ON LINK RELATIVE ESTIMATORS

Lillian H. Madow, U.S. Bureau of Labor Statistics
and William G. Madow, Washington, D.C.

1. Introduction

The type of estimator that forms the background for this paper on link relative estimators is one which uses a benchmark obtained periodically together with a survey estimate of change for time periods between benchmarks.

A benchmark is an essentially complete total for a population and may be obtained from administrative records, or censuses, or surveys large enough to be considered sample censuses. The information from which benchmarks are obtained may also provide the entire or a major part of the basic frame or list for the current (intervening) surveys.

By a link relative we mean the ratio of a total for a given period to the total for the same variable in the preceding period for units reporting in both periods. The link relative determines relative change from one time period to the next. The link relative estimator of a total is the product of a benchmark and the link relatives for the periods of time between the benchmark and the current period.

The Bureau of Labor Statistics Current Employment Statistics Program uses, essentially, a link relative estimator. This program provides monthly estimates of employment, hours and earnings of workers on nonagricultural establishment payrolls. Benchmark employment is obtained every year or so from Unemployment Insurance administrative records. Monthly estimates of change between benchmarks are obtained from a large voluntary monthly mail survey, known as the 790 Survey because of its schedule number. The 790 Survey data are obtained from cooperating establishments on a voluntary mail "shuttle" schedule. Descriptions of the Current Employment Statistics Program and the 790 Survey are available [1].

From the definition of link relative estimators, it follows that link relative estimators may yield biased estimators of change. Establishments change over time. They may go in and out of business. They may change through mergers, splits, purchases and sales (sometimes enough to change their industry classification). Unless the resulting biases are cumulatively very small, it is desirable to use a supplementary sample or an adjustment procedure or both. The 790 Survey currently depends primarily on an adjustment procedure with some updating largely at benchmark times. These topics are not further discussed in this paper. Rather, this paper is limited to the discussion of link relative estimators from the point of view of a simple statistical model.

Although the present paper is limited to statistical models, we emphasize that the user of a model must and does recognize that any model is unlikely to be correct. Even if initially correct, a model may become incorrect because of changes over time. Various industries and establishments of various sizes within an industry may be differentially affected by changes in the economy, perhaps because different regulations apply, perhaps because changes in technology affect some establishments early. For these and other reasons, models possibly applicable in one industry or one time period may not be applicable in another. A properly maintained probability sample design may have a larger or smaller mean square error (MSE) over time but will not become inconsistent because of such changes. However, probability sample designs also have problems. It is costly and not always possible to maintain probability samples of establishments, both because establishments change and because cumulative nonparticipation may occur over time. The effects of nonparticipation cannot be reduced in certainty strata or in strata with large sampling ratios. Rotation methods, for example, obviously cannot help in certainty strata.

If a large enough proportion of the benchmark total is in the sample, e.g., a high proportion of the large establishments report, and if the error in the model is not large for the remainder, then a model dependent design (when carefully monitored for changes) may provide good estimates of population characteristics, and of the associated mean square errors.

Our point of view also applies to the treatment of nonresponse or missing data. In general, establishments that do not respond should not be assumed to be a random sample from the probability sample. One may then ask, "Can a statistical model be formulated such that, for that model, the incomplete data are close enough to 'missing at random', for useful estimators and evaluations to be obtained? Such procedures can be studied and evaluated for surveys for which data for nonrespondents and benchmarks become available, from time to time, as in the "790" survey.

The methods used clearly generalize to the utilization of data for reporting units for which data are available at non-consecutive periods. These estimators will be discussed in a further paper.

2. Definitions and expected values. Model A.

The elements (establishments) of the population are denoted [2] by 1, 2, ..., N, and Y_{gi} is the random variable in which we are interested for element (establishment) i at time g. The population total for time g will be denoted by Y_g, where

$$(2.1) \quad Y_g = \sum_{i=1}^{N} Y_{gi} , \quad g = 0,1,\ldots, t$$

and it is desired to estimate Y_g or functions of $Y_0, Y_1, \ldots, Y_t$. (Since Y_g is a random variable, it might be preferable to speak of predictors rather than estimators; no confusion should result from our terminology.)

We suppose that at time 0, a benchmark, Y_0^*, is available, i.e. Y_0^* is "close" to Y_0, the total of $Y_{01}, \ldots, Y_{0N}$. The set of elements reporting at time, g, for which the Y_{gi} needed for estimation are available both for times g and g-1 is denoted by s_g, g = 1, 2,..., t.

Define

(2.2) $Y'_{gg} = \sum_{i \in s_g} Y_{gi}$

and

(2.3) $Y'_{g-1\ g} = \sum_{i \in s_g} Y_{g-1\ i}$.

Thus, Y'_{gg} and $Y'_{g-1\ g}$ are sums of the values of Y for the same establishments in periods, g-1 and g. For some elements of the sample, data may be available at one but not both times; such elements are excluded from the estimator in the present section, but should result in an improved estimator, depending on the cost, time and methodology used.

The sample link relative, R'_g, is defined by the equation,

(2.4) $R'_g = \frac{Y'_{gg}}{Y'_{g-1\ g}}$.

It is noted that

(2.5) $Y_g = Y_0 R_1 R_2 \ldots R_g$,

where

(2.6) $R_a = \frac{Y_a}{Y_{a-1}}$, a = 1, ..., g .

Because of the reasons discussed in the preceding section, i.e. the population's change over time, R'_a will usually be a biased estimator of R_a, and, hence, special samples or adjustment factors will, in practise, be needed to avoid or reduce the possibly large and cumulative biases. Generally, the benchmark, Y_0^*, will be a good estimate of Y_0.

We define

(2.7) $Y'_g = Y_0^* R'_1 \ldots R'_g = Y'_{g-1} R'_g$

to be the link relative estimator of Y_g, g=1, 2, ..., t. Thus, the link relative estimator is the product of the benchmark and ratios of random variables.

Realized values of Y will be denoted by y, i.e., when the random process having outcomes, $Y_{g1}, \ldots, Y_{gN}$, at time, g, is performed, the realized outcomes are denoted by $y_{g1}, \ldots, y_{gN}$.

The ascertainment of the values, y_{gi}, $i \in s_g$, leads to further sources of randomness, the nonsampling errors.

Model A. The random variables, Y_{gi}, g=1,..., t, i=1,...,N, and ξ, their joint distribution, are said to constitute Model A, if the random vector, $Y_g = (Y_{g1}, Y_{g2}, \ldots, Y_{gN})$, has distribution, ξ_g (a marginal distribtuion of ξ) satisfying the conditions

(a) $Y_{g1}, Y_{g2}, \ldots, Y_{gN}$ are uncorrelated

(b) $\mathcal{E}[Y_{gi} \mid (g-1)] = \beta_g Y_{g-1\ i}$

where β_g is a constant and (g-1) fixes the values of Y_{aj}, a=0,1,...,g-1; j=1,2,...,N.

(c) $\sigma^2_{Y_{gi}} = \sigma^2_g Y_{g-1\ i}$

where σ^2_g is a constant, g = 1,2,...,t.

Thus, if the model holds, then

(2.8) $\mathcal{E} Y'_{gg} \mid (g-1) = \beta_g Y'_{g-1\ g}$

and

(2.9) $\sigma^2_{Y'_{gg} \mid (g-1)} = \sigma^2_g Y'_{g-1\ g}$.

The simplicity of the model and the ease of applying it, as well as the frequently useful logic of assuming that the future is in large part proportionate to the past, make it a very tempting model to consider. Model A is a first step in finding a model that is both satisfactory for interpolation between benchmarks and yields satisfactory estimators in current periods.

3. Bias

Let us now consider the expected values of link relatives and link relative estimators.

Let $\sigma_{Y'_{a-1} R'_a} = 0$, a=1,...,g.

Then, writing Y_0^* in place of Y'_0 to indicate the benchmark, we have

(3.1) $\mathcal{E} Y'_g = \mathcal{E} Y_0^* \mathcal{E} R'_1 \ldots \mathcal{E} R'_g$.

More generally,

(3.2) $\mathcal{E} Y'_g = \mathcal{E} Y_0^* \mathcal{E} R'_1 \ldots \mathcal{E} R'_g + \sum_{a=0}^{g-1} \sigma_{Y'_a R'_{a+1}} \mathcal{E} R'_{a+2} \ldots \mathcal{E} R'_g$.

These results follow from

(3.3) $\mathcal{E} Y'_g = \mathcal{E} Y'_{g-1} R'_g = \mathcal{E} R'_g \mathcal{E} Y'_{g-1} + \sigma_{R'_g Y'_{g-1}}$.

If condition (b) of Model A is assumed, then

$\mathcal{E} Y'_{aa} = \beta_a \mathcal{E} Y'_{a-1\ a}$, $\mathcal{E} Y_a = \beta_a \mathcal{E} Y_{a-1}$,

and hence $\mathcal{E} R'_a = \mathcal{E} R_a = \beta_a$.

Also, under the same assumption, it follows that

$\sigma_{Y'_a R'_{a+1}} = \sigma_{Y'_{a-1\ a} R'_a} = 0$.

Hence, if it is assumed that $\mathcal{E}Y_0^* = \mathcal{E}Y_0 = \beta_0$, then Y_g' is an unbiased estimator of Y_g, i.e.,

$$\mathcal{E}Y_g' = \mathcal{E}Y_g = \beta_0 \beta_1 \cdots \beta_g .$$

However, if Y_0^* is biased, then Y_g' is biased and

$$\mathcal{E}Y_g' = \mathcal{E}Y_g (\mathcal{E}Y_0^*/ \beta_0) .$$

We note that if Y_a' is unbiased, a=0,1,...,t, then it follows that $Y_g' - Y_{g-1}'$ is an unbiased estimator of the change, $Y_g - Y_{g-1}$, and that $R_g' - 1$ is an unbiased estimator of relative change,

$$(Y_g - Y_{g-1})/Y_{g-1} .$$

Whether (b) holds is an empirical question and should be reexamined in any survey for which benchmarks and data for all establishments are frequently available in addition to the continuing samples. If (b) is incorrect,e.g.,$\mathcal{E}[Y_{gi}|(i-1)]$ is a polynomial in $Y_{g-1\ i}$, then the link relative estimator will become a biased estimator of Y_g. In this case, the balanced or over-balanced samples suggested by Royall and Herson (1973) and by Scott, Brewer and Ho (1978) for ratio estimators rather than link relatives are not practical for link relative estimators, since even if balanced samples are selected at time 0, they are most unlikely to be balanced at time g.

4. Mean Square Errors

In this section, benchmark comparisons are first discussed in subsection a. These comparisons provide "one degree of freedom" estimates of mean square errors between the estimates and the benchmarks. However, benchmarks may be available only some time after a survey is made and are available only at certain intervals; and in any case, it is desirable to have estimates of mean square error derived from the samples themselves.

In subsection b, variance estimators are derived based on Model A. These estimators are simple generalizations of Royall and Eberhardt (1975). The benchmark comparisons may be used to indicate whether the variance estimators are biased.

a. Benchmark Comparisons. If benchmarks are obtained at various times, e.g. at intervals of t, then at times t, 2t, ..., estimates of the mean square error of Y_{ct}' about Y_{ct} may be obtained, where Y_{ct} denotes the benchmark at time ct, c=1,2,... . Then,

$$(4.1) \quad M'^2 = (Y_{ct}' - Y_{ct}^*)^2$$

is an estimator of

$$(4.2) \quad M^2 = \mathcal{E}(Y_{ct}' - Y_{ct}^*)^2$$

where the asterisk (*) is used to indicate that the benchmark also may be in error.

If the benchmark is not in error, then M'^2 is a "one degree of freedom" unbiased estimator of the MSE.

If Y_{ct}' is an unbiased estimator of Y_{ct}^* and

$$\mathcal{E}(Y_{ct}'-Y_{ct})(Y_{ct}^*-Y_{ct}) = \mathcal{E}(Y_{ct}^*-Y_{ct})^2 ,$$

then

$$(4.3) \quad \mathcal{E}(Y_{ct}'-Y_{ct}^*)^2 = \mathcal{E}(Y_{ct}'-Y_{ct})^2 - \mathcal{E}(Y_{ct}^*-Y_{ct})^2$$

and M'^2 underestimates M^2, but ordinarily the second term on the right of (4.3) will be small compared to the first, unless Y_{ct}^* is badly biased. Thus, the importance of close agreement between the benchmark and Y_{ct} is emphasized.

On the other hand, if both Y_{ct}^* and Y_{ct}' are unbiased, estimators of Y_{ct} and the deviations of Y_{ct}^* and Y_{ct}' about Y_{ct} are uncorrelated, then

$$(4.4) \quad \mathcal{E}(Y_{ct}'-Y_{ct}^*)^2 = \mathcal{E}(Y_{ct}'-Y_{ct})^2 + \mathcal{E}(Y_{ct}^*-Y_{ct})^2;$$

the upward bias would be even larger, if errors in Y_{ct}' and Y_{ct}^* are negatively correlated. Again, the second term on the right of (4.4) should be small compared to the first.

When benchmarks are available for strata or subpopulations at times ct, c = 1,2, ..., additional benchmark comparisons may be made.

Thus, benchmark comparisons provide estimates of the mean squared errors, but may themselves be unreliable, because of being "one degree of freedom" comparisons. They also include the effects of nonsampling errors resulting from differences in procedures that may be used to obtain the estimators, Y_{ct}', and the benchmarks, Y_{ct}^* .

A major use for benchmark comparisons occurs when the s_g are de facto not probability samples. If enough benchmark comparisons, (4.1), are made, the averages and distributions of the comparisons provide useful evaluative information.

Benchmark comparisons may also be used to evaluate estimates of the mean square error based on the samples themselves in order to determine whether model-dependent biases are causing the mean square errors based on the models to have large errors. Choices may be made among alternative models, using agreement between benchmark comparisons and model dependent estimates of mean square errors as a basis for choice. Finally, there will be greater confidence in inferences concerning mean square errors for times between benchmark comparisons, if the model dependent mean square errors are sufficiently in agreement with the benchmark comparison mean square errors. When, in addition to the benchmarks, the data for the elements of the entire population at benchmark periods are available, the evaluation and revision of the model based procedures can be much more thorough.

b. Mean square error for a given sample. The sets s_1, s_2, ..., s_k, are considered fixed. Only the

Y_{gi}, g=1, ..., k, $i \in s_g$ are considered random.

Let us define

$$(4.5) \quad Y^*_g = \frac{Y^*_0}{Y_0} Y_g$$

and assume Model A.

Then,

$$(4.6) \quad \mathcal{E} Y'_g = \beta_0 \beta_1 \cdots \beta_g ,$$

if we define

$$\beta_0 = \mathcal{E} Y^*_0 .$$

If we assume, as before, that

$$\mathcal{E}[Y_g \mid (g-1)] = \beta_g Y_{g-1},$$

then

$$\mathcal{E} Y^*_g = \mathcal{E} Y'_g .$$

Hence, the mean square error of Y'_g about Y^*_g is

$$M^2_{Y'_g} = \xi \sigma^2_{Y'_g - Y^*_g} .$$

(since all variances in this section are with respect to ξ, the symbol, ξ, will be omitted.)

Then, from Model A, it follows that

$$(4.7) \quad RM^2_{Y'_g} = \frac{\sigma^2_{Y'_g - Y^*_g}}{\mathcal{E} Y^{*2}_g}$$

$$= \frac{\sigma^2_{Y^*_0}}{\mathcal{E} Y^{*2}_0} + \sum_{a=1}^{g} \frac{\mathcal{E}\sigma^2_{Y'_a - Y^*_a \mid (a-1)}}{\mathcal{E} Y^{*2}_a}$$

where $\sigma^2_{Y^*_0}$ is the variance of the benchmark, Y^*_0.

Let

$$(4.8) \quad u^2_a = \frac{\sum_{i \in s_a} \left(Y_{ai} - \frac{Y'_{aa}}{Y'_{a-1\,a}} Y_{a-1\,i}\right)^2}{(n_a - 1)\, \bar{Y}'^2_{a-1\,a} \,(1 - V^2_{a-1\,a})} ,$$

where n_a is the number of elements in s_a,

$\bar{Y}'_{a-1\,a} = Y'_{a-1\,a} / n_a$

and

$$(4.9) \quad V^2_{a-1\,a} = \frac{\sum_{i \in s_a} (Y_{a-1\,i} - \bar{Y}'_{a-1\,a})^2}{(n_a - 1)\, n_a \, \bar{Y}'^2_{a-1\,a}} .$$

Also, let

$$U^2_a = u^2_a \frac{Y'_{a-1}\,(Y'_{a-1} - Y'_{a-1\,a})}{Y'^2_{a-1\,a}} .$$

Then

$$\mathcal{E}[U^2_a \mid (a-1)] = \sigma^2_a \frac{Y'_{a-1}(Y'_{a-1} - Y'_{a-1\,a})}{Y'^2_{a-1\,a}}$$

$$= \sigma^2_{Y'_a - Y^*_a \mid (a-1)} .$$

Thus, a useful estimator of $RM^2_{Y'_g}$ is given by

$$(4.10) \quad \frac{s^2_{Y'_g - Y^*_g}}{Y'^2_g} = \frac{s^2_{Y^*_0}}{Y^{*2}_0} + \sum_{a=1}^{g} \frac{U^2_a}{Y'^2_a} ,$$

where $s^2_{Y^*_0}$ is assumed to be an unbiased estimator of $\sigma^2_{Y^*_0}$. The estimator (4.10) is consistent under general conditions. An unbiased estimator of $\sigma^2_{Y'_g - Y^*_g}$ is given by

$$(4.11) \quad s^2_{Y'_g - Y^*_g} = \prod_{a=1}^{g} (R'^2_a - \frac{u^2_a}{Y'^2_{a-1\,a}}) \, s^2_{Y^*_0}$$

$$+ \sum_{a=2}^{g} \prod_{f=a}^{g} (R'^2_f - \frac{u^2_f}{Y'^2_{f-1\,f}}) U^2_{a-1} .$$

Equation (4.10) generalizes the results obtained for ratio estimators by Royall and Eberhardt (1975).

Also, we note that, if it is desired to retain (b) of Model A, but define

$$\sigma^2_{Y_{gi}} = \sigma^2_g V_g(Y_{g-1\,i})$$

where $V_g(Y_{g-1\,i})$ is not $Y_{g-1\,i}$ but some other function of $Y_{g-1\,i}$, and to use estimators such as

$$Y'_g = Y_{gg} + R''_g (Y'_{g-1} - Y_{g-1\,g})$$

where

$$R''_g = \frac{\sum_{i \in s_g} \frac{Y_{gi}\, Y_{g-1\,i}}{V_g(Y_{g-1\,i})}}{\sum_{i \in s_g} \frac{Y^2_{g-1\,i}}{V_g(Y_{g-1\,i})}} ,$$

then, if (b) of Model A is retained, Y'_g is an unbiased predictor of Y_g, but the mean square error becomes intractable. Thus, if Model A is to be used, monitoring and possibly stratification to increase the likelihood that Model A holds within the resulting strata are essential.

The use of model-dependent estimators of mean squares, thus, requires frequent comparison with check data. One major reason for this is that sample surveys made over time are not controlled experiments, even less controlled than are clinical experiments. As a result, one cannot, as in the case of industrial products, establish a state of statistical control and then have some certainty that the assumption required by the statistical methods will continue to be sufficiently valid. The periodic availability of benchmark comparisons and data make it possible to determine whether the assumptions of the given model are resulting in gross under-estimation of the "true" MSE's. Between benchmarks, the assumptions made by the model can be examined for

the reporting establishments, although with less confidence in the conclusions.

When a survey produces many individually benchmarked estimates based on independent samples from many industries, the number of such benchmark comparisons may be large enough for confidence in the outcomes of studies such as we have suggested.

5. James-Stein Type Estimators

One important part of the "790" survey design is that estimators are prepared and benchmarked for each of 846 estimating cells corresponding to different industries or, in some cases, to industries classified by geographic region and size of establishment. Improvements in the estimates for individual industries may be attainable through the use of James-Stein type estimators, just as in other problems involving several estimators.

The results of this section are general and are not limited to link relative estimators. Also, space permits only a minimal listing of results.

Let Y'_{gm} be an unbiased estimator of Y_{gm}, where g identifies the time and m = 1,2,...,M for a specified M estimator, e.g., M may be the number of small areas or industries. It is not necessary that M be the total number of estimates to be made.

James-Stein type estimators are obtained below, assuming Y'_{gm} and Y_{gm} are random variables. Similar results have been obtained for realizations using a probability sample design approach and will be discussed in another paper.

The James-Stein type estimators will be obtained by first minimizing

$$(5.1) \quad \sum_{m=1}^{M} W_{gm} \mathcal{E}[Y_{gm}-Y'_{gm}-c_g (Y'_{gm} - \bar{Y}'_g)]^2$$

for c_g, where the weights W_{gm} satisfy

$W_{gm} \geq 0$, $\sum_{m=1}^{M} W_{gm} = 1$, and

$$(5.2) \quad \bar{Y}'_g = \sum_m W_{gm} Y'_{gm}$$

and then replacing the value $\hat{c}_g$ of c_g thus obtained by a ratio estimator of $\hat{c}_g$. As in regression, the mean square error, taking account of the ratio estimator, is complicated. Let us assume that Y'_{gm} is a ξ-unbiased estimator of Y_{gm}. Now, let us suppose that $\sigma_{Y'_{gm}Y_{gm}} = \sigma^2_{Y_{gm}}$, as is true for link relative estimators satisfying Model A, and define

$$(5.3) \quad s_1^2 = \sum_{m=1}^{M} W_{gm}(1-W_{gm})\, s^2_{Y'_{gm}-Y_{gm}}$$

$$(5.4) \quad s_2^2 = \sum_{m=1}^{M} W_{gm}(Y'_{gm} - \bar{Y}'_g)^2,$$

$$s_3^2 = \sum_{m=1}^{M} W_{gm}(Y_{gm} - \bar{Y}_g)^2$$

where, for link relative estimators, the $s^2_{Y'_{gm}-Y_{gm}}$ have been obtained in Section 4.

Then, $\mathcal{E}s_2^2 = \mathcal{E}s_1^2 + \mathcal{E}s_3^2$.

Also, the estimators

$$(5.5) \quad Y''_{gm} = Y'_{gm} - \frac{s_1^2}{s_2^2}(Y'_{gm}-\bar{Y}'_g), \quad m=1,2,\ldots,M,$$

will have smaller total mean square errors than the estimators, s, Y'_{gm}. In fact,

$$(5.6) \quad \sum_{m=1}^{M} W_{gm}\mathcal{E}(Y''_{gm}-Y_{gm})^2 = \sum_{m=1}^{M} W_{gm}\mathcal{E}(Y'_{gm}-Y_{gm})^2 - \frac{(\mathcal{E}s_1^2)^2}{\mathcal{E}s_2^2}$$

Often, it is desirable to have $W_{gm} = \frac{1}{M}$. Sometimes the W_{gm} are chosen to minimize $\mathcal{E}(\bar{Y}'_g-\bar{Y}_g)$. Then

$$W_{gm} = \frac{K}{\mathcal{E}(Y'_{gm}-Y_{gm})^2}$$

where $K^{-1} = \sum_m \frac{1}{\mathcal{E}(Y'_{gm}-Y_{gm})^2}$

Then

$$Y''_{gm} = Y'_{gm} - \frac{M-1}{\sum_{m=1}^{M} \frac{\mathcal{E}(Y'_{gm}-\bar{Y}'_g)^2}{\mathcal{E}(Y'_{gm}-Y_{gm})^2}}(Y'_{gm}-\bar{Y}'_g)$$

and

$$\sum_{m=1}^{M} W_{gm}\mathcal{E}(Y''_{gm}-Y_{gm})^2 = \frac{1}{K}\left[M - \frac{(M-1)^2}{\sum_{m=1}^{M} \frac{(Y'_{gm}-\bar{Y}'_g)^2}{(Y'_{gm}-Y_{gm})^2}}\right]$$

The expected reduction in total mean square errors will be somewhat less when the computable estimators[3], Y'''_{gm}, are used, where, if, $(s_1^2/s_2^2) \leq 1$, then, by definition,

$$(5.7) \quad Y'''_{gm} = Y'_{gm} - \frac{s_1^2}{s_2^2}(Y'_{gm} - \bar{Y}'_g),$$

and if $\frac{s_1^2}{s_2^2} > 1$, then by definition $Y'''_{gm} = \bar{Y}'_g$.

If, instead of minimizing the weighted sum of squares (5.1), we minimize

$$(5.8)\ \mathcal{E}\,[Y_{gm} - Y'_{gm} + c_{gm}(Y'_{gm} - \bar{Y}'_{g})]^2 ,$$

then,

$$(5.9)\quad \hat{c}_{gm} = \frac{\mathcal{E}s^2_{gm}}{\mathcal{E}s^2_{2m}} ,$$

where,

$$s^2_{gm} = (1 - W_{gm})\, s^2_{Y'_{gm} - Y_{gm}}$$

$$(5.10)\quad s^2_{2gm} = (Y'_{gm} - \bar{Y}'_{gm})^2 .$$

However, s^2_{2gm} is likely to have a large variance as an estimator of $\mathcal{E}s^2_{2m}$ and, hence, a hybrid of Y''' and Y^{IV} is suggested, namely,

$$(5.11)\quad Y^{V}_{gm} = Y'_{gm} - \frac{s^2_{gm}}{s^2_2}(Y'_{gm} - \bar{Y}'_{g}),$$

if $0 < f < 1$, where

$$(5.12)\quad f = \frac{s^2_{gm}}{s^2_2}\left[2(1 - W_{gm}) - \frac{s^2_{2gm}}{s^2_2}\right].$$

If $f \leq 0$, then by definition $Y^{V}_{gm} = Y'_{gm}$, and if $f \geq 1$, then by definition, $Y^{V}_{gm} = \bar{Y}'_{gm}$. The criterion, f, has been chosen by requiring that

$$(5.13)\quad \mathcal{E}\left[Y_{gm} - Y'_{gm} + \frac{\mathcal{E}s^2_{gm}}{\mathcal{E}s^2_2}(Y'_{gm} - \bar{Y}'_{g})\right]^2 \leq \mathcal{E}(Y_{gm} - Y'_{gm})^2$$

and replacing $\mathcal{E}s^2_{gm}$, $\mathcal{E}s^2_2$ and $\mathcal{E}s^2_{2gm}$ by s^2_{gm}, s^2_2 and s^2_{2gm} in the result.

We now show how to obtain approximate confidence intervals for Y_{gm}. Suppose that the variables have been defined so that for large enough samples and subpopulations, s^2_{gm} converges in probability to 0, $s_{gm}\frac{(Y'_{gm} - \bar{Y}'_{gm})}{s^2_2}$ converges in probability to 0, and

$$(5.14)\quad \frac{Y'_{gm} - Y_{gm}}{s_{gm}}$$

is approximately normally distributed. Then, ignoring the censoring in the definition of Y^{V}_{gm}, it follows that

$$(5.15)\quad Z = \frac{Y^{V}_{gm} - Y_{gm}}{s_{gm}(1 - f)^{\frac{1}{2}}}$$

is the sum of a random variable having normal limiting distribution (with 0 mean and variance, $[1 - f]$ and a random variable converging in probability to 0; hence, Z is approximately normally distributed for large samples (0,1). Thus, an approximate confidence interval for Y_{gm} may then be obtained from (5.15). Also, it is noted that under the assumptions made above, Y^{V} is consistent, if Y'_{gm} is consistent. Similar results may be obtained for Y'''_{gm} by making the same kinds of assumptions.

Such procedures are also available for probability sampling designs. It will be noted that, as the size of sample within a subpopulation or the sample proportion of the measured variable increases, e.g., employment for the subpopulation, increases, the benefits of the J-S estimators will decrease. This is expected, since we are attempting to estimate Y_{gm}, and the "within" sample is close to the population value, under these conditions. The samller the "within" variance of an unbiased estimator, the smaller the benefit to be gained from additional information such as the values of other estimators.

James-Stein estimators can also be stated for estimating change, either applying the approach to link relatives R'_{gm}, $m = 1,2,\ldots,M$, or to the differences, $Y'_{gm} - Y'_{g-1\,m}$. Both methods, however, have possible instabilities arising from small values of the terms corresponding to s^2_2 above. The use of either $Y'''_{gm} - Y'''_{g-1\,m}$ or $Y^{V}_{gm} - Y^{V}_{g-1\,m}$, where the estimates are independently calculated, may lead to changes of direction that are contrary to the evidence provided by the link relatives. To obtain James-Stein estimators that preserve direction may require the use of mathematical programming methods.

FOOTNOTES

1/ a. Bureau of Labor Statistics Bulletin 1910, BLS Handbook of Methods (1976) Ch. 3, "Employment, Hours and Earnings."
b. Employment and Earnings, BLS, Monthly, Explanatory Notes, Establishment Data.
c. Madow, L. H., "An Error Profile: Employment as Measured by the Current Employment Statistics Program," ASA Proceedings of the Social Statistics Section, 1977, Part I, pp. 35-41.

2/ The unrealistic assumption of an unchanging population is made to avoid many complications while disussing properties of link relative estimators.

3/ The replacement of expected values by their estimators will produce good or poor results depending on the closeness with which the estimators approximate their expected values.

REFERENCES

Royall, R.M. and Eberhardt, K.R. (1975) "Variance Estimates for the Ratio Estimator," Sankhya, Ser. C, 37, 43-52.

Royall, R.M. and Herson, J. (1973) "Robust Estimation in Finite Populations I," Journal of the American Statistical Association, 68, 880-889.

Scott, A.J., Brewer, K.R.W., and Ho, E.W.H. (1978) "Finite Population Sampling and Robust Estimation", Journal of the American Statistical Association, 73, 359-361.