# An Alternative Estimator for Multi-Frame Sample Designs

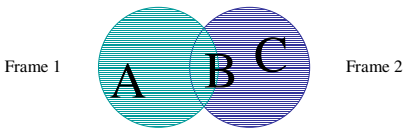
Charles D. Palit

Strictly speaking dual frame sampling refers to any sample design in which two different and usually overlapping frames are used to draw a sample for the same universe. Multi-frame designs are designs in which K frames are sampled where K is greater than 2 and and each frame contains some elements in common with another of the frames used to draw the sample. If there is no overlap in a dual or multiple frame situation then the design reduces to the familiar stratified sample design. Each frame being a stratum and the multi-frame designation is not normally used to describe the design.

In this paper the terms dual and multi frame design will be used to denote designs in which the frames overlap so that there are population elements which occur in both frames

In 1962 Hartley published/presented a paper which has served as the basis for dual frame estimators from 1962 to the present day. Hartley enumerated "Four principal situations " which can occur in dual frame surveys. The four situations are defined in terms of Domains. In a dual-frame sample there are three domains. If we call one frame 1 and the other frame 2, then elements which fall into frame 1 alone form a domain (A). Elements which fall into frame 2 alone form a domain (C) and elements which fall into the overlap form a third domain (B). Operationally every population element belongs to one of these domains.

Illustration of a Dual Frame System With Overlap Showing Domains A, B, and C




Case 1: Situations where all the domain sizes are known and it is possible to precisely control the sample size selected from each domain.
Case 2: Situations where all the domain sizes are known but precise control of the sample size is only possible at the frame level

Case 3: Situations where the domain sizes are not known, but frame sizes are known and it is only possible to control sample sizes at the frame level.

Case 4: Situations where "Neither domain sizes nor frame sizes are known, but the relative magnitude of the frame is known" and "prescribed sample sizes can only be allocated to frames".

The general form of Hartley's estimator for case 2 is :

$$\tilde{X}pop = \tilde{X}frame1\ only + P\tilde{X}frame1\ overlap + Q\tilde{X}frame2\ overlap + \tilde{X}frame2only$$

Where P+Q =1, and $\tilde{X}s$ is an estimator for the population total of the form

$$\tilde{X}_s = (N_s / n_s)\sum_{1}^{n_s} x_i$$

where $N_s$ is the size of domain s, and $n_s$ is the size of the sample from domain s.

The applicable domains are: Elements in Frame 1 only, Elements in Frame 1 and in Frame 2, and finally Elements in Frame 2 only.

For case 3 domain sizes are not known we have

$$\tilde{X}_S = (N_1 / n_1)\left\{\sum_1^{n_a} x_i + P\sum_1^{n_{b'}} x'_i\right\} + (N_2 / n_2)\left\{\sum_1^{nc} x_i + Q\sum_1^{nb''} x''_i\right\}$$

Where Ni is the size of frame i, and ni is the size of the sample drawn from frame i, 1=1,2. Where x' represents values for the sample elements drawn from Frame 1 but also exist in Frame 2, and x" represents values for those sample elments drawn from Frame 2 but also exist in Frame 1. Finally, b' and b" are the sample sizes for x' and x" respectively.

Hartley points out that
"In case 1 the estimation problem is reduced to the standard methodology for stratified sampling, whilst in case 4 it will only be possible to estimate means not totals."

Clearly for case 1 the situation reduces to the familiar stratified sample design if each domain is viewed as a stratum. For case 4 it turns out that population totals can be estimated even if we are using imperfect and other complicated frames. Because we can estimate totals in case 4 we can use this technology to deal with two important sampling problems, improving telephone surveys by combining cell phone frames with line phone frames and sampling rare populations using telephone frames.

This paper looks at two examples of dual frame sample designs which use both a cell phone frame and a line phone frame and two designs suitable for sampling Hispanic populations in the USA. One cell phone and line phone example samples the telephone household population, and the other samples the adult population. For the Hispanic sample examples one is a two frame design and the other is a three frame design.

Before we can discuss these designs we need to find an estimation procedure suitable for the cell phone and line phone frame combination. Even though we are not in case 2, we use Hartley's approach for case 2, and a dual frame situation to take us part of the way there. Think of the two frames as dividing the Universe into three disjoint sets of elements:

- A= All population elements in frame 1 only
- B=All population elements which occur in both frames.
- C =All population elements in frame 2 only

If we look on A, B, and C as not only as separate domains but also as super-strata in the sense that they define a stratification scheme , post or otherwise, which persists through all the frames, then we can use the samples from two frames to give independent samples for each super-stratum. One sample from A, two samples from B, and one sample from C for a total of 4 samples. Each sample delivers one estimate for its parent stratum.

For simplicity consider the case where the selection probability is a constant within each frame, but allowed to vary between frames. If we can produce four independent estimates $X_{1A}$, $X_{1B}$, $X_{2B}$, $X_{2C}$ for the three domains or super-stratum population totals. We have two estimates for B, one from each frame. We can use a weighted average to combine them into one estimate, $X_{12}$, and produce an estimate of the population total with

$$X_{1A} + X_{12} + X_{2C}$$

Where $X_{12} = (w_{1B} X_{1B} + w_{2B} X_{2B})$,
and $w_{1B} + w_{2B} = 1$

If we knew the domain sizes we would be in Hartley's case 2. However we are dealing with the situation where we do not know the domain sizes and precise sample sizes can only be controlled at the frame level which is Hartley's case 4. We can however use Horvitz-Thompson (H-T) estimators for each of the four super-stratum estimators.

Much of the prior work has dealt with case 2 and case 3 and been concerned with finding optimum values for $w_{1B}$ and $w_{2B}$ for estimating some specific variable. Since we are mostly concerned with omnibus surveys we are reluctant to optimize the survey for any specific variable at the possible expense of the other variables measured in the survey. Consequently we use a different criteria for choosing $w_{1B}$ and $w_{2B}$ . We choose practicality as our main criterion and choose $w_{1B}$ and $w_{2B}$ as:

$$w_{1B} = (\pi_1 / (\pi_1 + \pi_2) \text{ and } w_{2B} = (\pi_2 / (\pi_1 + \pi_2).$$

This choice when used in combination with the H-T estimator has interesting consequences. First the estimate $X_{12}$ becomes

$$X_{12} = (\pi_1 / (\pi_1 + \pi_2) X_{1B} + (\pi_2 / (\pi_1 + \pi_2) X_{2B}$$

Where $(\pi_1 / (\pi_1 + \pi_2) + (\pi_2 / (\pi_1 + \pi_2) = 1$

If the H-T estimator for $X_{1B}$ is

$$\sum_{j=1}^{n_{1B}} x_{1Bj} / \pi_1$$

then

$$(\pi_1/(\pi_1+\pi_2))X_{1B} = (\pi_1/(\pi_1+\pi_2))\sum_{j=1}^{n1B} x_{1Bj}/\pi_1$$

$$= \sum_{j=1}^{n1B} x_{1Bj}/(\pi_1+\pi_2)$$

Similarly the estimator $X_{2B}$ is :

$$X_{2B} = \sum_{j=1}^{n2B} x_{2Bj}/(\pi_1+\pi_2)$$

and so the estimator $X_{12}$ becomes

$$X_{12} = \sum_{j=1}^{n1B} x_{1Bj}/(\pi_1+\pi_2) + \sum_{j=1}^{n2B} x_{2Bj}/(\pi_1+\pi_2).$$

Using similar estimates for $X_{1A}$ and $X_{2C}$ the estimator for $X = X_{1A} + X_{12} + X_{2C}$ becomes

$$X = \sum_{j=1}^{n1A} x_{1Aj}/\pi_1 + \sum_{j=1}^{n1B} x_{1Bj}/(\pi_1+\pi_2)$$

$$+ \sum_{j=1}^{n2B} x_{2Bj}/(\pi_1+\pi_2) + \sum_{j=1}^{n2C} x_{2Cj}/\pi_2$$

Or

$$\sum_{S_A,S_B} x_j/(\prod_1+\prod_2)$$

where

$\prod 1 = \pi 1$ if xj $\in$ Frame 1 and 0 otherwise

$\prod 2 = \pi 2$ if xj $\in$ Frame 2 and 0 otherwise

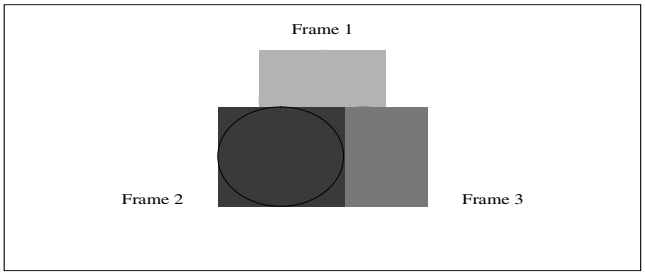
And

$$\sum_{s1,s2} \left(\begin{array}{l}\text{is the sum over both the sample from}\\ \text{Frame 1 and the sample from Frame 2}\end{array}\right)$$

**Extension to Multi-frame Sampling With more than Two Frames**

In the dual frame design we partitioned the Universe into three domains. In the three frame design we partition the Universe into seven domains, A, B, C, D, E, F, and G as shown in the Venn diagram following.

## A Three Frame Design




Looking at each domain as a super-stratum notice that we have:

Two samples from D, one from Frame 1 and one from Frame 2,

Two samples from E, one from Frame 2 and one from Frame 3,

Two samples from F, one from Frame 1 and one from Frame 3, and

Domain G contains sample from all three frames.

We already have seen that the estimators for A, B, and C are of the form

$$\sum_{j=1}^{n1A} x_{1Aj}/\pi_1\ ,\quad \sum_{j=1}^{n1B} x_{2Bj}/\pi_2\ ,\text{ and }\quad \sum_{j=1}^{n1B} x_{3Bj}/\pi_3$$

And the estimators for D, E, and F are of the form

$$\sum_{j=1}^{n1D} x_{1Dj}/(\pi_1+\pi_2) + \sum_{j=1}^{n2D} x_{2Dj}/(\pi_1+\pi_2),$$

$$\sum_{j=1}^{n2E} x_{2Ej}/(\pi_2+\pi_3) + \sum_{j=1}^{n3E} x_{3Ej}/(\pi_2+\pi_3),\text{ and}$$

$$\sum_{j=1}^{n1F} x_{1Fj}/(\pi_1+\pi_3) + \sum_{j=1}^{n3F} x_{3Fj}/(\pi_1+\pi_3)$$

Extending the previous argument it is easy to see that the estimator for G is of the form

$$\sum_{j=1}^{n1G} x_{1Gj}/(\pi_1+\pi_2+\pi_3) + \sum_{j=1}^{n2G} x_{2Gj}/(\pi_1+\pi_2+\pi_3)$$

$$+ \sum_{j=1}^{n3G} x_{3Gj}/(\pi_1+\pi_2+\pi_3)$$

When we use weights of the form

$\pi i/(\pi 1 + \pi 2 + \pi 3)$.

In fact the general solution for a domain covering k frames is

$$\sum_{i=1}^{k}\sum_{j=1}^{ny_i} y_{ij}/\sum_{i=1}^{k}\pi_i$$

Where $y_{ij}$ are the $x_{ij}$ observations from the $i^{th}$ frame and $y^{th}$ intersection

Adding the estimators for the totals in the super-strata A, B, C, D, E, F, and G, together produces an estimate for the Universe total.

**Dealing with unequal selection probabilities inside one or more frames**.

Sometimes the sample design for a frame results in unequal probabilities of selection for the sample selected from the frame. To resolve this situation we need only partition the sample into as many parts as there are different selection probabilities. These parts when crossed with the other frames in the system produce a number of cells and the same technique using these partitions to define super-strata can be used to produced estimates of the population totals for each cell. These estimates can then be added together to produce a population estimate for the entire universe.

Consider a dual frame design in which the sample from one frame is selected with two different selection probabilities, say $\pi_1$ and $\pi'_1 = m*\pi_1$

With equal selection probabilities in each frame the number of super-strata is three, one for each domain A, B, and C. If Frame 1 is split into two parts then the number of super-strata becomes five, say $A_1$, $A_2$, $B_1$, $B_2$, and C where $A_1$ and $B_1$ are sampled using $\pi_1$ , and A2 and B2 are sampled using $\pi'_1$. It is then only necessary to make independent estimates for each of the five super-strata and sum these estimates to arrive at an estimate for the universe as a whole.

$$\sum_{S_A, S_B} x_j /(\prod_1 + \prod_2)$$

where

$\prod_1 = z\pi_1$ if $x_j \in$ Frame 1 and 0 otherwise

z = 1 if $x_j$ is selected with probability $\pi_1$ and m otherwise

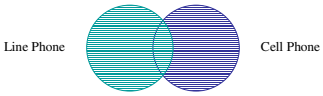
$\prod_2 = \pi_2$ if $x_j \in$ Frame 2 and 0 otherwise

And

$\sum_{S_1, S_2}$ is the sum over the sample from Frame 1 and Frame 2

**Sampling Telephone Households Using A Line and Cell Phone Dual Frame**

Dual Frame Situation for Telephone Households




In a kinder world telephone households would have one line phone and/or one cell phone apiece and this procedure could be applied without further thought. Unfortunately in our world households have variable numbers of line phones and variable numbers of cell phones and this needs to be taken into account in the estimation process because the number of line phone numbers and cell phone numbers attached to a household affects the probability of selecting the household into the sample when an RDD frame is used. So it is important to know the number of phone numbers attached to the household by type of phone (i.e., line or cell) because separate probabilities are needed for each frame.. With some planning, skill and luck this information can be obtained from the respondent during the course of the telephone interview.

With this information we can estimate totals for household characteristics with

$$\sum_{S_1, S_2} x_j /(\prod_l + \prod_c)$$

where

$c_i$ = the number of cell phones attached to a household

$l_i$ = the number of line phones attached to the household

$\pi_{li}$ = the probability of selecting the ith phone number from the line phone frame

$\pi_{ci}$ = the probability of selecting the ith phone number from the cell phone frame

$\prod_l = l_j\pi_{lj}$ if $x_j \in$ Line Frame (Frame 1) and 0 otherwise

$\prod_c = c_j\pi_{cj}$ if $x_j \in$ Cell Frame (Frame 2) and 0 otherwise

And $\sum_{S_1, S_2}$ is the sum over both samples

**Sampling Adults Using A Line and Cell Phone Dual Frame**

One approach to this problem is to apply the existing respondent selection rules for line phone households to cell phone households as well. There may be some operational problems with this approach when the respondent selection technique is applied to the cell phone sample. If this approach was used then the estimate for totals becomes

$$\sum_{s1,s2} x_j /(\prod_l + \prod_c)$$

where

$k_{lj}$ = # of adults in a line phone household

$k_{cj}$ = # of adults in a cell phone household

$\prod_l = (l_j \pi_{lj} / k_{lj})$ if $x_j \in$ Line Frame and 0 otherwise

$\prod_c = (c_j \pi_{cj} / kc_j)$ if xj $\in$ Cell Frame and 0 otherwise

$k_{lj}$ and $k_{cj}$ refer to the number of adults in household j

And

$\sum_{s1,s2}$ is the sum over both samples

A more practical approach is to determine if the cell phone is a personal phone or a household phone and interview the user of the personal phone. Most cell phones will be personal phones. The status and number of line phones in the household is still needed.

**Using Multi-frame Telephone Sample Designs for Rare populations.**

Multi-frame designs can be an effective way to improve the efficiency of sample designs for rare populations. These designs can be used to link a national RDD frame with lower coverage specialty frames such as lists of phone numbers which have a better than average chance of being attached to the rare population. These designs require the use of nested frame designs. The term "nested" is used to indicate that all of the elements in one frame are contained in one or more of the other frames.

This is essentially the situation when one frame is a list of phone numbers with a higher than normal probability of being attached to the rare population, e.g., Hispanic households and the other frame is the usual RDD frame of all possible phone numbers. Such phone number lists' are commercially available for the Hispanic population of the U.S., and can also be constructed using linkages between geography and exchange prefixes.

**Nested Dual Frame designs**

In the nested dual frame situation there are only two domains. B is used to denote the domain made up of the overlapping part of the two frames, and A to denote the domain made up of the part of frame 1 which is not included in frame 2. There are now three sample components one is the sample from A, one sample is the part of the sample drawn from frame 1 that are also elements from frame 2, and the other sample is the sample drawn only from frame 2.

Using the nomenclature of the non-nested design previously discussed the estimate for the population total for the variable x can be written as,

$$X_{1A} + X_{12}$$

Where $X_{12} = (w_{1B} X_{1B} + w_{2B} X_{2B})$, and $w_{1B} + w_{2B} = 1$.

Again if we set

$$w_{1B} = (\pi_1 /(\pi_1 + \pi_2) \text{ and } w_{2B} = (\pi_2 /(\pi_1 + \pi_2) = 1$$

And use an H-T estimator the estimate for the population total becomes

$$X = \sum_{j=1}^{n_{1A}} x_{1Aj} / \pi_1 + \sum_{j=1}^{n_{1B}} x_{1Bj} /(\pi_1 + \pi_2) + \sum_{j=1}^{n_{2B}} x_{2Bj} /(\pi_1 + \pi_2)$$

or

$$\sum_{S_A,S_B} x_j /(\prod_1 + \prod_2)$$

where

$\prod_1 = \pi_1$ if $x_j \in$ Frame 1 and 0 otherwise

$\prod_2 = \pi_2$ if $x_j \in$ Frame 2 and 0 otherwise

And

$\sum_{s1,s2}$ is the sum over the sample from Frame 1 and Frame 2

The adjustments for multiple phone lines and respondent selection if used remain the same as in the previous discussion.

**Sampling Hispanic Populations in the USA**

This design just described is appropriate if the Hispanic sample uses an equal probability RDD design and an equal probability sample from a list of phone numbers with a greater than normal

chance of being attached to Hispanic households. An example of this would be a list of phone numbers of households containing persons with Hispanic names. We can improve on this design by using more than one source for phone numbers which have a high likelihood of being attached to Hispanic households. A brief description of such a design is now presented.

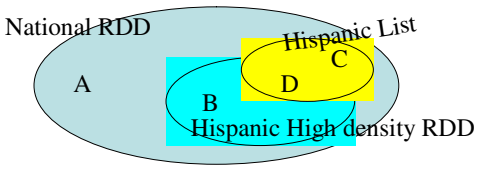
A Three Frame design for Sampling The U.S. Hispanic Population Via Telephone
This is a nested three frame design using unequal probabilities of selection within one of the RDD frames. The three frames are:

1. National RDD frame which is an Equal probability sample of US telephone numbers from 100 blocks with 1 or more listed phone numbers

2. A high density Hispanic RDD frame covering only high density Hispanic areas with three strata based on Hispanic Density. The three strata are High density, Medium density, and Low Hispanic density. Areas of very low Hispanic density 5% or less are not included in the frame.

3. A Hispanic list frame made up from a list of phone numbers with a higher than normal likelihood of being attached to Hispanic households. The phone numbers qualify for this list primarily on the basis of the ethnicity of the names of the persons believed to live in those households at the time the list is compiled.

The Composite Frame for This Hispanic Sample




As before estimates for universe totals are made by summing the estimates of the super-strata produced by the design.

**Accounting for Non-Response**
In practice of course there should be some adjustment for non-response. While there are different ways of making this adjustment
We prefer the use of a frame specific non-response adjustment for each element This means that the non-response adjustment for each completed element is based on the response rate performance of the frame from which it was selected. This recognizes that response rates are influenced both by the population characteristics and the characteristics of the calling facility and it's Interviewers

Finally all the other techniques used to improve single frame samples, such as ratio estimation, post-stratification, etc are applicable

**References**

**Hartley, H. O. (1962) : Multiple Frame Surveys, Proceedings of The Social Statistics Section, American Statistical Association, 203 ff.**