Working Paper 2005/1

Estimating Cell Adjustment Confidence Intervals

Paul Williamson

January 2005

Population Microdata Unit
Department of Geography
University of Liverpool

Estimating Cell Adjustment Confidence Intervals

Paul Williamson
Working Paper 2005(1), Population Microdata Unit, Department of Geography, University of Liverpool

January 2005

Contents

Acknowledgement

The general approach to the problem of estimating SDC confidence intervals presented here was inspired by initial suggestions made in Simpson L (2003) 'Are the census outputs fit for purpose', Royal Statistical Society/Office for National Statistics Conference 11-12 November 2003.

(1) General approach

(i) Variance for one cell

For each (potentially) perturbed cell, the variance,

$$s^2 = \sum_i^z \left(x_i p_i - \bar{x}_i p_i\right)^2 \quad (1.1)$$

where:
z = no. of possible post-modification cell-values
x_i = difference between the original (o_i) and post-modification (m_i) count, for possible post-modification cell-value i
$\bar{x}_i$ = mean size of cell perturbation
p_i = probability of a modification of size x_i

Assumption 1.1 All of the statistical disclosure methods to be considered are designed to be unbiased (i.e. mean cell perturbation = 0)

Given assumption 1.1, the formula for variance given above simplifies to:

$$s^2 = \sum_i^z \left(x_i p_i\right)^2 \quad (1.2)$$

(ii) Variance across multiple cells [1]

The variance associated with multiple (potentially) perturbed cells,

$$s_n^2 = s^2 n \quad (1.3)$$

Where n = no. of (potentially) perturbed cells

(iii) Standard deviation for one cell

For one cell, the standard deviation,

$$s = \sqrt{\sum_i^z \left(x_i p_i\right)^2} = \sqrt{s^2} \quad (1.4)$$

(iv) Standard deviation across multiple cells [1]

Following Simpson (2003), it is assumed that, by central limits theorem, as n increases, the distribution of errors approaches a normal distribution with mean 0 and standard deviation

$$s_n = \sqrt{s^2 n} = \sqrt{s^2}\sqrt{n} \quad (1.5)$$

(v) 95% Confidence interval [1]

$$95\%CI = \pm 1.96\sqrt{s^2}\sqrt{n} \quad (1.6)$$

[1] See section 9 for issues concerning error correction for small values of n

(2) The assumption of uniformly distributed cell counts

(i) The assumption of uniform distribution

In estimating SDC confidence intervals, the following key assumption is made:

Assumption 2.1: Prior to disclosure control, counts are uniformly distributed within a range of ±2 of each multiple of 3

(ii) Evidence in support of assumption

Evidence in support of *assumption 2.1* comes from analysis of 2001 Census outputs. For example, *Figure 2.1* plots the distribution of the 52,454 post-modification internal cell counts ≤20 reported in Key Statistics Table 2 (distribution of age across 16 age groups) for Merseyside output areas.

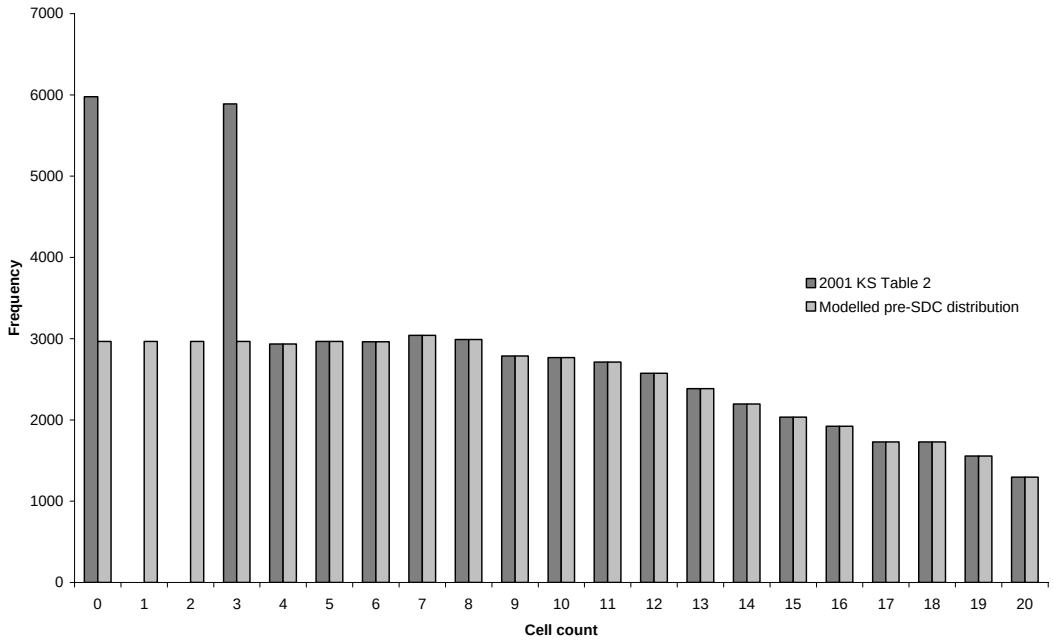

Figure 2.1 Distribution of census output cell counts ≤20

Across the 4586 output areas considered, the distribution of counts may be described as broadly uniform for all unmodified cell counts from 4 to 11. For counts > 11 (median count = 12) the assumption of a uniform distribution clearly breaks down, but even then the assumption that the distribution of counts *within ±2 of each multiple of 3* would appear to be at least a reasonable first approximation. The precise distribution of pre-modification 1s and 2s is unknown, but the modelled distribution, assuming a uniform distribution of 0s, 1s, 2s and 3s, appears to fit the trend observed for counts of 4-11. (Recourse to 1991 Census data would not help much in clarifying this point, as true 0s were not randomly modified.)

(iii) Exceptions to the assumption

Large multiples

To the extent the assumption does not hold for large multiples, as illustrated in *Figure 2.1*, nor will census agencies' claims that their rounding-based SDC methods are unbiased, as an excess of counts less than a given multiple will not be off-set by an equal but opposite deficit of counts

greater than the multiple. (i.e. on average, for larger multiples of 3, the net effect of SDC will be a slight increase in overall cell value.)

Counts of 0

Modified counts of 0 can only arise through rounding down; not through rounding up. Therefore, assuming a uniform distribution of counts between 0 and its nearest multiple, half of all counts of 0 will be unmodified.

(iv) Treatment of exceptions

In what follows the exceptions noted are ignored; in particular counts of 0 are treated as ordinary multiples of 3 (with the exception of an assessment of Small Cell Adjustment). More precise confidence interval estimates could be envisaged, requiring the user to count both the total number of potentially modified cells, and the number of 0s falling within this total, but the resulting increase in accuracy is thought unlikely, except for very sparse matrices, to off-set the increased burden of calculation this would impose.

(3) Rounding to Base 3: modified counts only

Assumption 3.1: Prior to disclosure control, counts are uniformly distributed within a range of ±2 of each multiple of 3
Assumption 3.2: If modified, probability of rounding to nearest multiple of 3 = 2/3; probability of rounding to next nearest multiple of 3 = 1/3
Assumption 3.3: All pre- and post-modification cell values differ (i.e. all cells have been modified)

Given *assumption 3.3*, a count, y_i, with a post-rounding value, m_i, of 15, will have had a pre-rounding value, o_i, of 13, 14, 16, or 17, giving rise to adjustments, x_i, of size -2, -1, 1 or 2 respectively.

Given *assumption 3.1*, an adjustment of size ±1 will have a probability of 2/3; an adjustment of ±2 a probability 1/3. Dividing the relevant probability by two gives the specific probability, p_i, of an adjustment of precise size x_i.

	Post-modification value *(i)*			
m_i	15	15	15	15
o_i	13	14	16	17
$x_i\ [o_i - m_i]$	-2	-1	1	2
p_i	1/6	2/6	2/6	1/6

From these values, and using *formula 1.2*, the variance for an individual perturbed cell may be calculated as follows:

	Post-modification value *(i)*				
m_i	15	15	15	15	
o_i	13	14	16	17	
$x_i\ [o_i - m_i]$	-2	-1	1	2	
p_i	1/3(1/2)	2/3(1/2)	2/3(1/2)	1/3(1/2)	
p_i	1/6	2/6	2/6	1/6	
$(x_i)^2$	4	1	1	4	
$(x_i)^2 p_i$	4/6	2/6	2/6	4/6	
$\Sigma(x_i)^2 p_i$					12/6

i.e. $s^2 = 12/6 = 2$

From this, using *formula 1.5*, it follows that the standard deviation for the post-modification sum of n modified cells,

$$s_n = \sqrt{2}\sqrt{n} = 1.41\sqrt{n} \qquad (3.1)$$

This in turn, using *formula 1.6*, gives a 95% confidence interval for the post-modification sum of n modified cells,

$$= \pm 1.96(1.41)\sqrt{n} = \pm 2.77\sqrt{n}$$

Therefore, the 95% confidence interval for the post-modification sum of 10 modified cells
$= \pm 2.77\sqrt{10} = \pm 8.76$

(4) Rounding to Base 3: all potentially modified counts

Assumption 4.1: Prior to disclosure control, counts are uniformly distributed within a range of ±2 of each multiple of 3
Assumption 4.2: If modified, the probability of rounding to nearest multiple of 3 = 2/3; probability of rounding to next nearest multiple of 3 = 1/3.
Assumption 4.3: Given assumption 3.1, post-modification 1/3 of all cells will remain unmodified (i.e. 1/3 of all original counts are multiples of 3 to start with)

Given *assumption 4.3*, a count, y_i, with a post-rounding value, m_i, of 15, will have had a pre-rounding value, o_i, of 13, 14, 15, 16, or 17, giving rise to adjustments, x_i, of size -2, -1, 0, 1 or 2 respectively.

Given *assumptions 4.1* and *4.3*, the specific probability, p_i, of an adjustment of precise size x_i may be calculated ($p_0 = 1/3$; $p_{\pm 1} = (2/3 \times 2/3)$; $p_{\pm 2} = (1/3 \times 2/3)$).

From these values, and using *formula 1.2*, the variance for an individual perturbed cell may be calculated as follows:

	Post-modification value (i)					
m_i	15	15	15	15	15	
o_i	13	14	15	16	17	
x_i $[o_i - m_i]$	-2	-1	0	1	2	
p_i	1/3(1/3)	2/3(1/3)	1/3	2/3(1/3)	1/3(1/3)	
p_i	1/9	2/9	1/3	2/9	1/9	
$(x_i)^2$	4	1	0	1	4	
$(x_i)^2 p_i$	4/9	2/9	0	2/9	4/9	
$\Sigma(x_i)^2 p_i$						12/9

i.e. $s^2 = 12/9 = 4/3 = 1.33$

From this, using *formula 1.5*, it follows that the standard deviation for the post-modification sum of n modified cells,

$$s_n = \sqrt{1.33}\sqrt{n} = 1.15\sqrt{n} \qquad (4.1)$$

This in turn, using *formula 1.6*, gives a 95% confidence interval for the post-modification sum of n modified cells,

$$= \pm 1.96(1.15)\sqrt{n} = \pm 2.26\sqrt{n}$$

Therefore, the 95% confidence interval for the post-modification sum of 10 modified cells
$= \pm 2.26\sqrt{10} = \pm 7.16$

(5) Small Cell Adjustment

Assumption 5.1: In Small Cell Adjustment only counts of 1 and 2 are modified
Assumption 5.2: The probability of rounding to nearest multiple of 3 = 2/3; probability of rounding to next nearest multiple of 3 = 1/3
Assumption 5.3: Prior to disclosure control, counts of 0, 1, 2 and 3 are uniformly distributed

Given *assumption 5.2*, a count of 1 or two will be adjustmed by ±1 with a probability of 2/3; and adjusted by ±2 with a probability 1/3.

Given *assumption 5.3*, half of all counts with pre-modification values of 0, 1, 2, or 3 will remain unmodified post-modification (because already a multiple of 3).

On this basis:

	Post-modification value *(i)*						
o_i	0	0	0	3	3	3	
m_i	2	1	0	3	2	1	
$x_i\ [o_i - m_i]$	-2	-1	0	0	1	2	
p_i	1/3(1/4)	2/3(1/4)	1/4	1/4	2/3(1/4)	1/3(1/4)	
p_i	1/12	2/12	3/12	3/12	2/12	1/12	
$(x_i)^2$	4	1	0	0	1	4	
$(x_i)^2 p_i$	4/12	2/12	0	0	2/12	4/12	
$\Sigma(x_i)^2 p_i$							12/12

From these values, and using *formula 1.2*, the variance for an individual perturbed cell may be calculated as follows:

i.e. $s^2 = 12/12 = 1$

From this, using *formula 1.5*, it follows that the standard deviation for the post-modification sum of *n* modified cells,

$$s_n = \sqrt{1}\sqrt{n} = 1\sqrt{n} \qquad (5.1)$$

This in turn, using *formula 1.6*, gives a 95% confidence interval for the post-modification sum of *n* modified cells,

$$= \pm 1.96(1)\sqrt{n} = \pm 1.96\sqrt{n}$$

Therefore, the 95% confidence interval for the post-modification sum of 10 modified cells
$= \pm 1.96\sqrt{10} = \pm 6.20$

(6) Rounding to Base 5

Assumption 6.1: Prior to disclosure control, counts are uniformly distributed within a range of ±4 of each multiple of 5
Assumption 6.2: If modified, the probability of rounding up to next multiple = ($|x_i|/b$), probability of rounding down to preceding multiple = ($1-|x_i|)/b$; where $|x_i|$ = absolute difference between original count and target multiple and b = rounding base. [This follows standard practice as outlined in statistical literature]
Assumption 6.3: Given assumption 4.1, post-modification 1/5 of all cells will remain unmodified (i.e. 1/5 of all original counts are multiples of 5 to start with)

Given *assumption 6.3*, a count, y_i, with a post-rounding value, m_i, of 15, will have had a pre-rounding value, o_i, of 11, 12, 13, 14, 15, 16, 17, 18 or 19 giving rise to adjustments, x_i, of size -4, -3, -2, -1, 0, 1, 2, 3, 4 or 5 respectively.

Given *assumptions 6.1* and *5.3*, the specific probability, p_i, of an adjustment of precise size x_i may be calculated ($p_0 = 1/5$; $p_{\pm 1} = 0.5(1/5 \times 4/5)$; $p_{\pm 2} = 0.5(3/5 \times 4/5)$; $p_{\pm 3} = 0.5(2/5 \times 4/5)$; $p_{\pm 4} = 0.5(1/5 \times 4/5)$)

From these values, and using *formula 1.2*, the variance for an individual perturbed cell may be calculated as follows:

	Post-modification value (i)									
o_i	11	12	13	14	15	16	17	18	19	
m_i	15	15	15	15	15	15	15	15	15	
x_i $[o_i - m_i]$	-4	-3	-2	-1	0	1	2	3	4	
p_i	0.5 [1/5(2/5)]	0.5 x [2/5(2/5)]	0.5 x 3/5(2/5)	0.5 x 4/5(2/5)	1/5	0.5 x 4/5(2/5)	0.5 x 3/5(2/5)	0.5 x [2/5(2/5)]	0.5 [1/5(2/5)]	
p_i	1/25	2/25	3/25	4/25	5/25	4/25	3/25	2/25	1/25	
$(x_i)^2$	16	9	4	1	0	1	4	9	16	
$(x_i)^2 p_i$	16/25	18/25	12/25	4/25	0	4/25	12/25	18/25	16/25	
$\Sigma(x_i)^2 p_i$										100/25

i.e. $s^2 = 100/25 = 4$

From this, using *formula 1.5*, it follows that the standard deviation for the post-modification sum of n modified cells,

$$s_n = \sqrt{4}\sqrt{n} = 2\sqrt{n} \qquad (6.1)$$

This in turn, using *formula 1.6*, gives a 95% confidence interval for the post-modification sum of n modified cells,

$$= \pm 1.96(2)\sqrt{n} = \pm 3.92\sqrt{n}$$

Therefore, the 95% confidence interval for the post-modification sum of 10 modified cells
$= \pm 3.92\sqrt{10} = \pm 12.40$

(7) Barnardisation (p=0.1)

Assumption 7.1: Prior to disclosure control, counts are uniformly distributed within a range of ±1 of potentially modified count
Assumption 7.2: The probability of cell count being subjected to a random change +1 = 0.1; the same probability applies to a random change in value of -1.
Assumption 7.3: Given assumption 7.1, post-modification 4/5 of all cells will remain unmodified

The source of this Barnardisation probability is based on analysis at ED-level of SAS Table S24 from the 1991 Census, comparing a free-standing and independently Barnardised count with a known non-Barnardised equivalent.

Given *assumption 4.3*, a count, y_i, with a post-rounding value, m_i, of 15, will have had a pre-rounding value, o_i, of 14, 15 or 16, giving rise to adjustments, x_i, of size -1, 0 or 1.

Given *assumptions 4.1* and *4.3*, the specific probability, p_i, of an adjustment of precise size x_i may be calculated ($p_0 = 8/10$; $p_{\pm 1} = (2/10)$).

From these values, and using *formula 1.2*, the variance for an individual perturbed cell may be calculated as follows:

	Post-modification value (i)			
o_i	15	15	15	
m_i	14	15	16	
x_i $[o_i - m_i]$	-1	0	1	
p_i	1/10	8/10	1/10	
$(x_i)^2$	1	0	1	
$(x_i)^2 p_i$	1/10	0	1/10	
$\Sigma(x_i)^2 p_i$				2/10

i.e. $s^2 = 2/10 = 0.2$

From this, using *formula 1.5*, it follows that the standard deviation for the post-modification sum of *n* modified cells,

$$s_n = \sqrt{0.2}\sqrt{n} = 0.45\sqrt{n} \qquad (7.1)$$

This in turn, using *formula 1.6*, gives a 95% confidence interval for the post-modification sum of *n* modified cells,

$$= \pm 1.96(0.45)\sqrt{n} = \pm 0.88\sqrt{n}$$

Therefore, the 95% confidence interval for the post-modification sum of 10 modified cells
$= \pm 0.88\sqrt{10} = \pm 2.77$

(8) Barnardisation (p=0.04)

Assumption 8.1: Prior to disclosure control, counts are uniformly distributed within a range of ±1 of potentially modified count
Assumption 8.2: The probability of cell count being subjected to a random change +1 = 0.02; the same probability applies to a random change in value of -1.
Assumption 8.3: Given assumption 7.1, post-modification 96/100 of all cells will remain unmodified

The source of this Barnardisation probability is based on analysis at ED-level of SAS Table S35 from the 1991 Census, in which 84 independently Barnardised cell counts are summed to give a table total which may be compared with a known non-Barndardised equivalent count. Even making allowance for a high proportion of structural and actual 0s, the distribution of errors observed is *far* less than would be expected given a ±1 adjustment probability of 0.2. (95^{th} percentile of observed distribution ≈ 2.5; theoretical 95^{th} percentile if 3/4 of cells are non-zero ≈ 6.8.) Instead, the observed distribution of errors, for this and other similar tables, loosely fits that which might be expected given a ±1 adjustment probability of 0.04.

Given *assumption 8.2*, a count, y_i, with a post-rounding value, m_i, of 15, will have had a pre-rounding value, o_i, of 14, 15 or 16, giving rise to adjustments, x_i, of size -1, 0 or 1.

Given *assumptions 8.1* and *8.3*, the specific probability, p_i, of an adjustment of precise size x_i may be calculated ($p_0 = 96/100$; $p_{\pm 1} = (4/100)$).

From these values, and using *formula 1.2*, the variance for an individual perturbed cell may be calculated as follows:

	Post-modification value (i)			
o_i	15	15	15	
m_i	14	15	16	
$x_i\ [o_i - m_i]$	-1	0	1	
p_i	2/100	96/100	2/100	
$(x_i)^2$	1	0	1	
$(x_i)^2 p_i$	2/100	0	2/100	
$\Sigma(x_i)^2 p_i$				4/100

i.e. $s^2 = 4/100 = 0.04$

From this, using *formula 1.5*, it follows that the standard deviation for the post-modification sum of n modified cells,

$$s_n = \sqrt{0.04}\sqrt{n} = 0.2\sqrt{n} \qquad (8.1)$$

This in turn, using *formula 1.6*, gives a 95% confidence interval for the post-modification sum of n modified cells,

$$= \pm 1.96(0.2)\sqrt{n} = \pm 0.39\sqrt{n}$$

Therefore, the 95% confidence interval for the post-modification sum of 10 modified cells

$$= \pm 0.39\sqrt{10} = \pm 1.24$$

(9) Error correction for small values of *n*

Comparing the estimated 95% confidence intervals to those found using exactly probabilities, it is clear that for small values of *n* (<30) the suggested formulae for *s* in sections 3 to 8 above provide an over-estimate of the actual confidence interval. Following the statistical convention of calculating standard deviations using *n*-1 rather than *n* gives better results for small values of n (<30). Consequently, the formulae recommended for use in estimating the standard deviation of SDC are as follows (with coefficients rounded to 1 d.p. for ease of use):

Rounding to Base 3

$$s_n = 1.2\sqrt{n-1} \quad (9.1)$$

Small Cell Adjustment

$$s_n = 1.0\sqrt{n-1} \quad (9.2)$$

Rounding to Base 5

$$s_n = 2.0\sqrt{n-1} \quad (9.3)$$

Barnardisation (p=0.2)

$$s_n = 0.5\sqrt{n-1} \quad (9.4)$$

Barnardisation (p=0.04)

$$s_n = 0.2\sqrt{n-1} \quad (9.5)$$

The above adjusted formulae provide effective estimates of 95% confidence intervals to the nearest integer, even for *n*=10. The accuracy of the estimated confidence interval diminishes slightly as confidence intervals narrow (i.e. 68%CI marginally less accurate than 95%CI, but still provides good estimate of nearest integer).