Simulating Capture-Recapture Data to Estimate the Correlation Between the Samples

Introduction

This report is designed to be read after a previous one 'Confidence Intervals for Capture-Recapture Data With Matching' and will assume familiarity with the previous work. This considered the problem of calculating confidence intervals for the estimate of the number of people excluded from both of two samples where it cannot be assumed that inclusion in the two samples are independent events. It has been shown that the confidence interval can be expressed as a function of ϕ, the dichotomous correlation coefficient between inclusion in the two samples. Clearly however, if ϕ is not known, the confidence intervals cannot be known either. This note reports the result of work done using simulated data to investigate what the value of ϕ might be in practice.

Method

We consider a population where each person has a propensity *p* to participate in public surveys such as the census and the census coverage survey. It is assumed (i) that *p* follows a beta distribution over the members of the population and (ii) that for a given person it is the same for the two surveys. The second assumption is not the case in practice so it is necessary first to examine the extent to which ϕ is sensitive to asymmetry in the observed data matrix (in which in practice *k* will not be observed).

		S1	
		in	out
S2	in	*A*	*C*
	out	*B*	*k*

We compare the calculated value of ϕ with that which would be calculated from the 'nearest' symmetric matrix. 'Nearest' in this case means leaving *A* and *k* as they are and replacing both *B* and *C* with some combined value such as the arithmetic or geometric mean (the latter being the square root of the product of *B* and *C*). The former has the advantage of retaining the total number of observations while the latter gives values of ϕ closer to the original, as shown by the following table:

A	*B*	*C*	*k*	ϕ	ϕ geo	ϕ arith
12,600	1,440	430	200	0.138	0.144	0.107
15	5	20	7	0.010	0.012	-0.096
250	80	160	60	0.034	0.035	0.009
25	2	5	1	0.124	0.128	0.099

This shows that for most data matrices likely to occur in practice, the symmetric matrix derived by replacing *B* and *C* with their geometric mean will give a value of ϕ which is close to the observed value and which errs on the side of caution (i.e. is too large and hence over-estimates the confidence intervals). The arithmetic mean gives errors which are larger and in the opposite direction.

The density function for the symmetric case, where the two samples have the same propensity for a given person, can be derived analytically.[1] To do this we first note that the probability that a given combination *A*, *B*, *C* and *k* will result from *N* trials is given by the multinomial likelihood

$$p(A,B,C,k \mid N) = \frac{N!}{A!\,B!\,C!\,k!}\, p_A^A p_B^B p_C^C p_k^k = \binom{N}{A,B,C} p_A^A p_B^B p_C^C p_k^k$$

where $k = N - A - B - C$. If *p* (without subscript) is the propensity generated from the beta distribution on a given trial then the multinomial probabilities can be expressed as $p_A = p^2$, $p_B = p_C = p(1-p)$ and $p_k = (1-p)^2$ which can be substituted into the multinomial likelihood. This must then be multiplied by the mass density function of the beta distribution and integrated over *p*:

$$p(A,B,C,k \mid N,\alpha,\beta) = \binom{N}{A,B,C} \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 p^{2A+B+C} (1-p)^{B+C+2k}\, p^{\alpha-1}(1-p)^{\beta-1} dp\,.$$

The integral is itself a beta function giving the final probability mass function as

$$p(A,B,C,k \mid N,\alpha,\beta) = \binom{N}{A,B,C} \frac{\mathrm{B}(2A+B+C+\alpha, B+C+2k+\beta)}{\mathrm{B}(\alpha,\beta)}\,. \qquad (1)$$

This argument is a slight generalisation of that used to derive the probability mass function of the beta binomial distribution. This has the same form as (1) which is in effect a beta multinomial distribution since on each trial there are four possible outcomes rather than two. Calculating the values of the mass function can be facilitated by substituting $k = 0$ in (1) at the first step and thereafter using the recursive relationship

$$\frac{p(A,B,C,k \mid N,\alpha,\beta)}{p(A,B,C,k-1 \mid N,\alpha,\beta)} = \frac{N}{k} \frac{(B+C+2k+\beta-1)}{(2N+\alpha+\beta-1)} \frac{(B+C+2k+\beta-2)}{(2N+\alpha+\beta-2)}\,.$$

Appendix A to this note lists the expected values of the first and second moments and joint moments of *A*, *B*, *C* and *D* for the beta multinomial distribution. One important result is that the expected values of *AD* and *BC* are equal. This implies that for this distribution, the expected value of the numerator of ϕ is zero. This does not imply that ϕ itself is exactly zero since the denominator also consists of combinations of variables but it suggests that ϕ will assume very small values.

This was confirmed by the following extension to the argument. For given values of α and β, the distribution of *N* conditional on *A*, *B* and *C* is given by Bayes' theorem as

$$p(N \mid A,B,C) = \frac{p(A,B,C \mid N)p(N)}{p(A,B,C)}\,. \qquad (2)$$

Footnote

1) The density function for the asymmetric case (where the *p* values for the two samples are not the same for each person) is given in appendix B but as it is very complicated and has broadly the same features as the symmetric version, it does not advance the argument significantly.

The denominator is a constant to be estimated by rescaling so that the probability mass sums to unity over *N*. We can assume that the prior distribution of *N* is a uniform distribution over a very wide range. This is safe because the uniform is the most uninformative prior of all and, conveniently, the range need not be specified. It is sufficient to know that the prior ordinate (the reciprocal of the range) is a constant which can be included in the constant of proportionality. This means that the posterior is simply proportional to the likelihood.

At each step of the calculation, *N* is known by hypothesis and *A*, *B* and *C* by observation and so the value of ϕ for that step, ϕ_N, can be calculated. The steps are for values of *N* starting with $A+B+C$ and continuing until a criterion is reached. The criterion used here was that the value of the likelihood was less than one millionth of the largest value encountered on any previous step. The estimate of ϕ at this point is then given by

$$\hat{\phi} = \frac{\sum_{N=A+B+C} \phi_N \, p(A,B,C \mid N)}{\sum_{N=A+B+C} p(A,B,C \mid N)}.$$

The next task is to estimate the parameters α and β of the beta distribution. It may be shown from (1) that the expected value of *A* is given by

$$E(A) = N \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}. \qquad (3)$$

The variables *B* and *C* have the same expected value which is given by replacing $(\alpha+1)$ with β in the numerator of (3). This yields $E(A)/E(B) = (1+\alpha)/\beta$ which we can use by inserting the observed value of *A* in the numerator and the mean of the observed values of *B* and *C* in the denominator.

Finding a second equation to provide unique parameter estimates has proved to be problematic in the extreme. Pragmatically therefore it is worthwhile investigating how sensitive $\hat{\phi}$ is to the value of α, given that the value of β satisfies the above constraint. The following table gives the values of $\hat{\phi}$ for various combinations of *A*, *B*, *C* and α.

Set	*A*	*B*	*C*	α = 0.01	α = 0.1	α = 1	α = 10	α = 100
1	15	5	20	.116	.116	.115	.106	.093
2	15	9	16	.042	.042	.041	.031	.016
3	15	13	12	.024	.024	.022	.013	-.003
4	250	80	160	.026	.026	.026	.026	.025
5	250	100	140	.007	.007	.007	.007	.007
6	250	120	120	.001	.001	.001	.001	.000
7	12,600	1,440	430	.022	.022	.022	.022	.022
8	12,600	1,190	680	.005	.005	.005	.005	.005
9	12,600	940	930	.000	.000	.000	.000	.000

The nine sets are grouped into three triplets which show the effect of increasing group size and increasing symmetry. Both noticeably reduce the value of $\hat{\phi}$. In no case however was the value of $\hat{\phi}$ as high as 0.12 and, apart from the smallest data sets, they were less than 0.03. These are smaller than the values of ϕ which would be expected in practice and confirms the above expectation that values of ϕ calculated from the beta multinomial distribution will be very small.

Appendix A

This lists the expected values of the first and second moments and joint moments of *A*, *B*, *C* and *D* for the beta multinomial distribution. In all cases, $S = \alpha + \beta$.

Expected values:

$$E(A) = N\frac{\alpha(\alpha+1)}{S(S+1)} \qquad E(B) = E(C) = N\frac{\alpha\beta}{S(S+1)} \qquad E(D) = N\frac{\beta(\beta+1)}{S(S+1)}$$

Expected values of squares and cross products of *A*, *B*, *C* and *D*:

$$E(A^2) = N(N-1)\frac{\alpha(\alpha+1)(\alpha+2)(\alpha+3)}{S(S+1)(S+2)(S+3)} + E(A)$$

$$E(B^2) = E(C^2) = E(BC) + E(B) = E(BC) + E(C)$$

$$E(D^2) = N(N-1)\frac{\beta(\beta+1)(\beta+2)(\beta+3)}{S(S+1)(S+2)(S+3)} + E(D)$$

$$E(AB) = E(AC) = N(N-1)\frac{\alpha\beta(\alpha+1)(\alpha+2)}{S(S+1)(S+2)(S+3)}$$

$$E(AD) = E(BC) = N(N-1)\frac{\alpha\beta(\alpha+1)(\beta+1)}{S(S+1)(S+2)(S+3)}$$

$$E(BD) = E(CD) = N(N-1)\frac{\alpha\beta(\beta+1)(\beta+2)}{S(S+1)(S+2)(S+3)}$$

Variances and covariances can be derived from these by subtracting squares and products of expected values from the appropriate second order term e.g. $\mathrm{cov}(B,C) = E(BC) - [E(B)]^2$. Note in particular that $E(AD - BC) = 0$ which suggests small values of ϕ unless *N* is small.

Appendix B

The asymmetric case is where, for each person, the value of *p* for one sample is taken to be the value for the other sample multiplied by a constant θ which lies between zero and one. Then the density function for *A*, *B* and *C* given the parameters *N*, α, β, and ϕ is given by

$$\binom{N}{A,B,C}\frac{\theta^{A+C}(1-\theta)^{B+D}}{\mathrm{B}(\alpha,\beta)}\sum_{k=0}^{B+D}\binom{B+D}{k}\frac{\mathrm{B}(2N-C-D-k+\alpha,C+D+k+\beta)}{(1-\theta)^{k}}.$$

It can be shown that for this distribution, the expected value of $AD-BC$ equals zero so the values of ϕ which will result will again be close to zero unless *N* is very small.