FUZZY-SET SAMPLING UNDER SEMI-DEFINITE CHAOS

Jimmy Thomas Efird
1430 Massachusetts Ave., Suite 306-51, Cambridge, MA 02138

Introduction

Estimating the population prevalence of a binary response variable when sample information is incomplete or inconsistent, poses immense problems to the applied statistician. Traditionally, the analyst simply deletes these values and treats the remaining data as a representative sample from the population in question. Providing that stable covariate fields exist and between-variable correlation is zero, then multivariate ANOVA and Chi-square techniques will yield a quick check of this strategy. However, if sampling irregularities plague covariate fields or confounding presents a problem, then the analyst must seek other means of verification. One possible solution, given the availability of population information, is to construct a test statistic using a population projection technique known as Bayesian Logistic Regression.

Bayesian Logistic Regression

Bayesian Logistic Regression is an estimation procedure that aims to eliminate within covariate group variance by utilizing the interrelational structure between the sample response variable and known population covariates. A new response vector with elements corresponding to the expected value of the underlying covariate patterns form the basis of Bayesian Logistic Regression. The expected values result from a bayesian expansion of the outcome variable jointly conditional on the set of model covariates.

Sampling begins by randomly selecting variables from a population data base. Aiming for a computationally ideal model having three covraiates, the next step involves reducing data set dimensionality via Stepwise Regression. Expressing in mathematical terms, let $T_n = (u, v, x, y, z)$ represent a n-observation sample, where u denotes a binary outcome variable with three model covariates v, x and y. The z-covariate represents a model-free population field having a relatively stable multilevel element space. On the other hand, the situation assumes instabilities exist among one or more of the model covariates v, x and y. Indexing as follows,

$i = 0, 1$

$j, k, h, m = 1, 2, 3, \ldots, (v_n, x_n, y_n, z_n)$

$z_n \rightarrow$ fixed population field

$s \rightarrow$ stable vector field

$c \rightarrow$ fuzzy-set/chaotic vector field

the expected response for a covariate pattern missing two elements (e.g. only the value for v is present) may be written as:

$$\Lambda_j = \hat{p}[\mu=i \mid v_c=j]$$

$$= \sum_{m=1}^{n_{z_s}} \hat{p}[\mu=i \mid v_c=j, z_s=m] \bullet \frac{\hat{p}[v_c=j \mid z_s=m] \bullet \hat{p}[z_s=m]}{\sum_{m=1}^{n_{z_s}} \hat{p}[v_c=j \mid z_s=m] \bullet \hat{p}[z_s=m]}$$

$$= \sum_{m=1}^{n_{z_s}} \hat{p}[v_c=j \mid \mu=i, z_s=m] \bullet \hat{p}[z_s=m \mid \mu=i] \bullet \hat{p}[\mu=i] \frac{\left[\frac{\hat{p}[z_s=m \mid v_c=j] \bullet \hat{p}[v_c=j]}{\sum_{j=1}^{n_{v_s}} \hat{p}[z_s=m \mid v_c=j] \bullet \hat{p}[v_c=j]} \bullet \hat{p}[z_s=m]\right]}{\sum_{m=1}^{n_{z_s}} \left[\frac{\hat{p}[z_s=m \mid v_c=j] \bullet \hat{p}[v_c=j]}{\sum_{j=1}^{n_{v_s}} \hat{p}[z_s=m \mid v_c=j] \bullet \hat{p}[v_c=j]}\right]}$$

Analogously,

$$\Lambda_{k,h} = \hat{p}[\mu=i \mid x_c=k, y_c=h]$$

$$= \sum_{m=1}^{n_{z_s}} \hat{p}[z_s=m \mid x_c=k, y_c=h] \bullet \hat{p}[\mu=i \mid x_c=k, y_c=h, z_s=m]$$

$$= \sum_{m=1}^{n_{z_s}} \frac{\hat{p}[x_c=k \mid y_c=h, z_s=m] \bullet \frac{\hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]}{\sum_{h=1}^{n_{y_c}} \hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]}}{\sum_{m=1}^{n_{z_s}} \hat{p}[x_c=k \mid y_c=h, z_s=m] \bullet \frac{\hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]}{\sum_{h=1}^{n_{y_c}} \hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]}} \bullet \hat{p}[\mu=i \mid x_c=k, y_c=h, z_s=m]$$

$$= \sum_{m=1}^{n_{z_s}} \gamma_1 \bullet \gamma_2 \bullet \left[\sum_{m=1}^{n_{z_s}} \gamma_1\right]^{-1}$$

where

$$\gamma_1 = \gamma_3 \bullet \gamma_4 \bullet \left[\sum_{k=1}^{n_{x_c}} \gamma_3\right]^{-1}$$

$$\gamma_2 = \frac{\hat{p}[x_c=k \mid \mu=i, y_c=h, z_s=m] \bullet \hat{p}[y_c=h \mid \mu=i, z_s=m] \bullet \hat{p}[z_s=m \mid \mu=i] \bullet \hat{p}[\mu=i]}{\sum_{i=1}^{n_\mu} \hat{p}[x_c=k \mid \mu=i, y_c=h, z_s=m] \bullet \hat{p}[y_c=h \mid \mu=i, z_s=m] \bullet \hat{p}[z_s=m \mid \mu=i] \bullet \hat{p}[\mu=i]}$$

$$\gamma_3 = \frac{\hat{p}[x_c=k \mid y_c=h, z_s=m] \bullet \hat{p}[y_c=h \mid z_s=m] \bullet \hat{p}[z_s=m]}{\sum_{k=1}^{n_{x_c}} \hat{p}[x_c=k \mid y_c=h, z_s=m] \bullet \hat{p}[y_c=h \mid z_s=m] \bullet \hat{p}[z_s=m]} \bullet \hat{p}[z_s=m \mid x_c=k] \bullet \hat{p}[x_c=k]$$

$$\gamma_4 = \frac{\hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]}{\sum_{h=1}^{n_{y_c}} \hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]}$$

and

$$\Lambda_{j,k,h} = \hat{p}[\mu=i \mid v_c=j, x_c=k, y_c=h]$$

$$= \sum_{m=1}^{n_{z_s}} \hat{p}[\mu=i \mid v_c=j, x_c=k, y_c=h, z_s=m] \bullet \frac{\hat{p}[v_c=j \mid x_c=k, y_c=h, z_s=m] \bullet \hat{p}[z_s=m \mid x_c=k, y_c=h]}{\sum_{m=1}^{n_{z_s}} \hat{p}[v_c=j \mid x_c=k, y_c=h, z_s=m] \bullet \hat{p}[z_s=m \mid x_c=k, y_c=h]}$$

Simplifying the above expression, $\Lambda_{j,k,h}$ becomes $\sum_{m=1}^{n_{z_s}} \frac{\xi_1 \bullet \xi_2 \bullet \xi_3 \bullet \left[\sum_{m=1}^{n_{z_s}} \xi_3\right]^{-1}}{\sum_{m=1}^{n_{z_s}} \xi_2 \bullet \xi_3 \bullet \left[\sum_{m=1}^{n_{z_s}} \xi_3\right]^{-1}}$

where ξ_1 is given by

$$\frac{\hat{p}[v_c=j|\mu=i,x_c=k,y_c=h,z_s=m] \bullet \hat{p}[y_c=h|\mu=i,x_c=k,z_s=m] \bullet \hat{p}[z_s=m|\mu=i,x_c=k] \bullet \hat{p}[x_c=k|\mu=i] \bullet \hat{p}[\mu=i]}{\sum_{i=1}^{n_\mu} \hat{p}[v_c=j|\mu=i,x_c=k,y_c=h,z_s=m] \bullet \hat{p}[y_c=h|\mu=i,x_c=k,z_s=m] \bullet \hat{p}[z_s=m|\mu=i,x_c=k] \bullet \hat{p}[x_c=k|\mu=i] \bullet \hat{p}[\mu=i]},$$

$$\xi_2 = \frac{\hat{p}[x_c=k \mid v_c=j, y_c=h, z_s=m] \bullet \hat{p}[y_c=h \mid v_c=j, z_s=m] \bullet \hat{p}[z_s=m \mid v_c=j] \bullet \hat{p}[v_c=j]}{\sum_{j=1}^{n_{v_c}} \hat{p}[x_c=k \mid v_c=j, y_c=h, z_s=m] \bullet \hat{p}[y_c=h \mid v_c=j, z_s=m] \bullet \hat{p}[z_s=m \mid v_c=j] \bullet \hat{p}[v_c=j]}$$

$$\xi_3 = \xi_4 \bullet \left[\sum_{k=1}^{n_{x_c}} \xi_4\right]^{-1} \bullet \xi_5$$

$$\xi_4 = \frac{\hat{p}[x_c=k \mid y_c=h, z_s=m] \bullet \hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]}{\sum_{h=1}^{n_{y_c}} \hat{p}[x_c=k \mid y_c=h, z_s=m] \bullet \hat{p}[z_s=m \mid y_c=h] \bullet \hat{p}[y_c=h]} \bullet \hat{p}[z_s=m \mid x_c=k] \bullet \hat{p}[x_c=k]$$

$$\xi_5 = \hat{p}[y_c=h \mid z_s=m] \bullet \hat{p}[z_s=m]$$

Replacing $(x_c=k, y_c=h)$ with $(v_c=j, x_c=k)$ or $(v_c=j, y_c=h)$ gives the equivalent forms $\Lambda_{y,k}$ and $\Lambda_{y,h}$, while Λ_j is similarly expressed as Λ_k and Λ_h upon substituting $(x_c=k)$ or $(y_c=h)$ for $(v_c=j)$.

A Homogeneity Test for Missing Response Elements

The statistical validity of discarding observations having missing or inconsistent response outcomes holds only if the true distribution of these items is indistinguishable from the set on non-missing elements. Unfortunately, routine testing of this assumption is often impossible due to covariate non-response and confounding.

An alternative approach involves partitioning the new response vector into two sets, depending upon whether or not the original response is missing. A test of homogeneity of the two groups, focusing on extreme deviation rather than on central tendency as a test criteria, is made possible by collapsing distinct covariate responses into single value blocks and subjecting these points to analysis using the parameter-free discrete range test. Designating non-missing element blocks as type-A and missing element blocks as type-B, combine values and rank accordingly, noting the relative position of each element block. Assuming element blocks to be numbered from 1 to N, define the range as the number on the highest type-B element block

minus the lowest type-B element. An unusually high or low range value raises suspicion that the two groups differ with respect to their tail regions. For illustrative purposes, consider the simple case where the number of type-A blocks (n_A) equals type-B blocks (n_B), e.g. $N=n_A+n_B=2n$. Letting ζ designate the range of type-B blocks among type-A blocks and defining

$$\Omega_i = \left\{ 1 - \left[1 - \sum_{\xi=1}^{\substack{i-1 \\ 3\le i\le 2}} \frac{\xi \cdot \prod_{z=0}^{1}(n-z) \cdot \prod_{\beta=\mathrm{int}(\xi/2)+1}^{\substack{\xi-2 \\ 3\le\xi\le 5}}(n-\beta)}{2^{\mathrm{int}(\xi/2)+1} \cdot \prod_{\delta=\xi}^{\substack{\xi \\ 3\le\xi\le 2}}(n-\delta-1) \cdot \prod_{v=0}^{\mathrm{int}\left(\frac{\xi+1}{2}\right)-1}[2(n-v)-1]} \right]^{-1} \cdot \left[\frac{i \cdot \prod_{z=0}^{1}(n-z) \cdot \prod_{\beta=\mathrm{int}(i/2)+1}^{\substack{i-2 \\ 3\le i\le 5}}(n-\beta)}{2^{\mathrm{int}(i/2)+1} \cdot \prod_{\delta=i}^{\substack{i \\ 3\le i\le 2}}(n-\delta-1) \cdot \prod_{v=0}^{\mathrm{int}\left(\frac{i+1}{2}\right)-1}[2(n-v)-1]} \right]^{-1} \right\}^{-1}$$

it follows that

$$\lim_{N\to\infty} P\{\zeta \le N-i-1\} = \lim_{N\to\infty} P\{\zeta \le N-i\} / \lim_{N\to\infty} \Omega_i .$$

Solving for i gives the probabilites and rejection regions listed in Tables 1 and 2. Proceeding in an analogous fashion will yield test values in the unequal group case. In Table 2, if ζ exceed ζ_0 for the indicated sample size then reject H_0 with Type-I error ≤ alpha (.05, .01).

Simulation

Preliminary simulation suggests that the above procedure performs well when both the number of elements within each distinct block and the overall number of blocks in each group remains high. This situation generally holds in large-scale sampling situation where the base population is greater than 100,000 with an initial sample greater than 10%.

Table 1: Exact Asymptotic Probabilities for the Range Distribution

i	$\lim_{N\to\infty} P\{\zeta \le N-i\}$	$\lim_{N\to\infty} P\{\zeta = N-i\}$
1	1.00000000	0.25000000
2	0.75000000	0.25000000
3	0.50000000	0.18750000
4	0.31250000	0.12500000
5	0.18750000	0.07812500
6	0.10937500	0.04687500
7	0.06250000	0.02734380
8	0.03515630	0.01562500
9	0.01953130	0.00878906
10	0.01074220	0.00488281
11	0.00585938	*********

Table 2: Rejection Regions for the Two-Sample ($n_1=n_2$) Range Test

Sup ζ_0: $P\{\zeta \le \zeta_0\} \le .05$	Sup ζ_0: $P\{\zeta \le \zeta_0\} \le .01$
$n \le 4$, $\zeta_0 = \{\emptyset\}$	$n \le 5$, $\zeta_0 = \{\emptyset\}$
$5 \le n \le 6$, $\zeta_0 = N - 6$	$n = 6$, $\zeta_0 = N - 7$
$7 \le n \le 26$, $\zeta_0 = N - 7$	$7 \le n \le 10$, $\zeta_0 = N - 8$
$n \ge 27$, $\zeta_0 = N - 8$	$11 \le n \le 20$, $\zeta_0 = N - 9$
	$21 \le n \le 204$, $\zeta_0 = N - 10$
	$n \ge 205$, $\zeta_0 = N - 11$

REFERENCES

Chernoff, H. and Savage, I.R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics* 29: 972-994.

Csorgo, M. and Guttman, I. (1962). On the empty cell test. *Technometrics* 4: 235-247.

David, F.N. (1950). Two combinatorial tests of whether a sample has come from a given population. *Biometrika* 30: 7-111.

David, F.N. and Barton, D.E. (1962). *Combinatorial Chance*. Hafner Publishing Company.

Efird, J.T. (1987). A Wald-Wolfowitz type test sensitive to outliers based on F.N. David's 'Ball and Cell' range statistic. *Joint Statistical Meeting, San Francisco, CA.*

Halperin, M. and Burrow, G.L. (1961). An asymptotic distribution for an occupancy problem with statistical applications. *Technometrics* 3: 79-86.

Jogdeo, K. (1968). Asymptotic normality in nonparametric methods. *Annals of Mathematical Statistics* 39: 905-922.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag.

Mood, A.M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. *Annals of Mathematical Statistics* 25: 514-522.

Mood, A., Graybill, F.A. and Boes, D.C. (1974). *Introduction to the Theory of Statistics*. McGraw-Hill, Inc.

Nicholson, W.L. (1961). Occupancy probability distribution critical points. *Biometrika* 58: 175-180.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics* 2: 147-162.