ECI VARIANCE ESTIMATION

Harry Marks and David Frevert, Bureau of Labor Statistics

The purpose of this paper is to describe the procedure being implemented for calculating estimates of the variance of the published ECI wage indexes. Also discussed are the rational for certain procedural decisions. A measurement of the components of variance are not covered here, but will be presented in subsequent papers.

The ECI is a survey employing a sampling design which involves several stages, including independent selections of occupations and establishments, and a final controlled selection process.[1/] To insure that the variance estimator captures all the effects of the design, it would have been necessary to form replicates of establishments crossed with occupations and within each of these replicates perform the controlled selection. This was not done. Therefore, replicates are being constructed that reflect primarily the first two independent stages of selection (i.e. occupation selection and establishment selection) without explicitly incorporating the controlled selection process. To the degree that the replicates do not reflect the distribution of quotes as they would if a controlled selection were done, the estimated variance will be biased (most likely an over-estimate). In many cases, the number of patterns selected was increased so that almost all possible quotes that could have been designated in the controlled selection were selected. Thus, the contribution to the total variance in these cases due to this stage is minimal.

The estimation is accomplished by producing replicates within each of the eight Subindustry Divisions defined as follows:

1. Mining
2. Construction
3. Manufacturing - Durable goods
4. Manufacturing - Nondurable goods
5. Transportation
6. Wholesale - Retail Trade
7. F.I.R.E.
8. Services

This level seems appropriate for the variance estimation because cell collapsing for the purpose of imputation was rarely done at levels more aggregate than Subindustry Division (SID) by Major Occupation Group (MOG). (The MOG's are the elements of a partition of the set of occupations and are the strata from which occupations were selected.)

The estimation process is accomplished by selecting two sets of half-samples, by establishment and by occupation, respectively. By crossing these two sets of half-samples, 4 "replicates" are created (i.e., a replicate consists of quotes designated by an establishment and occupation such that the establishment is in one of the establishment half-samples and the occupation is in one of the occupational half-samples.)

Diagram 1

	First Half-Sample of Occupations +	Second Half-Sample of Occupations -
First Half-Sample of Establishments +	++	+-
Second Half-Sample of Establishments -	-+	--

In view of the fact that expected numbers of establishments and occupations were fixed in the controlled selection, and thus for the final sample, the replicates were balanced with respect to the number of designated quotes. The establishments within each SID are classified and the classes are called Standard Industrial Classes (SIC). For estimation of variance at the national level it is desirable to balance these replicates within each two-digit SIC since this is the level at which the controlled selection was utilized. It may be desirable to balance combinations of SIC's in the case when SIC's can't be adequately balanced separately.

After "replicates" are generated within each SID, the usual imputation procedure is applied for missing values. Imputing is restricted to groupings no larger than SIC by MOG. To compute estimates of variance for the subindexes, the entire estimation procedure is applied separately to each subset of establishment/occupational quotes that are used for the subindex.

The following discussion presents the computational estimation procedure in detail. It should be noted that we are outlining more than one option for several phases of the procedure. The final method will result from decisions made during implementation. We first make the following definitions:

$$(1)\quad R_{t,t-1} = \frac{\bar{X}_t}{\bar{X}_{t-1}} = \frac{\sum_s X'_{ts}}{\sum_s X'_{t-1,s}} = \text{wage relative}$$

$$\text{where}\quad X'_{ts} = \sum_{k=1} C_{ks} \frac{1}{N_{ks}} \sum_{j=1}^{N_{ks}} \frac{\sum_i W_{ijs} X_{tijs}}{\sum_i W_{ijs}}$$

C_{ks} = census employment for k^{th} occupation stratum in the s^{th} SIC

N_{ks} = number of occupations in the k^{th} occupation stratum in the s^{th} SIC

W_{ijs} = est/occ weight for the j^{th} occupation in the i^{th} establishment in the s^{th} SIC

X_{tijs} = ij^{th} est/occ average wage level at time t in s^{th} SIC

Further, let c be the index to designate the SID's.

Now define, $X''_{tc} = \sum_{s\varepsilon c} X'_{ts}$

for the c^{th} SID.

And,

$$X'''_t = \sum_c X''_{tc} = \sum_s X'_{ts} .$$

Thus, we have

$$R_{t,t-1} = \frac{X'''_t}{X'''_{t-1}} = \frac{\sum_c X'_{tc}}{\sum_c X'_{t-1,c}}$$

Through use of the linear terms of the Taylor Series expansion of $R_{t,t-1}$ the following linear approximation is obtained:

$$(2)\quad \mathrm{VAR}\,(R_{t,t-1}) \doteq \frac{1}{(X'''_{t-1})^2}\left[\sum_c \mathrm{VAR}\,(X''_{tc} - R_{t,t-1}\,X''_{t-1,c})\right]$$

where $R_{t,t-1} = \dfrac{\sum_c X''_{tc}}{\sum_c X''_{t-1,c}}$

Now let

$$d_{ct,t-1} = (X''_{tc} - R_{t,t-1}\,X''_{t-1,c}),$$

Then (2) can be rewritten as:

$$(3)\quad \mathrm{VAR}\,(R_{t,t-1}) = \frac{1}{(X''_{t-1})^2}\left[\sum_c \mathrm{VAR}\,(d_{ct,t-1})\right] .$$

For estimating VAR $(d_{ct,t-1})$ three alternatives are available. All three methods involve using the four replicates constructed within each SID.

The assignment of quotes to these replicates is accomplished through an iterative computer routine. The program is built to first assign +'s and -'s alternatively to adjacent establishments after the establishments have been ordered as in Phase II (ie. by 2-digit SIC and within SIC by employment size). The program then ascribes +'s and -'s alternatively to each occupation of the pair within a MOG within a 2-digit SIC in such a way as to minimize the imbalance. Thus, the four cells are filled according to combinations of +'s and -'s. (See Diagram 1.) A test is then made to insure that the number of quotes within each cell is relatively constant across cells. In the case where the number is not relatively constant, the above assignment procedure is repeated with alternate patterns of +'s and -'s for establishments until the criteria for cell size is met. Certainty occupations receive both + and -, and consequently, are assigned to both occupational half-samples.

It should be noted here that if the cells cannot be balanced at the SID level, it may be unwise to estimate the variance at the SID level. In this case, the linear approximation is inappropriate. An alternate approach to estimating VAR $(R_{t,t-1})$ would be to form psuedo-replicates by collapsing across all SID's with different combinations of +'s and -'s. In either case the following three methods are available.

Method 1 - Assuming the replicates are balanced, $d_{ct,t-1}$ is computed separately for the set of quotes in each cell as shown in Diagram 2. Note that $R_{t,t-1}$ used in (3) is computed only once being based on all data from all four cells in all SID's.

Diagram 2

Within SID c for Time Interval (t-1,t)

ESTABLISHMENTS	OCCUPATIONS +	OCCUPATIONS -	
+	d_{11}	d_{12}	$\bar{d}_{1.}$
-	d_{21}	d_{22}	$\bar{d}_{2.}$
	$\bar{d}_{.1}$	$\bar{d}_{.2}$	$\bar{d}_{..}$

(. denotes averaging)

Using $\bar{d}_{..}$ as an estimate for $d_{ct,t-1}$, we have VAR $(\bar{d}_{..})$ as an estimate for VAR $(d_{ct,t-1})$. An unbiased estimator for VAR $(\bar{d}_{..})$ is

$$(4)\quad \widehat{\mathrm{VAR}}\,(\bar{d}_{..}) = 3/8\,(d_{.1} - d_{.2})^2 + 3/8\,(\bar{d}_{1.} - \bar{d}_{2.})^2 - 1/16\,[(d_{11} - d_{12})^2 + (d_{21} - d_{22})^2 + (d_{11} - d_{21})^2 + (d_{12} - d_{22})^2].$$

(See Appendix A for derivation.)

The proposed variance estimator is consequently of the following form:

$$(5)\quad \widehat{\mathrm{VAR}}\,(R_{t,t-1}) = \frac{1}{(X'''_{t-1})^2}\,[\Sigma\,\widehat{\mathrm{VAR}}\,(d_{ct,t-1..})],$$

where $d_{ct,t-1..}$ denotes $\bar{d}_{..}$ as defined above for stratum c in time interval (t-1,t).

Method II - The difference between Methods I and II lies in the way in which the marginal d's are computed. In Method II, marginal $\bar{d}'_{..}$ is the value of $X''_{tc} - R_{t,t-1}\,X''_{t-1,c}$ computed from all quotes in column 1.

Similarly, the row marginals are computed from all quotes in the two row half-samples, respectively. $\bar{d}'_{..}$ is computed similarly from all quotes in the entire stratum.

Diagram 3

ESTABLISHMENTS	OCCUPATIONS +	OCCUPATIONS -	
+	d_{11}	d_{12}	$\bar{d}'_{1.}$
-	d_{21}	d_{22}	$\bar{d}'_{2.}$
	$\bar{d}'_{.1}$	$\bar{d}'_{.2}$	$\bar{d}'_{..}$

Imputations are performed in a similar fashion, using all quotes within row and column half-samples, respectively.

(6a) Now, since $(\bar{d}_{1.} - \bar{d}_{2.})^2$ is analogous to $2 \sum_{i=1}^{2} (\bar{d}'_{i.} - \bar{d}'_{..})^2$ and similarly:

(6b) $(\bar{d}_{.1} - \bar{d}_{.2})^2 \longrightarrow 2 \sum_{j} (\bar{d}'_{.j} - \bar{d}'_{..})^2$

(6c) $(d_{11} - d_{12})^2 \longrightarrow 2 \sum_{j} (d_{1j} - \bar{d}'_{1.})^2$

(6d) $(d_{11} - d_{21})^2 \longrightarrow 2 \sum_{i} (d_{i1} - \bar{d}'_{.1})^2$

(6e) $(d_{21} - d_{22})^2 \longrightarrow 2 \sum_{j} (d_{2j} - \bar{d}_{2.})^2$

(6f) $(d_{12} - d_{22})^2 \longrightarrow 2 \sum_{i} (d_{i2} - \bar{d}'_{.2})^2$

we can substitute into VAR $(\bar{d}..)$ in (4) to obtain

$$\widehat{\text{VAR}}\,(d_{ct,t-1}) = 3/4 \sum_{i} (\bar{d}'_{.i} - \bar{d}'_{..})^2 + 3/4 \sum_{j} (\bar{d}'_{.j} - \bar{d}'_{..})^2 - 1/8 \left[\sum_{j} (d_{1j} - \bar{d}'_{1.})^2 + \sum_{i} (d_{i1} - \bar{d}'_{.1})^2 + \sum_{j} (d_{2j} - \bar{d}'_{2.})^2 + \sum_{i} (d_{i2} - \bar{d}'_{.2})^2 \right] \quad (7)$$

Method III - This method ignores the covariance terms by using the biased estimator,

$$\widehat{\text{VAR}}\,(d_{ct,t-1}) = 1/2 \left[W_1\, 1/4\, (d_{11} - d_{22})^2 + W_2\, 1/4\, (d_{12} - d_{21})^2 \right], \quad (8)$$

where w_i's adjust for cell size differences. (ie.: A weighted average of the squared differences of the replicates positioned on the diagonals in the Diagram 2.) See Appendix B for derivation.

From these three methods, three separate estimates will be obtained. The estimates produced by methods I and II will provide a measure of the impact of the survey in which imputation is done within replicates. Method II is expected to be the more appropriate estimator since it more closely reflects the sampling design imputation scheme. Method I, cheaper and easier to implement, will act as a check on the estimate produce by Method II. Method I will be used if resources prove insufficient for implementation of Method II. If the estimates computed from Methods I and II are negative, there would be evidence that the variances of these estimators are too large. Consequently, Method II will be the most appropriate method of estimation.

If the linearization,

$$\text{VAR}\,(R_{t,t-1}) = \frac{1}{(X''')^2_{t-1}} \sum \text{VAR}\,(d_{ct,t-1}) \quad ,$$

is judged to be inappropriate, psuedo-replicates will be formed to estimate VAR $(R_{t,t-1})$ directly as mentioned earlier. The three methods are again available for this purpose, substituting $R_{t,t-1}$ in place of $d_{ct,t-1}$ in the previous discussion. The functions and characteristics of these methods are the same as discussed before. The psuedo-replicates will be selected with replacement as a random sample from the universe of all possible psuedo-replicates.

There are 8 possible patterns for organizing the 4 cells:

a

1	2
4	3

b

1	4
2	3

c

3	2
4	1

d

3	4
2	1

e

2	1
3	4

f

2	3
1	4

g

4	1
3	2

h

4	3
1	2

A psuedo-replicate is formed by selecting one pattern from each of the 8 subindustry divisions. These patterns are then collapsed by combining quotes in similarly positioned cells to form four estimates of the relative. The estimate of the variance for the k^{th} replicate, $\widehat{\text{Var}}_k\,(\bar{R}_{t,t-1})$, is constructed by replacing the four d's with these four estimates of the relative in (4), (7) and (8).

The total number of ways to combine these 8 patterns is 8^8 since there are 8 subindustry divisions. Note, however; that each combination belongs to a group of 8 which produce the same estimate. For example, the following 8 combinations of patterns all yield the same result:

Combination	Sub Ind Group 1	2	3	4	5	6	7	8
1	a	b	c	d	a	b	a	g
2	b	a	d	c	b	a	b	h
3	c	d	a	b	c	d	c	e
4	d	c	b	a	d	c	d	f
5	e	g	f	h	e	g	e	b
6	f	h	e	g	f	h	f	a
7	g	e	h	f	g	e	g	d
8	h	f	g	e	h	f	h	c

Now computing the probability of selecting at least one duplicate psuedo-replicate, we have:

Prob (at least 1 duplicate) = 1 - P

where P = Prob (no duplicates)

Thus,

$$P = 1\left(1 - \frac{8}{8^8}\right)\left(1 - \frac{16}{8^8}\right) \ldots \left(1 - \frac{8(M-1)}{8^8}\right)$$

where M = number of psuedo-replicates generated

$$P = \prod_{k-1}^{M} \left(1 - \frac{8(k-1)}{8^8}\right)$$

To approximate this, we shall take logs of both sides:

$$\ln P = \sum_{k-1}^{M} \ln \left(1 - \frac{8(k-1)}{8^8}\right)$$

and since $\frac{8(k-1)}{8^8}$ is small,

$$\ln P = \sum_{k}^{M} \left[- \frac{8(k-1)}{8^8} - \frac{1}{2} \left(\frac{8(k-1)}{8^8}\right)^2 \right]$$

$$= - \frac{4M(M-1)}{8^8} \left[1 + \frac{4(2M-1)}{(3)\,8^8}\right] .$$

Thus,

$$P \text{ (at least 1 duplicate)} \doteq 1 - e^{-\frac{4M(M-1)}{8^8}\left[1 + \frac{4(2M-1)}{(3)8^8}\right]}.$$

For M = 500 the probability is approximately 0.058 and for M = 300 the probability is approximately 0.021.

An alternative to this selection process would be to create a set of replicates which is in some sense "balanced." Lack of resources has thus far precluded such an attempt. An interesting problem would be to derive an algorithm that would enable a balanced set of replicates to be generated efficiently.

In summary, the variance estimators are of the following forms:

With linearization ----------

$$\widehat{VAR}\ (R_{t,t-1}) = \frac{1}{(X''_{t-1})^2} \sum_{c=1}^{9} \widehat{VAR}\ (d_{ct,t-1}) \quad .$$

Without linearization ----------

$$\widehat{VAR}\ (R_{t,t-1}) = \frac{1}{k} \sum_{1}^{k} \widehat{VAR}_k\ (\bar{R}_{t,t-1})$$

where $\widehat{VAR}_k\ (\bar{R}_{t,t-1})$ is the estimate obtained from the k^{th} pseudo-replicate applying Method I, II or III after collapsing across all SID's.

APPENDIX A

The following discussion follows from the assumption that the distributions of d_{pq} (p=1,2 and q=1,2) are the same, and without lost of generality, $E(d_{pq})=0$.

We have:

$$(1) \quad VAR\ (d..) = VAR\ (1/4 \sum_{p=1}^{2} \sum_{q=1}^{2} d_{pq})$$

$$= 1/16\ [4VAR\ (d_{pq}) + 2(2\ Cov(d_{11},d_{12})) + 2(2Cov\ (d_{11},d_{21}))]$$

$$= 1/4\ [\sigma^2\ (1 + \rho_1 + \rho_2)],$$

where $\sigma^2 = VAR\ (d_{pq})$,

$$\rho_1 = \frac{Cov\ (d_{11}, d_{12})}{\sigma^2}$$

$$\rho_2 = \frac{Cov\ (d_{11}, d_{21})}{\sigma^2}$$

Now

$$(2) \quad VAR\ (\bar{d}_{1.} - \bar{d}_{2.}) = VAR\ 1/4\ (d_{11} + d_{12} - d_{21} - d_{22}).$$

Consequently,

$$VAR\ (\bar{d}_{1.} - \bar{d}_{2.}) = 1/4\ [4\ VAR\ (d_{pq}) + 2\ Cov\ (d_{11}, d_{12}) - 2\ Cov\ (d_{11}, d_{21}) - 2\ Cov\ (d_{12}, d_{22}) + 2\ Cov\ (d_{21}, d_{22})]$$

$$= 4\sigma^2\ (1 + \rho_1 - \rho_2).$$

Similarly,

$$E\ (\bar{d}_{.1} - \bar{d}_{.2})^2 = 4\sigma^2\ (1 + \rho_2 - \rho_1).$$

Also,

$$E\left[1/2\ \left[1/4 \sum_q \sum_p (d_{pq} - \bar{d}_{.q})^2 + 1/4 \sum_q \sum_p (d_{pq} - \bar{d}_{p.})^2\right]\right]$$

$$= E\left[1/8\ (d_{11} - d_{21})^2 + (d_{12} - d_{22})^2 + (d_{11} - d_{12})^2 + (d_{12} - d_{21})^2\right]$$

$$= 1/8\ (8\ \sigma^2 - 4\ \sigma^2\ \rho_1 - 4\ \sigma^2\ \rho_2)$$

$$= \sigma^2 - \frac{\sigma^2\ \rho_1}{2} - \frac{\sigma^2\ \rho_2}{2} = \sigma^2\left(1 - \frac{\rho_1 + \rho_2}{2}\right).$$

Thus, we have the following three unbiased estimators from which respective estimates of ρ_1, ρ_2, and σ^2 may be obtained.

$$(\bar{d}_{.1} - \bar{d}_{.2})^2 \xrightarrow{est} \sigma^2\ (1 - \rho_1 + \rho_2)$$

$$(\bar{d}_{1.} - \bar{d}_{2.})^2 \xrightarrow{est} \sigma^2\ (1 + \rho_1 - \rho_2)$$

$$1/8\ \left(\sum_q \sum_p (d_{pq} - \bar{d}_{.q})^2 + \sum_q \sum_p (d_{pq} - \bar{d}_{p.})^2\right) \xrightarrow{est} \sigma^2\ \left(1 - \frac{\rho_1 + \rho_2}{2}\right)$$

Now to estimate VAR $(\bar{d}..)$, we define

$$\underset{\sim}{X} = \begin{pmatrix} (\bar{d}_{.1} - \bar{d}_{.2})^2 \\ (\bar{d}_{1.} - \bar{d}_{2.})^2 \\ 1/8\ [(d_{11} - d_{12})^2 + (d_{21} - d_{22})^2 + (d_{11} - d_{21})^2 + (d_{12} - d_{22})^2] \end{pmatrix}$$

$$\underset{\sim}{g}^2 = \begin{pmatrix} \sigma^2 \\ \sigma^2\rho_1 \\ \sigma^2\rho_2 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1/2 & -1/2 \end{pmatrix}$$

Then

$$E\left[\underset{\sim}{X}\right] = D\ \underset{\sim}{g}^2 \quad \text{and} \quad E\left[D^{-1}\ \underset{\sim}{X}\right] = \underset{\sim}{g}^2$$

which implies

$$E\left[\underset{\sim}{C}\ D^{-1}\ \underset{\sim}{X}\right] = \underset{\sim}{C}\ \underset{\sim}{g}^2 = VAR\ (\bar{d}..),$$

where $\underset{\sim}{C}$ = (1/4 1/4 1/4) and $D^{-1} = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 3/4 & -1 \\ 3/4 & 1/4 & -1 \end{pmatrix}$.

Finally, we have an unbiased estimator of VAR(d..), namely,

$$\widehat{VAR}\ (\bar{d}..) = \underset{\sim}{C}\ D^{-1}\ \underset{\sim}{X}$$

$$= 3/8\ (\bar{d}_{.1} - \bar{d}_{.2})^2 + 3/8\ (\bar{d}_{1.} - \bar{d}_{2.})^2 - 1/16\ [(d_{11} - d_{12})^2 + (d_{21} - d_{22})^2 + (d_{11} - d_{21})^2 + (d_{12} - d_{22})^2].$$

APPENDIX B

We first define $\bar{d}.. = \frac{D_1 + D_2}{2}$,

where $D_1 = \frac{d_{11} + d_{22}}{2}$ and $D_2 = \frac{d_{12} + d_{21}}{2}$.

Now we have:

(1) $\text{VAR}(D_1) = 1/4\,[\text{VAR}(d_{11} + \text{VAR}(d_{22}))]$

$= 1/4\ \text{VAR}(d_{11} - d_{12})$

$= 1/4\ E(d_{11} - d_{22})^2$.

Similarly,

$$\text{VAR}(D_2) = 1/4\ E(d_{12} - d_{21})^2 .$$

Ignoring the covariance term, $\text{VAR}(\bar{d}..)$
$+ 1/4\,[\text{VAR}(D_1) + \text{VAR}(D_2)]$.

Thus, $\text{VAR}(\bar{d}..)$ will be estimated by

$\widehat{\text{VAR}}(\bar{d}..) = 1/4\,[W_1\ 1/4\ (d_{11} - d_{22})^2$

$= W_2\ 1/4\ (d_{12} - d_{21})^2]$

where W_1 and W_2 are determined in the following manner to adjust for differences in sample size.

Note that

$$E(d_{11} - d_{22})^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2} = \sigma_1^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

and $E(d_{12} - d_{21})^2 = \sigma_1^2 \left(\frac{1}{m_1} + \frac{1}{m_2}\right)$

where σ_1^2 is the unit variance within each of the four cells for each SID and n and m are the cell sizes:

n_1	m_1
m_2	n_2

.

Now,
$$\text{VAR}(d..) = \frac{\sigma_2^2}{n_1 + n_2 + m_1 + m_2}$$

where σ_2^2 is the unit variance for the entire SID, is estimated by $\text{VAR}(\bar{d}..)$.

Now,
$$E[\widehat{\text{VAR}}(\bar{d}..)] = \frac{\sigma_1^2}{4}\left[W_1\ 1/4\ \left(\frac{1}{n_1} + \frac{1}{n_2}\right) + W_2\ 1/4\ \left(\frac{1}{m_1} + \frac{1}{m_2}\right)\right].$$

In order for

$E[\text{VAR}(\bar{d}..)] = \text{VAR}(\bar{d}..)$ under the assumption that $\sigma_1^2 = \sigma_2^2$,

we must have

$$W_1 = \frac{\frac{8n_1n_2}{n_1 + n_2}}{n_1 + n_2 + m_1 + m_2} .$$

and $$W_2 = \frac{\frac{8m_1m_2}{m_1 + m_2}}{n_1 + n_2 + m_1 + m_2} .$$

FOOTNOTE

1/ See unpublished paper by Joseph Steinberg entitled "Sampling Aspects of the Employment Cost Index" and paper by E. Hoy "General Survey Design Aspects of the ECI."