

# Splitting and formatting data in a dual frame context

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

October 19, 2015

## Contents

<b>1</b>	<b>Data description</b>	<b>1</b>
<b>2</b>	<b>Formatting data</b>	<b>2</b>

## 1 Data description

To illustrate how to split and format a file including the information collected from a dual frame survey we will use data set *Dat* (included in the package). This data set includes some of the variables collected in a real dual frame opinion survey about immigration. This survey was conducted using telephone interviews using two sampling frames: one for landlines and another one for cell phones. From the landline frame, a stratified sample of size 1919 was drawn, while from the cell phone frame, a sample of size 483 was drawn using simple random sampling without replacement. Variables included in the data set are: **Drawnby**, which takes value 1 if the unit comes from the landline sample and value 2 if it comes from the cell phone sample; **Stratum**, which indicates the stratum each unit belongs to (for individuals in cell phone frame, value of this variable is NA); **Opinion** the response to the opinion question with value 1 representing a favorable opinion about immigration and value 0 representing a unfavorable opinion about immigration; **Landline** and **Cell**, which record whether the unit possess a landline or a cell phone, respectively. First order inclusion probabilities are also included in the data set.

Let see the first three rows of the data set:

```
> library (Frames2)
> data(Dat)
> head(Dat, 3)
```

	Drawnby	Stratum	Opinion	Landline	Cell	ProbLandline	ProbCell	Income
1	1	2	0	1	1	0.000673623	8.49e-05	1629.31
2	1	5	1	1	1	0.002193297	5.86e-05	2084.03
3	1	1	0	1	1	0.001831489	7.81e-05	1718.65

## 2 Formatting data

From the data of this survey we wish to estimate the number of people with a favorable opinion regarding immigration. In order to use functions of Frames2, we need to split this dataset. The variables we will use to do this are `Drawnby` and `Landline` and `Cell`. First step is to split the original data set in four new different data sets, each one corresponding to one domain.

```
> attach(Dat)
> DomainOnlyLandline <- Dat[Landline == 1 & Cell == 0,]
> DomainBothLandline <- Dat[Drawnby == 1 & Landline == 1 &
+                           Cell == 1,]
> DomainOnlyCell <- Dat[Landline == 0 & Cell == 1,]
> DomainBothCell <- Dat[Drawnby == 2 & Landline == 1 &
+                       Cell == 1,]
```

Then, from the domain datasets, we can easily build frame datasets

```
> FrameLandline <- rbind(DomainOnlyLandline, DomainBothLandline)
> FrameCell <- rbind(DomainOnlyCell, DomainBothCell)
```

Finally, we only need to label domain of each unit using "a", "b", "ab" or "ba"

```
> Domain <- c(rep("a", nrow(DomainOnlyLandline)), rep("ab",
+               nrow(DomainBothLandline)))
> FrameLandline <- cbind(FrameLandline, Domain)
> Domain <- c(rep("b", nrow(DomainOnlyCell)), rep("ba",
+               nrow(DomainBothCell)))
> FrameCell <- cbind(FrameCell, Domain)
```

Now dual frame estimators, as Hartley (1962, 1974) estimator, can be computed:

```
> Hartley(FrameLandline$Opinion, FrameCell$Opinion,
+          FrameLandline$ProbLandline, FrameCell$ProbCell,
+          FrameLandline$Domain, FrameCell$Domain)
```

Estimation:

```
      [,1]
Total 3.46686e+06
Mean  4.93861e-01
```

## References

- [1] Hartley, H.O. (1962). *Multiple Frame Surveys*. Proceedings of the American Statistical Association, Social Statistics Sections, 203 - 206.
- [2] Hartley, H.O. (1974). *Multiple frame methodology and selected applications*. Sankhya C., Vol. 36, 99 - 118.