# Examples of two-stage surveys

Thomas Lumley

March 24, 2005

Starting with version 2.9 the `survey` package can correctly analyse multistage samples without replacement. This is relatively rarely useful, as it requires more information than is often available and makes a substantial difference only when the first-stage sampling fraction is fairly large. As far as I am aware, SUDAAN is the only major commercial package that does these analyses, so there are fewer opportunities to validate results.

This document reproduces examples from two textbooks: *Sampling of Populations* by Levy and Lemeshow, and *Model Assisted Survey Sampling* by Särndal, Swensson, and Wretman.

```
> library(survey)
> library(foreign)
```

## Example 1

This is example 4.3.2 from Särndal, Swensson, and Wretman. They use a sample from the *MU284* population of municipalities. In the first stage of sampling 5 clusters were sampled from 50 in the population. In the second stage, a sample of 3 municipalities was taken from each cluster. The data are `data(mu284)` in the `survey` package.

First we define the design. We specify identifiers for each stage of clustering: `id1` identifiers a cluster and `id2` identifies a municipality. The `fpc` argument gives the population sizes, where `n1` is 50 and `n2` varies between clusters. There is no need to provide separate sampling weights, since these can be computed from the population sizes. (Separate sampling weights are still allowed, eg to handle non-response)

```
> data(mu284)
> dmu284 <- svydesign(id = ~id1 + id2, fpc = ~n1 + n2, data = mu284)
> dmu284

2 - level Cluster Sampling design
With (5, 15) clusters.
svydesign(id = ~id1 + id2, fpc = ~n1 + n2, data = mu284)
```

Note that the design object shows the number of units at each level of sampling.

Särndal et al estimate the total for a variable they call S82 and we call `y1`. They obtain an estimated population total of 15080 with a variance of 5172234.

```
> (ytotal <- svytotal(~y1, dmu284))

   total     SE
y1 15080 2274.3

> vcov(ytotal)

        [,1]
[1,] 5172234
```

## Example 2

Chapter 10 of Levy and Lemeshow gives both explicit calculations and the results of a number of packaged analyses for two examples. The explicit calculations are for a single-stage (ultimate cluster) analysis, but one that correctly accounts for finite population size.

The first example is a two-stage sample where all the first-stage clusters are the same size. The data are on page 284 and the results on page 287. All these analyses use the same probability weights, so they all obtain the same point estimates.

First we do an analysis for the correct design, two-stage sampling without replacemnt

```
> p284 <- read.table("LLp284.dat", header = TRUE)
> (dp284 <- svydesign(id = ~center + nurse, weight = ~w, fpc = ~m +
+     nbar, data = p284))

2 - level Cluster Sampling design
With (3, 6) clusters.
svydesign(id = ~center + nurse, weight = ~w, fpc = ~m + nbar,
    data = p284)

> svytotal(~nrefrred, dp284)

         total     SE
nrefrred    90 21.680

> svyratio(~nrefrred, ~npatnts, dp284)

Ratio estimator: svyratio.survey.design2(~nrefrred, ~npatnts, dp284)
Ratios=
           npatnts
nrefrred 0.2222222
SEs=
            npatnts
nrefrred 0.05837242
```

SUDAAN gives 21.68 and 0.058, as does the survey package. The explicit calculations give standard errors of 23.48 for the total and 0.0612 for the ratio, with the difference being due to the difference between a one-stage and two-stage analysis.

For an analysis pretending the data were sampled with replacement, SUDAAN and Stata, and the explicit calculations, give standard errors of 30.31 and 0.081. The survey package agrees. This analysis can be obtained by omitting the finite population (`fpc`) argument, in which case specifying the second stage of sampling becomes optional.

```
> (wrdp284 <- svydesign(id = ~center + nurse, weight = ~w, data = p284))

2 - level Cluster Sampling design
With (3, 6) clusters.
svydesign(id = ~center + nurse, weight = ~w, data = p284)

> svytotal(~nrefrred, wrdp284)

         total     SE
nrefrred    90 30.311

> svyratio(~nrefrred, ~npatnts, wrdp284)

Ratio estimator: svyratio.survey.design2(~nrefrred, ~npatnts, wrdp284)
Ratios=
           npatnts
nrefrred 0.2222222
SEs=
            npatnts
nrefrred 0.08127789

> (udp284 <- svydesign(id = ~center, weight = ~w, data = p284))

1 - level Cluster Sampling design
With (3) clusters.
svydesign(id = ~center, weight = ~w, data = p284)

> svytotal(~nrefrred, udp284)

         total     SE
nrefrred    90 30.311

> svyratio(~nrefrred, ~npatnts, udp284)

Ratio estimator: svyratio.survey.design2(~nrefrred, ~npatnts, udp284)
Ratios=
           npatnts
nrefrred 0.2222222
SEs=
            npatnts
nrefrred 0.08127789
```

**Example 3** In the data presented in table 10.10 of Levy and Lemeshow the first-stage clusters vary in size. Again, the book presents results from Stata and SUDAAN, and explicit computations using a one-stage analysis that correctly accounts for finite population size.

```
> hosp <- read.table("lltab10-10.dat", header = TRUE)
> (hdes <- svydesign(id = ~hospital + id, fpc = ~nhosp + admit,
+     data = hosp))

2 - level Cluster Sampling design
With (3, 708) clusters.
svydesign(id = ~hospital + id, fpc = ~nhosp + admit, data = hosp)

> svytotal(~dead, hdes)

      total     SE
dead 499.38 116.07

> svyratio(~dead, ~lifethrt, hdes)

Ratio estimator: svyratio.survey.design2(~dead, ~lifethrt, hdes)
Ratios=
      lifethrt
dead 0.1703017
SEs=
       lifethrt
dead 0.06389816
```

SUDAAN gives standard errors of 116.07 and 0.064, which the survey package duplicates. The explicit calculations give 114.65 and 0.072.

In Stata these data would be analysed as if they were a one-stage sample without replacement

```
> hosp$weight <- (10/3) * hosp$admit/hosp$admsamp
> (honestage <- svydesign(id = ~hospital, weight = ~weight, data = hosp))

1 - level Cluster Sampling design
With (3) clusters.
svydesign(id = ~hospital, weight = ~weight, data = hosp)

> svytotal(~dead, honestage)

      total     SE
dead 499.38 114.68

> svyratio(~dead, ~lifethrt, honestage)
```

```
Ratio estimator: svyratio.survey.design2(~dead, ~lifethrt, honestage)
Ratios=
      lifethrt
dead 0.1703017
SEs=
       lifethrt
dead 0.07227302
```

For this with-replacement design, Stata, SUDAAN, and the survey package all give standard errors of 114.68 and 0.0072.