european conference
on quality in
official statistics

Session on Statistical Literacy

JOHANNES KEPLER
UNIVERSITY LINZ

06-04-2014, Schönbrunn, Vienna

Statistical Education: The Teaching Concept of Pseudo-Populations

Andreas Quatember


1. The Horvitz-Thompson Estimator
2. The Ratio Estimator
3. Weighting Adjustment & Data Imputation
4. Examples of Further Applications

1. The Horvitz-Thompson Estimator

The concept of pseudo-populations helps us to teach the concepts of sampling theory:

The *Horvitz-Thompson* (HT) *estimator* of the total

$$t = \sum\nolimits_U y_k$$

of a study variable y in a population U of size N is given by

$$t_{HT} = \sum\nolimits_s y_k \cdot \frac{1}{\pi_k} = \sum\nolimits_s y_k \cdot d_k$$

with design weights d_k and first-order sample inclusion probabilities π_k.

The idea behind the expression of the HT estimator can be described by the generation of a pseudo-population U_{HT}^{*} as a set-valued estimator of U with respect to y:

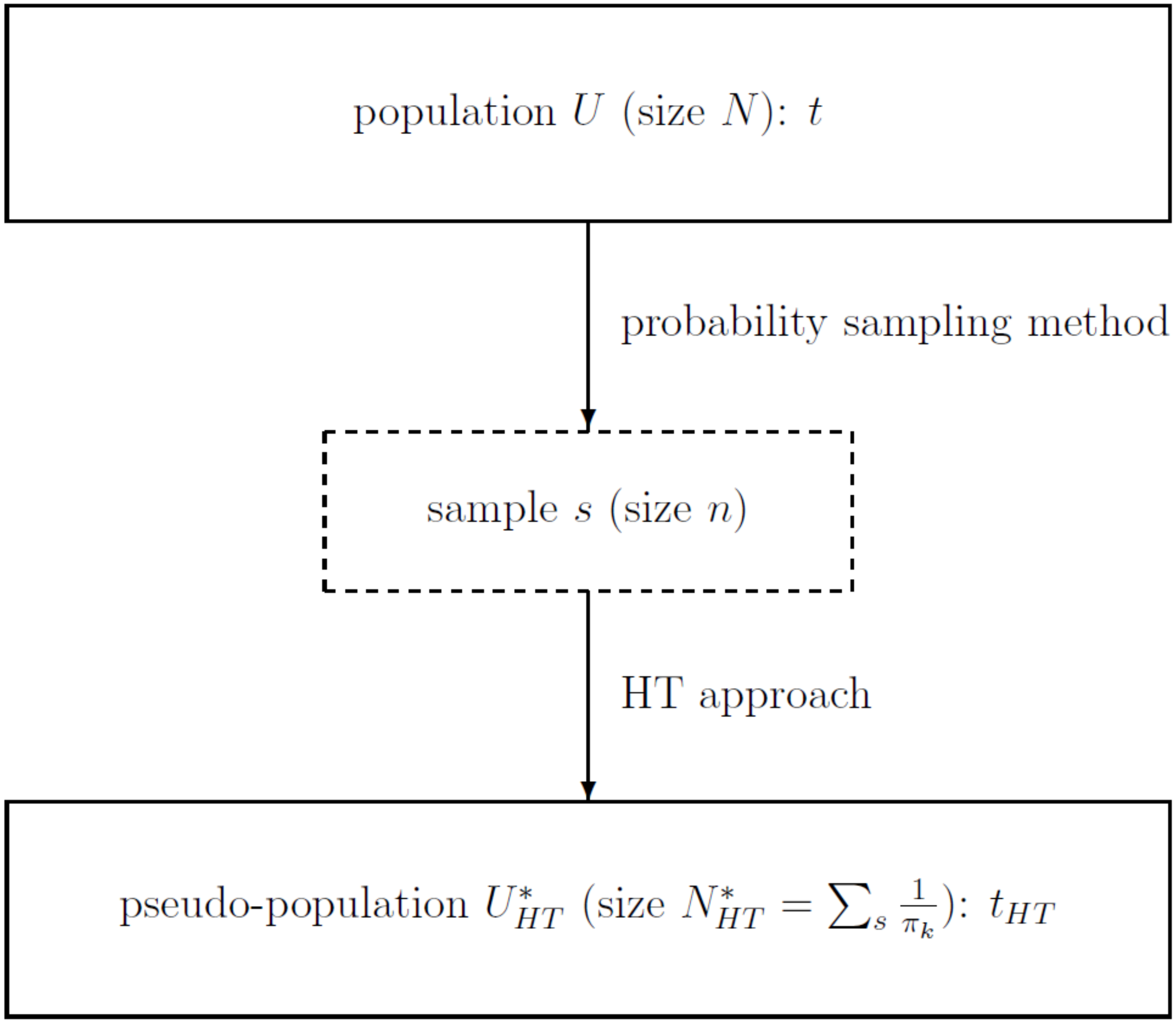
$$t_{HT} = \sum_{s} y_k \cdot d_k = \sum_{s} y_k \cdot \rho_{k(HT)}$$

For pseudo-population U_{HT}^{*}, sample value y_1 is replicated d_1 times, y_2 is cloned d_2 times, …, y_n is replicated d_n times.

With respect to U_{HT}^{*}, the d_k's can be interpreted as (not necessarily integer) *replication factors* $\rho_{k(HT)}$ of y_k (or *respondent's burden*) of the HT process.

The size N_{HT}^{*} of U_{HT}^{*} is given by $\sum_{s} \rho_{k(HT)}$

Hence, this results in a design-based estimation of t


population U (size N): t
probability sampling method
sample s (size n)
HT approach
pseudo-population U_{HT}^* (size $N_{HT}^* = \sum_s \frac{1}{\pi_k}$): t_{HT}

Rewriting the definition of the HT estimator:

The HT estimator of the total

$$t = \sum_U y_k$$

of study variable y in population U of size N is given by

$$t_{HT} = \sum_{U_{HT}^*} y_k^*,$$

the total of the replicated variable y^* in pseudo-population U_{HT}^* of size N_{HT}^*.

2. The Ratio Estimator

The *ratio estimator* (rat) of the total t is given by

$$t_{rat} = t_{HT} \cdot \frac{t^{(x)}}{t_{HT}^{(x)}} = \sum_s y_k \cdot \rho_{k(HT)} \cdot \frac{t^{(x)}}{t_{HT}^{(x)}} = \sum_s y_k \cdot \rho_{k(rat)}$$

The idea behind the ratio estimator can be described by the generation of a pseudo-population U_{rat}^* as a set-valued estimator of U with respect to t.

For pseudo-population U_{rat}^*, sample value y_1 is replicated $\rho_{1(rat)}$ times, y_2 is cloned $\rho_{2(rat)}$ times, …, y_n is replicated $\rho_{n(rat)}$ times.

The $\rho_{k(rat)}$'s are the *replication factors* of the ratio estimation process.

The size N_{rat}^* of U_{rat}^* is given by $\sum_s \rho_{k(rat)}$

Hence, this results in a model-assisted estimation of t

The *ratio estimator* of the total

$$t = \sum_U y_k$$

of study variable y in population U of size N is given by

$$t_{rat} = \sum_{U_{rat}^*} y_k^*,$$

the total of the replicated variable y^* in the pseudo-population U_{rat}^* of size N_{rat}^*.

Experience in teaching shows that this representation of different estimation methods helps a lot to improve the students' basic understanding of these concepts (especially of students with little knowledge on probability theory)

3. Weighting Adjustment & Data Imputation

The HT estimator in the presence of nonresponse:

$$t_{HT} = \sum_{s} y_k \cdot d_k = \sum_{s_r} y_k \cdot d_k + \cancel{\sum_{s_m} y_k \cdot d_k}$$

Weighting adjustment:

$$t_W = \sum_{s_r} y_k \cdot \rho_{k(HT)} \cdot \frac{1}{\widehat{\omega}_k} = \sum_{s_r} y_k \cdot \rho_{k(W)}$$

Replication factors $\rho_{k(rat)}$ contain the estimated response probabilities $\widehat{\omega}_k$

Pseudo-population U_W^* differs from U_{HT}^* in size and values

Data imputation:

$$t_I = \sum\nolimits_{s_r} y_k \cdot \rho_{k(HT)} + \sum\nolimits_{s_m} y_k^i \cdot \rho_{k(HT)}$$

Pseudo-population generated with HT replication factors $\rho_{k(HT)}$ for both the observed and imputed values

Pseudo-population U_W^* differs from U_{HT}^* in imputed values

Hence, both estimators of t are model-based

4. Examples of Further Applications

- General regression estimation
- Probability & non-probability sampling methods
- Poststratification & iIterative proportional fitting
- Associations in complex samples
- Capture-recapture
- Small area estimation
- Bootstrap method in finite population sampling
- Generalization of Randomized Response Questioning Designs to complex samples
- Techniques of Statistical Disclosure Control

Thank you very much for your non-pseudo-attention!