BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-88/11

A COMPARISON OF AGGREGATE AND PROPORTIONAL LOSS FUNCTIONS IN ADJUSTMENT USING ARTIFICIAL POPULATIONS

by

*Michael Lee Cohen and **Xiao Di Zhang

University of Maryland
and Bureau of the Census
ASA/Census Program
Washington, D.C.

Bureau of the Census
ASA/Census Program
Washington, D.C. 20233

*ASA/Census Fellow
**ASA/Census Associate

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Kirk M. Wolter

Report completed: March, 1988

Report issued: March 8, 1988

A Comparison of Aggregate and Proportional Loss Functions in Adjustment Using Artificial Populations

by

Michael Lee Cohen
University of Maryland and Bureau of the Census

and

Xiao Di Zhang
Bureau of the Census

Abstract: Aggregate loss functions are loss functions for counts of areas where the count itself is of interest. Proportional loss functions are loss functions for counts of areas where the count as a proportion of a total is of interest. To date, much of the work on adjustment has concentrated on aggregate loss functions. A question is whether this effort will necessarily produce estimated counts which provide a reduction with respect to the census for proportional loss functions. The importance of this question lies in the uses of the census counts for reapportionment of the House of Representatives and fund allocation, for which a relative loss function is natural.

Introduction

In the context of census counts, there are a substantial number of loss functions, i.e., methods for comparing two sets of counts to see which is preferred. Certainly, for every well-defined use of census counts there is an associated loss function, either implicit or explicit. In a series of recent reports, Diffendal, Isaki, Smith, Schultz, and Causey (see e.g., Isaki et al., 1987) have used as many as 11 different loss functions to compare adjusted counts with census counts.

Many of these loss functions can be categorized into two groups. The first group, referred to here as aggregate loss functions, are loss functions for counts of areas where the closeness of the counts to the true counts is key. The second group, referred to here as proportional loss functions, are loss functions where the closeness of the proportion of total to the true proportion is key.

To date, as described in various documents, especially Bureau of the Census (1987), the research effort examining adjustment at the Bureau of the Census has focused on techniques that bring the total population counts closer to the truth. Therefore the overall objective is to reduce aggregate loss. (For an overview of the likely components of adjustment error see Citro and Cohen, 1985 and Bureau of the Census, 1987.) For example, there has been an investigation into the causes of matching error, in order that the input of this factor into the dual system estimator will be as close as possible to the number of actual matches. Similarly, investigation of empirical Bayes regression models for use in adjustment implicitly makes use of aggregate loss functions. In fact, it is difficult to conceive of a research effort devoted to reducing proportional error.

On the other hand, two of the primary uses of census counts are to reapportion the House of Representatives and to distribute federal and state monies to local areas. Both of these uses imply a proportional loss function.

To demonstrate how contrary aggregate and proportional loss functions can be, think of a counting process that missed all men and counted women perfectly. For purposes of apportionment and fund allocation, such a process would very likely be preferred to the census counts. However, for other purposes, such as various programs with population threshold requirements, such a process would be drastically inferior to the census counts. Less fanciful is an evaluation of the effectiveness of various coverage improvement programs in the 1980 decennial census. It is certainly possible that some of these programs, though they very likely reduced the aggregate loss of the census counts, increased the proportional loss for certain demographic groups, and therefore increased the inequity of reapportionment and fund allocation with respect to those groups.

The research effort of the Bureau of the Census is thus targeted towards reducing aggregate loss while (at least) two primary uses of census counts are associated with proportional loss. It is therefore of interest to determine when reductions in aggregate loss coincide with reductions in proportional loss.

We take a first step towards answering this complex question. Consider estimates of the form:

$$(1) \qquad K_1 \, cm_i + K_2 \, cw_i$$

where cm_i and cw_i are, respectively, the census counts in area i for the minority population and the white and other population. This class of estimates includes the statistical synthetic estimate,

$$\hat{K}_1 \; cm_i + \hat{K}_2 \; cw_i \; ,$$

where $\hat{K}_1 = \sum m_i \; / \sum cm_i$, $\hat{K}_2 = \sum w_i \; / \sum cw_i$, and $\sum m_i$ and $\sum w_i$ are improved counts totalled over all areas for the minority population and the white and other population respectively. We will investigate the question of what pairs $(K_1 \; , \; K_2)$ produce estimates superior to the census for a proportional and an aggregate loss function.

Rather than investigating what occurs for the multi-dimensional collection of population distributions that could be envisioned, we made use of artificial populations which have been developed at the Bureau of the Census. These artificial populations are constructed to provide a true count and a census count for subregions and demographic groups and are believed to approximately exhibit many of the features of undercoverage. Artificial populations are fully described in Isaki, et al. (1986).

We focus on two artificial populations, AP2 and AP3. AP2 treats the undercoverage of Hispanics as identical to that for whites and others. AP3 treats the undercoverage of Hispanics as identical to that for blacks. For the purposes of this paper where we examine the case of two groups and estimates of the form (1), we have joined Hispanics with blacks and we call the resulting group minorities. For this reason, the approach used is more in harmony with AP3. The artificial populations, which can be examined at several levels of geographic aggregation, are investigated here at the level of states and the level of counties.

Proportional and Aggregate Loss

Probably the most cited aggregate loss function in the context of adjustment is the loss function:

$$\sum_{i=1}^{n} (a_i - t_i)^2/t_i$$

where t_i denotes the true count for area i and a_i denotes a proposed count for area i. This loss function, proposed by Fellegi (1980), Tukey (1983), and others before and after, is referred to here as the χ^2 loss function.

Probably the most cited proportional loss function in the context of adjustment is the loss function:

$$\sum_{i=1}^{n} \left(\frac{a_i}{t_i} - \frac{\sum_j a_j}{\sum_j t_j}\right)^2 t_i \ ,$$

again proposed by Tukey (1983), and others, and referred to here as the <u>share</u> loss function.

Let us examine some mathematical equivalences:

First:

$$\sum_i \left(\frac{a_i}{t_i} - \frac{\sum_j a_j}{\sum_j t_j}\right)^2 t_i = \sum_i [a_i^2/t_i - 2a_i \sum_j a_j/\sum_j t_j + (\sum_j a_j)^2 t_i/(\sum_j t_j)^2]$$

$$= (\sum_j a_j)^2 \sum_i [a_i^2/(\sum_j a_j)^2 - 2a_i t_i/\sum_j a_j \sum_j t_j + t_i^2/(\sum_j t_j)^2]/t_i$$

$$= (\sum_j a_j)^2 \sum_j [a_i/\sum_j a_j - t_i/\sum_j t_j]^2/t_i$$

which may be easier to interpret by some as a share loss function since it involves the difference of shares $a_i/\sum a_j$ and $t_i/\sum t_j$.

Also, we see that:

$$\sum_i (a_i - t_i)^2/t_i = \sum_i (a_i/t_i - t_i/t_i)^2 \; t_i$$

$$= \sum \left(\frac{a_i}{t_i} - \frac{\sum a_j}{\sum t_j} + \frac{\sum a_j}{\sum t_j} - 1\right)^2 t_i$$

$$= \text{Share Loss} + \left[\frac{\sum a_j}{\sum t_j} - 1\right]^2 \sum t_i \tag{2}$$

A few observations can be made from examining (2). First, if $\sum t_j = \sum a_j$ then the two loss functions agree. Second, the term $[\frac{\sum a_j}{\sum t_j} - 1]^2 \sum t_j$ can be rewritten as $(\sum a_j - \sum t_j)^2 / \sum t_j$, which is a type of global χ^2 loss function.

Ellipses of Equal Loss

Let us first consider which pairs (K_1 , K_2) have the property where the counts:

$$K_1 \, cm_i + K_2 \, cw_i$$

are superior to the census counts for the χ^2 loss function. This is the interior of the region defined by:

$$\sum (K_1 \, cm_i + K_2 \, cw_i - t_i)^2 / t_i = \sum (c_i - t_i)^2 / t_i$$

where $c_i = cm_i + cw_i$ is the census count for the i-th region.

Simplifying, we have:

$$K_1^2 \sum cm_i^2/t_i + K_2^2 \sum cw_i^2/t_i - 2\, K_1 \sum cm_i - 2\, K_2 \sum cw_i$$

$$+ 2\, K_1\, K_2 \sum cm_i\, cw_i/t_i - \sum c_i^2/t_i + 2 \sum c_i = 0\ .$$

By the Cauchy-Schwarz inequality we know that the discriminant of the above equation is negative. Therefore this equation describes a rotated ellipse. The center of this ellipse can be shown to be equal to:

$$K_1^* = \frac{(\sum cw_i\; cm_i/t_i)\ (\sum cw_i) - (\sum cw_i^2/t_i)\ (\sum cm_i)}{(\sum cm_i\; cw_i/t_i)^2 - (\sum cm_i^2/t_i)\ (\sum cw_i^2/t_i)}$$

$$K_2^* = \frac{(\sum cm_i\; cw_i/t_i)\ (\sum cm_i) - (\sum cm_i^2/t_i)\ (\sum cw_i)}{(\sum cm_i\; cw_i/t_i)^2 - (\sum cm_i^2/t_i)\ (\sum cw_i^2/t_i)}$$

which can also be found by minimizing:

$$\sum (K_1\, cm_i + K_2\, cw_i - t_i)^2/t_i$$

with respect to K_1 and K_2.

This point $(K_1^*\ ,\ K_2^*)$ is, of course, different from the point obtained when the single constraint

$$\sum (K_1\, cm_i + K_2\, cw_i) = \sum t_i$$

is applied along with the minimization, which in turn is different from the point obtained when the two constraints

$$\sum K_1 \, cm_i = \sum m_i \text{ and } \sum K_2 \, cw_i = \sum w_i$$

are applied, the later resulting in the statistical synthetic estimate. It is interesting to point out that these three points are quite close to one another for AP2 and AP3 at the state and county levels. This is not surprising when comparing the estimate obtained through optimization with the single constraint and the statistical synthetic estimate, as pointed out in Cohen and Zhang (1987). These three estimates for the four artificial population situations are given in Table 1.

Table 1. Comparison of Center of Ellipse to Statistical Synthetic Estimate for Artificial Populations

Population	Level	Center of Ellipse	One Constraint Point	Stat. Syn. Est.
AP2	State	(1.065, 1.005)	(1.068, 1.005)	(1.051, 1.009)
AP2	County	(1.065, 1.006)	(1.064, 1.006)	(1.051, 1.009)
AP3	State	(1.085, 1.002)	(1.085, 1.002)	(1.067, 1.005)
AP3	County	(1.085, 1.002)	(1.083, 1.002)	(1.067, 1.005)

Now let us consider which pairs (K_1, K_2) have the property where the counts

$$K_1 \, cm_i + K_2 \, cw_i$$

are preferred to the census counts for the share loss function. This is the interior of the region defined by:

$$\sum_i \left[\frac{K_1 \, cm_i + K_2 \, cw_i}{t_i} - \frac{\sum_j (K_1 \, cm_j + K_2 \, cw_j)}{\sum t_j}\right]^2 t_i$$

$$= \sum \left[(c_i/t_i) - \sum_j c_j/\sum_j t_j\right]^2 t_i$$

Simplifying, we have:

$$K_1^2 (\sum cm_i^2/t_i - (\sum cm_i)^2/\sum t_i) + K_2^2 (\sum cw_i^2/t_i - (\sum cw_i)^2/\sum t_i)$$
$$+ 2 K_1 K_2 (\sum cm_i cw_i/t_i - (\sum cm_i \sum cw_i) / \sum t_i)$$
$$- \sum c_i^2/t_i + (\sum c_i)^2 / \sum t_i = 0 .$$

Again, by the Cauchy-Schwarz inequality we know that the discriminant of the above equation is negative and therefore the equation describes a rotated ellipse, centered at (0, 0).

Pictures of these two ellipses and their intersection for artificial population AP3 at the state level are included as Figures 1 through 3. (The pictures of the ellipses for AP3 at the county level and for AP2 at both the state and county levels are similar.) In general, the ellipse corresponding to the χ^2 loss function is negatively sloped, centered near the statistical synthetic estimate, and fairly concentrated. The ellipse corresponding to the share loss function is positively sloped, and much longer than the ellipse for the χ^2 loss function.

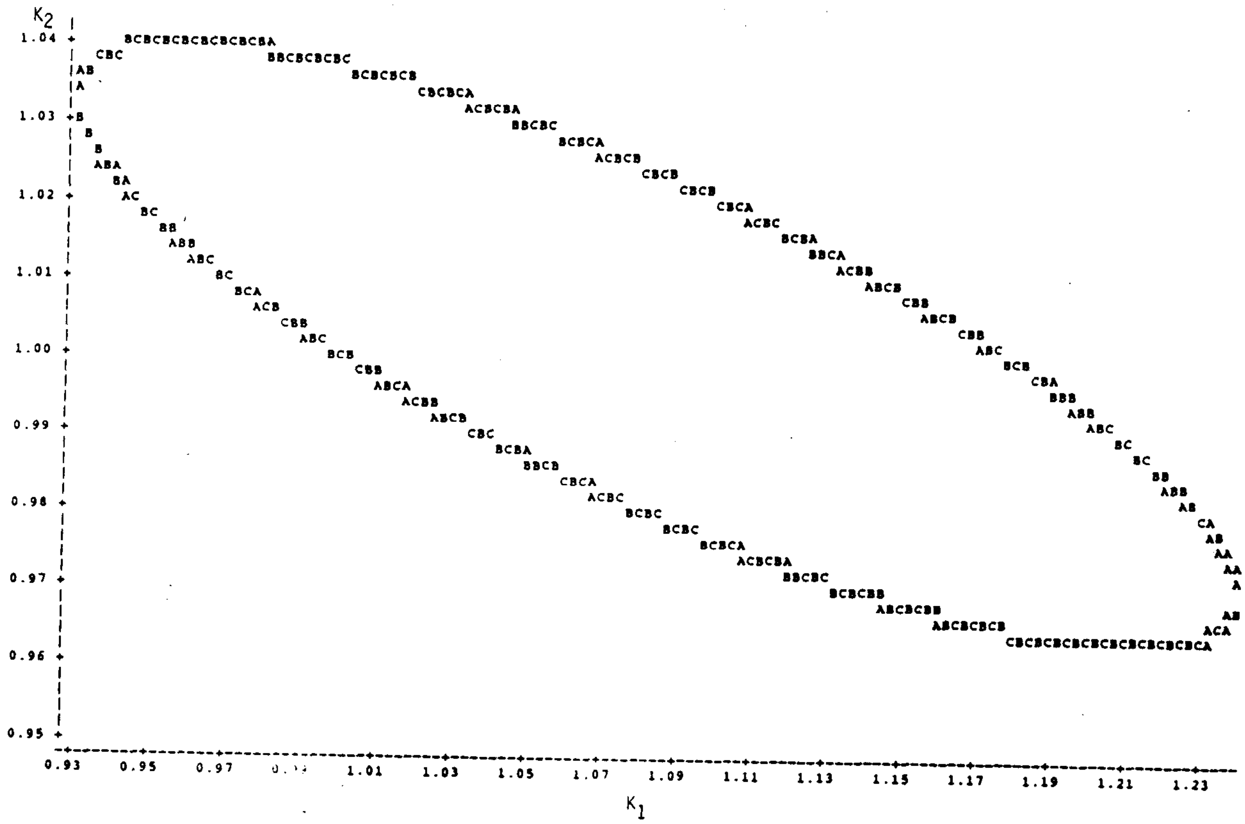


Figure 1. Ellipse Corresponding to χ^2 Loss Function for Artificial Population 3 at the State Level

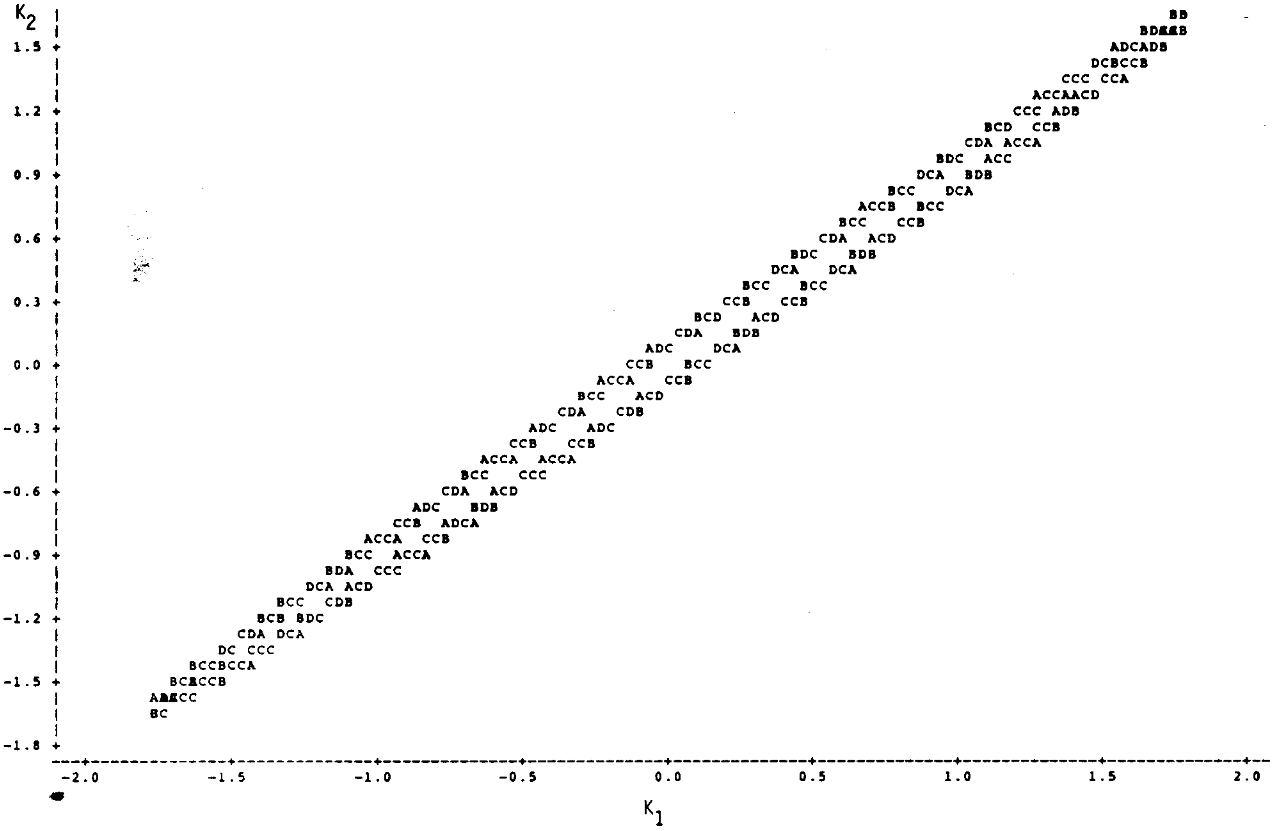
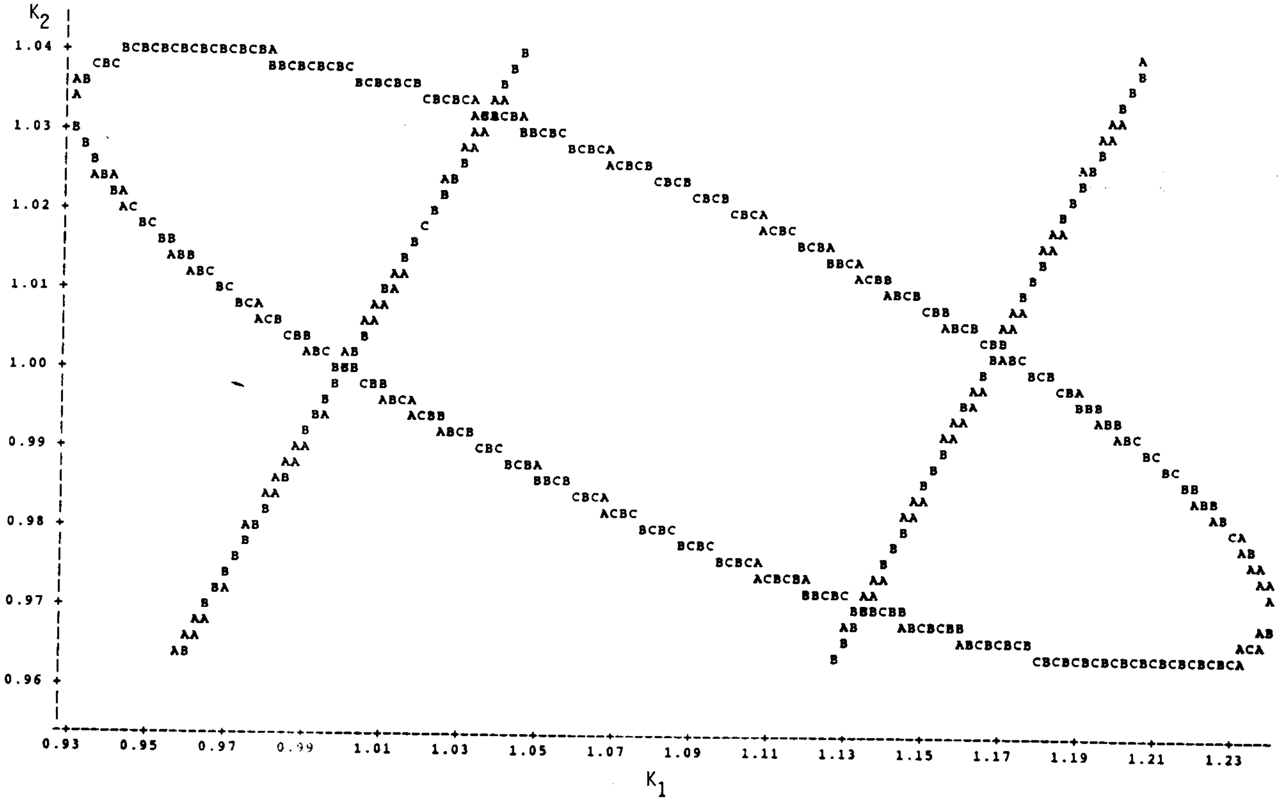


Figure 2. Ellipse Corresponding to Share Loss Function for Artificial Population 3 at the State Level



NOTE: 454 OBS HAD MISSING VALUES

Figure 3. Intersection of Ellipses Corresponding to χ^2 and Share Loss Functions for Artificial Population 3 at the State Level

Intersecting the Ellipses

A first question we might ask now, since we would likely determine K_1 and K_2 by minimizing an aggregate loss function, is what percentage of the ellipse is included in the intersection of the χ^2 ellipse and the share ellipse. In other words, given that our adjusted counts are preferred to the census counts for the χ^2 loss function, what is the probability that our adjusted counts are preferred to the census counts for the share loss function.

Rather than proceed analytically to determine the areas involved, 2000 pairs (K_1, K_2) were randomly generated, K_1 uniformly distributed on (a, b) and K_2 uniformly distributed on (c, d) where the four points (a, c), (a, d), (b, c), and (b, d) form the verties of a rectangle tightly enclosing the χ^2 ellipse. Then for each (K_1, K_2), inclusion in each of the two ellipses was ascertained. The bounds (a, b) and (c, d) are provided in Table 2. The conditional probabilities of inclusion in the share ellipse given inclusion in the χ^2 ellipse are provided in Table 3.

Table 2. Limits of Rectangles Generating K_1, K_2

Population	Level	(a,b)	(c,d)
AP2	State	(.92, 1.21)	(.97, 1.04)
AP3	State	(.93, 1.24)	(.96, 1.04)
AP2	County	(.97, 1.16)	(.98, 1.03)
AP3	County	(.98, 1.19)	(.97, 1.03)

Table 3. Conditional Probabilities of Inclusion in the Share Ellipse Given Inclusion in the χ^2 Ellipse

Population	Level	Conditional Probability
AP2	State	.41 (.016)
AP3	State	.52 (.017)
AP2	County	.64 (.013)
AP3	County	.74 (.012)

Another way to examine this issue is to assume that we have estimated K_1 and K_2 to some degree of accuracy and then to determine the probabilities of inclusion in the χ^2 ellipse, the share ellipse, and simultaneous inclusion in both ellipses.

It is only possible at present to hypothesize the accuracy of the estimation of K_1 and K_2. But assuming that the objective would be to choose K_1 and K_2 so that the adjusted counts are preferable to the census counts for the χ^2 loss function, it seems reasonable to hypothesize inaccuracies increasing up to but not exceeding the bounds determined by the rectangles tightly enclosing the χ^2 ellipses. To be precise, we used intervals of various widths, centered at the x or y- coordinate of the center of the χ^2 ellipse, with the largest intervals corresponding to the limits given in Table 2. Then K_1 and K_2 were generated according to independent uniform distributions over the two given intervals, and for each pair (K_1, K_2), inclusion in the χ^2 and share ellipses were determined. For each population and level of geography, for 6 different widths, 1000 points (K_1, K_2) were generated to estimate the inclusion probabilities. The results are given in Table 4.

Table 4. Inclusion Probabilities Assuming K_1 and K_2 Measured to Various Levels of Accuracy

AP2 State

(K_1- , K_1+)[a]	(K_2- , K_2+)[b]	Prob. χ^2	Prob. in Share	Prob. in Both	Prob. Agree
(1.05, 1.08)	(1.002, 1.009)	1.00	1.00	1.00	1.00
(1.04, 1.10)	(0.998, 1.013)	1.00	1.00	1.00	1.00
(1.01, 1.13)	(0.99, 1.02)	0.94	0.95	0.89	0.89
(0.98, 1.15)	(0.983, 1.027)	0.76	0.73	0.55	0.61
(0.95, 1.18)	(0.976, 1.035)	0.60	0.54	0.32	0.49
(0.92, 1.21)	(0.97, 1.04)	0.46	0.42	0.20	0.51

AP3 State

(K_1- , K_1+)	(K_2- , K_2+)	Prob. χ^2	Prob. in Share	Prob. in Both	Prob. Agree
(1.07, 1.10)	(1.00, 1.005)	1.00	1.00	1.00	1.00
(1.05, 1.12)	(0.993, 1.009)	1.00	1.00	1.00	1.00
(1.02, 1.15)	(0.986, 1.017)	0.95	1.00	0.95	0.95
(0.99, 1.18)	(0.978, 1.025)	0.72	0.86	0.59	0.60
(0.96, 1.21)	(0.971, 1.032)	0.60	0.64	0.37	0.49
(0.93, 1.24)	(0.963, 1.040)	0.47	0.55	0.26	0.49

AP2 County

(K_1- , K_1+)	(K_2- , K_2+)	Prob. χ^2	Prob. in Share	Prob. in Both	Prob. Agree
(1.05, 1.07)	(1.003, 1.009)	1.00	1.00	1.00	1.00
(1.04, 1.08)	(1.000, 1.012)	1.00	1.00	1.00	1.00
(1.03, 1.10)	(0.995, 1.018)	1.00	1.00	1.00	1.00
(1.01, 1.12)	(0.989, 1.023)	0.94	0.95	0.89	0.89
(0.99, 1.14)	(0.983, 1.029)	0.79	0.78	0.59	0.62
(0.97, 1.16)	(0.978, 1.035)	0.66	0.63	0.41	0.52

AP3 County

(K_1- , K_1+)	(K_2- , K_2+)	Prob. χ^2	Prob. in Share	Prob. in Both	Prob. Agree
(1.07, 1.09)	(0.999, 1.005)	1.00	1.00	1.00	1.00
(1.06, 1.10)	(0.996, 1.008)	1.00	1.00	1.00	1.00
(1.04, 1.12)	(0.989, 1.015)	1.00	1.00	1.00	1.00
(1.02, 1.15)	(0.983, 1.021)	0.96	0.99	0.96	0.96
(1.00, 1.17)	(0.977, 1.027)	0.82	0.92	0.74	0.74
(0.98, 1.19)	(0.971, 1.034)	0.64	0.80	0.48	0.53

[a] Lower bound of adjustment factors for minority population.
[b] Upper bound of adjustment factors for white and other population.

Conclusions

The objective of producing adjusted counts which are preferable to census counts for aggregate loss functions may not coincide to a substantial degree with the objective of producing adjusted counts which are preferable for proportional loss functions. Certainly, for artificial populations AP2 and AP3 at the state level, conditional probabilities of .41 and .52 for preferability for proportional loss given preferability for an aggregate loss function are not comforting.

On the other hand, when one examines situations where both of the parameters K_1 and K_2 are measured to reasonable degrees of accuracy, which we think are well-represented by the situations given in the third line of each section of Table 4, agreements of 89% and above are realized.

It is difficult to generalize beyond the artificial populations studied. However, the research effort of the Bureau of the Census has as its goal the estimation of each component of adjustment as precisely as possible, and not that the resulting adjusted counts be preferred to census counts for aggregate loss functions. Though the current adjustment procedures are far more complex than that considered here, it is not unreasonable to suspect that this goal will result in corresponding benefits for both aggregate and proportional loss functions. However, to be able to make more precise statements, more research needs to be done.

References

Bureau of the Census (1987), "On the Technical Feasibility on Adjustment of the 1990 Census," Unpublished paper presented to the Panel on Decennial Census Methodology, Committee on National Statistics, National Research Council, April 24, 1987.

Citro, C.F., and Cohen, M.L. (1985), The Bicentennial Census: New Directions for Methodology in 1990. Washington, DC, National Academy Press.

Cohen, M.L. and Zhang, X.D. (1988), "The Difficulty of Improving Statistical Synthetic Estimation," Unpublished manuscript.

Fellegi, Ivan (1980), "Should the census count be adjusted for allocation purposes--equity considerations," pp. 193-203 in Proceedings of the 1980 Conference on Census Undercount, Bureau of the Census, Washington, DC, U.S. Department of Commerce.

Isaki, Cary, Diffendal, Gregg, and Schultz, Linda (1986), "Statistical Synthetic Estimates of Undercount for Small Areas," pp. 557-568 in Proceedings of the Second Annual Research Conference, U.S. Bureau of the Census.

Isaki, C.T., Schultz, L.K., Smith, P.J., and Diffendal, G.J. (1987), "Small Area Estimation Research for Census Undercount-Progress Report," pp. 219-238 in Small Area Statistics: An International Symposium, John Wiley and Sons, New York, NY.

Tukey, John W. (1983), Affidavit Presented to District Court, Southern District of New York, Mario Cuomo, et al., Plaintiff(s), Malcolm Baldrige, et al., Defendants, 80 CIV. 4550 (JES).