SAMPLING FOR CHANGE

James M. Lowerre, Federal Trade Commission

Measurement of change in some characteristic of each element of a population over time is often a significant statistical problem. For example, some users of the Federal Trade Commission's publication, the Quarterly Financial Report, will bench mark on a large sample survey which is out of date. They adjust by adding quarter to quarter change estimates from the QFR. Standard references on sampling; Cochran (1977), Hansen, Hurwitz, and Madow (1953), and Raj (1968), treat the statistical aspects of change estimation when the population sizes and variances do not change from period to period, when the unmatched portion of the sample is the same from period to period, and the correlation of units over time is known.

The conditions under which the mathematical results are derived are not met in some surveys In the Quarterly Financial Report population counts must be estimated, and such estimates incorporate the same data used in estimating changes in the averages. The other statistical assumptions are also violated. The essential reason is that companies are bought, sold, merged, react differently to general economic conditions, have different technological problems and successes, etc. The purpose of this paper is to expand the general theory to handle these differences.

The basic problem can be handled through use of Aiken's regression estimate. It is assumes that a population has mean μ_1 in period one and mean μ_2 in periof two. Further, a sample of A units is drawn in period one and not repeated in period two, while B in period one are repeated in period two. C are sampled in period two, but not in period one. The data in period one may be represented by the vector (y_1', y_2'), with y_1' the data vector for the sample units not repeated in period two. Similarly, the data vector for period two may be represented by (z_1', z_2'), where z_1' is the data vector for units repeated in periods one and two. y_1' is 1xA, y_2' and z_1' are each 1xB, while z_2' is 1xC. The results can be summarized as

$$w = \begin{pmatrix} y_1 \\ y_2 \\ z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix} = Xb + e \qquad (1)$$

where the "1's" from topto bottom are Ax1, Bx1, Bx1, and Cx1 vectors, each of whose components are the integer one. A comparable size holds for the "0" vectors, and the μ's are the population means. The covariance of the error term is

$$\begin{pmatrix} \sigma_1^2 I_A & 0 & 0 & 0 \\ 0 & \sigma_1^2 I_B & \gamma I_B & 0 \\ 0 & \gamma I_B & \sigma_2^2 I_B & 0 \\ 0 & 0 & 0 & \sigma_2^2 I_C \end{pmatrix} = \Sigma \ . \qquad (2)$$

In practice, A,B, and C may be stochastic. This makes the row dimensions in X in (1) random variables.

GENERAL CHARACTERISTICS

If the population sizes are N_1 and N_2, the item to be estimated is $N_2\mu_2 - N_1\mu_1$. Using the popu lation estimates, designated by N_1 and N_2, the estimates here will be $\hat{N}'\hat{\mu} = (-\hat{N}_1, \hat{N}_2)\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix}$, and will be unbiased if $\hat{N}$ is unbiased and $\hat{\mu}$ is conditionally unbiased. The estimates of μ will be taken of the form Qw, where Q is a 2xk matrix. I f $QX = I$, and $\hat{N}$ and Q are indepenent then $E[\hat{N}Qw] = E[E[\hat{N}Qw|N,Q]] = N'\mu$. Also, $\mathrm{var}[\hat{N}'Qw] = E[\mathrm{var}[\hat{N}'Qw|N,Q]] + \mathrm{var}[E[\hat{N}'Qw|\hat{N},Q]] = E[\hat{N}'Q\Sigma Q'N] + \mu'\Sigma_N\mu$. This variance expression holds whether or not the variables $\hat{N}$ and Q are independent.

Now, a general formula can be derived for $E[\hat{N}'Q\Sigma Q'N]$, when $\hat{N}$ and Q are independent. It is

$$E[\hat{N}'Q\Sigma Q'\hat{N}] = \sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_{ij} \mathrm{tr}[\Sigma_N \mathrm{cov}(q_i,q_j) + N'\mathrm{cov}(q_i,q_j)N + E(q_i')(\Sigma_N + NN')E(q_j)] \qquad 3)$$

Equation (3) seems awkward to work with in practice, and, worse, $\hat{N}$ and Q may not be independent. This paper will develop properties of two estimates of change where N and Q are not independent, population parameters may vary from period to period, and the amount of data may vary stochastically from that planned.

TWO ESTIMATES OF FORM Qw

In (1) w is of the form $w = Xb + e$, ie the usual regression form. As a result the Q which minimizes $\mathrm{var}[\hat{N}'Qw|\hat{N},Q]$ is

$$Q = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} \ , \qquad (4)$$

and

$$\hat{N}'Qw = \hat{N}'(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}w \ . \qquad (5)$$

Using X in (1) and in (2), Qw becomes

$$Qw = h \cdot v \qquad (6)$$

where the scalar $h = [AC(1-\rho^2)+(A+C)B + B^2]^{-1}$, and the vector $\underset{\sim}{v}$ is

$$\begin{pmatrix} [B+C(1-\rho^2)]A & B(B+C) & \frac{-BC\sigma_1\rho}{\sigma_2} & \frac{BC\sigma_1\rho}{\sigma_2} \\ \frac{AB\sigma_2\rho}{\sigma_1} & \frac{-AB\sigma_2\rho}{\sigma_1} & B(A+B) & C[B+A(1-\rho^2)] \end{pmatrix}$$

times the vector $(\bar{y}_1, \bar{y}_2, \bar{z}_1, \bar{z}_2)'$ (6)

Another choice is

$$Qw = (X'X)^{-1}X'w = \begin{pmatrix} A/(A+B) & B/(A+B) & 0 & 0 \\ 0 & 0 & B/(B+C) & C/(B+C) \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} \quad (7)$$

In the special case that A=C, $\sigma_1 = \sigma_2$, $\mu = A/(A+B)$, $\lambda = B/(B+C)$, it is straight forward to show that (6) becomes

$$(1-\mu^2\rho^2)^{-1}\begin{pmatrix} \mu(1-\mu\rho^2)\bar{y}_1+\lambda\bar{y}_2+\mu\lambda\rho(\bar{z}_2-\bar{z}_1) \\ (\bar{y}_1-\bar{y}_2)\mu\lambda\rho+\lambda z_1+\mu(1-\mu\rho^2)\bar{z}_2 \end{pmatrix} \quad (8)$$

which are the standard results, Raj (1968).

THE EFFECT OF $\hat{\underset{\sim}{N}}$

In some series the population in the second period is not the same as in the first. This is because there are births and deaths. In an economic survey with fims classified by industry this will occur because companies are: bought, sold, formed, spin off subsidiaries, etc. The population in existence during period one can only decrease, although the total population count at time two may be larger due to births. The notation $\mu_{s(1)}$ and μ_b will be used for the mean of the survivors and the births. $N_{s(1)}$, N_b will be u sed for the corresponding population counts. Estimates of these will wear the usual "hats".

If at time one a sample of size R is planned with μR sample units not repeating in period two, and λR in the sample both times, with $\mu+\lambda = 1$, A,B, and C can be determined by what happens to these quantities.

If k_1 units were expected to be in the sample in period two, having been in the sample in period one, and died then

$$A = \mu R + k_1$$
$$B = \lambda R - k_1 \quad (9)$$

If k_2 units were expected at time one to yield valid data at time two but susequently died, then

$$C = \mu R - k_2 \quad . \quad (10)$$

If the population size at time one is estimated with $\hat{N}_1$, then the number of survivors at time two will be estimated with

$$\hat{N}_{s(1)} = \hat{N}_1[1-(k_1+k_2)/R] \quad (11)$$

The estimate in (11) is the same percentage of $\hat{N}_1$ as the perecentage of the sample actually in period two relative to what was planned in period two. Further, if k_1 and k_2 are a random sample of deaths then (11) is an unbiased estimate for the number of survivors if $\hat{N}_1$ is unbiased. Subsequently k_1 will be assumed to have a binomial distribution with parameters (λR, p), k_2 to be binomial with parameters (μR, p) independnet of k_1, and $N_{s(1)} \doteq N_1(1-p)$ Further, it can easily be shown that (11) can be written

$$\hat{N}_{s(1)} = \hat{N}_1(B+C)/R \quad . \quad (12)$$

If there is a population of births with $\hat{N}_b$ an unbiased estimate of the number and $\bar{x}_b$ the , independent of $\hat{N}_b$, average measured response on a sample of b births, then the total population estimate at time two can be taken to be $\hat{N}_2 = \hat{N}_{s(1)} + \hat{N}_b$. The estimate of $N_2\mu_2 - N_1\mu_1$ will be taken to be

$$-\hat{N}_1\hat{\mu}_1 + \hat{N}_{s(1)}\hat{\mu}_{s(1)} + \hat{N}_b\bar{x}_b = \hat{\underset{\sim}{N}}'Qw + \hat{N}_b\bar{x}_b \quad , \quad (13)$$

$$\text{with } \hat{\underset{\sim}{N}}' = \hat{N}_1(-1, \frac{C+B}{R}) \quad , \quad (14)$$

and Qw given by (6) or (7) .

Using Qw from (7) one finds $\hat{\underset{\sim}{N}}$ not independent of Q. However,

$$\hat{\underset{\sim}{N}}'Qw = \hat{N}_1(\frac{-A}{A+B} \quad \frac{-B}{A+B} \quad \frac{B}{R} \quad \frac{C}{R}) \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{z}_1 \\ \bar{z}_2 \end{pmatrix} ,$$

$$\begin{aligned} \text{so } E[\hat{\underset{\sim}{N}}'Qw] &= E[E[\hat{\underset{\sim}{N}}'Qw|A,B,C,\hat{N}_1]] \\ &= E[\hat{N}_1(-\mu_1 + (B+C)\mu_{s(1)}/R] \\ &= E[\hat{N}_1(-\mu_1 + (1-\frac{k_1+k_2}{R})\mu_{s(1)})] \\ &= -N_1\mu_1 + N_{s(1)}\mu_{s(1)} \quad . \end{aligned}$$

It follows that (13) is unbiased for $-N_1\mu_1 + N_2\mu_2$. A similar result holds on using Qw in (6), but showing its unbiased characteristic is somewhat different. Using Qw from (6)

$$\underset{\sim}{\hat{N}}'Q\underset{\sim}{w} = [N_1/(AC(1-\rho^2)+(A+C)B+B^2)]\cdot(-1\ ,\ \frac{B+C}{R})\ \cdot$$

$$\begin{pmatrix} [B+C(1-\rho^2)]A & B(B+C) & \frac{-BC\rho\sigma_1}{2} & \frac{BC\rho\sigma_1}{2} \\ \frac{AB\rho\sigma_2}{\sigma_1} & \frac{-AB\rho\sigma_2}{\sigma_1} & B(A+B) & [B+A(1-\rho^2)]C \end{pmatrix}\cdot$$

$$(\bar{y}_1,\ \bar{y}_2,\ \bar{z}_1,\ \bar{z}_2)'\ .$$

Again, the elements in $\underset{\sim}{\hat{N}}$ and Q are not independent. However,

$$E[\underset{\sim}{\hat{N}}'Q\underset{\sim}{w}] = E[E[\underset{\sim}{\hat{N}}'Qw|A,B,C,\hat{N}_1]]$$

$$= E[\hat{N}_1(-1\ ,\ \frac{(B+C)}{R})\begin{pmatrix}\mu_1 \\ \mu_{s(1)}\end{pmatrix}]$$

$$= -N_1\mu_1 + N_{s(1)}\mu_{s(1)}\quad .$$

Since $E[\underset{\sim}{\hat{N}}'Q\underset{\sim}{w}|A,B,C,\hat{N}_1]$ is the same for both Q's, it follows that as in the general characteristics section, that $\text{var}[\underset{\sim}{\hat{N}}'Q\underset{\sim}{w}] = E[\text{var}[\underset{\sim}{\hat{N}}'Qw|A,B,C,\hat{N}_1]] + \text{var}[E[\hat{N}'Qw|A,B,C,\hat{N}_1]]$ is smaller using $Q\underset{\sim}{w}$ from (6) rather than (7). However, the conclusion is different from that in the usual regression theory since even using $Q\underset{\sim}{w}$ from (6) it cannot be claimed that a BLUE has been obtained. Rather, it can only be concluded that such a choice yields an estimate with a smaller variance than is obtained by simply subtracting level estimates.

ALTERNATIVE ON BIRTHS

Although $\underset{\sim}{\hat{N}}'Qw$ in (13) and (14), and the subsequent development does estimate change from that extant at time one, it seems plausible that Aiken's estimate including births might do better still. However, if there are b births with the corresponding data vector $\underset{\sim}{x}_b$, then all the data can be assembled into the single vector

$$\underset{\sim}{w} = \begin{pmatrix} \underset{\sim}{y}_1 \\ \underset{\sim}{y}_2 \\ \underset{\sim}{z}_1 \\ \underset{\sim}{z}_2 \\ \underset{\sim}{x}_b \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \mu_1 \\ \mu_{s(1)} \\ \mu_b \end{pmatrix} + e\ = Xb + e \qquad (15)$$

with covariance matrix

$$\begin{pmatrix} \sigma_1^2 I_A & 0 & 0 & 0 & 0 \\ 0 & \sigma_1^2 I_B & \gamma I_B & 0 & 0 \\ 0 & \gamma I_B & \sigma_2^2 I_B & 0 & 0 \\ 0 & 0 & 0 & \sigma_2^2 I_C & 0 \\ 0 & 0 & 0 & 0 & \sigma_b^2 I_b \end{pmatrix} \qquad (16)$$

The population at time two, which is the survivors plus births, will have mean $\mu_2 = (N\mu_{s(1)} + N_b\mu_b)/(N_{s(1)} + N_b)$. It is straight forward, using (15) and (16), to show that

$$(-\hat{N}_1,\ \hat{N}_{s(1)},\ \hat{N}_b)\begin{pmatrix}\hat{\mu}_1 \\ \hat{\mu}_{s(1)} \\ \hat{\mu}_b\end{pmatrix} =$$

$$(-\hat{N}_1,\ \hat{N}_{s(1)},\ \hat{N}_b)\ (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\underset{\sim}{w}\ \text{is}$$

given by (13).

VARIANCES

In this section expressions for the variance of N'Qw are developed, using $\text{var}[\hat{N}'Qw] = E[\hat{N}'Q\Sigma Q'N] + \underset{\sim}{\mu}'\Sigma_N\underset{\sim}{\mu}$. The (1,1) element of Σ_N is $E[(\hat{N}_1-N_1)^2] = \sigma^2_{\hat{N}_1}$. The (1,2) and (2,1) elements of Σ_N are

$$E[(\hat{N}_1-N_1)(\hat{N}_1(\frac{B+C}{R}) - (1-p)N_1)]$$

$$= E[(\hat{N}_1-N_1)(\hat{N}_1(1-p) - (1-p)N_1)]$$

$$= \sigma^2_{\hat{N}_1}(1-p)$$

Similarly, the (2,2) element is

$$E[(\hat{N}_1(\frac{B+C)}{R} - (1-p)N_1)^2]$$

$$= R^{-1}[(\sigma^2_{\hat{N}_1} + N_1^2)p(1-p) + R(1-p)^2\sigma^2_{\hat{N}_1}]\ .$$

These can be substituted back to yield the exact expression for $\underset{\sim}{\mu}'\Sigma_N\underset{\sim}{\mu}$. A direct use of Q in (7) yields

$$E[N'Q\Sigma Q'N] = (\sigma^2_{\hat{N}_1} + N_1^2)[\sigma_1^2 - 2\lambda\rho\sigma_1\sigma_2(1-p)]$$

$$+ \sigma_2^2(1-p)]$$

. It follows that

$\text{var}\ [\underset{\sim}{\hat{N}}'Q\underset{\sim}{w}]$ is

$$R^{-1}(\sigma^2_{\hat{N}_1} + N_1^2)[\sigma_1^2 - 2\lambda\rho\sigma_1\sigma_2(1-p) + \sigma_2^2(1-p)]$$

$$+ \mu_1^2\sigma^2_{\hat{N}_1} + 2\mu_1\mu_2\sigma^2_{\hat{N}_1}(1-p)$$

$$+ R^{-1}\mu_2^2[(\sigma^2_{\hat{N}_1} + N_1^2)p(1-p) + R(1-p)^2\sigma^2_{\hat{N}_1}] \; . \qquad (18)$$

Q can also be substituted from (6), but it seems difficult to find a closed form expression as simple as (18). Of course, (18) is an upper bound. It is clear from the way that occurs in (18) that it should be as large as possible, as in the usual case, even if A,B, and C are stochastic.

CONCLUSIONS

Provided the estimate of the number of survivors is treated properly and the effect of births added in, Aiken's estimate from regression has been shown to yield an unbiased estimate of change with smaller variance than simply subtracting level estimates. The proof does not show that Aiken's estimate yields a BLUE, however, as would be the case if the population counts were known. An explicit formula for the variance of Aiken's estimate was not obtained. The variance for the difference in the levels was obtained and shows how the usual estimate of change has its variance inflated when population counts must be estimated.

REFERENCES

Cochran, W., (1977) "Sampling Techniques", 3rd Ed., John Wiley & Sons

Hansen, M., Hurwitz, W., and Madow, W., (1953) "Sampling Survey Methods & Theory", John Wiley & Sons

Raj, D., (1968) "Sampling Theory", McGraw Hill Book Co.

TABLE I: MEASURES OF RELIABILITY FOR THE ITERATIVE REGRESSION PROCEDURE

Class of Statistic	Number of Specified Domains (N)	Average Absolute Difference				Relative Average Absolute Difference			
		10% Samples		20% Samples		10% Samples		20% Samples	
		$\bar{A}_1$ (S.E.)	$\bar{A}_1$ (S.E.)	$\bar{A}_1$ (S.E.)	$\bar{A}_1$ (S.E.)	$\bar{A}_2$ (S.E.)	$\bar{A}_2$ (S.E.)	$\bar{A}_2$ (S.E.)	$\bar{A}_2$ (S.E.)
Narrow Range	604	.0336 (.0068)	.0254 (.0046)	.0601 (.0081)	.0462 (.0059)	1.9064 (.2560)	2.2854 (.2790	3.2310 (.3328)	3.3934 (.2990)
Med. Range	307	.0352 (.0086)	.0120 (.0031)	.0243 (.0049)	.0365 (.0070)	.7809 (.0813)	.4291 (.0435)	.6025 (.0531)	.8229 (.0754)
Wide Range	327	.0140 (.0029)	.0275 (.0065)	.0126 (.0022)	.0196 (.0033)	.5961 (.1314)	.8817 (.1802)	.5117 (.0890)	.8307 (.1210)

TABLE II: ESTIMATED REGRESSION COEFFICIENTS USING THE ITERATIVE PROCEDURE

Class of Statistic	Number of Specified Domains (N)	10% Samples				20% Samples			
		$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
Narrow Range	604	.000134	54.915	.002239	37.515	.-000291	110.049	.001520	82.955
Med. Range	307	.001517	23.474	.001771	13.6316	.001661	20.296	.001162	25.3201
Wide Range	327	.002434	1200.219	.001663	1979.673	.002447	1203.177	.002485	1140.034

TABLE III: MEASURES OF RELIABILITY FOR THE WEIGHTED LEAST SQUARES REGRESSION PROCEDURE

Class of Statistic	Number of Specified Domains (N)	Average Absolute Difference				Relative Average Absolute Difference			
		10% Samples		20% Samples		10% Samples		20% Samples	
		$\bar{A}_1$ (S.E.)	$\bar{A}_1$ (S.E.)	$\bar{A}_1$ (S.E.)	$\bar{A}_1$ (S.E.)	$\bar{A}_2$ (S.E.)	$\bar{A}_2$ (S.E.)	$\bar{A}_2$ (S.E.)	$\bar{A}_2$ (S.E.)
Narrow Range	604	.0239 (.0054)	.0225 (.0045)	.0273 (.0036)	.0268 (.0038)	1.5225 (.1489)	1.9194 (.2052)	2.5301 (.2625)	2.6902 (.2747)
Med. Range	307	.0187 (.0047	.0093 (.0024)	.0181 (.0036	.0238 (.0053)	.6080 (.0692)	.3952 (.0415)	.5136 (.0494)	.6308 (.0579)
Wide Range	327	.0134 (.0027)	.0225 (.0058)	.0121 (.0021)	.0184 (.0032).	.5430 (.1082)	.7146 (.1157)	.4803 (.0802)	.7515 (.0996)

TABLE IV: ESTIMATED REGRESSION COEFFICIENTS FOR THE WEIGHTED LEAST SQUARES PROCEDURE

Class of Statistic	Number of Specified Domains (N)	10% Samples				20% Samples			
		$\hat{\alpha}$ (S.E.)	$\hat{\beta}$ (S.E.)	$\hat{\alpha}$ (S.E.)	$\hat{\beta}$ (S.E.)	$\hat{\alpha}$ (S.E.)	$\hat{\beta}$ (S.E.)	$\hat{\alpha}$ (S.E.)	$\hat{\beta}$ (S.E.)
Narrow Range	604	.000916 (.0011)	33.248 (6.076)	.001732 (.000887)	31.312 (5.044)	.0030706 (.001455)	43.007 (7.639)	.003166 (.001575)	42.220 (8.383)
Med. Range	307	.002510 (.0012)	13.815 (2.137)	.001893 (.000454)	11.163 (.832)	.002369 (.000940)	15.079 (1.540)	.002489 (.001700)	16.666 (2.868)
Wide Range	327	.002920 (.0010)	1030.545 (126.594)	.003365 (.002540)	1428.750 (308.704)	.002819 (.000940)	1101.744 (102.682)	.003272 (.00158)	1228.045 (169.015)

TABLE V: PERCENT IMPROVEMENT IN RELIABILITY OF WLS OVER THE ITERATIVE REGRESSION PROCEDURE

Class of Statistic	Number of Specified Domains (N)	Improvement Relative to $\bar{A}_1$				Improvement Relative to $\bar{A}_2$			
		10% Samples		20% Samples		10 % Samples		20% Samples	
		I_1	I_1	I_1	I_1	I_2	I_2	I_2	I_2
Narrow Range	604	28.9	11.4	54.6	42.0	20.1	16.0	21.7	20.7
Med. Range	307	46.9	22.5	25.5	34.8	22.1	7.9	14.8	23.3
Wide Range	327	4.3	18.2	4.0	6.1	8.9	19.0	6.1	9.5