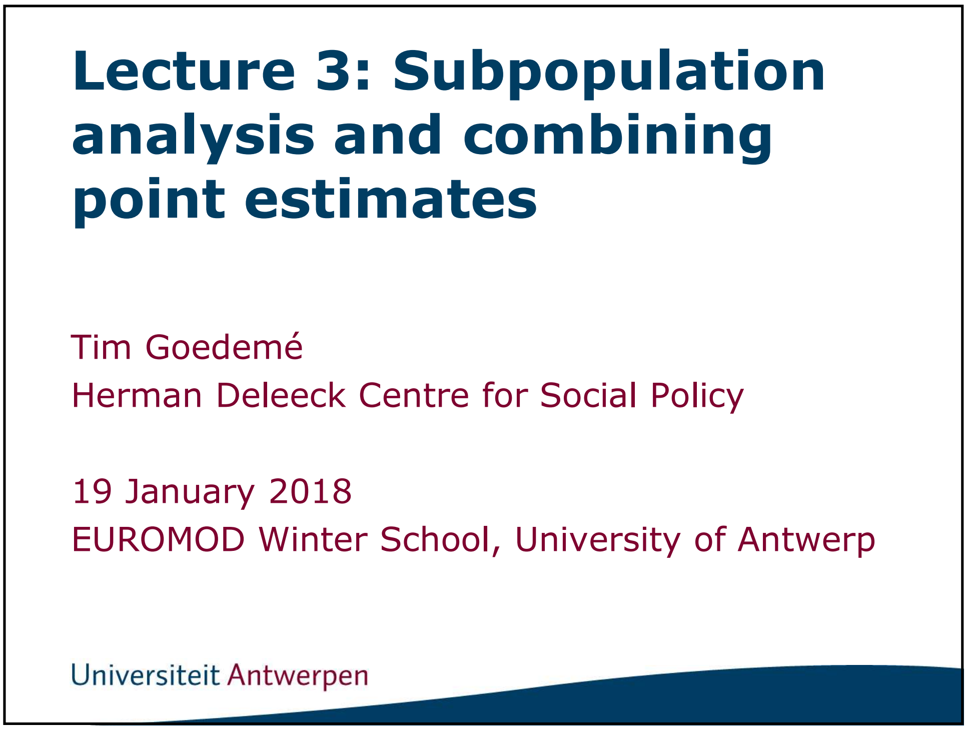
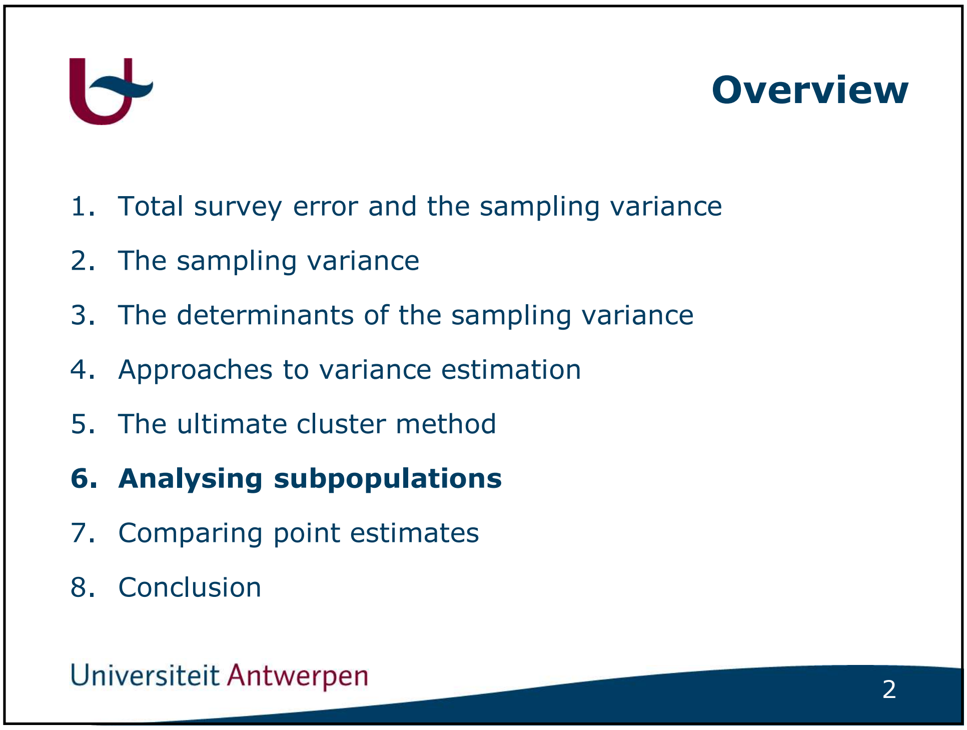


Lecture 3: Subpopulation analysis and combining point estimates

Tim Goedemé
Herman Deleeck Centre for Social Policy

19 January 2018
EUROMOD Winter School, University of Antwerp

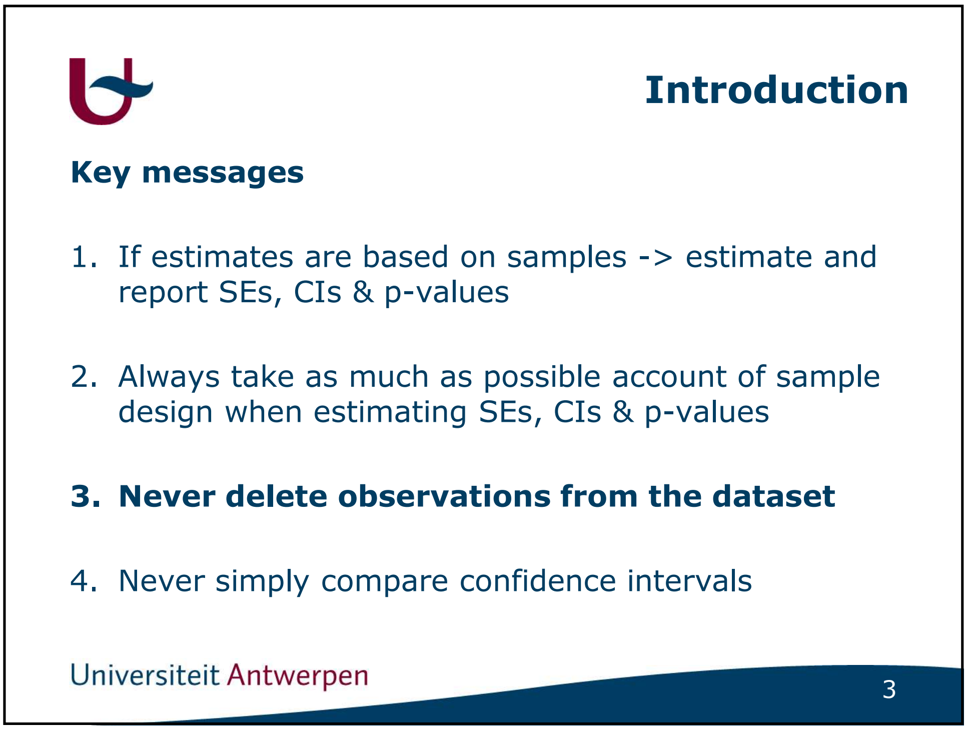
Universiteit Antwerpen



Overview

1. Total survey error and the sampling variance
2. The sampling variance
3. The determinants of the sampling variance
4. Approaches to variance estimation
5. The ultimate cluster method
6. **Analysing subpopulations**
7. Comparing point estimates
8. Conclusion

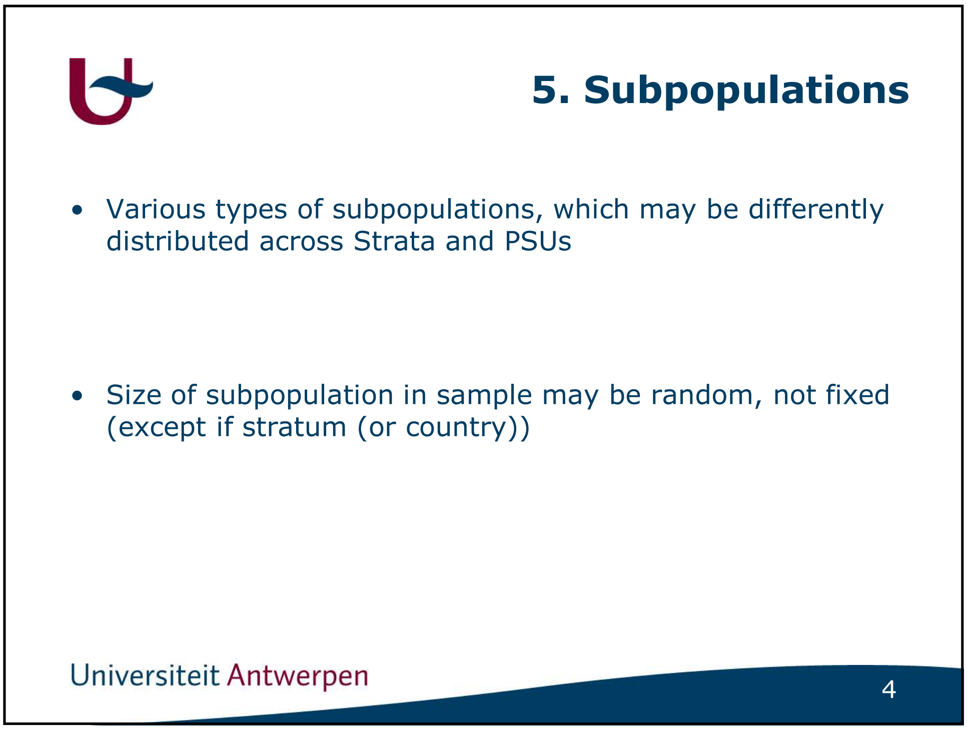
Universiteit Antwerpen



Introduction

Key messages

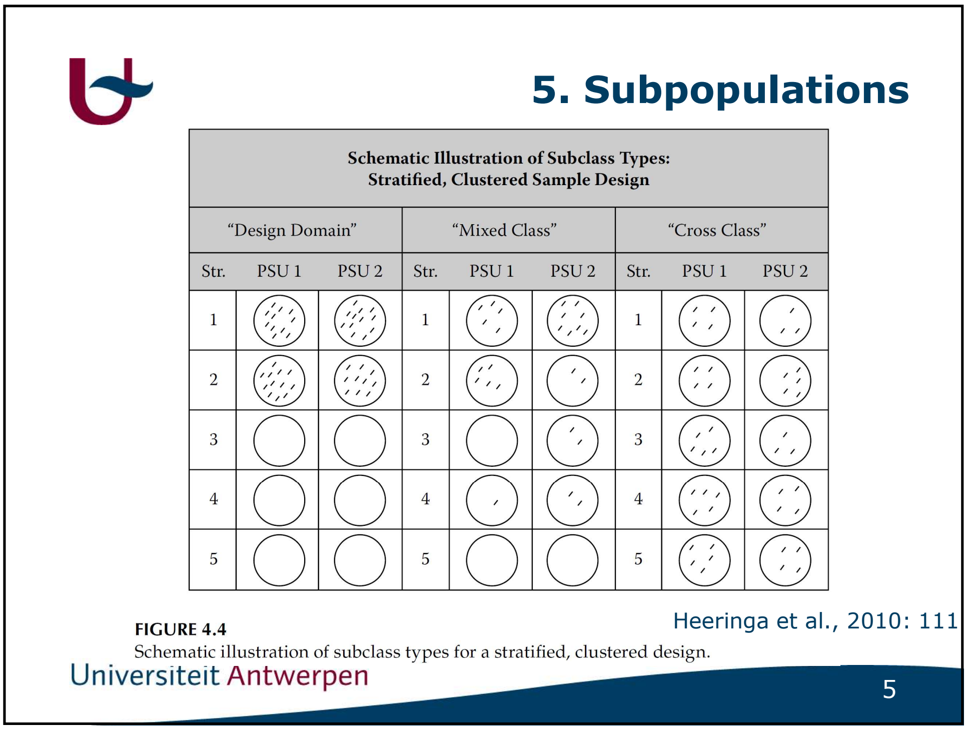
1. If estimates are based on samples -> estimate and report SEs, CIs & p-values
2. Always take as much as possible account of sample design when estimating SEs, CIs & p-values
3. **Never delete observations from the dataset**
4. Never simply compare confidence intervals



5. Subpopulations

- Various types of subpopulations, which may be differently distributed across Strata and PSUs

- Size of subpopulation in sample may be random, not fixed (except if stratum (or country))

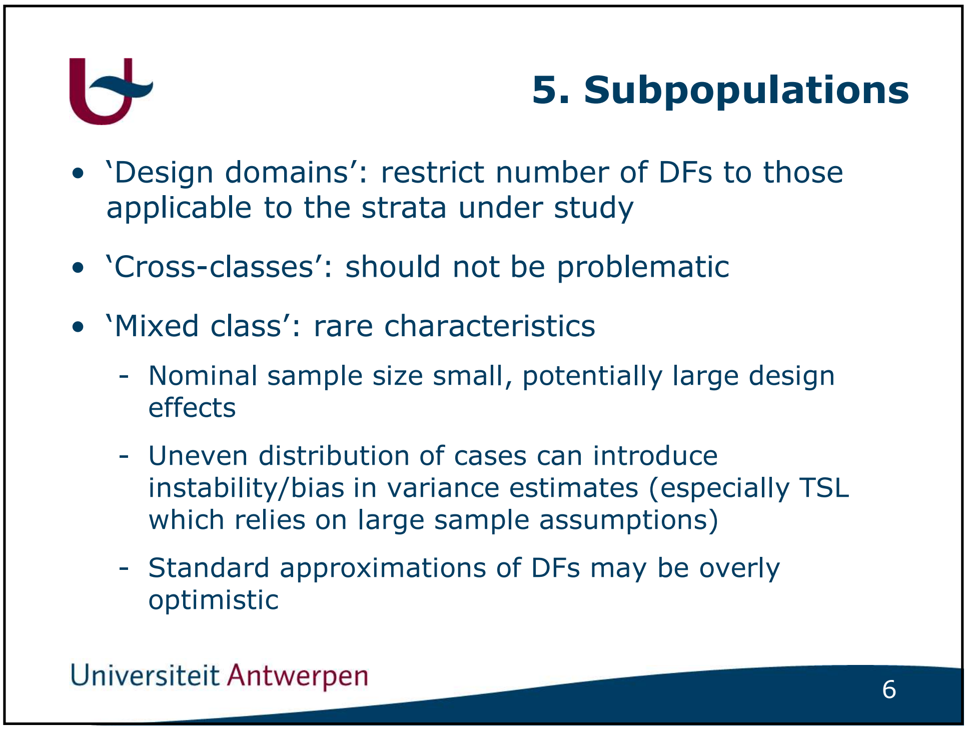


5. Subpopulations

Schematic Illustration of Subclass Types:
Stratified, Clustered Sample Design
"Design Domain"
"Mixed Class"
"Cross Class"
Str.
PSU 1
PSU 2
1
2
3
4
5

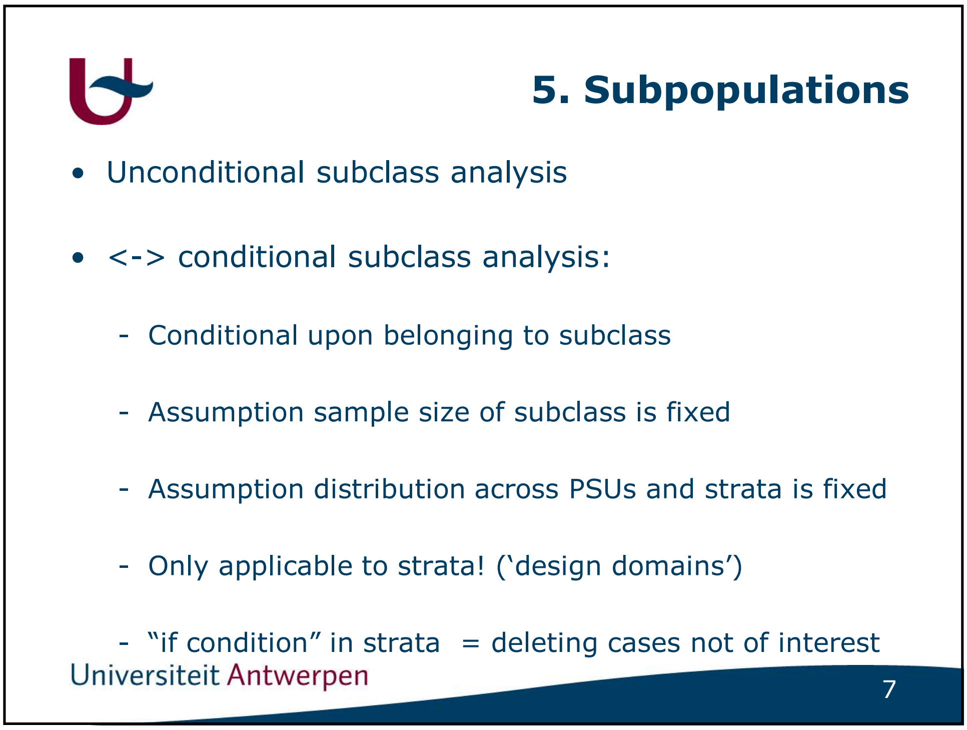
Heeringa et al., 2010: 111

FIGURE 4.4
Schematic illustration of subclass types for a stratified, clustered design.



5. Subpopulations

- 'Design domains': restrict number of DFs to those applicable to the strata under study
- 'Cross-classes': should not be problematic
- 'Mixed class': rare characteristics
  - Nominal sample size small, potentially large design effects
  - Uneven distribution of cases can introduce instability/bias in variance estimates (especially TSL which relies on large sample assumptions)
  - Standard approximations of DFs may be overly optimistic

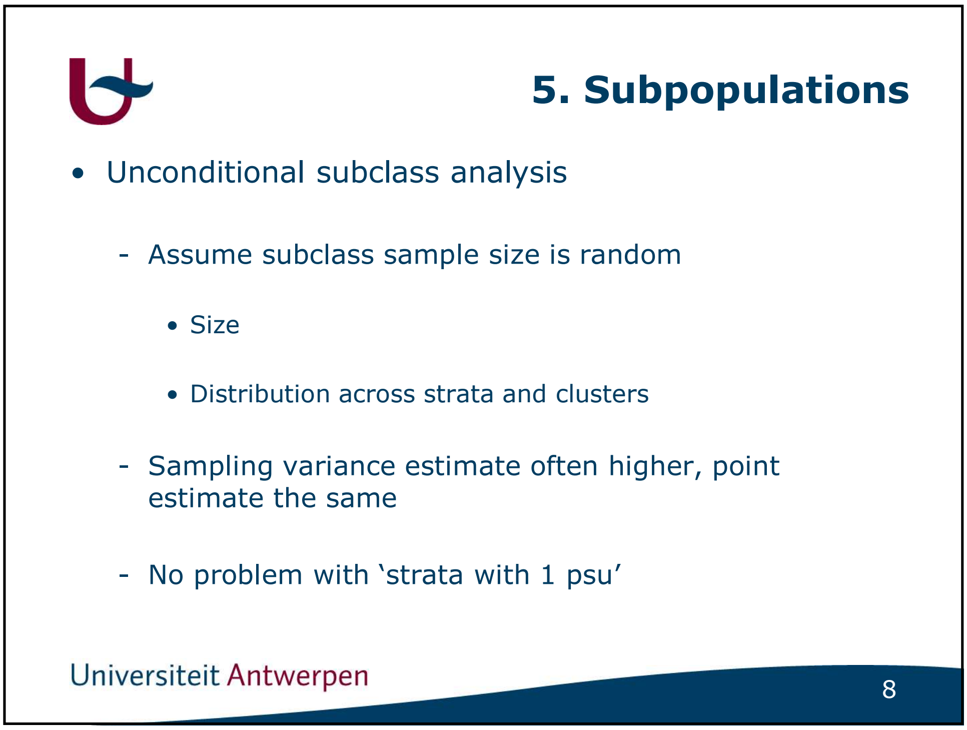


5. Subpopulations

- Unconditional subclass analysis
- <-> conditional subclass analysis:
  - Conditional upon belonging to subclass
  - Assumption sample size of subclass is fixed
  - Assumption distribution across PSUs and strata is fixed
  - Only applicable to strata! (‘design domains’)
  - “if condition” in strata = deleting cases not of interest

Universiteit Antwerpen

7

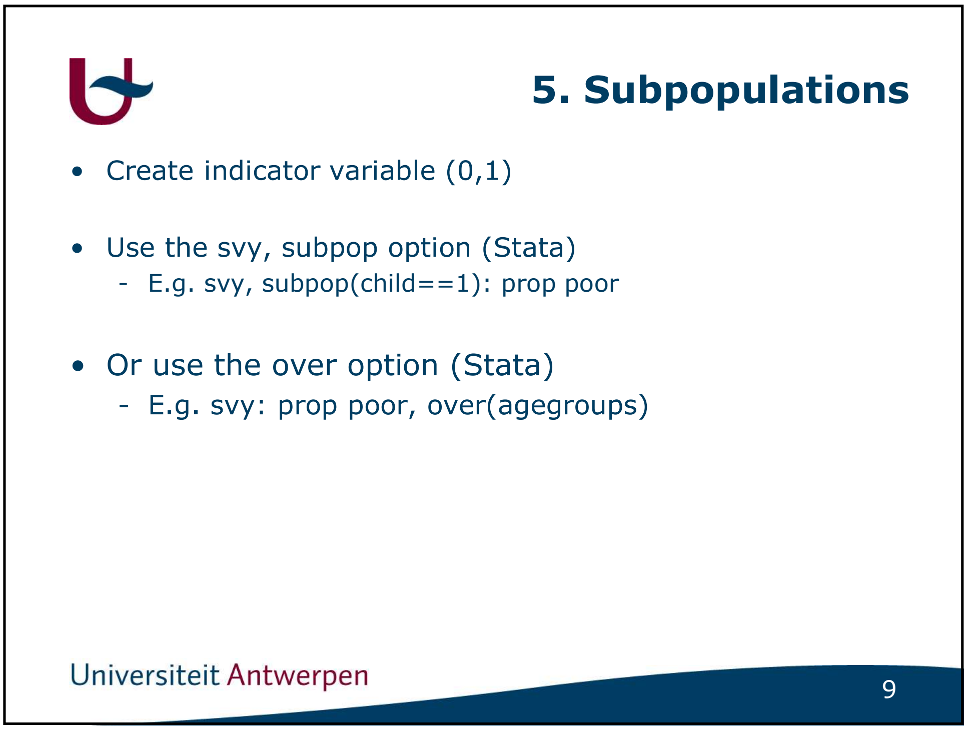


5. Subpopulations

- Unconditional subclass analysis
  - Assume subclass sample size is random
    - Size
    - Distribution across strata and clusters
  - Sampling variance estimate often higher, point estimate the same
  - No problem with ‘strata with 1 psu’

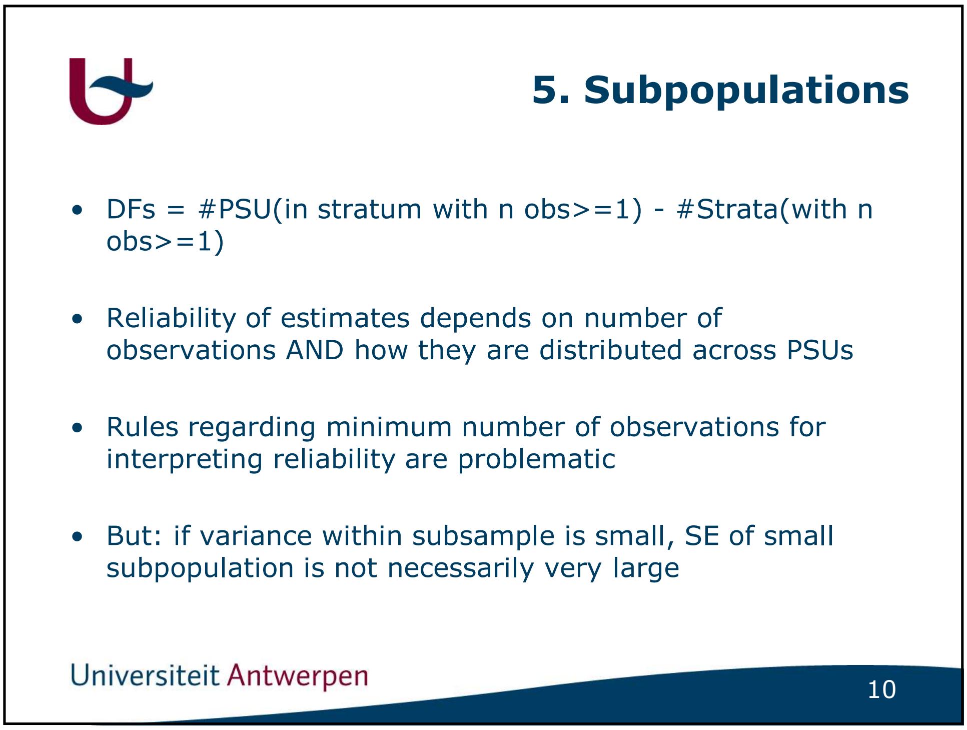
Universiteit Antwerpen

8



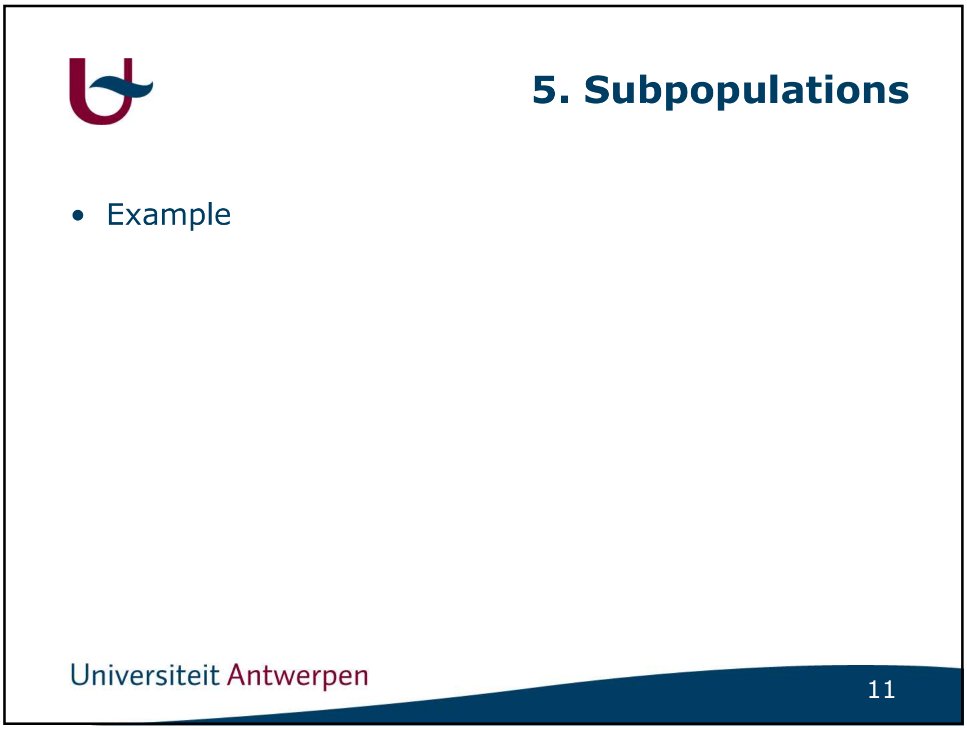
5. Subpopulations

- Create indicator variable (0,1)
- Use the svy, subpop option (Stata)
  - E.g. svy, subpop(child==1): prop poor
- Or use the over option (Stata)
  - E.g. svy: prop poor, over(agegroups)



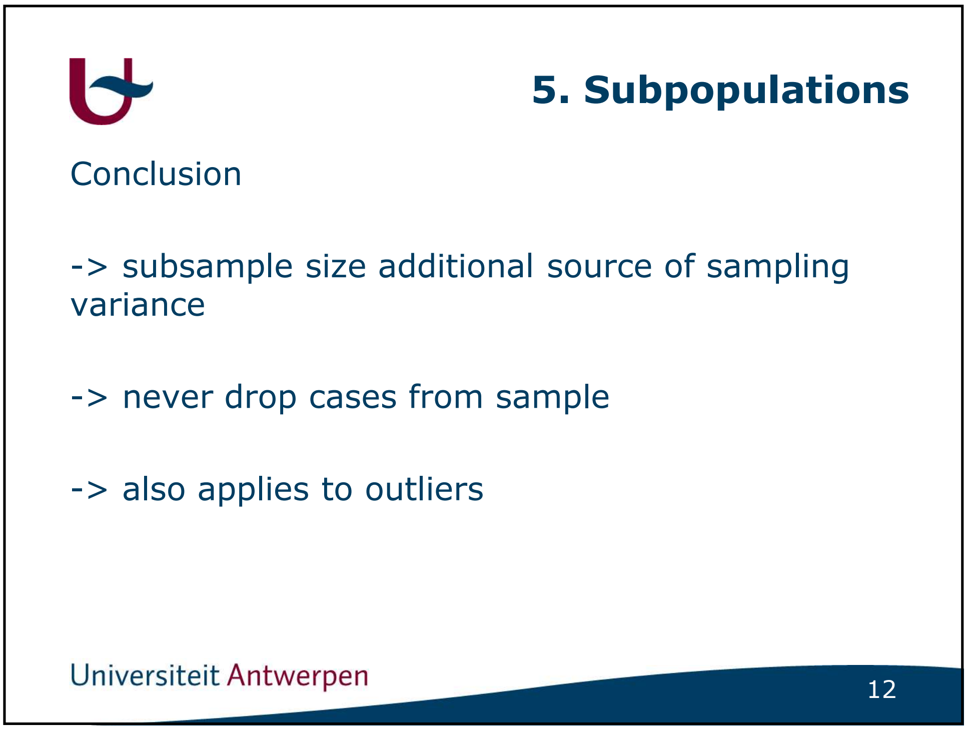
5. Subpopulations

- DFs = #PSU(in stratum with n obs>=1) - #Strata(with n obs>=1)
- Reliability of estimates depends on number of observations AND how they are distributed across PSUs
- Rules regarding minimum number of observations for interpreting reliability are problematic
- But: if variance within subsample is small, SE of small subpopulation is not necessarily very large



5. Subpopulations

- Example



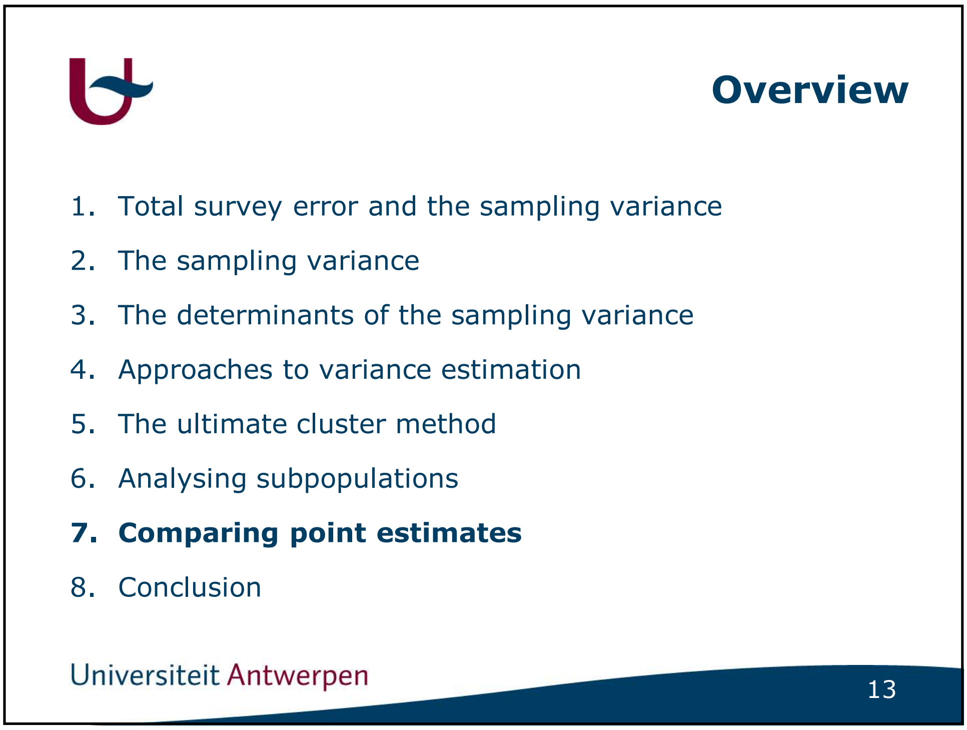
5. Subpopulations

Conclusion

-> subsample size additional source of sampling variance

-> never drop cases from sample

-> also applies to outliers

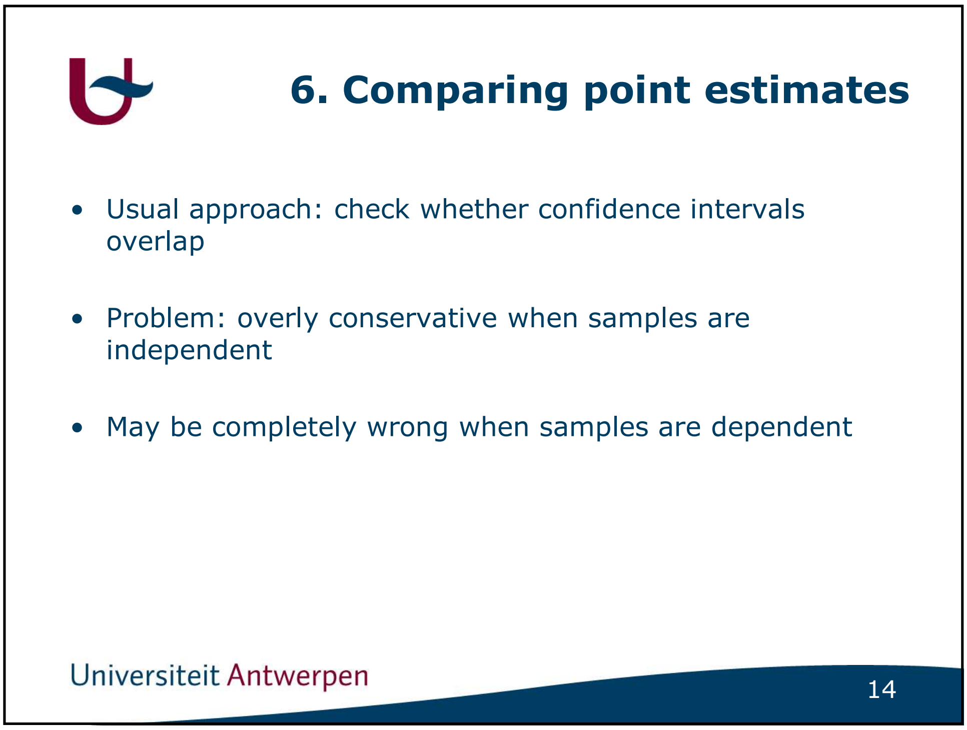


Overview

1. Total survey error and the sampling variance
2. The sampling variance
3. The determinants of the sampling variance
4. Approaches to variance estimation
5. The ultimate cluster method
6. Analysing subpopulations
7. **Comparing point estimates**
8. Conclusion

Universiteit Antwerpen

13

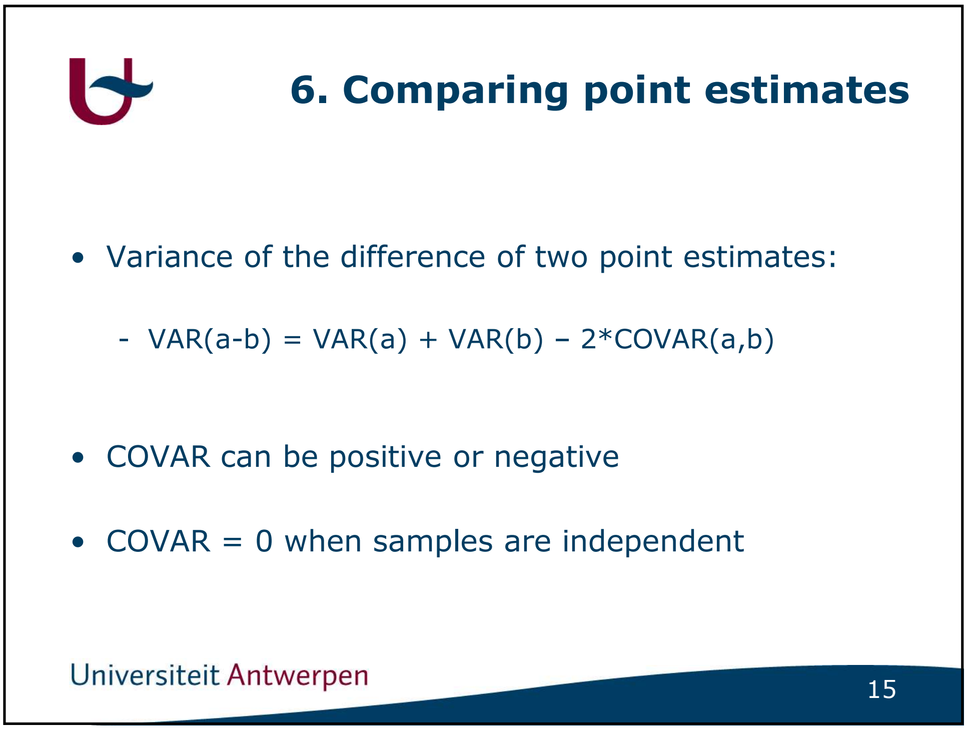


6. Comparing point estimates

- Usual approach: check whether confidence intervals overlap
- Problem: overly conservative when samples are independent
- May be completely wrong when samples are dependent

Universiteit Antwerpen

14

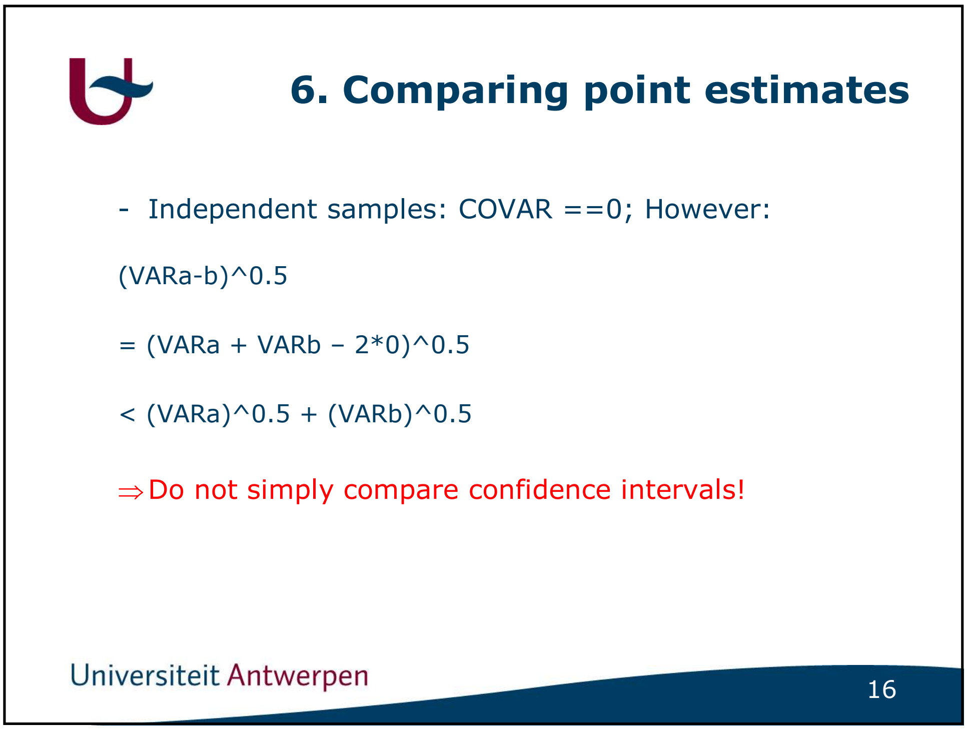


6. Comparing point estimates

- Variance of the difference of two point estimates:
  - VAR(a-b) = VAR(a) + VAR(b) – 2*COVAR(a,b)
- COVAR can be positive or negative
- COVAR = 0 when samples are independent

Universiteit Antwerpen

15



6. Comparing point estimates

- Independent samples: COVAR ==0; However:

(VARa-b)^0.5

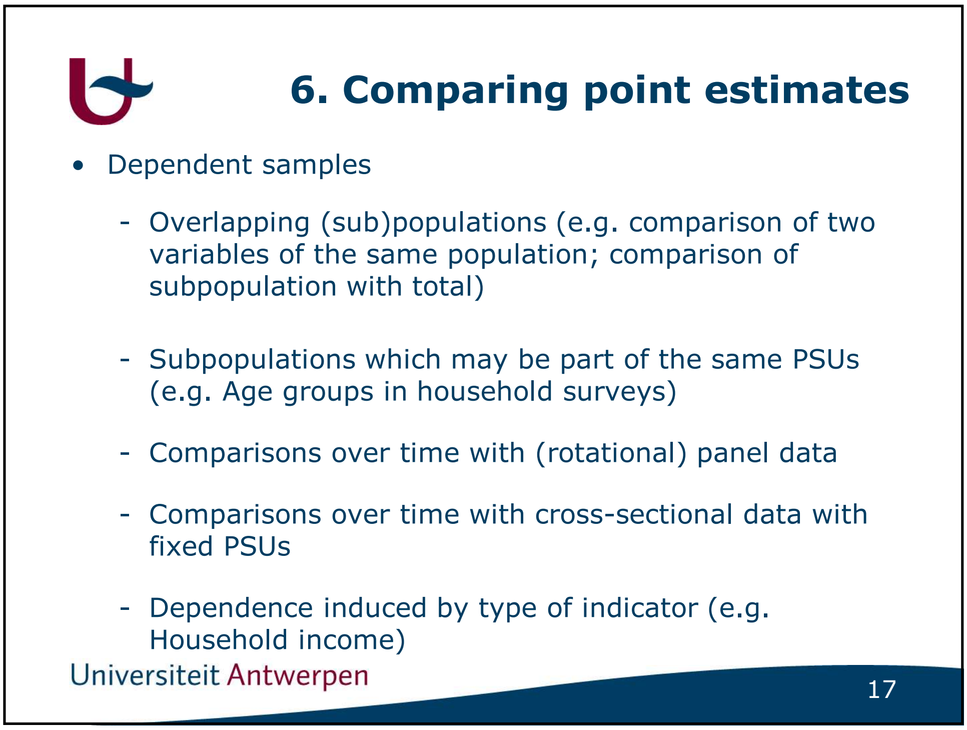
= (VARa + VARb – 2*0)^0.5

< (VARa)^0.5 + (VARb)^0.5

⇒Do not simply compare confidence intervals!

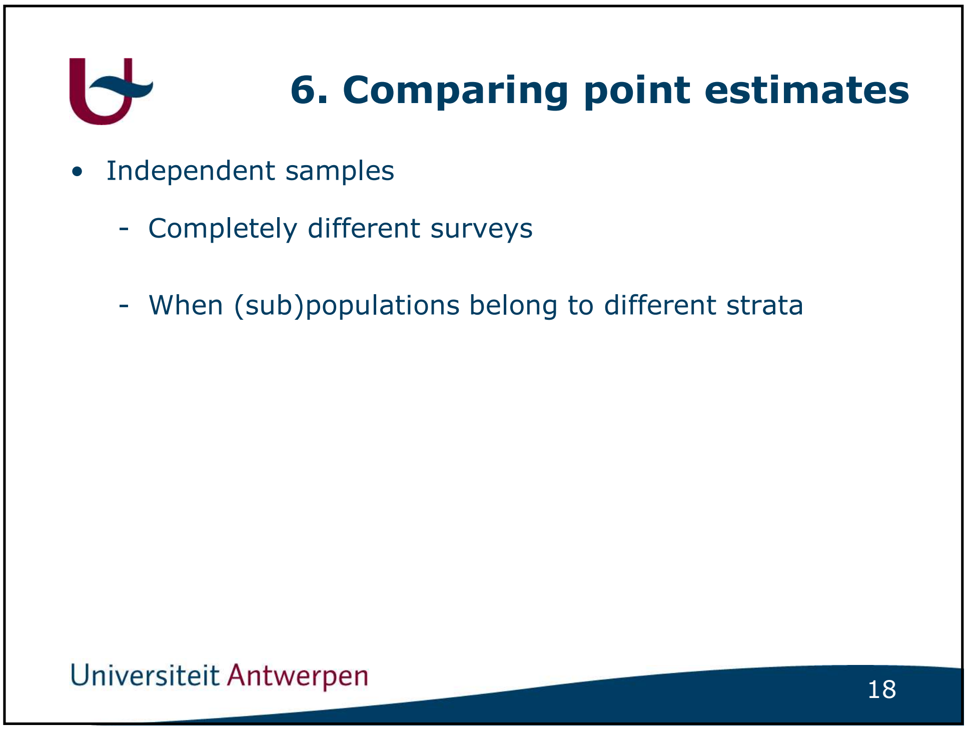
Universiteit Antwerpen

16



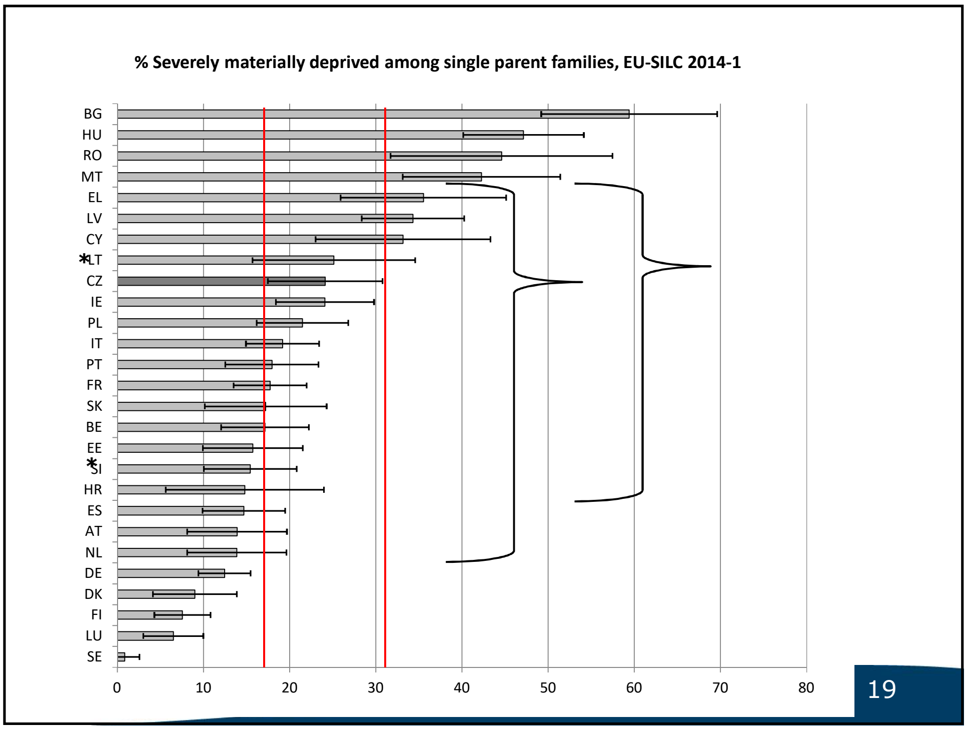
6. Comparing point estimates

- Dependent samples
  - Overlapping (sub)populations (e.g. comparison of two variables of the same population; comparison of subpopulation with total)
  - Subpopulations which may be part of the same PSUs (e.g. Age groups in household surveys)
  - Comparisons over time with (rotational) panel data
  - Comparisons over time with cross-sectional data with fixed PSUs
  - Dependence induced by type of indicator (e.g. Household income)

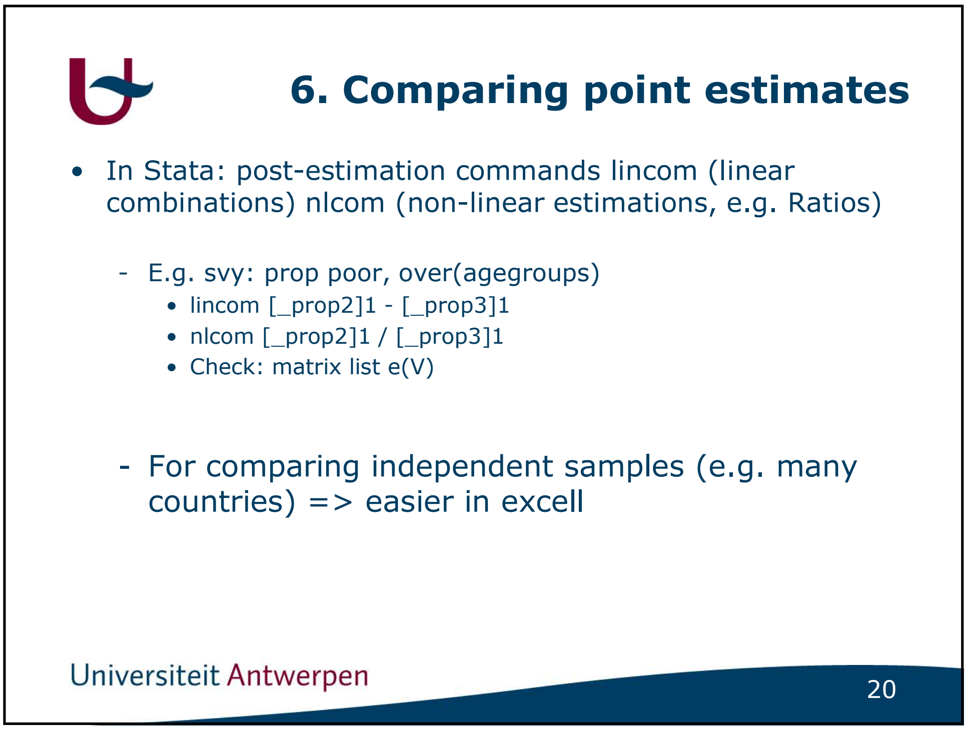


6. Comparing point estimates

- Independent samples
  - Completely different surveys
  - When (sub)populations belong to different strata



% Severely materially deprived among single parent families, EU-SILC 2014-1



6. Comparing point estimates

- In Stata: post-estimation commands lincom (linear combinations) nlcom (non-linear estimations, e.g. Ratios)
  - E.g. svy: prop poor, over(agegroups)
    - lincom [_prop2]1 - [_prop3]1
    - nlcom [_prop2]1 / [_prop3]1
    - Check: matrix list e(V)
  - For comparing independent samples (e.g. many countries) => easier in excell

Universiteit Antwerpen

6. Comparing point estimates

- An example, taken from

Goedemé, T., Van den Bosch, K., Salanauskaite, L., & Verbist, G. (2013). Testing the Statistical Significance of Microsimulation Results: A plea. The international journal of microsimulation, 6(3), 50-77.

- Three scenarios for system in LT:
  - Baseline
  - No family benefits
  - Estionan system of family benefits

Key Messages

So, do not fool yourself (and others):

1. If estimates are based on samples -> estimate and report SEs, CIs, p-values, …
2. Always take as much as possible account of sample design when estimating SEs and CIs
3. Never delete observations from the dataset (subpopulation analysis, negative incomes, …)
4. Never simply compare confidence intervals