Some Thoughts on Undercount Problem

Bikas K Sinha*
Retired Professor, Indian Statistcal Institute, Kolkata, India - bikassinha1946@gmail.com

Abstract

The undercount problem is well-known; so is also the ad-hoc solution provided in the literature. The "Text Book Series: Statistics for Social Science Public Policy" lists a book entitled *Statistics for Lawyers* by Finkelstein and Levin [1989/2001]. Chapter 9 [Subsection 9.1.8] of the book is devoted to a discussion on Current Population Survey [CPS]. It reads 'Since the 1980 census, the US Census Bureau has been the subject of intense litigation. The root of the problem...net undercount of the population ...notably Blacks and Hispanics are undercounted ...the Bureau made public its plan to adjust for the 2000 Census'. It continues to say...."two types of post-census sampling...Nonresponse Follow-up Program AND Post Enumeration Survey [PES]...the estimation of the undercount is based on capture/recapture technology:the census is the capture and the PES is the recapture".

As usual, a 2×2 joint response table is prepared: Census vs. PES: $[a \quad b; c \quad d]$ where a corresponds to the frequency count of those 'captured' during the Census and 'recaptured' during the PES; b = frequency count of those captured during the census but not found during the PES; c=frequency count of those missing during the census but found during the PES, and finally, d=frequency count of those totally missing during the entire study. The problem is to provide a 'reasonable estimate' of d, given the other three frequency counts. The problem is deceptively simple and no wonder, we have an ad-hoc solution: $\hat{d} = bc/a$ which is readily available. The fact is that most often this turns out to be an 'underestimate' of the actual missing count!
I propose to discuss this problem and present a direction towards achieving reasonable solution to it.

Keywords: Estimating equations; chi-square; Bayesian formulation.

1. Theoretical Considerations

We intend to discuss some salient features of this problem after a careful mathematical formulation. Let $N = a + b + c + d$ denote the total size of the 'closed' reference population under the per view of the study. Naturally, N is fixed but unknown. As it stands, based on a two-stage or two-phase sampling effort, we have available data on a, b, c and we need to provide an 'estimate' for d.

We denote by $\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}$ the cell probabilities in the 2×2 table. Under the assumption of independence of the two classifications, we have : $\pi_{11} = \pi_{1.}\pi_{.1}$ where $\pi_{1.} = \pi_{11} + \pi_{12}$; etc. Next note that the frequency counts a to d follow a 4-variate singular multinomial distribution with parameters $[N, \pi_{11}, ...]$. Further, under independence,

$$E[ad - bc] = N(N-1)[\pi_{11}\pi_{22} - \pi_{21}\pi_{12}] = 0.$$

Therefore, if $f(a, b, c, d)$ is a function for which $E[f(...)] = 0$, then 'f=0' may be viewed as an 'estimating equation' for d. It turns out that by choosing $f = ad - bc$, we end up with $E[f(..)] = 0$ under independence. Hence, we obtain the solution : $d = bc/a$. This shows that the ad-hoc solution has a statistical basis. Why then it leads to an underestimate?

Let us look at a more general choice of $f(..)$, viz.,

$$f = N(ad - bc)^2/[(a + b)(a + c)(d + b)(d + c)].$$

This is the Chi-square statistic with 1 df for testing independence in a 2-way frequency table. In large samples, under the assumption of independence, $E(\chi^2) = 1$ so that the choice $f(...) = \chi^2 - 1$ leads to the estimating equation: $\chi^2 - 1 = 0$. We consider a more general estimating equation :$f(...) = \chi^2 - \lambda = 0$, where $\lambda \geq 0$ is an arbitrary constant of our choice. Re-writing the equation, we have $N(ad-bc)^2 = \lambda(a+b)(a+c)(d+b)(d+c)$. The LHS is a cubic in d while the RHS is a quadratic in d. For $\lambda = 1$, the cubic equation in d is given by

$$d^3a^2 + d^2[a^3 + a^2(b+c-1) - 2abc - ab - ac - bc] +$$

$\mathrm{d}[\mathrm{b}^2c^2 - 2a^2bc - 2abc(b+c) - (a+b)(b+c)(c+a)] +$

$\mathrm{bc}[\mathrm{abc} + \mathrm{b}^2c + bc^2 - a^2 - ab - ac - bc] = 0.$
It may be noted that the cubic equation is symmetric in b and c. Traditional solution rests on the choice $\lambda = 0$ which is the minimum value and the mode of χ^2-distribution with 1 df. The choice $\lambda = 0.64$ would amount to choosing the median of the χ^2-distribution.

Illustrative Example. We take $a = 17, b = 60, c = 43$. Then $\hat{d} = 152$ is the ad-hoc solution available in the literature corresponding to $\lambda = 0$. For $\lambda = 1$, we derive $d = 209$. Note that under independence, χ^2 is central χ^2 with 1 df and hence it is the square of $Z = N(0,1)$ variable. Therefore, we can go for a choice of λ corresponding to as far as $Z = 1.96$ or, 2.0 to have a coverage of 95 percent. In Figure 1, we show a graph of d versus choices of λ in the closed interval $[0,4]$ based on the above data on a, b, c.

When the assumption of probabilistic independence is not tenable, χ^2 -defined above - follows non-central χ^2-distribution with 1 df and the non-centrality parameter [using delta method] is given by

$$\delta = N[\pi_{11}\pi_{22} - \pi_{12}\pi_{21}]^2[\pi_{1.}\pi_{.1}\pi_{2.}\pi_{.2}]^{(-1)}.$$

Writing Z for the underlying standard normal deviate so that $\chi^2 = Z^2$, it follows that the mean θ of Z is different from zero and as a matter of fact, $\theta^2 = \delta$. This time the mean of χ^2 is $1 + \delta$ and with 95 percent coverage probability, χ^2 will be less than $(\sqrt{\delta} + 2)^2 = 4 + \delta + 4\sqrt{\delta}$. Therefore, we will proceed to solve for d from the estimating equation derived from : $f(...) = 0$, or, equivalently, $d = h(a, b, c; \lambda)$ where this time λ will vary from $1 + \delta$ to $4 + \delta + 4\sqrt{\delta}$. We can make various choices of λ, starting from 0 [the null case] with increments of 0.05, for example. Once more, it is instructional to examine the behavior of the solution for d in terms of the known values of (a, b, c) derived from the experimental data.

2. Bayesian Analysis

The 2×2 cell probabilities π_{ij}'s are assumed to follow a Dirichlet prior distribution with parameters $\alpha, \beta, \gamma, \delta$ in the usual order. Then the joint marginal distribution of the cell frequencies a, b, c, d will be derived by integrating out the product of the multinomial density and the Dirichlet prior, integrated over the π_{ij}'s. The joint pmf has three components :

$$(i)\frac{\Gamma[a+\alpha]\Gamma[b+\beta]\Gamma[c+\gamma]\Gamma[d+\delta]}{\Gamma[n+\alpha+\beta+\gamma+\delta]};$$

$$(ii)\frac{\Gamma[n+1]}{\Gamma[a+1]\Gamma[b+1]\Gamma[c+1]\Gamma[d+1]};$$

$$(iii)\frac{\Gamma[\alpha+\beta+\gamma+\delta]}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)\Gamma(\delta)}.$$

The joint pmf has 4 product terms, each involving one of the random counts such as $\frac{\Gamma[a+\alpha]}{\Gamma(a+1)\Gamma(\alpha)}$.

Statistical Analysis: Since $E[a, given\ \pi_{ij}‘s] = n\pi_{11}$, it follows that marginally, $E[a] = n\alpha/[\alpha + \beta + \gamma + \delta]$.

We can still count on χ^2 as defined before and work out its model expectation under the Dirichlet distribution of the π_{ij}'s. Towards this, first note that $E[\chi^2] = 1 + \delta$ where δ is defined before. The above $E[]$ refers to the conditional expectation, conditional on the π_{ij}'s. Now we need to carry out the unconditional expectation by integrating δ wrt the joint distribution of the π_{ij}'s [which we have taken to be Dirichlet as described above]. For a given choice of the Dirichlet parameters, this calls for numerical integration. Once $E[\delta]$ has been computed, we go back to the defining equation $d = f(a, b, c; \lambda)$ where this time λ will vary from $1 + E[\delta]$ to $4 + E[\delta] + 4E[\sqrt(\delta)]$. We will make various choices of $E[\delta]$, via the defining parameters of the Dirichlet distribution.

Remark. In this context, it would be highly instructional to check the implications of $P[-\epsilon < \pi_{11}\pi_{22} - \pi_{12}\pi_{21} < \epsilon] > 1 - \eta$ on the choice of the parameters of the Dirichlet distribution and thereby examine the nature of the solution for d in terms of 'departure' from the assumption of probabilistic independence. In the above, $P[.]$ has to be evaluated wrt the parameters of the Dirichlet distribution. The above probability inequality is likely to give an upper bound to $E[\delta]$.
We now specialize to the case : Uniform prior for all the P's [denoted as x, y, u, v] in 4 cells. We set $h(x, y, u, v) = k$, for $0 < x, y, u, v < x + y + u + v = 1$.
It follows that $k = 1/\Gamma(4) = 1/3! = 1/6$.

Evaluation of $Pr.[-c < xv - yu < c]$ for $c > 0$

We proceed through the following steps:
1. Fix y and u and integrate out x and v subject to

$(i) x + v = 1 - (y + u); (ii) yu - c < xv < yu + c.$

Note that $(x - v)^2 = (x + v)^2 - 4xv = [1 - (y + u)]^2 - 4xv$.

Therefore, $[1 - (y + u)]^2 - 4yu - 4c < (x - v)^2$

i.e.,

$1 - 2(y + u) + (y - u)^2 - 4c < (x - v)^2 ....(*)$

and

$(x - v)^2 < [1 - (y + u)]^2 - 4yu + 4c$ i.e., $(x - v)^2 < 1 - 2(y + u) + (y - u)^2 + 4c.....(**)$

Combining $(*)$ and $(**)$, we may write

$[1 - (y + u) - [1 - 2(y + u) + (y - u)^2 + 4c] < 2x, 2v < [1 - (y + u) - [1 - 2(y + u) + (y - u)^2 - 4c]$.
Note that we have the same length of the interval for x and v and it is given by
$L = [1 - 2(y + u) + (y - u)^2 - 4c] - [1 - 2(y + u) + (y - u)^2 + 4c]...........(***)$.

Hence, the integral is given by $L^2/4$. We now simplify L^2.

$$L^2 = [1 - 2(y + u) + (y - u)^2 - 4c] +$$
$$[1 - 2(y + u) + (y - u)^2 + 4c] - 2[1 - 2(y + u) + (y - u)^2 - 4c][1 - 2(y + u) + (y - u)^2 + 4c].$$

Therefore, the integral of $h(...)$ simplifies to

$$h(y, u) = [1 - 2(y + u) + (y - u)^2]/2 -$$
$$[1 - 2(y + u) + (y - u)^2 - 4c][1 - 2(y + u) + (y - u)^2 + 4c]/2........(****).$$

For $'c = 0, h(y, u) = 0$ identically as it should. We need to integrate $h(y, u)$ wrt y and u over $0 < y, u < y + u < 1$.

More realistically, we may assume exchangeable distribution of y and u and, further, that P_{22} is the largest of all the cell probabilities. The parameters of the Dirichlet distribution have to be chosen accordingly. It would be interesting to check if the coverage probability towards independence is still substantial for some choices of the parameters of the Dirichlet distribution!

3. Acknowledgement I am thankful to Dr Tommy Wright of the US Bureau of Census and Professor Samindra Nath Sengupta of Calcutta University for some useful discussions on this problem. I also thank Dr Karabi Nandi of UCLA for taking interest in the Bayesian formulation.