ECI: GENERALIZED VARIANCES GOVERNMENT SECTOR

STEVEN KAUFMAN - BUREAU OF LABOR STATISTICS

INTRODUCTION

Over the last few years, the Employment Cost Index (ECI) Statistical Staff has developed and implemented variance estimation methodology using replication techniques. These procedures measure the variability of the published estimates with respect to their current sample sizes. How these variance estimates vary with different sample sizes is not readily obtainable using replication techniques. Since a variance based allocation scheme requires such information, a combination of replication and regression techniques will be used to approximate the variance estimates as a function with sample size as one variable. This function is called a generalized variance. The generalized variance can then be optimized with respect to sample size, providing the desired allocation. In addition to a variance based allocation procedure, the generalized variance can be used to estimate confidence intervals without directly using the replication variance system.

The ECI is a fixed weight index measuring the change in the total employer cost for wages, benefits and total compensation (wages plus benefits). Quarterly and annual change measures for both wages and total compensation, as well as the index are published. This paper will only analyze annual total compensation change measures. Results for wages and benefits are similar.

This paper will be restricted to determining a generalized variance for the ECI State and Local Government Sector. The Government sample design (except Hospitals and Local Governments with less than 100 employees), is based on a two stage establishment sample design. The first stage is a mail survey where employment counts for a set of probability selected occupations are collected within each establishment. Each occupation is a broad grouping of employees such as professors or managers and is called an Entry Level Occupation (ELO). The first stage occupational employment counts are used to produce measures of size to subsample establishments and occupations in the second stage. Next, after matching detailed establishment occupations (establishment/occupation) into ECI ELOs the field representative subsamples each second stage establishment/ELO to a detailed occupation using current employment counts. The employer's cost for wage and benefits are then collected for each selected detailed occupation within the selected establishment. The collected data for an establishment/occupation is considered one unit in the ECI sample size.

A single stage probability proportional to employment establishment selection procedure is employed in hospitals, and Local Governments with less than 100 employees. In this case too, the field representative, after the job match, subsamples each selected establishment/ELO into a more detailed occupation and collects employer's cost for wage and benefits.

The paper is divided into three main sections. The first section discusses a generalized variance formula for the government sector, describes how it is estimated, and evaluates how well it works. The second section applies the generalized variance to sample allocation. The third section derives confidence intervals from the generalized variances.

ECI GOVERNMENT SECTOR GENERALIZED VARIANCE MODEL

The relative from time a to time b is defined as

$$R_i^{a\,b} = \frac{X_{bi}}{X_{ai}},$$

where X_{bi} is an estimate of the total employer cost (cent/hr) for series i (Total Government, Education, etc.) at time b and X_{ai} is an estimate of the total employer cost for series i at time a .

Using a Taylor series first order approximation [1], the variance of $R_i^{a\,b}$ can be approximated as:

$$\bar{V}^2(R_i^{a\,b}) = \frac{1}{(\bar{X}_{ai})^2}\left[\bar{V}^2(X_{bi}) + \bar{V}^2(X_{ai})\,(\bar{R}_i^{a\,b})^2 - 2\,\bar{\rho}_{abi}\,\bar{R}_i^{a\,b}\bar{V}(X_{ai})\bar{V}(X_{bi})\right]. \quad (1)$$

where $\bar{X}_{ai}$ is the population employer cost for series i at time a,

$\bar{V}^2(X_{bi})$ is the variance of X_{bi},

$\bar{V}^2(X_{ai})$ is the variance of X_{ai},

$\bar{\rho}_{abi}$ is the correlation coefficient between X_{ai} and X_{bi}, and

$\bar{R}_i^{a\,b}$ is the population change for series i from time a to time b.

By replacing $\bar{X}_{ai}$, $\bar{V}(X_{bi})$, $\bar{V}(X_{ai})$, $\bar{\rho}_{abi}$ and $\bar{R}_i^{a\,b}$ with the corresponding sample estimates X_{ai}, $V(X_{bi})$, $V(X_{ai})$, ρ_{abi} and $R_i^{a\,b}$, an estimate of $V^2(R_i^{a\,b})$ is:

$$V^2(R_i^{a\,b}) = \frac{1}{(X_{ai})^2}\left[V^2(X_{bi}) + V^2(X_{ai})\,(R_i^{a\,b})^2 - 2\,\rho_{abi}\,R_i^{a\,b}\,V(X_{ai})\,V(X_{bi})\right]. \quad (2)$$

NOTE: ρ_{abi} is computed using a replication method.

Since $\bar{X}_{ai}$, $\bar{R}_i^{a\,b}$ and $\bar{\rho}_{abi}$ are not functions of the sample size, the sample estimates X_{ai}, $R_i^{a\,b}$ and ρ_{abi} will be considered independent variables in the model. Since $\bar{V}(X_{bi})$ and $\bar{V}(X_{ai})$ approach zero with

increasing sample size, an estimate will be developed which is a function of sample size and other variables not functions of sample size.

$\bar{X}_{bi}$ is not a function of sample size and is slowly increasing, but is stable across quarters. In addition, $V(X_{bi})$ should increase with larger employer costs for a fixed sample size. Keeping these two facts in mind, we propose the following model for the generalized standard error for a time t:

$$\hat{V}(X_{ti}) = \frac{\beta_1 X_{ti}^{\beta_2}}{n_i^{\beta_3}}, \qquad (3a)$$

which is equivalent to:

$$\ln \hat{V}(X_{ti}) = \ln \beta_1 + \beta_2 \ln X_{ti} - \beta_3 \ln n_i;$$

where β_1, β_2 and β_3 are positive unknowns to be estimated using least squares techniques,

X_{ti} is an estimate of the total employer cost for series i, at an arbitrary time t, and

n_i is the responding sample size (the number of establishment/occupations) within series i.

Let k be an industrial stratum (State Higher Education, Local Elementary Schools, etc).

Then, assuming the strata estimates are independent, the generalized variance for X_{ti} is defined to be:

$$\hat{V}^2(X_{ti}) = \sum_{k \in i} \hat{V}^2(X_{tk}), \qquad (4)$$

After substituting $\hat{V}^2(X_{ti})$ from (4) for $V^2(X_{ti})$ in (2), the generalized variance for R_i^{ab} becomes:

$$\hat{V}^2(R_i^{ab}) = \frac{1}{(X_{ai})^2}\left[\hat{V}^2(X_{bi}) + \hat{V}^2(X_{ai})\,(R_i^{ab})^2 - 2\,\rho_{abi}\,R_i^{ab}\,\hat{V}(X_{ai})\,\hat{V}(X_{bi})\right]. \qquad (5)$$

NOTE: Because imputations are made across industrial strata, (4) is not strictly correct. However, the assumption does improve the overall model by improving the industrial stratum estimates, while only minimally hurting combined cells (total Governments, Education, etc.).

ESTIMATING $\hat{V}(X_{ti})$

Since (4) expresses $\hat{V}(X_{ti})$ in terms of $\hat{V}(X_{tk})$, estimation of the model (3a) can be restricted to the use of only industrial strata (k). The unknowns β_1, β_2 and β_3 from $\hat{V}(X_{tk})$ must be chosen to 'best' fit the replication total employer cost standard error estimates. Since the log of $\hat{V}(X_{tk})$ is a linear function of $\ln X_{tk}$ and $\ln n_k$ (see 3b), linear least squares[2/] can be used to find a 'best' fit for the log of the replicated standard error in terms of $\ln X_{tk}$ and $\ln n_k$. Also, since the replicated standard error estimates based on larger samples are more reliable than those based on smaller samples, the replicated standard error estimates based on larger samples should have more weight in the least squares estimation. This will be accomplished by weighting the residual sums of squares by the sample size used to calculate the replicated total employer cost standard error.

$\ln \beta_1$, β_2, and β_3, can be estimated by using Government total compensation data from December 1981 to September 1983. The estimates are given below.

$$\ln \beta_1 = -4.445$$

$$\beta_2 = 1.310 \quad \text{or} \quad V(X_{tk}) = \frac{0.0117\,X^{1.31}}{n_i^{0.65}}$$

and $\beta_3 = 0.65$

This model fits the data very well with an R-square of 0.98 and residuals which are reasonably uniformly distributed. One point to note is β_3 being greater than 0.5 which implies that the survey design is more efficient than a simple random sample.

The appropriateness of the model for $\hat{V}(R_i^{ab})$ can now be checked graphically using proportional error. Here (5) will be used to compute generalized standard errors for all industrial annual compensation estimates from December 1982 to March 1983.

Define Proportional Error to be

$$\frac{\hat{V}(R_{ti}^{ab}) - V(R_{ti}^{ab})}{V(R_{ti}^{ab})},$$

where

$V(R_{ti}^{ab})$ is the replicated standard error for the annual change, R_{ti}^{ab} and $\hat{V}(R_{ti}^{ab})$ is the generalized standard error for the annual change R_{ti}^{ab} both at time t.

Graph 1 shows these Proportional Errors by sample size. The proportional error decreases as the sample size increases. The error is small (less than 12%) when the sample size is above 600, indicating that for industries with large employment, the generalized variance estimates are about as good as those from the replication.

Overall, $\hat{V}^2(R_i^{ab})$ works well. Sample allocation and confidence intervals from the model will be discussed in the next sections.

OPTIMUM SAMPLE ALLOCATION

In this section, an optimal generalized variance sample allocation procedure will be developed. Before the allocation can be derived, the following must be determined: 1) a particular variance estimate to be optimized, 2) a set of strata for which the optimum sample sizes are desired, and 3) a cost function describing the unit collection cost.

The generalized variance estimate for the overall government annual change will be optimized. The annual change estimate is used instead of the index generalized variance to provide a more stable allocation by quarter. This estimate is more stable because the annual ρ's are more clustered than the continually decreasing index ρ's.

Let g stand for the overall governments. Then since

$$\hat{V}^2(X_{tg}) = \sum_{k \in g} V^2(X_{tk})$$

and

$$\rho_{abg}\, \hat{V}(X_{ag})\, \hat{V}(X_{bg}) = \sum_{k \in g} \rho_{abk}\, \hat{V}(X_{ak})\, \hat{V}(X_{bk}),$$

the overall government annual change generalized variance can be expressed as:

$$V(R_g^{a\,b}) = \frac{1}{X_{tg}^2} \sum_{k \in g} \left[\hat{V}^2(X_{bk}) + \hat{V}^2(X_{ak})\,(R_g^{ab})^2 - 2\,\rho_{abk} R_g^{ab} \hat{V}(X_{ak})\, \hat{V}(X_{bk}) \right] \qquad (6)$$

using (5), where g (for overall governments) replaces i and the time period a to b represents a year.

The strata used for this analysis are Industry/ownership (k) cells described in table 1. These are the strata used in the survey design and, therefore are the most appropriate for this analysis.

The cost function is

$$C = C_o + \sum_k C_k n_k, \qquad (7)$$

where: C_o is a fixed overhead cost,

C_k is the unit collection cost within stratum k, and

n_k is the sample size within stratum k.

The generalized variance estimate (6) can be minimized with respect to the sample size, given the cost function (7) as a constraint[3/], using the Cauchy-Schwarz inquality.

The optimum allocation yields:

$$n_k \sim (S_k/C_k)^{1/(1+2\beta_3)} \quad \text{or} \quad n_k \sim (S_k/C_k)^{0.435} \qquad (8)$$

where S_k is the unit generalized variance for stratum k (i.e., the stratum generalized variance assuming a sample size of 1).

$$S_k/C_k = X_{tk}^{2\beta_2} \left[\left((R_k^{ab})^{\beta_2} - \rho_{abk} R_g^{ab}\right)^2 + (R_g^{ab})^2 (1 - \rho_{abk}^2) \right] / C_k \qquad (9)$$

The allocation will be large when S_k/C_k is large, i.e. from (9), when:

1) X_{tk} is large. Strata with large employer employer costs have a large impact on the estimates and should have a greater portion of the sample;

2) R_k^{ab} is different from R_g^{ab}. Strata with exceedingly large increases or decreases have a large impact on the estimates, so strata with change (R_k^{ab}) which deviate greatly from the overall change R_g^{ab} should have a greater portion of the sample;

3) ρ_{abk}^2 is small. Since ρ's for ECI are always positive, the Taylor Series approximation for the variance of $R_{k,}^{ab}$ (see (5)), shows that the larger the ρ_{abk} the smaller the variance. Therefore, strata with small ρ_{abk} require more sample.

and 4) when the unit collection cost is small.

One sample allocation problem with a periodic survey is that the estimates and the optimum allocations will change with each publication. Hopefully, the allocations will be relatively stable. To avoid having to choose a specific time period to determine the optimum allocation which may work poorly for other quarters, an average stratum allocation across the available quarters can be used to obtain the average optimum allocation. This allocation (see table 1) will not be optimum for any quarter, but the efficiency should be very close to optimum for some quarters and reasonably close to optimum for the other quarters.

The optimum or average optimum allocation, assuming equal collection costs (C_k), can be compared with a simple, intuitively reasonable allocation to get some measure of efficiency. Allocating the sample proportionately to the stratum total employer cost is a simple allocation which should provide reasonable results, since the employer cost is a measure of stratum importance within the change estimate (R_g^{ab}). Table 2 provides these efficiencies with respect to the total employer cost allocation.

The optimum allocation described in this section provides good results. When the employer cost is large or the stratum change is greatly different from the overall change or the correlation is relatively small or the unit collection cost is small, the allocation is increased as expected. When compared with the simpler employer cost allocation, the average optimum allocation provides consistent gains in precision.

CONFIDENCE INTERVALS

This section will discuss how the generalized variance formula can be used to develop confidence intervals. The first step is to verify that our estimates are normally distributed. Next, three generalized variance estimates, dependent only on historical estimates from the replicated variance system, will be produced. The two standard deviation interval for each estimate about the published estimate should be an approximate 95% confidence interval. Finally, these three intervals can be compared using the proportional error.

Normality of ECI Estimates

To analyze the distribution of the change estimates, we assume: 1) the numerator and denominator of the estimate are both normally distributed with correlation correlation coefficient ρ and 2) the mean of the denominator is much larger than its standard error. (Since the estimated employer costs are a weighted average, the central limiting theorem provides the rational for the first assumption. In addition, the overall government cost weight is 50 times larger than its standard error, providing the rational for the second assumption, since 3 times is sufficient.) Using these assumptions, Dr. Sandra West of BLS's Office of Research and Evaluation derived the density of the ratio of the numerator to the denominator.[4] Using estimated government data to estimate the parameters of this theoretical distribution, this density can be compared to an appropriate normal density.

To make this comparison, the mean and variance of the theoretical distribution, described above, are numerically calculated. The normal density with this mean and variance is compared to the theoretical distribution by computing the maximum absolute error between the distributions for a Kolmogorov test. From the examples investigated, the maximum absolute error ranged from 0.0014 to 0.05. Using a Kolmogorov test for equality of distributions, it would take sample sizes of 1,000,000 and 600, respectively, to detect a 0.0014 and 0.05 distributional difference at the 95% confidence level.

This indicates that the theoretical distribution closely approximates a normal distribution. The largest differences occur when ρ is large (greater than 0.99), in which case, the theoretical distribution's tails are contained within the normal distribution. Therefore, assuming a normal distibution should be conservative.

Generalized Variance Estimate

The generalized variance estimate is a function of sample size, two quarters of employer cost estimates and the correlation between the employer costs (ρ). Sample size and employer costs are outputs from the estimation system and are readily available. The correlation coefficient is an output of the replication variance system and is not available during the production cycle. Therefore, if estimated confidence intervals are required prior to or concurrently with publication, an estimated correlation coefficient must be found. Using the eight quarters of ρs used in this analysis, the minimum ρ, the median ρ and the mean ρ within each industrial stratum can be calculated and used as an estimated correlation coefficient for all quarters.

From (5), since the correlation coefficients are always positive for the ECI, the minimum correlation coefficient provides the most conservative estimate because the variance reduction effects of the correlation will be minimal. The median correlation will provide overestimates half the time and under estimates the other half.

The various confidence intervals can be compared by computing the proportional error.

$$\text{Proportional error} = \frac{Z}{L}$$

where Z is the length of generalized variance confidence interval, and;

L is the length of replicated variance confidence interval.

Graphs 2 and 3 represent the proportional error for the annual and quarterly compensation change, using the median ρ. Graphs using mininum ρ and average ρ are available upon request.[3] However, the results for all three ρ estimates are summarized below:

1) All of the confidence interval methods work reasonably well for annual change. The median ρ confidence interval may be slightly better than the other intervals.

2) Confidence intervals for quarterly change do not work as well as annual change intervals. The graphs have a much larger spread. Even when the interval is based on a large sample size, the error may still be large. To improve the quarterly change intervals, better ρ estimates must be found. Estimation of ρ may be complicated by seasonal relationships. After a few more years, if we can detect seasonal relationships, it may be possible to improve the confidence intervals. Until then only annual change generalized variance confidence intervals should be used.

CONCLUSIONS

A generalized variance model for the ECI government sector works very well. An optimum allocation based on the generalized variance can be derived. For a more stable allocation, the average optimum allocation (a simple average of the optimum allocations) can also be analyzed. These allocations provide consistent improvements in the standard errors as compared with the employer cost allocation (See table 2).

Generalized variances can also be used to produce confidence intervals. They work reasonably well only for annual change estimates. Generalized confidence intervals for quarterly change estimates do not work well and more analysis is required to improve them.

REFERENCES

1/ Introduction to the Theory of Statistics, A. Mood, F. Graybill and D. Boes, McGrawHill, 1984, pg. 181.

2/ Applied Regression Analysis, Draper and Smith, J Wiley and Sons, 1966 pg. 9, 58.

3/ Employment Cost Index Generalized Variance Government Sector, Steven Kaufman, unpublished BLS paper, 1984.

4/ Distribution of the Ratio to Two Normal Variables, Sandy West, unpublished BLS paper, 1984.

TABLE 1 — Average Annual Optimum Allocation* (Proportion of total sample)

Industry	Ownership	Compensation
Elem. and Sec Schools	Local	0.349
Higher Ed.	State	0.089
Higher Ed.	Local	0.029
Hospitals	State	0.032
Hospitals	Local	0.034
Rest of Services	State	0.073
Rest of Services	Local	0.038
Construction	State	0.047
Construction	Local	0.015
Transportation	State	0.014
Transportation	Local	0.042
Public Admin.	State	0.024
Public Admin.	Local	0.204
All other Industries	State	0.004
All other Industries	Local	0.007

* This allocation assumes the stratum collection cost (C_k) are equal.

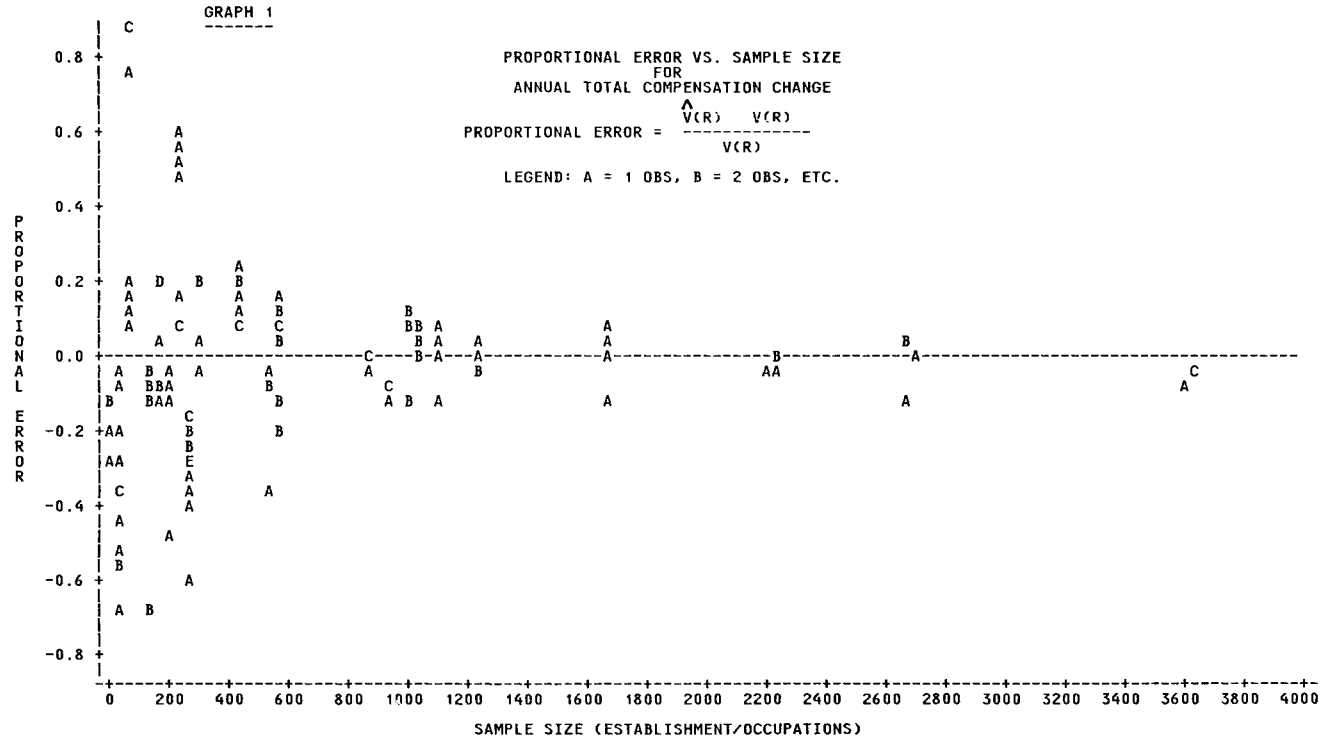
TABLE 2 — Allocation Efficiencies* (Percent decrease in standard error)

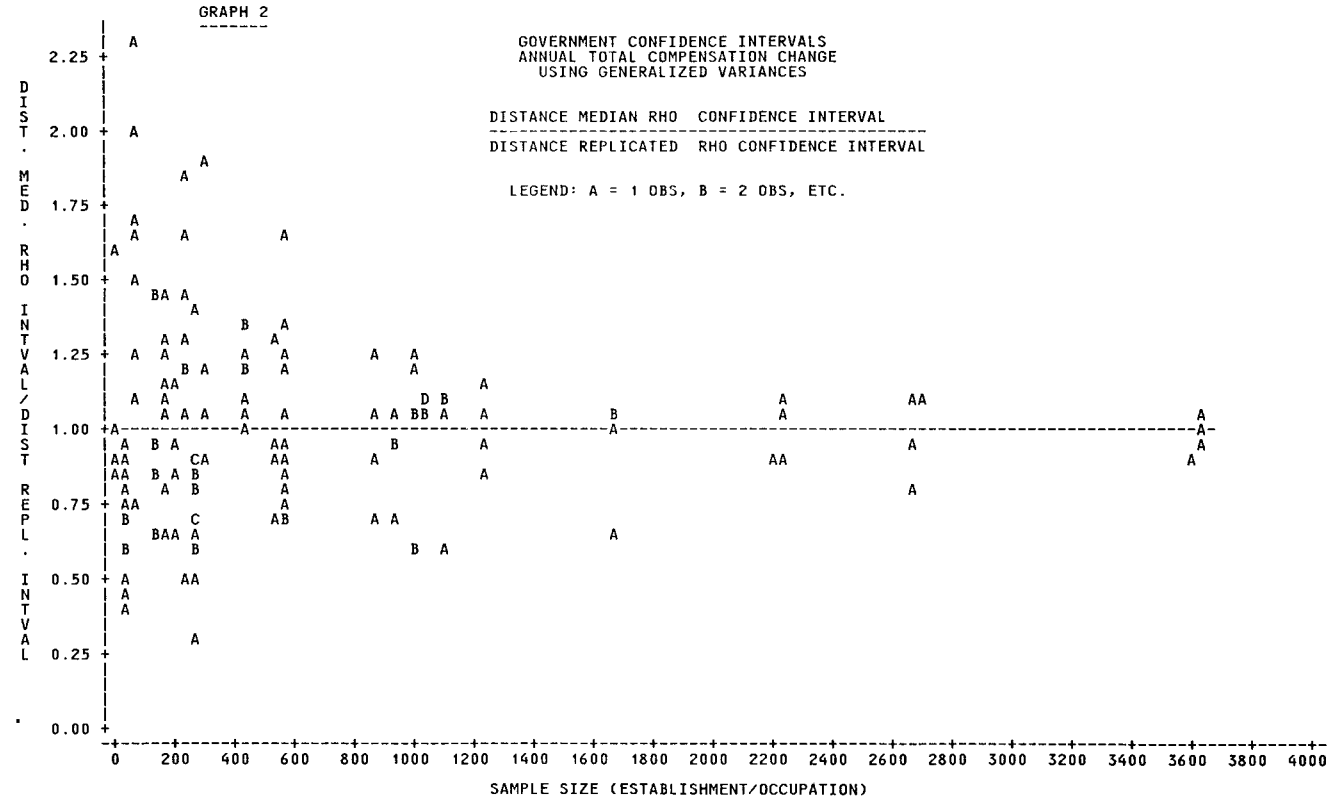
Quarter	Employer Cost Alloc. Std. Error	Opt. Alloc.** Std. Error	Opt.** Alloc. Effic.	Ave. Opt.** Std. Error	Ave.** Opt. Effic.
Dec. 82	0.0025	0.00232	6.9	0.0024	3.5
March 83	0.0029	0.00255	11.5	0.0026	9.8
June 83	0.0026	0.00234	8.1	0.00237	7.2
Sept. 83	0.0028	0.00244	11.0	0.00261	4.8

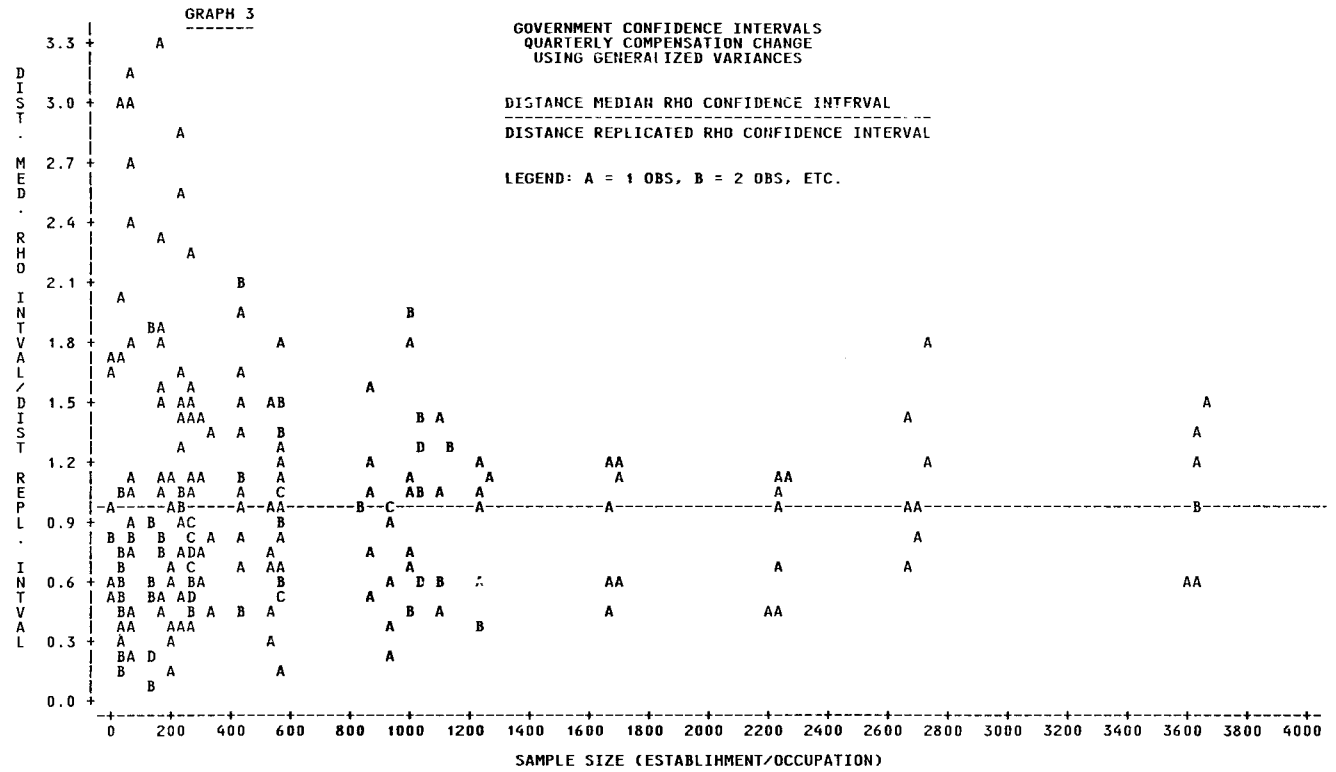
* These efficiencies are computed with respect to the Generalized Standard Error using the total employer cost allocation.

** The optimum allocations assume equal stratum collection costs (C_k)

GRAPH 1




GRAPH 2
GOVERNMENT CONFIDENCE INTERVALS
ANNUAL TOTAL COMPENSATION CHANGE
USING GENERALIZED VARIANCES
DISTANCE MEDIAN RHO CONFIDENCE INTERVAL
DISTANCE REPLICATED RHO CONFIDENCE INTERVAL
LEGEND: A = 1 OBS, B = 2 OBS, ETC.
DIST. MED. RHO INTVAL/DIST REPL. INTVAL
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 4000
SAMPLE SIZE (ESTABLISHMENT/OCCUPATION)


GRAPH 3
GOVERNMENT CONFIDENCE INTERVALS
QUARTERLY COMPENSATION CHANGE
USING GENERALIZED VARIANCES
DISTANCE MEDIAN RHO CONFIDENCE INTERVAL
DISTANCE REPLICATED RHO CONFIDENCE INTERVAL
LEGEND: A = 1 OBS, B = 2 OBS, ETC.
DIST. MED. RHO INTVAL/DIST REPL. INTVAL
3.3
3.0
2.7
2.4
2.1
1.8
1.5
1.2
0.9
0.6
0.3
0.0
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800 3000 3200 3400 3600 3800 4000
SAMPLE SIZE (ESTABLIHMENT/OCCUPATION)