Simple Random Samples vs. Cluster Samples

FERSI Training Seminar on Evaluation, 19-21 May 2010, Bern

Steffen Niemann, M.A., Research Associate
s.niemann@bfu.ch – www.bfu.ch

Ready to study ...

- Your targets are defined
- Outcome indicators are found
- Your population is defined

➔ Calculate sample size

➔ Draw your sample

➔ Baseline survey

➔ Follow up survey

➔ Analyze data, calculate tests and confidence intervals

Simple random sample (SRS)

- All elements of the target population are independent and have equal probability to be sampled
- That is the theory, so far...

SRS - Problem

- In most research topics, simple random sampling is hard to realise
- All standard computations estimating population parameters assume SRS
- If the data generating process is not SRS, standard calculations underestimate standard errors

Cluster sampling

- More convenient than SRS, saving time and costs
- Selection of primary sampling units (PSU)
  e.g. hospitals, schools, communities, regions, …
- Select all units = simple cluster sampling
- Random sample of units = two stage cluster sampling

Problem

- Two stage sampling is less efficient than a simple random sample of the same size
- Units in one cluster can be „more similar“ than units between clusters
- Intraclass correlation = ratio of variation between clusters to total variation
- Loss of efficiency can be quantified by the “design effect”

Definition of design effect (deff)

$$DEFF(T) = \frac{\sigma(T)^2_{CS}}{\sigma(T)^2_{SRS}}$$

$\sigma(T)^2_{CS}$ = variance of parameter T under two stage cluster sampling

$\sigma(T)^2_{SRS}$ = variance of parameter T under simple random sampling

DEFT

$$DEFT(T) = \sqrt{DEFF(T)}$$

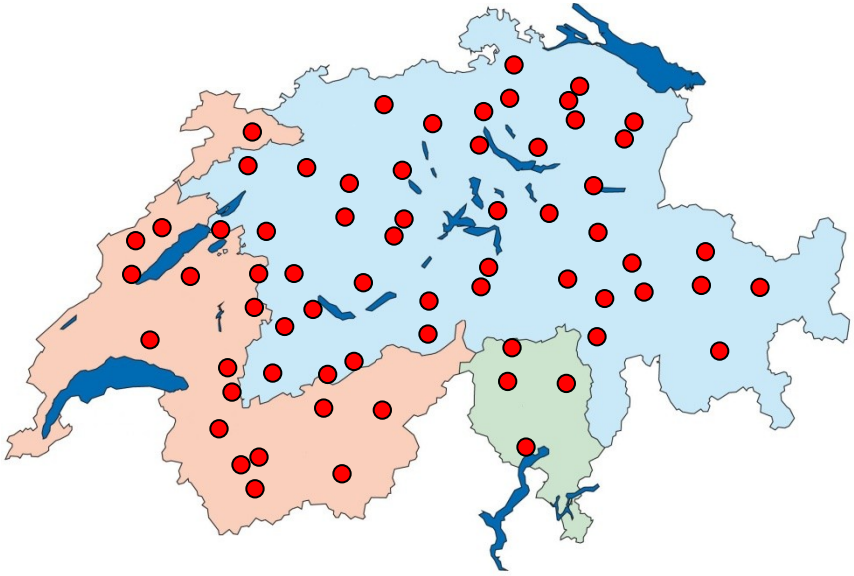
DEFT = ratio of standard errors

Example: bfu-survey on seat belt use

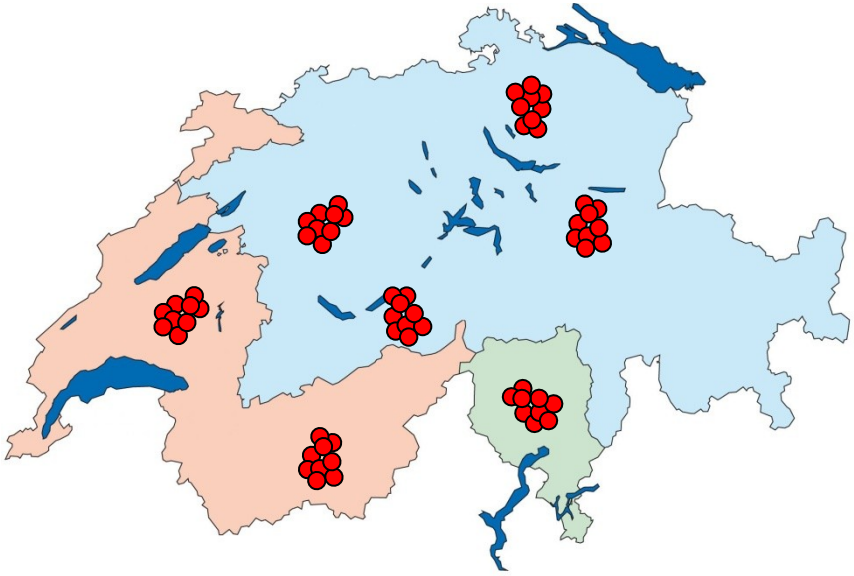
- Stratified, two stage clusterdesign
- Precision of estimates depends on:
  - Observed rate
  - Sample size
  - **Number of observation sites**
  - **Variance between observation sites**

Sampling frame

Simple Random Sample



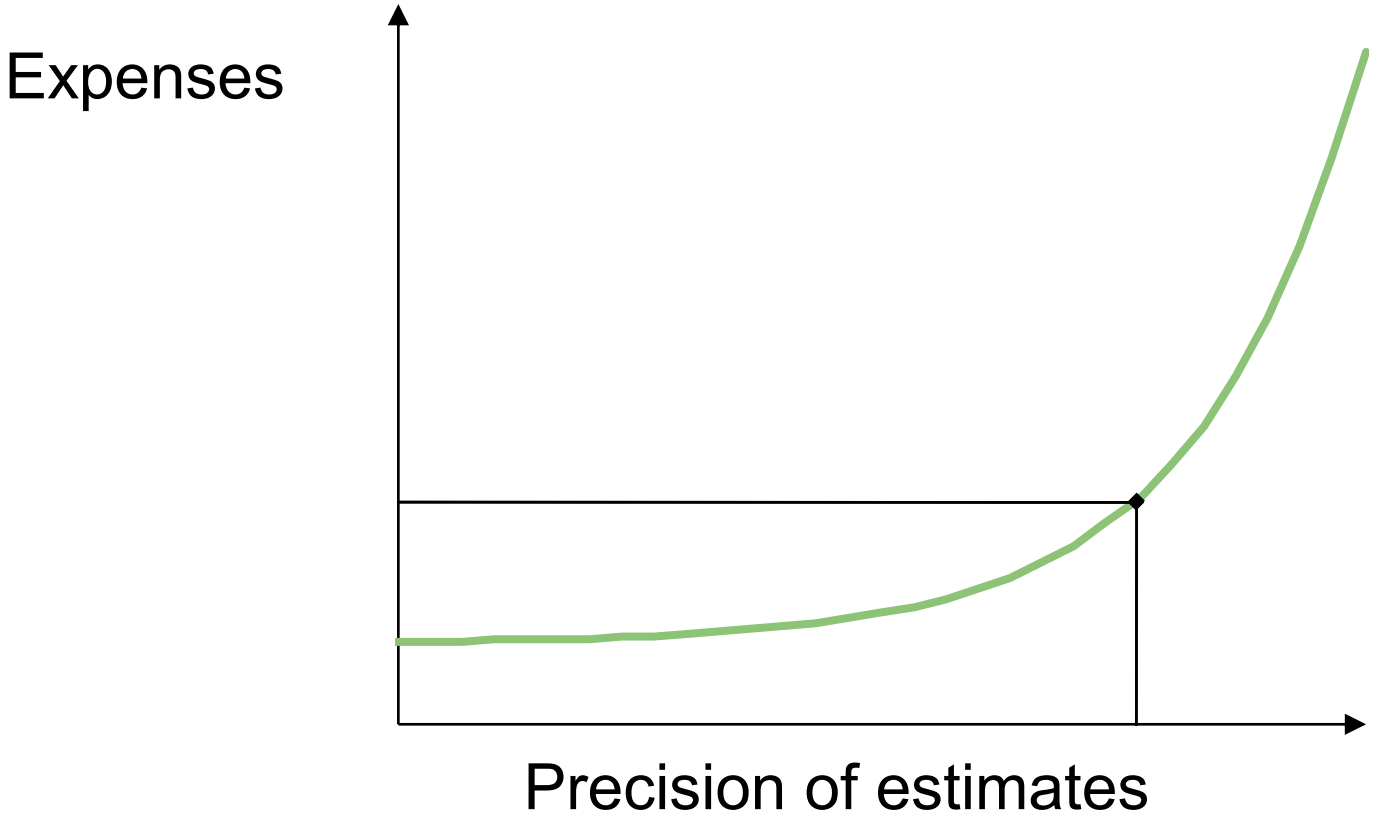
Cluster Sample



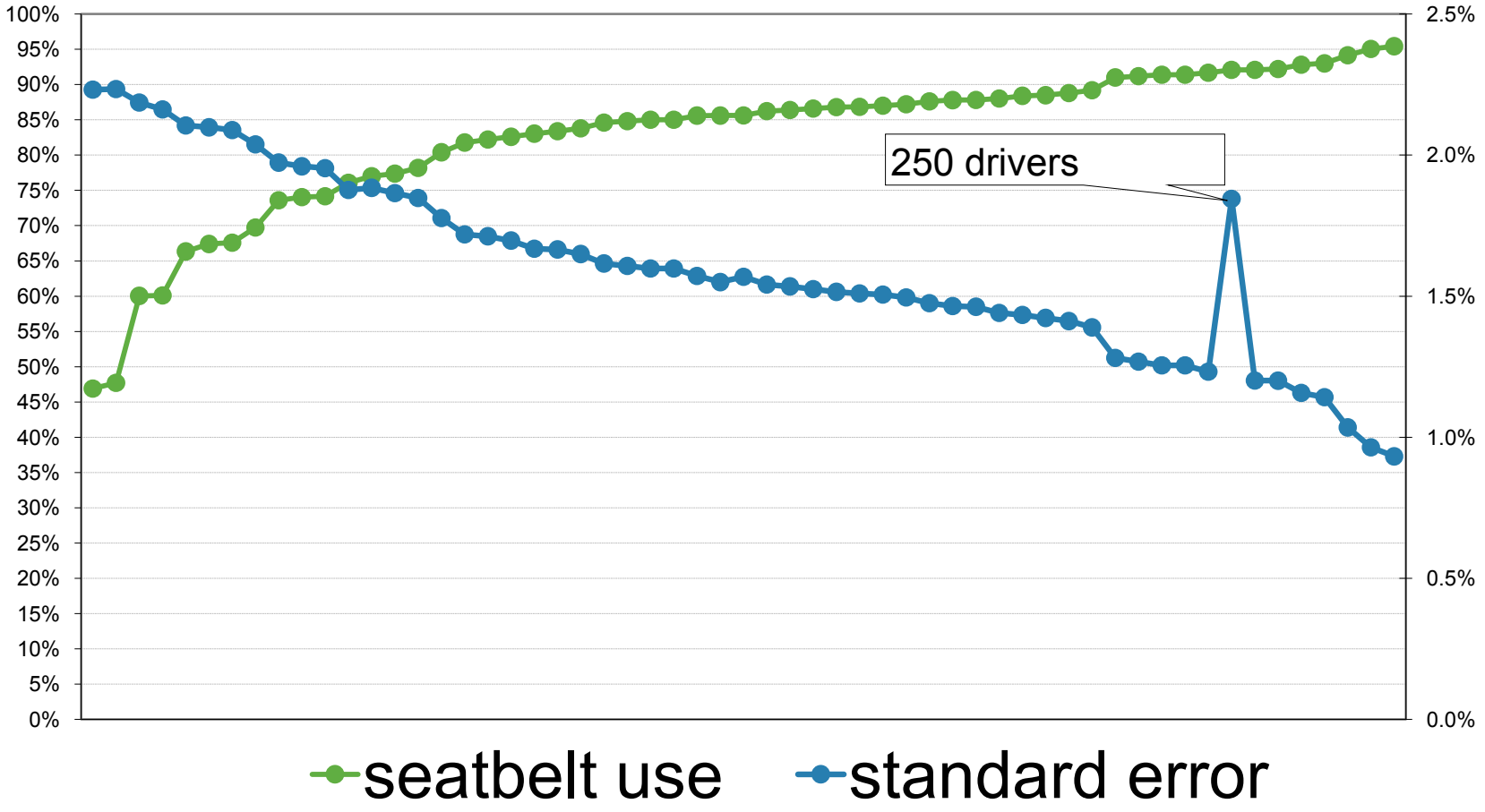
Example: bfu-survey on seat belt use

Observation sites (OS)	59
Total sample size (SS)	30 000
Ratio SS/OS	500
Sampling of sites	Pre-assigned
Observation on sites	Fixed number

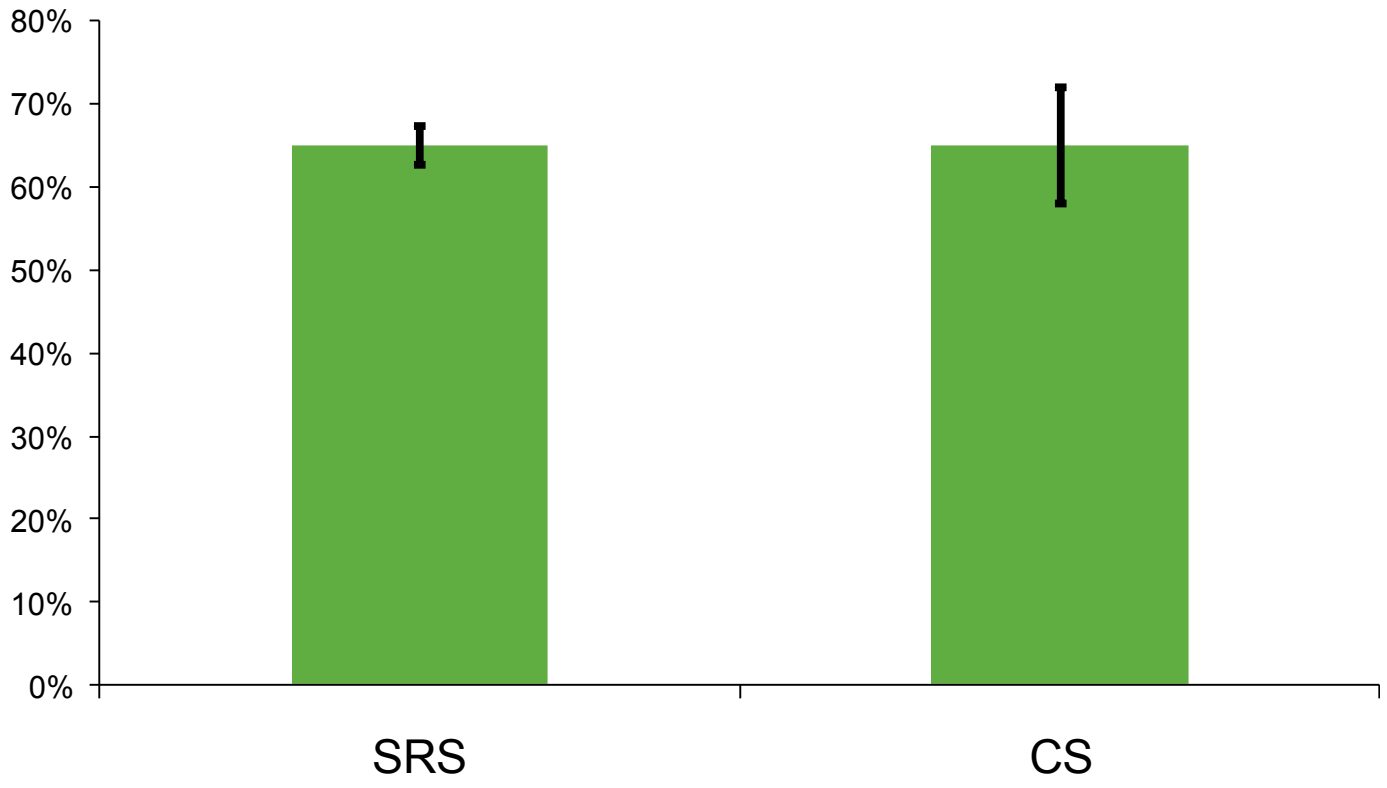
Example: bfu-survey on seat belt use



Estimates and standard errors by observation sites



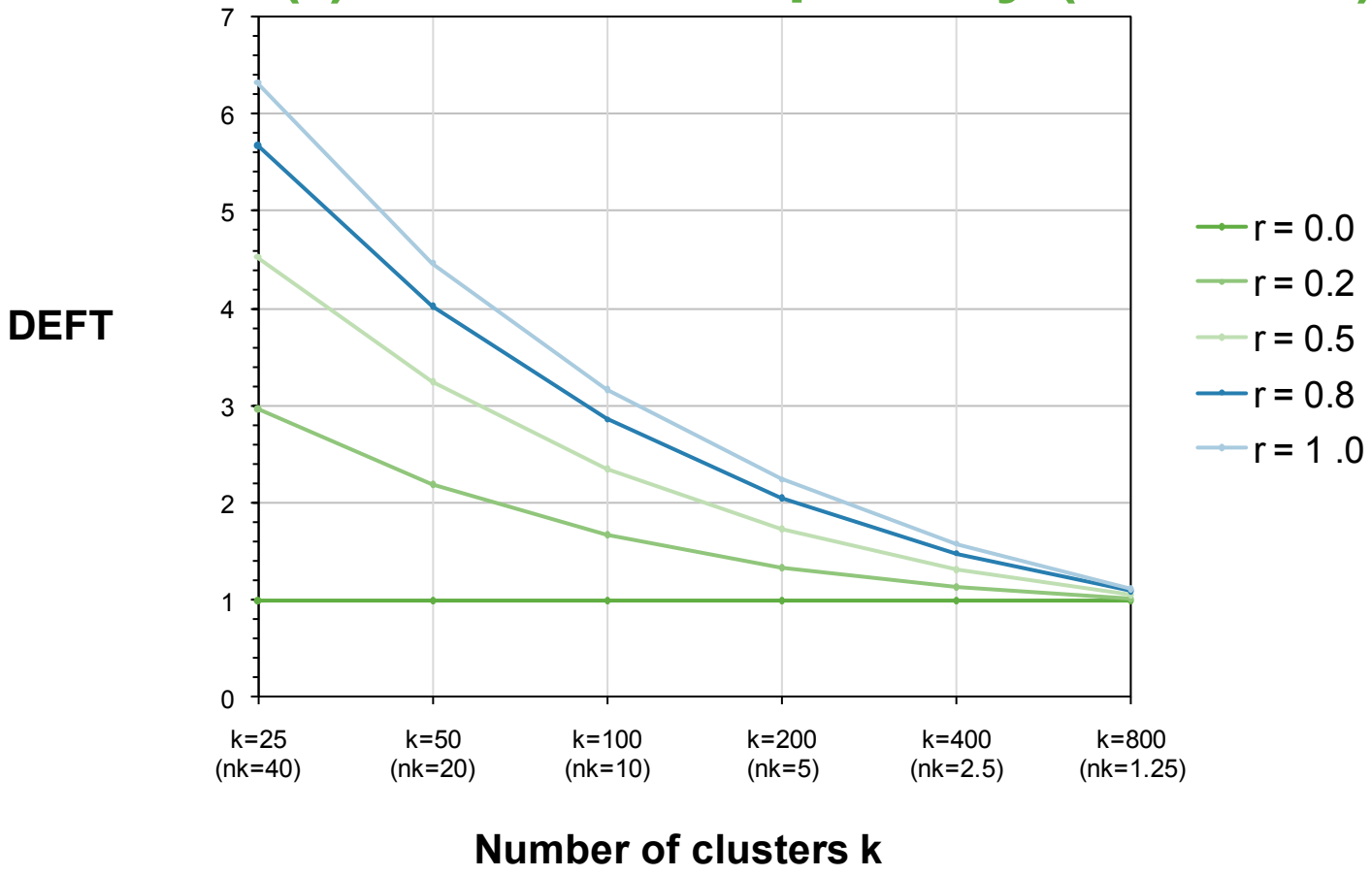
CIs under SRS and CS



Design effects in bfu-survey

Region	Road type	Estimates	Std. Err.	95% Conf. Intervall		Design Effect (DEFT)
German	urban	0.83	0.02	0.80	0.86	4
	rural	0.89	0.01	0.87	0.91	2
	motorway	0.92	0.01	0.89	0.94	3
French	urban	0.69	0.06	0.58	0.81	7
	rural	0.80	0.03	0.74	0.87	5
	motorway	0.87	0.01	0.85	0.89	1
Italian	urban	0.59	0.07	0.46	0.72	5
	rural	0.78	0.04	0.70	0.86	3
	motorway	0.87	0.00	0.86	0.87	0

Design effect (DEFT) by varying intraclass correlations (r) and cluster quantity (n = 1000)



How to deal with design effects:

- Sample size calculations
  If you don't have an idea about the size of the design effect, conduct a pilot
- Increase number of PSU
- Data analysis
  Use software, that account for complex sample designs
- Do not ignore the design effect. Live with it!