L11

Basic statistics for economists

Spring 2020

Department of Statistics

Stockholm University

Today

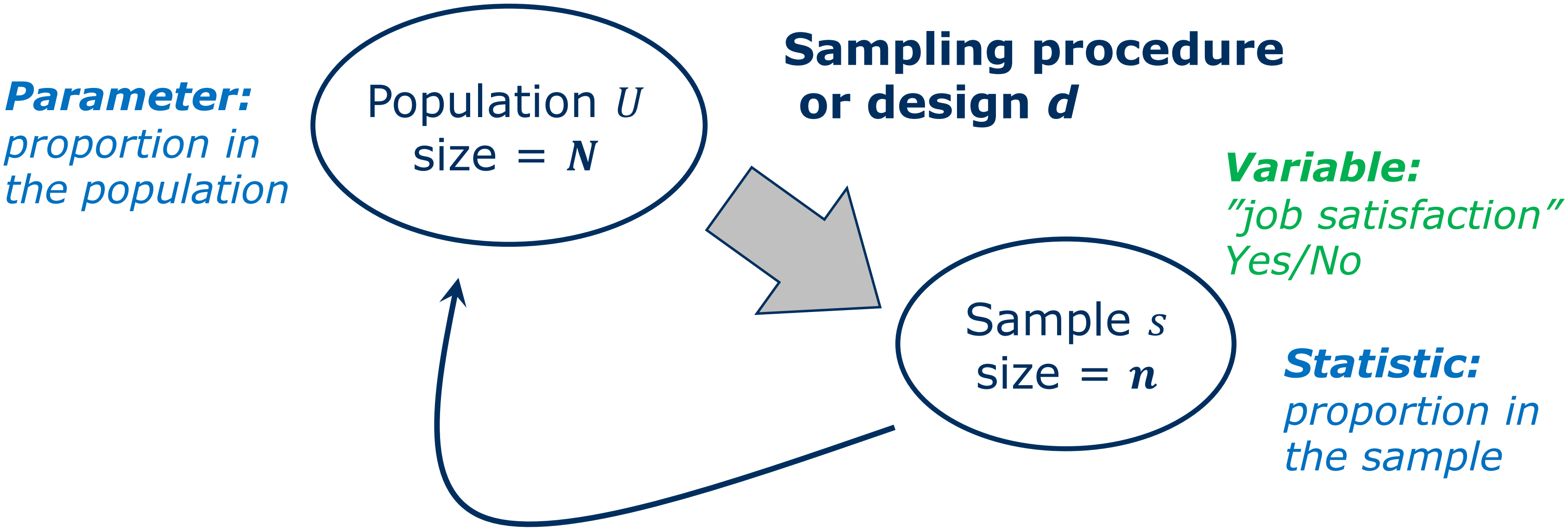
Survey methodology - <u>how</u> to collect data

- specifically, sampling from **finite populations**
- based on course literature **JB**
- examined through **Home assignment 2**

- Random sampling – sampling designs
  - **Simple random sampling (SRS)**
  - **Stratified random sampling (Strat)**
  - A little about **cluster sampling**

- Errors in surveys, collection methods, questionnaires

Inference for <u>finite</u> populations



Inference: to say something about the properties of a finite population using the information from a sample obtained through a (random) process; estimation of **population parameters**.

Stockholm University

Population parameters

Typically calculated in the same as for **descriptive statistics:**

Examples:

- **Mean**: $\boldsymbol{\mu} = \frac{1}{N}\sum_{k=1}^{N} y_k$

Assignment 2B - focus is on μ

- **Total**: $\boldsymbol{\tau} = \sum_{k=1}^{N} y_k = \boldsymbol{N\mu}$

- **Proportion**: $\boldsymbol{P} = \frac{1}{N}\sum_{k=1}^{N} y_k, \quad y_k = 0 \text{ or } 1$

  Note! Special case of the mean

- **Variance**: $\boldsymbol{\sigma^2} = \frac{1}{\boldsymbol{N}}\sum_{k=1}^{N}(y_k - \mu)^2,$ StDev: $\boldsymbol{\sigma} = \sqrt{\sigma^2}$

  Note! The variation among <u>all</u> $\boldsymbol{N}$ objects in the entire population

Stockholm University

Estimating population parameters

- Unknown **parameter**, e.g. $\boldsymbol{\mu}$ - a property of the population
- $\bar{Y}$ and $\bar{Y}_{\mathbf{str}}$ are **estimates** of $\boldsymbol{\mu}$ - they are **statistics** and as such they are **random variables**
- Distribution: $f(\bar{Y})$ - sampling distribution (CLT?)

The exact distribution of the estimator depends on the sampling design and on the distribution of the finite population.

- **Expected value**: $E(\bar{Y})$ - mean of <u>all</u> possible $\bar{y}$
- **Biased or Unbiased**: - unbiased if $E(\bar{Y}) = \mu$

(*sv. väntevärdesriktig*) $\boldsymbol{Bias}(\bar{Y}) = E(\bar{Y}) - \mu$

Stockholm University

Estimating population parameters, cont.

- **Variance**: $V(\overline{\boldsymbol{Y}})$

  Depends on the sampling design and population distribution

  - variance of all possible $\overline{\boldsymbol{y}}$
  - $V(\overline{\boldsymbol{Y}}) = E[(\overline{\boldsymbol{Y}} - \boldsymbol{\mu})^2]$
  - depends on $\boldsymbol{\mu}$, the population mean, which is unknown

- **Variance <u>estimation</u>**: $\widehat{\boldsymbol{V}}(\overline{\boldsymbol{Y}})$
  - an estimate of the estimator's variance
  - root of this number is the **standard error**: $\boldsymbol{SE}(\overline{\boldsymbol{Y}}) = \sqrt{\widehat{\boldsymbol{V}}(\overline{\boldsymbol{Y}})}$

- **Mean square error** – combination of variance and bias:

  $$\boldsymbol{MSE}(\overline{\boldsymbol{Y}}) = \boldsymbol{V}(\overline{\boldsymbol{Y}}) + [\boldsymbol{Bias}(\overline{\boldsymbol{Y}})]^2$$

Stockholm University

Random sampling designs

- We want the sample to “represent” the entire population
- **Representative** samples – what do we mean with that?
  - Deliberately selecting objects that you might think are representative can cause serious problems (see p. 31).
  - What you think and assume may not be the whole truth – not even partially.
- **Random sampling** guarantees **unbiased** or close to unbiased estimates with **controllable precision** (variance)
  - if the random procedure is repeated, the average of all outcomes will be equal to or close to the truth.
  - on average you also capture properties that you weren’t aware of and that may be important.
  - random sampling results in representative samples – on average.

Stockholm University

Inclusion probability

Probability of being included in the sample

- **Inclusion probability must be > 0 for all objects in the population**
  - All objects must be able to be included in the sample.
  - Inference only applies to those objects that have a chance of being drawn to the sample.
- **Inclusion probability must be <u>known</u> for all objects**
  - Do not have to be equal, they just need to be known.
  - If they are unknown you can't assess the precision (variances, standard errors).

Stockholm University

SRS = Simple random sampling design

- Sv. *OSU* **=** *obundet slumpmässigt urval*

- Draw $\boldsymbol{n}$ from $\boldsymbol{N}$ without replacement (wor), order doesn't matter

- No. of possible SRS samples without replacement: $\boldsymbol{C_n^N} = \binom{\boldsymbol{N}}{\boldsymbol{n}}$

Ex. $\binom{\mathbf{290}}{\mathbf{30}} = \underbrace{\mathbf{5\,936\,798\,537\ \ldots 000}}_{\text{41 digits}}$ different samples possible

- All $\boldsymbol{N}$ objects in the population have equal inclusion probabilities i.e. they all have the same chance of being drawn to the sample
  - inclusion probability = $\dfrac{\boldsymbol{n}}{\boldsymbol{N}}$

Estimating μ with SRS

Sample mean: $\bar{Y} = \frac{1}{n}\sum_{k=1}^{n} y_k$

- Expected value: $\boldsymbol{E(\bar{Y}) = \mu}$

$\bar{Y}$ is an unbiased estimator for $\boldsymbol{\mu}$

- Variance of $\bar{Y}$
  - wor: $\boldsymbol{V(\bar{Y})} = \boxed{\frac{N-n}{N-1}} \cdot \boxed{\frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}}$

What happens when $\boldsymbol{n \to N}$? If $\boldsymbol{n = N}$?

- **Finite population correction** (fpc) when σ^2 is known: $\boxed{\frac{\boldsymbol{N-n}}{\boldsymbol{N-1}}}$

see p. 251 in NCT

Stockholm University

Estimating μ with SRS, cont.

- Typically $\boldsymbol{\sigma}^2$ is unknown; **estimate** it with the **sample variance**:

$$\boldsymbol{s}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2$$

- **Estimate the variance of $\bar{Y}$:**

Adjusted fpc

$$\hat{\boldsymbol{V}}(\bar{\boldsymbol{Y}}) = \left(\frac{N-n}{N}\right)\cdot\frac{s^2}{n} = \left(\mathbf{1}-\frac{\boldsymbol{n}}{\boldsymbol{N}}\right)\cdot\frac{\boldsymbol{s}^2}{\boldsymbol{n}}$$

- **Adjusted fpc** when we use $\boldsymbol{s}^2$ instead of $\boldsymbol{\sigma}^2$

$$\boxed{\left(\mathbf{1}-\frac{\boldsymbol{n}}{\boldsymbol{N}}\right)}$$

NCT p. 310 is
<u>not</u> correct!

Stockholm University

How to draw an SRS

One way of doing it (see Assignment 2 instructions for another):

- Each object in the population gets a random number between 0-1
  - the numbers should be totally random and independent of each other
  - ideally no ties, unique random number for everyone
- Select the n objects with the smallest (or largest) random numbers
  - sort the population units according to size of the random number and take the first (or last) n objects

Stockholm University

Assignment 2B

Problem 1: SRS

1. Draw a random sample size $\boldsymbol{n = 30}$
2. Calculate sample mean $\bar{\boldsymbol{y}}$ and sample variance $\boldsymbol{s^2}$
3. Calculate the estimated variance of the sample mean and the standard error

$$\hat{\boldsymbol{V}}(\bar{\boldsymbol{Y}}) = \left(\mathbf{1} - \frac{\boldsymbol{n}}{\boldsymbol{N}}\right) \cdot \frac{\boldsymbol{s^2}}{\boldsymbol{n}} \qquad \boldsymbol{SE}(\bar{\boldsymbol{Y}}) = \sqrt{\hat{\boldsymbol{V}}(\bar{\boldsymbol{Y}})}$$

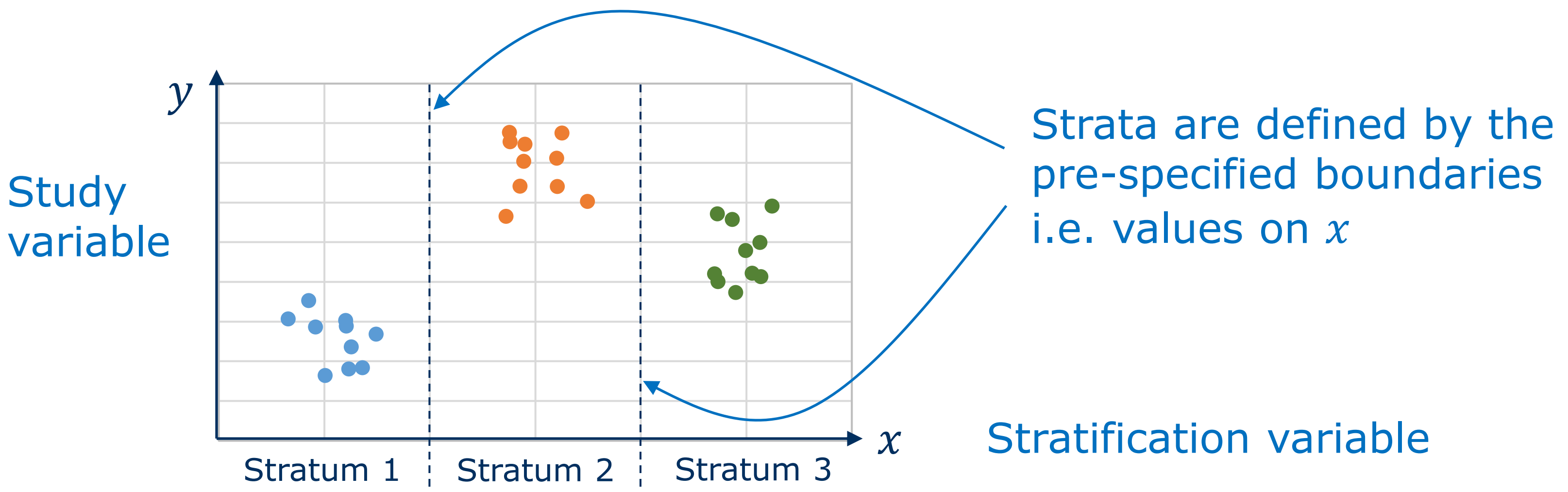
- $\bar{\boldsymbol{y}}$ is the point estimate of $\boldsymbol{\mu} = \frac{1}{N}\sum_{k=1}^{N} y_k$ and $\boldsymbol{SE}(\bar{\boldsymbol{Y}})$ is the standard error of $\bar{\boldsymbol{y}}$

Stratified random sampling

- The population is **partitioned** into L groups/**strata**
  - every object belongs to exactly one and only one stratum
  - which stratum an object belongs to is determined by its value on one or several **stratification variables**
- The value(s) of the stratification variable(s) are known for all
- From **each stratum** k, draw an **SRS** (wor) of size n_k
  - the sample sizes n_k will typically differ between strata
- Samples from each stratum are drawn independently
  - what happens in one stratum does not affect the others

Stockholm University

Stratified sampling



- Typically different means $\boldsymbol{\mu_k}$ across strata and small variances $\boldsymbol{\sigma_k^2}$ within strata, not necessarily equal, but all or most, hopefully smaller than $\boldsymbol{\sigma^2}$
- Draw independent SRS samples from each stratum

Stockholm University

Stratified random sampling, cont.

- The population is split into $\boldsymbol{L}$ groups defined by one or more variables $\boldsymbol{X_1, \ldots, X_p}$ that co-vary with the study variable $\boldsymbol{Y}$:

Partitioned over **one stratification variable**, e.g. X = age

Age	$15-17$	$18-24$	…	>65	
Stratum no.	1	2	…	L	
Population size	$\boldsymbol{N_1}$	$\boldsymbol{N_2}$	…	$\boldsymbol{N_L}$	$\sum_k \boldsymbol{N_k} = \boldsymbol{N}$
Relative size	$\boldsymbol{W_1 = N_1/N}$	$\boldsymbol{W_2 = N_2/N}$	…	$\boldsymbol{W_L = N_L/N}$	$\sum_k \boldsymbol{W_k = 1}$
Stratum mean	$\boldsymbol{\mu_1}$	$\boldsymbol{\mu_2}$	…	$\boldsymbol{\mu_L}$	$\sum_k \boldsymbol{W_k \mu_k = \mu}$

All the means are unknown!

Stockholm University

Allocating the sample to the strata

- How do we decide the sample sizes n_k for each stratum?
- **Proportional allocation** – based on the stratum **size** N_k

$$\boxed{W_k = \frac{N_k}{N} \approx \frac{n_k}{n} \quad \Rightarrow \quad n_k \approx n \cdot W_k}$$

Use this method in Assignment 2

- ***Neyman allocation*** (sometimes referred to as optimal)
  - Takes into account both ***size*** and ***variance*** with-in strata
  - Strata with large with-in variance – sample more; strata with small with-in variance – sample less

- ***Optimal allocation***
  - Takes into account ***size***, ***variance*** and ***cost***
  - Expensive strata – sample less; cheap strata – sample more

Stockholm University

Stratified random sampling, cont.

- Draw an SRS from each stratum and calculate the stratum sample mean and stratum sample variance:

Age:	$15-17$	$18-24$	...	>65	
Stratum no.	1	2	...	L	
Population size	$\boldsymbol{N_1}$	$\boldsymbol{N_2}$	...	$\boldsymbol{N_L}$	
	SRS ⇓	SRS ⇓		SRS ⇓	
Sample size	$\boldsymbol{n_1 \approx nW_1}$	$\boldsymbol{n_2 \approx nW_2}$	...	$\boldsymbol{n_L \approx nW_L}$	$\boldsymbol{\sum_k n_k = n}$
Mean, variance	$\boldsymbol{\bar{y}_1,\ s_1^2}$	$\boldsymbol{\bar{y}_2,\ s_2^2}$	...	$\boldsymbol{\bar{y}_L,\ s_L^2}$	

Stockholm University

Stratified random sampling, cont.

- Proportion of the population that belong to stratum $\boldsymbol{k}$: $W_k = \frac{N_k}{N}$
- Population mean, unknown parameter:

$$\mu_y = W_1\mu_1 + W_2\mu_2 + \cdots + W_L\mu_L$$

Weighted mean of means

- Estimation of unknown population mean $\boldsymbol{\mu}$:

$$\bar{Y}_{\text{str}} = W_1\bar{Y}_1 + W_2\bar{Y}_2 + \cdots + W_L\bar{Y}_L$$

Unbiased estimator for $\boldsymbol{\mu}$

- Estimation of variance of $\bar{Y}_{\text{str}}$:

$$\hat{V}(\bar{Y}_{\text{str}}) = \sum_{i=1}^{L} \underbrace{W_i^2}_{\text{Weight factor}} \underbrace{\left(1 - \frac{n_i}{N_i}\right)\frac{s_i^2}{n_i}}_{\text{Same formula as SRS}}$$

Same formula as SRS

Weight factor

Stockholm University

Why stratified sampling?

- By sampling from each stratum we hope to get a sample guaranteed to cover most of the observable region of $\boldsymbol{Y}$

- Assumes that the stratification variables $\boldsymbol{X}$ and the study variable $\boldsymbol{Y}$ are related (linear or non-linear relationship)
  - will almost always provide better estimates compared to pure SRS, even when the relationship is moderate

- Strictly speaking, we're hoping for small variances within strata
  - objects are similar within strata, different between strata
  - **homogeneous within strata, heterogeneous between strata**

Stockholm University

Why stratified sampling, cont.?

- Often we want statistics for sub-populations or domains of the population
  - areas, industries, size, gender …

- By planning for this in advance we can get better estimates for each sub-population and typically better estimates for the whole population (i.e. smaller standard errors)

Comment: we can still provide statistics for sub-domains even if they are not used in a stratification design – called **post-stratification** – but we will get estimates with larger variance (smaller precision).

Stockholm University

How should we stratify?

- Which variables should we use for the stratification?
  - those that are related to and co-vary with the study variable (linearly or non-linearly, doesn't matter)
- How many strata? $\boldsymbol{L = 2, 3, 4, \ldots}$?
- How do we define the strata, where do we set the bounderies?
  - e.g. Age: 15-17, 18-24, 25-35, …?

Addressing these issues requires a more comprehensive treatment and is out of scope for this course.

- For Assignment 2, choose sensibly; deliberately defining strata so that they are equally sized is <u>not</u> a sensible choice! Why?

Stockholm University

Assignment 2B

Problem 2: Stratified sample

1. Select a stratification variable, you have 5 to choose from
   - which of the available variables do you think affect taxes most?
2. Define the boundaries for $\boldsymbol{L = 3}$ strata (e.g. low, medium, high)
3. Count the stratum sizes $\boldsymbol{N_k}$ and calculate the stratum weights $\boldsymbol{W_k}$
4. Allocate the sample i.e. determine $\boldsymbol{n_1}$, $\boldsymbol{n_2}$ and $\boldsymbol{n_3}$
5. From each strata, draw an SRS of size $\boldsymbol{n_k}$
6. For each stratum $\boldsymbol{k}$, calculate the sample mean $\boldsymbol{\bar{y}_k}$ and variance $\boldsymbol{s_k^2}$
7. Finally, calculate the estimate $\boldsymbol{\bar{y}_{\mathrm{str}}}$ of the population mean $\boldsymbol{\mu}$ and the standard error for the estimate

Assignment 2B, cont.

	Stratum 1	Stratum 2	Stratum 3
Stratum boundaries	*you choose*	*you choose*	*you choose*
Stratum size	N_1	N_2	N_3
Stratum weight	W_1	W_2	W_3
Sample size	n_1	n_2	n_3
Selected sample (the numbers of the sampled municipalities)			
Sample mean	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
Sample variance	s_1^2	s_2^2	s_3^2

$$\bar{y}_{\text{str}} = W_1\bar{y}_1 + W_2\bar{y}_2 + W_3\bar{y}_3$$

$$SE(\bar{Y}_{\text{str}}) = \sqrt{\sum_{k=1}^{3} W_k^2\left(1-\frac{n_k}{N_k}\right)\frac{s_k^2}{n_k}}$$

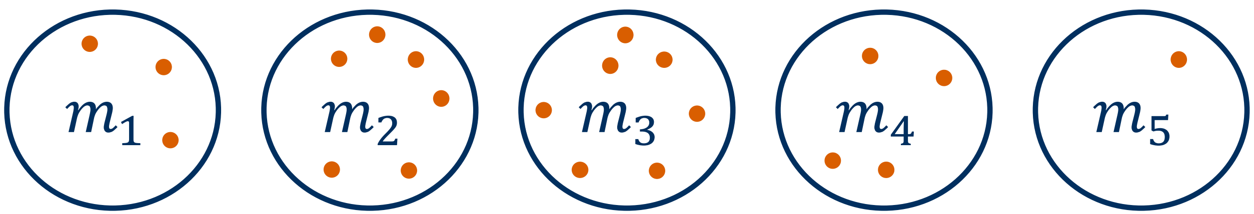
Stockholm University

Cluster sampling – brief overview

Instead of drawing individual objects, draw entire groups/**clusters** of objects

- The population is divided into $\boldsymbol{N}$ clusters (a partition)
- Choose $\boldsymbol{n}$ clusters, e.g. using SRS
- In total there are $\boldsymbol{M}$ objects in the population
- In cluster $\boldsymbol{i}$ there are $\boldsymbol{m_i}$ objects, $\sum \boldsymbol{m_i} = \boldsymbol{M}$

No need to learn any formulas!



$N = 5$ $\quad\quad$ $M = m_1 + \cdots + m_5 = 21$

Cluster sampling, cont.

Opposite of stratified sampling

- Choose say $\boldsymbol{n} = \mathbf{2}$ clusters randomly with SRS

m_1 $\boldsymbol{m_2}$ m_3 m_4 $\boldsymbol{m_5}$

$$N = 5 \qquad M = m_1 + \cdots + m_5$$
$$n = 2 \qquad m = m_2 + m_5$$

- Cluster sampling works if all clusters are similar to each other; with-in variance is large and between variance is small
  - **Homogeneous between, heterogeneous with-in**
- **Every cluster should be a miniature of the population**

Cluster sampling, cont.

- First stage – primary sample units (PSU)
- Second stage – secondary sample units (SSU)

Why cluster sampling? Example:

- Schools (PSU), students within schools (SSU)
  - we have a list of all schools but not of all students

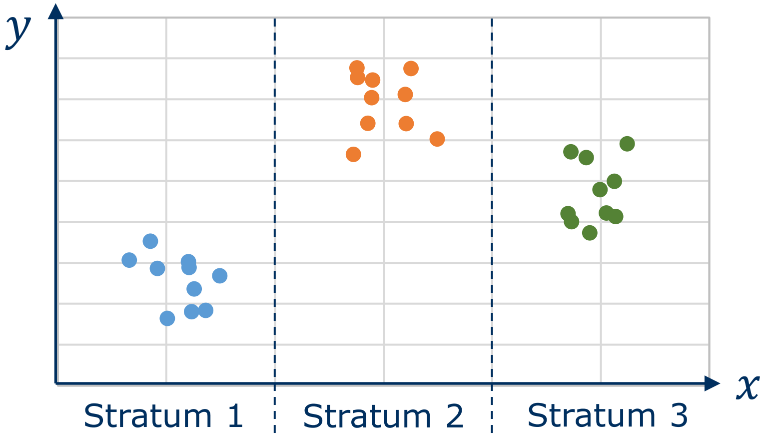
Real ex. Structural wages (SCB, National Mediation Office):

- Choose companies (= clusters, PSU)
  - stratified random sample (size, industry etc.)
- Within each selected company all employed are chosen (SSU)

Stockholm University

Stratified and Cluster sampling

Strata



Different means and

small variances within strata

(not necessarily equal).

- Sample from each stratum

Clusters

Approximately equal means and

same (large) variances in all clusters;

each cluster mimics the population.

- Sample a few clusters

Stockholm University

Assignment 2B

Problem 4: Cluster sampling

- You are not required to do any sampling or calculations, only reasoning is required!
- Given the data set, how could we define our clusters?
  - you need to select one the available variables in the data file to define your clusters.
- Is cluster sampling suitable in this case?
  - Are each of the clusters miniature representations of Sweden?
  - Are they equally sized or do they differ? Means and variances?
  - Can we expect the clusters to resemble each other, are they homogeneous clusters?

Other sampling designs

Random sampling

- Systematic sampling – select a starting point e.g. in the interval 1-100 and then select every 100:th observation
- Multistage – sample clusters, then sample from each cluster (2-stage)
- Multiphase – large sample first, then sample from the sample
- pps and πps – proportional to size – with or without replacement

Non-random sampling

- Quota sampling – next slide
- Snowball sampling – *”hard to reach populations*”

Non-random designs

An example: **Quota sampling**

- Non-probabilistic version of stratified sampling:
  - interviewer is told to interview ”5 young women, 5 young men, 10 older women, …” and so on
  - these numbers (quotas) may represent the corresponding population proportions (proportional allocation)
- High risk of non-random selection by the interviewers; convenience selection, accidental selection etc. may cause systematic errors i.e. **bias**
- True **inclusion probabilities** are totally **unknown**

Error types

Sampling error

- Not all objects are included in the sample, only a subset
- Assessed with the standard errors $\sqrt{\left(1-\frac{n}{N}\right)\frac{s}{\sqrt{n}}}$

Non-sampling errors, systematic errors

- Non-response
- Frame errors, coverage errors
- Measurement errors
- Processing errors
- …

Non-response

- If the propensity to be a non-responder depends on the study variable, systematic errors are introduced, i.e. **bias**
  - e.g. younger people tend to be non-responders more often than middle aged; we are studying wages, how people vote …

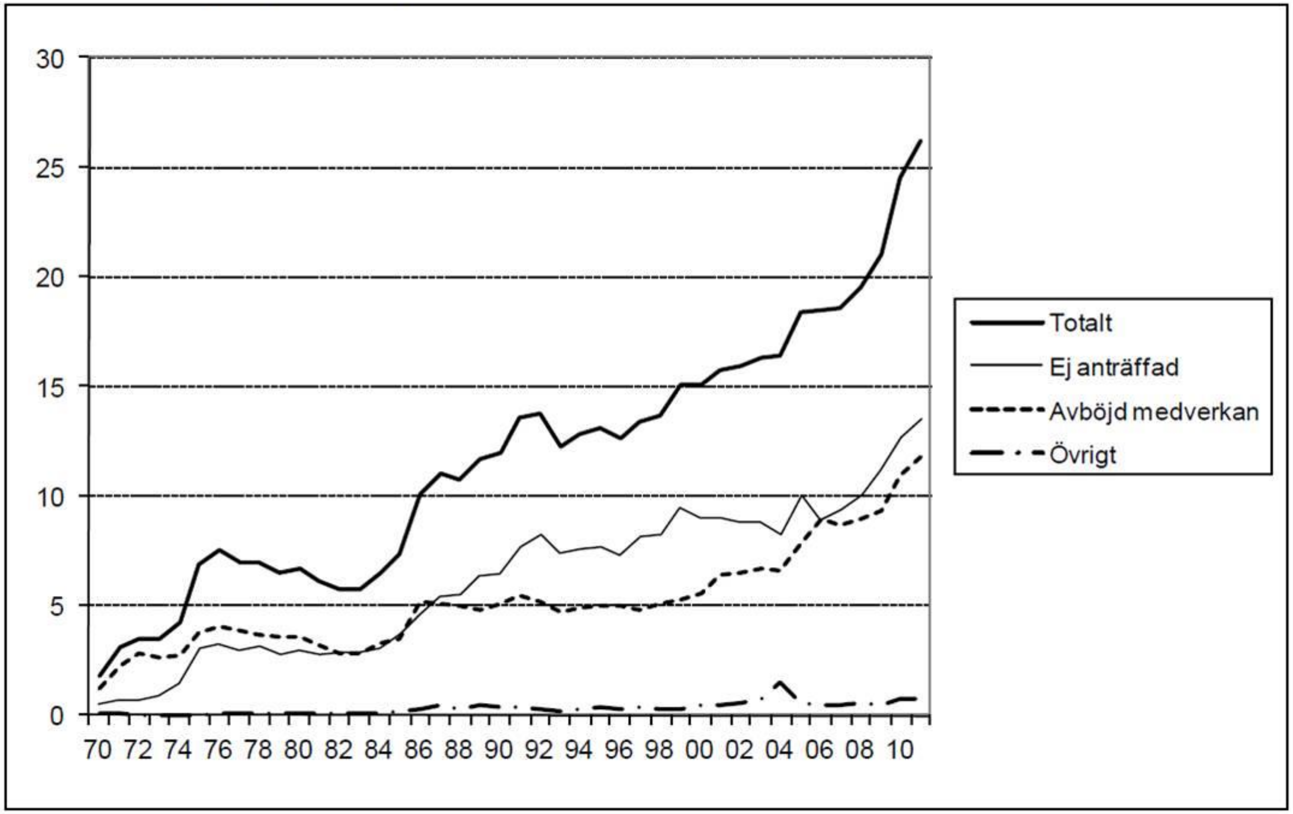
- Compensating for non-response requires assumptions about the study variable and how it relates to the propensity of being a non-responder that are difficult (impossible) to verify.

- Small non-response rates (<10%) may be acceptable
- Many surveys today have non-response rates at >50%

Stockholm University

Ex. Non-response

Labor force survey (LFS, sv. AKU), 1970-2011, age group 16-64, non-response rates (%)



Stockholm University

Effect of non-response

- We draw a sample and study the variable $\boldsymbol{Y}$ but we will only observe values on $\boldsymbol{Y}$ for those who respond

- Applying the **law of total probability** (see L5):

$$\boldsymbol{P}(\boldsymbol{Y}) = \boldsymbol{P}(\boldsymbol{Y}|\text{response}) \cdot \boldsymbol{P}(\text{response}) + \boldsymbol{P}(\boldsymbol{Y}|\text{no response}) \cdot \boldsymbol{P}(\text{no response})$$

- We are missing $\boldsymbol{P}(\boldsymbol{Y}|\text{no response})$, no data = we can't estimate it

- If $\boldsymbol{Y}$ and response are statistically dependent and we substitute $\boldsymbol{P}(\boldsymbol{Y})$ for $\boldsymbol{P}(\boldsymbol{Y}|\text{response})$ we introduce **BIAS**!

- Ex. if dependent we may have that $\bar{\boldsymbol{y}}_{\textbf{reponse}} \neq \bar{\boldsymbol{y}}_{\textbf{non response}}$

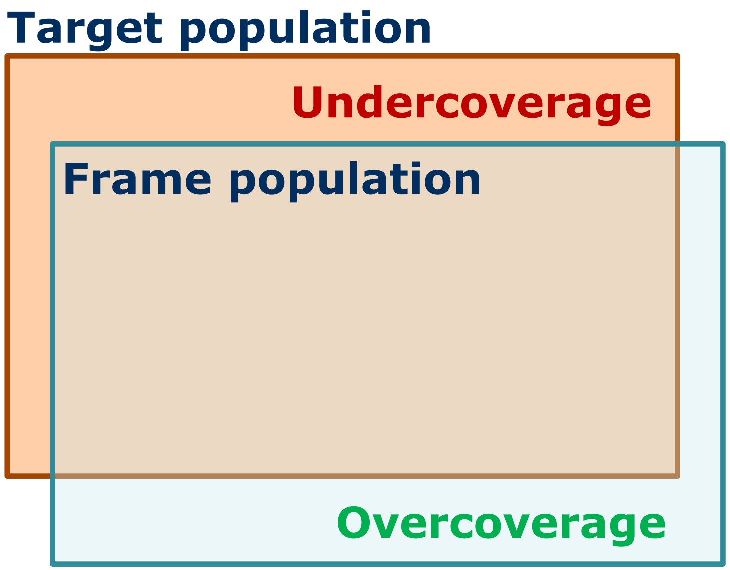
Stockholm University

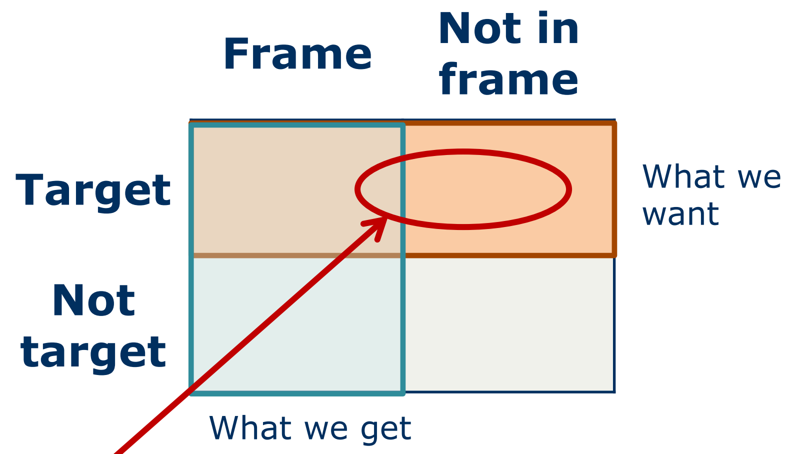
Frame and coverage errors

- **Sampling frame** or just **frame** (sv. *ram*)
  - A list, data file, register or similar, that lists all objects in the **finite** population
  - the sample is drawn from the frame
  - should ideally match the population you are interested in i.e. the **target population**
- Sometimes the frame is out of date; or it's just a bad frame
  - objects that belong to the target population but are missing in the frame will have inclusion probability 0
  - objects that are in the frame but do not belong to the target population can often be identified and set aside

Stockholm University

Coverage problems





Our study variable of interest
- skewed representation
- **BIAS**

Ex. Non-response & frame issues

A classic:

- Landon – Roosevelt,

  US presidential election 1936

- *Literary Digest* successfully predicted the winner for the past five elections
  - 10M questionnaires, 2.3M returned (**67 % non-response**)
  - Frames: *Literary Digest* subscribers, motor vehicle registries and telephone registries
    - Obvious **coverage problems** due to poor frames

Stockholm University

Data collection methods - *mode*

- **Mode should be chosen to suit the respondents**
  - companies, organizations, real people (adults, children)
- Telephone interviews (computer aided = CATI)
- Personal interviews (computer aided = CAPI)
- Snail mail (postal) - questionnaires
- Electronically via web sites or e-mail – questionnaires
- On-line (automatized) electronic “dumps” from systems
  - registers, data bases, accounting software
- Mixed mode – combining several methods
- **Mode affects quality – non-response, measurement error**

Questionnaire design

- Formulating the question – clarity, easily understood by everyone
- Closed questions (multiple choices)
  - scale (nominal, ordinal), number of alternatives, exhaustive, non-overlapping, midpoint, don't know alternative, …
- Open questions, answer in your own words
  - how to operationalize open answers to variables?
- Factual (e.g. age) vsv non-factual (e.g. opinions)
- Sensitive or embarrassing questions: "Have ever spent time in a jail?"
- Avoid negatives and negations: "Do you dislike …?", "Don't you dislike …?"
- Does the question work? Test your questions on each other!
- **Bad questions = bad data! Measurement errors!**

Stockholm University

Assignment 2A

Things to address:

See the instructions!

- Target and frame populations
  - how are the study objects and the survey units defined?
- Potential coverage issues with the frame that you are using
- Sampling and collection methods, what are the possible effects?
- Non-response?
- Measurement errors?

Can you anticipate any problems with your survey scenario?

You'll never design a perfect survey!

- **Precision (standard errors) is of course a quality issue as well but by choosing good sample sizes and designs we can control and calculate these**

Stockholm University