Minimizing sampling error

Preliminary findings of sampling design simulations using survey and administrative data

Gró Einarsdóttir, Ph.D.*

Specialist at Statistics Iceland

Anton Örn Karlsson, Head of unit

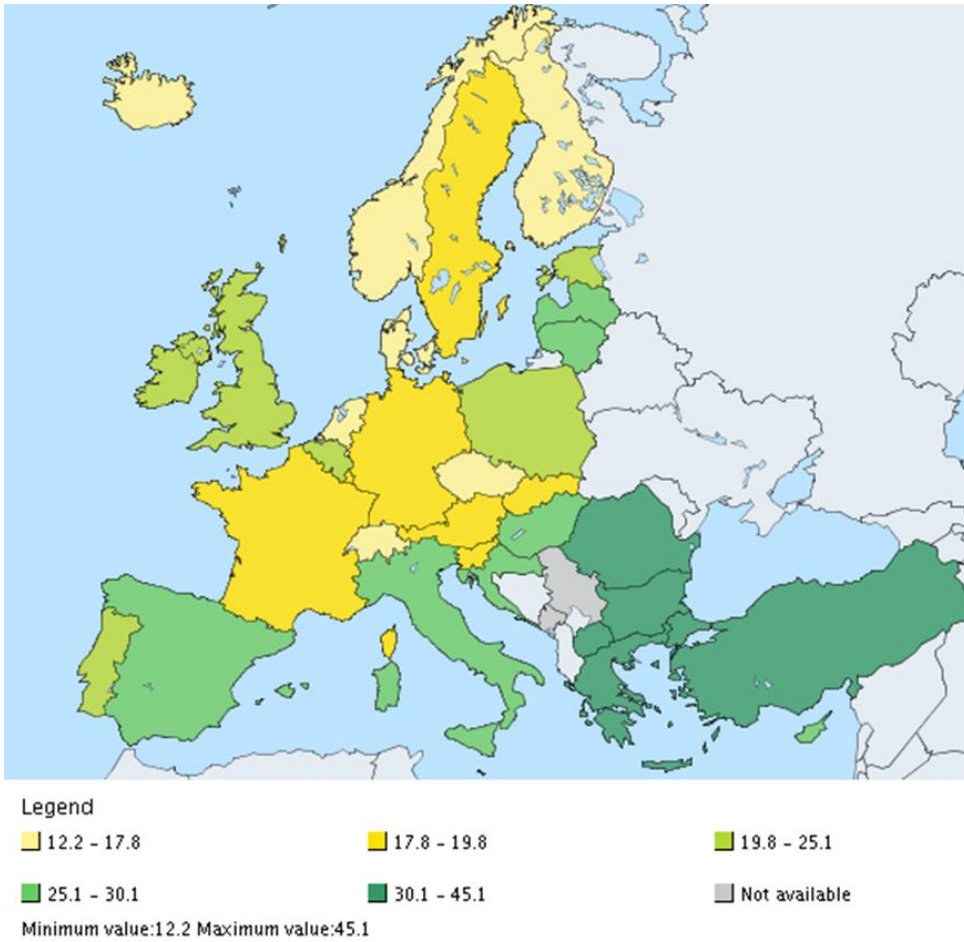
Statistics Iceland

eurostat Statistics Iceland

IS-SILC

- Annual houshold survey on income and living condition
- Eurostat survey conducted in Iceland
- Based on telephone interviews + administrative data
- Small population + small sample = large standard error

At risk of poverty and social exclusion



Standard error

- According to Weisberg, 2005 there are four sources of standard error
- Sample size
  - large sample size = small standard error
- Variance of the estimate
  - Small variance = small standard error
- Sample fraction
  - If the fraction is a large share of the population then small standard error
- Sample design

The four sources of standard error

Cutting SE in half means quadrupled sample size = Too costly

- Sample size
  - large sample size = small standard error
- Variance of the estimate
  - Small variance = small standard error
- Sample fraction
  - If the fraction is a large share of the population then small standard error
- Sample design

The four sources of standard error

- Sample size
  - large sample size = small standard error
- Variance of the estimate
  - Small variance = small standard error
- Sample fraction
  - If the fraction is a large share of the population then small standard error
- Sample design

Eurostat decides the questions so hard to influence variance

The four sources of standard error

- Sample size
  - large sample size = small standard error
- Variance of the estimate
  - Small variance = small standard error
- Sample fraction
  - If the fraction is a large share of the population then small standard error
- Sample design

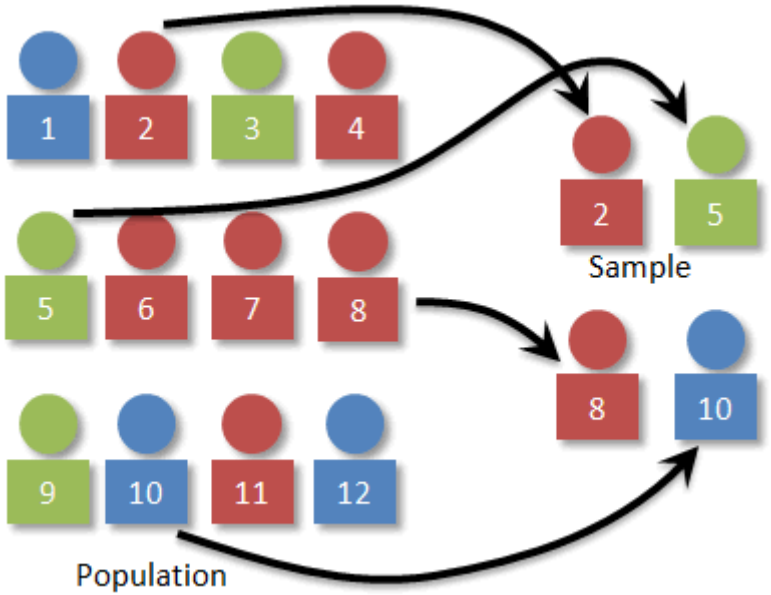
To costly to increase by a very large margin

That leaves: sampling design!!!

Current sample design

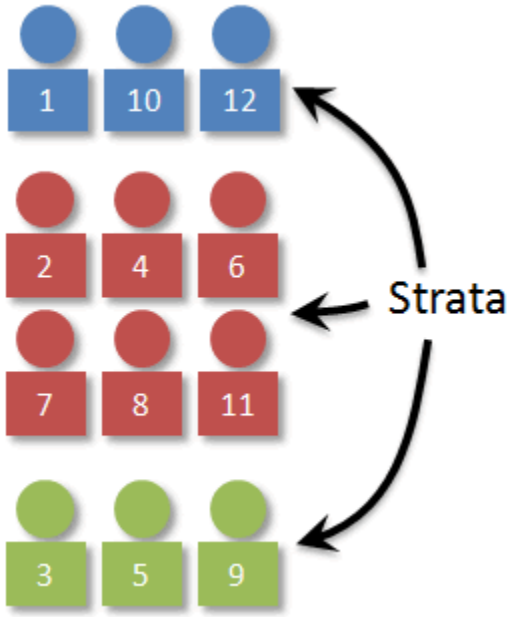
- Multistage random sampling

1. 5.000 housholds a year
2. Random sample of family numbers of individiuals
3. Family numbers used to search for phone numbers
4. Family numbers of individuals used to cluster around the houshold
5. Participants are all those living in the houshold of the chosen individuals

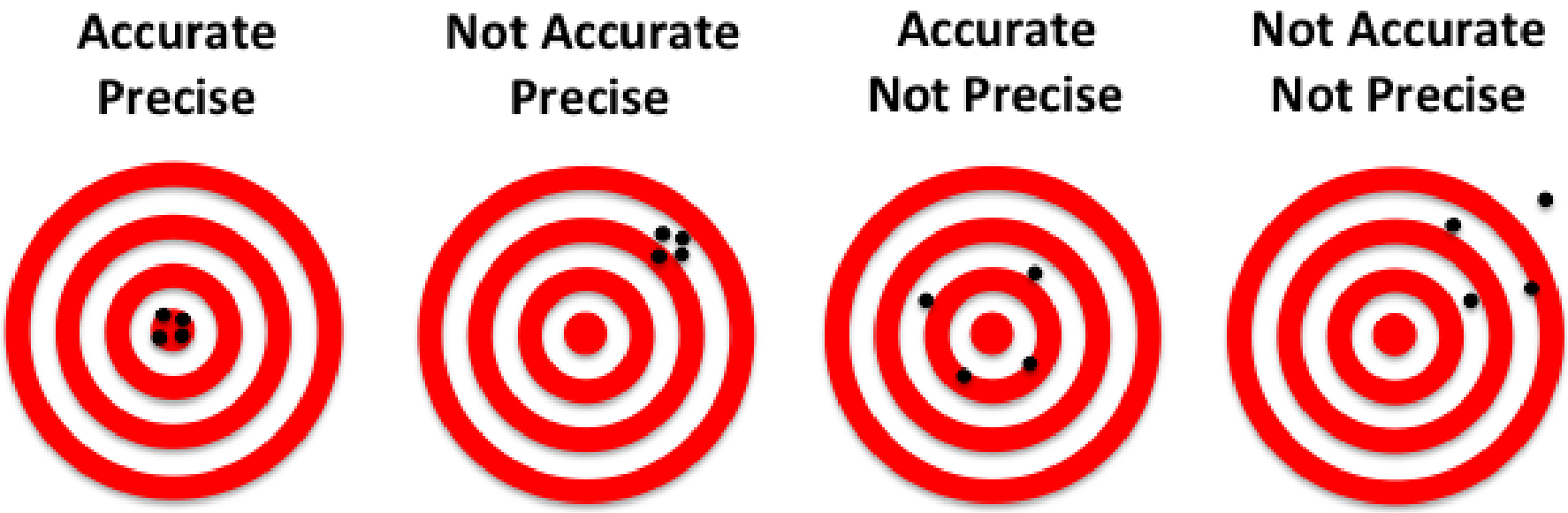


Stratifying samples

- Stratifying samples can reduce sampling error by sampling the right proportion of people
- Categories for stratification must be carefully chosen
- Important to compare stratifyed sample with the population
- Since total populationa are seldom available we use simulations

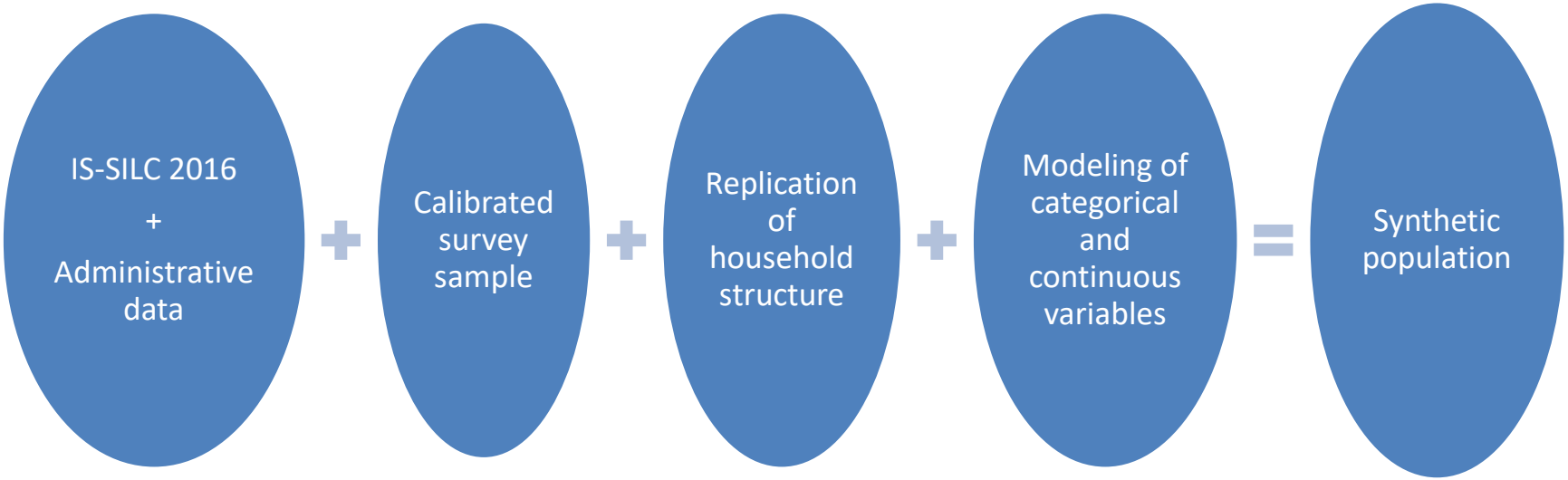


Aim



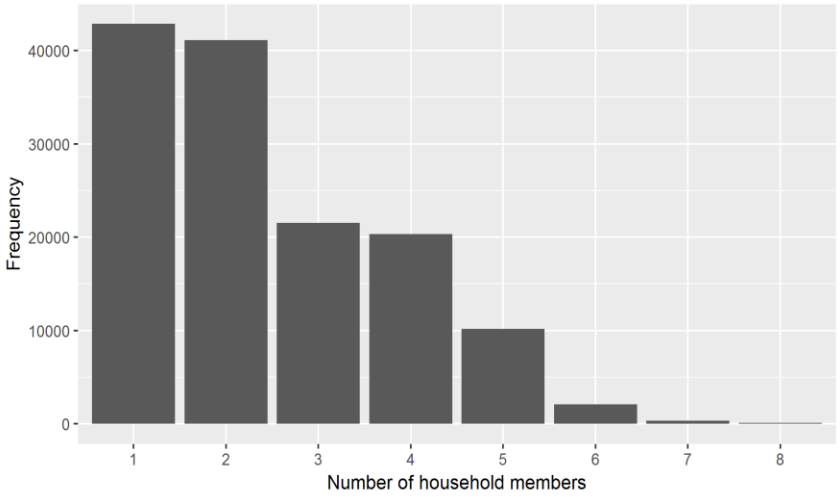
- The aim of the pilot study was to identify which sample stratification lead to most accurate and precise estimates of ability to make ends meet and the rate of overcrowded housing.

The workflow of creating the synthetic population



eurostat Statistics Iceland

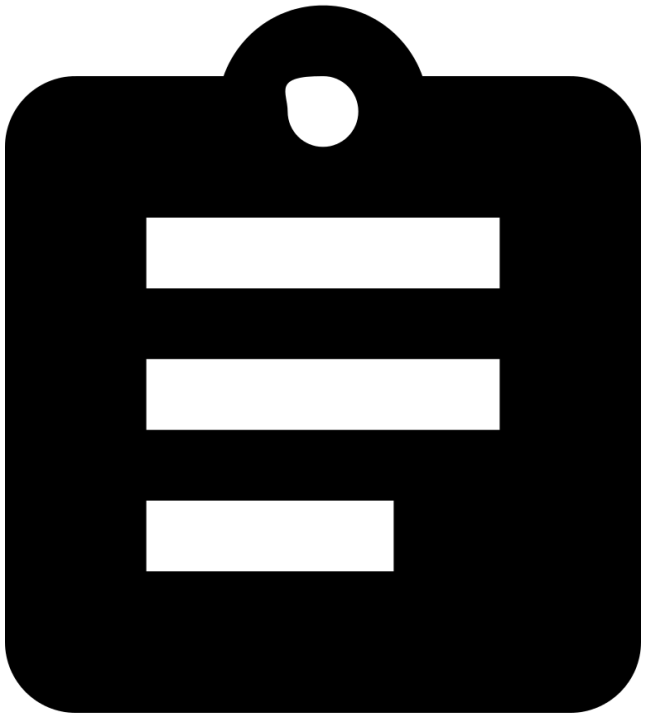
Baseline sample



- The simulated dataset was treated as the complete population
- The complete population was 138,493 housholds
- Then we drew a simple random sample of 5,200 housholds
- This was used as a baseline sample to compare the findings to

Predicting a non-response pattern

- We created a non-response pattern in the data using a binomial regression model predicting response propensity
- We used age, gender and nationality as predictors
- We only chose participants with response probabilities above .6

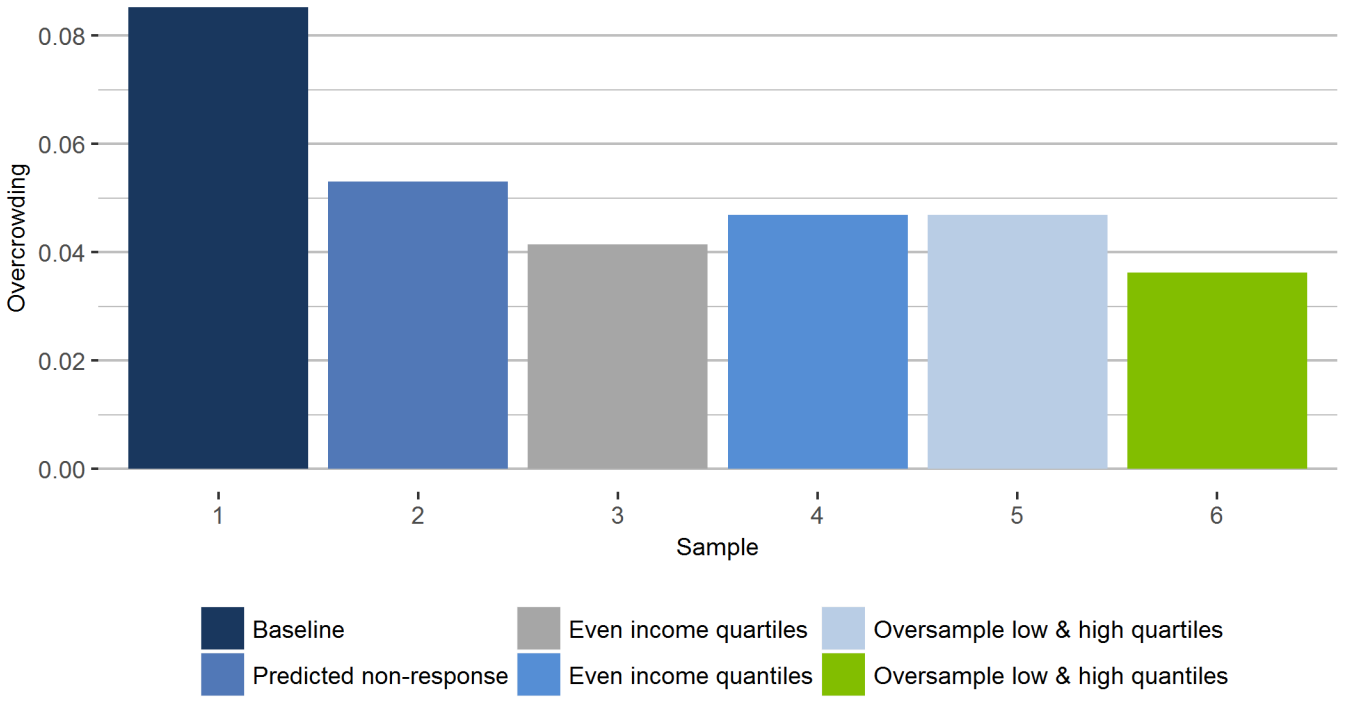


Stratification

- Income distribution used to build strata
- Good category
  - Since ability to make ends meet and overcrowding likely linked to income
  - Statistics Iceland has access to the administrative data from the tax registry
- We used income quartiles and quantiles
- We varied the distribution in each strata

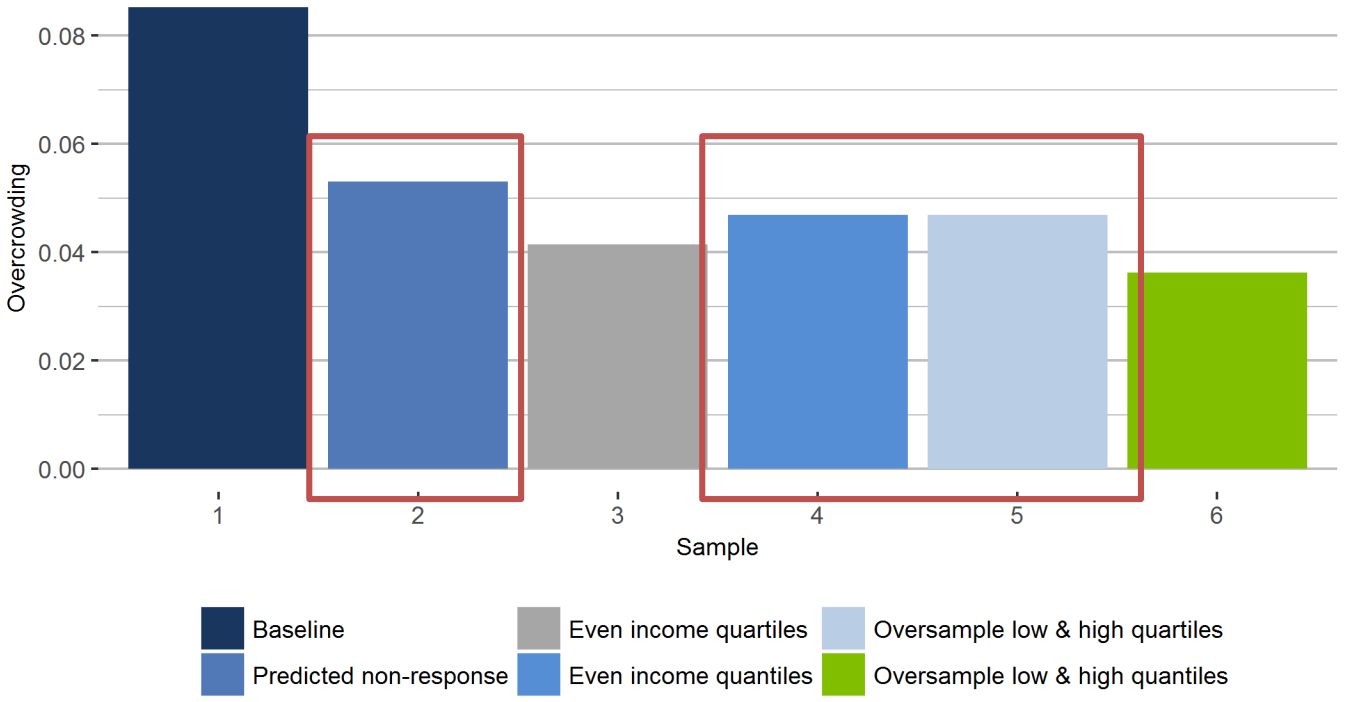
eurostat Statistics Iceland

Accuracy
Overcrowding



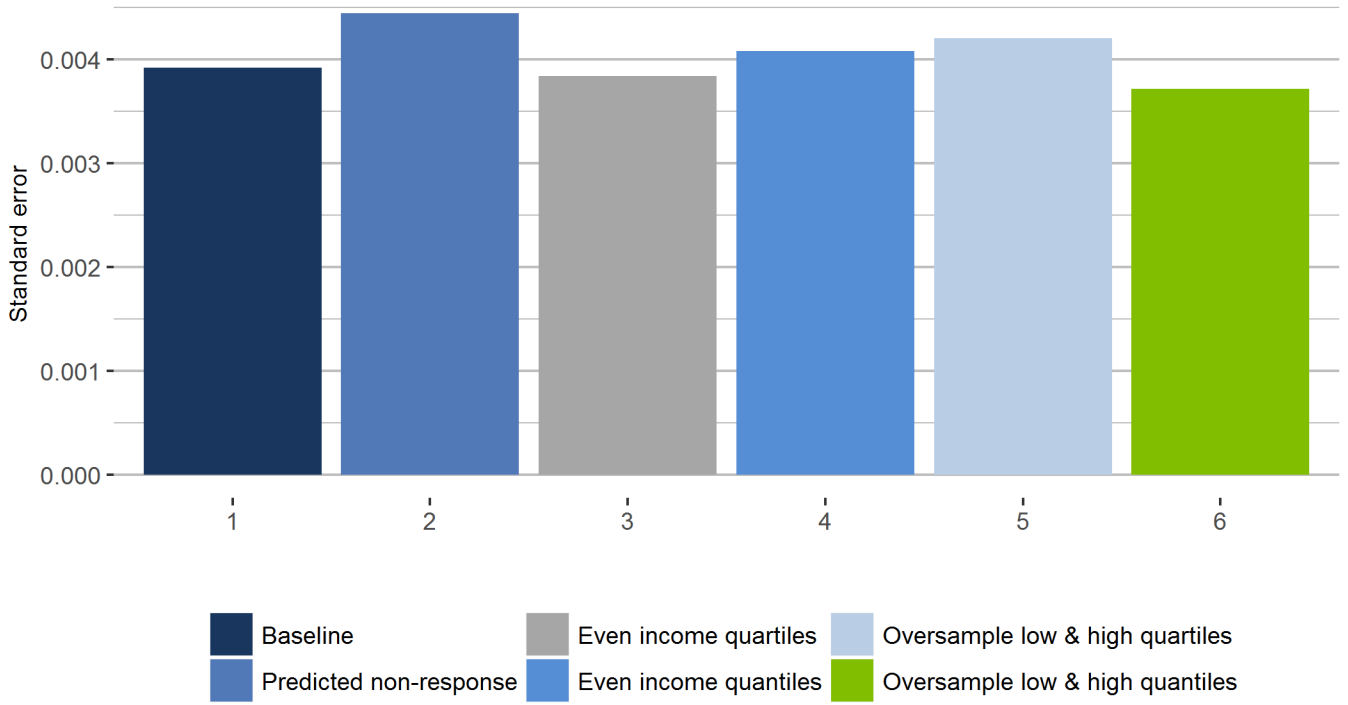
eurostat Statistics Iceland

Accuracy
Overcrowding



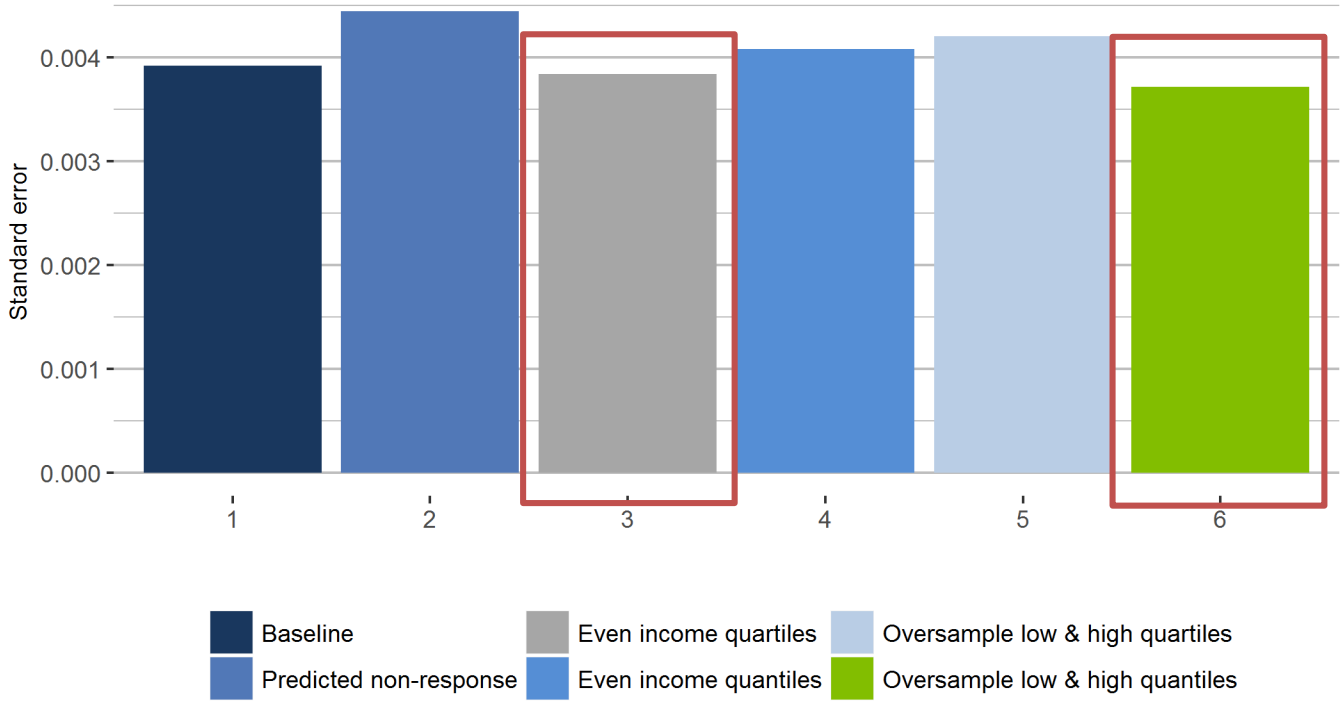
eurostat Statistics Iceland

Precision
Overcrowding



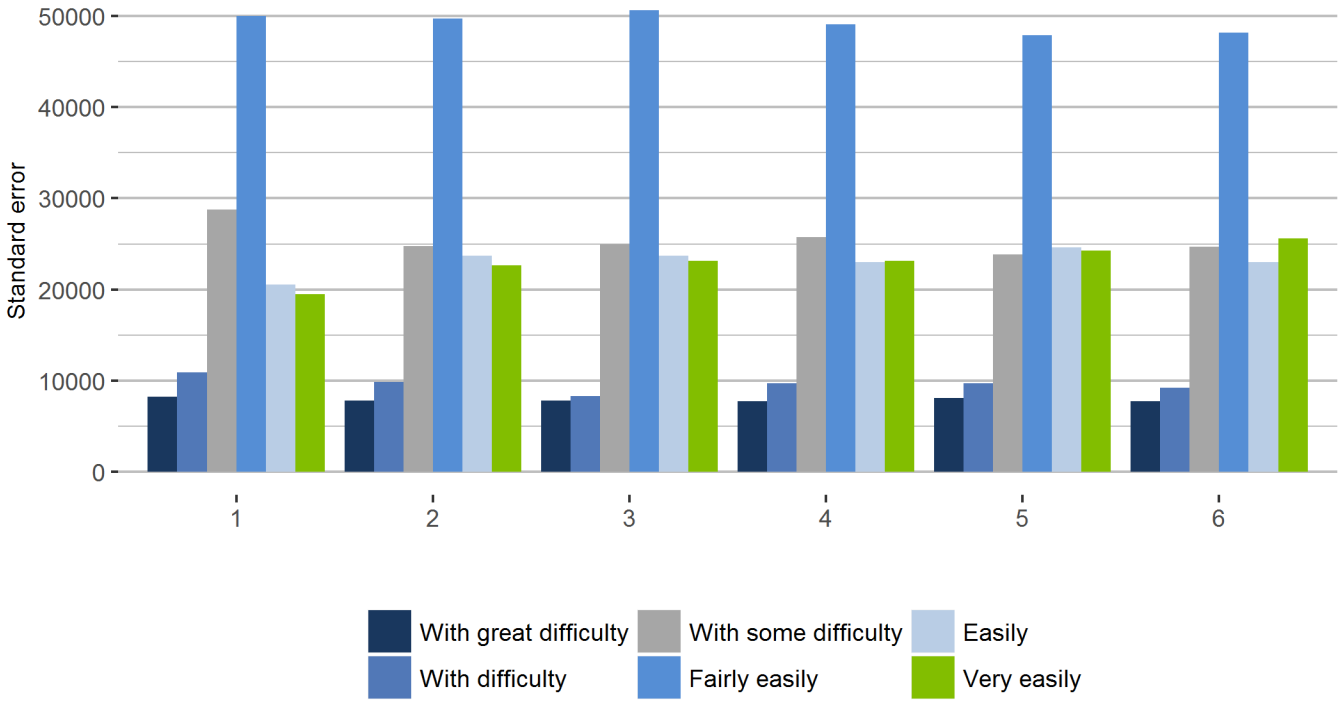
eurostat
Statistics Iceland

Precision
Overcrowding



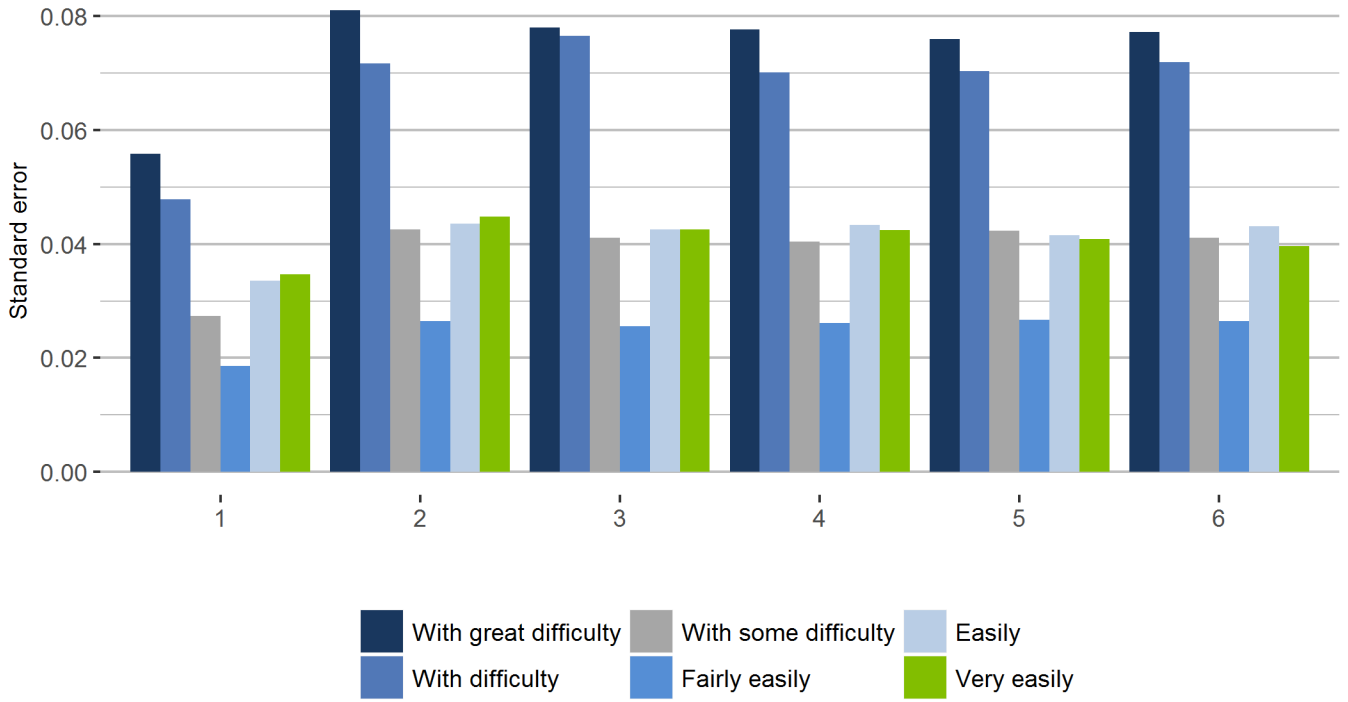
eurostat Statistics Iceland

Accuracy
Ability to make ends meet



1 = "Baseline", 2 = "Predicted non-response", 3= "Even income quantiles", 4= "Even income quartiles", 5= "Oversample low & high quartiles", 6="=Oversample low & high quantiles"

Precision
Ability to make ends meet



1 = "Baseline", 2 = "Predicted non-response", 3= "Even income quantiles", 4= "Even income quartiles", 5= "Oversample low & high quartiles", 6="=Oversample low & high quantiles"

eurostat Statistics Iceland

Next steps

- We intend apply bootstrapping to the stratification samples in order to get a distribution of results instead of the current point estimate
- Test other models of non-response, e.g. add education
- Create and tests more versions of strata
- Test the influence on more variables e.g. at risk of poverty rate

Pre-prepaired questions

- Do any of you use stratification for your sample design and what is your experience of it?
- Do you have any suggestions on how to improve the stratification?