

Estimation in a dual frame context

A. Arcos, M. Rueda, M. G. Ranalli and D. Molina

September 16, 2024

Contents

| | | |
|----------|---|----------|
| 1 | Introduction | 1 |
| 2 | Data description | 2 |
| 3 | Estimation with no auxiliary information | 3 |
| 4 | Estimation using frame sizes as auxiliary information | 6 |
| 5 | Estimation using frame and overlap domain sizes as auxiliary information | 8 |
| 6 | Estimation using additional variables as auxiliary information | 8 |

1 Introduction

Classic sampling theory usually assumes the existence of one sampling frame containing all finite population units. Then, a probability sample is drawn according to a given sampling design and information collected is used for estimation and inference purposes. But in practice, the assumption that the sampling frame contains all population units is rarely met.

Dual frame sampling approach solves this issue by assuming that two frames are available for sampling and that, overall, they cover the entire target population. The most common situation is the one represented in Figure 1 where the two frames, say frame A and frame B , show a certain degree of overlapping, so it is possible to distinguish three disjoint non-empty domains: domain a , containing units belonging to frame A but not to frame B ; domain b , containing units belonging to frame B but not to frame A and domain ab , containing units belonging to both frames.

Then, independent samples s_A and s_B of size n_A and n_B are drawn from frame A and frame B and the information included is suitably combined to provide results.

This vignette shows the way package `Frames2` operates and their wide options to work with data coming from a dual frame context.

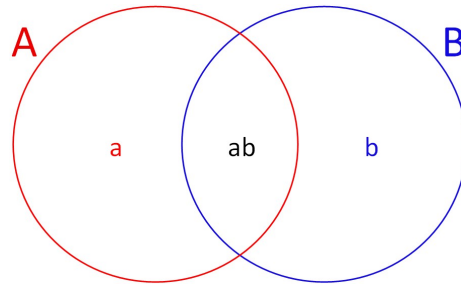


Figure 1: Two frames with overlapping.

2 Data description

To illustrate how functions of the package operate, we will use data sets *DataA* and *DataB* (included in the package) as sample data from frame *A* and frame *B*, respectively. *DataA* contains information about $n_A = 105$ individuals selected through a stratified random sampling design from the $N_A = 1735$ individuals composing frame *A*. Sample sizes by strata are $n_{hA} = (15, 20, 15, 20, 15, 20)$. On the other hand, a simple random without replacement sample of $n_B = 135$ individuals has been selected from the $N_B = 1191$ included in frame *B*. The size of the overlap domain for this case is $N_{ab} = 601$.

Let see the first three rows of each data set:

```
> library (Frames2)
> data(DataA)
> data(DataB)
> head (DataA, 3)
```

| | Domain | Feed | Clo | Lei | Inc | Tax | M2 | Size | ProbA | ProbB |
|---|--------|--------|-------|-------|---------|--------|--------|------|------------|-----------|
| 1 | a | 194.48 | 38.79 | 23.66 | 2452.07 | 112.90 | 0.00 | 0 | 0.02063274 | 0.0000000 |
| 2 | a | 250.23 | 16.92 | 22.68 | 2052.37 | 106.99 | 0.00 | 0 | 0.02063274 | 0.0000000 |
| 3 | ab | 199.95 | 24.50 | 23.24 | 2138.24 | 121.16 | 127.41 | 2 | 0.02063274 | 0.1133501 |

Stratum

| | |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |

```
> head (DataB, 3)
```

| | Domain | Feed | Clo | Lei | Inc | Tax | M2 | Size | ProbA | ProbB |
|---|--------|--------|-------|-------|---------|--------|--------|------|------------|-----------|
| 1 | ba | 332.42 | 38.42 | 21.12 | 3109.75 | 148.07 | 186.46 | 3 | 0.02063274 | 0.1133501 |
| 2 | b | 222.47 | 19.94 | 19.74 | 0.00 | 0.00 | 126.79 | 2 | 0.00000000 | 0.1133501 |
| 3 | b | 215.43 | 35.13 | 20.17 | 0.00 | 0.00 | 148.67 | 3 | 0.00000000 | 0.1133501 |

Each data set incorporates information about three main variables: Feeding, Clothing and Leisure. Additionally, there are two auxiliary variables for the

units in frame A (Income and Taxes) and another two variables for units in frame B (Metres2 and Size). Corresponding totals for these auxiliary variables are assumed known in the entire frame and they are $T_{Inc}^A = 4300260$, $T_{Tax}^A = 215577$, $T_{M2}^B = 176553$ and $T_{Size}^B = 3529$. Finally, a variable indicating the domain each unit belongs to and two variables showing the first order inclusion probabilities for each frame complete the data sets.

Numerical square matrices $PiklA$ and $PiklB$ (also included in the package), with dimensions $n_A = 105$ and $n_B = 135$, are used as probability inclusion matrices. These matrices contains second order inclusion probabilities and first order inclusion probabilities as diagonal elements.

3 Estimation with no auxiliary information

When there is no further information than the one on the variables of interest, one can calculate some estimators, as Hartley (1962, 1974) or Fuller-Burmeister (1972) estimators

```
> data(PiklA)
> data(PiklB)
> yA <- with(DatA, data.frame(Feed, Clo, Lei))
> yB <- with(DatB, data.frame(Feed, Clo, Lei))
> Hartley(yA, yB, PiklA, PiklB, Data$Domain, DatB$Domain)
```

```
Estimation:
      Feed      Clo      Lei
Total 586959.9820 71967.62214 53259.86947
Mean   246.0429   30.16751   22.32556
```

```
> FB(yA, yB, PiklA, PiklB, Data$Domain, DatB$Domain)
```

```
Estimation:
      Feed      Clo      Lei
Total 591665.5078 72064.99223 53034.09810
Mean   248.0153   30.20832   22.23092
```

Results show, by default, estimations for the population total and mean for the considered variables. If only first order inclusion probabilities are available, estimates can also be computed

```
> Hartley(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain)
```

```
Estimation:
      Feed      Clo      Lei
Total 570867.8042 69473.86532 51284.2727
Mean   247.9484   30.17499   22.2746
```

```
> FB(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 571971.9511 | 69500.11448 | 51210.03819 |
| Mean | 248.4279 | 30.18639 | 22.24236 |

Further information about estimation process (as variance estimations or values of parameters involved in estimation, if any) can be displayed by using function `summary`

```
> summary(Hartley(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain,
+              DatB$Domain))
```

Call:

```
Hartley(ysA = yA, ysB = yB, pi_A = Data$ProbA, pi_B = DatB$ProbB,
        domains_A = Data$Domain, domains_B = DatB$Domain)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|------------|
| Total | 570867.8042 | 69473.86532 | 51284.2727 |
| Mean | 247.9484 | 30.17499 | 22.2746 |

Variance Estimation:

| | Feed | Clo | Lei |
|------------|--------------|--------------|--------------|
| Var. Total | 9.050344e+08 | 1.550443e+07 | 6.977928e+06 |
| Var. Mean | 1.707326e+02 | 2.924874e+00 | 1.316370e+00 |

Total Domain Estimations:

| | Feed | Clo | Lei |
|---------------|----------|----------|----------|
| Total dom. a | 263233.1 | 31476.84 | 22839.95 |
| Total dom. ab | 166651.7 | 21494.96 | 15984.64 |
| Total dom. b | 164559.2 | 20451.85 | 15693.59 |
| Total dom. ba | 128704.7 | 15547.49 | 11112.38 |

Mean Domain Estimations:

| | Feed | Clo | Lei |
|--------------|----------|----------|----------|
| Mean dom. a | 251.8133 | 30.11129 | 21.84909 |
| Mean dom. ab | 241.6468 | 31.16792 | 23.17791 |
| Mean dom. b | 242.2443 | 30.10675 | 23.10221 |
| Mean dom. ba | 251.5291 | 30.38466 | 21.71707 |

Parameters:

| | Feed | Clo | Lei |
|-------|-----------|-----------|-----------|
| theta | 0.3787075 | 0.3358878 | 0.3362615 |

Results slightly change when a confidence interval is required. In that case, user has to indicate the confidence level desired for the interval through argument `conf_level` (default is `NULL`) and add it to the list of input parameters.

In this case, default output will show 6 rows for each variable, lower and upper boundaries for confidence intervals are displayed together with estimates. So, one can obtain a 95% confidence interval for estimations computed using Hartley and Fuller-Burmeister estimators in this way

```
> Hartley(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain,
+         DatB$Domain, 0.95)
```

Estimation and 95 % Confidence Intervals:

| | Feed | Clo | Lei |
|-------------|-------------|-------------|-------------|
| Total | 570867.8042 | 69473.86532 | 51284.27265 |
| Lower Bound | 511904.6588 | 61756.37677 | 46106.87729 |
| Upper Bound | 629830.9496 | 77191.35387 | 56461.66802 |
| Mean | 247.9484 | 30.17499 | 22.27460 |
| Lower Bound | 222.3386 | 26.82301 | 20.02587 |
| Upper Bound | 273.5582 | 33.52697 | 24.52333 |

```
> FB(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain,
+    0.95)
```

Estimation and 95 % Confidence Intervals:

| | Feed | Clo | Lei |
|-------------|-------------|-------------|-------------|
| Total | 571971.9511 | 69500.11448 | 51210.03819 |
| Lower Bound | 513045.7170 | 61802.57411 | 46036.74627 |
| Upper Bound | 630898.1852 | 77197.65484 | 56383.33011 |
| Mean | 248.4279 | 30.18639 | 22.24236 |
| Lower Bound | 222.8342 | 26.84307 | 19.99541 |
| Upper Bound | 274.0217 | 33.52971 | 24.48930 |

When, for units included in overlap domain, first order inclusion probabilities are known for both frames, estimators as the one proposed by Bankier (1986), Kalton and Anderson (1986) can be computed. To do this, numeric vectors `pik_ab_B` and `pik_ba_A` of lengths n_A and n_B should be added as arguments. While `pik_ab_B` represents first order inclusion probabilities according to sampling design in frame B for units belonging to overlap domain selected in sample drawn from frame A , `pik_ba_A` contains first order inclusion probabilities according to sampling design in frame A for units belonging to overlap domain selected in sample drawn from frame B .

```
> BKA(yA, yB, Data$ProbA, DatB$ProbB, Data$ProbB, DatB$ProbA,
+     Data$Domain, DatB$Domain)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 566434.3200 | 68959.26705 | 50953.07583 |
| Mean | 247.8845 | 30.17814 | 22.29822 |

These examples include just a few of the estimators that can be used when no auxiliary information is known. Other estimators, as the pseudo-empirical likelihood estimator (Rao and Wu, 2010) or the dual frame calibration estimator (Ranalli et al., 2014), can be also calculated in this case. In this context, function `Compare` is quite useful, since it returns all possible estimators that can be computed according to the information provided as input

```
> Compare(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain)
```

```
$Hartley
```

```
Estimation:
```

| | Feed | Clo | Lei |
|-------|-------------|-------------|------------|
| Total | 570867.8042 | 69473.86532 | 51284.2727 |
| Mean | 247.9484 | 30.17499 | 22.2746 |

```
$FullerBurmeister
```

```
Estimation:
```

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 571971.9511 | 69500.11448 | 51210.03819 |
| Mean | 248.4279 | 30.18639 | 22.24236 |

```
$PEL
```

```
Estimation:
```

| | Feed | Clo | Lei |
|-------|-------------|------------|-------------|
| Total | 591956.1900 | 72391.7894 | 53396.32780 |
| Mean | 247.5017 | 30.2676 | 22.32544 |

```
$Calibration_DF
```

```
Estimation:
```

| | Feed | Clo | Lei |
|-------|-------------|-------------|------------|
| Total | 595162.2604 | 72214.13351 | 53108.5059 |
| Mean | 248.8422 | 30.19332 | 22.2051 |

4 Estimation using frame sizes as auxiliary information

Some of the estimators defined for dual frame data, as raking ratio (Skinner, 1991) or pseudo-maximum likelihood estimators (Skinner and Rao, 1996), require the knowledge of frame sizes to provide results. So, frame sizes need to be incorporated to the function through two additional input arguments, `N_A` and `N_B`. There is also a group of estimators, including pseudo-empirical likelihood

and calibration estimators, that even being able to provide estimations without the need of auxiliary information, can use frame sizes to improve their precision

```
> SFRR(yA, yB, Data$ProbA, DatB$ProbB, Data$ProbB, DatB$ProbB,
+       Data$Domain, DatB$Domain, N_A = 1735, N_B = 1191)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|------------|-------------|
| Total | 596147.5461 | 72584.5907 | 53527.39414 |
| Mean | 248.1584 | 30.2148 | 22.28185 |

```
> CalSF(yA, yB, Data$ProbA, DatB$ProbB, Data$ProbB, DatB$ProbB,
+        Data$Domain, DatB$Domain, N_A = 1735, N_B = 1191)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 595996.8469 | 72566.50183 | 53513.52578 |
| Mean | 248.1587 | 30.21495 | 22.28174 |

Previous estimators need probabilities of inclusion in both frames for the units in the overlap domain to be computed. This condition may be restrictive in some cases. As an alternative, in cases where frame sizes are known but this condition is not met, it is possible to calculate dual frame estimators as pseudo-maximum likelihood, pseudo-empirical likelihood and dual frame calibration estimators

```
> PML(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain,
+      N_A = 1735, N_B = 1191)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 594400.6320 | 72430.05834 | 53408.30337 |
| Mean | 248.0934 | 30.23115 | 22.29178 |

```
> PEL(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain,
+      N_A = 1735, N_B = 1191)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|------------|-------------|
| Total | 591956.1900 | 72391.7894 | 53396.32780 |
| Mean | 247.5017 | 30.2676 | 22.32544 |

```
> CalDF(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain,
+        N_A = 1735, N_B = 1191)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 588416.4644 | 71432.53671 | 52520.31623 |
| Mean | 248.8131 | 30.20539 | 22.20832 |

5 Estimation using frame and overlap domain sizes as auxiliary information

In addition to the frame sizes, in some cases, it is possible to know the size of the overlap domain, N_{ab} . Generally, this highly improves the precision of the estimates. Functions implementing pseudo-empirical likelihood and calibration estimators can incorporate overlap domain size to the estimation procedure through parameter N_{ab} , as shown below

```
> PEL(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain,
+     N_A = 1735, N_B = 1191, N_ab = 601)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 575289.2187 | 70429.95641 | 51894.32490 |
| Mean | 247.4362 | 30.29245 | 22.32014 |

```
> CalSF(yA, yB, Data$ProbA, DatB$ProbB, Data$ProbB, DatB$ProbB,
+       Data$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+       N_ab = 601)
```

Estimation:

| | Feed | Clo | Lei |
|-------|------------|-------------|------------|
| Total | 577071.959 | 70294.89095 | 51771.9309 |
| Mean | 248.203 | 30.23436 | 22.2675 |

```
> CalDF(yA, yB, Data$ProbA, DatB$ProbB, Data$Domain, DatB$Domain,
+       N_A = 1735, N_B = 1191, N_ab = 601)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|------------|-------------|
| Total | 578895.6961 | 70230.1131 | 51570.55683 |
| Mean | 248.9874 | 30.2065 | 22.18088 |

6 Estimation using additional variables as auxiliary information

Some of the estimators are defined such that, in addition to frame sizes, they can incorporate auxiliary information about extra variables to the estimation process. This is the case of pseudo-empirical likelihood and calibration estimators. Functions implementing them are also able to manage auxiliary information. To achieve maximum flexibility, these functions are prepared to deal with auxiliary information when it is available only in frame A , only in frame B or in both frames.

For instance, auxiliary information collected from frame A should be incorporated to functions through three arguments: `xsAFrameA` and `xsBFrameA`,

numeric vectors, matrices or data frames (depending on the number of auxiliary variables in the frame); and `XA`, a numeric value or vector of length indicating population totals for the auxiliary variables considered in frame *A*. Similarly, auxiliary information in frame *B* is incorporated to each function through arguments `xsAFrameB`, `xsBFrameB` and `XB`. If auxiliary information is available in the whole population, it must be indicated through parameters `xsT` and `X`. Let see some examples

```
> PEL(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735,
+     N_B = 1191, xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc,
+     XA = 4300260)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 588917.7336 | 72077.37877 | 53263.75154 |
| Mean | 246.8638 | 30.21355 | 22.32722 |

```
> CalSF(yA, yB, PiklA, PiklB, DatA$ProbB, DatB$ProbA, DatA$Domain,
+       DatB$Domain, N_A = 1735, N_B = 1191, xsAFrameB = DatA$M2,
+       xsBFrameB = DatB$M2, XB = 176553)
```

Estimation:

| | Feed | Clo | Lei |
|-------|------------|-------------|-------------|
| Total | 581539.671 | 70735.99535 | 52208.48996 |
| Mean | 247.159 | 30.06336 | 22.18902 |

```
> CalDF(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735,
+       N_B = 1191, xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc,
+       xsAFrameB = DatA$M2, xsBFrameB = DatB$M2, XA = 4300260,
+       XB = 176553)
```

Estimation:

| | Feed | Clo | Lei |
|-------|-------------|-------------|-------------|
| Total | 585185.4497 | 71194.61148 | 52346.43878 |
| Mean | 247.8075 | 30.14866 | 22.16705 |

While pseudo-empirical likelihood estimator has been computed considering only auxiliary information in frame *A*, single frame calibration estimator has been calculated considering auxiliary information in frame *B*. For the dual frame calibration estimator, auxiliary information in both frames has been taken into account.

References

- [1] Arcos, A., Molina, D., Rueda, M. and Ranalli, M. G. (2015). *Frames2: A Package for Estimation in Dual Frame Surveys*. The R Journal, 7(1), 52 - 72.

- [2] Bankier, M.D. (1986). *Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys*. Journal of the American Statistical Association, Vol. 81, 1074 - 1079.
- [3] Fuller, W.A. and Burmeister, L.F. (1972). *Estimation for Samples Selected From Two Overlapping Frames* ASA Proceedings of the Social Statistics Sections, 245 - 249.
- [4] Hartley, H.O. (1962). *Multiple Frame Surveys*. Proceedings of the American Statistical Association, Social Statistics Sections, 203 - 206.
- [5] Hartley, H.O. (1974). *Multiple frame methodology and selected applications*. Sankhya C., Vol. 36, 99 - 118.
- [6] Kalton, G. and Anderson, D.W. (1986) *Sampling Rare Populations*. Journal of the Royal Statistical Society, Ser. A, Vol. 149, 65 - 82.
- [7] Ranalli, M.G., Arcos, A., Rueda, M. and Teodoro, A. (2014). *Calibration estimators in dual frames surveys*. arXiv:1312.0761 [stat.ME].
- [8] Rao, J. N. K. and Wu, C. (2010) *Pseudo Empirical Likelihood Inference for Multiple Frame Surveys*. Journal of the American Statistical Association, Vol. 105, 1494 - 1503.
- [9] Skinner, C.J. (1991). *On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys*. Journal of the American Statistical Association, Vol. 86, 779 - 784.
- [10] Skinner, C. J. and Rao, J. N. K. (1996). *Estimation in Dual Frame Surveys With Complex Designs*. Journal of the American Statistical Association, Vol. 91, 349 - 356.