

Comparing the Horvitz-Thompson estimator and the Hajek estimator

June 25, 2015

Consider a finite population $U = \{1, 2, \dots, N\}$. Suppose $y_k, k \in U$ are values of the variable of interest in the population. We wish to estimate the total $\sum_{k=1}^N y_k$ based on a sample s taken from the population U . Assume that the sample is taken according to a sampling scheme having inclusion probabilities $\pi_k = Pr(k \in s)$. When the π_k is proportional to a positive quantity x_k available over U , and s has a predetermined sample size n , then

$$\pi_k = \frac{nx_k}{\sum_{i=1}^N x_i},$$

and the sampling scheme is said to be probability proportional to size (πps). Under this scheme, the Hájek estimator of the population total is defined by

$$\hat{y}_{Hajek} = N \frac{\sum_{k \in s} y_k / \pi_k}{\sum_{k \in s} 1 / \pi_k}.$$

Särndal, Swenson, and Wretman (1992, p. 182) give several cases for regarding the Hájek as ‘usually the better estimator’ comparing to the Horvitz-Thompson estimator

$$\hat{y}_{HT} = \sum_{k \in s} y_k / \pi_k :$$

- a) the $y_k - \bar{y}_U$ tend to be small,
- b) sample size is not fixed,
- c) π_k are weakly or negatively correlated with the y_k .

Monte Carlo simulations are used here to compare the accuracy of both estimators for a sample size (or expected value of the sample size) equal to 20. Four cases are considered:

- Case 1. y_k is constant for $k = 1, \dots, N$; this case corresponds to the case a) above;
- Case 2. Poisson sampling is used to draw a sample s ; this case corresponds to the case b) above;
- Case 3. y_k are generated using the following model: $x_k = k, \pi_k = nx_k / \sum_{i=1}^N x_i, y_k = 1 / \pi_k$; this case corresponds to the case c) above;

Case 4. y_k are generated using the following model: $x_k = k, y_k = 5(x_k + \epsilon_k), \epsilon_k \sim N(0, 1/3)$; in this case the Horvitz-Thompson estimator should perform better than the Hájek estimator.

Tillé sampling is used in Cases 1, 3 and 4. Poisson sampling is used in Case 2. The `belgianmunicipalities` dataset is used in Cases 1 and 2 with $x_k = Tot04_k$. In Case 2, the variable of interest is `TaxableIncome`. The mean square error (MSE) is computed using simulations for each case and estimator. The Hájek estimator should perform better than the Horvitz-Thompson estimator in Cases 1, 2 and 3.

```
> data(belgianmunicipalities)
> attach(belgianmunicipalities)
> # sample size
> n=20
> pik=inclusionprobabilities(Tot04,n)
> N=length(pik)
>
>
```

Number of simulations (for an accurate result, increase this value to 10000):

```
> sim=10
> ss=ss1=array(0,c(sim,4))
>
```

Defines the variables of interest:

```
> cat("Case 1\n")
> y1=rep(3,N)
> cat("Case 2\n")
> y2=TaxableIncome
> cat("Case 3\n")
> x=1:N
> pik3=inclusionprobabilities(x,n)
> y3=1/pik3
> cat("Case 4\n")
> epsilon=rnorm(N,0,sqrt(1/3))
> pik4=pik3
> y4=5*(x+epsilon)
>
```

Simulation and computation of the Horvitz-Thompson estimator and Hájek estimator:

```
> ht=numeric(4)
> hajek=numeric(4)
> for(i in 1:sim)
+ {
```

```

+ cat("Simulation ",i,"\n")
+ cat("Case 1\n")
+ s=UPtille(pik)
+ ht[1]=HTestimator(y1[s==1],pik[s==1])
+ hajek[1]=Hajkeestimator(y1[s==1],pik[s==1],N,type="total")
+ cat("Case 2\n")
+ s1=UPpoisson(pik)
+ ht[2]=HTestimator(y2[s1==1],pik[s1==1])
+ hajek[2]=Hajkeestimator(y2[s1==1],pik[s1==1],N,type="total")
+ cat("Case 3\n")
+ ht[3]=HTestimator(y3[s==1],pik3[s==1])
+ hajek[3]=Hajkeestimator(y3[s==1],pik3[s==1],N,type="total")
+ cat("Case 4\n")
+ ht[4]=HTestimator(y4[s==1],pik4[s==1])
+ hajek[4]=Hajkeestimator(y4[s==1],pik4[s==1],N,type="total")
+ ss[i,]=ht
+ ss1[i,]=hajek
+ }
>
>

```

Computation of the MSE and the ratio $\frac{MSE_{HT}}{MSE_{Hajek}}$:

```

> #true values
> tv=c(sum(y1),sum(y2),sum(y3),sum(y4))
> for(i in 1:4)
+ {
+ cat("Case ",i,"\n")
+ cat("The mean of the Horvitz-Thompson estimators:",mean(ss[,i])," and the true value:",tv[i],
+ MSE1=var(ss[,i])+(mean(ss[,i])-tv[i])^2
+ cat("MSE Horvitz-Thompson estimator:",MSE1,"\n")
+ cat("The mean of the Hajek estimators:",mean(ss1[,i])," and the true value:",tv[i],"\n")
+ MSE2=var(ss1[,i])+(mean(ss1[,i])-tv[i])^2
+ cat("MSE Hajek estimator:",MSE2,"\n")
+ cat("Ratio of the two MSE:", MSE1/MSE2,"\n")
+ }

```

Case 1

```

The mean of the Horvitz-Thompson estimators: 1705.734 and the true value: 1767
MSE Horvitz-Thompson estimator: 177685.5
The mean of the Hajek estimators: 1767 and the true value: 1767
MSE Hajek estimator: 3.446586e-26
Ratio of the two MSE: 5.155406e+30

```

Case 2

```

The mean of the Horvitz-Thompson estimators: 124993652699 and the true value: 121128481686
MSE Horvitz-Thompson estimator: 1.008334e+21
The mean of the Hajek estimators: 118180918507 and the true value: 121128481686
MSE Hajek estimator: 4.691477e+20
Ratio of the two MSE: 2.149289

```

Case 3

The mean of the Horvitz-Thompson estimators: 21582116 and the true value: 60436.25

MSE Horvitz-Thompson estimator: 1.21279e+15

The mean of the Hajek estimators: 1561012 and the true value: 60436.25

MSE Hajek estimator: 3.689927e+12

Ratio of the two MSE: 328.6759

Case 4

The mean of the Horvitz-Thompson estimators: 862928.3 and the true value: 868797.8

MSE Horvitz-Thompson estimator: 118447606

The mean of the Hajek estimators: 167945 and the true value: 868797.8

MSE Hajek estimator: 515200314326

Ratio of the two MSE: 0.0002299059

>