Content

- Methods of estimation
- Variance estimation
- Treatment of non-responses
- Panel surveys
- Micro census as example

Methods of estimation

- Opposite of selection procedure, conclusion from the sample to the population
- Condition of each estimation:
  For each sample unit *i* , the *inclusion probability* π_i must be known !
  (probability of being included in the sample)
- Inclusion probabilities are usually known from the sample plan.

Free Estimation

- Extrapolation with reciprocal values of the inclusion probabilities

$$\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

(„Horwitz-Thompson-Estimator“)

- Most simple possibility
- Information comes only from the sample design

Free Estimation

- Example:
  - Survey of professional training in companies 2005 (CVTS 3)
  - Random sampling stratified by 30 business branches and 6 employee size ranges of enterprises from the business register

CVTS 3 – Survey 2005
Business Branch (NACE 30)
Coal mining and extraction of stones and earth

Employee size ranges		Number of businesses in the population	STIA - sampling		Number of selected businesses	
			Size of zones	Sampling per zone	Altogether according to sampling protocol	Not part of target population
nr.	from.. until… employees [1)]					
1	10 – 19	375	4	1	93	13
2	20 – 49	228	4	1	57	1
3	50 – 249	129	5	4	103	6
4	250 – 499	7	1	1	7	-
5	500 – 999	6	1	1	6	1
6	1000 plus	8	1	1	8	-
altogether		753			274	21

[1)] Employees being subject to social insurance contributions on 31.12.2003

Loose extrapolation

- Task 1:
  - Which kind of information would you consider for a loose extrapolation?
  - Determine the inclusion probability for one business of each stratum!

Loose Extrapolation

- The estimator of loose extrapolation is linear, i.e. in the form of

$$\hat{Y} = \sum_{i=1}^{N} w_i y_i$$

- Simple computational implementation:
  add the extrapolation factor to the individual material.
- Analysis with SAS:
  WEIGHT Statement

Individual Material with Extrapolation Factor

Business Number	Stratum		Sampling characteristics		. . .	. . .	Extrapolation factor
	NACE 30	Size class	Employees on 31.12.2005	Participant of training costs			
. . .	1	1	. . .	. . . .			4
. . .	1	2	. . .	. . . .			4
. . .	1	3	. . .	. . . .			1,25
. . .	1	4	. . .	. . . .			1
. . .	1	5	. . .	. . . .			1
. . .	1	6	. . .	. . . .			1
. . .	. . .	. . .	. . .	. . .			. . .
. . .	. . .	. . .	. . .	. . .			. . .
. . .	. . .	. . .	. . .	. . .			. . .
. . .	. . .	. . .	. . .	. . .			. . .
. . .	. . .	. . .	. . .	. . .			. . .
. . .	. . .	. . .	. . .	. . .			. . .
. . .	. . .	. . .	. . .	. . .			. . .

Use of Auxiliary Information

Adjusted Weighting:

- Target: More precise results as opposed to loose extrapolation
- Basically two kinds of auxiliary information:
  - Information from the sampling frame
    (as long as not being used for sample plan): generally total value of a characteristic X in the sampling population
    ⇨ ratio estimation, difference estimation, regression estimation
  - Information from a different source:
    e.g. total value of a characteristic X at the survey point of time from a register
    ⇨ adaptation, calibration

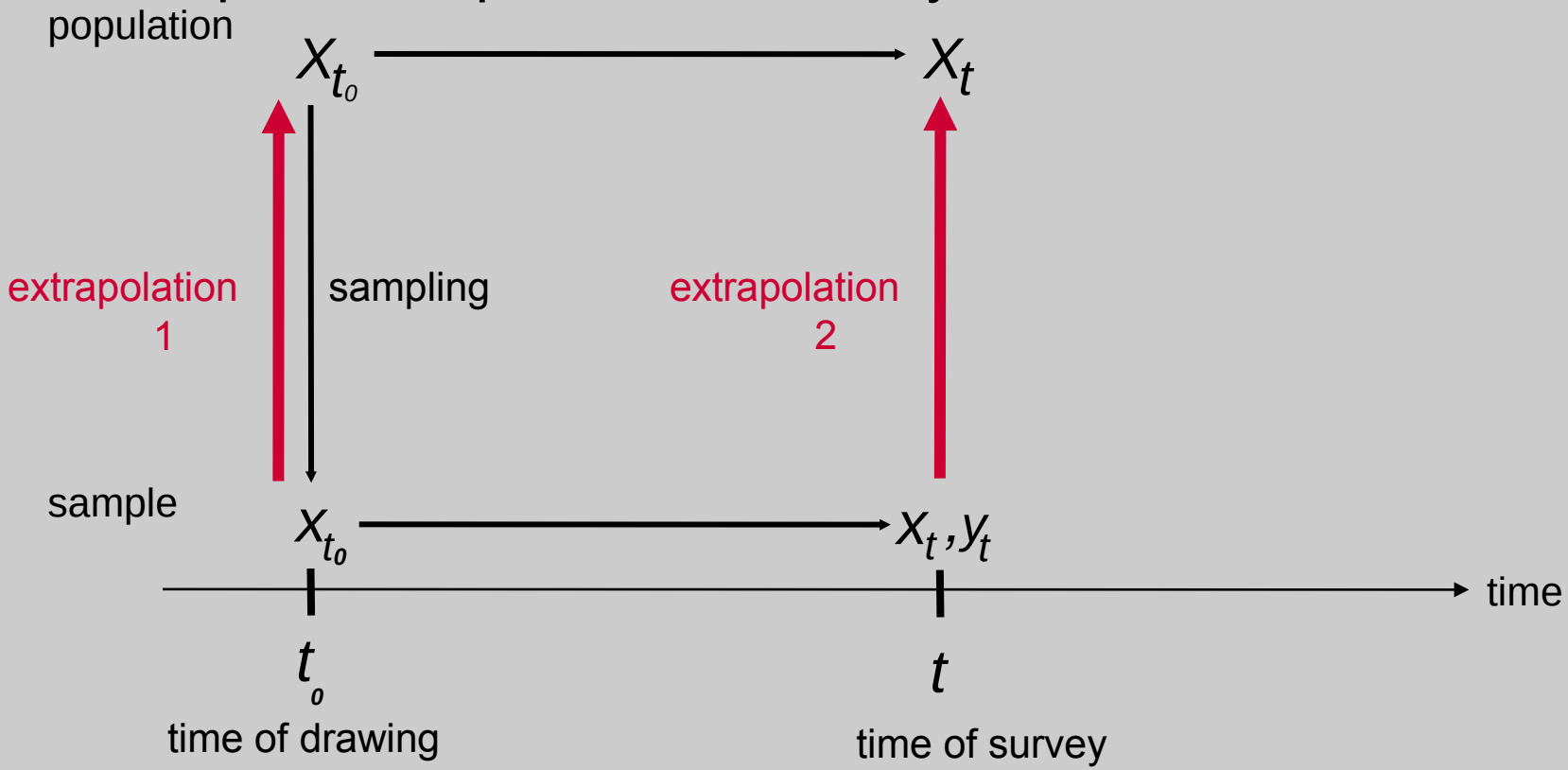
Use of Auxiliary Information

Adjusted Weighting

- For the auxiliary characteristic the values of sample units
  x_i (i = 1,….., n) must be known
- Efficiency depends on correlation between sampling- and auxiliary characteristic

Use of Auxiliary Information

past vs. up-to-date auxiliary information



Use of Auxiliary Information

- Task 2:
  - Is it better to use past or up-to-date information? Discuss advantages and disadvantages!

Ratio Estimation

- Idea 1:
  - “Adjust” total value of an auxiliary characteristic of the sampling population with the ratio survey characteristic/auxiliary characteristic from the sample
- Idea 2:
  - Extrapolate the total value of the auxiliary characteristic from the sample loosely, compare with the “true” total value from the sampling population. “Improve” the estimation via the ratio of “true” to estimated total value.

Ratio Estimation
Coal mining and extraction of stones and earth
CVTS 3 – survey 2005

Employee size classes	Sampling population		sample		
	Number of businesses	employees 31.12.2003	Sampling fraction in %	employees 31.12.2003	employees31.12.2005 according to survey
1	375	5128	25	706	632
2	228	6896	25	714	749
3	129	12428	80	3374	2662
4	7	2935	100	896	473
5	6	4083	100	1855	1126
6	8	52921	100	50404	41812
altogether	753	84391	-	57949	47454

Ratio Estimation

- Task 3:
  - How would you picture a ratio estimation in the CVTS 3 - survey?
  - The aim is an extrapolated number of employees on 31.12.2005 in the mining industry for one size class

Ratio Estimation

- Estimator for the total value Y for simple random sampling

$$\hat{Y}_{VH} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} X = \frac{X}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} y_i$$

with

X : Total wert of the auxiliary characteristic

x_i : Value of the auxiliary characteristic of sampling unit i (i =1, … , n)

y_i : Value of the survey characteristic of the sample unit

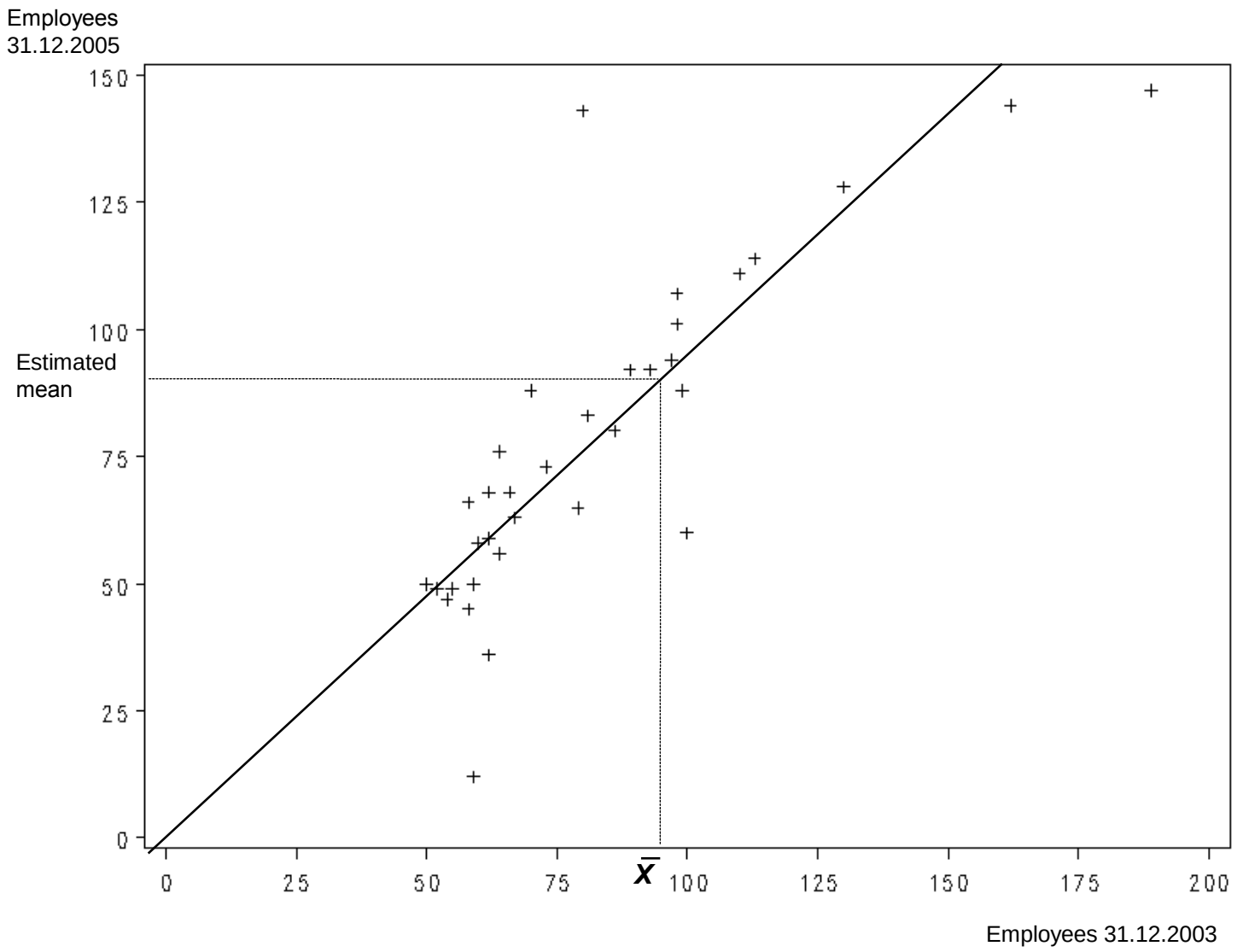
Ratio Estimation

- Error variance for $\hat{Y}_{VH}$ can be estimated from the sample by

$$\hat{V}\left(\hat{Y}_{VH}\right) = \frac{N^2}{n}\left(1 - f\right)\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \hat{R}x_i\right)^2$$

with

$$\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$: ratio value in the sample



CVTS 3-survey mining size-class 3

Ratio Estimation

- Ratio estimation is the more precise, the smaller the sum of squared differences of (x_i,y_i) of the straight line with gradient through the zero-point.
- Ratio estimation is biased, yet the bias converges towards zero with increasing sample size.
- Rule of thumbs:
  - Correlation coefficient $\geq$ 0,5 or 0,6
  - $n \geq 30$

Ratio Estimation

- Two methods for stratified samples
  - Separate ratio estimation:
    One ratio estimation per stratum, after summate across the stratums
  - Combined ratio estimation:
    Generate ratio value (if applicable weighted) across all stratums
- Preference of separate ratio estimation whenever the stratum- ratio values differ strongly

Regression Estimation

- Estimate for the total value Y in a simple random sampling

$$\hat{Y}_{Reg} = \frac{N}{n}\sum_{i=1}^{n} y_i + b\left(X - \frac{N}{n}\sum_{i=1}^{n} x_i \right)$$

whereas

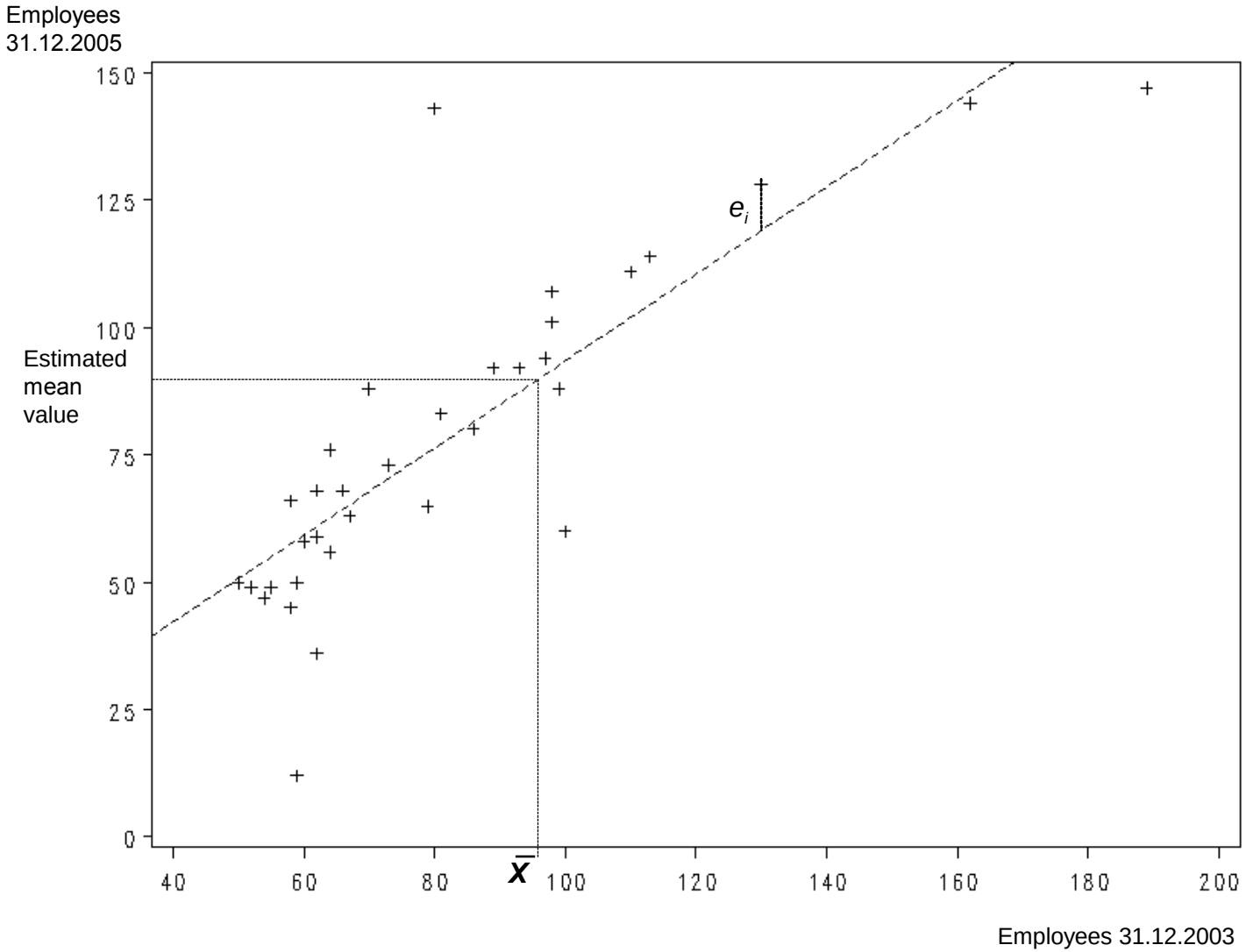
$$b := \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

is the regression coefficient.

(„least-squares-estimation“)

Regression Estimation

- b is gradient of the regression line
- $b = r_{xy} \times \frac{s_x}{s_y}$, whereas r_{xy} is correlation coefficient between x und y.



CVTS 3-survey mining size-class 3

Regression Estimation

- The regression estimator is biased, yet the bias converges towards zero with increasing sample size.
- It is at least as precise as loose extrapolation and (almost always) as the ratio estimation
- Rule of thumb: sample size at least 30 !
- For stratification: separate and combined regression estimation
- Extrapolation of the auxiliary characteristic yields exactly the total value of the sampling population

Regression Estimation

- The regression estimate is linear:

  extrapolation factor

$$w_i = \frac{N}{n}\left(1 + \frac{X - \frac{N}{n}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\left(x_i - \bar{x}\right)\left(1 - \frac{1}{n}\right)\right)$$

Regression Estimation

- Variance estimation of the regression estimate in simple random sampling

$$\hat{V}\left(\hat{Y}_{Reg}\right) = \frac{N^2}{n}\left(1-f\right)s_y^2\left(1-r_{xy}^2\right)$$

$$= \frac{N^2}{n}\left(1-f\right)\times\frac{1}{n-1}\sum_{i=1}^{n} e_i^2$$

whereas

$e_i := y_i - \hat{y}_i \; (i = 1,\ldots,n)$ are the residuals and

$\hat{y}_i := \bar{y} + b\left(x_i - \bar{x}\right)$ is the value estimated from the regression line of y_i .

Regression Estimation

- Regression estimation is more efficient, the larger the correlation coefficient between survey- and auxiliary characteristic.
- Error calculation can be carried out –for any sample design- according to the formula for loose extrapolation, you only have to replace the y_i by the residuals.

Stratification according to Sampling (Post-Stratification)

- Condition: the sampling population can be divided into classes, according to which was not stratified.
- Idea: we pretend that the stratification was done according to these classes and use the formula for loose extrapolation, ratio estimation or regression estimation.
- The sample size which falls upon one class depends on chance $\Rightarrow$ don't divide classes as profoundly as in a stratification!

Adaptation Methods

- Use of up-to-date auxiliary information with total values of the target population from a different source („benchmarks“)
- Possible sources: register, extrapolation, larger samplings
- Implicit treatment of under-coverage (missing accesses, errors in the sampling population) and non-response

Adaptation Methods

- Conditions:
  - Benchmarks must be correct, i.e. they may not show any (or just insignificantly small) systematic or random errors
  - Auxiliary characteristics in the sample must be available equally divided by definition
  - Existence of original extrapolation factors (e.g. factor of loose extrapolation)

Adaptation Methods

- Target:
  - New extrapolation factor, which “meets” the benchmarks („adaptation“, „calibration“)
  - New extrapolation factor should differ “preferably little” from the original factor
    - „ preferably little“ is measured by distance functions
    - Several distance functions are possible, e.g. sum of squared deviations: general regression estimator

Adaptation Methods

- Generally not only one, but several (often categorical) auxiliary characteristics
- Separate adaptation to marginal distribution possible

Working with Extrapolation Factors

- Imagination: extrapolation factor of one unit tells, how many units of the population are represented by this unit
- Extrapolation of case numbers for sub-groups:
  - Extrapolation factor summate across all data-sets of the sub-group
  - Sample methodological is $y_i = 1$, if unit i belongs to the sub-group and $y_i = 0$ else wise
  - The smaller the sub-groups, the smaller a possible gain in accuracy of a bounded extrapolation

Working with Extrapolation Factors

- Extrapolation of total values of continuous variables:
  - Multiply characteristic value by extrapolation factor, then sum up
  - Extrapolation of ratios, means, shares
  - Extrapolate nominator and denominator separately, divide only afterwards

Working with Extrapolation Factors

- Task 4:
  - In a very skew distribution, e.g. for income, the median is often of interest.
  - How do you proceed, with given extrapolation factors, to estimate the median from the sample?

Working with Extrapolation Factors

- Often uniform extrapolation methods for all survey characteristics
  - Inconsistencies get avoided
  - Not optimal for all characteristics
- It is common to weight with the factor of loose extrapolation in complex analyzing procedures

Variance Estimation (Error Calculation)

- Error calculation is very important for assessing the quality (quality reports). Therefore it should be an integral part of each sample survey!
- For the calculation of the (error-)variance of an estimation function you actually need all the values of the survey characteristic in the population
- Variance must be estimated out of the sample

  ⇨ not reliable for small sample sizes

Variance Estimation

- Extend of the random sampling error depends primarily on
  - the sample size
  - the dispersion of the characteristic values (for continuous variables)
  - The share value of the sub-group (for categorical values)

Variance Estimation

- For demonstration, the following measurements of random sampling error are better suited than the variance:
  - Standard error: square root of variance
    - For variables with a higher mean, generally the standard deviation of the individual values and thus the standard error higher are higher, too.
  - Relative standard error: standard error related to the extrapolated value (often given in percent)
  - Confidence interval
    - Often at a level of 90% or 95%
    - Lower- and upper level- or ±

Variance Estimation
CVTS 3 – Survey
Mining and manufacture of stones and earth

employees on 31.12.2005	Variance	Standard-error	Relative standard-error (%)	Confidence interval at the level 90%	
				from	until
73289	1,79 Mill.	1338	1,8	71094	75484

- Combined regression estimation
- For loose extrapolation relative standard error 24,4 %

Variance Estimation

Praxis

- Request on accuracy must be determined from a professional point of view
- Results with a relative standard error of 15% or more often don't get published (slash),
  of 10% until less than 15% get marked as afflicted with a high error by bracketing
- Tag with letters for size classes of the relative standard error
- For selected important results, the random error should be directly indicated at the result

Variance Estimation

Praxis

- For categorical variables case number criteria:
  - A table element with 50 sample units has a relative standard error of 15%, in case of simple random sampling and loose extrapolation
  - If sub-groups are very small, stratification and bounded extrapolation loose their efficiency increasingly
  - Examine a relation between case number and random error out of the error calculation and deduce a case number criteria from this (example micro census)

Variance Estimation

Praxis

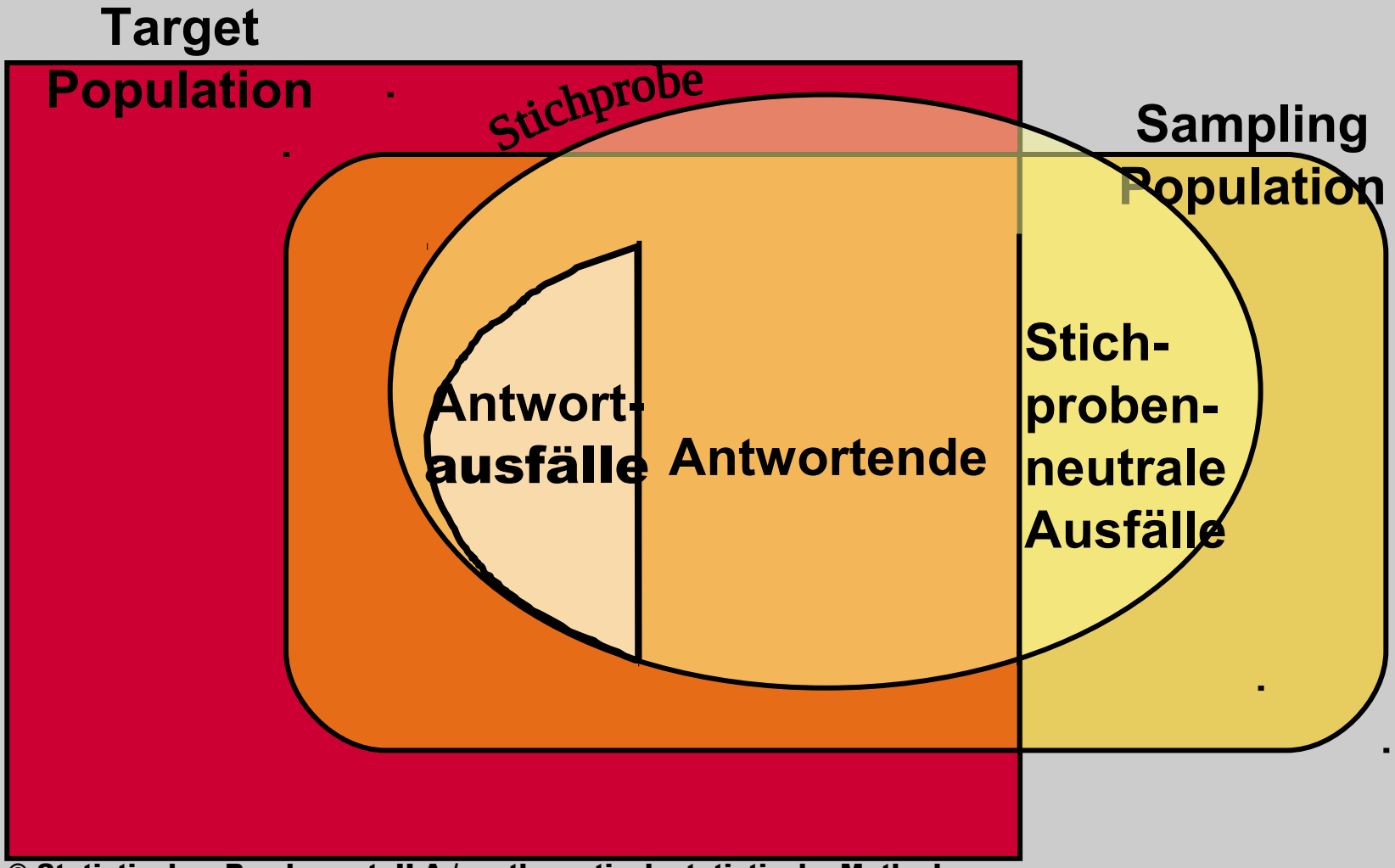
- Design effect = $\frac{\textit{Variance}}{\text{Variance in simple random sample}}$

calculate from recent survey or for several selected characteristics.

Treatment of Nonresponses

- Nonresponses can cause biases (also a higher random sampling error because of less usable sample units)
- For its treatment a model is required, which explains the failure rate
- Two kinds of nonresponses:
  - Unit- Nonresponse (all survey characteristics are missing): Treatment within the framework of the extrapolation $\Rightarrow$ usually larger extrapolation factors
  - Item- Nonresponse (only several survey characteristics are missing): Treatment by imputation

Treatment of Nonresponses



CVTS 3 – Survey 2005
Coal mining and extrapolation of stones and earth

Employee size classes	Sampling population		Sample						
			According to survey protocol		Belonging to target population		Respondents		
	Number of businesses	employees 31.12.2003	Number of businesses	employees 31.12.2003	Number of businesses	employees 31.12.2003	Number of businesses	employees 31.12.2003	employees 31.12.2005
1	375	5128	93	1261	80	1102	41	547	632
2	228	6896	57	1703	56	1675	24	686	749
3	129	12428	103	9705	97	9130	34	2799	2662
4	7	2935	7	2935	7	2935	2	896	473
5	6	4083	6	4083	5	3559	2	1331	1126
6	8	52921	8	52921	8	52921	6	50404	41812
altogether	753	84391	274	72608	253	71322	109	56663	47454

Treatment of nonresponses

- Task 5:
  - How could a extrapolation which considers nonresponses look like?
  - Extrapolate the number of employees on 31.12.2005 for size class 1!

Treatment of Nonresponses

- Most important: Differentiation between nonresponses (“real losses“) and sample neutral losses („unreal losses“)
- Sample neutral losses don’t cause biases; sample-methodological reply with $y_i = 0$
- What to do if it is not known, if the missing unit belongs to the target population?

Treatment of Nonresponses

- Idea: Consider respondents as random sub-sample with (unknown) inclusion possibilities. Estimation by a model.
- Most simple model:
  - All units have the same answer-possibility θ_i
  - No use of auxiliary means for the nonresponses.
  - Estimation of

$$\hat{\theta}_i = \frac{n - n_a}{n}$$

with

n_a: number of nonresponses

Treatment of Nonresponses

- Most simple model:
  - Adjusted extrapolation factor in simple random sample

$$\frac{N}{n} \times \frac{n}{n - n_a} = \frac{N}{n - n_a}$$

Treatment of Nonresponses

- „multiplicative addition“:
  - Random granting of response in each stratum
  - Adjusted extrapolation factor

$$\frac{N_h}{n_h} \times \frac{n_h}{n_h - n_{h,a}} = \frac{N_h}{n_h - n_{h,a}}$$

with $n_{h,a}$ = number of nonresonses in stratum h

Treatment of Nonresponses

- „multiplicative addition“ (continued):
  - Also works with classes instead of stratums
  - Uses auxiliary information about the nonresponses
- Estimation of individual response possibilities $\hat{\theta}_i$ for example by Logit-Model

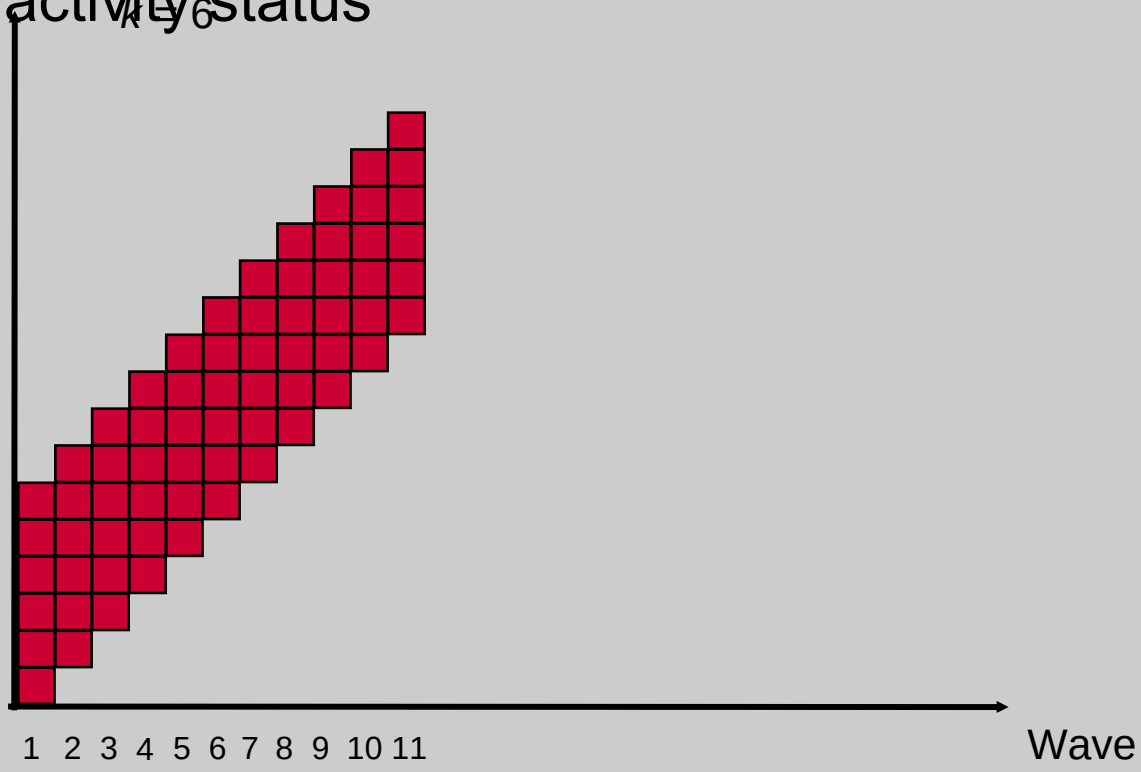
Replace inclusion probability π_i by $\pi_i \hat{\theta}_i$.

Panel Surveys

- At different points of time (e.g. monthly or annual; „waves“) interviewing the same sample units about the same survey characteristics
- In praxis, a rotating panel is commonly used: each sample unit is interviewed in *k* consecutive waves and is released according to plan.

Panel Surveys

- Example: Monthly telephone survey of the ILO-activity status

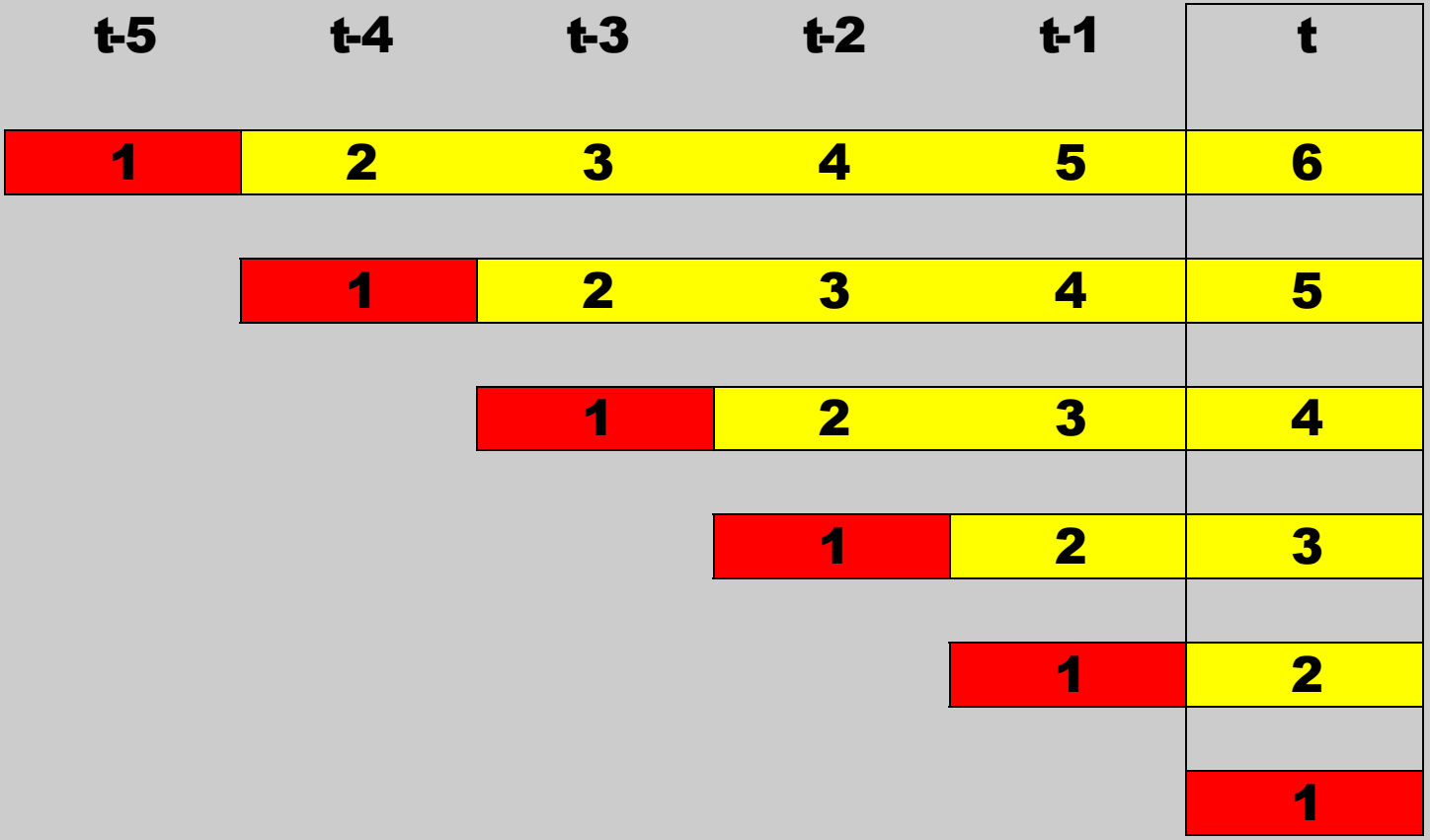


Panel Surveys

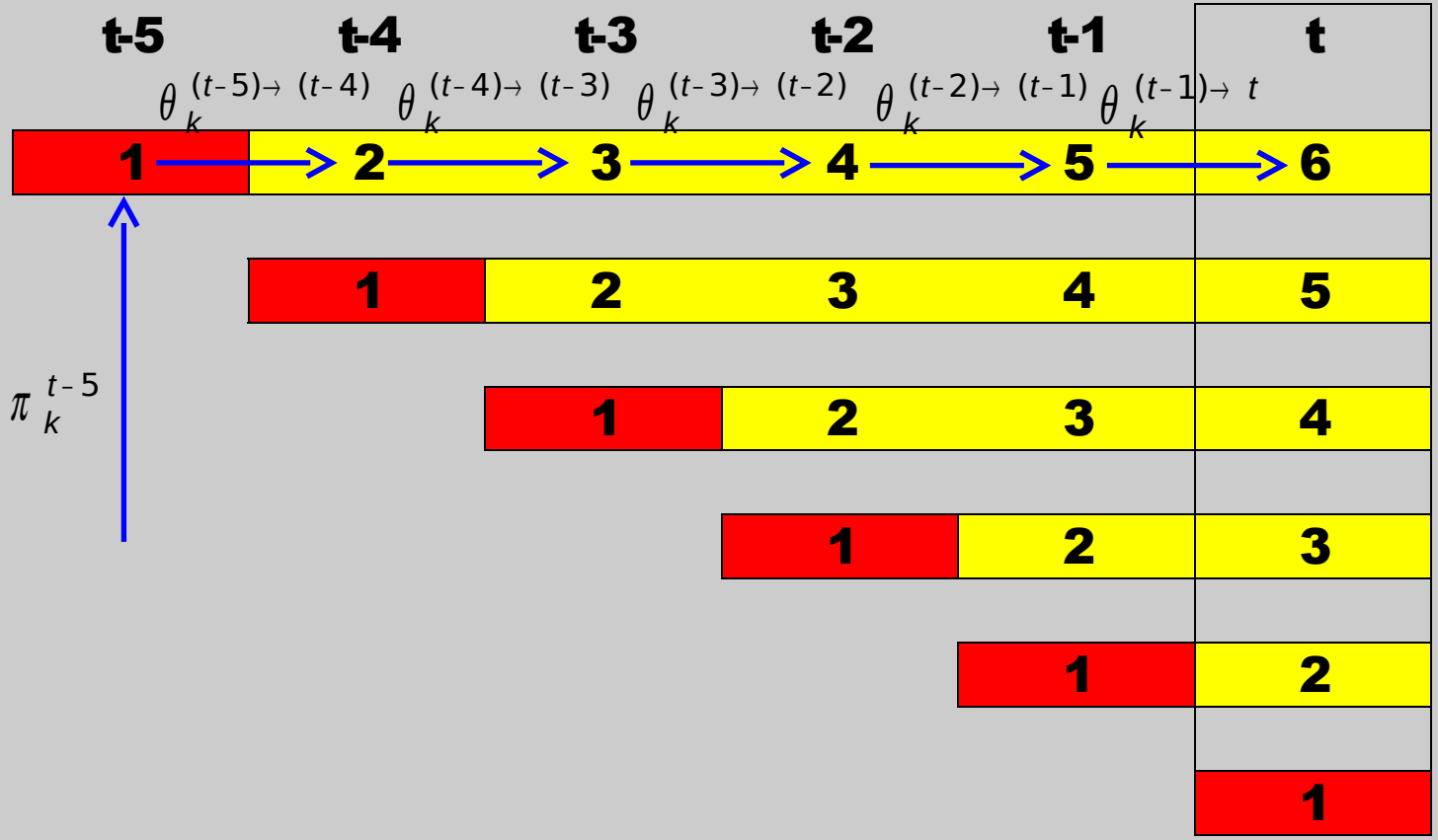
- Advantages of Panel Surveys:
  - Longitudinal analysis possible without retrospective surveys
  - More precise estimations of changes possible
  - Treatment of nonresponses: survey characteristics form recent waves as auxiliary characteristics

    example: telephone survey of ILO- activity status

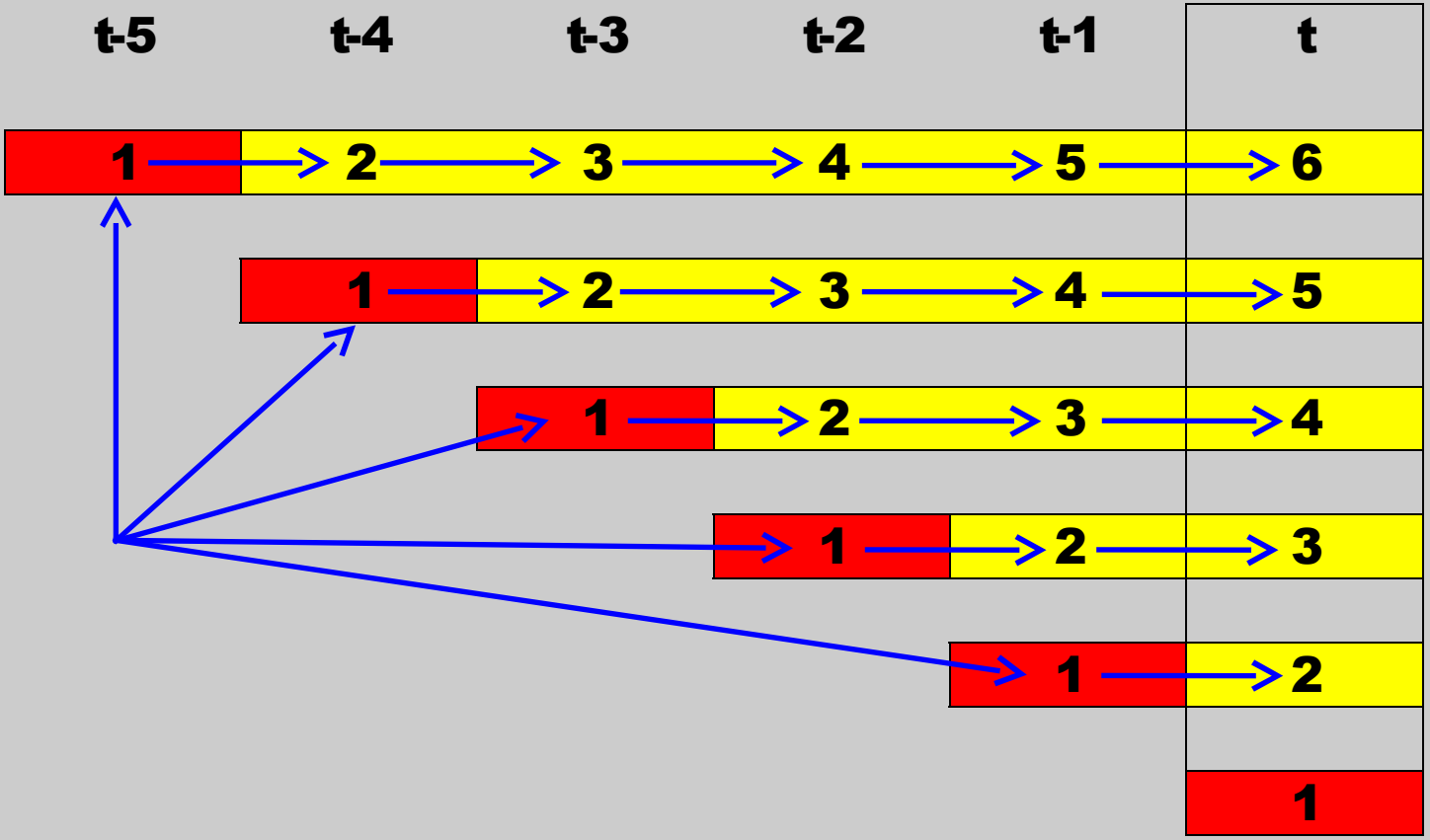
Composition of Sample in Time t



Inclusion Possibility for Panel Cases in Time t



Inclusion Possibility for Panel Cases in Time t



Nonresponses in Longitudinal Analysis

Estimation of re-participation probability by Logit-Model

$$\log\left(\frac{\theta_k^{(t-1)\to t}}{1-\theta_k^{(t-1)\to t}}\right) = \alpha + \sum_i \beta_i x_{ik}$$

with

$\theta_k^{(t-1)\to t}$ *: value of auxiliary characteristic* x_i *in previous* x_{ik} *month of person k*

Nonresponses in Longitudinal Analysis

Auxiliary characteristics X_i with highest influence on re-participation:

- Sex
- Age
- Nationality
- Activity status in previous month
- Residence (BIK-region size classes as well as East/ West)
- Qualifications, education
- Period of participation in the survey

Nonresponses in the Longitudinal Data

- Separate estimation of a Logit-Model

  for each intersection from month to month

- For panel cases calculation (estimation) of the probability to be included in the sample

  with the help of

  - Selection probability of the first-questioned
  - Participation probability for the following surveys

Panel Survey

- Disadvantages
  - Less precise estimation of total values as in independent cross-sectional survey of same size
  - Panel effects: survey characteristics depend on the amount of which a unit has been interviewed ⇒ bias
  - Panel mortality: data basis for longitudinal analysis decreases because of nonresponses in the course of time
  - For cross-sectional surveys additional samples from the inflows, because the population changes over time by in- and outflows.

Sampling Plan Microcensus

Sampling Frame

- Old Federal States
  population census 1987: number of apartments and persons,
  divided in municipality, street and house number
  (i.e. depending to buildings)

- New Federal States (since 1991)
  population register statistics of the DDR: number of households and persons per house number

Selection Units

- So called selection districts:
  artificially separated areas, which contain several buildings in generally close neighborhood; in cases of big buildings it contains only one building -complete or partial-

- For that reason allocation depends on building sizes
  - 1 to 4 apartments
  - 5 to 10 apartments
  - 11 and more apartments
  - Presumably shared accommodation
    (transferred from proportion of persons/apartments in the building)

Selection Units

- Depending on the size of the building, a different average number of apartments in each selection district were aspired

Building Size	Formation of the Selection District	Benchmark
1 – 4 apartments	several buildings	12 apartments
5 – 10 apartments	single buildings	7 apartments
11 and more apartments	building part	6 apartments
shared accommodation	single buildings or building part	15 persons

Selection Units

- Accrual in split building: generally floor by floor, in shared accommodation by name
- Cluster sample (area sample): all persons and households, which live in the area of the selected district, are statistically collected

Cluster effect

- For many survey characteristics higher random sample error than in simple random sample of persons / households
- Determining factors
  - Average cluster size
  - Dispersion within the clusters
  - Dispersion between the clusters

Sample Sizes

- Selection ratio of 1% each year
- About 45 000 selection districts
  (about 800 000 persons in more than 350 000 households)

Stratification

- Regional
  summary of counties or single counties, also single cities or parts of it.
  215 regional stratums
- 4 building size categories

Allocation of a Sample

- Proportional, i.e. an uniform selection of 1% in all stratums

same externalization probability for households and persons

Rotation

- Each year one quarter of samples is changed, i.e. one selected selection district remains 4 years in the sample

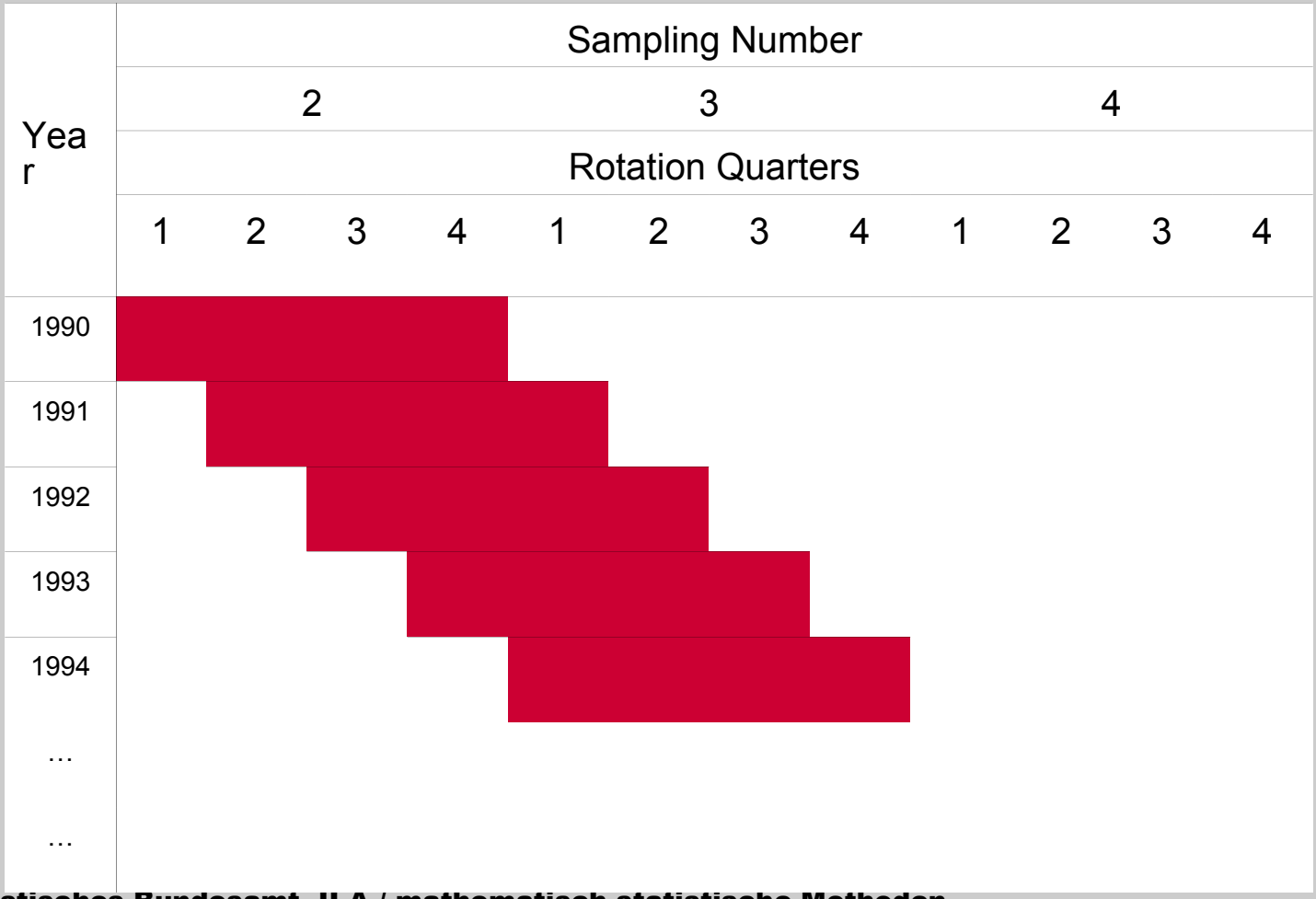
Arrangement

- Before the selection: regional arrangement of the selection districts in every stratum

Selection Method

- Decomposition of the selection entity in 100 1%-samples:
  each 100 succeeded selection districts („zones“) are provided with a random permutation of numbers from 0 to 99
- Decomposition of the selection entity in 4 rotation quarters:
  each 4 succeeded zones are provided with a random permutation of numbers from 1 to 4
- Decomposition of the selection entity in 48 „rotation weeks“:
  starting with a random start in each stratum the zones are consecutively modulo 48 numbered
- 20 samples are stored as „reserve samples“
- Starting with a random sample number reserve samples are consecutively depleted

Selection Method



Allocation of the Selected Districts during the Period

1. Quarter	
January	1, 13, 25, 37
February	5, 17, 29, 41
March	9, 21, 33, 45

4. Quarter	
October	4, 16, 28, 40
November	8, 20, 32, 44
December	12, 24, 36, 48

48 weekly marks in the inventory samples

2. Quarter	
April	2, 14, 26, 38
May	6, 18, 30, 42
June	10, 22, 34, 46

3. Quarter	
July	3, 15, 27, 39
August	7, 19, 31, 43
September	11, 23, 35, 47

yellow : **Interviewer package 1 (for 1. half of t**

white : **Interviewer package 2 (for 2. half of the month)**

Selection Method

- No random selection of fixed reporting weeks
- No fixed reporting week for the household: calendar week before interview respectively self-completion

Updating the Selection

- Annual additional selection from new buildings (construction activity statistics)
- Benchmark for accrual of sampling districts in each buildings uniformly 6 apartments
- No stratification in building size categories per regional stratum, i.e. only one stratum („new construction stratum“)
- Systematic selection with random start, sorting in updating year and regional aspect

Extrapolation Method

- Quarterly extrapolation in the federal states
- Bounded extrapolation: usage of auxiliary information for
  - Reduction of sample errors
  - Reduction of systematic errors, particularly nonresponse
- 2 steps :
  - Handling of nonresponse
  - Adaption to known benchmarks of the population („calibration“)

Method

- Regression estimation
  („Generalized Regression Estimator“, GREG)
  linear estimator for a total value: t_Y

$$\hat{t}_y = \hat{t}_{y,HT} + \hat{\mathbf{B}}' \times (\mathbf{t}_x - \hat{\mathbf{t}}_{x,HT})$$

$$= \sum_{k=1}^{n} \left(1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x,HT})' \left(\sum_{k=1}^{n} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k \right) d_k y_k$$

$$= \sum_{k=1}^{n} w_k y_k$$

Method

with

y_k : value of the survey characteristic ***y*** for person ***k***

n : (net) sample size

$$\hat{t}_{y,HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k \hat{\theta}_k} = \sum_{k=1}^{n} d_k y_k$$

π_k: inclusion probability for person k (π_k = 0,25%)

$\hat{\theta}_k$: estimated response probability

Method

$\mathbf{x}_k$: vector of all values of the calculation auxiliary characteristics for person k

$$\hat{\mathbf{t}}_{\mathbf{x},HT} = \sum_{k=1}^{n} d_k \mathbf{x}_k$$

$\mathbf{t}_x$: vector of total value of the auxiliary characteristics („benchmarks“)

$$\hat{\mathbf{B}} = \left(\sum_{k=1}^{n} d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k=1}^{n} d_k \mathbf{x}_k y_k \right)$$

Method

benchmarks are achieved, i.e. $\hat{\mathbf{t}}_x = \mathbf{t}_x$

Limitation of the correction factors g_k: $0{,}01 \leq g_k \leq 5$

(iterative method with damping factors)

Auxiliary Variables for Extrapolation

- Age (under 15/15 - 44/ ⌑ 45)
- Citizenship (German/ Turkish/ European/ non-European)
- Soldiers (soldiers/ people doing basic military service/ civil population)
- Regional accommodation stratum: nationwide 132 regional units with generally 500 000 inhabitants at least
- Month (from 2. quarter 2005 on)

Auxiliary Variables for Extrapolation Model

regional level	extrapolation terms	Quelle
federal state	▪ age (under 15, 15 – 44, 45 and older) differentiated according to sex	LBF
	▪ citizenship (german, turkish, EU, non EU) differentiated according to sex	AZR
	▪ number of soldiers incl. Federal Border Guard and riot police, number of persons doing basic military service, civil population	BMV‘g
	▪ total population per month	BMI BGS LBF
administrative district	▪ citizenship (german, non-german) differentiated according to sex	LBF
reg. accommodation stratum	▪ total population	LBF

Auxiliary Variables for Extrapolation

- For person ***k*** : mean value of the household, to which person ***k*** belongs

Same extrapolation factor for all persons of a household

Treatment of the Unit-Nonresponse

(„compensation of nonresponses“)

- because of obligation to provide information only 5% Unit-nonresponse in the annual average
- nonresponse mechanism at household level
- auxiliary variables at Unit-Nonresponse (observations of the interviewer)
  - household size (1/2/3+); from 2. quarter 2005 on
  - household German/ non-German
  - Only for 1-Person households: sex and age (<60/60+); from 2. quarter 2005 on
  - main residence/ secondary residence
- Item-Nonresponse in the auxiliary variables !

Treatment of the Unit-Nonresponse

- Auxiliary variables in Unit-Nonresponse
  - regional subgroup (nationwide 379 regional units)
  - new construction stratum (yes/no)
  - rotation quarter
- Regression estimation by summating the auxiliary variables across the gross sample as benchmarks (extrapolation from net to gross sample)

Treatment of the Unit-Nonresponse

- The inverse value of calculated „compensation factor“ can be

  interpreted as an estimate for the response probability $\hat{\theta}_k$

⇒ **Input weight for actual extrapolation**

Compensation Model for the known Nonresponses

1. Households

regional level	compensation terms
federal state	▪ **rotation quarter** ▪ **new construction stratum (yes/no)**
regional accommodation stratum	▪ **quantity of household (1, 2, > 3-Person household)** ▪ **citizenship of household reference person (german/non-german)** ▪ **residence of household reference person (main res./sec. res.)** **additional for 1-Person households:** ▪ **sex**
regional subgroups	▪ **age (under 60 / 60 years and older)** ▪ **private households**

Compensation Model for the known Nonresponses

2. **Collective Accommodations**

regional level	compensation terms
administrative district	▪ **entire people**

Extrapolation Factors

- Extrapolation factors quarter (EF951)
  - Only for quarterly analysis;
    Filter respectively classification variable reporting quarter (EF12)
  - In 1000; for estimations of total values multiply by 1000
- Extrapolation factor year (EF952)
  - For annual evaluations
  - EF952 = 1/4 * EF951
  - In 1000; for estimations of total values multiply by 1000

Variance Estimation

Conversion of the regression estimation:

$$\hat{t}_y = \sum_{i=1}^{n} w_k y_k = \sum_{k \in U} \hat{\mathbf{B}}' \mathbf{x}_k + \sum_{k=1}^{n} w_k \left(y_k - \hat{\mathbf{B}}' \mathbf{x}_k \right)$$

consequent:

$$\mathrm{Var}\left(\hat{t}_y \right) = \mathrm{Var}\left(\sum_{i=1}^{n} d_k \left(\frac{w_k \varepsilon_k}{d_k} \right) \right) = \mathrm{Var}\left(\sum_{i=1}^{n} d_k z_k \right)$$

⇒ **Variance formulas applicable for free extrapo**

Variance Estimation

$$\mathrm{Var}\left(\hat{t}_y\right) = \sum_h \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\frac{1}{n_h - 1}\left[\sum_{i \varepsilon S_h} z_{hi}^2 - \frac{1}{n_h}\left(\sum_{i \varepsilon S_h} z_{hi}\right)^2\right]$$

With

N_h : Number of selection districts of stratum h in the population (stratum extend)

n_h : Number of selection districts of stratum h in the sample (sample size)

S_h : Amount of sampling selection districts in stratum h

z_{hi} : Sum of weighted residues of all persons in the selection district i of stratum h

Variance Estimation

- Model for the relation between the extrapolated total value $\hat{n}_g$ of a table cell g and its squared relative standard error

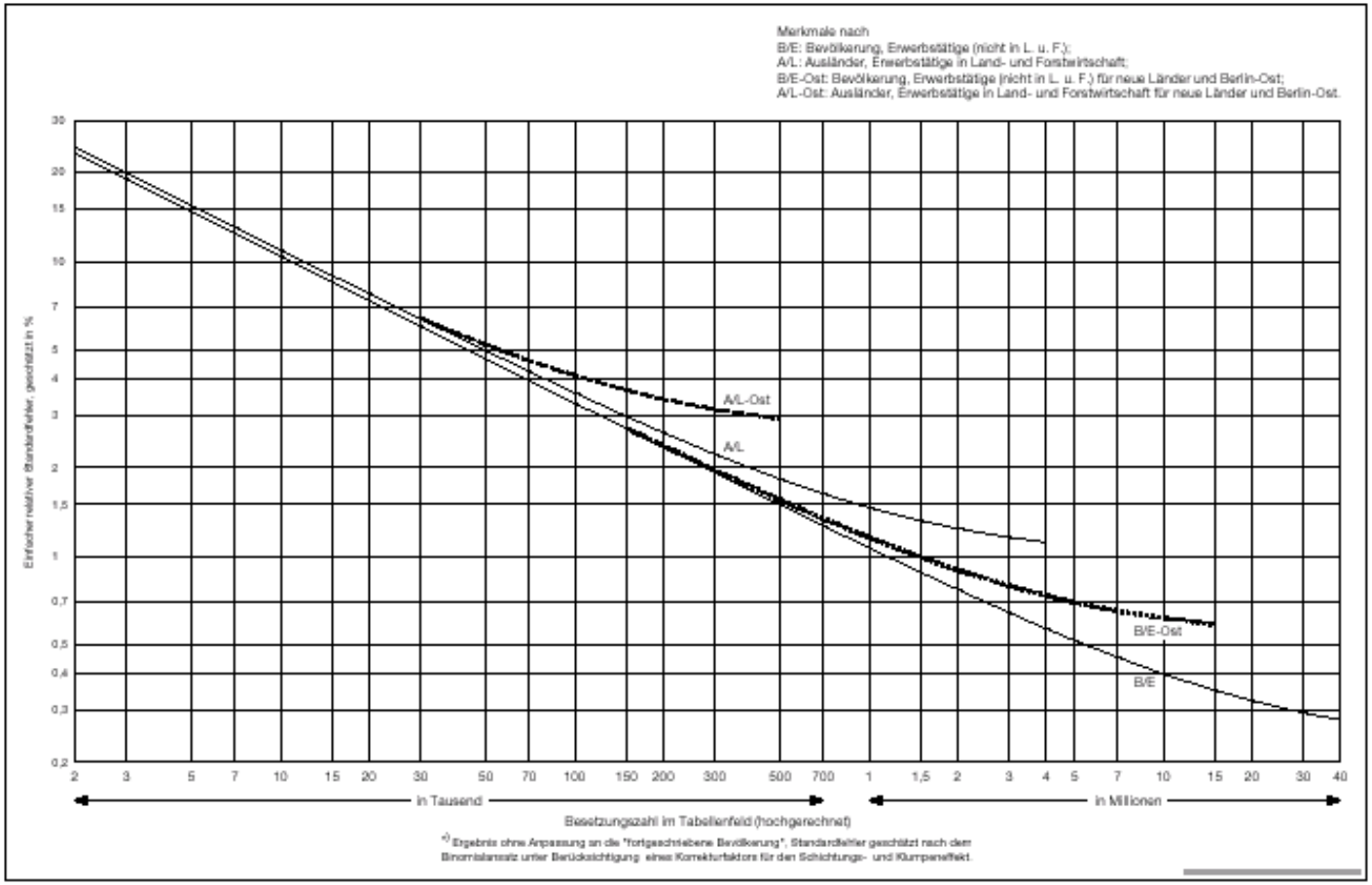
$$\hat{v}_g^2 = a + \frac{b}{\hat{n}_g} + e_g$$

- Estimation of the parameters a and b by regression with the results of the error calculation
- Regression divided by table cells for
  - foreigners or employees in agriculture and forestry
  - other population

  in each case for federation/west/east

Übersicht 4

Einfacher relativer Standardfehler einer 1%-Mikrozensusstichprobe*)



*) Ergebnis ohne Anpassung an die "fortgeschriebene Bevölkerung", Standardfehler geschätzt nach dem Binomialansatz unter Berücksichtigung eines Korrekturfaktors für den Schichtungs- und Klumpeneffekt.