Urvalsmetoder och Estimation 5

Sampling and Estimation 5

2011-02-14

- Next time the last assignment will be dealt out. I.e. you will get a list of papers to choose from (at first come first served basis).
- 27 first assignments → about 15-20 presentations á 10 minutes → 4-5 hours altogether. (2 h is planned for March 14). We need another 3 hours for presentations When do you want it?

- February 28, Frida Videll from Statistics Sweden will come and talk about LFS (The Labour Force Survey).

5.4
Are sampling with varying probabilities and HT-estimation any good

?

Basu´s elefants

- A circus-owner arrives a railway station with his herd of five elephants. He must estimate the weight in order to pay for the freight. There is only one scale and each weighting costs 1 rupie, but the owner has only one coin. He decides to weight Mumbo, the middle elphant and multiply by 5.
- The circus statistician protests. This is not a probability sample! Every elephant must have a positive probability.
- But he admits that it should be more sensible to weight the middle elephant than the two extremes so they decide on a compromise: Weight the biggest Colonel Hathi or the smallest Jumbo with probabilities 0.01 and the next biggest or smallest, Tvumbo and Dumbo with probability 0.04 each.
- After looking into a random number table Mumbo is chosen and the director is happy: What did I say? Mumbo is weighted and the result is 2 ton. The director estimates the total weight to 10 ton.
- The statistician, however, says No! No! That is not an unbiased estimate; use the HT-estimate 2/0.9=2.22 ton.

(What would the the estimate have been, if Colonel Hathi had been weighed (His weight is 5 ton?). The other weights are 3, 1.5 and 1)

When does HT behaves badly?

1. If the sample size varies the HT-estimate may be severely affected (or rather if $\Sigma_S\ 1/\pi_i$ is a bad estimator of the total population size).
   - One often uses a ratio estimate instead $N(\Sigma_S\ y_i/\pi_i)/(\Sigma_S\ 1/\pi_i)$.
   - As a rule of thumbs: Use HT-estimates only when the sample size is fixed or varies marginally.
   - Sample size may in this discussion be replaced by other size measures. E.g. the number of employees in a firm. Use ratio estimators also in this case (or regression or prediction estimators see below)
2. Model-based methods are usually preferable for small sample sizes (cf Basu's elephants)

MSE of Basu’s elephants

- MSE of median elephant 6.25
- Of HT-estimator 72000
- Of HT-ratio estimator 9

6. Cluster sampling

6.1 Introduction

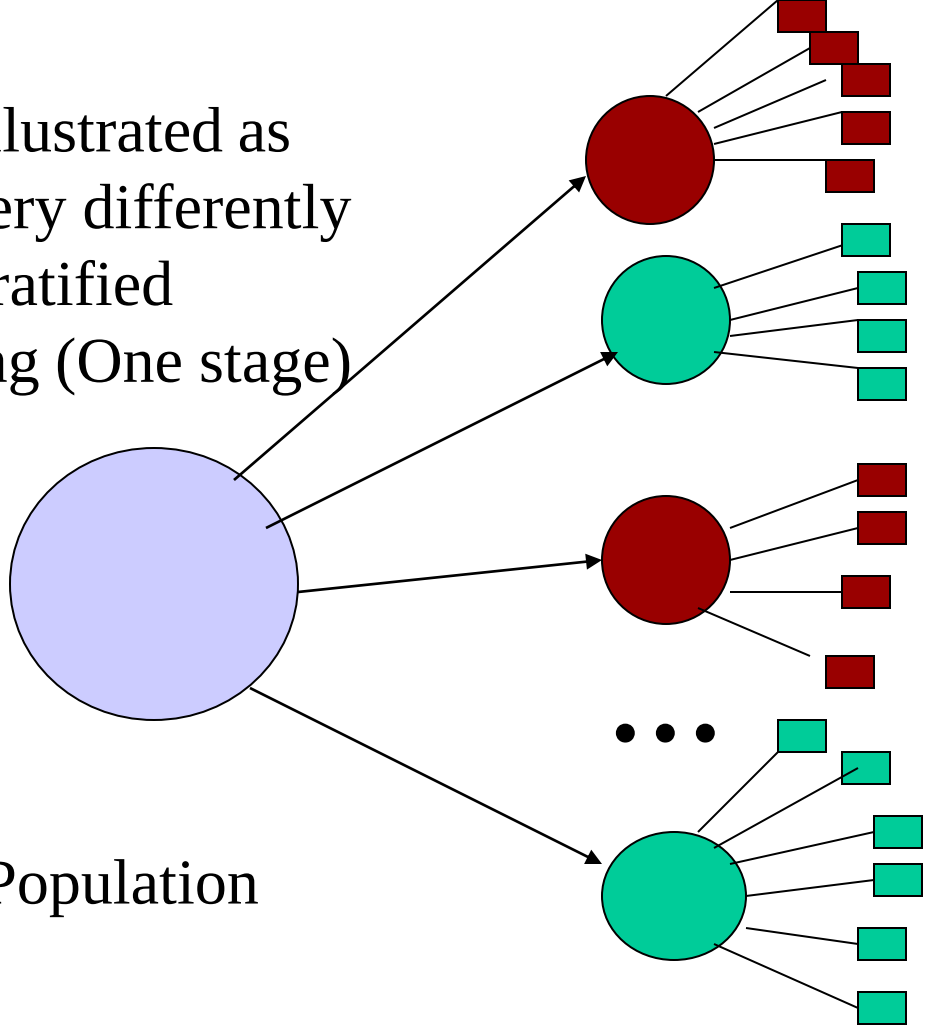
6.1.1 Description

- The population consists of a number of groups of units. Only a sample of these groups are selected.
  - These groups are called clusters or primary sampling units, psu.
  - The units are called secondary sampling units, ssu.

- In stratified sampling a few units from all groups/strata were selected.

- Example: Study the teachers at Stockholm university. Let departments be clusters and select a few departments and interview some of the teachers at that department

Often illustrated as here. Very differently from stratified sampling (One stage)



In sample

Not in sample

In sample

• • •

• • •

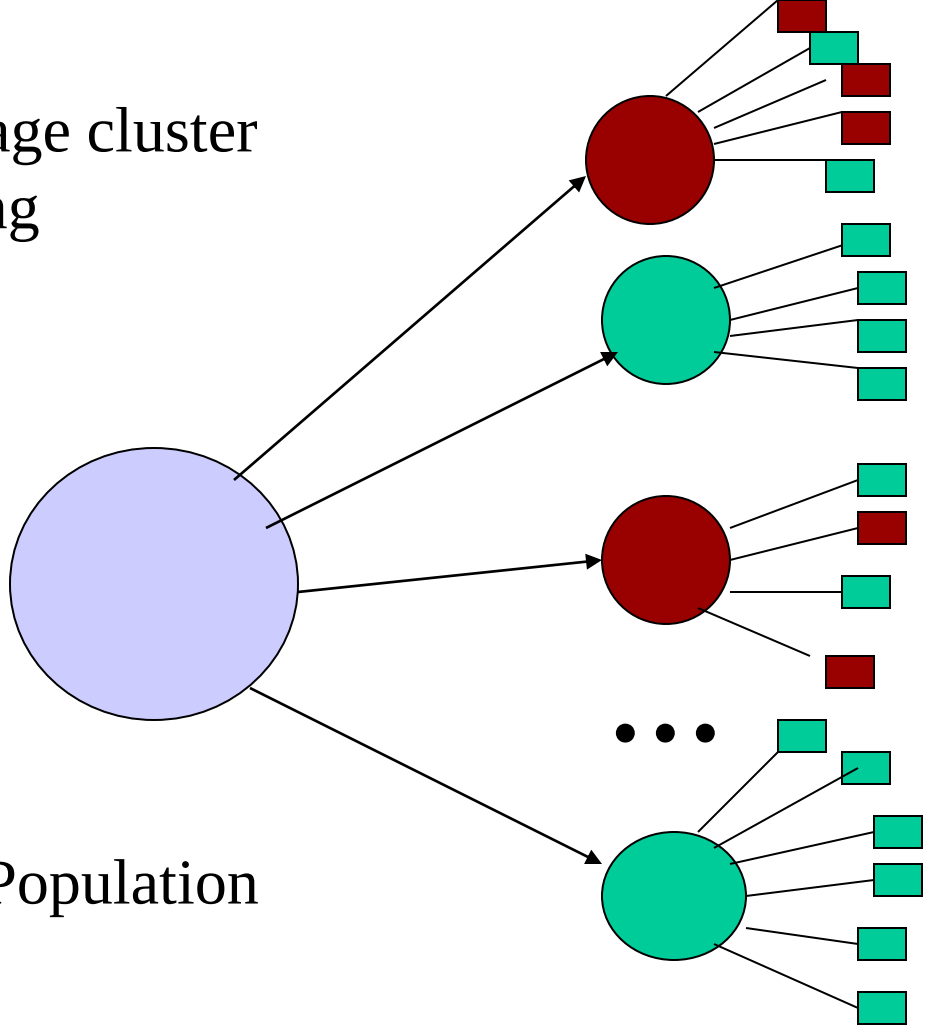
Not in sample

Population

n out of N clusters or psu:s

$K = \Sigma M_i$ ssu:s

Two-stage cluster sampling



In sample (red)

Not in sample

In sample (red)

• • •

• • •

Not in sample

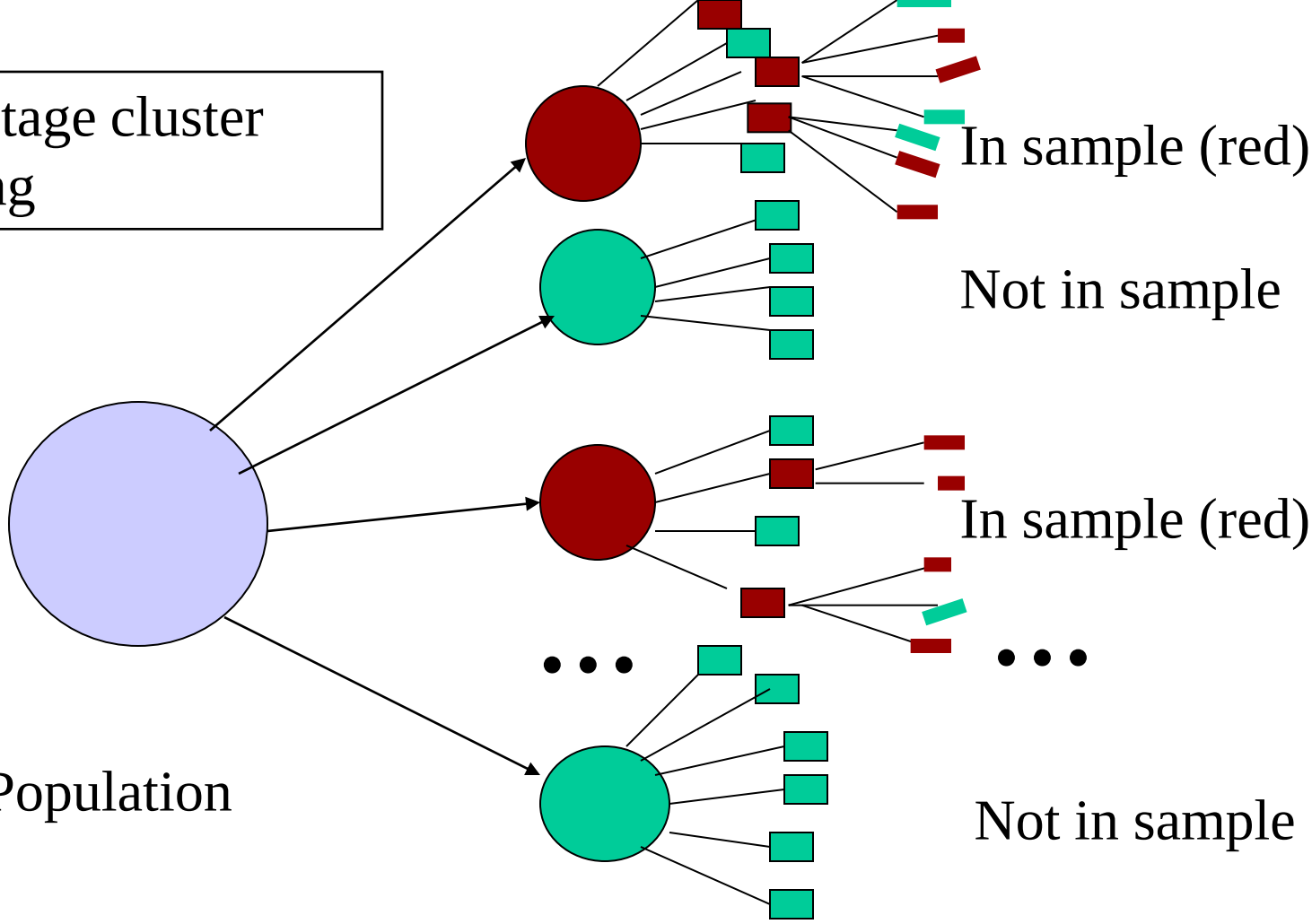
Population

n out of N clusters or psu:s

$K = \Sigma M_i$ ssu:s

$n = \Sigma m_i$ in sample

Three-stage cluster sampling



Four and more stages accordingly

n out of N clusters or psu:s

$K = \Sigma M_i$ ssu:s

$n = \Sigma m_i$ in sample

Where have you seen this situation in other parts of statistics?

Where have you seen this situation in another part of statistics?

- In variance analysis with nested factors (Hierarchical design. Variance components). You can look it up).
- In multilevel models (modern techniques/ models for such data, also discrete. A good program is Mlwin)

6.1.2 Examples

- Study pupils in fifth grade. Select a number of schools as clusters and all fifth grade pupils at these schools
- Four stage cluster sampling. Select first a number of municipalities, second some schools, third some classes, fourth some pupils.
- Multicenter studies. Recruit patients in five different clinics.

Further examples

- Auditing: Select some days during the year. Study some of the transactions during those days.
- Forest inventory (”Riksskogstaxeringen”): Select some random points in the forested area of Sweden. Select 16 four square meter areas around this point (1/2 km apart on a square. A one days work to investigate all these areas). One of Swedens oldest surveys.

Still further examples

- We have talked about villages in a development country or blocks in a town with households as ssu:s.
- Systematic sampling with several start values may be thought of as one stage cluster analysis. Each cluster is given by one starting value.
- RDD. Select all but the last two digits in a telephone number. Draw one to five numbers from each 100-series (used for substitution).

Ordinary variance analysis with hierarchical models/variance components/multilevel models under normality

$Y_{kji} = \alpha_k + \beta_{kj} + \varepsilon_{kji}$, *where*

$\alpha_k \in N(m, \sigma_1^2)$ *is the primary (school) effect*

$\beta_{kj} \in N(0, \sigma_2^2)$ *is the secondary (class) effect*

$\varepsilon_{kji} \in N(0, \sigma_3^2)$ *is the tertiary (pupil) effect*

all variables are independent. Estimate all variances by the corresponding sums of squares

$$s_3^2 = \sum_{kji} (y_{kji} - \overline{y_{kj.}})^2 / \sum_{kji} (n_{kji} - 1)$$

$$s_2^2 + s_3^2 / n_{kji} = \sum_{kji} (\overline{y_{kj.}} - \overline{y_{k..}})^2 / \sum_{kj} (n_{kj} - 1)$$

$$s_1^2 + s_2^2 / n_{kj} + s_3^2 / \Sigma_i n_{kji} = \sum_{kji} (\overline{y_{k..}} - \overline{y_{...}})^2 / \sum_{k} (n_k - 1)$$

(*formulas when all n are equal at the same level*)

- There are three variance components here. For modelling it is easiest to use σ^2, but in the estimation phase the sums are the natural expressions to work with. (They follow χ^2-distributions) (We will see the same effect later here in cluster sampling)

- Cluster sampling
  - No assumption on normality (only design variance)
  - No assumption on independence
  - Finite population (corrections) at all levels (interested in size weighted mean)
  - Varying sample sizes

- But still estimate the variances by dividing the observed variance into components

6.2 Analysis
6.2.1 One-stage cluster sumpling

- As usual assume
  - Objective of our analysis is estimation of population total
  - Selection at all stages is SRSwor
- In one stage cluster sampling, all units in the n selected clusters are sampled

Notation, One-stage SRS

- N = number of clusters (psus) in frame
- n = number of clusters in sample
- M_i = number of units in cluster i
- Y_{ij} = value of unit j in cluster i
- $T_i = \Sigma_{ij}^{Mi} Y_{ij}$
- $T = \Sigma_i T_i = \Sigma\Sigma_{ij} Y_{ij}$

Estimation

- Pretend that the clusters i = 1, …, N are the units in the frame and that the cluster totals T_i are the study variables.

- Use estimates from SRS. All formulas carry over

- $t_{Ti} = (N/n) \Sigma_{i€S} T_i$
  $Var(t_{Ti}) = N(N-n) \sigma_{Ti}^2 /n$
  $Var^*(t_{Ti}) = N(N-n) s_{Ti}^2 /n$

Remarks

- This method works well if the clusters are (almost) equally large.

- With varying sample sizes one may have better to use ratio estimates e.g. $(\Sigma_{i€S} T_i / \Sigma_{i€S} M_i)*\Sigma_{i€U} M_i$ (if the cluster sizes are known)

- The variance can be estimated in the same way as for the ordinary ratio estimator. E.g. by writing

$RelVar(t_{yR}) \sim RelVar(t_y) + RelVar(t_x) - 2RelCov(t_y t_x)$

where the relaitive variances/covariances are estimated as we said before. (We gave another expression too. That works equally well but is more difficlult to generalise into two and more stage sampling.)

- In stratified sampling we wanted the groups/strata to be as different as possible - in cluster sampling we want the groups/clusters to be as similar as possible.

- An extreme case (never encountered in practice) illustrates this: The population consists of married couples (psus) (no unisexmarriages) and you want to estimate the proportions of females in that population. It is sufficient to take 1 cluster, which gives a perfect estimate (Variance 0). (A SRS from the full population would have given variance ((2N-n)/(2N))*(1/(4n)).)

One stage cluster sampling with varying inclusion probabilities

- Suppose that cluster i is chosen with probability π_i
- Then we may estimate the total with the HT-estimator $\Sigma_i \, T_i/\pi_i$.
- We can use the ordinary formulas for variance estimation e.g. SYG (if fixed number of clusters)

  $\frac{1}{2} \, \Sigma\Sigma_{SS} \, ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) \, (T_i/\pi_i - T_j/\pi_j)^2$
- Or use a HT ratio estimator

  $(\Sigma_{i \in U} \, M_i) \, (\Sigma_{i \in S} \, T_i/\pi_i)/(\Sigma_{i \in S} \, M_i/\pi_i)$

6.2.2 Two-stage cluster sampling

Total or mean estimator

We have already seen how we can obtain an unbiased estimate, its variance and an unbiased variance estimator using the HT- and the SYG -estimators. Thus we could have stopped here. However, we give also the usual direct treatment in the following.

- Look at the case when only a sample (m_i) is observed in each selected cluster (drawn by SRS within the selected clusters).
- The total T_i within cluster is estimated as for SRS and we get estimates t_i with variances $Var(t_i)= M_i (M_i-m_i) \sigma_i^2 /m_i$.

- If we had observed every unit in the selected clusters the estimator of the total would have been

$$t_{Ti} = (N/n) \Sigma_i T_i$$

Now just replace every T_i by its estimate t_i, giving

$$t_{Ti} = (N/n) \Sigma_i t_i$$

Variance

- The estimator was $t_{Ti} = (N/n) \Sigma_{h€S} t_i$ where the t_i estimate the cluster totals. It is unbiased.

- Its variance is the sum of
  - The variance of $(N/n) \Sigma_i T_i$ if all T_i were known (see one stage cluster sampling, above)
  - The variance of the estimates within all the strata (due to the replacement of T_i by their estimates t_i $(N/n)^2 \Sigma_{h€S} Var(t_i - T_i)$

$$Var(t_{Ti}) = Var((N/n) \Sigma_i T_i) + (N/n)^2 \Sigma_{h€S} Var(t_i - T_i) =$$
$$N(N-n) \Sigma_{h€U} \sigma_{Ti}^2/n + (N/n)^2 \Sigma_{h€S} M_i (M_i-m_i) \sigma_i^2 /m_i$$

- One problem remains, though: the sample is random. So we take the expectation over the last term

$$Var(t_{Ti}) = N(N-n) \Sigma_{h€U} \sigma_{Ti}^2 /n + (N/n) \Sigma_{h€U} M_i (M_i-m_i) \sigma_i^2 /m_i$$

Variance estimator

- The situation is now, as we saw, more complicated than those we have previously encountered. There are two variance components and two levels of correction for a finite population M_i <-> m_i and N <-> n.

- What about the variance estimator? We use a new trick; not the same but similar to that for the variance.
  - Suppose that t_i were the true cluster totals, then the variance (from one stage sampling) would be
    $N(N-n)\,\sigma_{ti}^2/n$
    with the estimator
    $N(N-n)\,s_{ti}^2/n$

- The difference between this total and the true total is
  $\Sigma_{i\in U}(T_i-t_i)$

with variance (formula: stratified sampling for each cluster)

$\Sigma_{i\in U}(M_i-m_i)\,M_i\,\sigma_i^2/m_i$

which can be estimated by

$(N/n)\,\Sigma_{i\in S}(M_i-m_i)M_i\,s_i^2/m_i$

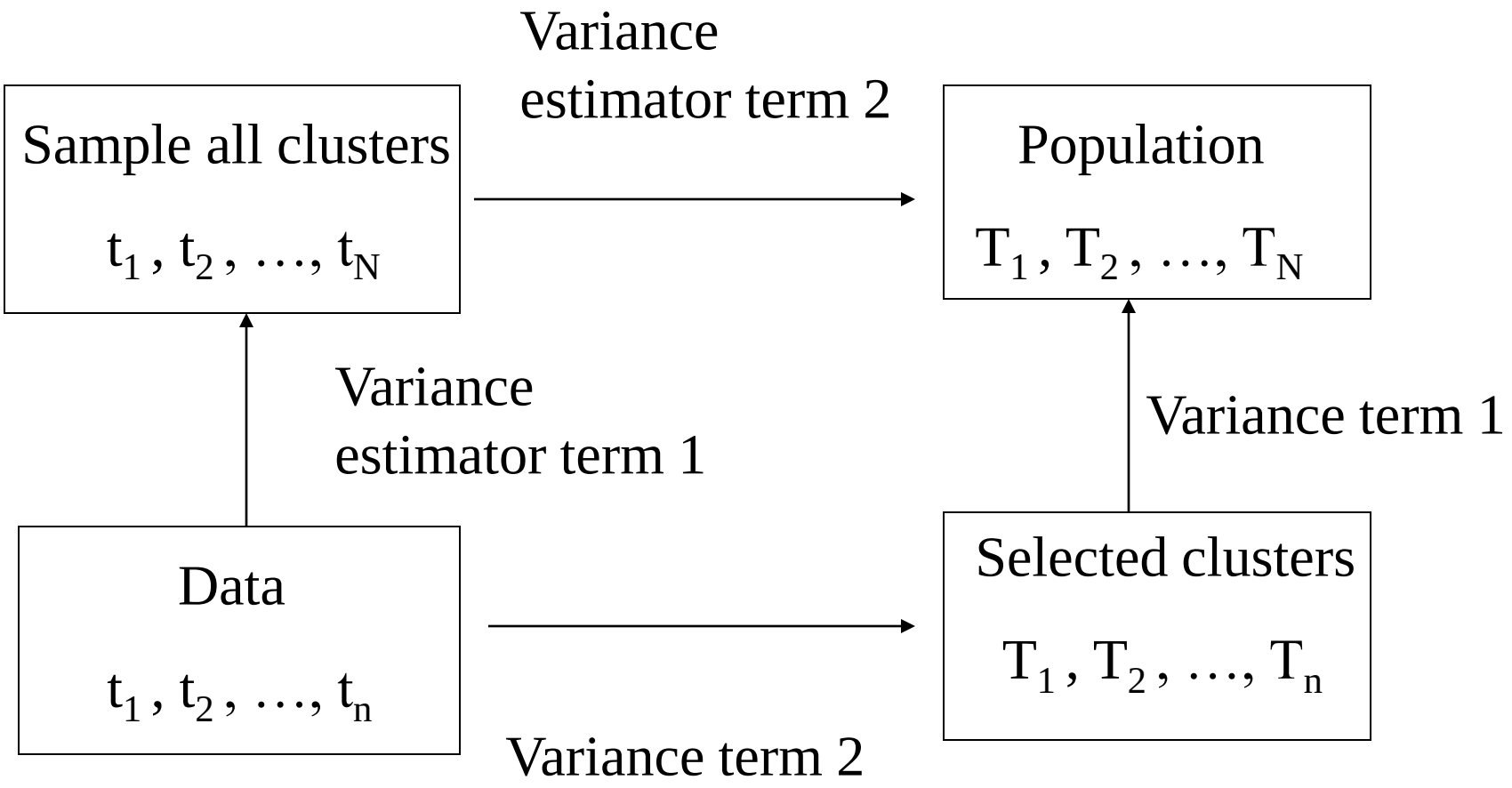
- The total variance estimator is thus the sum of these two terms

$$N(N-n)\, s_{ti}^2 /n + (N/n) \Sigma_{i \in S} (M_i - m_i) M_i s_i^2 / m_i$$

(Since the two error sources are independent, our sampling within a cluster does not depend on the sampling of clusters)

- Note that the terms here do not correspond exactly to the two terms in the variance expression on the previous page. The expected value of the first term is larger here and of the second term it is smaller.

Description of variance terms



An example – A school

- In a school there are 8 classes at level 9. (With 32, 27, 14, 28, 22, 13, 25, and 30 pupils). Three classes are selected and five pupils are selected in each class. They are asked how they have liked the school. Among other things on an ordinal discrete scale with ten levels ranging from highly dislike (1) to liked very much (10). (Not a question in this course but what are the properties of such a scale?). Estimate the average score for all pupils in the school and give a precision interval!

Data and mean estimate

- Data
  - Group/class 1 (28 p.) 7 8 3 6 9
  - Group/class 2 (14 p.) 1 3 6 2 - (assume randomly missing i.e set $m_2 = 4$)
  - Group/class 3 (25 p.) 6 10 9 5 6
- Total estimate

  [28*(7+8+3+6+9)/5 + 14*(1+3+6+2)/4 + 25*(6+10+9+5+6)/5]*(8/3) = 1084.8
- Mean estimate 1084.8/191=5.68

Ratio estimator

- One might argue that one should use a ratio estimator when the cluster sizes vary much.
- [28*(7+8+3+6+9)/5 + 14*(1+3+6+2)/4 + 25*(6+10+9+5+6)/5]*((32+27+14+…+30)/(28+14+25) = 1159.7
- This is not an exactly unbiased estimator. The previous estimator is the standard one.
- Mean estimator 6.07

Variance estimate

- Variance within clusters s_i^2 = 5,3 , 4,67 resp. 4,7
- The between clusters variance s_{ti}^2 is based on the three cluster total estimates 184,8 , 42 and 180 giving s_{ti}^2 = 6584
- Now we have all ingredients of the formula:
  $Var^*(t) = N(N-n)\, s_{ti}^2/n + (N/n)\, \Sigma_{i \in S} (M_i - m_i) M_i\, s_i^2/m_i =$
  $8*5*6584/3 + (8/3)[28*(28-5)*5,3/5 + 14*(14-4)*4,67/4 + 25*(25-5)*4,7/5] = 94896$
- Standard deviation 308
- Interval 5,68 +/- 2*308/191 = 5,68 +/- 3,23 = (2,4, 8,9)
- (but three groups are not enough with these differences between clusters. (Using standard statistical theory based on normal distribution the factor 2 should be replaced by a t-value with 2 degrees of freedom which is much higher for a confidence level of 0,95 (but who said anything about the level?)))

Comments

- This was a description for two-stage sampling with SRS at both stages.
- These methods can be generalised
  - to other sampling schemes e.g. πps at the first stage
  - to other estimation techniques e.g. ratio estimates
  - to more than two stages
- But formulas will soon become hideous!
  - It would have been complicated enough to derive variance estimates for two stage sampling without our tricks.

πps-sampling

- Cluster sampling may be done with varying selection probabilities (Leads to more complicated formulas, e.g. see next page)
- Self-weighted sampling (probability proportional to size, i.e. number of units in the cluster (ssu:s in the psu))
- Lohr discusses formulas for sampling with replacement – but there is a general theorem saying that

Sampling with replacement is always less efficient than without replacement.

- Thus: never sample with replacement if you can avoid it. (You may use the variance formulas for with replacement sampling and be on the safe side even if you sample without replacement).

Variance estimator

- The situation is now, as we saw, more complicated than those we have previously encountered. For the variance estimator? We use a the same trick as before.
  - Suppose that t_i were the true cluster totals, then the variance estimator (SYG) would be
  - $\Sigma\Sigma_{SS} ((\pi_i\pi_j - \pi_{ij})/\pi_{ij}) (t_i/\pi_i - t_j/\pi_j)^2$

- The difference between this total and the true total is

$$\Sigma_{i€U} (T_i - t_i)$$

with variance (formula: stratified sampling for each cluster)

$$\Sigma_{i€U} (M_i - m_i) M_i \sigma_i^2 / m_i$$

which can be estimated by

$$\Sigma_{i€S} (M_i - m_i) M_i s_i^2 / (\pi_i m_i)$$

The total variance is estimated by the sum of these two.

6.3 Comments

6.3.1 Is cluster sampling any good?

- Cluster sampling can seldom be motivated from variance aspects only. The main reason is administrative or cost efficiency.
- Sometimes one may also be interested in the relations between objects e.g. between class mates or within a village or family. In that case one may need to observe the full cluster
- In stratified sampling we wanted the groups/strata to be as different as possible - in cluster sampling we want the groups/clusters to be as similar as possible. ”As Karlstad votes - Sweden votes”. If that were true it would be sufficient to study one cluster: Karlstad.
- With varying sample sizes SRS may not be good at the first level. One may have better use πps-sampling at first stage.

6.3.2 Design aspects

- Often design is chosen so that one cluster is a practical unit like one day's work for one person or one box of papers …
- Try to make clusters as similar as possible.
- One can try to find a cost function and find optimal cluster sizes. However costs function are seldom continuous for cluster sampling as they were for stratification.

.

- Cluster analysis can be combined with stratification.
  - E.g. stratify after county or in urban-rural areas and take clusters within strata. In that case try to make strata as different as possible and clusters as similar as possible within strata.
  - or schools after stage and private-municipal.

- When the design is determined, estimate total cost and total variance and decide whether these are acceptable. (Cost not too high, precision not too low or unnecessarily high).

6.3.3 Master samples (basurval)

- A master sample is a sample that is taken once and after that used in many different surveys.
- A first stage in a cluster analysis can be suitable as a master sample. Select a number of municipalities as permanent clusters. Recruit and train an interviewer living at that location and take new ssu:s for each new study.
- Other examples: Web panels, Persons born on the 15th (c.f. Metropolit)
- The term panel formally used for a sample that is followed over time in a longitudinal study but is often used also when it is used only for ordinary one time studies