Inferences on the binomial parameter of a finite population

Executive summary

A problem which arises in survey design is as follows. There is a population of finite and known size *M* and an attribute (speaks Welsh; is Hindu; thinks there is too much dog mess) which is possessed by a proportion *p* of the population. It is required to draw a sample from the population in order to estimate *p*. Clearly, the larger the sample, the smaller will be the confidence interval *c* around the estimate. For design purposes, it is necessary to know how large the sample must be to ensure a given confidence interval *c* for a hypothesised value of *p*. For example, the designer might require that if the proportion is one in 200 or 0.005 then the 95% confidence interval should be ±0.002. How big a sample is required in order to achieve this?

The key unknown variable is the number of people in the non-sampled part of the population who have the attribute. In this paper it is proposed that this variable follows a beta-binomial distribution. This has a known density function with three parameters which can be calculated from the known and assumed values of *M*, *p* and *c*. This enables the calculation for a given sample size of the 2.5% and 97.5% points of the distribution of the estimate of *p* without making any approximations. The calculation can be repeated until the required confidence is located (this process can be automated). The method is illustrated by relating *M* and sample size for two combinations of *p* and *c*. Computer code, sample output and tips for improving efficiency are presented both for the calculation of *c* for a given sample size and for the automatic calculation of sample size for a given *c*.

Introduction

Consider a population of size M, which is known. The population consists of a proportion p of individuals who possess an attribute which is, or can be treated as if it were, dichotomous. These are denoted positives, the remaining proportion $1-p$ being negatives. It is required to draw a sample of size S such that, if the true value of the proportion is p_h, the 95% confidence interval for the estimate of the proportion is of a specified width. This problem may arise in the design of attribute surveys where each member of a finite population (of a geographical area for example) either has or does not have the attribute and the sample size for the area should be chosen on the basis of the required precision of the estimate around a hypothesised value.

The solution offered below is a general one but it is illustrated by calculations which cover two cases. These are (i) where p_h has the value 0.005 and the 95% confidence interval has width 0.004 (i.e. the estimate is of the form 0.005 ± 0.002) and (ii) where p_h has the value 0.05 and the 95% confidence interval has width 0.04 (i.e. the estimate is of the form 0.05 ± 0.02).

A solution

Table 1: Notation for the sample and population			
	Positive	Negative	Total
Sample[1]	x	$S-x$	S
Non-sample	k	$M-S-k$	$M-S$
Total	$x+k$	$M-x-k$	M

The population and sample are represented in table 1 where the unknown variable is k (the number of positives in the non-sampled part of the population) and $p=(x+k)/M$. The estimate taken directly from the sample is $p_s = x/S$ which tells us something about the value of p but also leaves some uncertainty. We can represent this by taking p to follow a beta distribution with parameters *a* and *b* which we estimate using the method of matching of moments. If the mean and variance of the beta distribution are matched to those of the sample then the distribution and the sample are consistent both in terms of our knowledge of p (as represented by the mean) and in terms of our uncertainty about p (as represented by the variance). The variable k then follows a binomial distribution with a known integer parameter $M-S$ and a probability parameter which follows a beta distribution with parameters $a=(S-1)p_s$ and $b=(S-1)(1-p_s)$, these being the values which match the two moments. Combining these results, k follows a beta binomial distribution with parameters $n=M-S$, $a=(S-1)p_s$ and $b=(S-1)(1-p_s)$.

Having the density function for k allows us to calculate the 2.5% and 97.5% points of its distribution and so express the confidence interval of the estimate of p as a function of the sample size S. In fact, what is required here is the opposite - to find the sample size as a function of the confidence interval - so the calculation must be done repeatedly with different sample sizes in order to arrive at the desired outcome.

The mean of the beta binomial is $na/(a+b)$ which in this case equals $(M-S)p_s$. Substituting this into the estimate for p yields p_s, so the expected value of the estimate of p is simply the observed proportion in the sample. The variance of p is not important for the present argument but for completeness it is given by

$$\left(1-\frac{S}{M}\right)\left(1-\frac{1}{M}\right)\frac{p_s(1-p_s)}{S}.$$

From this we can see that as M tends to infinity, the first two terms vanish and the sole source of uncertainty about p is the finite sample size. The first term is the usual finite population correction while the second term will in practice be negligible.

Footnote

1) This table assumes a zero rate of non-response. In practice of course this is unlikely to happen and the calculations which follow will be based on the numbers of positives and negatives in the achieved sample. To the extent that non-response correlates with the attribute under investigation, the results will be biased. If the extent of the correlation is known, *x* could be adjusted to reflect its likely value if all the sample had responded. This problem of course is common in inferential statistics and is not specific to the present method.

Application

The form of a beta-binomial density function is a combinatorial term multiplied by the ratio between two beta functions. With the above parameters, it is

$$f(k \mid a,b,n) = \binom{n}{k} \frac{\mathrm{B}(a+k, n+b-k)}{\mathrm{B}(a,b)}.$$

Computationally, the most efficient way to calculate each term is to update it from the previous one as this allows certain properties of the combinatorial and the beta functions to be exploited. Specifically, we note that, for $k > 0$,

$$\binom{n}{k} = \binom{n}{k-1}\left(\frac{n+1}{k} - 1\right) \quad \text{and} \quad \frac{\mathrm{B}(a+k, n+b-k)}{\mathrm{B}(a+k-1, n+b-k+1)} = \frac{a+k-1}{n+b-k}.$$

From this it easily follows that

$$\frac{f(k \mid a,b,n)}{f(k-1 \mid a,b,n)} = \left(\frac{n+1}{k} - 1\right)\frac{a+k-1}{n+b-k}.$$

The use of beta functions can be eliminated altogether by expressing the term for $k = 0$ as

$$\frac{\mathrm{B}(a, n+b)}{\mathrm{B}(a,b)} = \prod_{i=1}^{n} \frac{n+b-i}{n+a+b-i},$$

although the code runs a little faster if the beta function in Enterprise Guide 4 is used for the $k = 0$ step. A version of the code used to apply the beta-binomial method (using a ratio of beta functions for the $k = 0$ term) and a sample output are given in appendix A to this note. The accuracy with which the terms are calculated can be assessed by taking the cumulative total of the whole series which should of course equal one. In the calculations reported below, deviation from one occurred around the thirteenth or fourteenth place of decimals, indicating that an adequate degree of accuracy was achieved even where large numbers were involved.

Results (i)

Assuming that $p_s = 1/200$ and using a variety of values for M, it is possible using the above method to calculate the values of S which cause the central 95% of the distribution of p to cover an interval equal to, or very close to 0.004, as required. The results are given in table 2.

Table 2 : Required values of s for various values of M (case i)				
M	S	2.5% value	97.5% value	CI
1,000,000	4,723	0.0032	0.0072	0.0040
750,000	4,715	0.0032	0.0072	0.0040
598,830	4,707	0.0032	0.0072	0.0040
400,000	4,695	0.0032	0.0072	0.0040
220,420	4,658	0.0032	0.0072	0.0040
100,000	4,550	0.0032	0.0072	0.0040
75,000	4,480	0.0032	0.0072	0.0040
50,770	4,370	0.0032	0.0072	0.0040
35,000	4,200	0.0032	0.0072	0.0040
20,000	3,900	0.0032	0.0072	0.0040
12,000	3,450	0.0033	0.0073	0.0040
5,000	2,600	0.0034	0.0074	0.0040
2,000	1,450	0.0036	0.0076	0.0040
1,000	850	0.0043	0.0083	0.0040
600	557	0.0046	0.0080	0.0033
600	556	0.0046	0.0096	0.0050

The third, fifth and eighth rows of data are respectively City of Glasgow, City of Aberdeen and Clackmannanshire. The other rows are simply representative values. The 2.5% and 97.5% values are not the same distance from the criterion value of 0.005 since the distribution is somewhat skewed with such a small hypothesised proportion of positives. Also, when this is combined with a very small value for M, it may not be possible to find a value of S which gives the required confidence interval width. A value for M of 600 for example means that just three positives are hypothesised in the population and although the beta-binomial argument remains valid in theory, its application in practice is highly ‘pixellated’ and it may not be possible to find a solution which satisfies the criteria exactly. This is shown by the large shift which occurs between values of 556 and 557 for S.

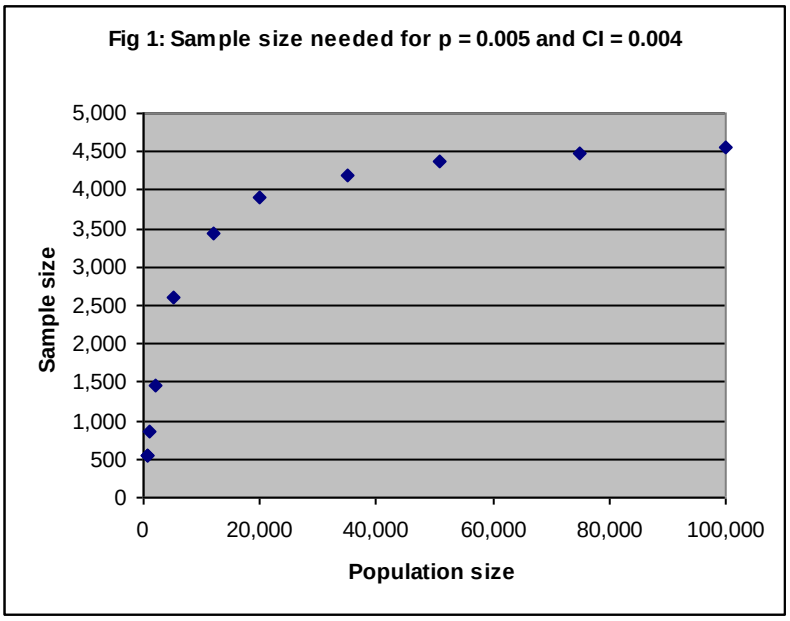


Fig 1 represents the data graphically and shows the consistent relationship between M and S. As noted above, as M tends to infinity, the normal approximation to the confidence interval tends to $\pm 1.96\sqrt{p_s(1-p_s)/S}$. Setting this interval equal to 0.004, using 0.005 for p_s and solving for S yields the value 4,778 which is close to, though probably slightly greater than, the value towards which *s* is converging in table 2 and fig 1.

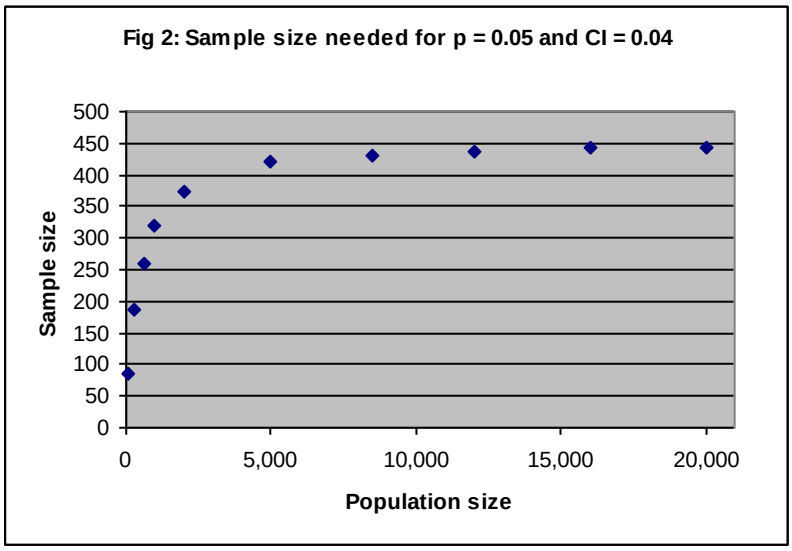
Results (ii)

Table 3: Required values of S for various values of M (case ii)

M	S	2.5% value	97.5% value	CI
1,000,000	453	0.032	0.072	0.040
598,830	453	0.032	0.072	0.040
400,000	453	0.032	0.072	0.040
220,420	452	0.032	0.072	0.040
100,000	451	0.032	0.072	0.040
50,770	450	0.032	0.072	0.040
20,000	444	0.032	0.072	0.040
16,000	442	0.032	0.072	0.040
12,000	438	0.032	0.072	0.040
8,500	430	0.032	0.072	0.040
5,000	420	0.032	0.072	0.040
2,000	375	0.032	0.072	0.040
1,000	320	0.033	0.073	0.040
600	260	0.033	0.073	0.040
300	186	0.034	0.074	0.040
100	85	0.043	0.083	0.040

A second set of calculations was undertaken using a value of 0.05 for p_s. The confidence interval width used was 0.04 (so that the estimate was approximately of the form 0.05 ± 0.02). The results are given in table 3 and fig 2. They are very similar

in shape to those reported above but the value towards which S was trending was about 453 (as opposed to a normal approximation asymptote of 456). Note also that as M increases, the value for S approaches its asymptote much earlier – at about 10,000 as opposed to 60,000 previously.



Conclusion

The method given above based on the beta-binomial distribution gives results which show a consistent relationship between M and S. It would be possible to calculate S for every possible value of M from say 100 (below which values would be meaningless) to 1,000,000 (above which they would not be needed). Appendix B to this paper gives the code to find the correct value of S automatically given an approximate starting point though it can take a lot of computer time if the starting value is not accurate. In practice values for S which are accurate enough can be obtained by visual interpolation from figs 1 and 2, or the search can be done manually for other values of M, p_s and required confidence intervals.

Appendix A : The code used to calculate the 2.5% and 97.5% points of the distribution of *p* for a single given value of S .

```
data _null_;
   M = 5000; s = 420; p = 0.05; RCI = 0.04;
   x = s * p; * hypothesised number of positives in the sample;
   a = (s - 1) * p; b = (s - 1) * (1 - p); n = M - s;
   * the three parameters of the beta binomial distribution;
   term = beta(a, n + b) / beta(a, b);  * the term for k = 0;
   if term = . or term = 0 then do;
      put " This combination of a, b and n has caused an overflow or
      underflow. ";  stop;
   end;
   sofar = term; * hence the sum of the series so far;
   do k = 1 to M - s;
      * k is the number of positives in the non-sampled population;
      term = term * ((n + 1) / k - 1) * (a + k - 1) / (n + b - k);
      * the current term in the series;
      if sofar lt 0.025 and sofar + term ge 0.025
      then p025 = (x + k) / M;
      * do not include the next term as it would take you over the boundary;
      if sofar lt 0.975 and sofar + term ge 0.975
      then p975 = (x + k + 1) / M;
      * do include the next term so that it does take you over the boundary;
      sofar = sofar + term;
   end;
   CI = p975 - p025;
   if sofar ne 1 then error = int(-log(max(sofar, 2 - sofar) - 1) / log(10));
   put " Lower bound = " p025 " Upper bound = " p975;
   put " CI = " CI " Series sum is accurate to " error " places of decimals.";
   * At the end, sofar is the sum of the whole beta binomial series, which should equal
one. Error measures the number of decimal places to which it is accurate;
   if CI lt RCI then put " Decrease s ";
   else if CI gt RCI then put " Increase s ";
run;
```

Sample output for the above settings:

```
 Lower bound = 0.032  Upper bound = 0.072
 CI = 0.04 Series sum is accurate to 14 places of decimals.
```

Appendix B: The code used to calculate the value of s which achieves the confidence interval of RCI given a starting value for S.

```
data _null_;
   M = 5000; s = 420; p = 0.95; RCI = 0.04;
   nc = 0; * number of cycles completed;
   start: * s cycle starts here;
   nc = nc + 1; if nc gt 1000 then goto finish;
   * abort after 1000 cycles;
   a = (s - 1) * p; b = (s - 1) * (1 - p); n = M - s;
   * the three parameters of the beta binomial distribution;
   term = beta(a, n + b) / beta(a, b);  * the term for k = 0;
   if term = . or term = 0 then do;
      put " This combination of a, b and n has caused an overflow or
      underflow. ";  stop;
   end;
   sofar = term; * hence the sum of the series so far;
   do k = 1 to M - s;
      * k is the number of positives in the non-sampled population;
      term = term * ((n + 1) / k - 1) * (a + k - 1) / (n + b - k);
      * the current term in the series;
      if sofar lt 0.025 and sofar + term ge 0.025
      then p025 = (s * p + k) / M;
      * do not include the next term as it would take you over the boundary;
      if sofar lt 0.975 and sofar + term ge 0.975
      then p975 = (s * p + k + 1) / M;
      * do include the next term so that it does take you over the boundary;
      sofar = sofar + term;
   end;
   LCI = TCI; lp975 = p975; lp025 = p025;
   TCI = p975 - p025; * the value of CI this cycle;
   if LCI = . then LCI = 2 * TCI - RCI;
   * only needed at the end of the first cycle;
   if TCI = RCI then do; CI = TCI; goto finish; end;
   if LCI lt RCI and TCI gt RCI then do;
      CI = LCI; p975 = lp975; p025 = lp025;
      s = s + 5; goto finish;
   end; * straddle from above;
   if LCI gt RCI and TCI lt RCI then do;
      CI = TCI; goto finish;
   end; * straddle from below;
   if LCI le TCI and TCI lt RCI then s = s - 5; * approach from above;
   if LCI ge TCI and TCI gt RCI then s = s + 5; * approach from below;
   goto start;
   finish: * this is the best value you're going to get for CI;
   if sofar ne 1 then error = int(-log(max(sofar, 2 - sofar) - 1) / log(10));
   put " Sample size = " s;
   put " No of cycles  = " nc;
   put " Lower bound = " p025 " Upper bound = " p975;
   put " CI = " CI " Series sum is accurate to " error " places of decimals.";
   * At the end, sofar is the sum of the whole beta binomial series, which should equal
one.  Error measures the number of decimal places to which it is accurate;
run;
```

Sample output for the above settings:
Sample size = 420
No of cycles = 2
Lower bound = 0.03195 Upper bound = 0.07215
CI = 0.04 Series sum is accurate to 13 places of decimals.