Confidence Intervals for Capture-Recapture Data With Matching

Executive summary

Capture-recapture data is often used to estimate populations. The classical application for animal populations is to take two samples and count how many individuals are present in both. The classical method involves assuming stochastic independence (i.e. that having been in one sample has no effect on the probability of being in the other). For animals this may a reasonable assumption but for humans it is much less credible. People who do not return their census forms (and hence are not in sample 1) are more likely to be less than co-operative when the enumerator calls (and hence less likely to be in the coverage survey which is sample 2). This paper presents a Bayesian method which expresses the probability distribution of the population not only as a function of the three numbers of people (those in both samples, in sample 1 only and in sample 2 only) and also of the correlation coefficient ϕ between the samples. The median estimate of the population size and the upper and lower 2½ per cent points can be modelled as functions (i) of ϕ (to which they are quite sensitive) and (ii) of the form of the prior distribution (to which they are not very sensitive). The method is illustrated using Scottish data from the 2011 Census. Computer code, sample output and tips for improving efficiency are presented.

1. Introduction

The use of capture-recapture data to estimate animal populations dates back to the early 20th century. One current example of its use to estimate human populations is in the census when it is common to follow up a general census enumeration (which is known not to cover all members of a population) with a census coverage survey (in which intensive effort is expended in achieving accurate counts in a minority of geographical areas). If this minority is representative, comparing the census and coverage counts allows an impression to be formed of the adjustments which must be made to the census counts to make them better reflect the actual population.

However to do this, it is necessary to count how many persons were in both the census and the coverage survey and how many were in one but not the other. Finally it is necessary to estimate how many persons did not feature in either enumeration. Classically, this was done by assuming that inclusion in the two samples were independent of each other, even though there are good grounds for assuming that this is not the case and that in particular, those with an incentive not to participate in the census are also more likely to be unwilling to take part in the coverage survey.

In this report, a Bayesian argument is used to model how the estimate of the number of people missing from both samples, and its credible interval, change as a function of the extent of stochastic dependence between the two samples. First the problem is stated formally. Then the Bayesian solution is stated and the various components needed to implement the argument are developed, along with a discussion of matters relevant to computational efficiency. Finally the method is applied to sample data sets with different assumptions about the Bayesian prior distribution.

2. The problem: A random sample of size $A+B$ cases is taken. A second random sample of size $A+C$ is taken of which A were also in the first sample. What is the probability that there are exactly $N=A+B+C+k$ cases in the total population?

		S1	
		In	Out
S2	In	A	C
	Out	B	k

3. A solution*:* A complete solution is not possible solely on the basis of the information given. However a partial solution can be developed as follows. First, we note that *N* differs from *k* only by the sum of the three known constants *A*, *B* and *C*. We therefore require the probability density of *k* conditional on *A*, *B* and *C*. We can write

$$p(k \mid A,B,C)=\frac{p(A,B,C \mid k)p(k)}{p(A,B,C)} \propto p(A,B,C \mid k)p(k). \qquad (1)$$

This is the classical Bayesian solution in which the posterior probability is proportional to the product of the likelihood and the prior probability.

3a. The likelihood function: The vector $\{A,B,C,k\}$ follows a multinomial distribution with integer parameter $A+B+C+k$ of the form

$$p(A,B,C \mid k)=\frac{(A+B+C+k)!}{A!B!C!k!}\,p_A^A p_B^B p_C^C p_k^k. \qquad (2)$$

The four probabilities must be constant over *k* since, while the cell entries vary probabilistically, the parameters on which they are based are functions of the samples and of the stochastic relationship between them and these are fixed. The probabilities are unknown and must be based on assumptions. The method will be as credible as these assumptions. It seems credible to assume that p_A, p_B and p_C will stand in the same proportions as *A*, *B* and *C* which gives us two constraints. A third is provided by the requirement that the four probabilities sum to one. Instead of imposing a fourth constraint, we express the result in terms of the degree of stochastic dependence between the samples. This is measured by the Pearson correlation coefficient for dichotomous variables ϕ which we can define in terms of the four probabilities as

$$\phi=\frac{p_A p_k-p_B p_C}{\sqrt{(p_A+p_B)(p_C+p_k)(p_A+p_C)(p_B+p_k)}}.$$

Substituting the three constraints into the expression for ϕ and solving for p_A yields the rather ungainly result

$$p_A=\frac{1}{(1+B')(1+C')}\left\{1-\frac{\phi^2(B'+C')+\phi\sqrt{4B'C'+\phi^2(B'-C')^2}}{2(1-\phi^2)}\right\} \qquad (3)$$

where $B'=B/A$ and $C'=C/A$. Then $p_B=p_A B'$, $p_C=p_A C'$ and $p_k=1-p_A-p_B-p_C$. These can then be substituted directly into (2) to calculate the likelihood.

We assume that there is positive correlation between the samples so that if a case has been included in one of the samples, the probability of inclusion in the other is increased. The value of ϕ can be varied systematically from zero upwards to explore the effect on the confidence intervals of changing the degree of stochastic dependence between the samples.

To implement these calculations it is safer to use the natural logarithms of the likelihood as this removes the danger of numeric overflows and underflows. It is useful to note that the natural logarithm of *k*! is returned by the function gammaln(k+1) in Excel and the function lgamma(k+1) in Enterprise Guide. The value for the first term of the likelihood where $k = 0$ can be found by substituting this value in (2). Thereafter it is easier to calculate each term as a function of the preceding one using the relationship

$$\frac{p(A,B,C,k \mid N)}{p(A,B,C,k-1 \mid N-1)} = p_k \frac{N}{k}.$$

3b. The prior distribution*:* Use of the multiplication rule for probabilities in (1) implies that the prior and the likelihood must be stochastically independent and so the numbers *A*, *B* and *C* cannot be used in the definition of the prior. A safe approach would be to adopt an uninformative prior with a large variance representing little prior knowledge about the value of *N*. However it may be argued that this does not use all the information available and that other relevant data (particularly previous estimates of *N* including Mid Year Estimates or MYEs) may be used. This approach is adopted in the next section.

Priors can be divided into one-parameter models such as the Poisson distribution, and two-parameter models such as the normal distribution. The difference in practice is that two-parameter models allow the mean and variance of the prior to be controlled independently. We now illustrate these points with hypothetical, though realistic, data.[1] In all cases the algorithm was run until the log likelihood fell below a value equal to one millionth of its highest value.

4a. An example using a Poisson prior: This illustration of the method uses a one-parameter Poisson prior.[2] The prior density is given by $f(N) = e^{-\lambda}\lambda^N / N!$ and the mean and variance are both equal to the parameter λ. In this example it was assumed that 388 cases were found in both samples, 56 in the first one only and 75 in the second one only. It was also assumed that a previous study had returned a population estimate of 550, so a Poisson prior with parameter 550 was used.

Table 1: Using a Poisson prior with □ = 550			
	2.5%	50%	97.5%
Complete independence (ϕ = 0.00)	4	11	18
Low dependence (ϕ = 0.10)	12	21	31
Low dependence (ϕ = 0.20)	22	33	46
Low dependence (ϕ = 0.30)	36	49	64
Medium dependence (ϕ = 0.40)	54	71	89
Medium dependence (ϕ = 0.50)	81	101	122
Medium dependence (ϕ = 0.60)	123	146	171

Footnotes

1) The code for all the calculations in this section is given in appendix B to this note.

2) Implementing the Poisson form for the prior is eased by noting that each term can be derived simply by multiplying the previous one by λ/N.

High dependence (ϕ = 0.70)	193	222	253
High dependence (ϕ = 0.80)	336	374	413
High dependence (ϕ = 0.84)	444	487	532

Table 1 gives the 2.5%, 50% and 97.5% points of the posterior distribution of the number *k* of cases included in neither sample. The rows represent increasing degrees of dependence. The table shows that the degree of dependence makes a large difference to the outcome throughout the range of ϕ values and that the effect increases the higher the value. As ϕ assumes higher values[3], the calculation becomes unstable and the results cannot be relied on for values above 0.80. Informal experience suggests that the dependence is in the low range (ϕ equals 0.4 or below) but even within this range there is considerable variation.

4b. An example using a normal prior*:* This requires a both mean and a variance. The MYE can serve as the mean, as it did for the Poisson. Confidence intervals are not published for MYEs (otherwise, the variance parameter could be derived from them) so another approach must be used. One such approach is to link the variance to the mean. The one-parameter Poisson prior above in effect did this because, with a parameter is high as 550, the Poisson closely approximates the normal distribution with mean and variance both equal to 550. Table 1 was recompiled using the normal density with these parameters but the result was too close to Table 1 to merit replication. Only at the bottom end of the table did the results differ perceptibly with the normal giving slightly lower cut-offs. For ϕ = 0.8 for example, the figures were 301, 334 and 367. Setting the variance equal to the mean is equivalent to assuming that the estimate of 550 has a standard deviation of 23.4 or 4.3%, which could well be reasonable for MYEs.

To explore the effect of changing the variance parameter, two further simulations were undertaken using the same mean of 550 but variances of 450 and 700 (equivalent to standard deviations of 3.9% and 4.8% respectively). The results are given in Table 2 which shows that the sensitivity of the posterior distribution cut-offs to changes in the variance of the normal prior increases with the degree of dependency between the samples but is only important for degrees which are higher than those likely to be met in practice.

Table 2: Using a normal prior with μ = 550

	σ^2 = 450			σ^2 = 700		
	2.5 %	50 %	97.5 %	2.5 %	50%	97.5 %
Complete independence (ϕ = 0.00)	4	11	18	4	10	18
Low dependence (ϕ = 0.10)	12	21	31	12	21	31
Low dependence (ϕ = 0.20)	22	33	46	22	33	46
Low dependence (ϕ = 0.30)	35	49	64	36	50	65
Medium dependence (ϕ = 0.40)	54	70	87	55	72	90
Medium dependence (ϕ = 0.50)	79	98	118	83	103	124
Medium dependence (ϕ = 0.60)	116	138	161	126	150	176
High dependence (ϕ = 0.70)	174	199	226	199	228	258
High dependence (ϕ = 0.80)	274	304	335	336	372	409

Footnote

3) While the upper bound of ϕ is in theory one, for some combinations of *A*, *B*, *C* and *k* it may be significantly less than this. This can lead to results which are 'out of bounds' such as producing negative values for probabilities where the bracketed denominator in (3) assumes a negative value.

High dependence (ϕ = 0.84)	338	370	403	429	468	508

Implementing the normal form for the prior is eased by expressing each term as a function of the previous one using the relationship

$$\ln \frac{\phi\left(N \mid \mu, \sigma^2\right)}{\phi\left(N-1 \mid \mu, \sigma^2\right)} = \frac{\mu - N + ½}{\sigma^2}.$$

A diagnostic statistic of interest concerns the compatibility between the prior and posterior distributions. In Bayesian analysis it is common (and safe) to use an uninformative prior as this will be compatible with any posterior distribution which is likely to occur in practice. However if an informative prior is used, as here, it is useful to be aware of the extent of the compatibility. For large values of *A*, *B* and *C*, an approximate method of doing this is to divide the difference between the means of the prior and posterior distributions by the standard deviation of the difference, the latter being the square root of the sum of their variances. The ratio is approximately normally distributed with zero mean and unit variance. Formally,

$$z = \frac{\mu_{posterior} - \mu_{prior}}{\sqrt{\sigma^2_{posterior} + \sigma^2_{prior}}}.$$

For the data given above, *z* took values from -0.83 where $\phi = 0.0$ to +0.11 for $\phi = 0.2$ and +1.60 for $\phi = 0.4$ to +10.50 for $\phi = 0.8$. The values varied very little from table to table.

5. Conclusion*:* The method outlined in section 3 and illustrated in section 4 can be applied in practice to capture-recapture data. There are two elements of uncertainty which must be addressed to apply it successfully. The first element concerns the form of the prior and the parameters used for it. The central limit theorem shows that the distributions of sums and means of samples tend to the normal distribution as sample size increases. This is true for all distributions, subject only to the observations being stochastically independent. In this sense the normal is the limiting case of all other distributions. It is a common choice for the prior in Bayesian analyses and can be used here. MYEs can be used for the mean and any reasonable value for the variance, as the results are not sensitive to this at low levels of dependency.

The second element concerns the likelihood, where an assumption must be made about the degree of stochastic dependency between the two samples. Section 4 suggests that the results are sensitive to this and it will be important to have a rationale for how the question is dealt with. It is not necessary to assume a single value for the correlation coefficient. We might use the knowledge we have about the correlation to assume that it is not less than 0.10 but that it is not greater than 0.3. A distribution over the remaining range would then express what we do know about ϕ and what we do not know. This could be implemented in the WinBUGS software by expressing ϕ as following a beta distribution with the bulk of the probability mass concentrated in the interval from 0.3 to 0.1. In practice however a close approximation to this could be obtained by taking a weighted average of the relevant rows in Table 2. If for example we are content to assume that ϕ equals 0.3, 0.2 and 0.1 with probabilities 0.3, 0.4 and 0.3 respectively then the percentage points (rounded to the nearest integer) would be 23, 34 and 47. In summary, subject to a satisfactory decision about the degree of dependency between the samples, this method gives technically very defensible results for the confidence intervals of population estimates based on capture-recapture data.

Appendix A: Comparison of the method with the Chapman estimator and variance.

The Chapman method assumes that the two samples are stochastically independent. The estimate of the population and its approximate variance are given by

$$\hat{N} = \frac{(A+B+1)(C+B+1)}{B+1} - 1 \quad \text{and} \quad \operatorname{var}(\hat{N}) = \frac{(A+B+1)(C+B+1)AC}{(B+1)^2(B+2)}.$$

In the case of the above example, this gives an estimate of 530 with a variance of 14.7 which, assuming a normal approximation, gives upper and lower cut-offs of 515 and 545. The ‘Complete independence’ rows of the three tables in section 2 give corresponding values of 523, 530 and 537. Thus while the estimate is the same, the above method offers slightly smaller confidence intervals than the Chapman method.

Appendix B: Code used for the calculations in section 2

```
* The first step calculates and stores the product terms which are proportional to the
posterior distribution and the proportionality constant. The second step calculates the
posterior distribution and the three cut-offs;

data products (keep = k loglk logpr) const (keep = const);
   N1 = 56; N2 = 75; N12 = 388;
   * Numbers of cases in sample 1 only, sample 2 only, and both samples;
   ncyc = 5000; * The maximum number of cycles;
   crit = -10000; * The criterion below which no further cycles are run;
   const = 0; * The constant of proportionality;
   sum = 0; * The sum of the prior distribution;
   phi = 0.20; * The dichotomous correlation;
   R1 = N1 / N12; R2 = N2 / N12;
   t1 = (1 + R1) * (1 + R2);
   t2 = sqrt(4 * R1 * R2 + phi**2 * (R1 - R2)**2);
   p12 = (1 / t1) * (1 - (phi**2 * (R1 + R2) + phi * t2) / (2 * (1 - phi**2)));
   p1 = p12 * R1; p2 = p12 * R2; pk = 1 - p1 - p2 - p12;
   * The probability parameters;
   k = 0;  * k is the number of cases not in either sample;
   label1:       N = k + N1 + N2 + N12;
   if k = 0 then do;
      loglk = lgamma(1 + N) - lgamma(1 + N1) - lgamma(1 + N12)
       - lgamma(1 + N2) + N1 * log(p1) + N12 * log(p12) + N2 * log(p2);
      * For the normal prior we have;
      mu = 550; var = 450; z2 = (N - mu) ** 2 / var;
      logpr = -(log(2 * 3.1415926 * var) + z2) / 2;
      /** For the Poisson prior we have;
      lamda = 550; logpr = - lamda + N * log(lamda) - lgamma(N + 1);*/
   end;
   else do;
      loglk = loglk + log(N / k) + log(pk);
      logpr = logpr + (mu - N + 0.5) / var;
      * For the Poisson prior, logpr = logpr + log(lamda / N);
   end;
   const = const + exp(loglk + logpr);
   sum = sum + exp(logpr);
   output products;
   k = k + 1; if k le ncyc and loglk gt crit then goto label1;
   output const;
   const = const * 10 ** 20;
   put " Constant of proportionality * 10 ** 20 = " const;
   put " Number of cycles run = " k;
   acc = int(-log(max(sum, 2 - sum) - 1) / log(10));
   put " Sum of prior probabilities is " sum " and equals one to " acc " places of decimals. ";
run;

data _null_;
   set const;
   call symput('const', const);
   put " const = " const;
run;
```

```
data _null_;
   retain s0 s1 s2 0 p975 p500 p025;
   const = symget('const');
   set products end = eof;
   term = exp(loglk + logpr) / const;
   if s0 lt 0.025 and s0 + term ge 0.025
   then p025 = max(k - 1, 0);
   * do not include the next term as it would take you over the boundary;
   if s0 lt 0.5 and s0 + term ge 0.5 then do;
      if s0 + term / 2 gt 0.5 then p500 = k - 1; else p500 = k;
   end;
   if s0 lt 0.975 and s0 + term ge 0.975
   then p975 = k;
   * do include the next term so that it does take you over the boundary;
   s0 = s0 + term; s1 = s1 + term * k; s2 = s2 + term * k * k;
   if eof then do;
      CI = p975 - p025;
      acc = int(-log(max(s0, 2 - s0) - 1) / log(10));
      put " Percentage point of the number of cases in neither sample:";
      put " 2.5% point = " p025 " median = " p500 " 97.5% point = " p975;
      put " Sum of posterior probabilities is " s0 " and equals one to " acc " places of
decimals. ";
      * At the end, s0 is the sum of the whole posterior distribution,
      which should equal one.  Acc measures how close it gets;
      z = (388 + 56 + 75 + s1 - 550)/sqrt(550 + s2 - s1**2);
      put " Measure of the consistency between the prior and the posterior = " z;
   end;
run;
```