Weighted multiple - recapture

How to correct for linkage errors?

Daan Zult

In cooperation with Bart Bakker, Peter-Paul de Wolf, Jan van der Laan and Peter van der Heijden

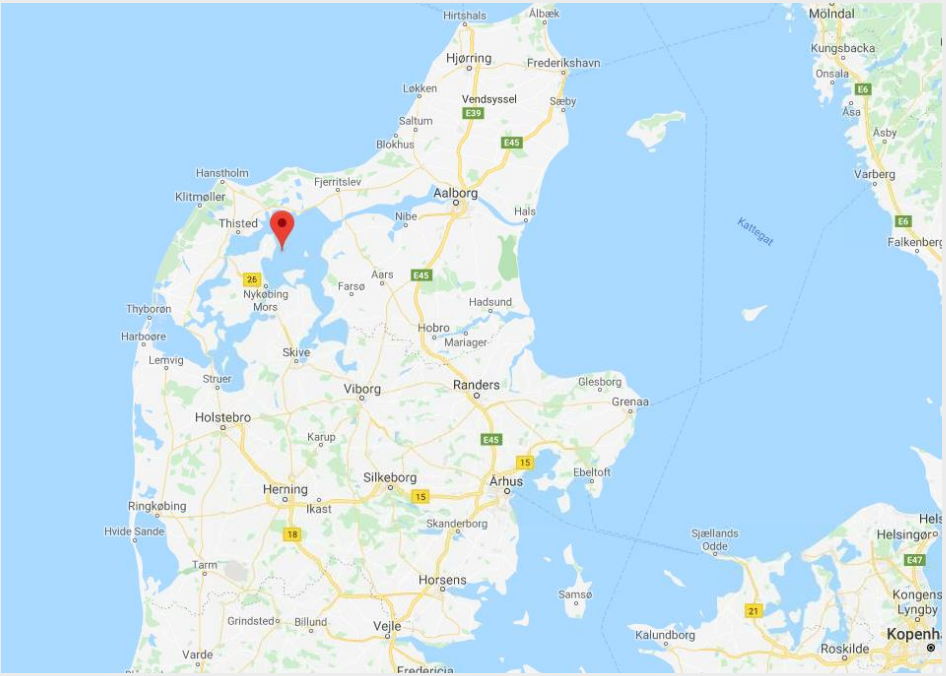
Innovatiedag, Den Haag, November 26

Statistics Netherlands

The problem: How many fish are in the pond?

Classic solution: Capture - recapture

– First applied by Johannes Petersen in 1896 when he was investigating the migration of young plaice (schol in Dutch) into the Limfjord from the German sea (nowadays North Sea).



Simple example

Frequency table

Capture 1	Capture 2	Number of fish
1	1	100
1	0	200
0	1	50
0	0	?

$? = \frac{200*50}{100} = 100$ More general for the total number of fish: $\widehat{N} = \frac{n_{1+} n_{+1}}{n_{11}}$

Or equivalent use **log - linear Poisson regression**, i.e. fit:

$$\text{Number of fish} = \exp(\beta_0 + \beta_1 \text{Capture1} + \beta_2 \text{Capture2})$$

$$? = \exp(\beta_0)$$

Advantage: easy to **add captures and covariates**.

Example of linkage errors

Petersen made small holes in the fins of the plaice to mark them.
Problem: hard to see -> linkage errors

- A hole may be missed (missed match)
- A natural hole may be identified as a mark (mismatch)

Capture 1	Capture 2	Number of fish, real	Number of fish, observed
1	1	100	90
1	0	200	210
0	1	50	60
0	0	?	?*

$$?^* = \frac{210 * 60}{90} = 140 \neq 100$$

Our problem: How many people are in the Netherlands?

- Captures are registers
- **Multiple registers** due to register dependence
- **Use of covariates** (age, sex, etc.) due to different capture probabilities
- **Linkage errors** due to wrong or missing information

A linkage error correction method.
by Ding & Fienberg (1994) and Di Consiglio and Tuoto (2015)

- Idea: Use small audit sample and apply both probabilistic and deterministic linkage.

  - Calculate probability of missed match (α)

  - Calculate probability of mismatch (β)

  - Use α and β to correct population size estimate.

Three problems

1. Very complex, hard to grasp
2. Does not consider covariates
3. Can only be applied with two captures

– Step 1: Simplify
  - From pages of formulas to: $\widehat{N}_{corrected} = \frac{n_{1+}n_{+1}}{\boldsymbol{E}[\boldsymbol{n_{11}}]}$

Step 2: Add covariates

Audit sample:

C1	C2	x	n^*	m^*
1	1	1	n^*_{111}	m^*_{111}
1	0	1	n^*_{101}	m^*_{101}
0	1	1	n^*_{011}	m^*_{011}
1	1	0	n^*_{110}	m^*_{110}
1	0	0	n^*_{100}	m^*_{100}
0	1	0	n^*_{010}	m^*_{010}

Frequency table

C1	C2	x	n	$\hat{m} = E[n]$
1	1	1	n_{111}	$n_{111}m^*_{111}/n^*_{111}$
1	0	1	n_{101}	$n_{101}m^*_{101}/n^*_{101}$
0	1	1	n_{011}	$n_{011}m^*_{011}/n^*_{011}$
1	1	0	n_{110}	$n_{110}m^*_{110}/n^*_{110}$
1	0	0	n_{100}	$n_{100}m^*_{100}/n^*_{100}$
0	1	0	n_{010}	$n_{010}m^*_{010}/n^*_{010}$

$$\hat{m} = \exp(\beta_0 + \beta_1 \mathrm{C1} + \beta_2 \mathrm{C2} + \boldsymbol{\beta_3 x})$$

Obtain individual weights

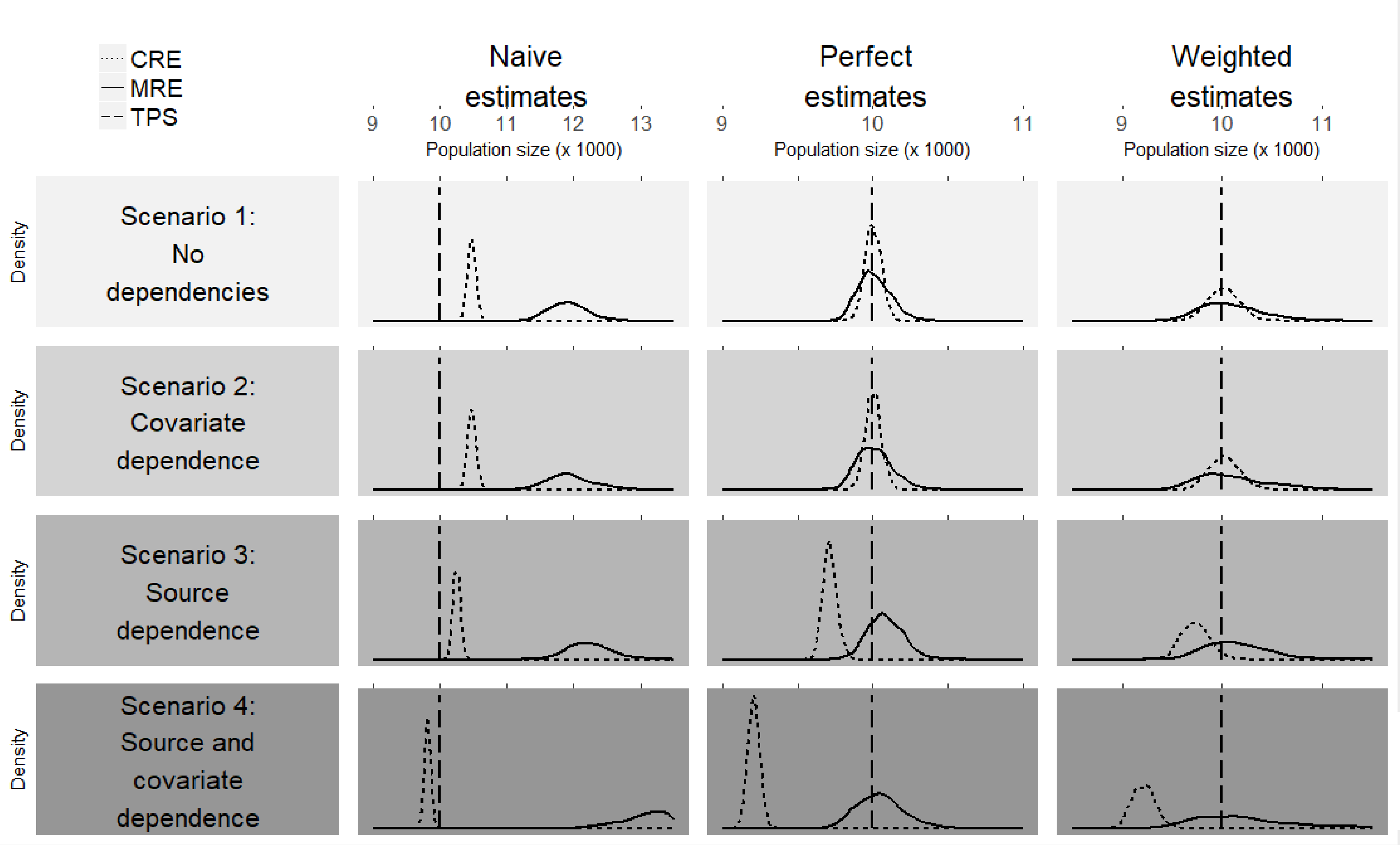
- $w_i = \frac{\hat{m}_{111}}{n_{111}}$
- Aggregate over w_i to get linkage error corrected frequency table.

Step 3: Add captures by updating w_i

- $w_{i,t} = w_{i,t-1} \dfrac{\widehat{m}_{111,t}}{n_{111,t}}$
- $w_{i,t}$ has interpretation of regular sample weight
- Aggregate over $w_{i,t}$ to get linkage error corrected frequency table with multiple captures.
- $\widehat{m} = \exp(\beta_0 + \beta_1 C1 + \beta_2 C2 + \beta_3 C3)$
- $\widehat{m}_{000} = \exp(\beta_0)$

C1	C2	C3	$\widehat{m}$
1	1	1	$\sum_{i\epsilon 111} w_{i,t}$
1	1	0	$\sum_{i\epsilon 110} w_{i,t}$
1	0	1	$\sum_{i\epsilon 101} w_{i,t}$
1	0	0	$\sum_{i\epsilon 100} w_{i,t}$
0	1	1	$\sum_{i\epsilon 011} w_{i,t}$
0	1	0	$\sum_{i\epsilon 010} w_{i,t}$
0	0	1	$\sum_{i\epsilon 001} w_{i,t}$
0	0	0	?

Nice theory, but does it work?
2 models, 3 estimates and 4 scenarios.



Thank you for your attention!

Extensive treatment on this subject can be found at:
https://www.cbs.nl/en-gb/background/2019/19/correcting-for-linkage-errors-in-the-multiple-capture

Any further questions?

Contact information: db.zult@cbs.nl