Bayesian Methods for Disclosure Risk Assessment

Jon Forster
(University of Southampton, UK)

Sample data consists of values of categorical variables, recorded for each individual in the sample, expressed as a multiway contingency table.

$f_1, \ldots, f_K$ are cell counts in the released contingency table of key variables.

$F_1, \ldots, F_K$ are the corresponding population cell counts.

n and N are the sample and population totals respectively.

Record-level measures of disclosure risk

If E_j represents disclosure event in (sample non-empty) cell j

$$P(E_j|\boldsymbol{F}) = \frac{1}{F_j}$$

(Benedetti and Franconi, 1998)

Alternatively,

$$P(F_j = 1|\boldsymbol{F}) = I[F_j = 1]$$

is the probability of uniqueness.

Bayesian inference

The prior for the population is constructed as a parametric model

$$P(\boldsymbol{F}) = \int P(\boldsymbol{F}|\boldsymbol{\beta})P(\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta}$$

Then, for the unreleased population, assuming $\boldsymbol{F} - \boldsymbol{f} \perp\!\!\!\perp \boldsymbol{f}|\boldsymbol{\beta}$

$$P(\boldsymbol{F} - \boldsymbol{f}|\boldsymbol{f}) = \int P(\boldsymbol{F} - \boldsymbol{f}|\boldsymbol{\beta})P(\boldsymbol{\beta}|\boldsymbol{f})\ \mathrm{d}\boldsymbol{\beta}$$

where

$$P(\boldsymbol{\beta}|\boldsymbol{f}) \propto P(\boldsymbol{f}|\boldsymbol{\beta})P(\boldsymbol{\beta})$$

$$\text{Posterior} \propto \text{likelihood} \times \text{prior}$$

Bayesian disclosure risk assessment

We calculate Bayesian predictive probabilities as the posterior expectations

$$P(\text{event}|\boldsymbol{f}) = E[P(\text{event}|\boldsymbol{F})|\boldsymbol{f}]$$

Hence our risk measures become

$$P(E_j|\boldsymbol{f}) = E[1/F_j|\boldsymbol{f}]$$

and

$$P(F_j = 1|\boldsymbol{f}).$$

which are calculated using $P(\boldsymbol{F} - \boldsymbol{f}|\boldsymbol{f})$.

Bayesian inference for contingency tables

$\boldsymbol{F}$ has a multinomial(N, π) distribution.

π has a Dirichlet distribution (Takemura, 1999, Omori, 1999)

Related to Poisson-Gamma models and Negative binomial models (Bethlehem et al, 1990, Benedetti and Franconi, 1998, Rinott, 2003, Polettini and Stander, 2004).

Underlying cell exchangeabilty $\Rightarrow$ similar inference for any two sample uniques.

Highly restricted 'borrowing of strength' across cells.

Log-linear models

A log-linear model for π utilises the natural (multivariate) structure of the data.

$$\log \pi = \boldsymbol{X}_m \boldsymbol{\beta}_m$$

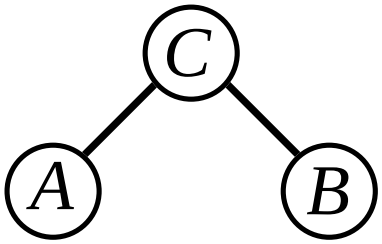
Prior is completed by adding a further hierarchy, consisting of a prior distribution for $\boldsymbol{\beta}_m$.

Skinner and Holmes (1998), Elamir and Skinner (2004) adapt the Poisson-Gamma approach to a log-linear model.

But which model is best?

Let the data 'choose' incorporating uncertainty when present.

Decomposable graphical models and the hyper-Dirichlet



A is independent of B given C, so that

$$P(A = i \text{ and } B = j|C = k) = P(A = i|C = k)P(B = j|C = k)$$

or

$$\pi_{ijk} = \pi^A_{i|k}\ \pi^B_{j|k}\ \pi^C_k$$

Independent Dirichlet priors for $\{\pi^A_{i|k}\}$, $\{\pi^B_{j|k}\}$, for each k, and for $\{\pi^C_k\}$.

Tractable computation of $P(\boldsymbol{F} - \boldsymbol{f}|\boldsymbol{f})$, and associated functions.

Bayesian inference under model uncertainty

Model search and uncertainty is incorporated through a discrete prior (posterior) distribution $p(m)$ over the models $m \in M$, some set of log-linear models.

Then, posterior predictive expectations of any function of $\boldsymbol{F}$ will be a *model average*

$$E[g(\boldsymbol{F})|\boldsymbol{f}] = \sum_{m \in M} P(m|\boldsymbol{f}) E[g(\boldsymbol{F})|\boldsymbol{f}, m].$$

where by Bayes theorem

$$P(m|\boldsymbol{f}) = \frac{P(m)P(\boldsymbol{f}|m)}{\sum_{m \in M} P(m)P(\boldsymbol{f}|m)}$$

and $P(\boldsymbol{f}|m) = \int P(\boldsymbol{f}|m, \boldsymbol{\beta}_m) P(\boldsymbol{\beta}_m|m) \mathrm{d}\boldsymbol{\beta}_m$.

Computation

Computational difficulties

1. Evaluating integrals – may be mathematically intractable
2. Number of models is large.
3. Number of possible values of (multivariate) $\boldsymbol{F}$ is large.

Details in the paper.

Example

Six potential key variables from the 3% Individual SAR for the 2001 UK Census (`http://www.ccsr.ac.uk/sars/2001`).

Restricted to 154295 individuals living in South West England

Sex (2 categories)
Age (coded into 11 categories)
Accomodation type (8 categories)
Number of cars owned or available for use (5 categories)
Occupation type (11 categories)
Family type (10 categories)

The full table has 96800 cells of which 3796 are uniques.

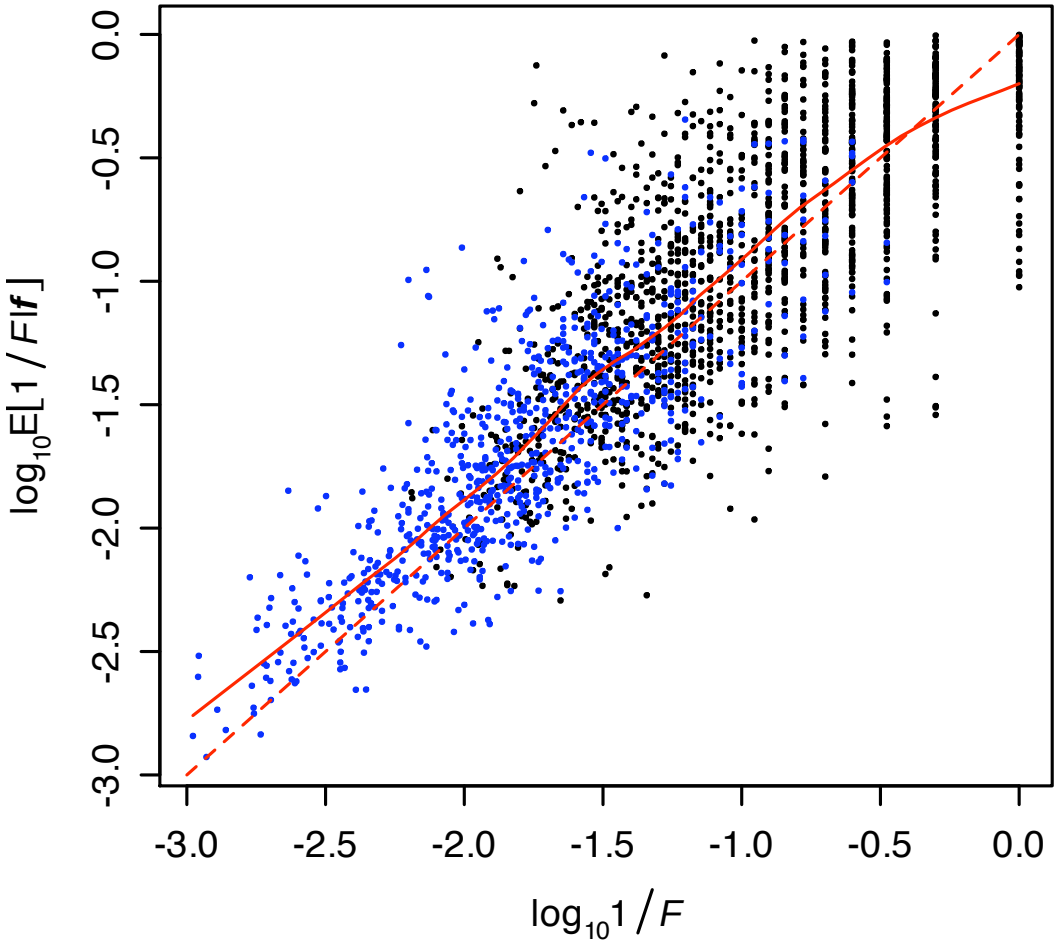
This is our 'population', from which we took a 3% subsample.

Sample data contains 4761 individuals in 2330 cells.

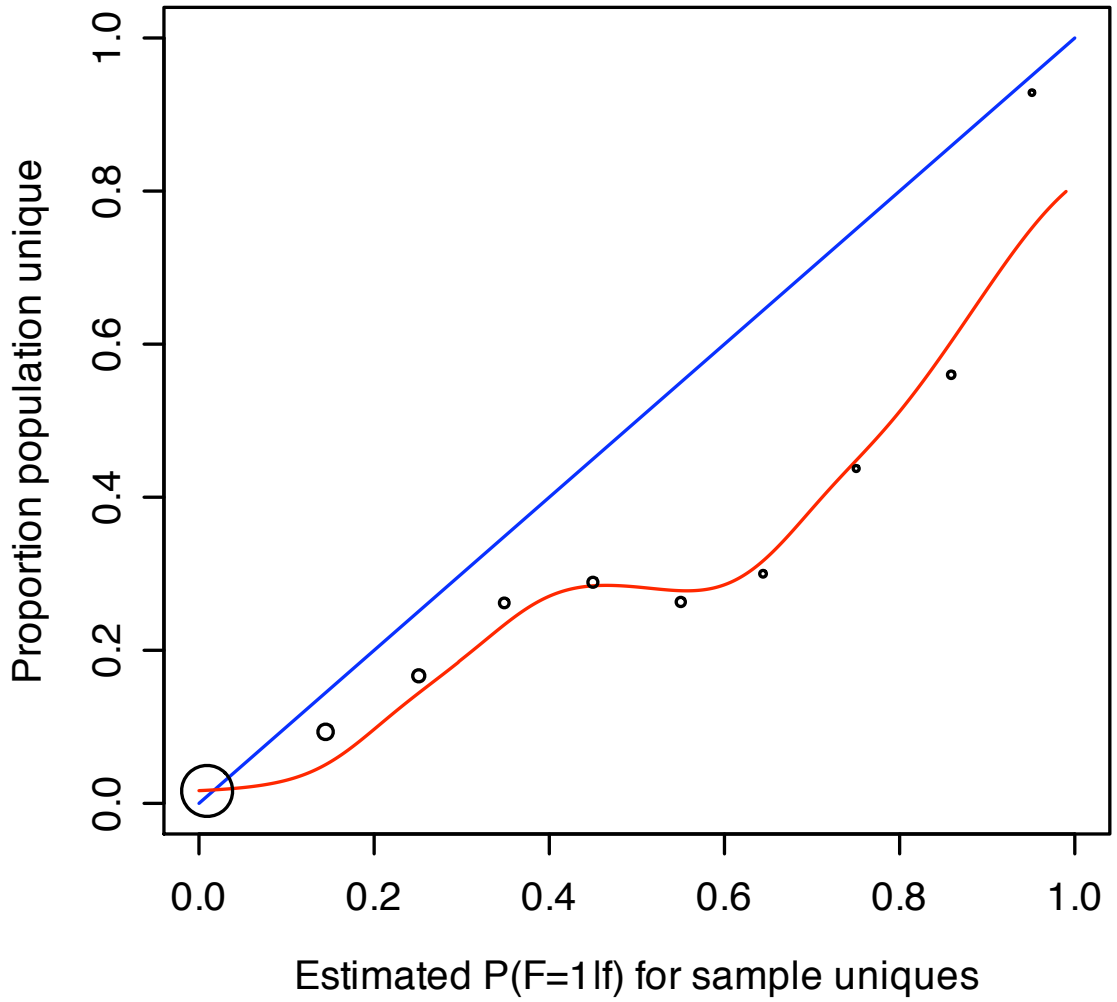
1543 (32%) are uniques, of which 114 (7%) are population uniques. Average population total in a sample unique cell is 17.

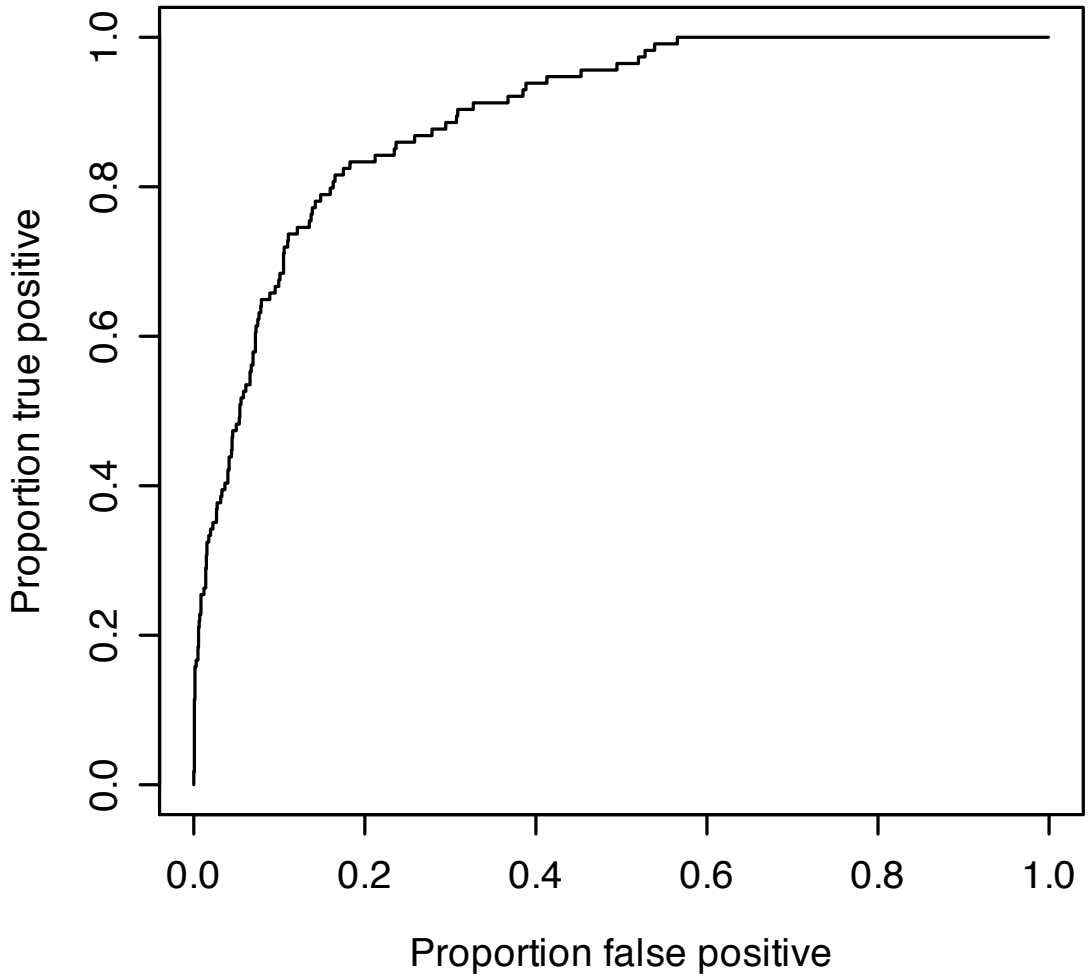
		Population								
		0	1	2	3	4	5-9	10-19	20+	Total
Sample	0	84867	3682	1694	967	631	1482	757	390	94470
	1	—	**114**	**110**	**118**	**104**	**313**	**322**	**462**	**1543**
	2	—	—	0	2	5	28	67	266	368
	3	—	—	—	0	0	1	15	140	156
	4	—	—	—	—	0	0	0	76	76
	5-9	—	—	—	—	—	0	0	125	125
	10-19	—	—	—	—	—	—	0	48	48
	20+	—	—	—	—	—	—	—	14	14
	Total	84867	3796	1804	1087	740	1824	1161	1521	96800

Estimated v. True Disclosure Risk



True v. Estimated Probability of Uniqueness




ROC curve for uniqueness detection
Proportion true positive
0.0
0.2
0.4
0.6
0.8
1.0
Proportion false positive
0.0
0.2
0.4
0.6
0.8
1.0