# Survey Methodology


- Lasse Sluth, lbs@dst.dk
- Søren Kühl, ska@dst.dk


STATISTICS DENMARK

# Contents

- Populations
- Stratification and allocation
- Rotating panel designs
- Estimation
- Quality indicators

# Contents

- Populations
- Stratification and allocation
- Rotating panel design
- Estimation
- Quality indicators



STATISTICS DENMARK

# The different populations

- *Target population* is the population the survey seeks to provide information about
  - *eg all enterprises in the Retail industry with an annual turnover of more than 350k Euro*
- *Sampling Frame* is the population from where the sample is drawn
  - *eg a subset of a Statistical Business Register (SBR)*

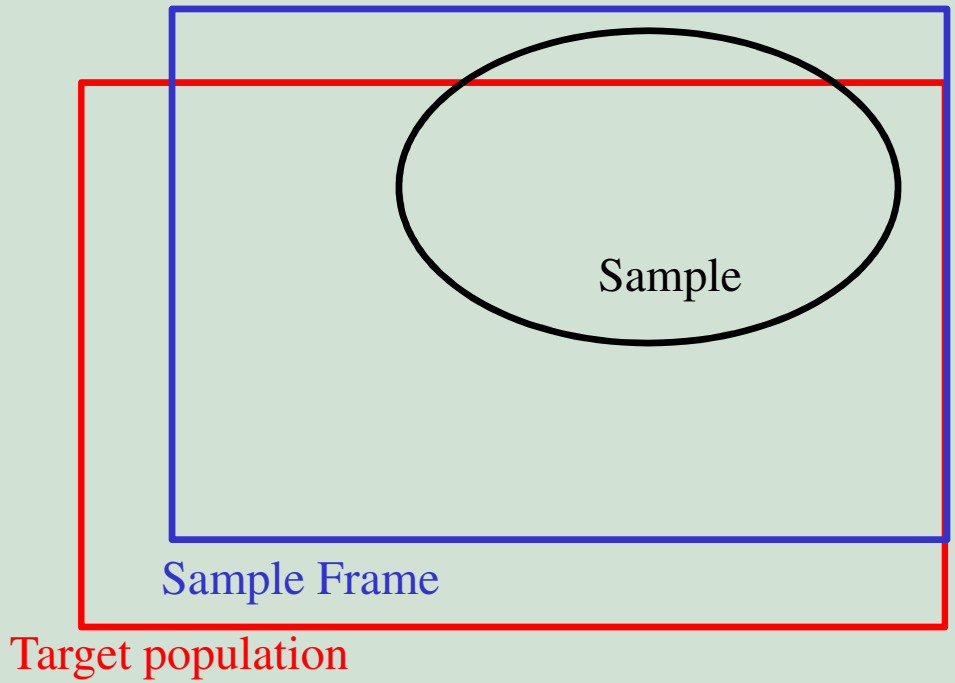


STATISTICS DENMARK

# Frame imperfections

- Coverage problems, both under- and overcoverage
- Misclassifications
- Time-lags
- Variables that could be used for size-measures? eg. employment and/or turnover

# Illustration of Coverage errors




- If the sample is probalistic and representative then it's possible to asses the overcoverage
- Undercoverage is a more serious problem
- Both will lead to bias if not attended to



STATISTICS DENMARK

# From Danish RTI

Target population:

- Retail trade enterprises with more than 2.500.000 DKK yearly turnover.

Frame:

- From SBR all active retail trade enterprises are drawn – with monthly turnover from VAT.
- Yearly turnover is grossed up (missing months)
- Yearly turnover is used for cut-off (and later for stratification)

# From Danish RTI

- Coverage issues:
  - SBR is synchronized with CBR (Central Business Register)
  - Enterprises can change their code of activity in CBR -> SBR
  - Misclassifications
    - Cause both overcoverage and undercoverage
    - Can cause bias in the estimation
- Time-lag from time of drawing the frame to actual use:
  - Births and deaths
  - Corrections (activity, turnover)

# Contents

- Populations
- Stratification and allocation
- Rotating panel design
- Estimation
- Quality indicators

# Stratification and allocation

Focus will be on the most commonly used design in business surveys - *stratified simple random sampling (STSI)*

- *Use classification variables (eg. size and industry) to divide the frame into disjoint subsets called strata*
- *Choose a simple random sample from each strata*

Given a total sample size *n* - How large should the sample size in each strata be?

# Stratification and allocation

## Proportional allocation:

The sample size *n* is distributed among strata propotional with the number of enterprises in the *H* strata:

$$n_i = n \cdot \frac{N_i}{\sum_{h=1}^{H} N_h}$$

# Stratification and allocation

Optimal allocation (Neymann allocation).

The sample size *n* is distributed among strata proportional with both the number of enterprises and standard deviation of the strata

$$n_i = n \cdot \frac{N_i \cdot S_i}{\sum_{h=1}^{H} N_h \cdot S_h}$$

# Stratification and allocation

Why is it called optimal?

-because it minimizes the expected sample variance. An important quality indicator

A good stratification ensures that the enterprises respond as homogenousley as possible within strata, such that most of the variance occurs between  - and not with in -  strata.

# Stratification and allocation

The probability that an enterprise will be included in the sample is called *the inclusion probability*

In take-all strata the inclusion probability is 1 for all enterprises

In other strata it will be $^{n_h}/_{N_h}$

If the allocation is done proportional all enterprises will have the same inclusion probability. This is called a balanced design

# Stratification and allocation

In Statistics Denmark allocation is done by weighting different allocations.

- More than one variable – a trade off
- Domains and total – a trade off
- Stratification is done by employment or turnover
- Analysis of where to select all
- Gradual transition based on more years

The gradual transition is important. Otherwise outliers and external events can skew the allocation.

# From Danish RTI

- Sample is renewed yearly by methodology dep.
- Stratification

| Stratum | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Yearly turnover (mill. DKK/year) | <2.5 | 2.5-5 | 5-10 | 10-20 | >20 |

There is an alternative strata 1 (1-5 mill. DKK/year) for certain industries.

# From Danish RTI

- Sample allocation:

| Stratum | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Coverage (% by number of ent.) | 0 | 15-35 | 15-55 | 40-100 | 100 |

- Total coverage is about 35% (3.500 out of 10.000)
- Weighted coverage is larger than 80%



STATISTICS DENMARK

# Contents

- Populations
- Stratification and allocation
- Rotating panel design
- Estimation
- Quality indicators

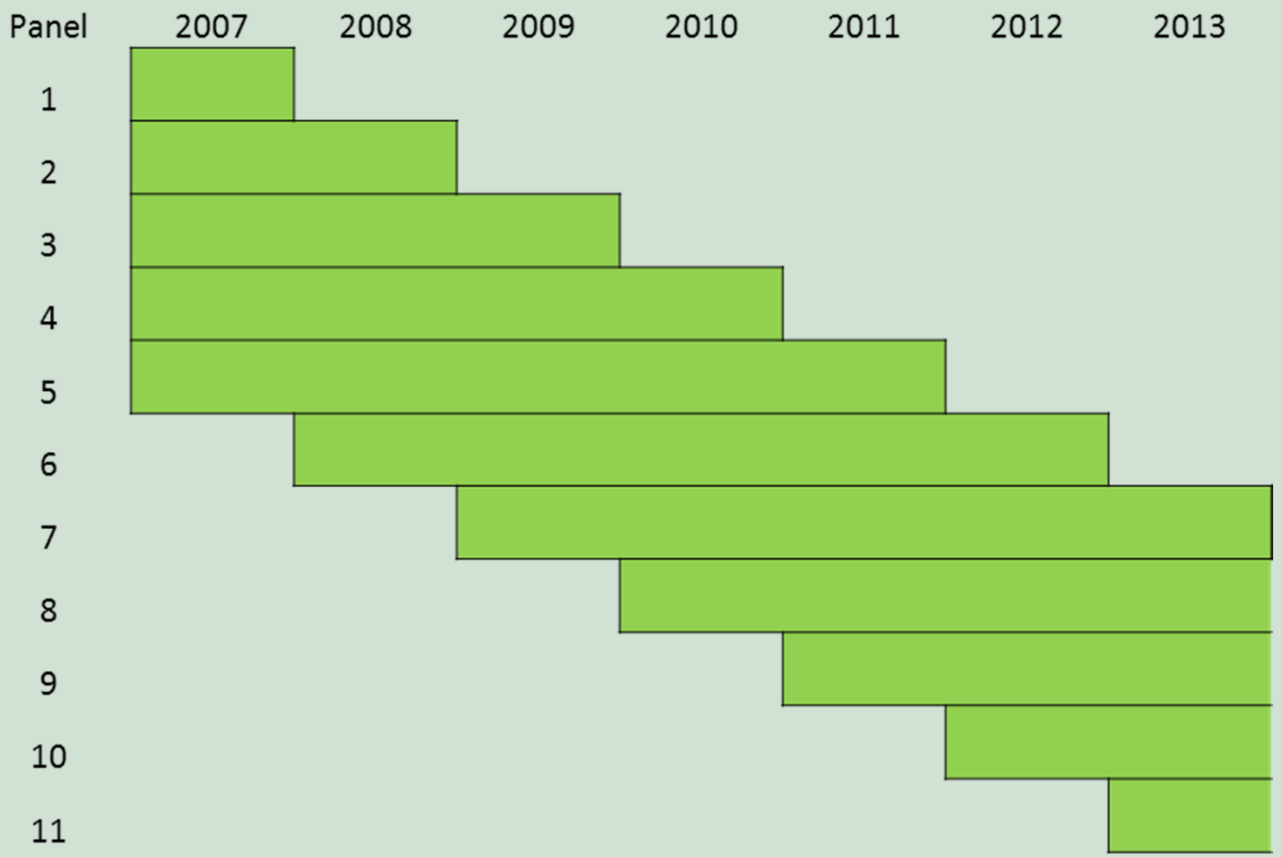
# Rotating panel design

A rotating panel design is a design, where you re-select a pre-determined proportion of the sample from the previous period of time.

Panels are used to optimize estimation of changes, reduce sampling costs and reduce response burdens

The challenge in having such a design is to keep the sample representative.

# Rotating panel designs






STATISTICS DENMARK

# Rotating panel design

Such a rotation scheme is only unbiased if the frame is constant over time.

There are several methods to account for this. One of them is called *the permanent random number technique* and has recently been implemented in danish RTI.

# From Danish RTI

Rotation:

- All enterprises get a random number $r$ between 0 and 1
- And they get a random rotation group number (1,2 or 3)
- When sampling in a stratum, all enterprises with $r$ lower than the inclusion probability $p$ of that given stratum are included in the sample
- Every year the enterprises of one rotation group (starting with group 1) have 1/3 added to $r$.
- New enterprises get a random $r$ and rotation group number.

# From Danish RTI

Consequences:

- In strata with $p$ lower than 1/3, enterprises are in the sample for 3 years and then out for at least 6
- If $p$ is between 1/3 and 2/3, enterprises are in the sample between 3 and 6 six years and out for at least 3.
- If $p$ is larger than 2/3, enterprises are in for more than 6 years but still out for at least 3.

# Contents

- Populations
- Stratification and allocation
- Rotating panel design
- Estimation
- Quality indicators

# Estimation

The usual technique for estimating a population total consists in summing appropiately weigthed variable values for the responding enterprises in a sample

Different weigthing systems can come into consideration

# Estimation

One can use the design weights, given by inverting the inclusion probabilities. This gives the Horwitz-Thompson (HT) estimator.

With a STSI-design every enterprise have equal inclusion probabilities within strata, namely $n_h/N_h$ . If the variable of interest is denoted *y*, the estimator is given by:

$$\hat{Y}_{HT} = \sum_{h \in H} {}^{N_h}\!/_{n_h} \sum_{i \in h \cap S} y_i$$

Which is unbiased if all sampled enterprises respond



STATISTICS DENMARK

# Estimation

Suppose you have an auxiliary variable, *x*, which is known for every enterprise in the frame then the *ratio estimator* is given by

$$\hat{Y}_R = \frac{X}{\hat{X}_{HT}} \hat{Y}_{HT}$$

This leads to a weighting scheme, where each design weight is adjusted with an equal factor. It is possible to do the ratio adjustment per strata or industry. This is often the best solution and is called the *seperate ratio estimator*

This estimator is only effective if the auxiliary variable is correlated with the variable of interest – *eg VAT reports from administrative sources and retail trade turnover*

# Estimation

The ratio estimator is a popular choice in business surveys.

- It is easy to calculate
- If the correlation between the auxiliary variable and the variable of interest is strong it reduces the sample variance
- In case of non-response it decreases bias if the auxiliary variable is correlated with the response probability

# Estimation

In the case of non-response bias will occur, espicially if the respone probability is correlated with the variable of interest

*-eg enterprises with decreasing (or a small amount of) turnover might not find the time to answer the survey*

In that case you have to adjust the weights for non-response.

# Estimation

A simple solution is assuming that every enterprise responds with the same probabilty within strata, and thus modify the design weights proportionally.

More advanced methods use auxiliary information to model the response probalities, and then adjust the design weights accordingly.

*The ratio estimator* is a simple example of this.

# From Danish RTI

- Ratio estimate with grossing up population updated quarterly.
- VAT turnover as auxiliary variable
- Ratio estimation by industries (not strata)
- Imputation only for special units

# From Danish RTI

Month-to-month chain-linking

- Re-estimation of previous month
- Same grossing up population
- Calculation of monthly growth rate
- Eliminates impact of structural changes (in frame, grossing up population, sample)

# Contents

- Populations
- Stratification and allocation
- Rotating panel design
- Estimation
- Quality indicators



STATISTICS DENMARK

# Quality indicators

There exists a vast number of quality indicators for surveys.

Some of them are simple, such as *response rate* or *weighted response rate*

# Quality indicators

Focus will be on 2 important indicators

- Sampling error
- Mean square error

# Quality indicators

If the design is unstratified simple random sampling (SRS) and the estimator is the HT-estimator. Then the sampling error is estimated by:

$$\sqrt{\hat{V}\left(\hat{Y}_{HT}\right)} = N \cdot \sqrt{\left(1 - {}^{n}/_{N}\right)} \cdot {}^{S_y}/_{\sqrt{n}}$$

# Quality indicators

- The quotient between the sampling error and the estimated total is called the coefficient of variation (CV)
- In take-all strata there is no sampling error
- The standard deviations are the only stochastic elements under a fixed design and allocation
- The ratio estimator reduces the standard deviations

# Quality indicators

The sampling error doesn't account for bias.

In fact an estimtate corrected for non-response bias will typically have a higher sampling error than the uncorrected biased estimate

The mean square error is defined as:

$$MSE(\hat{Y}) = sample\ error + (bias)^2$$

The magnitude of bias will often be a jugdement call.

# Quality indicators

- Many more quality indicators are available through ESS-guidelines