The Two Sample Problem  October 2008

Alan H. Dorfman  
dorfman.alan@bls.gov

Abstract
It sometimes happens that two separate samples from a population, having perhaps quite distinct designs
and mode of sampling, contribute information on the same variable of interest, and it becomes an important
question how to combine the data from the two samples. An example is the Occupational Employment
Statistics survey (OES) and the National Compensation Survey (NCS), carried out by the Bureau of Labor
Statistics, both contributing information on occupational wages. We discuss some new options for
combining data from two samples and achieving unified estimation.

Key Words: Missing Information Principle, Post-stratification, Pseudo-likelihood

1. Introduction – The Question of Combining Two Samples

Suppose two distinct surveys gather related information on a single population \( U \). How best to combine
data from the two surveys to yield a single set of estimates?

This situation is fairly common. For example, the National health Interview Survey (NHIS) and the
Behavioral Risk Factor Surveillance System (BRFSS) both collect data on risk factors for serious illness.
The Occupation Employment Statistics program (OES) and the National Compensation Survey (NCS) both
collect data on occupational employment and wages.

The typical approach is to get separate estimates of the target parameter from the two surveys and weight
them together with weights the inverse of their estimated variances. For example, see (Merkouris 2004)
and the references therein.

Another possibility, however, is to combine the two data sets into a single data set and “weight up”
appropriately. This can be awkward in the standard design based framework, but nevertheless can pay
dividends.

In this paper we consider a simple example where this turns out to be the case. This example is a
simplified version of the OES/NCS situation, and was constructed in the hopes of shedding light on what
might best work in unifying those two surveys.

2. Study Example

Consider a population \( U \) of \( N \) establishments \( e \), each with \( O \) occupations \( o \). Two samples of
establishments, \( s_1 \) of size \( n_1 \) and \( s_2 \) of size \( n_2 \), with data collected for all occupations in \( s_1 \) but only for a sub-
sample of \( k \) occupations in \( s_2 \). Let \( E_{oe} \) be the employment in \( o \) in \( e \), and \( E_e = \sum_o E_{oe} \) the employment in \( e \).

The goal is to estimate the proportion \( p_o = \frac{E_{oe}}{E_{ee}} \) of employment in each of the \( O \) occupations. Here
\( E_{oe} = \sum_e E_{oe} \) is total employment in occupation \( o \) and \( E_{ee} = \sum_o E_{oe} \) is total employment.

In this context, it is natural to consider a Multinomial model, where \( \{E_{oe}\} \sim \text{Multinom}(E_{ee}, \{p_o\}) \). Then
the population log likelihood would be \( \log L = \sum_e \left( \sum_{o=1}^{O-1} E_{oe} \log p_o + E_{oe} \log p_o \right) \), with score functions
\[ \frac{\partial L}{\partial p_o} = \sum_v \left( \frac{E_{ov} - E_{Ov}}{p_o} \right) = \frac{E_{ov}}{p_o} - \frac{E_{Ov}}{p_O}, \quad o = 1, 2, ..., O - 1. \]  

(1)

Thus, if we had data for whole population, we would set the score function to zero and solve, getting estimates

\[ \hat{E}_o = \frac{\hat{E}_{ov}}{\hat{E}_{ov}^{(O)}}, \quad o = 1, 2, ..., O, \]

so that, not surprisingly, the estimates would equal the targets: \( \hat{E}_o = p_o \). But, of course, we do not have the population data, only what has been collected in the two samples, plus some possible population auxiliary information. We shall assume that overall establishment employment \( E_e \) is available for all establishments, in and out of sample.

2.1 Pseudo-likelihood

One commonly accepted way to handle maximum likelihood problems in the survey context is to use Pseudo-likelihood. Replace population sums in the population score function by sample estimates of those sums, using the inverse of selection probabilities, the design weights \( w_e \):

\[ \frac{\partial L}{\partial p_o} = \sum_v \left( w_v E_{ov} - w_v E_{Ov} \right) = \frac{\hat{E}_{ov}}{p_o} - \frac{\hat{E}_{Ov}}{p_O}, \quad o = 1, 2, ..., O - 1. \]  

(2)

Setting these pseudo-likelihood score functions to zero and solving leads to \( \hat{E}_o = \frac{\hat{E}_{ov}}{\hat{E}_{ov}^{(O)}} \).

If, for example, on sample 1, units are selected \( pps(E_e) \), we could get an estimate from that sample using

\[ \hat{p}_o^{(1)} = \frac{1}{n_1} \sum_{\text{exc}1} \frac{E_{ov}}{E_e}, \text{the pseudo-likelihood estimator for sample 1}. \]

For combining the surveys we could, similarly, get an estimate from second sample, and then weight these together based on our estimate of the variances of the two estimates. This would be in keeping with the approach usually suggested. Can we use pseudo-likelihood on a combined version of the data from the two surveys?

2.2 Dual sample pseudo-likelihood estimator

Suppose the two samples are independent, with inclusion probabilities \( \pi_{oe}^{(1)}, \pi_{oe}^{(2)} \).

Then the probability that \( oe \) gets into the combined sample is \( \pi_{oe}^{(\text{dual})} = \pi_{oe}^{(1)} + \pi_{oe}^{(2)} - \pi_{oe}^{(1)} \pi_{oe}^{(2)} \).

Let \( w_{oe}^{(\text{dual})} = 1 / \pi_{oe}^{(\text{dual})} \). Then we can use these weights in (1) to get a pseudo-likelihood estimator

\[ \hat{p}_o^{(\text{dual})} = \frac{\sum_{oe, \text{comb}} w_{oe}^{(\text{dual})} E_{oe}}{\sum_{o} \sum_{oe, \text{comb}} w_{oe}^{(\text{dual})} E_{oe}} \]

This idea of using a combined inclusion probability underlies a lot of the work on the problem of dual frame estimation.
2.3 Missing Information Principle

Another approach, quite different from pseudo-likelihood, is the application of the Missing Information Principle (Breckling, et al. 1994). This gives the actual maximum likelihood estimates of the targets, given all the data (from each sample and the auxiliary data) and the truth of the model, in this case the simple multinomial model. Here is a sketch of the principles involved:

Suppose \( \frac{\partial \log L(x; \omega)}{\partial \omega} \) is the score function if we have “full information” \( x \), and let \( D \) represent the portion of \( x \) for which we actually have data, then the score function given that data can be written as the expectation of the original score function conditional on \( D \),

\[
\frac{\partial \log L(D; \omega)}{\partial \omega} = E(\frac{\partial \log L(x; \omega)}{\partial \omega} | D).
\]

This is a strict application of maximum likelihood and depends strongly on the truth of the (multinomial) model. Note that there is no use of inclusion probabilities. We spell out how this works in the present example:

Units \( e \) fall into three categories, depending on what we know about them, as described in Table 1.

**Table 1:** Classification of units (“establishments”) in the Example

<table>
<thead>
<tr>
<th>( o = 1 )</th>
<th>( o = 2 )</th>
<th>...</th>
<th>( o = O )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_1 )</td>
<td>( E_{1e} )</td>
<td>( E_{2e} )</td>
<td>( E_{Oe} )</td>
</tr>
<tr>
<td>( s_{2:1} )</td>
<td>*** or ( E_{1e} )</td>
<td>*** or ( E_{2e} )</td>
<td>*** or ( E_{Oe} )</td>
</tr>
<tr>
<td>( r = U - (s_1 \cup s_2) )</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>( r = U - (s_1 \cup s_2) )</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
</tbody>
</table>

In the table, “***” means “data is missing”. Thus, the units that are in neither sample, are always missing employment on occupations, those in \( s_1 \) always have occupational employment (whether or not they are also in \( s_2 \)), and those in \( s_2 \) but not in \( s_1 \) will, for a given occupation, possibly have the occupations employment, but possibly not.

Thus the expectation of the population score function (1), conditional on the sample data, takes the form

\[
E\left(\frac{\partial L}{\partial p_o} | D\right) = \sum_{e \in s_1} E\left(\frac{E_{we} - E_{Oe}}{p_o} \bigg| D\right) + \sum_{e \in s_2} E\left(\frac{E_{we} - E_{Oe}}{p_o} \bigg| D\right) + \sum_{e \in r} E\left(\frac{E_{we} - E_{Oe}}{p_o} \bigg| D\right)
\]

The last term is zero:

\[
E\left(\frac{E_{we} - E_{Oe}}{p_o} \bigg| E_{we}\right) = E\left(\frac{E_{we} - E_{Oe}}{p_o} \bigg| E_{we}\right) = 0
\]

The second term is messy:

\[
E\left(\frac{E_{we} - E_{Oe}}{p_o} \bigg| D\right) = E\left(\frac{E_{we} - E_{Oe}}{p_o} \bigg| E_{we} | E_{we}\right)
\]

where \( s_e \) is the sample of occupation in establishment \( e \).
For example, if both $o$ and $O$ are in $s_e$, then the term looks like terms in the $s_1$ sum. If $o$ is not in $s_e$, then, letting $E_{oe} = E_{oe} - \sum_{o' \in s_e} E_{o'e}$ and $p_o' = \frac{P_o}{1 - \sum_{o' \in s_e} p_o'}$, we have $E(E_{oe} | E_{oe'}, E_{o'e}) = E_{oe'} p_{o'}$

The net result is the sample score function

$$E \left( \frac{\partial L}{\partial p_o} | D \right) = \frac{1}{p_o} \left\{ \sum_{e \in s} E_{oe} + \sum_{e \in s_2} E_{oe} + \sum_{e \in s_2} E_{oe} p_{o'} \right\} - \frac{1}{p_o} \left\{ \sum_{e \in s} E_{oe} + \sum_{e \in s_2} E_{oe} + \sum_{e \in s_2} E_{oe} p_{o'} \right\}.$$  

Setting this equal to zero to get estimates of the $p_o$ leads to

$$\hat{p}_o = \frac{\bar{E}_o}{\sum_{o \in s} \bar{E}_o}, \quad (3)$$

where $\bar{E}_o = \sum_{e \in s} E_{oe} + \sum_{e \in s_2} E_{oe} + \sum_{e \in s_2} E_{oe} p_{o'}$. This looks simple, but actually is not, since the $p_o$ we are solving for are components of the terms on the right side of (3). However, we can solve by iteration, plugging in initial values for $p_o$ into the expressions for the $p_o'$, to calculate the right side of (3), giving to get new values $\hat{p}_o$. These new values can in turn to be plugged in, and the process repeated. This works well.

2.4 Iterated post-stratified estimator

What if, in estimating, we do not wish to assume the multinomial model? One approach involves the well known sampling technique of poststratification. In post-stratification, the population is divided into strata after the sample data are collected. Estimation proceeds as if these were original design strata.

If the stratification variable $x$ is continuous, then it often seems to be arbitrary where stratum boundaries are drawn. A way around this arbitrariness, is to do several poststratifications, using systematically different boundaries. In the iterated post-stratified estimator several post-stratifications are constructed, each leading to a single estimate. The final estimate is taken as a simple average of these.

For the simplified NCS-OES like population with two samples, this would be accomplished by the following sequence:

(1) we take the stratification variable to be $E_e$.

(2) For each occupation $o$, let $s(o)$ be the set of $e$ for which the combined sample data provide $E_{oe}$. (that is, all sample 1 $e$’s and sample 2 $e$’s, for which $E_{oe}$ is sub-sampled) Let $n_o$ be the size of this set.

(3) Let $k$ be number of post-stratifications. This is at the discretion of the analyst. How to choose a best $k$ is an open question.

(4) for the lower bound of lowest stratum of any of the poststratifications, we can take any $E^*$ that is less than all $E_e$ in the population. Likewise, the upper bound of highest stratum is any $E^{**}$ greater than all $E_e$.

(5) In first post-stratification take intermediate bounds of post-strata halfway between values of $E_e$ at $jk$th and $jk + 1$ sample points in $s(o)$, for $j = 1, \ldots, [n_o/k]$.

For the second post-stratification, take intermediate bounds between the $jk$th +1 and $jk$th + 2 sample points in $s(o)$, etc.
This puts \( n_{h(i,j)} = k \) sample units in each post-strata \( h(o,i) \) of \( i \)th poststratification, except possibly in the boundary strata.

The number of population units \( N_{h(i,j)} \) with \( E \) within the stratum boundaries will in general vary from stratum to stratum. The following Figure depicts schematically the choice of stratum boundaries for the several poststrata for a given occupation.

**Poststratification Schema -- Two Poststrata**

(6) Then estimate the employment in \( o \) in the stratum by

\[
\hat{E}_{o,h(i,j)} = \frac{N_{h(i,j)}}{n_{h(i,j)}} \sum_{e \in h(o,i) \cap s(o)} E_{e} ,
\]

and get overall

\[
\hat{E}_{o(i)} = \sum_{i} \hat{E}_{o,h(i,j)} ,
\]

the estimated employment in \( o \), based on the \( i \)th postratification.

(7) Take the average of these as the overall estimate of employment in the occupation:

\[
\hat{E}_{o} = \frac{\sum_{i} \hat{E}_{o(i)}}{k} .
\]

**Note:** The implicit weight \( w_{oe} \) on sample value \( E_{oe} \) is

\[
w_{oe} = \frac{1}{k} \sum_{i=1}^{k} I(e \in h(o,i) \cap s(o)) \frac{N_{h(i,j)}}{n_{h(i,j)}} .
\]

(7) Repeat for the other occupations \( o' \).

(8) Take \( \hat{p}_{o} = \hat{E}_{o} / \sum_{o'} \hat{E}_{o'} \) as the iterated post-stratified estimator of \( p_{o} \).
3. Simulation Studies

3.1 Simulation Study 1

3.1.1 Population 1 - Description
Using a lognormal random generator, we created 500 “establishments” with sizes $E_e$ ranging from 12 to 385. Their median size was 97. Each establishment’s employment was divided among 5 “occupations” using a multinomial distribution. In particular, occupational employments in $e$ were generated by

$$\text{Multinom}(E_e, \mathbf{p} = (.05, .1, .23, .27, .35))$$

The corresponding proportion of employment in the population turned out to be

$$\mathbf{p}_u = (.051, 0.098, 0.230, 0.271, 0.350)$$

The goal is to estimate $\mathbf{p}_u$, based on data from two samples.

3.1.2 Sampling Methodology and Estimators

500 pairs of samples were taken:

Sample 1 $\text{pps}(E_e)$ $n_1 = 50$ – with a census of $e$’s Occupations
Sample 2 $\text{pps}(E_e)$ $n_2 = 30$ – in this case a sub-sample $s_o$ of 2 occupations was sampled $\text{ppswr}(E_o)$.

We calculated the following estimators:

(i) the pseudo-likelihood estimator

$$\hat{\mathbf{p}}^{[1]}_o = \frac{1}{n_1} \sum \frac{E_{oe}}{E_e}$$

(ii) the pwr estimator on sample 2, based on with replacement sampling of occupations:

$$\hat{\mathbf{p}}^{[2]}_o = \frac{1}{2n_2} \sum \sum I_o k$$

where $I_o k = 1$, if $o$ selected at $k$th selection of occupation in $e$, and $I_o k = 0$, otherwise.

(iii) A weighted combination of the above estimators:

Let $w_i = \text{var}^{-1}(\hat{\mathbf{p}}^{[i]}_o)$, estimated empirically over the 500 samples (actually used median of these across the 5 occupations)

$$\hat{\mathbf{p}}^{[w]}_o = \frac{w_1 \hat{\mathbf{p}}^{[1]}_o + w_2 \hat{\mathbf{p}}^{[2]}_o}{w_1 + w_2}$$

(iv) $\hat{\mathbf{p}}^{[\text{mip}]}_o$, the MIP max’m likelihood estimator

(v) $\hat{\mathbf{p}}^{[\text{mip}]}_o$, the MIP max’m likelihood estimator, using only data from sample 1

(vi) $\hat{\mathbf{p}}^{[\text{dual}]}_o$ the dual estimator, i.e. the pseudo-likelihood estimator, using two-sample $\pi$’s

(vii) $\hat{\mathbf{p}}^{[\text{iter.ps}]}_o$ the iterated post-stratified estimator

3.1.3 Results

Relative biases and root mean square errors are given in Table 2. There is not too much to distinguish among the estimators with respect to Bias.

With respect to root mean square error: There is little difference between pseudo -1 (the pseudo-likelihood estimator based on just sample 1) and the combined estimator. In other words, combining the results of estimation from the two surveys does not improve on just using the big survey. The mip is distinctly best, and does improve on mip -1. That is, using this estimator—which combines the data from the two surveys
Table 2: Simulation Results – Population 1

<table>
<thead>
<tr>
<th></th>
<th>Occ1</th>
<th>Occ2</th>
<th>Occ3</th>
<th>Occ4</th>
<th>Occ5</th>
<th>Occ1</th>
<th>Occ2</th>
<th>Occ3</th>
<th>Occ4</th>
<th>Occ5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>pseudo -1</strong></td>
<td>-2.5</td>
<td>1.4</td>
<td>0.7</td>
<td>-0.3</td>
<td>-0.2</td>
<td>2.7</td>
<td>3.7</td>
<td>5.5</td>
<td>5.6</td>
<td>6.2</td>
</tr>
<tr>
<td><strong>pwr -2</strong></td>
<td>-11.8</td>
<td>8.8</td>
<td>4.2</td>
<td>3.7</td>
<td>-6.4</td>
<td>30.0</td>
<td>38.2</td>
<td>56.2</td>
<td>56.4</td>
<td>63.4</td>
</tr>
<tr>
<td><strong>combined</strong></td>
<td>-2.5</td>
<td>1.6</td>
<td>0.7</td>
<td>-0.3</td>
<td>-0.3</td>
<td>2.7</td>
<td>3.7</td>
<td>5.5</td>
<td>5.6</td>
<td>6.2</td>
</tr>
<tr>
<td><strong>mip</strong></td>
<td>7.5</td>
<td>-11.2</td>
<td>1.1</td>
<td>4.2</td>
<td>-1.9</td>
<td><strong>2.4</strong></td>
<td><strong>3.1</strong></td>
<td><strong>3.9</strong></td>
<td><strong>4.2</strong></td>
<td><strong>4.7</strong></td>
</tr>
<tr>
<td><strong>mip - 1</strong></td>
<td>7.9</td>
<td>-11.2</td>
<td>1.4</td>
<td>3.6</td>
<td>-1.7</td>
<td>2.4</td>
<td>3.3</td>
<td>4.7</td>
<td>4.9</td>
<td>5.5</td>
</tr>
<tr>
<td><strong>dual</strong></td>
<td>-1.9</td>
<td>3.1</td>
<td>-0.8</td>
<td>1.2</td>
<td>-1.0</td>
<td>3.3</td>
<td>5.1</td>
<td>10.8</td>
<td>11.5</td>
<td>13.5</td>
</tr>
<tr>
<td><strong>iterated ps</strong></td>
<td>2.0</td>
<td>5.0</td>
<td>1.7</td>
<td>-0.9</td>
<td>-2.1</td>
<td>2.9</td>
<td>3.8</td>
<td>5.2</td>
<td>5.1</td>
<td>5.8</td>
</tr>
</tbody>
</table>

-- does capitalize on the extra bit of information in sample 2. The iterated post-stratified estimator is here sometimes better, sometimes marginally worse than the pseudo-likelihood estimator. The dual estimator based on overall inclusion probabilities is considerably worse than the other estimators relying on both samples.

### 3.2 Simulation Study 2

**3.2.1 The Question: What if the Multinomial Model is Wrong?**

Consider the following Population: Workers in e with large $E_e$ have probability $p_{o}^l$ of being in $o$. Workers in e with small $E_e$, have probability $p_{o}^s$ of being in $o$. Suppose $p_{o}^l < p_{o}^s$ and we aim to estimate $p_o = \sum_{e, E_e \in \mathcal{E}} E_e / \sum_{e, E_e \in \mathcal{E}} E_e$ based on a single sample, where sampling was pps, with size measure $E_e$. In this, single sample case, the pseudo-likelihood and the mip estimators take the forms $\hat{p}_{o}^{\text{ml}} = \frac{1}{n_1 \sum_{o, E_e \in \mathcal{E}}} \sum_{e, o} E_e$ and $\hat{p}_{o}^{\text{mp}} = \frac{\sum_{o, E_e \in \mathcal{E}} E_e}{\sum_{o, E_e \in \mathcal{E}}}$. On the one hand, $\hat{p}_{o}^{\text{ml}}$ is design-unbiased for $p_o$. But in the mip $\hat{p}_{o}^{\text{mp}}$, large $E_e$ dominate sums in numerator and denominator. Under the assumptions above, one can anticipate that $p_{o}^{\text{mp}} \approx p_{o}^l < p_{o}^s$, so that the mip will be seriously biased down. Thus the success of the mip depends on getting the model right.

**3.2.2 Population 2 - Description**

For Population 2 we use the same establishment employments $E_e$ as Pop 1, and same first stage samples. But $E_e$ has distinct model for 250 smallest and 250 largest $E_e$, as in Table 3. Second stage sampling is carried out as it was in Simulation Study 1.

Table 3: Probabilities in Population 2 of Allocation of Employment to Different occupations

<table>
<thead>
<tr>
<th>Size\occupation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small Establishments</strong></td>
<td>0.05</td>
<td>0.10</td>
<td>0.23</td>
<td>0.27</td>
<td>0.35</td>
</tr>
<tr>
<td><strong>Large Establishments</strong></td>
<td>0.35</td>
<td>0.27</td>
<td>0.23</td>
<td>0.10</td>
<td>0.05</td>
</tr>
</tbody>
</table>

For occupation 4 and 5, MIP based on overall multinomial model should be biased low and for 1 and 2, high.
Table 4: Overall Population 2 Proportions to be Estimated

<table>
<thead>
<tr>
<th>Occupation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>0.25</td>
<td>0.21</td>
<td>0.23</td>
<td>0.16</td>
<td>0.15</td>
</tr>
</tbody>
</table>

3.2.3 Simulation Study 2 Results

Relative biases and root mean square errors are given in Table 5. The relative biases do show a striking differences between the mip and the other estimators, as was predicted:

Table 5: Simulation Results for Population 2

<table>
<thead>
<tr>
<th>Relative Bias X 1000</th>
<th>Root Mean Square Error X 1000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Occ1</td>
</tr>
<tr>
<td>pseudo -1</td>
<td>5.2</td>
</tr>
<tr>
<td>pwr -2</td>
<td>-11.0</td>
</tr>
<tr>
<td>combined</td>
<td>3.5</td>
</tr>
<tr>
<td>mip</td>
<td>193.9</td>
</tr>
<tr>
<td>dual</td>
<td>-0.1</td>
</tr>
<tr>
<td>iterated ps</td>
<td>-4.1</td>
</tr>
</tbody>
</table>

Its bias leads to the mip having large rmse. Indeed, the pseudo-likelihood estimator based just on sample 1 does much better than mip based on both samples. The combined estimator is marginally better than sample 1 pseudo-likelihood. The dual estimator is sometimes better than sample 1 pseudo-likelihood estimator. But the clear winner is the iterated post-stratified estimator.

4. Discussion

Success of mip depends on model being correct. However, it may be that Population 2 is an extreme case. We might be able in practice to use prior knowledge or diagnostics to handle this problem – this is something to investigate. In neither population, did the design based ways of combining estimates or data yield best results. The iterated post-stratified estimator looks to be a very promising non-parametric way to combine data from two sources.

Acknowledgements

Thanks to Ronald Johnson of BLS for comments improving the presentation of this paper.

References


The views expressed are the author’s and do not reflect Bureau of Labor Statistics policy.