Correspondence: Using recently developed software on a 2×2 table of matched pairs with incompletely classified data

P. M. E. Altham and Robin K. S. Hankin
University of Cambridge

Abstract

Recent work by Lin, Lipsitz, Sinha, Gawande, Regenbogen, and Greenberg proposed a Bayesian analysis of a 2×2 table including incompletely classified data. Here, we subject the same dataset to further analysis using recently developed techniques and software written in the R programming language.

This vignette is based on a manuscript submitted to the Journal of the Royal Statistical Society Series C, as correspondence.

For reasons of performance, this vignette uses a preloaded dataset ('here's one I prepared earlier'). To calculate the dataset from scratch, set variable `calculate_from_scratch` to `TRUE` in the first chunk.

Keywords: Aylmer test, computational combinatorics, R, hyperdirichlet distribution, Bayesian analysis.

Lin *et al.* (2009) propose a Bayesian analysis of an interesting dataset which included incompletely classified data; they extend a result published by Altham (1971). Here, we subject the same dataset to further analysis using recently developed techniques and software written in the R programming language (R Development Core Team 2008).

The dataset is given here for convenience as Table 1. It arises from 69 medical malpractice claims, and are the two Surgeon Reviewers' answers to the question: was there a communication breakdown in the hand-off between physicians caring for the patient? The rows of the Table correspond to the answers given by Reviewer 1, and the columns to the answers given by Reviewer 2.

Following Lin *et al.* (2009), we adopt the notation given in Table 2 for the corresponding observed frequencies.

We now assess whether Reviewer 2 is giving significantly higher proportion of 'Yes' responses than is Reviewer 1. Although the McNemar test is applicable to the 2×2 table of complete observations [the exact one-sided p-value is $\frac{7}{2^6} \simeq 0.1094$], we suggest using the 'Aylmer test' (West and Hankin 2008). The **aylmer** R package is available at CRAN, `http://cran.r-project.org/`.

The Aylmer test is a generalization of the Fisher Exact test which allows for the possibilities of structural zeros; the figure in the third row, third column of table 1 is effectively a structural zero because we are not interested in cases not missed by both reviewers.

Reviewer 1	Reviewer 2			
	Yes	No	Missing	Total
Yes	26	1	2	29
No	5	18	9	32
Missing	4	4	0	8
Total	35	23	11	69

Table 1: Two surgeon reviews of malpractice claims data

Reviewer 1	Reviewer 2			
	Yes	No	Missing	Total
Yes	y_{11}	y_{10}	z_{1+}	$y_{1+} + z_{1+}$
No	y_{01}	y_{00}	z_{0+}	$y_{0+} + z_{0+}$
Missing	u_{+1}	u_{+0}	0	u_{++}
Total	$y_{+1} + u_{+1}$	$y_{+0} + u_{+0}$	z_{++}	n

Table 2: Notation for the data

In this case, the statistic of interest is the difference between row 1, column 2 and row 2, column 1 (=5-1=4):

```
> a

          Reviewer 2
Reviewer 1 yes no missing
   yes      26  1       2
   no        5 18       9
   missing   4  4      NA

> aylmer.test(a, alternative = function(x) x[1, 2] - x[2, 1])

        Aylmer functional test for count data

data:  a
p-value = 0.1690
alternative hypothesis: test function exceeds observed
```

and thus the p-value of 0.169 would indicate failure to reject the null hypothesis. We suggest our analysis is superior to the McNemar test because it does not disregard the partially classified data points.

Cases missing at random

Lin *et al.* assess the hypothesis that the cases are missing at random, and use Fisher's exact test in a 'somewhat informal way' to compare the marginal proportions of the 2×2 table

of complete cases with the marginal proportions in those missing a row or column variable; a p-value of 0.046 is reported.

We again suggest using the Aylmer test. In this case, the statistic of interest is row 2, column 3, which corresponds to the number of cases missed by reviewer 2 but were classified as "no" (as opposed to "yes") by reviewer 1. The R idiom is straightforward:

```
> f1 <- function(a) a[2, 3]

> aylmer.test(a, alternative = f1)

        Aylmer functional test for count data

data:  a
p-value = 0.03202
alternative hypothesis: test function exceeds observed
```

The resulting p-value is 0.032, considerably lower than that from the Fisher test; this is consistent with the Aylmer test's using more data than Chen and Little.

The hyperdirichlet distribution

The likelihood function of the data D may be taken as

$$L(\theta|D) \propto \prod_{ij} \theta_{ij}^{\,y_{ij}} \prod_{i} \theta_{i+}^{\,z_{i+}} \prod_{j} \theta_{+j}^{\,u_{+j}} \tag{1}$$

for $\sum\sum \theta_{ij} = 1$, all taken to be non-negative. Here $\theta = (\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ and $i, j = 0, 1$. Subscripts match those of table 2; in computational work we identify $(\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00})$ with `p1, p2, p3, p4` respectively.

If we take a Dirichlet prior for (θ_{ij}), $i = 0, 1$, $j = 0, 1$ with parameters α_{ij}, then the posterior density of θ induced by the data D is

$$\mathsf{P}(\theta|D) \propto \prod_{ij} \theta_{ij}^{\,(y_{ij}+\alpha_{ij}-1)} \prod_{i} \theta_{i+}^{\,z_{i+}} \prod_{j} \theta_{+j}^{\,u_{+j}}. \tag{2}$$

Then we seek $\mathsf{P}(\theta_{+1} > \theta_{1+}|D)$, equivalently we will find $\mathsf{P}(\theta_{01} > \theta_{10}|D)$. We will do this for the special case of $\alpha_{ij} = 1$ for all i, j, which corresponds to a uniform prior density for θ.

We thus seek the posterior probability $\mathsf{P}\,(\theta_{01} > \theta_{10}|D)$ given that

$$\mathsf{P}(\theta|D) \propto \theta_{11}^{26+1-1}\theta_{10}^{1+1-1}\theta_{01}^{5+1-1}\theta_{00}^{18+1-1}(\theta_{11}+\theta_{10})^2(\theta_{01}+\theta_{00})^9(\theta_{11}+\theta_{01})^4(\theta_{10}+\theta_{00})^4. \tag{3}$$

The expression for $\mathsf{P}(\theta|D)$ is a special case of the 'hyperdirichlet' distribution (Hankin 2009). The R idiom is straightforward:

```
> b
```

```
     p1 p2 p3 p4 params powers
[1]   0  0  0  0      0      0
[2]   0  0  0  1     19     18
[3]   0  0  1  0      6      5
[4]   0  0  1  1      9      9
[5]   0  1  0  0      2      1
[6]   0  1  0  1      4      4
[7]   0  1  1  0      0      0
[8]   0  1  1  1      0      0
[9]   1  0  0  0     27     26
[10]  1  0  0  1      0      0
[11]  1  0  1  0      4      4
[12]  1  0  1  1      0      0
[13]  1  1  0  0      2      2
[14]  1  1  0  1      0      0
[15]  1  1  1  0      0      0
[16]  1  1  1  1      0      0

Normalizing constant not known
```

is the appropriate hyperdirichlet distribution. The hyperdirichlet R package gives $\mathsf{P}(\theta_{01} > \theta_{10}|D) = 0.969$ (further computational details are given online by Altham (2009)):

```
f3 <- function(x){x[2]>x[3]}
probability(b,disallowed=f3,eps=1e-2)

  [1] 0.9686342
```

Thus Reviewer 2 is more likely to give a 'Yes' answer than is Reviewer 1. This agrees well with the value of 0.968 given by Lin *et al.* in their Table 3.

Figure 1 shows some numerical results made using the hyperdirichlet R package (Hankin 2009).

Likelihood

The above techniques used a Bayesian approach in which integration was used to calculate the p-value. Here we show how the method of support (Edwards 1992) may be used instead. This is numerically advantageous because multidimensional integration is not needed.

First, find the maximum likelihood estimate for the distribution:

```
    maximum_likelihood(b)

$MLE
        p1         p2         p3         p4
0.45121066 0.01799551 0.11125458 0.41953925
```

```
$likelihood
[1] 1.386809e-28

$support
[1] -64.14538
```

The maximum likelihood estimate of the four parameters given by the PDF of equation 3 is thus $\hat{\theta}^{\mathrm{f}} = (0.451, 0.018, 0.111, 0.42)$, with a corresponding support of $\mathcal{S}^{\mathrm{f}} = -64.15$ (superscript 'f' means that the optimization proceeded **f**reely over the domain).

Now the maximum likelihood estimate under the restriction that $\theta_{01} < \theta_{10}$ is given by

```
f3 <- function(x){x[2]<x[3]}
maximum_likelihood(b,disallowed=f3)
```

```
$MLE
        p1         p2         p3         p4
0.45344213 0.06029619 0.06029619 0.42596548

$likelihood
[1] 1.878493e-29

$support
[1] -66.1445
```

Thus the restricted MLE is $\hat{\theta}^{\mathrm{r}} = (0.453, 0.06, 0.06, 0.426)$, with a corresponding support of $\mathcal{S}^{\mathrm{r}} = -66.14$. Observe that $\hat{\theta}^{\mathrm{r}}_{01} = \hat{\theta}^{\mathrm{r}}_{10}$ as the numerical optimization routine finds a point on the boundary of the admissible region.

The difference $\mathcal{S}^{\mathrm{f}} - \mathcal{S}^{\mathrm{r}} \simeq 1.9991$, suggests that one may increase the support from *any* point consistent with $\theta_{01} < \theta_{10}$ by (almost) two units of support by the expedient of not restricting the search to regions where $\theta_{01} < \theta_{10}$.

Conclusions

Our analysis has added to the techniques which practising statisticians may bring to bear on the analysis of this type of 2×2 table, and we hope to stimulate interest in the aylmer and hyperdirichlet R packages.

References

Altham PME (1971). "The analysis of matched proportions." *Biometrika*, **58**(3), 561–576.

Altham PME (2009). Worksheet 20, URL http://www.statslab.cam.ac.uk/~pat/misc.ps.

Chen HY, Little RJA (1999). "A test of missing completely at random for generalized estimating equations." *Biometrika*, **86**, 1–13.

Edwards AWF (1992). *Likelihood (Expanded Edition)*. John Hopkins.

Hankin RKS (2009). *hyperdirichlet: A generalization of the Dirichlet distribution*. R package version 1.1-6; vignette therein based on a manuscript under review at the *Journal of Statistical Software*, URL http://www.R-project.org.

Lin Y, Lipsitz S, Sinha D, Gawande AA, Regenbogen SE, Greenberg CC (2009). "Using Bayesian p-values in a 2×2 table of matched pairs with incompletely classified data." *Journal of the Royal Statistical Society, Series C*, **58**(2). doi:10.1111/j.1467-9876.2008.00645.x. Published Online: Jan 22 2009 5:54PM.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

West L, Hankin RKS (2008). "Exact tests for two-way contingency tables with structural zeros." *Journal of Statistical Software*, **28**(11).

Affiliation:

P. M. E. Altham
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 0WB

Robin K. S. Hankin (corresponding author)
Cambridge Centre for Climate Change Mitigation Research
University of Cambridge
19 Silver Street
Cambridge CB3 9EP
United Kingdom
E-mail: hankin.robin@gmail.com
URL: http://www.landecon.cam.ac.uk/staff/profiles/rhankin.htm

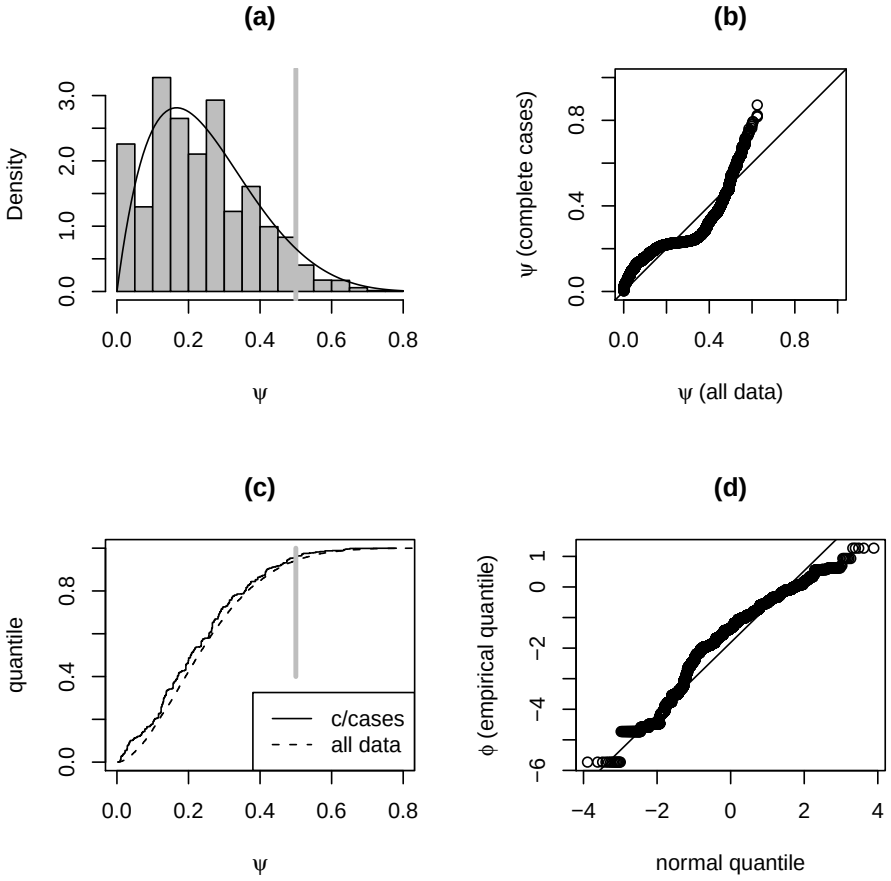

Figure 1: The distribution of $\psi = \theta_{10}/(\theta_{10} + \theta_{01})$ under the posterior distribution induced by the complete cases [a beta distribution with parameters 2, 6; the 'complete cases PDF'] and the whole dataset [empirically sampled using `rhyperdirichlet()` with 30000 samples; the 'all data PDF']; $\psi > 0.5$ means $\theta_{01} > \theta_{10}$ and the gray lines mark $\psi = 0.5$. (a), histogram of complete cases PDF together with the analytically determined all data PDF (b), qqplot; (c) empirical CDF ; (d) normal quantile plot for $\phi = \log(\theta_{01}/\theta_{10})$ with straight line corresponding to asymptotic mean and variance