A PREDICTION THEORETIC VIEW OF COMPOSITE ESTIMATION

Stephen M. Woodruff, Bureau of Labor Statistics

1. INTRODUCTION

The use of supplementary information to increase the accuracy of estimates made from sample survey data is an important foundation of both superpopulation prediction theory (also refered to as model based sampling theory) and composite estimation. The background for superpopulation prediction theory is covered in papers by Royall, Cassel, Särndal, and Wretman in the reference list. Composite estimation is discussed in the papers by Wolter and Woodruff.

Composite estimation (as discussed by Wolter, Woodruff, and others) is used in repeated surveys where data is collected at regular intervals (each month, quarter, or year). In composite estimation, sample data from previous time periods is used to improve survey estimates for the current time. Implicit in the theory which justifies the use of composite estimation is the assumption that data collected at time t from a sample unit is highly correlated with the data from the same unit at time t-1. This assumption allows an alternate estimate of level at time t to be constructed from an update of the estimate at time t-1. This alternate estimate of level at time t can then be averaged in an appropriate way with the usual survey estimate at time t to get what is known as a composite estimator of level at time t. This work on composite estimation is founded on sampling distributions alone and does not directly use superpopulation models. The theory of survey sampling based on artifical randomization in the sample selection process (also refered to as design based sampling theory) is covered in the texts by Hansen, Hurwitz, & Madow, Kish, Raj, and Cochran which appear in the reference list.

In prediction theory a superpopulation model based on auxillary information or variables is used to improve the accuracy of survey estimates. In this case the auxillary information is available for each unit in the universe and is related to the sample data by the superpopulation model. A linear superpopulation model will often imply a correlation structure between the survey variable and the auxiliary variable which is very similar to the correlation between the data from adjacent time periods that composite estimation depends upon. In this paper we will treat the data from the preceding time period as an auxiliary variable which is related to the data for the current time period by a superpopulation model.

An important difference between superpopulation prediction theory and composite estimation is that in prediction theory the value of the auxiliary variable is known for each unit in the population while in composite estimation the correlated data from the preceeding time period is known for only a sample of units in the population. Another important difference is that composite estimators are analyzed with respect to the sampling distribution while in prediction theory the sample is conditioned upon and the analysis is conducted with respect to the superpopulation distribution.

The purpose of this paper is to analyze composite estimation using not only the sampling distribution but also the implicit superpopulation distribution. Another way to think of this is as an extension of prediction thoery to situations where the auxiliary variable is known for only a sample of units from the universe. When the auxiliary variable is available for only a sample of units then it is no longer feasible to condition on the sampling outcome since without knowledge of the auxiliary variable for all units in the universe we cannot know which sample was selected. That is, inferences in prediction theory can condition on the sample because by comparing the sample values of the auxiliary variable with the nonsample values of the auxiliary variable we essentially know which sample was selected.

Therefore, in a prediction theory approach to composite estimation, estimators must be analyzed with respect to both the sampling distribution and the superpopulation distribution. The superpopulation distribution will usually suggest a best linear unbiased estimator (BLUE) for the updated component of the composite estimator. This model based composite estimator which uses the BLUE will in most cases be superior to the composite estimators suggested by design based sampling theory alone.

Section two of this paper defines a superpopulation model, derives a model based composite estimator based on this model, and analyzes its theoretical properties. A ratio composite estimator founded on design based sampling theory is also proposed as a standard for comparison with the model based composite estimator. In section three, these two composite estimators are compared both theoretically and empirically and the results of this comparison are tabulated.

2. SAMPLING SCHEME AND ESTIMATORS

Let Y_i denote a random variable attached to the i^{th} population unit where $1 \le i \le N$ and N is the size of population. Let y_i denote the realization of Y_i and let U denote this population (universe). In this paper we will be concerned with estimating the population mean, denoted $\bar{y}$, of the $\{y_i\}$.

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

For each of these random variables Y_i let x_i be the auxiliary variable such that the following linear model holds:

$$\mathcal{E}(Y_i) = \beta x_i \quad \text{for all} \quad 1 \le i \le N$$

$$\mathcal{V}(Y_i) = \sigma^2 x_i^2 \quad \text{for all} \quad 1 \le i \le N$$

where β and σ are positive constants independent of i, $\mathcal{E}(\ )$ denotes expectation, and $\mathcal{V}(\ )$ denotes variance. The $\{Y_i\}$ are uncorrelated.

Let s_x be the sample of population units from which the auxiliary variable, x, is observed and let s_y be the sample of units from which the Y variable is observed. In terms of a continuing survey y is the variable of interest for the current estimate of the population mean and x is the same variable for the preceeding time period. Thus s_y is the current sample and s_x is the sample for the preceeding time period. We will further assume that s_x is a simple random sample without

replacement and s_y consists of two components as follows. The first component is a simple random sample without replacement of size m from s_x and the second component is a simple random sample without replacement of size n-m from $\tilde{s}_x$. Where n is the sample size of both s_x and s_y. $\tilde{s}_x$ denotes the compliment of s_x in U. Note that s_y is unconditionally also a simple random sample without replacement although conditional on s_x, s_y is a stratified sample.

If the $\{x_i\}$ were known for all units in U then the BLUE for $\bar{Y}$ would be:

$$f\bar{y} + (1-f)\check{\beta}\bar{x}_c$$

where: $\check{\beta} = (1/n) \sum_{i\varepsilon s_y} (y_i/x_i)$,

$$\bar{x}_c = (1/(N-n)) \sum_{i\varepsilon U-s_y} x_i ,$$

and $f = n/N$.

$\bar{x}_c$ is unknown but $\hat{x}_c$ is, conditional on s_y, an unbiased estimate of $\bar{x}_c$.

where $\hat{x}_c = (1/(n-m)) \sum_{i\varepsilon \tilde{s}_y \cap s_x} x_i$

(Note $U-s_y = \tilde{s}_y$ = the complement of s_y in U)

Similarly, $\check{\beta}$ is unknown but $\hat{\beta}$ is, conditional on s_y, an unbiased estimate of $\check{\beta}$.

where $\hat{\beta} = (1/m) \sum_{i\varepsilon s_y \cap s_x} (y_i/x_i)$

If we replace $\check{\beta}$ and $\bar{x}_c$ in the model BLUE with their sampling estimates, $\hat{\beta}$ and $\hat{x}_c$, we get an estimator for $\bar{Y}$ with variability which comes from both the sampling distribution and the super-population distribution. In order to reflect this increase in variance we replace f with α where α is a real number chosen to minimize $E(\hat{e} - \bar{Y})^2$ where $\hat{e}$ is our prediction theory composite estimator and is defined as:

$$\hat{e} = \alpha\bar{y} + (1-\alpha)\hat{\beta}\hat{x}_c$$

Let:

$$E_1 = E(\hat{\beta}\hat{x}_c - \bar{Y})^2$$

$$E_2 = E(\bar{y} - \bar{Y})^2$$

$$E_3 = E[(\hat{\beta}\hat{x}_c - \bar{Y})(\bar{y} - \bar{Y})]$$

Then $E(\hat{e} - \bar{Y})^2 = \alpha^2E_2 + (1-\alpha)^2E_1 + 2\alpha(1-\alpha)E_3$ and the value of α that minimizes this expression is denoted α^o and is found to be:

$$\alpha^o = (E_1 - E_3)/(E_1 + E_2 - 2E_3)$$

Recall that the above expectations are taken with respect to the unconditional sampling distribution and the superpopulation distribution where the $\{x_i: 1\le i\le N\}$ are the auxiliary variables. Thus the samples and sample intersections contained in this estimator are simple random samples without replacement. Therefore, the expectations E_1, E_2, and E_3 are:

$$E_1 = (\{\sigma^2/m\} + \beta^2)[(\{1/[n-m]\}-\{1/N\})S_x^2 + \bar{X}^2]$$
$$+ (\{N-1\}/N^2)\sigma^2S_x^2 + \bar{X}^2(\{\sigma^2/N\} + \beta^2)$$
$$- 2\bar{X}^2(\{\sigma^2/N\} + \beta^2).$$

$$E_2 = (1/n - 1/N)(S_x^2\{\beta^2 + [(N-1)/N]\sigma^2\} + \sigma^2\bar{X}^2).$$

$$E_3 = (1/n - 1/N)\sigma^2\bar{X}^2$$

Where $S_x^2 = (1/\{N-1\}) \sum_{i\varepsilon U} (x_i - \bar{X})^2$

and $\bar{X} = (1/N) \sum_{i\varepsilon U} x_i$.

For N sufficiently large these three expectations can be closely approximated by:

$$E_1 \doteq (\{\sigma^2/m\} + \beta^2)(1/[n-m])S_x^2 + (\sigma^2/m)\bar{X}^2.$$

$$E_2 \doteq (1/n)(S_x^2\{\beta^2 + \sigma^2\} + \sigma^2\bar{X}^2).$$

$$E_3 \doteq (1/n)\sigma^2\bar{X}^2$$

$E(\hat{e} - \bar{Y})^2$ evaluated at $\alpha=\alpha^o$ is equal to:

$$(E_1E_2 - E_3^2)/(E_1 + E_2 - 2E_3)$$

If $\bar{y}$ and $\hat{e}$ are compared as estimators of $\bar{Y}$ then a measure of the relative improvement of $\hat{e}$ over $\bar{y}$ is the ratio:

$$E(\hat{e} - \bar{Y})^2/E_2 \text{ evaluated at } \alpha = \alpha^o$$

This ratio along with α^o are functions of two papameters. These papameters are $a = \sigma^2/\beta^2$ and $r = S_x^2/\bar{X}^2$. a measures the dispersion of the data generated by the superpopulation model given the auxiliary variable and r measures the dispersion of the auxiliary variable (x). The approximate sampling correlation between the pairs $\{(x_i,y_i): 1\le i\le N\}$ is also a function of a and r.

Recall that for a simple random sample without replacement from which two variables, (z,w), are observed for each sample unit we have that the

sampling correlation between these variables is:

$$\rho_{zw} = S_{zw}/[S_z^2 S_w^2]^{\frac{1}{2}}$$

Where $S_{zw} = (1/[N-1]) \sum_{i\varepsilon U} (z_i - \bar{Z})(w_i - \bar{W})$

If the variables z and w are replaced in the above expression by x and y then the sampling correlation between the auxiliary variables $\{x_i\}$ and the realizations of the random variables $\{Y_i\}$ is obtained. Denote this sampling correlation as S_{xy}. Then S_y^2 and S_{xy} (where S_y^2 is defined as S_{yy}) are themselves random variables through the superpopulation distribution on the set $\{Y_i: 1\leq i\leq N\}$. If S_{xy} and S_y^2 are replaced with their expected values in the formula for the sampling correlation, ρ_{xy}, and N is sufficiently large so that $(N-1)/N \doteq 1$, then this correlation can be approximated as:

$$\rho_{xy} \doteq \beta S_x^2/(S_x^4[\beta^2 + \sigma^2] + \sigma^2\bar{X}^2 S_x^2)^{\frac{1}{2}}$$

Rewriting ρ_{xy} in terms of a and r we get:

$$\rho_{xy} = \{1 + a + a/r\}^{-\frac{1}{2}}$$

This measure of sampling correlation will be useful later on for comparing the model based approach to composite estimation with the usual approach where only sampling variability is considered.

The optimal value of α, the relative gain, and the optimal value for the size of the overlap, m, will now be written as functions of a and r. Let $L=a(1/m - 1/n) + r(1/\{n-m\}) + ar(1/\{m[n-m]\})$ Then $\alpha^o = L/(L + r/n + ar/n)$ The relative gain of the composite estimator was defined as the ratio of the expected squared error of $\bar{y}$ to the expected squared error of $\hat{e}$ evaluated at α^o and this is:

$$E(\hat{e} - \bar{Y})^2/E(\bar{y} - \bar{Y})^2 =$$

$$[GH - a^2/n^2]/[H(G + r/n - a/n + ar/n)]$$

Where $G = a/m + r/(n-m) + ar/\{m(n-m)\}$ and
$H = (1/n)(r + ar + a)$
The value of m that minimizes this ratio is found to be:

$$m^o = [-a(r + n) + \{ra[r + n][a + n]\}^{\frac{1}{2}}]/[r-a]$$

This largely completes the theoretical analysis of $\hat{e}$ with respect to the superpopulation model.

Next we will compare $\hat{e}$ to the usual composite estimator for a sampling problem where a ratio estimator of change between adjacent time periods is suggested by the sampling correlation. This estimator is denoted $\hat{d}$ and is defined as follows:

$$\hat{d} = \lambda\bar{y} + (1-\lambda)\mathring{\beta}\bar{x}$$

Where $0 < \lambda < 1$, $\bar{x}=(1/n) \sum_{i\varepsilon s_x} x_i$,

and $\mathring{\beta}= \left[\sum_{i\varepsilon s_x \cap s_y} y_i\right]\left[\sum_{i\varepsilon s_x \cap s_y} x_i\right]^{-1}$

The optimal value of λ for minimizing $E_s(\hat{d} - \bar{Y})^2$, where E_s denotes the expectation with respect to the sampling distribution, is:

$$\lambda = \lambda^o = (A_1 - A_3)/(A_1 + A_2 - 2A_3)$$

Where $A_1 = E_s(\mathring{\beta}\bar{x} - \bar{Y})^2$

$$A_2 = E_s(\bar{y} - \bar{Y})^2$$

$$A_3 = E_s(\bar{y} - \bar{Y})(\mathring{\beta}\bar{x} - \bar{Y})$$

Note that as with design based finite population sampling theory the stochastic structure comes from the sampling distribution only. Thus the optimal value for λ is evaluated with respect to the sampling distribution and conditional on the outcome of the superpopulation distribution.

Let $R = \bar{Y}/\bar{X}$, where $\bar{X} =(1/N) \sum_{i\varepsilon U} x_i$, then:

$$A_1 = (\{1/m\} - \{1/N\})S_y^2 + (\{1/m\} - \{1/n\})R^2 S_x^2$$

$$+ 2(\{1/n\} - \{1/m\})RS_{xy}$$

$$A_2 = (\{1/n\} - \{1/N\})S_y^2$$

$$A3 = (\{1/n\} - \{1/N\})S_y^2 + RS_{xy}(\{m/n^2\} - \{1/n\})$$

Note that all three of these quantities are functions of things that are random variables with respect to the superpopulation model and therefore A_1, A_2, and A_3 are random under the model. This will present a small problem when we try to compare $\hat{e}$ and $\hat{d}$.

In order to slightly simplify a composite estimator like $\hat{d}$ it is often assumed that $S_x^2 = S_y^2 = S^2$ and if this is done in the case of $\hat{d}$, λ^o can be written as:

$$\lambda^o = w/z$$

Where $w = \left(1 + R^2 - (2 - \{m/n\})R\rho_{xy}\right)$

and $z = \left(1 + R^2 - (2 - \{2m/n\})R\rho_{xy}\right)$

With this value of λ^o, $E_s(\hat{d} - \bar{Y})^2$ is minimized for each outcome of the superpopulation model (y_1, y_2, ... y_N). Note that λ^o is itself a random variable through the superpopulation model. Thus

$E(\hat{d} - \bar{Y})^2 = \mathcal{E}E(\hat{d} - \bar{Y})^2$ is minimized for these values of λ^o which are dependent on $(y_1, y_2, \ldots\ldots y_N)$. In practice the composite estimator is relatively robust against misspecification of λ^o and the variability of λ^o over the different superpopulation outcomes is small, thus for purposes of comparing $\hat{e}$ and $\hat{d}$, λ^o will be choosen to minimize $E(\hat{d} - \bar{Y})^2$ rather than the more usual procedure of minimizing $E_s(\hat{d} - \bar{Y})^2$ for each different outcome of the superpopulation distribution. Technically the latter procedure is preferable but as stated above minor deviations from the optimal weight will have little effect on accuracy in most practical situations and with the λ^o based on minimizing $E(\hat{d} - \bar{Y})^2$ we can compare $\hat{e}$ and $\hat{d}$ for various superpopulation distributions. The primary reason for the superiority of the model based composite estimator, $\hat{e}$, over a more usual composite estimator like $\hat{d}$ is the form of the second component of $\hat{e}$ which is suggested by the BLUE under the given model.

The value of λ which minimizes $E(\hat{d} - \bar{Y})^2$ is:

$$\lambda' = \left[\mathcal{E}(A_1) - \mathcal{E}(A_3)\right]\left[\mathcal{E}(A_1) + \mathcal{E}(A_2) - 2\mathcal{E}(A_3)\right]^{-1}$$

If the universe size is sufficiently large and the population U can be thought of as increasing in such a way that each set $\mathcal{A}_N = \{x_i^3/\bar{X}^3: 1 \le i \le N\}$ is bounded above by a positive constant which is independent of N then for N sufficiently large we have:

$$\mathcal{E}(A_1)/\beta^2\bar{X}^2 = (a/m)(r + 1) + r/n$$

$$\mathcal{E}(A_2)/\beta^2\bar{X}^2 = (a/n)(r + 1) + r/n$$

$$\mathcal{E}(A_3)/\beta^2\bar{X}^2 = (a/n)(r + 1) + (m/n^2)r$$

With these expected values standardized by $\beta^2\bar{X}^2$ the optimal value for λ is:

$$\lambda' = \frac{a(r + 1)(1/m - 1/n) + r(1/n - \{m/n^2\})}{a(r + 1)(1/m - 1/n) + r(2/n - \{2m/n^2\})}$$

The relative gain from using $\hat{d}$ in place of $\bar{y}$ is:

$$E(\hat{d} - \bar{Y})^2/E(\bar{y} - \bar{Y})^2 = (\tau - \omega)/(\zeta\psi)$$

Where: $\tau = ((a/m)(r+1)+(r/n))((a/n)(r+1)+(r/n))$

$\omega = ((a/n)(r+1)+(mr/n^2))^2$

$\zeta = a(r+1)((1/m) - (1/n)) + r((2/n) - (2m/n^2))$

$\psi = (a/n)(r+1) + r/n$

In the next section we compare the relative gains of $\hat{d}$ and $\hat{e}$ for various values of a and r.

3. TABULAR RESULTS; THEORETICAL AND SIMULATED

In this section the relative gains of $\hat{d}$ and $\hat{e}$ will be compared for various values of a, r, and m. The theoretical relative gains of $\hat{d}$ and $\hat{e}$ at $m=m^o$. were derived in the previous section and are tabulated for certain pairs, (a,r). These theoretical relative gains are supported by an empirical study using a Bureau of Labor Statistics employment data set. Although the theoretical results may be legitimately criticized because they depend on an assumed superpopulation model the empirical results substantially support the theory even though the simulation data sets are clearly the result of a process which is only a very crude approximation of the superpopulation model of section two.

In the previous section the relative gains of $\hat{e}$ and $\hat{d}$ were derived as functions of a and r but these functions are a little to complex to be very informative. In order to see how these relative gains behave as functions of a and r they are tabulated in Table one for values of a and r between zero and .5. In these tabular comparisons $m=m^o$, the optimal value of m for $\hat{e}$, was used in the formula for the relative gains of both $\hat{e}$ and $\hat{d}$. It was found that m^o is an upper bound for the optimal overlap of $\hat{d}$ and that the relative gain of $\hat{d}$ at m^o was, for all practical purposes, the same as its relative gain at its optimal value of m (within one percent). Although the relative gain of $\hat{d}$ was reduced very little by using the optimal overlap of $\hat{d}$ in place of m^o these two optimal overlaps were occasionally quite different from one another (difference of 10 or more units). In addition ρ_{xy} was included for each value of a and r.

It was found that these relative gains are almost unaffected by misspecification of the population parameters used to estimate the optimal weights λ^o and α^o. Since, in practice, these parameters must be estimated this robustness property is important. In addition, small deviations of m from m^o were also found to have relatively little affect on these relative gains. Thus, the most important sourse of variation in these relative gains comes from the variation of the pair (a,r) and this is what is tabulated in Table one. In this section n=50.

It is immediately evident from examining Table one that $\hat{e}$ is superior to $\hat{d}$ under the criterion of relative gain. The definition of relative gain implies that the smaller the relative gain the better the estimator. The relative gain of $\hat{e}$ is generally more than 10% less than the relative gain of $\hat{d}$ except in the extreme cases where either no composite estimator should be considered because ρ_{xy} is very small or either composite estimator will work well because ρ_{xy} is very close to one.

Table 1. Theoretical Relative Gains

		r					
		.01	.11	.21	.31	.41	.46
	a						
Gain $\hat{e}$	.001	.65	.55	.53	.53	.52	.52
Gain $\hat{d}$		.71	.57	.55	.55	.54	.54
ρ_{xy}		.95	.99	1.00	1.00	1.00	1.00
Gain $\hat{e}$	.10	.99	.85	.78	.74	.71	.70
Gain $\hat{d}$		1.00	.94	.89	.86	.84	.83
ρ_{xy}		.30	.70	.80	.84	.86	.87
Gain $\hat{e}$	.20	.99	.90	.84	.80	.76	.75
Gain $\hat{d}$		1.00	.97	.95	.92	.91	.90
ρ_{xy}		.22	.57	.68	.74	.77	.78
Gain $\hat{e}$	.30	1.00	.93	.87	.83	.79	.78
Gain $\hat{d}$		1.00	.99	.97	.95	.94	.93
ρ_{xy}		.18	.50	.60	.66	.70	.72
Gain $\hat{e}$	.40	1.00	.94	.89	.85	.81	.80
Gain $\hat{d}$		1.00	.99	.98	.97	.96	.95
ρ_{xy}		.16	.45	.55	.61	.65	.66
Gain $\hat{e}$	.45	1.00	.94	.89	.85	.82	.81
Gain $\hat{d}$		1.00	.99	.98	.97	.96	.96
ρ_{xy}		.15	.42	.53	.59	.63	.64

If useful relative gain is defined as any relative gain less than .9 then that region of the (a,r)-plane ($0<a<.45$, $0<r<.46$) where $\hat{e}$ achieves a useful relative gain is more than twice the area of the region of this plane where $\hat{d}$ achieves a useful relative gain. Therefore there are many situations where the model based composite estimator, $\hat{e}$, will still be useful when the design based composite estimator, $\hat{d}$, would not be considered because of its relatively puny improvement over $\bar{y}$. Note also that $\hat{e}$ attains a useful relative gain for values of ρ_{xy} of (0.7) or greater and in some cases for values of ρ_{xy} as low as .53 (see Table 1 a=.45 and r=.21). $\hat{d}$ generally requires a value of ρ_{xy} greater than .8 to achieve a useful relative gain.

Table 2 is concerned with the behavior of $\hat{e}$ and $\hat{d}$ on real data where our hypothetical superpopulation model is but a crude explanation for the process which generated the data. This data set consists of employment for the months of March 1983 and January 1984 for 2000 hospitals with March 1983 employment between 50 and 250. From this universe of 2000 hospitals the sampling was done as described in section two with n=50 and m=20. The x-variable was the March 1983 employment and the y-variable was the January 1984 employment. The sample data was used to estimate the various parameters needed for $\hat{e}$ and $\hat{d}$. These are λ^o, α^o, and certain other parameters needed in their estimators.

For λ^o, we use $(a_1 - a_3)/(a_1 + a_2 - 2a_3)$

where $a_1=(1/m)s_y^2 + (\{1/m\} - \{1/n\})\mathring{\beta}^2 s_x^2 + 2(\{1/n\} - \{1/m\})\mathring{\beta}s_{xy}$

$$a_2=(1/n)s_y^2$$

$$a_3=(1/n)s_y^2 + \mathring{\beta}s_{xy}(\{m/n^2\} - \{1/n\})$$

$$s_y^2 = \sum_{s_y \cap s_x} (y_i - \bar{y})^2$$

$$s_x^2 = \sum_{s_x \cap s_y} (x_i - \bar{x})^2$$

$$s_{xy} = \sum_{s_x \cap s_y} (x_i - \bar{x})(y_i - \bar{y})$$

The estimator for the optimal α to be used in $\hat{e}$ is:

$(e_1 - e_3)/(e_1 + e_2 - 2e_3)$ where:

$$e_1 = \left[\hat{\sigma}^2/m + \hat{\beta}^2\right]\left[1/(n-m)\right]s_x^2 + \left[\hat{\sigma}^2/m\right]\left[\bar{x}^2 - s_x^2/n\right] .$$

$$e_2 = (1/n)\left[s_x^2(\hat{\beta}^2 + \hat{\sigma}^2) + \hat{\sigma}^2(\bar{x}^2 - s_x^2/n)\right]$$

$$e_3 = (1/n)\hat{\sigma}^2\left[\bar{x}^2 - s_x^2/n\right]$$

$$\hat{\sigma}^2 = (1/\{n-m\}) \sum_{s_x \cap s_y} (y_i - \hat{\beta}x_i)^2/x_i^2$$

$$\hat{\beta}^2 = \hat{\beta}^2 - \hat{\sigma}^2/(n-m)$$

With these sample data based estimates for the superpopulation parameters and the population

Table 2. Simulated Relative Gains

Population Correlation	$\hat{\delta}$	RG($\hat{e}$)	RG($\hat{d}$)
.96	1.53	.65	.74
.95	1.80	.73	.88
.90	1.84	.69	.83
.83	1.85	.93	.94
.76	1.82	.79	.89
.71	1.83	.88	.97
.83	2.22	.71	.87
.83	2.17	.85	.97
.83	2.03	.79	.92
.82	2.01	.93	.95
.82	1.96	.82	.91
.81	2.16	.89	.97
.94	1.86	.73	.89
.87	1.82	.71	.85
.78	1.84	.95	.97
.69	1.82	.83	.93
.61	1.84	.90	.98

parameters the estimates, $\hat{e}$ and $\hat{d}$ were constructed. Then their squared errors, $(\hat{e} - \bar{Y})^2$ and $(\hat{d} - \bar{Y})^2$, were computed. This was replicated 150 times and these squared errors were averaged over these 150 replications to obtain the empirical relative gains for $\hat{e}$ and $\hat{d}$. These empirical relative gains are contained in Table 1 and denoted as $RG(\hat{e})$ and $RG(\hat{d})$ where:

$$RG(\hat{e}) = \left[\sum_{\mathcal{Q}} (\hat{e} - \bar{Y})^2 \right] \left[\sum_{\mathcal{Q}} (\bar{y} - \bar{Y})^2 \right]^{-1}$$

$\mathcal{Q}$ is the set of 150 replications. $RG(\hat{d})$ is similarly defined. For this hospital data set the correlation, ρ_{xy}, was .96. In order to compare these two composite estimators on a greater variety of populations where this correlation was much lower than .96, noise was injected into the y-variable (January 1984 data) and the simulation was rerun. These derived populations are classified by their values for ρ_{xy}.

The entries in Table 2 substantially support the conclusions from Table 1. The prediction theory based composite estimator remains a great improvement over the more classical approach even when the hypothesized superpopulation model is a very rough description of the data. In particular, the superpopulation model hypothesized in section two assumes that the unit variance of the i^{th} unit is proportional to x_i^2. As a test of this variance assumption the universe data was used to estimate δ where δ was the exponent of x_i under the model that assumes the variance of the i^{th} unit is proportional to x_i^{δ}. The estimates of δ for each of the simulated populations are given in the column labeled $\hat{\delta}$. Note that $\hat{d}$ is similar to the BLUE under a superpopulation model with $\delta=1$.

As is seen by examining table two these estimated values of δ all lie between about 1.5 and 2.2 but even when $\hat{\delta}$ is 1.53 the prediction theory composite estimator, $\hat{e}$, still is strikingly superior to $\hat{d}$.

4. CONCLUSIONS

The purpose of this paper is to suggest ways of using superpopulation prediction theory to derive superior estimators for classical sampling problems where the sampling distribution still plays an essential role. The composite estimation problem which this paper analyzes is only one example of this technique. In nearly all sample survey situations where some kind of supplementary information is available about the population being studied a superpopulation model may be hypothesized and used to derive an improved estimator (i.e. the BLUE). The usual techniques of superpopulation prediction theory may not be directly applicable (i.e. conditionality) but slight generalizations of these techniques can produce improved estimators. The composite estimation problem which this paper studies shows that this method of combining both these approaches to sampling problems can be very benificial.

This approach to an improved sampling strategy can be summarized in three steps. In step one we use supplementary information to hypothesize a superpopulation model. In step two we derive the BLUE under this model and in step three we use the sampling distribution (if necessary) to estimate the unknown superpopulation parameters in the BLUE. Thus, variability in this estimator will, in general, come from both the sampling distribution and the superpopulation distribution. For example, in the composite estimator, $\hat{e}$, the x-mean for all the non- sample units in the population was estimated by its conditionally (given s_y) sampling unbiased estimator, the x-mean of the units in s_x but not in s_y.

Both the theoretical and the simulation results of section three demonstrate an improved approach to composite estimation. These results can be roughly summarized by saying that the prediction theory approach to composite estimation is as superior to the more classical approach as the classical approach is superior to the basic estimator $\bar{y}$. In addition,the prediction theory composite estimator may give substantial improvements in precision in many situations where the classical approach to composite estimation would give such a very marginal improvement that it would not be considered worthwhile to use.

The problems of generalizing this approach to composite estimation to a larger set of superpopulation models, of estimating variances for these hybrid estimators, and of extending this approach to other inference problems beyond composite estimation may be other directions in which to continue this work.

REFERENCES

Cassel,C. , Särndal, C. and Wretman, J.H. (1977), Foundations of Inference in Survey Sampling, John Wiley & Sons.

Cochran, W.G. (1966), Sampling Techniques, (Second Edition), John Wiley & Sons, Inc.

Hansen, M.H. , Hurwitz, W.N. , Madow, W.G. (1953) Sample Survey Methods and Theory, John Wiley & Sons, Inc.

Kish, Leslie (1965), Survey Sampling, John Wiley & Sons, Inc.

Raj, D. (1968), Sampling Theory, McGraw-Hill.

Royall, R.M. (1971) "Linear Regression Models in Finite Population Sampling Theory," in Godambe, V.P. and Sprott, D.A. (eds), Foundations of Statistical Inference, Toronto: Holt, Rinehart and Winston of Canada, Ltd.

Särndal, Carl-Erik (1978), "Design-based and Model-based Inference in Survey Sampling," Scandinavian Journal of Statistics, Vol 5, 1.

Wolter, K.M. (1979), "Composite Estimation in Finite Populations," Journal of the American Statistical Association, 74, 604-613.