chao (1987, 1989), Chao, et al. (1992), and Pollock (1991), built on this approach incorporating heterogeneity into multiplicative models for the p_{ij}, e.g., $p_{ij} = \phi_i \psi_j$. Unfortunately, this literature provides little recognition of the special statistical features that require attention when the number of parameters included for heterogeneity increases in direct proportion to the population size N.

Interest in the heterogeneity problem arose again in connection with discussions about the use of capture-recapture methods in the context of the 1990 U.S. decennial census, and Darroch, et al. (1993) presented a model for heterogeneity based on a log-linear representation of the Rasch model, which they then combined with log-linear models for dependence. Their approach had been anticipated in part by Cormack (1989) but without the link to the Rasch model framework, and then suggested separately by Agresti (1994). The International Working Group on Disease Monitoring and Forecasting (1995a, b), provided a simple discussion of these approaches and their application to a problem of estimating the size of a population of diabetics. The introduction of the Rasch model representation for heterogeneity in this example, however, had apparently strange and not totally satisfactory consequences, and we return to their problem in Section 2 below.

1.2 Bayesian Approaches

Roberts (1967) presented the earliest known Bayesian approach to the simple capture-recapture problem, i.e., with two lists, and with constant capture probabilities. Freedman (1972) introduced a Bayesian approach to sequential estimation, and then later (Freedman, 1973) contrasted this to the capture-recapture model, using constant capture probabilities in both settings. He also introduced the role of different loss functions. Castledine (1981) generalized Robert's and Freedman's approach to multiple-recapture studies and derived the marginal posterior distribution of N by assuming that prior probability distribution on N and the probabilities $\mathbf{p} = \{p_{ij}\}$ is of the form $\pi(N, \mathbf{p}) = \pi(N)\pi(\mathbf{p})$, where $p_{ij} = p_j \overset{\text{iid}}{\sim} \text{Beta}(a, b)$ and $\pi(N) \propto 1$ and $\pi(N) \propto N^{-1}$, which is the Jeffreys prior. Smith (1991) found the posterior distribution of N in this case using both empirical-Bayes, and Bayes/empirical-Bayes approaches. Garthwaite, et al. (1995) extended these results to allow for random sample sizes, and explored the sensitivity of the posterior distribution for N to the prior specification.

Smith (1988) also estimated the unknown population size N by using a Poisson approximation to the hypergeometric distribution of marked items in the sample, and a gamma density on $\omega = N^{-1}$. Smith found the formal Bayes rule for (a) 0-1, (b) quadratic, (c) chi-squared, and (d) squared proportional losses. He then showed that under certain conditions all of the estimates but the one from the 0-1 loss approach are equivalent to the Geiger-Werner-Schnabel multiple-recapture estimate.

George and Robert (1992) were the first to bring the modern Bayesian technology of Markov chain Monte Carlo estimation to bear on the capture-recapture problem. They began with the Casteldine formulation in which N and $p_1, \ldots, p_J$ are *a priori* independent, and suggested the use of either a Poisson prior distribution or Jeffreys prior distribution $\pi(N) \propto N^{-1}$ on N, and either a $\text{Beta}(a, b)$ for each p_j or a logit model for p_j, that is $\text{logit}(p_j) \sim \text{N}(\mu_j, \sigma^2)$ and beta or normal log-odds priors for the p_i's. To simulate from the conditional posterior distribution for N they used the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992). They also discussed hierarchical extensions to the estimation procedure, e.g., using hyperpriors on a and b when a beta prior is used on $\mathbf{p}$, and on μ_j and σ when the logit model is used for $\mathbf{p}$, and suggested an estimation approach that utilized Gibbs sampling.

Basu (1998) considers log additive mixed effects models for the p_{ij} with random effects for the the objects whose population size is of interest, and a fixed effect for each list, i.e., $\log(p_{ij}) = \theta_i + \beta_j$, or $p_{ij} = e_i r_j$ (c.f., Pollock, 1991). Basu gives both the catchability and catch efforts discrete prior distributions, i.e., $e_i \sim F_e$ where $F_e = \sum_{v=1}^{V} \alpha_v I_{\{\varepsilon_v\}}(e)$, with $0 \leq \alpha_v \leq 1$, $\sum_{v=1}^{V} \alpha_v = 1$, $0 \leq \varepsilon_1 \leq \ldots \leq \varepsilon_V \leq 1$, where I is the indicator function, and a similar prior for r_j. With this setup Basu finds complete conditional distributions for each parameter and hyperparameter, and implements a Gibbs sampling scheme using two-point prior distributions on the e_i and r_j parameters.

Madigan and York (1995, 1997) pursued the route of hierarchical Bayesian models for log-linear dependencies for the multiple-recapture problem with covariates, using the subclass of decomposable graphical models. Instead of estimating N based on a single model, they use Bayesian model averaging. While we do not pursue the model-averaging approach in this paper, it represents a sensible way to extend our approach in order to account for model uncertainty.

2 Three Examples

In this paper, we explore different approaches to the multiple-recapture problem using three examples. The first is based simulated data linked to the Bayesian hierarchical Rasch model described in Section 4 below. The other two examples relate to actual problems of population size estimation, one from the area of public health and the other information sciences. Preliminary analyses of the data in these examples, using what are demonstrably inappropriate models, lead to erroneous inferences. This is related to some of the observations in Hay (1997), for example.

2.1 Simulated Data

Using the basic Rasch model of (1), we randomly drew independent results for the presence of $N = 2000$ individuals from each of $J = 6$ lists. We simulated the values of the individual parameters θ_i, for $N = 2000$ subjects from a $N(0, 4)$ distribution, and their presence or absence from each of six lists according to list parameters $\beta = (-1, -.5, -.25, .25, .5, 1)$. The result was a 2000×6 array of ones and zeros. We summarized this information according to the presence or absence of individuals in the 6 lists, yielding the 2^6 cross-classification of Table 1. When we analyze these data we will treat the number of individuals falling into no lists as if it were unobserved and to be estimated.

In Table 2, we present the classical capture-recapture estimates for N for each pair of lists, using the Petersen estimator:

$$\hat{N} = \left\lfloor \frac{n_{1+} n_{+1}}{n_{11}} \right\rfloor, \tag{2}$$

where n_{1+} is the number of objects in list 1, n_{+1} is the number of objects in list 2, n_{11} is the number of objects in both lists and $\lfloor x \rfloor$ is the greatest integer contained in x. Notice that all 15 estimates of N, which assume the lists are pairwise independent and the objects homogeneous, lie below the true value of 2000, but more importantly below the observed number of objects in all 6 lists, i.e., $n = 2000 - 331 = 1669$. This gives fairly strong evidence about the positive dependence among the lists induced by the heterogeneity. One of the benchmarks of the analyses to come on these data will be the extent to which they adequately deal with this dependence in a parsimonious fashion.

Table 1: 2^6 Table of 2000 Individuals Simulated From a Rasch Model.

					x6	
x1	x2	x3	x4	x5	0	1
0	0	0	0	0	304	108
0	0	0	0	1	70	55
0	0	0	1	0	37	42
0	0	0	1	1	37	50
0	0	1	0	0	30	26
0	0	1	0	1	16	24
0	0	1	1	0	16	25
0	0	1	1	1	17	46
0	1	0	0	0	21	21
0	1	0	0	1	15	37
0	1	0	1	0	9	20
0	1	0	1	1	10	52
0	1	1	0	0	10	10
0	1	1	0	1	8	28
0	1	1	1	0	8	23
0	1	1	1	1	15	97
1	0	0	0	0	11	10
1	0	0	0	1	6	12
1	0	0	1	0	8	11
1	0	0	1	1	7	28
1	0	1	0	0	5	4
1	0	1	0	1	9	18
1	0	1	1	0	1	12
1	0	1	1	1	6	58
1	1	0	0	0	1	4
1	1	0	0	1	6	12
1	1	0	1	0	1	7
1	1	0	1	1	6	58
1	1	1	0	0	2	3
1	1	1	0	1	3	30
1	1	1	1	0	4	25
1	1	1	1	1	14	331

Table 2: Traditional capture-recapture estimates for N using pairs of lists from Table 1.

Lists		$\hat{N}$
x1	x2	1253
x1	x3	1254
x1	x4	1335
x1	x5	1394
x1	x6	1472
x2	x3	1347
x2	x4	1416
x2	x5	1457
x2	x6	1512
x3	x4	1431
x3	x5	1515
x3	x6	1564
x4	x5	1534
x4	x6	1572
x5	x6	1623

2.2 Multiple Sources For Diabetes Ascertainment

Bruno, et al. (1994) used multiple sources to identify known cases of diabetes among the residents of the area of Casale Monferrato in northern Italy on October 1, 1988. They had four sources for their data:

Clinics: List of all patients with a previous diagnosis of insulin-dependent diabetes mellitus (IDDM) or non-insulin dependent mellitus (NIDDM), via diabetes clinic and/or family physicians;

Hospitals: List of all patients discharged with a primary or secondary diagnosis of diabetes in all public and private hospitals in the region;

Prescriptions: Computerized database list of insulin and oral hypoglycemic prescriptions for 1988;

Reimbursements: List of all residents of region who requested a reimbursement for insulin and reagent strips.

We reproduce the data here as Table 3. Bruno, et al. (1994) described an in depth analysis using log-linear models, including the using of stratification to reduce heterogeneity. Their best estimates for N remained in the neighborhood of 2700, substantially in excess of the total observed number of cases in Table 3, i.e., $n = 2069$. When we look at sources in pairs and compute the standard capture-recapture estimates for N as we did in the previous example, we get the results in Table 4. This time three of the six pairwise estimates fall below 2,069 and the other three also lie well below the value reported by Bruno, et al. (1994), which is quite an unsatisfactory situation, indicative of the failure of the assumptions of independent lists and homogeneous objects. Unlike our simulation

Table 3: Data from prevalent cases of known diabetes mellitus for residents of Casale Monferrato, Italy, on October 1, 1988, according to four sources of ascertainment.

		Clinics			
		yes		no	
		Hospitals		Hospitals	
Prescriptions	Reimbursements	yes	no	yes	no
yes	yes	58	46	14	8
yes	no	157	650	20	182
no	yes	18	12	7	10
no	no	104	709	74	?

Table 4: Traditional capture-recapture estimates for N using pairs of sources from Table 3.

Lists	$\hat{N}$
Clinics, Hospitals	2,351
Clinics, Prescriptions	2,185
Clinics, Reimbursements	2,262
Hospitals, Prescriptions	2,052
Hospitals, Reimbursements	803
Prescriptions, Reimbursements	1,555

example, however, we have wide variation in the estimates of N and we may need to cope with both heterogeneity and dependence. Subsequent analyses of these data appeared in International Working Group for Disease Monitoring and Forecasting (1995a), and Biggieri, et al. (1999), some of which we describe below, explore the issue of heterogeneity in different ways.

2.3 The Number of Pages on the World Wide Web

Lawrence and Giles (1998) studied the coverage and recency of six major and widely-available World Wide Web search engines by submitting 575 queries on various scientific topics. We list the six engines in Table 5, along with estimated coverage within their universe of approximately 190×10^6 pages obtained by aggregating across all six search engines and all 575 queries.

Working with Lawrence and Giles directly, we have obtained a nearly-equivalent data set of web pages matching the same 575 queries they used, with the same six search engines. Our data set and the one reported on in Lawrence and Giles (1998) are based on the same raw data, but evolution and fine-tuning of the aggregation algorithms used to obtain the $575 \times (2^6 - 1)$ cross-classifying table lead to minor differences in the two data sets.

These six search engines have built-in positive and negative associations with one another based on how they parse the queries, on how web pages come to be in each search engine's database, and on dependencies that would be induced for example if some engines were to use other engines, or

Table 5: Six major search engines and their estimated coverages within the sample of ca. 190×10^6 pages found by 575 web search queries. *Source:* Lawrence and Giles (1998).

Search Engine	Coverage (%)	95% CI (%)
HotBot (HB)	57.5	± 1.3
AltaVista (AV)	46.5	± 1.3
NorthernLight (NL)	32.9	± 1.1
Excite (Ex)	23.1	± 0.86
InfoSeek (Is)	16.5	± 1.0
Lycos (Ly)	4.41	± 0.42

were to work from a common set of index pages, to develop part of the set of pages that match a particular query. Some of this information is proprietary with the search engine provider, hence unknown to us, and therefore it cannot be directly incorporated into a statistical model for the multiple-recapture data. Indeed one of our central goals is to develop statistical models that are sufficiently flexible that they can capture dependence due to such hidden relations (if they exist), yet sufficiently parsimonious to produce useful estimates of the missing number.

It is also important to realize that, in principle, each of the 575 queries defines a different population of pages, some of which are observed in the corresponding $2^6 - 1$ layer of the full $575 \times (2^6 - 1)$ table. Thus, we tentatively think of query as a stratifying variable to go with our models. Keeping this stratification in mind, we restrict our attention for now to one layer of the $575 \times (2^6 - 1)$ table, corresponding to a single query, #535, listed in Table 6. Further analysis of these data, which is ongoing, could and should try to incorporate query as a multi-level stratifying variable in the models.

As with the other two examples, we begin our analyses by considering the lists in a pairwise fashion and computing the traditional capture-recapture estimates for the population size, which we give here in Table 7. The total number of observed pages across all 6 search engines is $n = 305$, and only 5 of the 15 estimates in Table 7 exceed this value. Thus there is evidence from these marginal calculations suggesting that 10 of the 15 pairs of search engines are positively dependent; there may be some counterbalancing negative dependence among the remaining five pairs. In any case, it seems unlikely that a model asserting joint independence for all six lists will produce very satisfactory population size estimates.

3 The Rasch Model and Quasi-Symmetric Log-Linear Models

We can think of the Rasch model, which we introduced in Section 1, as a mixed-effects generalized linear model that allows for object heterogeneity, and list heterogeneity. For object i, we model the probability of inclusion on list j, as in equation (1), where θ_i is the random catchability effect for object i, which is distributed as a random variable from F_Θ, and the β_j are fixed parameters representing the penetration of list j into the target population. The heterogeneity of capture

Table 6: Multiple list data for Query 535, obtained from Lawrence and Giles (priv. comm.).

					Northern Light	
Alta Vista	Infoseek	Excite	Hot Bot	Lycos	0	1
0	0	0	0	0	–	35
0	0	0	0	1	2	2
0	0	0	1	0	79	13
0	0	0	1	1	0	3
0	0	1	0	0	21	5
0	0	1	0	1	0	1
0	0	1	1	0	3	2
0	0	1	1	1	0	1
0	1	0	0	0	8	0
0	1	0	1	0	11	5
0	1	0	1	1	1	0
0	1	1	0	0	4	2
0	1	1	0	1	1	0
0	1	1	1	0	2	2
0	1	1	1	1	1	0
1	0	0	0	0	33	7
1	0	0	0	1	1	0
1	0	0	1	0	14	7
1	0	0	1	1	1	4
1	0	1	0	0	6	1
1	0	1	1	0	0	2
1	0	1	1	1	0	1
1	1	0	0	0	0	2
1	1	0	0	1	1	2
1	1	0	1	0	5	5
1	1	0	1	1	0	1
1	1	1	1	0	2	5
1	1	1	1	1	1	0

Table 7: Traditional capture-recapture estimates for total N web pages matching Query 535, using pairs of search engines.

WWW Search Engines		$\hat{N}$
AltaVista	Infoseek	256
AltaVista	HotBot	359
AltaVista	NorthernLight	294
AltaVista	Excite	353
AltaVista	Lycos	202
Infoseek	HotBot	254
Infoseek	NorthernLight	274
Infoseek	Excite	192
Infoseek	Lycos	183
HotBot	NorthernLight	362
HotBot	Excite	489
HotBot	Lycos	293
NorthernLight	Excite	309
NorthernLight	Lycos	172
Excite	Lycos	252
Total observed in $2^6 - 1$ table:		$n = 305$

probabilities across objects is therefore influenced by the distribution of θ, F_Θ.

Let $X_1, \ldots, X_J$ be the variables cross-classified in $2^J - 1$ tables like Tables 1 and 6. Mixed-effects models for such tables with a random individual effect θ are a way of thinking about disaggregating the table according to values of θ, and then re-aggregating. In the disaggregated table, let

$$P_j(\theta) = P[X_j = 1|\theta];$$

usually we assume that the lists are independent given θ, so that the probability of observing a count in cell $k_1 \cdots k_J$, given fixed θ, is

$$\pi_{k_1 \ldots k_J|\theta} = P[X_1 = k_1, \ldots, X_J = k_J|\theta] = \prod_{j=1}^{J} P_j(\theta)^{k_j}[1 - P_j(\theta)]^{1-k_j}. \tag{3}$$

Re-aggregating into the $2^J - 1$ table is just integrating over θ, thus the marginal probability of observing a count in cell $k_1 \cdots k_J$ is simply

$$\pi_{k_1 \ldots k_J} = P[X_1 = k_1, \ldots, X_J = k_J] = \int \prod_{j=1}^{J} P_j(t)^{k_j}[1 - P_j(t)]^{1-k_j} dF_\Theta(t) \tag{4}$$

The Rasch model specifies additive logits, $\log P_j(\theta)/[1 - P_j(\theta)] = \theta + \beta_j$, so that

$$\pi_{k_1 \ldots k_J} = \int \exp\left\{\sum_{j=1}^{J} k_j(t + \beta_j)\right\} \prod_{j=1}^{J} \frac{1}{1 + e^{t+\beta_j}} dF_\Theta(t) \tag{5}$$

Integrating with respect to the distribution for θ in equations (4) and (5) turns the Rasch model into a multi-level log-linear model of the form:

$$\log(\pi_{k_1 \dots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_K + \gamma(k_+) \tag{6}$$

where

$$k_+ = \sum_{j=1}^{J} k_j \quad \text{and} \quad \gamma(s) = \log E\left[e^{s\theta}|\mathbf{X} = \mathbf{0}\right] \tag{7}$$

(Cressie and Holland, 1983; Fienberg and Meyer, 1983; Holland 1990; Darroch, et al. 1993). The term $\gamma(k_+)$ models a specific kind of dependence in the $2^J - 1$ table cross-classifying the lists: this dependence is not due to associations between the lists, but rather it arises directly from aggregating across the strata indexed by θ in the model. The value of $\gamma(k_+)$ is not affected by permutations of $k_1, \dots, k_J$ and hence we have a quasi-symmetry model (e.g., see Bishop, et al. 1975), which we can fit to the $2^J - 1$ table of observed counts using standard software for fitting log-linear and generalized linear models. Unfortunately, this transformed version of the Rasch model ignores the moment inequalities implicit in equation (7) (e.g., see Cressie and Holland, 1983), a point to which we return below.

For $J = 3$ lists, the quasi-symmetry model is equivalent (ignoring moment restrictions) to the constraints:

$$\pi_{011}\pi_{100} = \pi_{101}\pi_{010} = \pi_{110}\pi_{001}. \tag{8}$$

These constraints do not relate the probability of the unobserved cell, π_{000}, to the other probabilities, and hence an additional assumption such as no J-way interaction is needed. Estimation can then be carried out using traditional methods for log-linear models, and N can be estimated by

$$\hat{N} = n + \frac{\hat{m}_{odd}}{\hat{m}_{even}}, \tag{9}$$

where $\hat{m}_{odd}$ is a product of estimated expected cell values over all cells whose subscripts sum to an odd value and $\hat{m}_{even}$ is a product of estimated expected cell values over all cells whose subscripts sum to an even value (e.g., see Fienberg, 1972).

Incorporating 2- and 3-way interactions into a log-linear model may not be necessary, and so, as an alternative to the log-linear model in (6), we can consider only lower order symmetries, and simply set the higher-order log-linear interaction terms equal to zero. Again, we can use standard log-linear model or GLM software to fit such models and then project the model to the missing cell using (9).

All log-linear models in this paper were fitted with the `glm()` function in S-Plus (Mathsoft, 1996). The software package `intersect` from the `S` archives at Carnegie Mellon's StatLib (`http://lib.stat.cmu.edu`) was used to construct the quasisymmetry terms and combine them with various linear constraints such as no highest-order interaction. See Darroch, et al. (1993) for further details.

Interval estimates for the total number of cases N were computed using the profile likelihood methods of Cormack (1992): The profile likelihood estimate of the $1 - \alpha$ confidence set for N is defined to be $\{N : G^2(N - n) - G^2(\hat{N} - n) < \chi^2_{(1),1-\alpha}\}$ where G^2 is the model deviance, and $\chi^2_{(1),1-\alpha}$ is the $1 - \alpha$ quantile of a $\chi^2_{(1)}$ distribution. Because S-Plus's `glm()` function estimates the multinomial capture-recapture model using a Poisson likelihood, we must approximate the multinomial deviance , $\widehat{G}^2$, from the Poisson fit. The approximation we use was suggested by Cormack (1992); it is $\widehat{G}^2(u) = D(u) - \log\left\{\frac{u^u(n+u)!}{(n+u)^{n+u}u!}\right\}$, where $D(u)$ is the model deviance for a log-linear Poisson model fit to the 2^J contingency table with u in the missing cell.

4 Hierarchical Bayes Formulation of the Rasch Model

4.1 Model Formulation

If one takes subject heterogeneity as modeled by (4) at all seriously, then the log-linear quasi-symmetry approach has two deficiencies: First, the moment constraints implicit in (7) are not easily incorporated into GLM fits, and hence are usually ignored. Second, the need for and use of the "no k-way interactions" assumption (for $2 < k \leq J$) does not translate into a natural condition on the conditional capture probabilities $P_j(\theta)$ or the catchability distribution $F_\Theta(t)$.

An alternative approach is to estimate the parameters β_j and any parameters of $F_\Theta(t)$ directly from the marginal likelihood (5), by maximum likelihood say, and use the constraints implicit in this formulation to project an estimate onto the missing cell. Coull and Agresti (1999) do exactly this, for example, replacing $F_\Theta(t)$ with a discrete distribution motivated from Gaussian quadrature.

We prefer to work with a fully Bayesian hierarchical specification of the Rasch model. This allows us to lay out all of the pieces of the model, and modify exactly those parts that need adjustment to reflect the dependency in the data. We can use MCMC computing methodology to give essentially exact inferences for remarkably complex models in which the log-linear and marginal maximum likelihood approaches become unwieldy.

We begin our formulation of the basic Rasch model as a hierarchical Bayes model as follows:

$$\left.\begin{array}{rcl} X_{ij} & \overset{indep}{\sim} & Bern(p_j|\theta_i); \quad \log\dfrac{P_j(\theta_i)}{1-P_j(\theta_i)} = \theta_i + \beta_j, \\ & & i = 1,\ldots,N, j = 1,\ldots,J \\ \theta_i & \overset{iid}{\sim} & F_\Theta(\theta_i), i = 1,\ldots,N \\ \beta_j & \overset{iid}{\sim} & G_\beta(\beta_j), j = 1,\ldots,J \end{array}\right\} \tag{10}$$

Note that we only need to add prior distributions $G_\beta(\cdot)$ for the β_j's to the development of the marginal/mixed effects model of equation (5).

When N is unknown and the objects in the $0\ldots0$ cell of the table are missing, we treat N as a parameter in the likelihood for the $2^J - 1$ cross classifying the n observed objects

$$L(N,\theta,\beta;X) = \binom{N}{n}\prod_{i=1}^{N}\prod_{j=1}^{J}\left(\frac{e^{\theta_i+\beta_j}}{1+e^{\theta_i+\beta_j}}\right)^{x_{ij}}\left(\frac{1}{1+e^{\theta_i+\beta_j}}\right)^{1-x_{ij}} \tag{11}$$

where $\mathbf{X}$ is an $N \times K$ matrix, with the (i,j) element $x_{ij} = 1$ if object i is on list j and 0 otherwise. We shall denote the n rows of observed data in $\mathbf{X}$ as $\mathbf{X_1}$. Note that the remaining rows $n+1$ to N of $\mathbf{X}$ are all $\mathbf{0}$'s, vectors of zeroes.

4.2 Estimation by Markov Chain Monte Carlo (MCMC)

Given N and the complete 2^J table, MCMC estimation of the posterior distributions for the parameters in the Rasch model is a straightforward exercise (Patz and Junker, 1997). Here we propose an extension to the MCMC procedure for the Rasch model for multiple-recapture population estimation that is similar to the extensions of the binomial-logit model by George and Robert (1992), and the log-additive model by Basu (1998).

To simplify the MCMC algorithm we have assumed the vector of list parameters, β, is independent of the vector (θ, σ^2, N), and that the list parameters are distributed as $\beta \sim \mathrm{N}_K(\mu, \tau_b \cdot \mathrm{I}_K)$. In the Rasch model described in (10) we have also assumed that conditional on the variance σ^2 and the population size N, that the catchability parameter vector is distributed as $\theta \sim \mathrm{N}_N(\mathbf{0}, \sigma^2 I)$, and that $\sigma^2 \sim \Gamma^{-1}(\alpha, \eta)$.

As usual, the hyperparameters of the list vector β can be chosen to reflect any available prior information. For our analyses, we have chosen $\beta \sim \mathrm{N}_K(\mathbf{0}, 10 \cdot \mathrm{I}_K)$ because we have little prior information about the lists. Similarily we may fix the hyperparameters for the variance of the catchability parameters according to any prior knowledge we may have solicited from the researcher; we have set $\alpha = \eta = 1$. Note that if the prior for σ^2 concentrates mass near zero the model will perform similar to the independent list model.

To further simplify the MCMC algorithm, we assume that the prior distribution of N is

$$f_{I\!N}(N) \propto \frac{(N-l)!}{N!} \mathrm{I}_{\{n<N<N_{max}\}}.$$

This form of prior ensures that the conditional posterior for N is a truncated negative binomial; see equation (12) below. As l increases the prior distribution for N is more concentrated near the observed number n. We have chosen to use $l = 1$, or Jeffrey's prior $f_{I\!N}(N) \propto \frac{1}{N}$. In principle, we can use any proper prior on N but we have found that restricting N to have finite support on the integers, say $[n, N_{max}]$ for some value N_{max}, is helpful in the MCMC simulation. For the examples we present below, we have typically taken N_{max} to be 10,000. In our examples, truncating at $N_{max} = 10,000$ gives up essentially no posterior probability. The specification of the shape of the prior distribution on N, is somewhat more important and we shall return to it below in Section 6.5.

From these model specifications we can readily deduce that the complete conditional posterior distribution for each $\beta_1, \ldots, \beta_J$ is

$$f_B(\beta_j | N, \theta, \mathbf{X}) \propto \frac{e^{\beta_j x_{+j}}}{\prod_{i=1}^N (1 + e^{\theta_i + \beta_j})} f_B(\beta_j)$$

where $x_{+j} = \sum_{i=1}^n x_{ij}$, and $f_B()$ is the normal prior density indicated above. Similarly for σ^2, the complete conditional posterior is

$$\sigma^2 | N, \theta \sim \Gamma^{-1}\left(\alpha + n, \eta + \frac{\sum_{i=1}^N \theta_i^2}{2}\right)$$

Notice that, given θ, σ^2 is drawn independent of the data $\mathbf{X}$.

As Basu (1998) notes, a problem occurs when N is conditioned on θ since $length(\theta) = N$ would tell us N, and we would not have to estimate it. For this reason we follow Basu (1998) in computing a joint conditional posterior distribution $f_{I\!N,\theta}(N, \theta | \mathbf{X_1}, \beta, \sigma)$ for (N, θ) together, and then breaking this apart as $f_{I\!N,\theta}(N,\theta|\mathbf{X_1},\beta,\sigma) = f_{I\!N}(N|\mathbf{X_1},\beta,\sigma)\cdot f_{\theta|I\!N}(\theta|N,\mathbf{X_1},\beta,\sigma)$ for the purposes of simulation: we first draw N from $f_{I\!N}(N|\mathbf{X_1},\beta,\sigma)$, and then we draw θ from $f_{\theta|I\!N}(\theta|N,\mathbf{X},\beta,\sigma) = \prod_{i=1}^{N} f_{\theta_i}(\theta_i|\mathbf{X},\beta,\sigma)$. The (incomplete) conditional posterior for N is

$$\begin{aligned} f_{I\!N}(N|\mathbf{X_1},\beta,\sigma) &\propto \binom{N}{n} f_{I\!N}(N) \prod_{i=n+1}^{N} \int P[\mathbf{0}|\beta,\theta_i] f_{\theta_i}(\theta_i|\sigma)d\theta_i \\ &\propto \binom{N}{n} f_{I\!N}(N) P[\mathbf{0}|\beta,\sigma]^{N-n} \end{aligned} \tag{12}$$

We can determine the probability $P[\mathbf{0}|\beta,\sigma]$, which is the the probability associated with the unobserved cell, i.e., $\mathbf{X}=\mathbf{0}$, analytically for some priors $f_\theta(\theta|\sigma)$, but in most cases we must approximate or compute it by numerical integration.

If $f_{I\!N}(N) \propto I_{\{N>n\}}$, $\frac{I_{\{N>n\}}}{N}$, or $\frac{I_{\{N>n\}}}{N(N-1)}$, etc., then $N-n$ has a truncated negative-binomial conditional posterior. After we have simulated N, we now simulate from the complete conditional posterior for θ, satisfying

$$f_{\theta_i|I\!N}(\theta_i|\boldsymbol{\beta},\sigma,N,\mathbf{X}) \propto \frac{\exp(-\frac{\theta_i^2}{2\sigma^2}+\theta_i x_{i+})}{\prod_{j=1}^{J}(1+e^{\theta_i+\beta_j})}$$

for $i=1,\ldots,N$, where $x_{i+}=\sum_{j=1}^{J} x_{ij}$.

The time taken to complete a sufficiently long run for this algorithm is highly dependent on the data set. This is because each step of the Markov chain is $O(N)$ and N is being estimated. The speed of the algorithm is approximately 10,000 simulations per hour for the world wide web data set, on a Hewlett Packard 9000/770 UNIX workstation. We have taken 50,000 simulations after a burn-in period of 10,000 simulations in our analyses. Long runs are necessary because the model parameters are highly correlated. For example, in the diabetes analysis the posterior correlation between the population size N and a given list parameter β_j is around -0.8. The correlation between N and σ^2 is 0.82. For this reason the Markov Chain is very slow to mix.

An alternative to the Basu (1998) approach is to treat the distribution of the observed data $P(\mathbf{X_1}|N,\beta,\sigma)$ implied by (11) as a different model for $\mathbf{X_1}$, for each different N. This leads to the formulation of the problem of selecting N as a model selection problem for $\mathbf{X_1}$; and the relevant MCMC technique is Green's (1995) "reversible jump" approach for randomly selecting models from a well-defined model space. We have also implemented this approach; the results are quite similar to the Basu approach outlined above—and may be useful for more complex models—but in the models we have compared the approaches with, the Basu technique produces faster and more stable MCMC runs. Fortran programs implementing both the Basu and Green approaches are available from us (contact `masjohns@stat.cmu.edu`).

5 Initial Analyses of the Three Examples

5.1 The Simulated Data

We applied the basic log-linear models and Bayesian Rasch model to the simulated data for six lists that we presented earlier in Table 1. Table 8 contains our various estimates for N, the size

Table 8: Estimates of the population size for 2000 objects and six lists; data simulated from a Rasch model. (Observed $n = 1696$.)

Model	df	Deviance	Point Estimate	95% Interval
Independence:	56	1335.44	1701	[1697,1707]
QS with no 3-way or higher	55	50.16	1974	[1913,2046]
QS with no 4-way or higher	54	42.05	1859	[1796,1946]
QS with no 5-way or higher	53	41.46	1932	[1772,2326]
QS with no 6-way	41.45	52	1904	[1697,6536]
Bayesian Rasch model			Median 2019	[1939,2128]

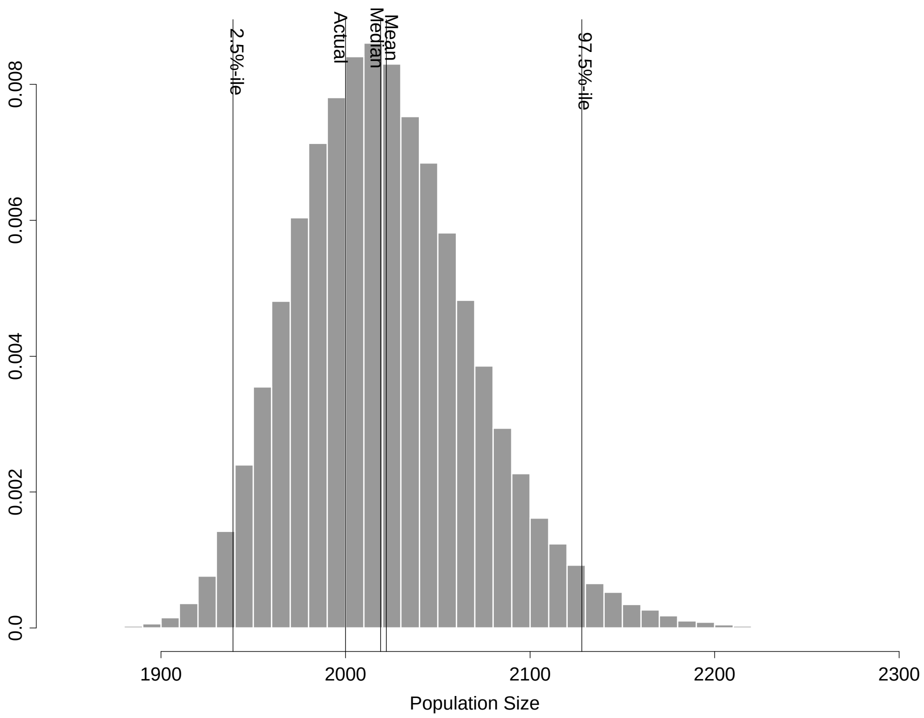

Figure 1: Posterior Distribution for Unknown Sample Size for Simulated Data (Actual=2000).

Table 9: MCMC estimated posterior mean and and quantiles for the list parameters, $\{\beta_j\}$, and prior standard deviation σ on the random catchability effects, $\{\theta_i\}$, based on 2000 objects simulated from the Rasch model, but where $n = 1696$. Actual parameters used in the simulation of the data are given in the rightmost column.

Name	mean	2.5%ile	median	97.5%ile	actual
List 1	-1.03	-1.27	-1.02	-0.81	-1.00
List 2	-0.40	-0.64	-0.40	-0.19	-0.50
List 3	-0.29	-0.53	-0.29	-0.08	-0.25
List 4	0.24	0.00	0.24	0.45	0.25
List 5	0.58	0.33	0.58	0.79	0.50
List 6	0.95	0.70	0.94	1.17	1.0
σ	2.10	1.90	2.10	2.32	2.00
N	2022	1939	2019	2128	2000

of the simulated population. The 95% interval for the classical models is a 95% profile-likelihood based interval (see Cormack, 1992), whereas the 95% interval for the Bayes Rasch model is an equal-tailed posterior probability interval. We recall that the true total is $N = 2000$. The independence model fits the data poorly (as indicated by the value of the deviance), and it underestimates the true value substantially as well. All of the quasi-symmetry log-linear models fit the data reasonably well but they too underestimate the true value. The quasi-symmetry model 95% confidence intervals illustrate somewhat erratic behavior. The quasi-symmetry model with no second-order interaction is well-behaved and has a relatively tight confidence interval which includes the true value. Allowing for a third-order interaction leads to a much lower estimate and an interval which does not include the true value. Finally the estimate with no fifth-order interaction is reasonable but the confidence interval "explodes" suggesting some specification problem or a ridge in the likelihood function. See Section 6.4 for further examples of this sort of behavior.

The Bayesian Rasch model, which we display in Figure 1, yields a well-behaved posterior distribution, centered close to the true value and a reasonably tight 95% posterior interval. Table 9 contains the estimates of list parameters, or catch efforts, $\{\beta_j\}$, and the standard deviation of the random catchability effects $\{\theta_i\}$.

5.2 The Diabetes Data

The International Working Group (1995a) gives a detailed treatment of the estimation of the log-linear and quasi-symmetry log-linear models for the diabetes data. Thus in Table 10, we simply provide some illustrative models in the class outlined in Section 3. As in Table 8, the 95% interval for the classical models is a 95% profile-likelihood based interval, whereas the 95% interval for the Bayesian Rasch model is an equal-tailed posterior probability interval. The independence model fits the data poorly, and the confidence bounds are tight and relatively close to the observed value of $n = 2069$. The second model in the table, involving all first-order interactions except the interaction between reimbursements and clinics, was one we chose based on a stepwise procedure using

Table 10: Estimates for the number of diabetes mellitus cases in Casale Monferarato, Italy, on October 1, 1988 using various methods. ($n = 2069$)

Model	df	Deviance	Point Estimate		95% Interval
Independence	10	217.48		2250	[2216,2288]
All 1st-order interactions, except Reimbursements × Clinics	5	7.62		2771	[2533,3112]
QS no 3-way or higher	9	105.63		2669	[2525,2846]
QS no 4-way	8	93.95		2238	[2142,2425]
Saturated	−−	−−		5367	
Bayesian Rasch model			Median	2703	[2567,2906]
			Mode	2693	
			Mean	2712	

the Bayesian information criterion or BIC (e.g., see Kass and Wasserman, 1995) and provides an extremely good fit to the data. The results for this model are similar to those for the "best" model reported by the International Working Group (1995a), and Bruno, et al. (1994).

Both of the quasi-symmetry models improve substantially upon the fit of the independence model, and the quasi-symmetry model with no second order interactions produces an estimate of N which is reasonably close to the value of the best model above, and has tighter intervals. But the fit of the quasi-symmetry model *with* the second-order interactions included seems to be "off," much like that of the independence model.

The Bayesian Rasch model produces results that are remarkably close to those from the best fitting log-linear model but with a much more parsimonious model. The posterior 95% probability interval is in fact much tighter that the corresponding classical confidence interval for the best fitting log-linear model.

5.3 The World Wide Web Data

The classical multiple-recapture independence model fits the data in our WWW example for Query #535 quite poorly, and it projects only an additional 75 unseen web pages. We chose the second log-linear model in Table 11 using a stepwise procedure and the BIC criterion as in the Diabetes example above. It includes 8 first-order interaction terms (using the notation of Table 5):

$$(\text{AV:Is})+(\text{AV:HB})+(\text{AV:Ly})+(\text{AV:NL})+(\text{Is:Ex})+(\text{Is:HB})+(\text{Ex:NL})+(\text{Ly:NL}).$$

Not surprisingly, it fits the data considerably better as measured by the deviance, although the goodness-of-fit improvement is not as striking as was the case for the best standard log-linear model in the diabetes example. What is especially interesting here is that, while we observe only 305 web pages in total for the 6 search engines combined, our best estimate for the total population size is 602, with a fair bit of variability about this value.

The quasi-symmetry model restricted to first-order interactions also provides a reasonable fit to the data, with estimates of N close to those from the "best" standard log-linear model comes

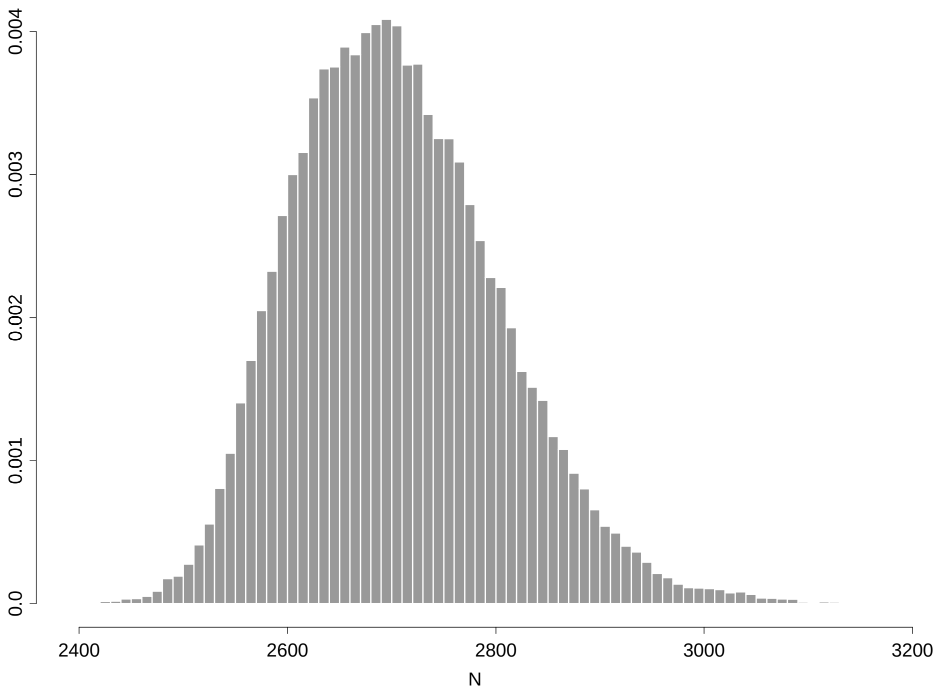

Figure 2: Posterior distribution of the number of individuals with diabetes in Casale Monferrato, Italy.

Table 11: Estimates of the number of world wide web pages on the topic "Query #535" using various estimation methods. ($n = 305$.)

Model	df	Deviance	Point Estimate	95% Interval
Independence	56	148.73	373	[350,400]
BIC-based log-linear model	46	65.62	602	[481,793]
QS no 3-way or higher	55	88.61	614	[498,778]
QS no 4-way or higher	54	83.33	1266	[624,3296]
QS no 5-way or higher	53	82.43	508	[305,4875]
QS no 6-way	52	81.71	861882	[305,∞]
Bayesian Rasch model			Median 798 Mode 685 Mean 861	[554,1525]

close to the best, and the quasi-symmetry models with higher order interactions appear to blow up again. We interpret this as evidence that the likelihood function is not well behaved, and it may also provide diagnostic information to suggest that the positive dependence presumed by the underlying Rasch model is not well satisfied by the data.

From this estimated posterior distribution displayed in Figure 3, we see that the mode for the Bayesian Rasch model is close to that of the best classical log-linear model estimate. Clearly the data in these problems is generally not as informative as one would like about the missing count, but the dramatically greater spread of the Bayesian Rasch model credible interval for N (vs. the interval estimate from the log-linear model) hints at inadequacies of this basic Rasch model. On the other hand, we have a clear standard against which to calibrate our search for alternative models: the basic loglinear model with a selection of first order terms. A careful perusal of the fits of the other models in Table 11 (and several more complex loglinear models with 2nd-order interactions for which we do not present detailed results) makes clear that adding higher-order terms in a standard loglinear fashion is unlikely to make major gains in fit and estimation of N. It is only through something akin to a generalization of the full hierarchical Rasch model that we could hope to simultaneously reduce the parametrization and improve the fit.

5.4 Discussion of Initial Analyses

There are some common features in our initial analyses of the examples in this section.

First, we have seen that relying upon the assumption of independence of lists is misleading, and sometimes badly so. In none of the cases did the model fit the data well. In the simulated example, we knew this to be true by design, and the goodness-of-fit tests verify that fact. But in the other two examples, we have poor fits as well and other evidence pointing to a serious underestimate of the population size, and some earlier indication that this was a result of positive association among at least a subset of the lists. Fitting log-linear models to account for dependencies, one at a time, provides one possible fix to this mis-estimation.

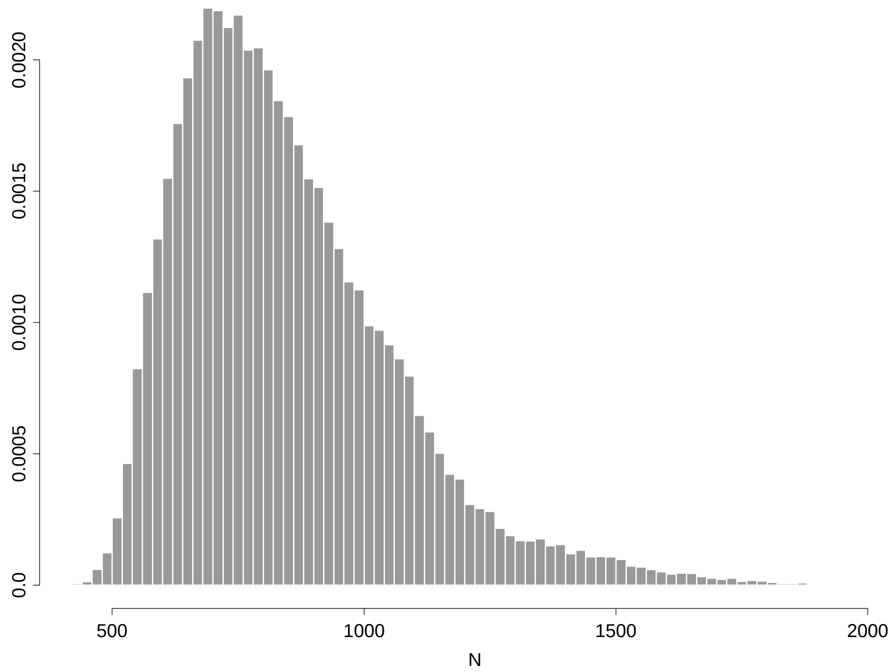

Figure 3: Histogram of the posterior distribution of the number of web pages of corresponding to query # 535.

Second, the log-linear quasi-symmetry version of the Rasch model, which tries to provide a broad fix for positive dependence among lists due to heterogeneity in object catchability, combined with the assumption of no highest order interaction, can often "blow up"–as it did in the simulation and WWW examples—even though the model provides a clear improvement over independence. Setting additional higher-order terms equal to zero seems to counter act this anomalous behavior, and it yields remarkably good estimates and with fewer parameters than the standard log-linear model approach.

Finally, the approach based on a hierarchical Bayesian formulation of the Rasch model does at least as well as the log-linear and quasi-symmetry log-linear approaches, and this appears to be through its replacement of the log-linear "no highest-order interaction" assumption with the hidden moment constraints indicated in equation (7) above. The skewed posterior distributions for the Bayesian Rasch model, and the not-so-great fit of the quasi-symmetry models suggest that perhaps we can do better through a relaxation of the strong assumptions of the Rasch model. We explore this notion in the next section.

6 List and Latent Variable Interactions

The log-linear representation of the Rasch model, given in equation (6), invites a variety of extensions of the model within the space of log-linear models for the marginal distribution of the data, $\pi_{k_1\cdots k_J}$. Starting from the basic model

$$\log(\pi_{k_1\ldots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_J + \gamma(k_+),$$

and treating the k_1 and k_+ terms as identifying various margins of the table $\{n_{k_1\cdots k_J}\}$, we can immediately consider adding interactions among the terms in hierarchical fashion, e.g.

$$\log(\pi_{k_1\ldots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_J + \gamma(k_+) + \sum_{j_1\neq j_2}\sum \beta_{j_1 j_2} k_{j_1} k_{j_2} + \sum_{\ell} \gamma'(k_\ell, k_+). \qquad (13)$$

We believe that these so-called "generalized log-linear Rasch models" were first considered in detail by Kelderman (1984), whose interest in them was to develop hierarchically nested alternatives to the null hypothesis that the data follows the log-linear Rasch model of equation (6). Cormack (1989) provides an independent, alternative development of these ideas. They have also proven useful in extending the log-linear Rasch model to accommodate dependence in the table $\{n_{k_1\cdots k_J}\}$ that is Rasch-like but more general than the "exchangeable higher moments" structure of the Rasch model (Darroch and McCloud, 1990; Carriquiry and Fienberg, 1998; and Biggeri et al., 1999). In particular, the list-by-list interactions allow for negative dependence between some pairs of lists that are excluded by the basic Rasch model's assertion of equal positive associations among all the lists. It is also worth noting that models of the form (13) may contain interactions of all orders, but they may be severely constrained both by the form of the $\gamma'(\cdot)$ terms in (13) and by moment inequality constraints of the type discussed after (7) above.

In this section, we sketch some connections between the generalization in (13) of the log-linear Rasch model on the one hand, and the hierarchical Bayesian Rasch model considered in Section 4 on the other. In order to do this, we find it convenient to break the generalized Rasch model of equation (13) into two parts:

$$\log(\pi_{k_1\cdots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_J + \gamma(k_+) + \sum_{j_1\neq j_2}\sum \beta_{j_1 j_2} k_{j_1} k_{j_2}, \tag{14}$$

and

$$\log(\pi_{k_1\cdots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_J + \gamma(k_+) + \sum_{\ell} \gamma'(k_\ell, k_+), \tag{15}$$

corresponding to list-by-list interactions, i.e., as in equation (14), and list-by-total-captures interactions, i.e., as in equation (15).

6.1 List-By-List Interactions

We may write the likelihood (3) for the observed table of counts, given θ as

$$\begin{aligned}\pi_{k_1\cdots k_J|\theta} &= \prod_{j=1}^{J} P_j(\theta)^{k_j}[1-P_j(\theta)]^{1-k_j} = \prod_{j=1}^{J}\left[\frac{P_j(\theta)}{1-P_j(\theta)}\right]^{k_j}\prod_{j=1}^{J}[1-P_j(\theta)] \\ &= \exp\{\sum_j \lambda_j(\theta)k_j\}\prod_{j=1}^{J}[1-P_j(\theta)],\end{aligned}$$

where $\lambda_j(\theta) = \log\{P_j(\theta)/[1-P_j(\theta)]\}$. Hence

$$\log \pi_{k_1\cdots k_J|\theta} = \alpha(\theta) + \sum_j \lambda_j(\theta)k_j. \tag{16}$$

where $\alpha(\theta) = \prod_{j=1}^{J}[1-P_j(\theta)]$, and for the Rasch model in particular, $\lambda_j(\theta) = \theta + \beta_j$.

Following the discussion of Darroch, et al. (1993), we note that interdependencies among the lists in the observed table $\{n_{k_1\cdots k_j}\}$ may either be the result of collapsing over θ (a version of Simpson's paradox, see Holland and Rosenbaum, 1986, or Kadane, et al. 1999) or they may be a consequence of the lists being truly interdependent even when the data are disaggregated to the person or object level. As an example of this latter type of dependence, consider two web search engines that draw from the same pool of web pages (perhaps because of a common indexing strategy): a web page may be more likely to show up in one, given than it is in the other, quite apart from the visibility of the page to search engines in general. Or, in the diabetes example, clinical records of prescriptions and medical reimbursements almost always occur together in areas where drugs expenses are part of health coverage, inducing a positive dependence that does not depend on averaging over heterogeneous visibility of patients. Similarly, lists that by their nature penetrate nearly disjoint subpopulations induce a tendency toward negative dependence in the marginal distribution $\pi_{k_1\cdots k_J}$.

To model list-by-list dependencies that are not artifacts of aggregating over θ, we add to equation (16) the two-way interactions in the conditional (fixed θ) model:

$$\log \pi_{k_1\cdots k_J|\theta} = \alpha(\theta) + \sum_j \lambda_j(\theta)k_j + \sum_{j_1\neq j_2}\sum \lambda_{j_1 j_2}(\theta) k_{j_1} k_{j_2}, \tag{17}$$

where $\alpha(\theta)$ is simply the usual log-linear model normalizing constant (sum of model terms over all values of $k_1, \ldots, k_J$). We assume as before that $\lambda_j(\theta) = \theta + \beta_j$, and now we also assume that $\lambda_{j_1 j_2}(\theta) = \theta + \beta_{j_1 j_2}$. This leads to the form

$$\log \pi_{k_1 \cdots k_J|\theta} = \alpha(\theta) + \sum_j \beta_j k_j + \theta k_+ + \sum_{j_1 \neq j_2}\sum \beta_{j_1 j_2} k_{j_1} k_{j_2} + \theta \sum_{j_1 \neq j_2}\sum k_{j_1} k_{j_2}. \tag{18}$$

Jannarone (1986) and Jannarone, Yu and Laughlin (1990) developed extensions of the Rasch model similar to this in order to model dependence between examination items in educational testing.

Exponentiating, integrating with respect to the distribution of the random catchability effects, θ, and taking the logarithm again, we readily obtain the model

$$\log(\pi_{k_1 \cdots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_J + \sum_{j_1 \neq j_2}\sum \beta_{j_1 j_2} k_{j_1} k_{j_2} + \gamma^*(k_+ + \sum_{j_1 \neq j_2}\sum k_{j_1} k_{j_2}). \tag{19}$$

Now since the k_j's are binary and $k_+ \geq 0$, it follows that

$$\gamma^*(k_+ + \sum_{j_1 \neq j_2}\sum k_{j_1} k_{j_2}) = \gamma^*([k_+]^2) = \gamma(k_+),$$

subject to moment constraints (that, as before, are usually suppressed in log-linear fitting). The result is that we obtain the log-linear model of equation (14).

6.2 List-By-Total Interactions

We can also use the ideas leading to the model (19) to introduce list-by-total interactions in the marginal model. If we add terms of the form

$$\sum_j \lambda_{j+}(\theta) k_j k_+ \tag{20}$$

to the local dependence model (17), where $\lambda_{j+}(\theta)$ has the now familiar form $\lambda_{j+}(\theta) = \theta + \beta_{j+}$, we introduce additional terms of the form

$$\sum_j \beta_{j+} k_j k_+ + \theta \cdot [k_+]^2$$

in equation (18); and, after exponentiating, integrating and simplifying, we obtain a version of the full model (13):

$$\log(\pi_{k_1 \cdots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_J + \sum_{j_1 \neq j_2}\sum \beta_{j_1 j_2} k_{j_1} k_{j_2} + \sum_j \beta_{j+} k_j k_+ + \gamma(k_+),$$

subject again to moment constraints on $\gamma(k_+)$ which are usually ignored in the log-linear fits. This model is not well identified, but sensible submodels of it involving a small number of list-by-list or list-by-total interactions may be useful in providing well-fitting, parsimonious log-linear models for multiple recapture problems. It is difficult, however, to interpret the local—conditional on θ—interaction terms in expression (20). An alternative development, which we now present, does not lead to as general a model, but does allow us to give a sensible interpretation to list-by-total interactions, in terms of heterogeneity of catchability (visibility) of the objects being counted.

Let us begin again with the basic likelihood (3) in Section 3, and suppose now that θ is multi-dimensional, i.e.,

$$\theta = (\theta_1, \theta_2, \dots, \theta_q).$$

Moreover, suppose that different sets of lists depend on different θ_i's through the Rasch model. For example, suppose that $\theta = (\theta_1, \theta_2)$ and we can partition the lists into I lists that depend only on θ_1 and $J - I$ lists that depend only on θ_2. Then, after permuting list indices, the likelihood given θ becomes

$$\pi_{k_1 \cdots k_J | \theta_1 \theta_2} = \prod_{j=1}^{I} P_j(\theta_1)^{k_j} [1 - P_j(\theta_1)]^{1-k_j} \prod_{j=I+1}^{J} P_j(\theta_2)^{k_j} [1 - P_j(\theta_2)]^{1-k_j}. \tag{21}$$

This sort of structure was employed by Darroch, et al. (1993) to model the different visibility of persons in administrative lists, versus their visibility in U.S. Census lists and in a post-enumeration survey also conducted by the U.S. Census Bureau. Here θ_1 and θ_2 are both random effects for catchability; separating them out in this way allows for some list-by-person or list-by-latent-variable interactions, that are not otherwise easy to model.

If, as would usually seem reasonable, the density $f(\theta_1, \theta_2)$ does not factor, then a derivation similar to that leading from equation (18) to equation (19) above now leads us to

$$\log(\pi_{k_1 \dots k_J}) = \alpha + k_1\beta_1 + \cdots + k_J\beta_J + \gamma(k_+^{(1)}, k_+^{(2)}), \tag{22}$$

where $k_+^{(1)}$ is the number of captures in the first I lists, and $k_+^{(2)}$ is the number of captures in the remaining lists. Such general partial quasi-symmetry terms are not usually considered in log-linear modeling of multiple-recapture data, and they suggest new ways to expand the basic Rasch quasi-symmetry log-linear model (6) to account for "extra-Rasch" variability in the catchability random effect.

For now, we show how to exploit the model of equation (22) to motivate adding list-by-total interactions to the basic Rasch quasi-symmetry model. In particular, we can use the construction above to add a single interaction term, say $\gamma(k_1, k_+ - k_1)$, to the basic Rasch quasi-symmetry model. This is equivalent to adding the term $\gamma(k_1, k_+)$ to the model, since the pair $(k_1, k_+ - k_1)$ and the pair (k_1, k_+) have the same number of levels (due to the constraint $k_+ = k_1 + \cdots + k_J$); and this provides some extra theoretical grist for Carriquiry and Fienberg's (1998) somewhat informal interpretation of this term. In their work they were attempting to interpret a component of a model proposed by Darroch and McCloud (1990) for an epidemiological example which was, in essence, allowing for something like a list-by-latent-variable interaction. Of course we may also add list-by-list interactions to this model, by analogy with equation (17).

6.3 Hierarchical Bayesian Approaches to List-by-List and List-by-Total Interactions

The structure of the hierarchical Bayes form of the models discussed in Sections 6.1 and 6.2 is exactly of the same form as discussed in equation (10). In the case of the two-dimensional Rasch model (21) underlying the list-by-total interaction submodel of (22), we defined θ_1 to underlie $J-1$ lists, and θ_2 to underlie the list for which we want a list-by-total interaction. The prior distribution

Table 12: Estimates of the number of diabetes mellitus cases in Casale Monferarato, Italy from Section 5, and with list-by-list and list-by-total interactions.

Model	df	Deviance	Point Estimate	95% Interval
BIC[1]	5	7.62	2771	[2533, 3112]
QS2[2]	9	105.64	2669	[2522, 2846]
QS3	8	93.95	2239	[2142, 2425]
QS2 + BIC[3]	6	8.32	2752	[2531, 3060]
QS3 + BIC[4]	5	2.04	4152	[2843, 7388]
QS2+$k_+\times$Prescriptions[5]	6	67.39	4367	[3111, 6893]
QS2+$k_+\times$Clinic	6	66.48	7796	[3793, 12068]
QS2+$k_+\times$Reimbursements	6	77.03	2411	[2306, 2552]
QS2+$k_+\times$Hospitals	6	65.86	2381	[2257, 2568]
Bayesian Rasch	−−	−−	2693[6]	[2567, 2906]
Bayesian $k_+\times$Prescriptions	−−	−−	2743[6]	[2558, 2993]
Bayesian $k_+\times$Reimbursements	−−	−−	2556[6]	[2446, 2750]
Bayesian $k_+\times$Clinic	−−	−−	2537[6]	[2426, 2831]
Bayesian $k_+\times$Hospitals	−−	−−	2845[6]	[2664, 3210]

[1]Stepwise BIC selects: independence + reimburse:hospitals + presc:reimburse + presc:clinic + presc:hospitals + clinic:hospitals.

[2]QS2 indicates the Rasch quasi-symmetry model with no 3- or higher-way interactions. Similarly QS3 indicates Rasch quasi-symmetry with no 4- or higher-way interactions.

[3]Stepwise BIC starts with QS2 and adds presc:reimburse + presc:clinic + reimburse:hospitals.

[4]Stepwise BIC starts with QS3 and adds presc:reimburse + presc:clinic + reimburse:hospitals.

[5]$k_+\times$Prescriptions indicates one list-by-total interaction involving the prescriptions list. Similarly for the other $k_+\times$List interaction models shown.

[6]Posterior mode.

for (θ_1, θ_2) was taken to be iid $\mathrm{N}_2(\mathbf{0}, \Sigma)$, with $\Sigma \sim$ Inv-Wishart$_\nu(\Sigma_0)$ and $\nu = 1$, $\Sigma_0 = I$, the identity matrix. The population size N is assumed to be independent of Σ, the covariance matrix of the catchability distributions. The prior distributions on N and on the list parameters β_j were the same as in the unidimensional model described in (10). The MCMC development is exactly analogous to the development in Section 4, except that the Γ^{-1} complete conditional distribution for σ^2 is replaced by a corresponding inverse Wishart distribution for Σ.

6.4 Some Illustrative Fits for the Examples

To illustrate some of the flexibility that these models provide we present here classical log-linear fits for models that add list-by-list and list-by-total interactions to the basic log-linear quasi-symmetry Rasch model.

6.4.1 Diabetes

We have reproduced some of our earlier analysis results for the diabetes example in Table 12 along with a summary of results for models from this section. Earlier, we had seen that the quasi-symmetry models by themselves did not provide an adequate fit to the data although the quasi-symmetry with no 2nd-order interactions had a reasonable estimate of N. These results are reproduced as the first panel in Table 12. In the second panel of Table 12, we show what happens when we combine the QS2 and QS3 models with list-by-list interactions selected by BIC. The QS2 plus interactions model produces a result essentially the same as the "best" log-linear model selected using the BIC criterion alone. The QS3 plus interactions model, while fitting the data extremely well, "blows up" and produces what is a substantively nonsensical result—that there are more than double the total number of observed diabetics in the population! This behavior, similar to what we observed earlier in the other two examples, is indicative of the model overfitting a near-zero margin of the $(2^4 - 1) \times (total\ captures)$ table of counts contributing to the denominator of (9).

In the third panel of Table 12 we see that QS2 models that add a single list-by-total interaction provide substantially better fit than QS2 alone, for example, but they do not fit well relative to the saturated model. Moreover the point estimates are not particularly consistent with one another, and all the interval estimates seem "wrong" in that they do not contain the baseline BIC estimate of 2771.

As a first step toward improving these generalized Rasch models, we also tried a model that adds to QS2 both of the "list $\times$ total" interactions from the two best-fitting cases: "Hospitals" $\times$ total and "Reimbursements" $\times$ total, and a model that included these interactions as well as freeing the corresponding list $\times$ list interaction ("Hospitals" $\times$ "Reimbursements"). These models produced marginally better fits, but in the latter case the point estimate was quite high, again indicating overfit to a near-zero margin in the $(2^4-1)\times(total\ captures)$ table. Moreover in both cases the fitted model was deficient in degrees of freedom—again, we suspect, due to the sparseness of the table—precluding the possibility of profile likelihood calculation of the corresponding interval estimates. Illustrations of estimates based on models of these types will be given in the World Wide Web example in the next subsection, where interval estimates can be calculated.

In the fourth panel we have repeated the Bayesian Rasch estimate we obtained earlier, and in the fifth panel we have provided estimates for the Bayesian form of the "Rasch plus list $\times$ total" models outlined at the end of Section 6.2, again for all four lists. It can be seen that the point estimates and interval estimates are much more stable and mutually consistent, and the corresponding interval estimates all contain the BIC point estimate of 2771, except for the "Reimbursements $\times$ total" model, which just misses it.

This example suggests that the Bayesian Rasch model alone does quite well, and the Bayesian "Rasch plus list $\times$ total" models do a better job of smoothing over the sparse table than do the corresponding log-linear "Rasch plus list $\times$ total" models. There seems to be little in the data to cause us to prefer one model over the other, but it does not matter since these models all give answers consistent with one another, and consistent with our best log-linear model, the baseline BIC model.

The parsimony of the Bayesian Rasch and Rasch plus list $\times$ total models is attractive. In addition, we can examine and compare posterior information on the catchability effects θ in these models. In the simple unidimensional Bayesian Rasch model for example, the posterior mode for the variance of θ is approximately 1.24, with an equal tailed 95% credible interval running from 1.04 to 1.45. When $\theta = (\theta_1, \theta_2)$ is two dimensional, the variance of θ_1 (the catchability parameter underlying $J - 1$ lists) increases slightly to the range of 1.38 to 1.51 (depending on the particular model), with interval estimates similar to the unidimensional case. The posterior mode for the variance of θ_2 (the catchability parameter underlying the list in the list $\times$ total interaction) ranges from approximately 3.39 to 11.36 (depending on the model), with 95% intervals ranging from a lower bound of approximately 1.95 to 4.09 to an upper bound of approximately 20.06 to 34.16 (again depending on the model). The correlation between θ_1 and θ_2 has a posterior mode ranging from 0.64 to 0.75, with 95% intervals bounded below by about 0.4. Clearly there isn't much information available to identify the distribution of θ_2 separately in the model, but including θ_2 does provide the model with a little bit more freedom to fit the observed table of counts, while still smoothing over sparse margins in the $(2^4 - 1) \times (total\ captures)$ table.

6.4.2 World Wide Web

In Table 13, we report on relevant models from the earlier analyses of the WWW data as well as models with list-by-list and list-by-total interactions. What is clear from a perusal of the table is that the log-linear models are very unstable when any interactions higher than first order (two-way interactions) are included in the model. On the other hand, all of the two-way interaction models perform similarly, suggesting a population total estimate of approximately 600 with a CI that runs from approximately 500 to 800.

By analogy with the Diabetes analyses above, we also constructed a model that adds the two best list$\times$total interactions to the QS2 model; but this produces only a minor improvement in fit and no qualitative improvement in point or interval estimates of the population total. In the fourth panel of Table 13, this model is compared with several other models that add two or three list $\times$ total terms to the basic QS2 model. We see some moderate increase in the point estimates and in the range of the interval estimates, as can be expected when more of the positive dependence in the lists is modeled, but these results are broadly supportive of our general impression about the point and interval estimates for the population total garnered from the two-way interaction models above.

The simple Bayesian Rasch model can be seen to perform intermediately between most of the log-linear two-way interaction models and the QS3 and higher-order log-linear models, with a fairly reasonable point estimate but a wide 95% credible interval indicating a long right tail in the posterior distribution for N. The three Bayesian Rasch plus list $\times$ total models included in the last panel of Table 13 show that the right list $\times$ total interaction helps the interval estimate settle down to something reasonable, but there is little hint to guide us in selecting the list $\times$ total interaction that we want.

Again, sparseness in the table seems to be the culprit. With 305 observations spread across 6 lists there is an average of less than 5 observations per cell, and in fact there are many empty cells in the table. This produces severe non-smoothness in the $2^6 - 1$ table—let alone the $(2^6 - 1) \times (total\ captures)$ table—and higher-order models tend to track this non-smoothness; they are consquently misled in their estimation of dependence in the table. The hierarchical Bayesian model allows some restricted fitting of interactions of all orders (recall equation (7) for example) and it may be that unless further natural restrictions are placed on these interactions the Bayesian model is similarly confused by the relatively sparse table for Query #535.

In the simple unidimensional Bayesian Rasch model, the posterior mode for the variance of θ is approximately 1.65, with an equal tailed 95% credible interval running approximately from 1.20 to 2.23. The posterior mode of the variance for θ_1 (the catchability parameter for $J-1$ lists in the two-dimensional model) increases to the range 1.97 to 2.75 (depending on the model) and for θ_2 ranges from 6.15 to 12.35, with correspondingly wide 95% intervals. The posterior mode of the correlation between θ_1 and θ_2 ranges from 0.63 to 0.88, with 95% intervals bounded below by approximately 0.39. In the Web data therefore, there was little difference between the simple Bayesian Rasch model and the two-dimensional models underlying the Bayesian Rasch plus list $\times$ total models. This explains the similarity in the point estimates among the various Bayesian Rasch models listed at the end of Table 13.

6.5 Analysis of the Sensitivity of Inferences to Prior Specifications

We have explored the sensitivity of the Bayesian model based estimates of N to the choice of prior distributions and found that the posterior 95% intervals for the population size N are somewhat affected by the choice of prior distribution for N and for σ^2, but less so by the choice of prior for the β's. Summaries of the sensitivity analyses for N and σ^2 for the Diabetes and World Wide Web data are contained in Tables 14 and 15.

For example, applying the simple Bayesian Rasch model to the World Wide Web data, we find that changing the prior distribution on N from a truncated uniform distribution, $\pi(N) \propto 1$, to a truncated Jeffreys prior, $\pi(N) \propto \frac{1}{N}$ [which was used throughout the paper], while holding the prior on $\sigma^2 \sim \Gamma^{-1}(1,1)$, unchanged, changes the 95% posterior interval from [569,1662] to [554,1525] (see Table 15). These changes are not large relative to the sizes of the estimates involved. Moreover, increasing the high truncation point we chose, $N_{max} = 10,000$, did not affect the posterior distribution for N in any significant way.

Tables 15 also shows that, as the prior distribution on σ^2 places more and more mass near zero, the estimate for N decreases towards the independent-lists estimate—as we would expect. For example the interval estimate for N using Jeffreys prior for N and $\Gamma^{-1}(1,1)$ for the variance gives a posterior 95% interval of [554,1525], and we find an interval of [464,961] when using a prior that concentrates mass at zero for the variance, $\sigma^2 \sim \Gamma^{-1}(5,.05)$.

7 Discussion

In this paper we have reviewed and extended the by-now long history of statistical modeling of multiple-recapture or multiple-list census data for the purpose of estimating population totals. In any situation in which there may be heterogeneity in the lists' penetration into the target population of objects to be counted, as well as heterogeneity in the catchability of individual objects, the modeling problem is inherently multi-level: there is a level of fixed effects for lists, and a level of random effects for objects.

We have also argued that the Rasch model, borrowed from the educational testing literature, is a natural place to start when seeking to model the multi-level structure of this problem. We may also easily incorporate additional multi-level structure into the Rasch model based on observed

Table 13: Estimates of the number of web pages of type "Query #535" from Section 5 and using quasi-symmetry models with list-by-list and list-by-total interactions.

Model	df	Deviance	Point Estimate	95% Interval
BIC[1]	46	65.62	602	[481, 793]
QS2[2]	55	88.61	614	[498, 778]
QS3	54	83.33	1266	[624, 3296]
QS4	53	82.43	508	[309, 5778]
QS5	52	81.71	861,882	[306, ∞]
QS2 + Dep[3]	41	59.14	598	[473, 799]
QS2 + BIC[4]	44	60.69	588	[468, 778]
QS3 + BIC[5]	44	55.99	1370	[650, 3829]
QS4 + BIC[6]	43	55.09	526	[309, 6570]
QS2 + $k_+ \times$AltaVista[7]	50	81.96	660	[474, 1003]
QS2 + $k_+ \times$Infoseek	50	83.29	810	[570, 1227]
QS2 + $k_+ \times$Excite	50	76.74	687	[514, 978]
QS2 + $k_+ \times$HotBot	50	80.70	591	[409, 993]
QS2 + $k_+ \times$Lycos	50	74.23	737	[558, 1021]
QS2 + $k_+ \times$NorthernLight	50	81.04	739	[507, 1175]
QS2 + $k_+ \times$Infoseek + $k_+ \times$Excite	46	68.09	857	[566, 1436]
QS2 + $k_+ \times$Excite + $k_+ \times$Lycos	46	68.35	756	[531, 1180]
QS2 + $k_+ \times$Infoseek + $k_+ \times$Lycos	46	64.36	829	[559, 1344]
QS2 + $k_+ \times$Infoseek + $k_+ \times$HotBot + $k_+ \times$Lycos	42	63.14	634	[393,1364]
Bayesian Rasch	– –	– –	685[8]	[554,1525]
Bayesian $k_+ \times$ Infoseek	– –	– –	685[8]	[531, 996]
Bayesian $k_+ \times$ Lycos	– –	– –	695[8]	[526, 1071]
Bayesian $k_+ \times$ Excite	– –	– –	666[8]	[559, 1319]

[1] Stepwise BIC selects: (AV:Is) + (AV:HB) + (AV:Ly) + (AV:NL) + (Is:Ex) + (Is:HB) + (Ex:NL) + (Ly:NL).

[2] QS2 indicates the Rasch quasi-symmetry model with no 3- or higher-way interactions. Similarly QS3 indicates Rasch quasi-symmetry with no 4- or higher-way interactions, etc.

[3] Dep indicates that all two-way interactions were also fitted, in addition to the quasi-symmetry model.

[4] Stepwise BIC starts with QS2 and adds AltaVista:Infoseek + AltaVista:Excite + AltaVista:HotBot + AltaVista:NorthernLight + Infoseek:Lycos + Infoseek:NorthernLight + Excite:HotBot + Excite:Lycos+ Excite:NorthernLight + HotBot:Lycos + HotBot:NorthernLight.

[5] Stepwise BIC starts with QS3 and adds AltaVista:Infoseek + AltaVista:Excite + AltaVista:HotBot + AltaVista:NorthernLight + Infoseek:NorthernLight + Excite:HotBot + Excite:Lycos + Excite: NorthernLight + HotBot:Lycos + HotBot:NorthernLight.

[6] Stepwise BIC starts with QS4 and adds AltaVista:Infoseek + AltaVista:Excite + AltaVista:HotBot + AltaVista:NorthernLight + Infoseek:NorthernLight + Excite:HotBot + Excite:Lycos + Excite:NorthernLight + HotBot:Lycos + HotBot:NorthernLight.

[7] $k_+ \times$AltaVista indicates one list-by-total interaction involving the AltaVista list. Similarly for the other $k_+ \times$List interaction models shown.

[8] Posterior mode.

Table 14: Sensitivity table for the diabetes data set.

	Priors on σ^2	
Priors on N	Inv-χ_1^2	$\Gamma^{-1}(1,1)$
$\pi(N) \propto 1$	2664 [2560,2917]	
$\pi(N) \propto \frac{1}{N}$		2693 [2567,2906]

Table 15: Sensitivity table for the world wide web data set.

	Priors on σ^2			
Priors on N	Inv-χ_1^2	$\Gamma^{-1}(1,2)$	$\Gamma^{-1}(1,1)$	$\Gamma^{-1}(5,0.05)$
$\pi(N) \propto 1$	680 [573,1684]		754 [569,1662]	
$\pi(N) \propto \frac{1}{N}$	721 [565,1498]	752 [587,1520]	685 [554,1525]	579 [464,961]

object- or list-covariates.

Application of the Rasch model to multiple-list census problems goes back at least to Sanathanan (1972), but only recently have we begun to understand how the Rasch model can be modified to accommodate list-by-list dependence and/or list-by-total interactions. We may incorporate these interactions directly into the likelihood that relates capture history to the random catchability effect which we view as a latent variable, or we may interpret them as a kind of stratification of the latent variable into multidimensional components by particular captures or lists. Biggeri, et al. (1999) have also tried to interpret list-by-total interactions as manifestations of a stratification of the latent variable by one or more captures, and this remains an interesting and active area of research. We have shown here how to convert these models into extensions of the log-linear quasi-symmetry model that has been associated with the Rasch model since at least Cressie and Holland (1983) and Fienberg and Meyer (1983). Frequentist analyses of these models using relatively standard GLM programs provide a useful first approximation to a fully Bayesian fit of the Rasch model to multiple-list data.

When the basic log-linear quasi-symmetry model holds, we have illustrated that the fully-Bayesian hierarchical formulation of the Rasch model provides at least as good a population total estimate, and does so more parsimoniously (exploiting a few hyperprior parameters rather than a full set of quasi-symmetry terms in the log-linear model). An important open question in comparing these two approaches is understanding the interplay between the Bayes model's relaxing of the "no highest-order interaction" assumption needed in the log-linear model to project an estimate onto the missing cell count in the $2^J - 1$ table cross-classifying list membership for all objects, and the Bayes model's imposition of moment constraints on the quasi-symmetry terms in the log-linear model that are usually not imposed in frequentist GLM fits of the model.

When the basic log-linear quasi-symmetry model does not hold, adding list-by-list or list-by-total interactions, as outlined in the previous section, can greatly improve the log-linear model fit and the population total estimates based on the log-linear models. These are naturally seen as log-linear manifestations of an underlying hierarchical Bayes model, and we demonstrated how to derive them as such. These models appeared to lend some additional flexibility to the Bayesian analyses of our examples and produced somewhat more stable estimates than the basic Bayesian Rasch model did.

References

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, **50**, 494–500.

Agresti, A. and Lang, J. B. (1993) Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*, **49**, 131–139.

Basu, S. (1998). Bayesian estimation of the number of undetected errors when both reviewers and errors are heterogeneous. Unpublished manuscript.

Biggeri, A., Stanghellini, E., Merletti, M. and Marchi, M. (1999). Latent class models for varying catchability and correlation among sources in Capture-Recapture estimation of the size of a human population. *Statistica Applicata* (to appear).

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, Mass.

Bruno, G., Laporte, R., Merletti, E., Biggieri, A., McCarty, D. and Pagano, C. (1994). National diabetes programs: Application of capture-recapture to “count” diabetes. *Diabetes Care*, **17**, 548–556.

Burnham, B. K. and Overton, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika,* **65,** 625–633.

Carriquiry, A. and Fienberg, S. E. (1998). Rasch models. In *Encyclopedia of Biostatistics*, **5**, Wiley, New York, 3724–3730.

Castledine, B. J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika,* **67,** 197-210.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, **45**, 427–438.

Chao, A., Lee, S.-M. and Jeng, S.-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, **48**, 201–216.

Cressie, N. and Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, **48**, 129–141.

Cormack, R. (1989). Log-linear models for capture-recapture. *Biometrics*, **45**, 395–413.

Cormack, R. (1992). Interval estimates for mark-recapture studies of closed populations. *Biometrics*, **48**, 567-576.

Coull, B. A., and Agresti, A. (1998). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, to appear.

Darroch, J. N. and Fienberg, S. E., and Glonek, G. F. V. and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, **88**, 1137–1148.

Darroch, J. N. and McCloud, P. I. (1990). Separating two sources of dependence in repeated influenza outbreaks. *Biometrika*, **77**, 237–243.

Fienberg, S. E. (1972). Multiple-recapture census for closed populations and incomplete contingency tables. *Biometrika*, **59**, 591–603.

Fienberg, S. E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, **18**, 143–154.

Fienberg, S. E. and Meyer, M. M. (1983). Loglinear models and categorical data analysis with psychometric and econometric applications. *Journal of Econometrics*, **22**, 191–214.

Freedman, P. R. (1972). Sequential estimation of the size of a population. *Biometrika*, **59**, 9–18.

Freedman, P. R. (1973). A numerical comparison between sequential tagging and sequential recapture. *Biometrika*, **60**, 499–508.

Garthwaite, P. H., and Yu, K. and Hope, P. B. (1995). Bayesian analysis of a multiple-recapture model. *Communications in Statistics, Part A–Theory and Methods*, **24**, 2229–2247.,

Geiger, H., and Werner, A. (1924). Die zahl der ion radium ausgesandsena-teilchen. *Zeitschrift fúr Physik*, **21**, 187–201.

George, E. I. and Robert, C. P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika,* **79,** 677-683.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics,* **41,** 337–348.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hay, G. (1997). The selection from multiple data sources in epidemiological capture recapture studies. *The Statistician*, **46**, 515–520.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, **55**, 577–601.

Holland, P. W. and Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, **14**,1523–1543.

International Working Group for Disease Monitoring and Forecasting (1995a). Capture-recapture and multiple-record systems estimation, I: History and theoretical development. *American Journal of Epidemiology*, **141**, 1047–1058.

International Working Group for Disease Monitoring and Forecasting (1995b). Capture-recapture and multiple-record systems estimation, II: Applications in human diseases. *American Journal of Epidemiology*, **141**, 1059–1088.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51,* 357–373.

Jannarone, R. J., Yu, K. F. and Laughlin, J. E. (1990). Easy Bayes estimation for Rasch-type models. *Psychometrika, 55,* 449–460.

Johnson, M. S., Cohen, W. M., and Junker, B. W. (1998). Measuring appropriability in research and development with item response models. Advanced Data Analysis Paper, Department of Statistics, Carnegie Mellon University, Pittsburgh PA.

Kadane, J. B., Meyer, M. M., and Tukey, J. W. (1999). Yule's association paradox and ignored stratum heterogeneity in capture-recapture studies. *Journal of the American Statistical Association*, to appear. [*Carnegie Mellon University Technical Report No.* **682**].

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, **49**, 223–245.

Lawrence, S. and Giles, C. L. (1998). Searching the world wide web. *Science*, **280**, 98–100.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, **63**, 215–232.

Madigan, D. and York, J. (1997) Bayesian methods for estimating the size of a closed population. *Biometrika*, **84**, 19-31.

Mathsoft, Inc. (1996). S-Plus Version 3.4 Release 1 for HP 9000 Series, HP-UX 9.0.5: Author.

Patz, R. J. and Junker, B. W. (1997). A straightforward approach to Markov chain Monte Carlo methods for item response models, *Journal of Educational and Behavioral Statistics,* to appear. [*Carnegie Mellon University Technical Report No.* **658**.]

Petersen, C. (1896). The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station to the Ministry of Fisheries*, **6**, 1–48.

Pollock, K. H. (1991). Modeling capture, recapture, and removal statistics for estimation of demographic parameters for fish and wildlife populations: Past, present, and future. *Journal of the American Statistical Association*, **86**, 225–238.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Niesen and Lydiche, Copenhagen. [expanded English edition (1980). University of Chicago Press.]

Roberts, H. V. (1967). Informative stopping rules and inferences about population size. *Journal of the American Statistical Association*, **62**, 763–775.

Sanathanan, L. (1972). Models and methods in visual scanning experiments. *Technometrics*, **14**, 813–830.

Sanathanan, L. (1973). A comparison of some models in visual scanning experiments. *Technometrics*, **15**, 67–78.

Schnabel, Z. (1938). The estimation of the total fish population of a lake. *American Mathematical Monthly*, **45**, 348–352.

Smith, P. J. (1988). Bayesian methods for multiple capture-recapture surveys. *Biometrics,* **44,** 1177–1189.

Smith, P. J. (1991). Bayesian analyses for a multiple capture-recapture model. *Biometrika,* **78,** 399–407.

Wu, M. L., Adams, R. J., and Wilson, M. R., (1997). *ConQuest: Generalized item response modeling software.* Australian Council on Educational Reasearch.

A Notation

Here is a summary of notation used in the paper.

- N objects in population, labelled $i = 1, \ldots, N$.
- J lists, labelled $j = 1, \ldots, J$.
- n objects have been found in the union of the lists so far.
- $N - n$ objects have not been observed.
- For individual i and list j, let

  $X_{ij} = 1$ if individual i on list j, considered as a random variable;

  x_{ij} are the observed values of X_{ij};

  $p_{ij} = P[X_{ij} = 1]$, the probability that individual i appears on list j;

- In the cross-classification of the j lists

  $k_j = 1$ or 0 to indicate presence or absence on list j;

  $(k_1, k_2, \cdots, k_J)$ indicates a pattern of presence/absence on the J lists, a "capture pattern" on the lists;

  $n_{k_1 k_2 \cdots k_J}$ is the number of individuals with capture pattern $k_1 k_2 \cdots k_J$, so that

  $$\sum_{k_1=0}^{1} \sum_{k_2=0}^{1} \cdots \sum_{k_J=0}^{1} n_{k_1 k_2 \cdots k_J} = n$$

  $\pi_{k_1 k_2 \cdots k_J}$ is the (multinomial) probability of appearing with capture pattern $(k_1, k_2, \cdots, k_J)$

  $X_j = 1$ or 0 to indicate presence or absence on list j of a randomly-chosen object under multinomial sampling;

  $m_{k_1 k_2 \cdots k_J} = E[n_{k_1 k_2 \cdots k_J}] = N\pi_{k_1 k_2 \cdots k_J}$.

  $\hat{m}_{odd}$, $\hat{m}_{even}$ MLE's of products (or products of MLE's) of $m_{k_1 k_2 \cdots k_J}$

- In the hierarchical Bayes formulation
  - θ is a random effect for catchability of objects;
  - β_j is a fixed effect for the penetration of list j into the population of objects.