MPIDR-NES Training Programme

Moscow, New Economic School,14th January - 1st February 2013

Population and Health

Лекция 10: Измерение межгрупповых различий в смертности с использованием связанных с переписью данных.

Lecture 10: Inter-group mortality differences based on the census-linked data

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC
RESEARCH

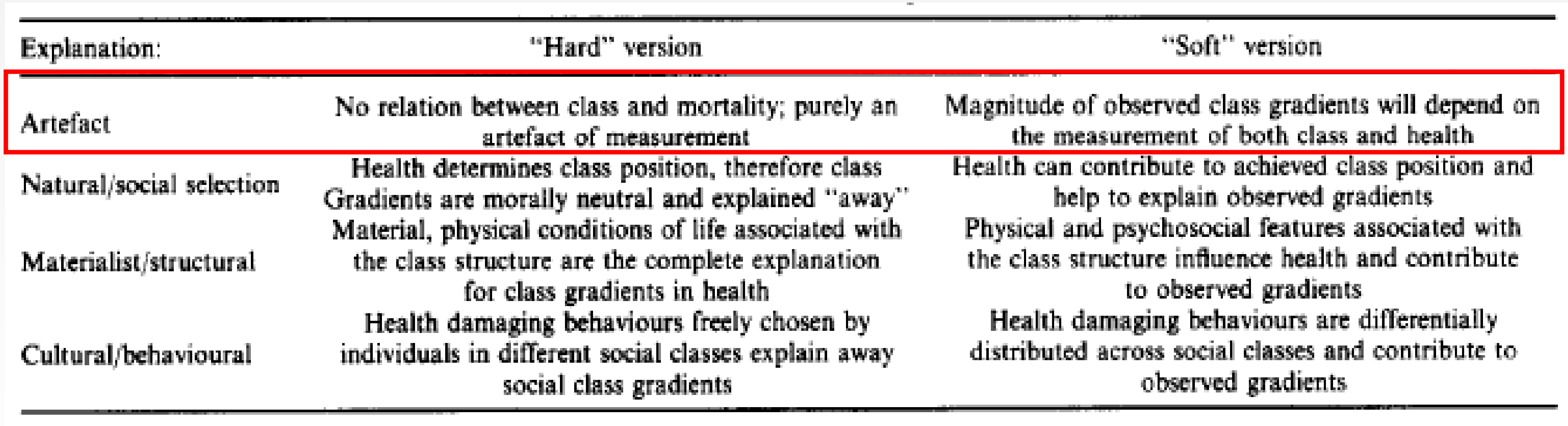
MAX-PLANCK-INSTITUT
FÜR DEMOGRAFISCHE
FORSCHUNG

РЭШ
Российская
экономическая
школа

Outline of Lecture 10

- ❖ **Measurement issues**
- ❖ **Illustrations of various problems related to measuring inter-group health inequalities**
- ❖ **Census-linked design**
- ❖ **Methods of analyses**

Explanations of health inequalities: “artefact” explanation



Explanation:	"Hard" version	"Soft" version
Artefact	No relation between class and mortality; purely an artefact of measurement	Magnitude of observed class gradients will depend on the measurement of both class and health
Natural/social selection	Health determines class position, therefore class Gradients are morally neutral and explained "away"	Health can contribute to achieved class position and help to explain observed gradients
Materialist/structural	Material, physical conditions of life associated with the class structure are the complete explanation for class gradients in health	Physical and psychosocial features associated with the class structure influence health and contribute to observed gradients
Cultural/behavioural	Health damaging behaviours freely chosen by individuals in different social classes explain away social class gradients	Health damaging behaviours are differentially distributed across social classes and contribute to observed gradients

Source: Macintyre, 1997.

Study designs used to measured health inequalities

Valkonen (1993):

1. Cross-sectional "unlinked" studies

This approach is based on separate tabulations of deaths and population at risk:

a) deaths are obtained from vital registry and are classified according to the last socio-economic status stated on the death certificates.

b) the data population at risk by socio-economic status are obtained from censuses taken in the middle of the period covered by death records.

2. Census-linked records studies

This approach is based on the linking the death records with the records from an earlier census.

a) deaths and population at risk (person years) by socio-economic group are classified according to the uniform source – population census.

b) availability of census and precise survivorship data (exact dates of death and emigration) allows estimating exact numbers of person-years lived during the period of observation.

3. Prospective epidemiological surveys

These surveys are designed to study specific risk factors of mortality and morbidity. They can also be used to study health inequalities (and their determinants). Usually small number of deaths, nationally not representative.

Difficulties of measuring health inequalities

Vallin (1979):

1. Imprecise nature of the object to be measured.
2. Changes in characteristics used for classification.
3. Discrepancies between sources for establishing numerator and denominator.
4. Selection of individuals by status (occupational, marital, etc.).

Valkonen (1993); Kunst et al. (1998):

5) Changes and differences in classifications.

6) Exclusion of economically inactive population (housewives, disabled, unemployed, retired).

Difficulties of measuring health inequalities

Vallin (1979):

1. Imprecise nature of the object to be measured.

There is no clear and uniform across country and in time definition of socio-economic class. There is no single *criterion* to classify population by socio-economic class.

2. Changes in characteristics used for classification.

Socio-economic status is variable characteristic. Three intermingled effects need to be measured: selection, effect of change in status, and effect of status itself.
The time before and after the change in status is of vital importance.

3. Discrepancies between sources for establishing numerator and denominator of the group-specific death rates (will be discussed further).

4. Selection of individuals by status (occupational, marital, etc.).

Part of the differences in mortality associated with certain statuses may be attributed to the bias due to selection phenomena. This is due to the fact that possession of a certain status is associated to a choice predetermined by individual's health.

Difficulties of measuring health inequalities

Valkonen (1993); Kunst et al. (1998):

5) Classifications and definitions.

Across different countries and time, various schemes have been used to classify people according to their socio-economic status (education, occupation, employment status, …). Results may depend on the classification scheme used.

Example: Age standardized mortality rates for external causes of death of skilled and of unskilled manual workers as estimated by the two classification schemes, Sweden, 1980-86.

SMRs estimated by the EGP algorithm		SMRs estimated by the GLT algorithm	
Skilled workers	Unskilled workers	Skilled workers	Unskilled workers
110	**133**	**124**	**117**

6) Exclusion of economically inactive population (disabled, unemployed, retired, housewives).

Their exclusion from analysis often leads to an underestimation of differentials, because economically inactive men frequently belong to lower occupational classes. Possible solutions: classifying economically inactive females according to the socioeconomic status of their husbands, using prior occupation information, adjustment procedures, etc.

Difficulties of measuring health inequalities

Exclusion of economically inactive population: an illustration

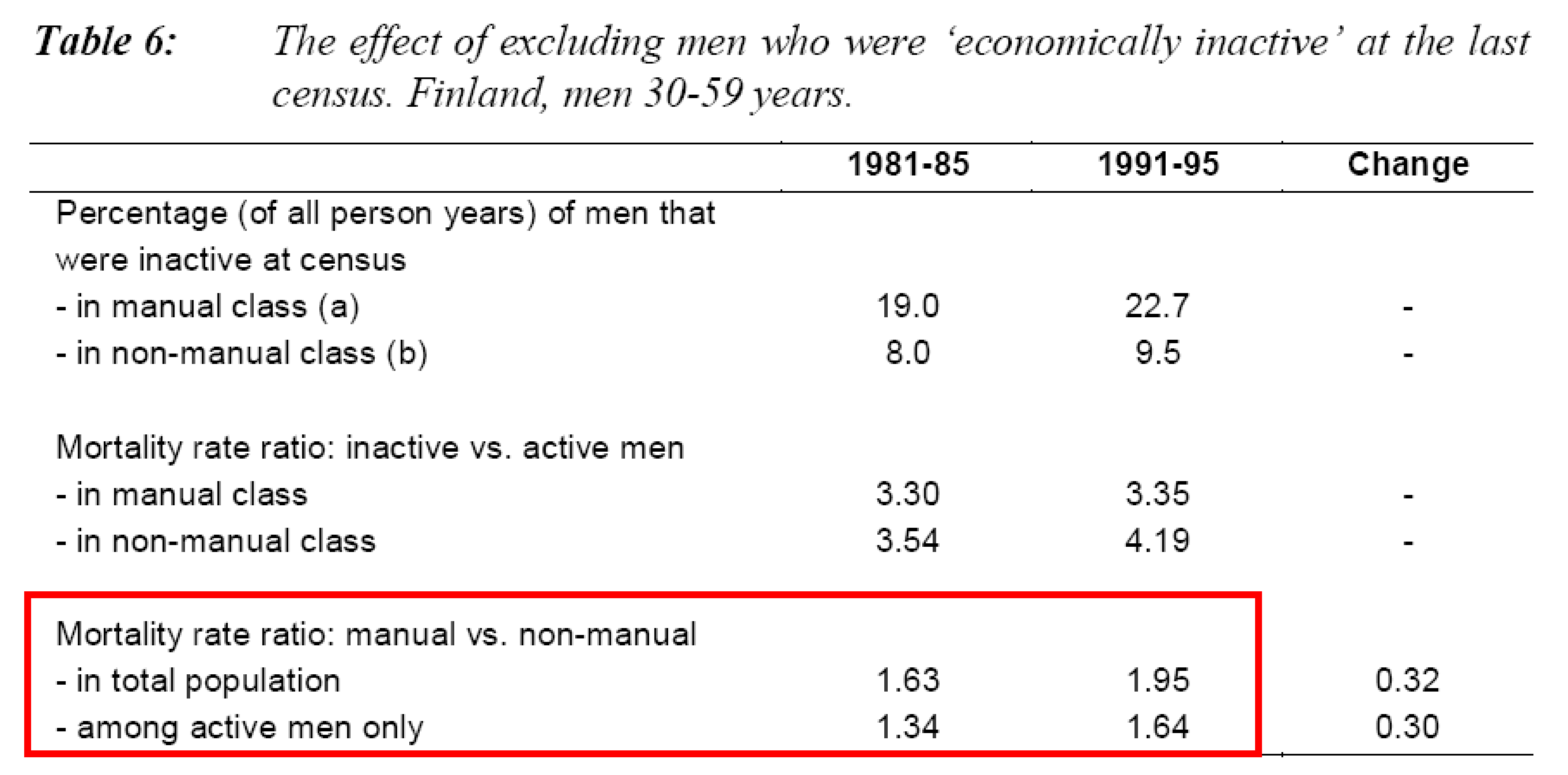


Table 6: *The effect of excluding men who were 'economically inactive' at the last census. Finland, men 30-59 years.*

	1981-85	1991-95	Change
Percentage (of all person years) of men that were inactive at census			
- in manual class (a)	19.0	22.7	-
- in non-manual class (b)	8.0	9.5	-
Mortality rate ratio: inactive vs. active men			
- in manual class	3.30	3.35	-
- in non-manual class	3.54	4.19	-
Mortality rate ratio: manual vs. non-manual			
- in total population	1.63	1.95	0.32
- among active men only	1.34	1.64	0.30

Source: Kunst et al., 2004.

Problems around “unlinked” cross-sectional data: numerator – denominator bias

Unlinked cross-sectional studies are the most widely used source of data on socio-economic inequalities in health. These studies typically use two sources of information on socio-economic status:

1. Deaths by socio-economic status are distributed according to the information given in the death certificates (based on the reports of proxy informants – relatives, officials, …);
2. Population at risk by socio-economic status is distributed according to the information given in the census records (based on self-reported information by individual him-/herself).

The major difference – in the census information is always reported by individual himself, whereas at death it always made by a third party.

The numerator – denominator bias occurs when the information about the status indicated by individual himself in the census and reported information about the same individual (e.g. by relative) in the death certificate is different.

Usually it is assumed that self-reported census information is more reliable:

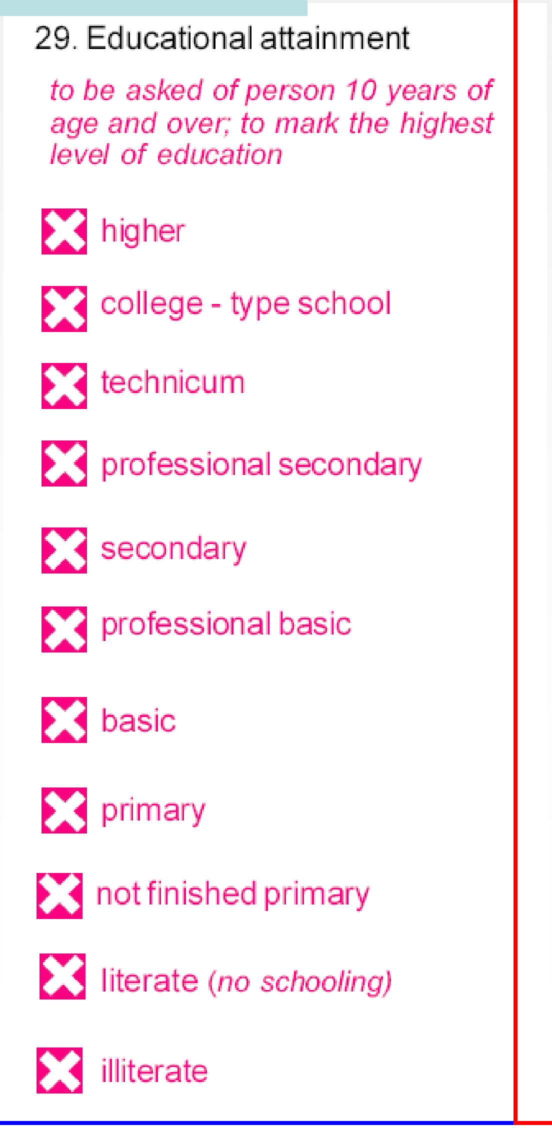
- because census questions about socio-demographic status are usually more detailed and accurate.
- because information reported in the death certificate may be biased due to variety of reporting inaccuracies such as “promoting the dead” phenomenon.

Problems around “unlinked” cross-sectional data: numerator – denominator bias

Census information

- To be reported by individual him/herself;

- refers to the highest attained education (confirmed by diploma).



29. Educational attainment

to be asked of person 10 years of age and over; to mark the highest level of education

☒ higher

☒ college - type school

☒ technicum

☒ professional secondary

☒ secondary

☒ professional basic

☒ basic

☒ primary

☒ not finished primary

☒ literate *(no schooling)*

☒ illiterate

Death record information

Education:

1. Higher
2. Secondary or specialized secondary
3. Basic
4. Primary
5. Unknown

- To be reported by proxy informant;

- the person just should mark one of the five categories (there is no requirement about the highest attained or completed education).

Problems around “unlinked” cross-sectional data: numerator – denominator bias

Example 2. Effect of misreporting in death records on education-specific life expectancy estimates

Male life expectancy at age 30 by education in Lithuania, 2001-2004. “Unlinked” vs. “linked” estimates

	"Unlinked"	"Linked"
High	47.0	45.5
Secondary	39.4	39.4
Lower than secondary	32.3	34.2
Difference	***14.7***	***11.3***

Source: Shkolnikov, Jasilionis, Andreev et al., 2007A.

Problems around “unlinked” cross-sectional data: numerator – denominator bias

Example 3. Effect of misreporting in death records on occupation-specific mortality estimates

Kunst et al. (1998)

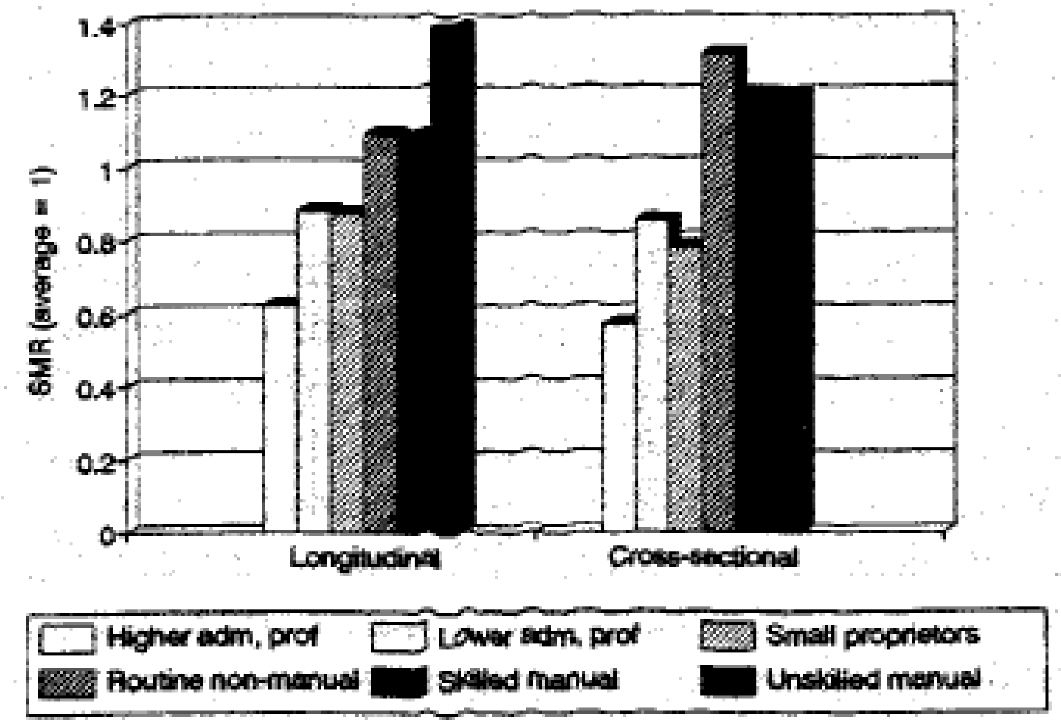


FIG. 1. — Age standardised mortality rate by occupational class. Estimates based on longitudinal and unlinked cross-sectional studies. France, circa 1980-1984, men circa 45-59 years at death.

Problems around “unlinked” cross-sectional data: numerator – denominator bias

The Hispanic paradox in the USA

Despite disadvantages in economic and living conditions, the Hispanic population shows similar or even better mortality situation than non-Hispanic White population (Rosenwaike, 1987; Elo et al., 2004).

Using the National Mortality Follow-Back Survey data, Swallen and Guend (2003) estimated that underreporting of the Hispanic ethnicity in death certificates was about 15% (based on the reports of next-of-kin informants of almost 23 thou. deceased in 1993). According to this study, adjusted life expectancy at birth for the Hispanic group was by almost 2 years lower than for the non-Hispanic White group.

But mortality study (linking the Current Population Survey records and death records for 1979-1998) confirmed that despite misreporting of ethnicity in death records, the Hispanic population had about 20% lower age-standardized death rates than non-Hispanic white population (Arias et al., 2010).

Problems around “unlinked” cross-sectional data: numerator – denominator bias

Effect of misreporting in death records on ethnicity-specific mortality differentials

Poisson regression mortality rate ratios for females aged 30 and over, calculated from the census-linked and unlinked mortality data

	“Unlinked”	“Linked”
Lithuanian	1.00	1.00
Polish	0.92* *(0.89-0.95)*	1.21* *(1.18-1.24)*

*p≤0.001. All models are adjusted for age.

Source: Jasilionis, D., Stankūnienė, V., Ambrozaitienė, D., Jdanov, D.A., Shkolnikov, V.M. (2011). Ethnic mortality differentials in Lithuania: contradictory evidence from census-linked and unlinked mortality estimates. *Journal of Epidemiology and Community Health*, doi:10.1136/jech.2011.133967 (in press).

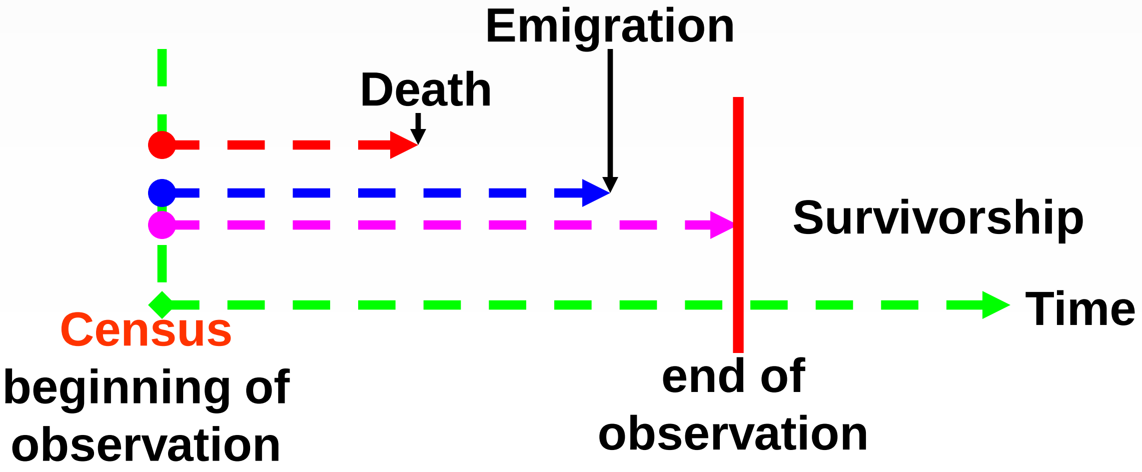
Census-linked data: advantages and design

The main advantage of the linkage between the census and death records:

Group-specific mortality indicators are calculated using an uniform source of information on socio-demographic status:

- Both deaths and population exposures are grouped according to the socio-demographic status indicated in the census.

- Census-based socio-demographic information for deaths are obtained through the linkage of death and census records. Sometimes, additional information is obtained from other registers.



Follow-up design

Census-linked data: some technical aspects

The means of linkage between death and census records:

- Ideally: personal identification numbers;
- Other personal data: name/surname, address, socio-demographic characteristics.

The success of linkage may vary across countries:

- From a low of 70% in Spain (Madrid region) and 90% in Austria to a high of 99% in Scandinavian countries.

Unlinked death records usually tend to concentrate among low socio-economic groups (with high mortality), therefore it is not recommended to exclude these deaths from analyses:

Lithuanian study shows that if 5% unlinked deaths were excluded from calculations, life expectancy at age 30 is overestimated by 1 year for men and by 0.5 year for women.

Source: Menvielle et al., 2008; Shkolnikov et al., 2007; Jasilionis et al., 2007.

Some disadvantages of the “linked” studies

1. **The major obstacle to conduct census-linked study: it involves work with individual-level confidential data (both census and death records).**

   Many countries (e.g. Germany) have strict regulations regarding protection of personal data (such as personal numbers or addresses). In many cases even statistical agencies cannot use such data to perform the linkage between census and death records.

   The census-linked data are available:
   Covers whole national populations: Finland, Sweden, Norway, Denmark, Belgium, Austria, Bulgaria, Lithuania, Slovenia.

   Covers part of the population: Switzerland (German speaking), Italy (Turin), Spain (Madrid region, Barcelona, Basque country).

   Representative samples: France and England & Wales (1% of population).

2. **Remaining methodological challenges (except numerator – denominator bias):** classifications, exclusion of economically inactive population, …

Census-linked datasets: individual and frequency formats

1. **Individual data format:** each row of the dataset corresponds to an individual from the census and includes all information on his / her socio-demographic status (date of birth, date of death (if occurred), education, marital status, place of residence, etc.). Theoretically it is possible to identify an individual from such information, therefore such individual data are usually unavailable for research.

2. **Frequency data format:** each raw represents combinations of available categories of socio-demographic variables. In addition, numbers of deaths and person years are given for each combination.

Examples of combination of socio-demographic variables:

1. Age 30, male, high education, married, living in the urban area.

2. Age 35, female, secondary education, divorced, living in the urban area.

Such datasets do not disclose personal information. All works with individual data are performed at statistical offices, whereas researchers get access only to the frequency (aggregated) datasets.

Census-linked dataset in frequency format

From individual to frequency data format (an example). The period of observation: 01.01.2000-31.12.2000

Individual data

Ind. Nr.	Year of birth	Month of birth	Year of death	Month of death	Sex	Education	Marital status
1	1950	6	.	.	Male	High	Married
2	1950	6	.	.	Male	High	Married
3	1950	6	.	.	Male	High	Married
4	1950	6	.	.	Male	High	Married
5	1950	6	2000	3	Male	High	Married
6	1950	6	2000	3	Male	High	Married
7	1950	6	.	.	Female	Secondary	Divorced
8	1950	6	2000	4	Female	Secondary	Divorced

Frequency data

Comb. Nr.	Age	Sex	Education	Marital status	Deaths	Person Years
1	49	Male	High	Married	2	2.25
2	50	Male	High	Married	0	2.17
3	49	Female	Secondary	Divorced	1	0.75
4	50	Female	Secondary	Divorced	0	0.54

Methods of analyses of the census-linked frequency data

1. **Classic measures of absolute and relative inequalities based on group-specific mortality rates** (non-parametric approach):

1. Group-specific life tables

2. Standardized death rates, rate differences, and mortality rate ratios

3. More sophisticated measures of inter-group inequalities:

Gini (inter-group)
Average inter-group difference (AID)
Index of Dissimilarity
…

To be discussed in the forthcoming lectures.

Methods of analyses of the census-linked frequency data

2. Regression-based (parametric) approach (mortality differentials expressed using regression coefficients)

Advantage: allows to control for confounding factors
(e.g. part of the differences in mortality by ethnicity maybe explained by other factors such as differences in education or income (between ethnic groups)).

Poisson regression model for count (e.g. frequency) data

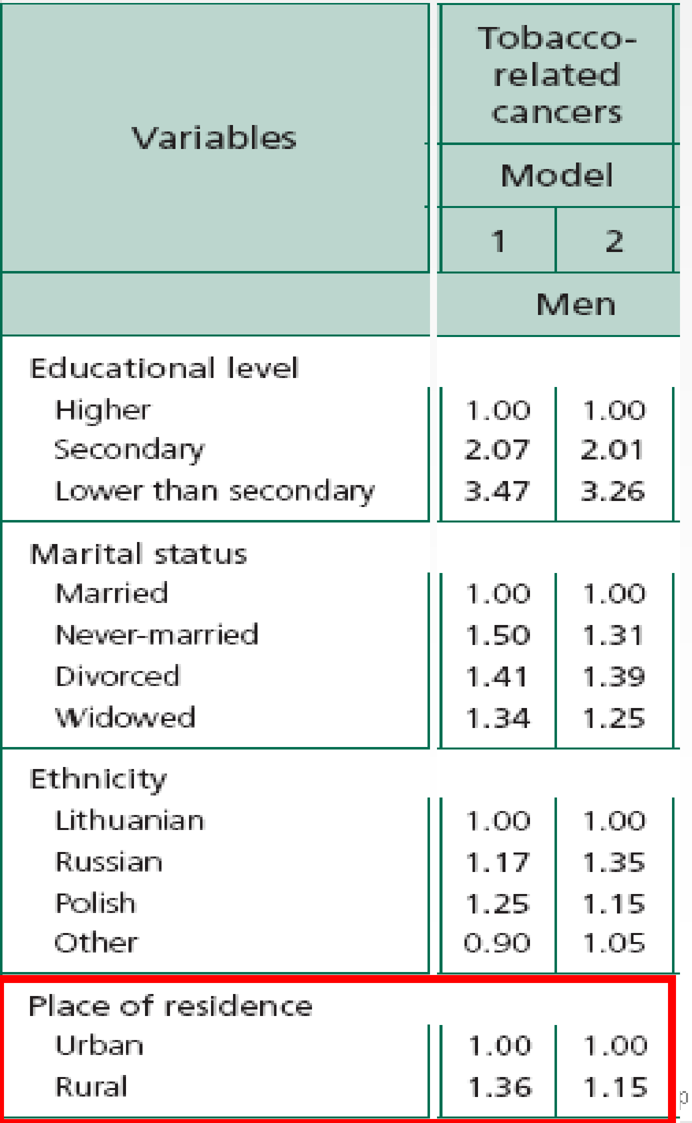
$$D_j = E_j e^{\beta_0+\beta_1 x_{1,j}+\ldots+\beta_k x_{k,j}} = e^{\ln(E_j)+\beta_0+\beta_1 x_{1,j+\ldots+}\beta_k x_{k,j}}$$

D – expected number of deaths, *E* - exposure where *j* is observation (e.g. combination in the frequency data) number.

The results are usually reported by rate ratios comparing mortality risk in the group under consideration to the mortality risk in the reference group (usually the highest socio-economic group).

Methods of analyses of frequency data

Results of Poisson regression on four variables for risk of dying from tobacco-related cancers. Lithuanian males, 2001-2004.



Variables	Tobacco-related cancers	
	Model	
	1	2
	Men	
Educational level		
Higher	1.00	1.00
Secondary	2.07	2.01
Lower than secondary	3.47	3.26
Marital status		
Married	1.00	1.00
Never-married	1.50	1.31
Divorced	1.41	1.39
Widowed	1.34	1.25
Ethnicity		
Lithuanian	1.00	1.00
Russian	1.17	1.35
Polish	1.25	1.15
Other	0.90	1.05
Place of residence		
Urban	1.00	1.00
Rural	1.36	1.15

Model 1: controlling for age.
Model 2: controlling for all the variables.

Jasilionis et al., 2007.