# Computational example with direct estimation under planned and unplanned domain structures

We demonstrate with real data the direct Horvitz-Thompson and Hájek estimation of totals for domains. The data set contains disposable income of households in $D = 12$ regions of Western Finland. The population consists of $N = 431{,}000$ households. In addition to the income data, the record of a household shows the number of household members who had higher education (variable EDUC) and the number of months in total the household members were employed (EMP) during last year. EDUC and EMP will be used later in GREG estimation. All three variables were determined using administrative registers maintained by Statistics Finland. Thus, for this computational exercise we had access to population level information also for our y-variable of interest (income). This gives a possibility to compare sample estimates to the known population values.

We estimate the yearly total disposable income $t_d = \sum_{k \in U_d} y_k$ in regions (domains) $U_d$, $d = 1,...,12$.

A sample of 1,000 households was drawn from the population by using stratified πPS (without-replacement type probability proportional to size sampling) with household size as the size variable. To demonstrate estimation for planned domains, we interpret here the sample as a stratified sample where the regions constitute the strata and a separate sample is drawn from each stratum. The domain structure is of planned type, where the regional sample sizes are considered fixed by the sampling design. For HT we also consider the sample as a single sample drawn from the population.

In Table 1, we grouped the domains by sample size into minor ($8 \le n_d \le 33$), medium-sized ($34 \le n_d \le 45$) and major ($46 \le n_d \le 277$) domains, where $n_d$ is the observed domain sample size in domain $U_d$. There were four domains in each domain size class.

Results are in Table 1. *Absolute relative error* ARE of an estimator in domain *d* is calculated as $\left|\hat{t}_d - t_d\right| / t_d$ and MARE in a domain group is the mean of absolute relative errors over domains in the group. MCV is the *mean coefficient of variation* of the estimate, over domain group. Coefficient of variation is calculated as $\text{s.e}(\hat{t}_d) / \hat{t}_d$, where s.e refers to the estimated standard error of an estimator.

HT estimated by (1) and for Hájek by (6). For variance estimation we approximated the design by with-replacement type PPS. Variance estimators for HT were defined by (4) for planned domains and by (5) for unplanned domains (using *extended domain variables* $y_{dk} = 1$ if $k \in s_d$, 0 otherwise). For Hájek estimator in planned domains case the variance is estimated by (7).

Hájek estimator, which contains known domain sizes $N_d$, yielded better results than HT. Note that MCV of HT is much larger in the unplanned domains case than for planned domains. In large domains, MCV figures were usually smaller than in small domains.

NOTE: Example is adapted from Lehtonen and Veijanen (2009) Section 3.5.

**Table 2**. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and Hájek estimators of totals for minor, medium-sized and major domains for planned domains (HT and Hájek) and unplanned domains (HT).

| Domain sample size class | HT | | | Hájek | |
|---|---|---|---|---|---|
| | Auxiliary information | | | | |
| | None | | | Domain sizes | |
| | MARE % | MCV1 % | MCV2 % | MARE % | MCV1 % |
| Minor $8 \le n_d \le 33$ | 11.5 | 11.9 | 28.3 | 5.3 | 10.9 |
| Medium $34 \le n_d \le 45$ | 7.6 | 9.0 | 20.3 | 6.4 | 9.0 |
| Major $46 \le n_d \le 277$ | 12.5 | 5.2 | 9.6 | 4.7 | 5.6 |
| MCV1: Assuming planned domains for HT and Hájek<br>MCV2: Assuming unplanned domains for HT | | | | | |

Reference

Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis*. Vol. 29B. New York: Elsevier.