**Question 1**

Good morning, I would like to ask something about the exercise of the last video and package of slide. With some colleagues we were wondering about the computation of probability of inclusion of the household and the column to be filled. The first probability is the one regarding the inclusion of the village and for the north part should be 10/50, right? the 2nd probability is 10/200?

**Answer 1**

Yes, right. The probability of inclusion is the same for each village in the stratum because the Primary Sampling Units (the villages) are selected by simple random sampling without replacement:

πih=nah/Nah where nah is the number of villages selected in the stratum h and Nah is the number of villages in the stratum.

The probability of inclusion at the second stage is: πk|ih=nih/Nih because the Secondary Sampling Units (the households) are selected by simple random sampling without replacement. This probability is the same for each household in the selected village. As nih and Nih do not change changing stratum, it is always equal to 10/200.

The probability of inclusion of k-th household is equal to πk= πih· πk|ih

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | North | Centre | East | West | South |
| πih | 10/50 | 13/50 | 6/50 | 7/50 | 14/50 |
| πk|ih | 10/200 | 10/200 | 10/200 | 10/200 | 10/200 |

**Question 2.**

**2.1** My understanding about HT estimator is, suppose we have a data set, but we are interested in one specific domain and we are not going to conduct another survey focused on the domain. In that case we estimate the domain total and variance by HT estimation, using the existing data set. Is this a correct understanding?

**2.2** And, what is the purpose of estimating the domain total and variance? Is it to estimate the distribution of the data in the specified domain?

**Answer 2.1**

Yes, it is correct. The HT estimator is called a direct “expansion” estimator of the total T of the study variable, based only on survey data. I imagine that the variance you are mentioning is the variance of the estimator,



that measures its sampling error (expression 2 or 3 in the slides).

The variance of the study variable y, indicated by σ2, can be estimated from the sample as well – by the way it is more difficult -and measures the variability of the study variable in the population.

**Answer 2.2**

The distribution of the data in the domain is a very relevant “target parameter”.

Here you see the consumption expenditure distribution in Italy in 2012 (data from HBS).

You can estimate it by the HT estimator, estimating the frequencies associated to the classes of consumption, when the domain is planned et equal to the Italian Regions, as in this figure. In the slides we focus on the estimates of a synthesis of this distribution: the total consumption expenditure, the mean consumption expenditure, and P0 and P1, once defined the poverty line *z*. It is possible to estimate also percentiles and quantiles in the domain.

**Question 3.**

Could you explain again why we need two different ways for the variance estimation, for “planned domain” and “unplanned domain”?

The variability of the estimator HT is function of the sample size in the domain. This sample size is fixed when the domain is planned, while it is a random variable when the domain is not planned. So in this case we need another estimator of it.

In case of unplanned domains you cannot fix the size of the sample before the selection of it. Imagine to size a sample in Tuscany to obtain accurate estimates of total consumption expenditure at regional level. When the survey is done, a policy maker requires additional estimates of the total consumption expenditure at Municipality level. While the accuracy of the estimate at regional level is planned, the accuracy at Municipality level is not, as the sample size you have in each Municipality is equal to the number of observations, which, by chance, are in each Municipality. You have not planned this number and it can be very small, even zero. It is the value of the random variable “sample size in the Municipality”.

**Question 4.**

When will this HT estimator not applicable anymore? For example, if you want to analyze domains that you have to be relevant in the study but have really small sample size. Is there a limit to these small area estimators?

Yes, right. They are not always accurate for unplanned domains, the domains for which you do not control the sample size, and by chance you can have very few observations.