STATISTICS IN MEDICINE
Statist. Med. 2005; **1**:1–29 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Addendum: Bayesian analysis of paired count data using the Poisson difference distribution

D. Karlis *and I. Ntzoufras

Department of Statistics, Athens University of Economics and Business
76, Patission Str., 10434, Athens, Greece

SUMMARY

Paired count data usually arise in medicine when before and after treatment measurements are considered. In the present paper we assume that the correlated paired count data follow a bivariate Poisson distribution in order to derive the distribution of their difference. The derived distribution is shown to be the same as the one derived for the difference of the independent Poisson counts, thus recasting interest on the distribution introduced by Skellam (1946). Using this distribution we remove correlation, which naturally exists in paired data, and we improve the quality of our inference by using exact distributions instead of normal approximations. The zero-inflated version is considered to account for an excess of zero counts. Bayesian estimation and hypothesis testing for the models considered are discussed. An example from dental epidemiology is used to illustrate the proposed methodology. Copyright © 2005 John Wiley & Sons, Ltd.

7. Chib's 'Marginal Likelihood' method for Poisson Difference Data

7.1. Calculation of Marginal Likelihood

The calculation of marginal likelihood is based on the equation

$$f(\boldsymbol{y}|m) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m)}{f(\boldsymbol{\theta}_m|\boldsymbol{y},m)}$$

which holds for every $\boldsymbol{\theta}$. Here we consider $\boldsymbol{\theta}^*$ as the posterior mode or mean. Hence the above can be rewritten as

$$\log f(\boldsymbol{y}|m) = \log f(\boldsymbol{y}|\boldsymbol{\theta}^*_m) + \log f(\boldsymbol{\theta}^*_m) - \log f(\boldsymbol{\theta}^*_m|\boldsymbol{y},m).$$

The quantities $f(\boldsymbol{y}|\boldsymbol{\theta}^*_m)$ and $f(\boldsymbol{\theta}^*_m)$ can be calculated directly but $f(\boldsymbol{\theta}^*_m|\boldsymbol{y},m)$ should be estimated by the MCMC output.

*Correspondence to: Department of Statistics, Athens University of Economics and Business, 76, Patission Str., 10434, Athens, Greece, email: karlis@aueb.gr

Received 17 May 2004
Revised ?? February 2005

In our example we have $f(\theta_1^*, \theta_2^*, p^*|\boldsymbol{y})$ which can be given by

$$f(\theta_1^*, \theta_2^*, p^*|\boldsymbol{y}) = f(\theta_1^*|\theta_2^*, p^*, \boldsymbol{y}) f(\theta_2^*|p^*, \boldsymbol{y}) f(p^*|\boldsymbol{y})$$

which can be estimated by

$$f(\theta_1^*|\theta_2^*, p^*, \boldsymbol{y}) = G^{-1} \sum_{g=1}^{G} Gamma(\theta_1^*; n\bar{v}^{(g)} - S_{\delta v}^{(g)} + a_1, n - n\bar{\delta}^{(g)} + b_1) \quad (1)$$

$$f(\theta_2^*|p^*, \boldsymbol{y}) = G^{-1} \sum_{g=1}^{G} Gamma(\theta_2^*; n\bar{u}^{(g)} - S_{\delta u}^{(g)} + a_2, n - n\bar{\delta}^{(g)} + b_2) \quad (2)$$

$$f(p^*|\boldsymbol{y}) = G^{-1} \sum_{g=1}^{G} Beta(p^*; n\bar{\delta}^{(g)} + a_3, n - n\bar{\delta}^{(g)} + b_3) \quad (3)$$

where $S_{\delta v}^{(g)} = \sum_{i=1}^{n} \delta_i^{(g)} v_i^{(g)}$ and. $S_{\delta u}^{(g)} = \sum_{i=1}^{n} \delta_i^{(g)} u_i^{(g)}$.

Analytically, for each model we have

Model 1: $f(\boldsymbol{\theta}_{m_1}^*|\boldsymbol{y}) = f(\theta^*|\boldsymbol{y}) = G^{-1} \sum_{g=1}^{G} Gamma(\theta^*; n\bar{v}^{(g)} + n\bar{u}^{(g)} + a, 2n + b)$.

Model 2: $f(\boldsymbol{\theta}_{m_2}^*|\boldsymbol{y}) = f(\theta_1^*, \theta_2^*|\boldsymbol{y}) = f(\theta_1^*|\theta_2^*, \boldsymbol{y}) f(\theta_2^*|\boldsymbol{y})$
$= G^{-1} \sum_{g=1}^{G} Gamma(\theta_1^*;\ \ n\bar{v}^{(g)} - S_{\delta v}^{(g)} + a_1,\ \ n - n\bar{\delta}^{(g)} + b_1)$
$\times G^{-1} \sum_{g=1}^{G} Gamma(\theta_2^*;\ \ n\bar{u}^{(g)} - S_{\delta u}^{(g)} + a_2,\ \ n - n\bar{\delta}^{(g)} + b_2)$

Model 3: $f(\boldsymbol{\theta}_{m_3}^*|\boldsymbol{y}) = f(\theta^*, p^*|\boldsymbol{y}) = f(\theta_1^*|p^*, \boldsymbol{y}) f(p^*|\boldsymbol{y})$
$= G^{-1} \sum_{g=1}^{G} Gamma(\theta_1^*;\ \ n\bar{v}^{(g)} + n\bar{u}^{(g)} - S_{\delta v}^{(g)} - S_{\delta u}^{(g)} + a,\ \ 2n - 2n\bar{\delta}^{(g)} + b)$
$\times G^{-1} \sum_{g=1}^{G} Beta(p^*;\ \ n\bar{\delta}^{(g)} + a_3,\ \ n - n\bar{\delta}^{(g)} + b_3)$.

Model 4: $f(\boldsymbol{\theta}_{m_4}^*|\boldsymbol{y}) = f(\theta_1^*, \theta_2^*, p^*|\boldsymbol{y}) = f(\theta_1^*|\theta_2^*, p^*, \boldsymbol{y}) f(\theta_2^*|p^*, \boldsymbol{y}) f(p^*|\boldsymbol{y})$
$= G^{-1} \sum_{g=1}^{G} Gamma(\theta_1^*;\ \ n\bar{v}^{(g)} - S_{\delta v}^{(g)} + a_1,\ \ n - n\bar{\delta}^{(g)} + b_1)$
$\times G^{-1} \sum_{g=1}^{G} Gamma(\theta_2^*;\ \ n\bar{u}^{(g)} - S_{\delta u}^{(g)} + a_2,\ \ n - n\bar{\delta}^{(g)} + b_2)$
$\times G^{-1} \sum_{g=1}^{G} Beta(p^*;\ \ n\bar{\delta}^{(g)} + a_3,\ \ n - n\bar{\delta}^{(g)} + b_3)$.

Prepared using simauth.cls

Table I. Comparison of Estimated $\log B_{21}$ Using RJMCMC and Chib's Marginal Likelihood Approach; $\log B_{21}$: log Bayes factor of m_2 vs m_1, *Monte Carlo Error*: estimated by the standard deviation of the $\log B_{21}$ for 50 sub-samples, *In brackets*: number of total iterations kept.

	Estimated $\log B_{21}$			Monte Carlo Error		
	RJMCMC	Chib's	Chib's	RJMCMC	Chib's	Chib's
School	(10,000)	(10,000)	(25,000)	(10,000)	(10,000)	(25,000)
1	37.96	38.51	38.61	0.19	1.21	0.48
2	20.00	20.59	20.55	0.14	1.25	0.43
3	31.49	32.00	31.92	0.20	0.58	0.35
4	14.69	15.52	15.48	0.14	0.70	0.33
5	40.31	40.92	40.88	0.13	0.74	0.39
6	17.87	18.32	18.36	0.11	0.54	0.37
All	168.60	168.88	169.07	0.13	0.92	0.39

7.2. Results from the Comparison of $PD(\theta_1, \theta_2)$ and $PD(\theta, \theta)$ for DMFT data.

In order to compare the efficiency of RJMCMC and Chib's marginal likelihood approach we have calculated the logarithm of B_{21} using output of length 5000 iterations and additional 500 burn-in each (resulting to 10,000 iterations in total which is equal to the number of iterations we considered in RJMCMC). We have separated the output in 50 batches and estimated in each of them the $\log B_{21}$ using the Chib's marginal likelihood approach. The standard deviation of the estimated quantities in each batch measures the Monte Carlo error. This is directly comparable to the corresponding Monte Carlo error we have calculated using RJMCMC. In the Table which follows we clearly see that Chib's method has considerably higher Monte Carlo error (about 3–9 times higher). Finally, we have calculated Monte Carlo error using 12,500 iterations with additional 2,500 iterations as burn-in (which mean 25,000 iterations kept in total) but still Monte Carlo error was found higher than the corresponding error of RJMCMC with 10,000 iterations.