The effect of 'ignoring the pairing' on the distribution of the log crossratio

P.M.E.Altham, Statistical Laboratory, University of Cambridge,
p.altham@statslab.cam.ac.uk

May 26, 2004

This short note was produced in response to the paper (2004) by Marchant, who criticises the statistical analyses in recent work by Farrington and others.

a) The **crime prevalence comparisons**

For example, Painter and Farrington (1999) show their Table 2 'Changes in the prevalence of Crime' and use the log crossratio to examine the effect of the 'Treatment': in this case improved street lighting, on reduction in particular types of crime. We wish to compare data from the 'Experimental' households with corresponding data from the 'Control' households, by means of the log-crossratios of frequencies.

A difficulty for Painter and Farrington was that when the time came to analyse their data, the information on the exact **pairing** of the households was unavailable. Thus, for example, although in the Experimental group, 431 households were measured before the 'treatment', and 372 households were measured after the 'treatment', we do not know the pairing for the 372 households which were quite possibly observed both before and after. We compare the reduction in crime rate for the 'Experimental' group with the corresponding reduction for the 'Control' by looking at the difference between log crossratios for the 2 groups, and its estimated variance in the 2 cases.

In this short note, we examine carefully the effect of 'ignoring the pairing' on the distribution of the log crossratio.

Consider first the Experimental data alone, which we assume to be summarised by Table 1.

Suppose we have data on n households,

with $x_i = 1$ if the ith household is burgled *before*, $x_i = 0$ if the ith household is NOT burgled *before*, and $y_i = 1$ if the ith household is burgled *after*, $y_i = 0$ if the ith household is not burgled *after*.

Assume $E(x_i) = \pi_x, E(y_i) = \pi_y$. Our object is to compare π_x, π_y.

Then $a = \Sigma_i x_i,\ a + b = n$, and

$c = \Sigma_i y_i,\ c + d = n$.

Thus a/n is the proportion burgled before, c/n is the proportion burgled after, and $E(a/n) = \pi_x,\ E(c/n) = \pi_y$. We will assume

$(x_1, y_1), \ldots, (x_n, y_n)$ are mutually independent pairs. We will evaluate the large-sample distribution of

$$\log(ad/bc)\ .$$

There are 2 cases to consider

i) x_i is independent of y_i, for $i = 1, \ldots, n$.

In this case, a and c are independent variables, and a is distributed as Binomial, parameters n, π_x, and c is distributed as Binomial, parameters n, π_y. Thus $var(a/n) = \pi_x(1 - \pi_x)/n$. Thus for any function $g()$, by writing

$$g(a/n) \simeq g(\pi_x) + (a/n - \pi_x) g'(\pi_x)$$

Table 1: The 2-way table for the Experimental Data

	yes	no
before	a	b
after	c	d

Table 2: The 2-way table for the Control Data

	yes	no
before	a'	b'
after	c'	d'

(ie using a standard Taylor series expansion), we see that

$$var(g(a/n)) \simeq \pi_x(1-\pi_x)(g'(\pi_x))^2/n$$

and so, if we put

$$g(u) = log(u/(1-u))$$

by noting that $g'(u) = 1/u(1-u)$, we see that

$$var(log(a/b)) \simeq \frac{1}{n\pi_x(1-\pi_x)},$$

which is estimated by

$$var(log(a/b)) \simeq (1/a) + (1/b).$$

Thus, since (a/b) and (c/d) are independent, we see that

$$var(log(a/b) - log(c/d)) = var(log(a/b)) + var(log(c/d)),$$

and this is estimated as

$$1/a + 1/b + 1/c + 1/d.$$

ii) What happens to the argument above if the observations on a given household are not independent: specifically if we assume that

$$cor(x_i, y_i) = \rho \text{ for each } i \text{ ?}$$

We seek

$$var\ log(ad/bc) = var(g(a/n) - g(c/n)) = var(g(a/n)) - 2cov(g(a/n), g(c/n)) + var(g(c/n))$$

where the function $g()$ is defined as above. Now, as above,

$$g(a/n) \simeq g(\pi_x) + (a/n - \pi_x)g'(\pi_x)$$

and similarly

$$g(c/n) \simeq g(\pi_y) + (c/n - \pi_y)g'(\pi_y).$$

Thus we see that

$$cov(g(a/n), g(c/n)) \simeq cov(a/n, c/n)g'(\pi_x)g'(\pi_y)$$

and, clearly

$$cov(a/n, c/n) = cov(\Sigma x_i/n, \Sigma y_i/n) = \Sigma cov(x_i/n, y_i/n).$$

Hence

$$\begin{aligned} var\ log(ad/bc) &= var(log(a/b)) + var(log(c/d)) - 2cov(x_1, y_1)g'(\pi_x)g'(\pi_y)/n \\ &= var(log(a/b)) + var(log(c/d)) - 2\frac{\rho}{(n^2\pi_x(1-\pi_x)\pi_y(1-\pi_y))^{1/2}}. \end{aligned}$$

Thus if $\rho > 0$ (as would surely be the case in practice) we will find that

$$var\ log(ad/bc) < var(log(a/b)) + var(log(c/d)).$$

This means that the estimate $1/a + 1/b + 1/c + 1/d$ is an UPPER bound for the variance of the log-crossratio.
For example, if $\pi_x = \pi_y = \pi$ say, then

$$var\ log(ad/bc) = \frac{2(1-\rho)}{(n\pi(1-\pi))}.$$

Turning now to the Control data, which we take to be summarised by Table 2, we see that again, if there is positive dependence between the paired observations on a given household, then

$$var\ log(a'd'/b'c') < var(log(a'/b')) + var(log(c'/d')).$$

Now the two Tables corresponding to Experimental and Control are independent, thus

$$var\ (log(ad/bc) - log(a'd'/b'c')) = var\ log(ad/bc) + var\ log(a'd'/b'c')$$

and so

$$var\ (log(ad/bc) - log(a'd'/b'c')) \leq 1/a + 1/b + 1/c + 1/d + 1/a' + 1/b' + a/c' + 1/d'.$$

Hence, if we have positive dependence (in fact not necessarily the same ρ for each household) between observations on a given household, the procedure which ignores the pairing is actually **conservative** with respect to detecting a difference between the logcrossratios in the Experimental and Control two-way tables.
b) The **crime incidence comparisons**

What can we say about crime incidence, as illustrated for example in Table 3 of Painter and Farrington (1997)?
For example,
in the E group, 431 households experienced 495 crimes before the treatment,
and 372 households experienced 253 crimes after the treatment.
The corresponding figures for the control group were
448 households experienced 368 crimes before the treatment, and
371 households experienced 259 crimes after the treatment.
Once again we have the difficulty that any information on pairing of before/after is unavailable. The situation is just a little more complicated than the one given in (a), but as above, we start with the really simple case.
Let Y_{ij} be the number of crimes experienced by the n_{ij} households, where $i = 1, 2$ corresponds to Experimental, Control respectively, and $j = 1, 2$ corresponds to Before, After respectively. Then (Y_{11}, Y_{12}) will be independent of (Y_{21}, Y_{22}).
i) The simple case.
We assume the 4 variables Y_{ij} are mutually independent, with Y_{ij} distributed as Poisson, mean $n_{ij}\mu_{ij}$. Thus Y_{ij}/n_{ij}, which we will refer to as r_{ij} (for observed rate) is approximately Normal, mean μ_{ij}, variance μ_{ij}/n_{ij}, and so $logr_{ij}$ is approximately Normal, mean $log\mu_{ij}$, and variance $1/(\mu_{ij}n_{ij})$, which of course is estimated by $1/Y_{ij}$. In order to make the comparison of interest, we refer

$$\frac{[log(r_{11}/r_{12}) - log(r_{21}/r_{22})]^2}{\sum_{ij}(1/Y_{ij})}$$

to χ^2_1, thus we are testing for a significant interaction. For the 'all crime' data this gives us the test statistic of 10.40.
ii) More realistically.
$E(Y_{ij}) = \mu_{ij}$ as before, but

$$cor(Y_{1j}, Y_{2j}) = \rho \text{ for each } j$$

because we may have repeated observations on some or all of the households. Further,

$$var(Y_{ij}) = \phi\mu_{ij}$$

because of over-dispersion relative to the Poisson. Thus

$$cov(Y_{i1}, Y_{i2}) = \rho\phi\sqrt{\mu_{i1}\mu_{i2}}.$$

By an argument similar to that given in (a) above, we see that

$$var(log(r_{11}/r_{12})) \simeq v_1 = \phi(1/Y_{11} - 2\rho/\sqrt{(}Y_{11}Y_{12}) + 1/Y_{12}).$$

Similarly,

$$var(log(r_{21}/r_{22})) \simeq v_2 = \phi(1/Y_{21} - 2\rho/\sqrt{(}Y_{21}Y_{22}) + 1/Y_{22}).$$

Once again,

$$var(log(r_{11}/r_{12}) - log(r_{21}/r_{22}))$$

will be the sum of the variances of $(log(r_{11}/r_{12})), (log(r_{21}/r_{22}))$. In this case it is $v_1 + v_2$.
Let us consider what values these variances might have for the given (Y_{ij}) and for various values of ϕ, ρ. (To a large extent the overdispersion with respect to the Poisson, causing $\phi > 1$, is due to an 'excess' of households in the zero-crimes category: we could model this more formally if required.)
Clearly $\phi > 1$ and ρ (assumed > 0) are 'pulling in different directions' in their effect on the estimated variance.
Using the expressions for v_1, v_2 given above, we note that while

$$v_1 + v_2 = .01255 = (.1120)^2 \text{ for } \phi = 1, \rho = 0,$$

we see that

$$v_1 + v_2 = .01297 = (.1139)^2 \text{ for } \phi = 2, \rho = .5,$$

$$v_1 + v_2 = .01946 = (.1395)^2 \text{ for } \phi = 3, \rho = .5,$$

$$v_1 + v_2 = .01218 = (.1104)^2 \text{ for } \phi = 3, \rho = .7.$$

It is remarkable that the effect of having an over-dispersion factor of 3, combined with a 'before-after' correlation of $\rho = .7$, gives an estimated variance which is just a little less than we would have for the simple case of independent Poisson observations discussed in b i) above.

References

Farrington, D.P. and Welsh, B.C. (2004) 'Measuring the effects of improved street lighting on crime: a reply to Dr Marchant' Brit J Criminology, 44, 448-467.

Marchant, P.R. (2004) 'A demonstration that the claim that brighter lighting reduces crime is unfounded' Brit J Criminology, 44, 441-447.

Painter, K. and Farrington, D.P. (1997) The Crime Reducing Effect of Improved Street Lighting: The Dudley project, in 'Situational Crime Prevention: Successful case studies' ed R.V.Clarke.

Appendix: allowing for the 'excess of zero entries'.
Take the problem of crime prevalence, outlined in (a).
With the notation introduced in (a), let us now assume that the first m households are 'burglar-proof', both before and after the 'treatment'. In this case $x_1 = \ldots = x_m = 0$ and similarly $y_1 = \ldots = y_m = 0$. Thus we find a is now binomial, parameters $n - m, \pi_x$, and c is binomial, parameters $n - m, \pi_y$. Hence $E(a/n) = (n - m)\pi_x/n$, and $var(a/n) = (n - m)\pi_x(1 - \pi_x)/n^2$.

$$cov(a/n, c/n) = \frac{(n - m)cov(x_n, y_n)}{n^2}.$$

This gives us, with $g(u) = log(u/(1 - u))$ as before,

$$E(g(a/n)) \simeq log((n - m)\pi_x/(n - (n - m)\pi_x))$$

which is estimated, as before, by $log(a/b)$, and similarly we could find an expression for $var\ log(ad/bc)$.
Note that the true variance of a/n will in fact be SMALLER than the value it would have if a was actually the sum of n independent binomial variables.