MPIDR-NES Training Programme

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

Population and Health

Лекция 7. Качество данных и статистические ошибки.
Lecture 7. Data quality and statistical errors.

MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH
MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG

РЭШ Российская экономическая школа

Data quality and statistical errors

PART 1

❖ Availability of mortality data across the globe

❖ Potential data problems:
-- coverage of the population;
-- numerator problems: completeness of death recording;
-- denominator problems: under- or over-stated population;
-- age misreporting.

❖ How to treat defective data? Model life tables

PART 2

❖Standard error and CI for age-specific death rates and age-specific death probabilities

❖ Standard error and CI for linear aggregates of age-specific death rates

❖ Standard errors of life- and health expectancies

Part 1. Data quality

Demographers have always paid a considerable attention to the data sources and quality of the data:

Three levels of explanation of a difference in health outcomes between two populations (Vaupel, 1995):

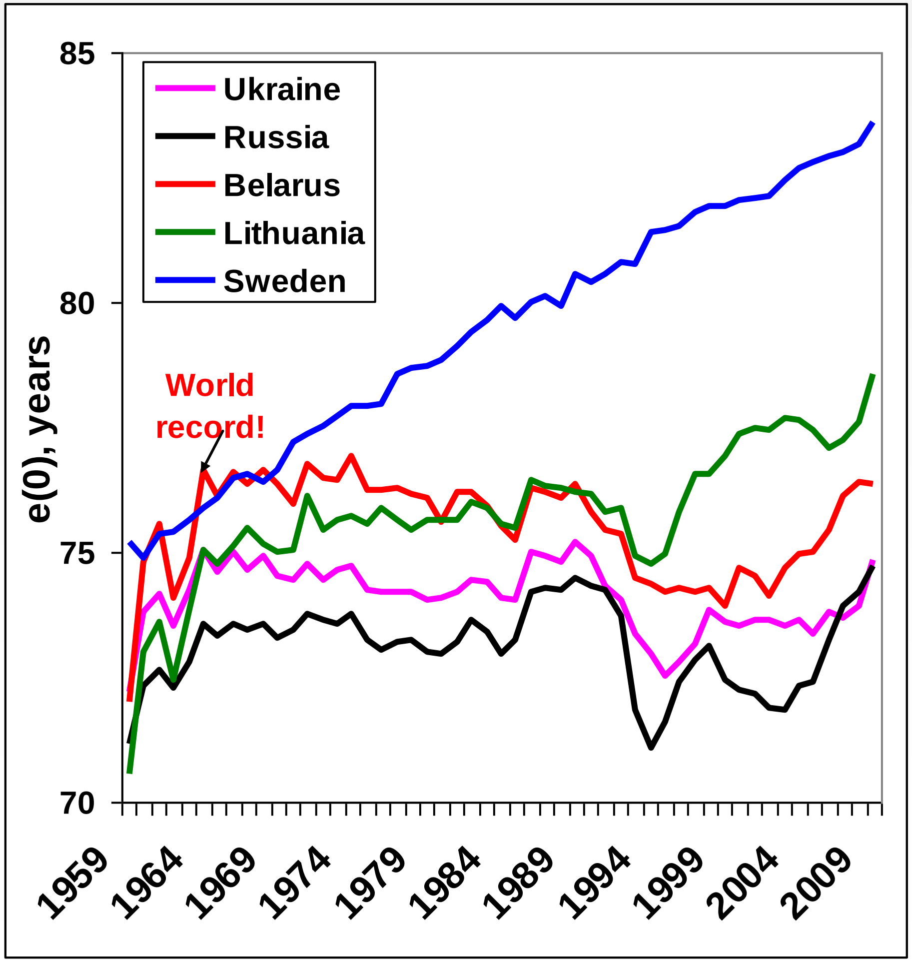
- "Level-0: explanation that the data are erroneous"!

- "Level-1: explanation that the observed population difference is produced by a corresponding difference at the individual level in the characteristics of interest;

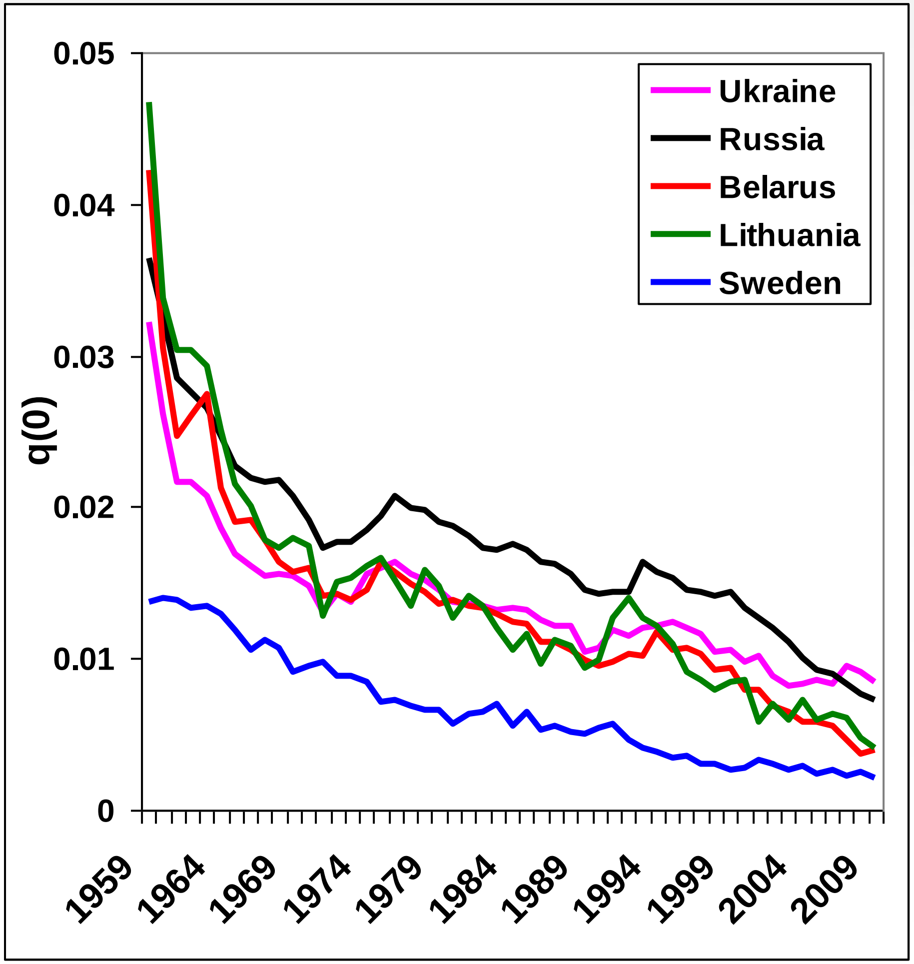
- "Level-2: explanation that the observed difference is attributable to a difference in a population structure, i.e. to a difference in the composition of the two populations with regard to characteristics other than the characteristics of interest.

Record levels of life expectancy: True or data quality issues?

Female life expectancy at birth



Probability of dying at age 0



Availability of mortality data across the globe

For most of the world population, complete and accurate data on mortality are not available.

To produce such data an expensive and well-organized system for registration of deaths and also censuses or population registers to count population are needed. This is something that majority of developing nations have been unable to achieve. Good death registration does not exist or is very fragmentary in most of the developing world including its most populated parts (China, India, Indonesia) and also in countries that are facing the greatest health challenges (Sub-Saharan Africa).

Vital registration that can be used to calculate life tables over the whole range of ages exist in about 55 countries. In about 15 to 20 of these countries, quality of these data is a serious concern. For other 35-40 countries, data quality can still be problematic during some time periods or at some ages. That is why one should care about data quality even when working on data from an industrialized country.

Types of data problems

$$M(age, country) = \frac{D(age, country)}{P(age, country)}$$

Coverage: a problem with „country“. Vital registration may not cover entire population.

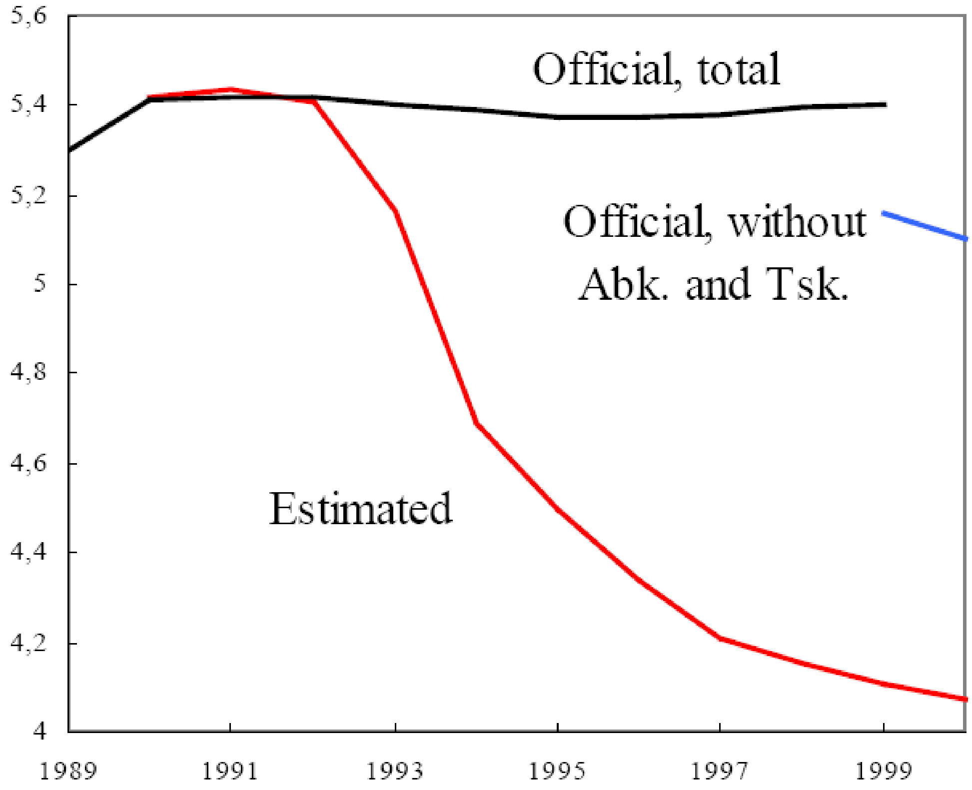
Completeness: problems with „D“. Not all deaths are registered. The under-registration is more likely among infants and very old people.

Denominator problems: problems with „P“. Population can be under- or over-estimated. Often it happens because of imprecise registration of migration.

Age misreporting: problems with „age“. Sometimes people misreport their age. It can happen if they there are no documents certifying date of their birth. In some cultures, old people tend to overstate their real age. Age of some deceased can be unknown or known only approximately.

Data quality: limited coverage evidence from data quality studies

Estimated trends in Georgian population size compared with the official data in 1989-2000 (in millions).



Alternative population estimates and additional data are used to detect the problem.

Overestimation of the population of Georgia in the 1990s due to massive unregistered out-migration and due to de-facto missing territories (Abkhasia, S. Osetia).

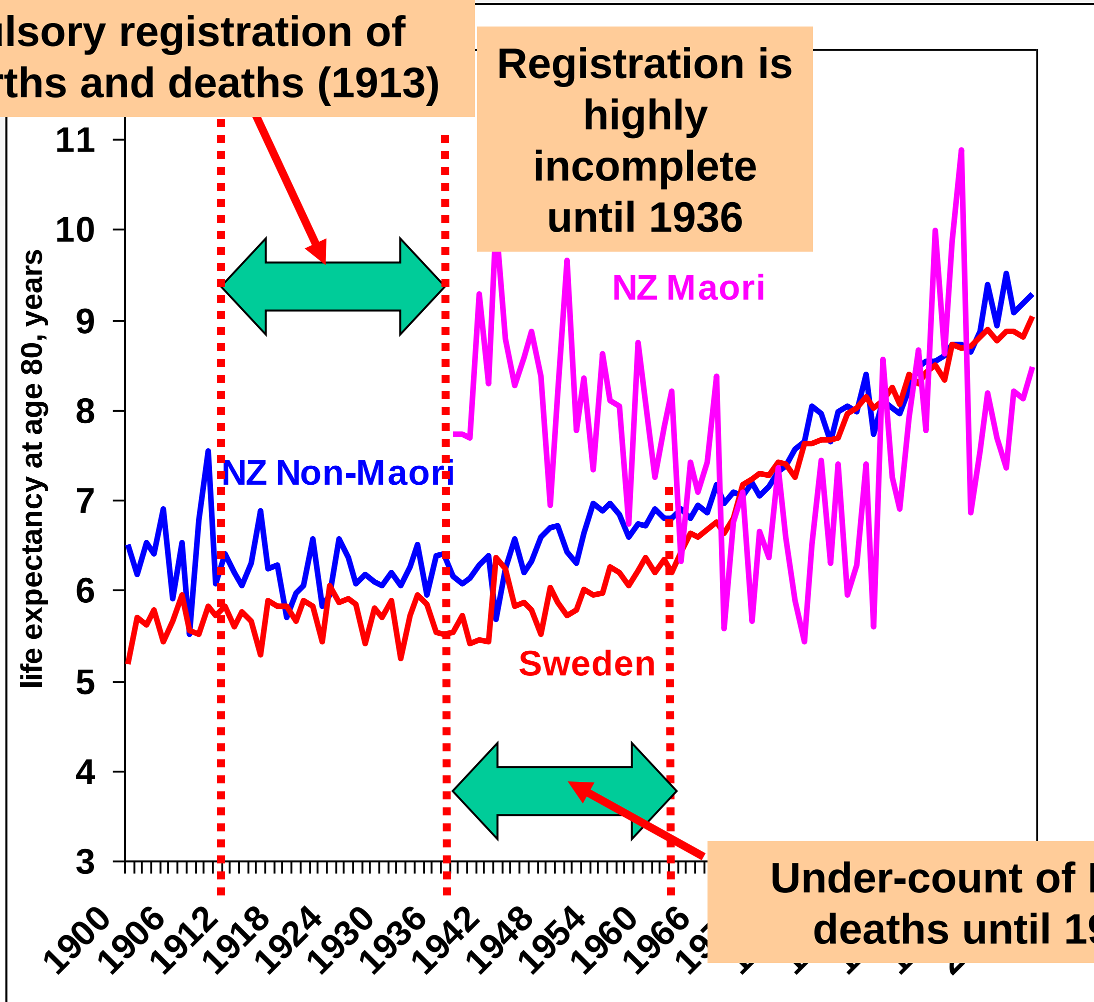
Source: Yeganyan et al., 2001.

Data quality checks: using additional information

Female life expectancy at age 80 for New Zealand Maori, New Zealand Non-Maori, and Sweden, 1900-2003.

Compulsory registration of Maori births and deaths (1913)

Registration is highly incomplete until 1936



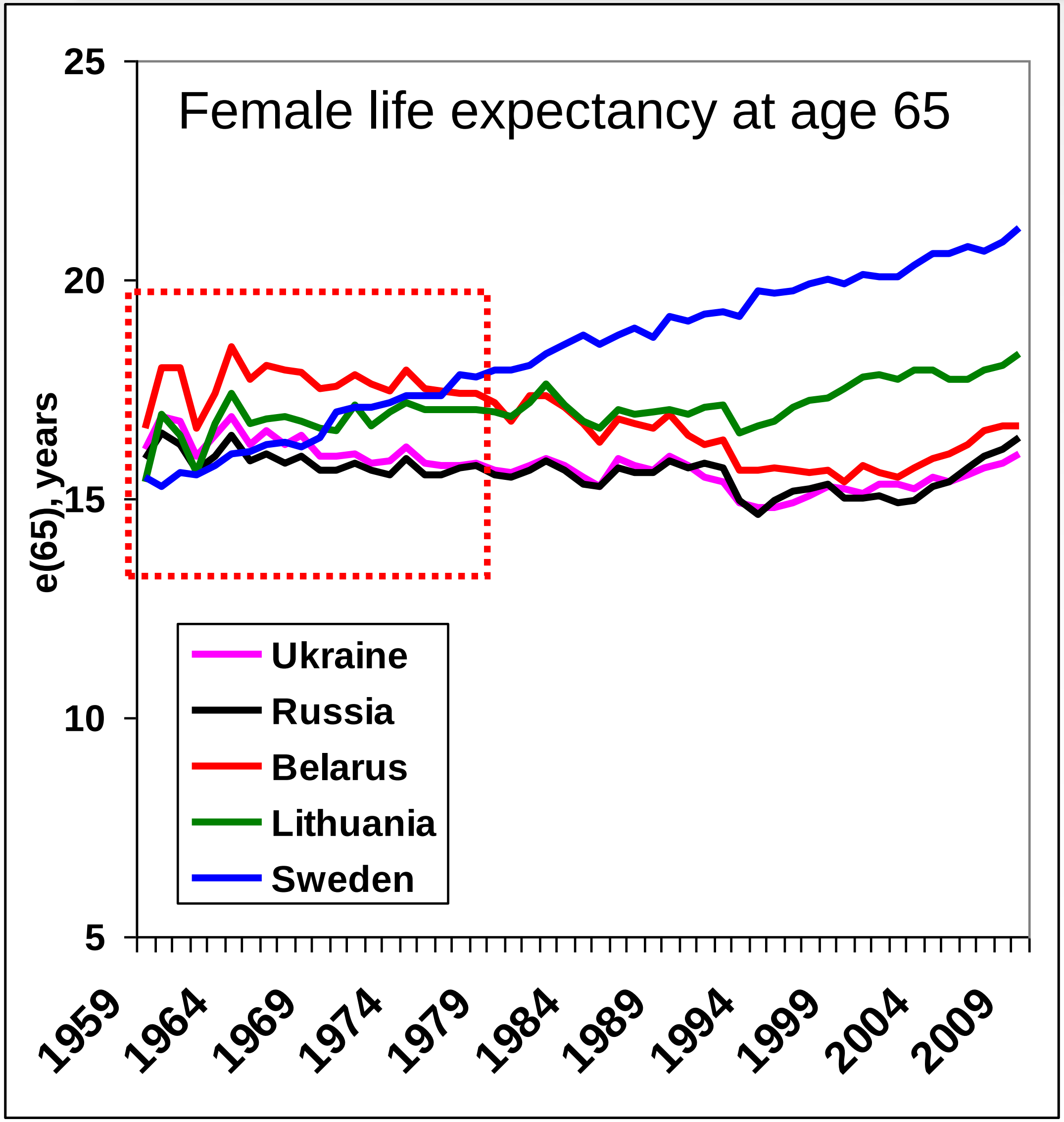
Under-count of Maori deaths until 1960

Until 1936,
New Zealand official statistics covered only Non-Maori population

To detect the problem, comparisons with reliable data and knowledge of relevant facts is used.

Source: The Human Mortality Database, 2007 (www.mortality.org).

Data quality checks: comparison to country with high quality data

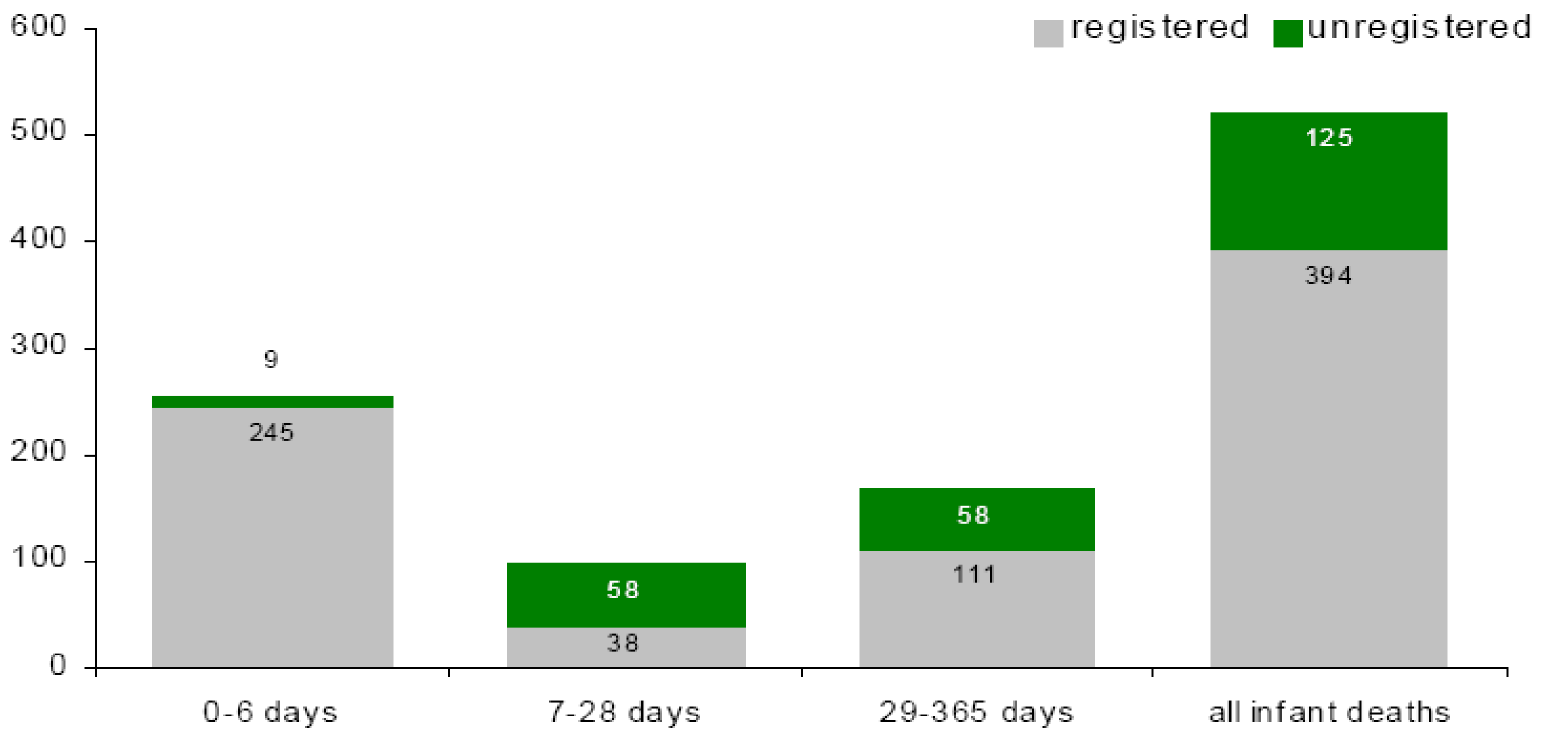


Comparison to country with high quality data helps to detect a problem.

Source: The HMD project, 2012.

Data quality checks: using alternative data sources

Figure 3: *Infant deaths in Armenia registered at medical facilities but unregistered at Civil Acts Registration Bureau, 2000 (absolute number from survey of 519 deaths)*



Source: Aleshina & Redmond, 2003.

An alternative and more precise data source is used.

Data quality issues: restrictive definition of live birth in USSR

Table 2: *Soviet and WHO definitions of live birth*

	Infant born after the end of the 28th week of pregnancy			
	No signs of life	No breath but other signs of life	Died during the first 7 days	Survived the first 7 days
USSR	**Stillbirth**		**Live birth**	
WHO	**Stillbirth**	**Live birth**		

	Infant born before the end of the 28th week of pregnancy, or with weight under 1,000 gr. or length under 35 cm			
	No sign of life	No breath but other signs of life	Died during the first 7 days	Survived the first 7 days
USSR	**Miscarriage**			**Live birth**
WHO	**Stillbirth**	**Live birth**		

Source: Anderson and Silver (1986).

This part of newborns were not counted as infant deaths nor as live births

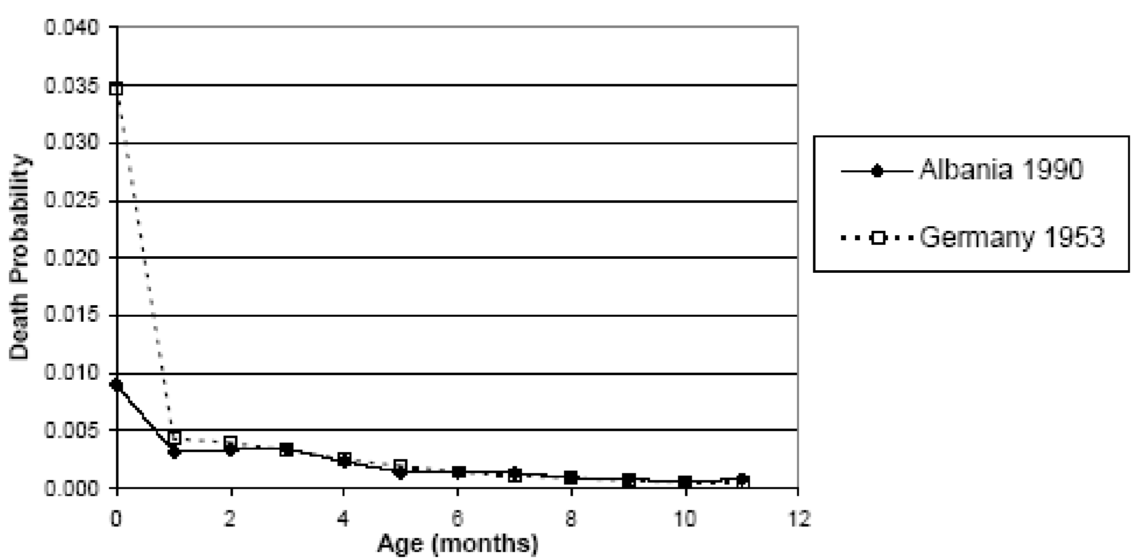
Information about functioning of registration system functions shows that part of infants with especially high risk of death is excluded from calculations.

Data quality: underestimation of infant mortality due to a restrictive definition of live birth and death undercount

	Average official infant mortality rate 1987-2000	Adjustment factor (per cent)	Adjusted infant mortality rate 1987-2000
Albania (4)	28.4	+110.9	59.83
Bulgaria (5)	15.1	+56.6	23.64
Croatia (6)	9.8	+0.8	9.90
FYR Macedonia (8)	27.9	+32.9	37.05
Romania (9)	23.5	+81.2	42.57

Adjustment is made by correction of the monthly mortality curves. The adjustment brings these curves to certain "golden-standard" curve.

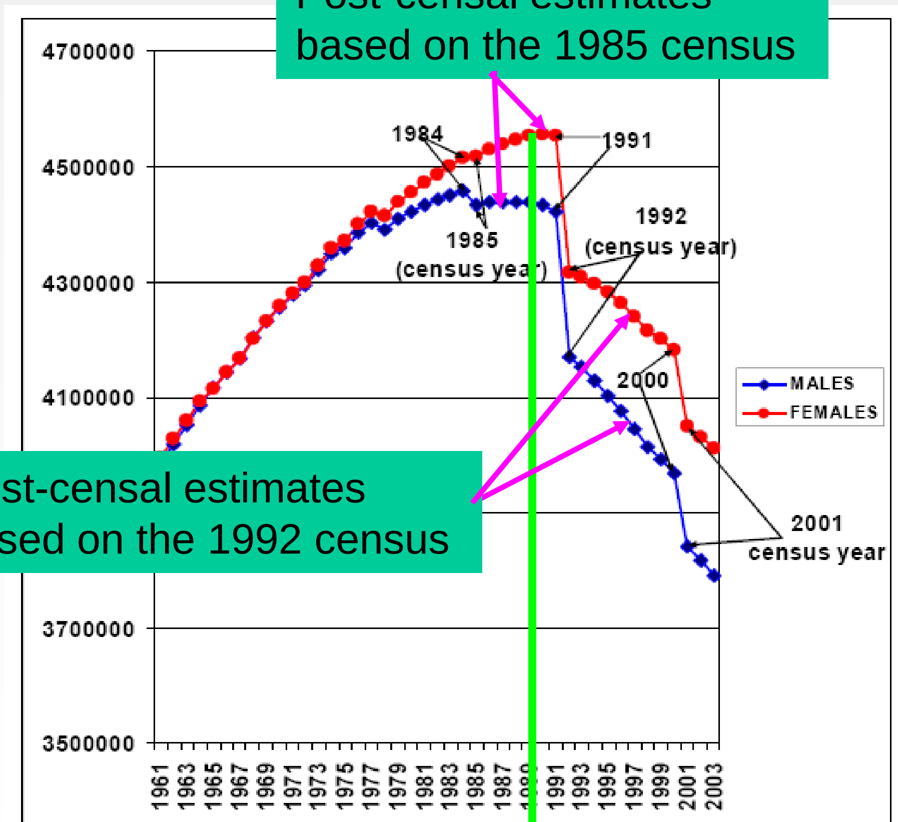
Figure 3. Monthly Probabilities of Dying in First Year of Life
Albania vs Germany



Source:
Kingkade& Sawyer, 2001.

Data quality: population overstatement of population produced by unregistered migration (Bulgaria in the HMD)

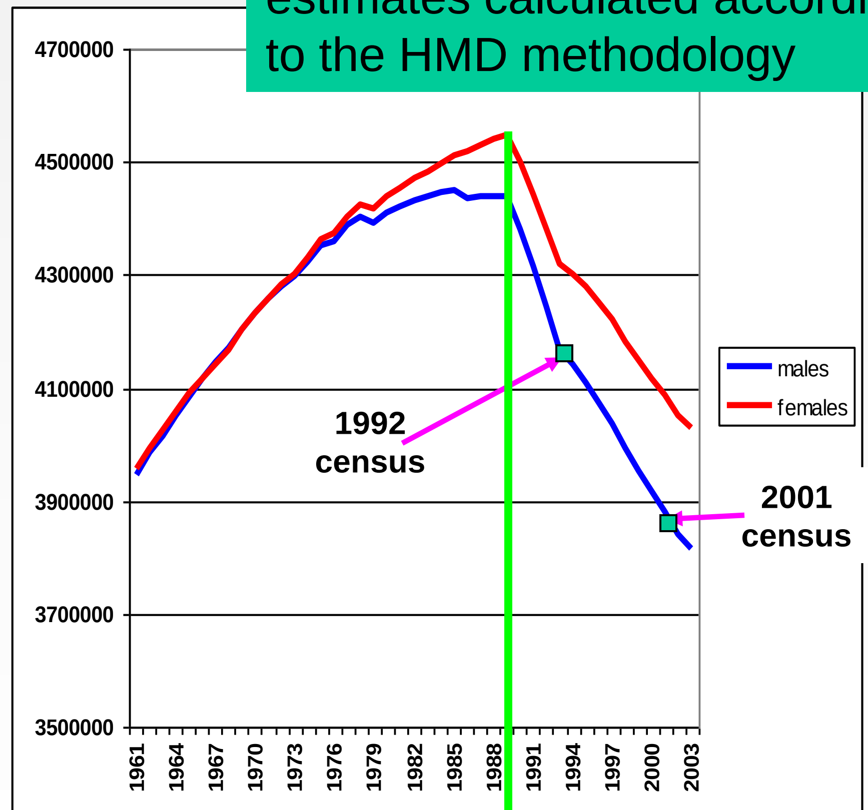
Post-censal estimates based on the 1985 census



Post-censal estimates based on the 1992 census

Release of restrictions for international migration in 1989 led to large (unregistered) emigration waves in the subsequent years 1990-1991.

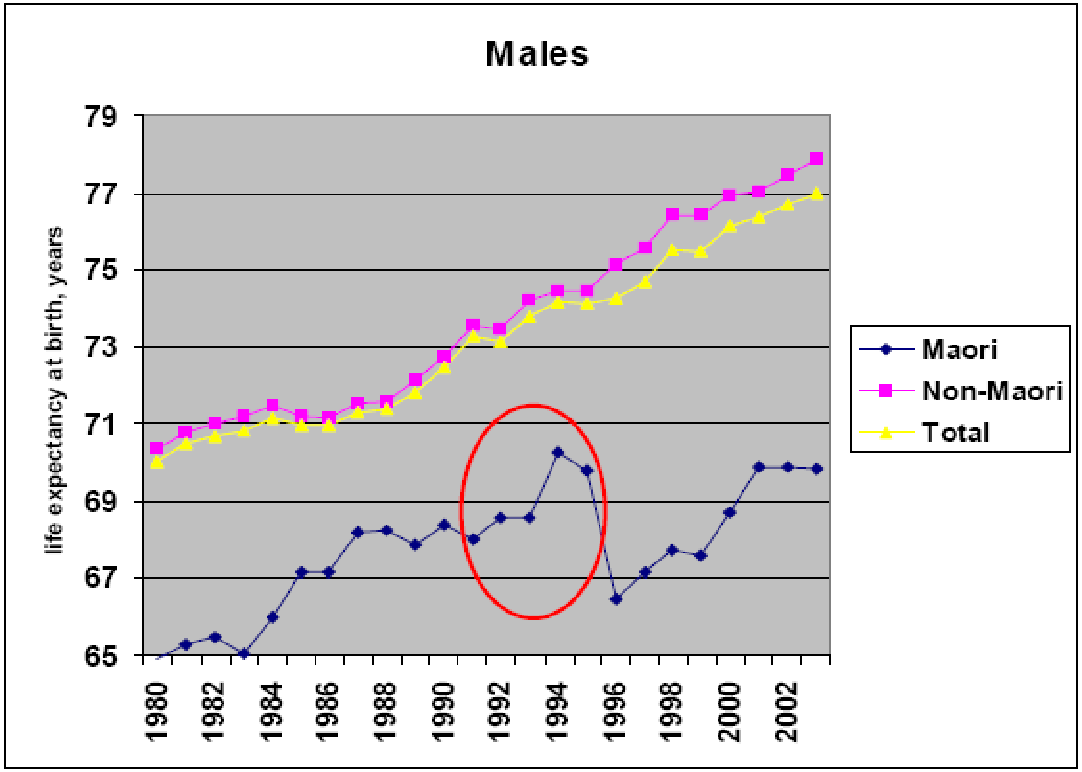
New ***inter-censal*** population estimates calculated according to the HMD methodology



Release of restrictions for international migration in 1989

Data quality: problem with inter-censal population estimates due to a change in definition of ethnicity (New Zealand Maori in the HMD)

Figure 4. Life expectancy at birth of Māori population calculated from the official (unadjusted) data.



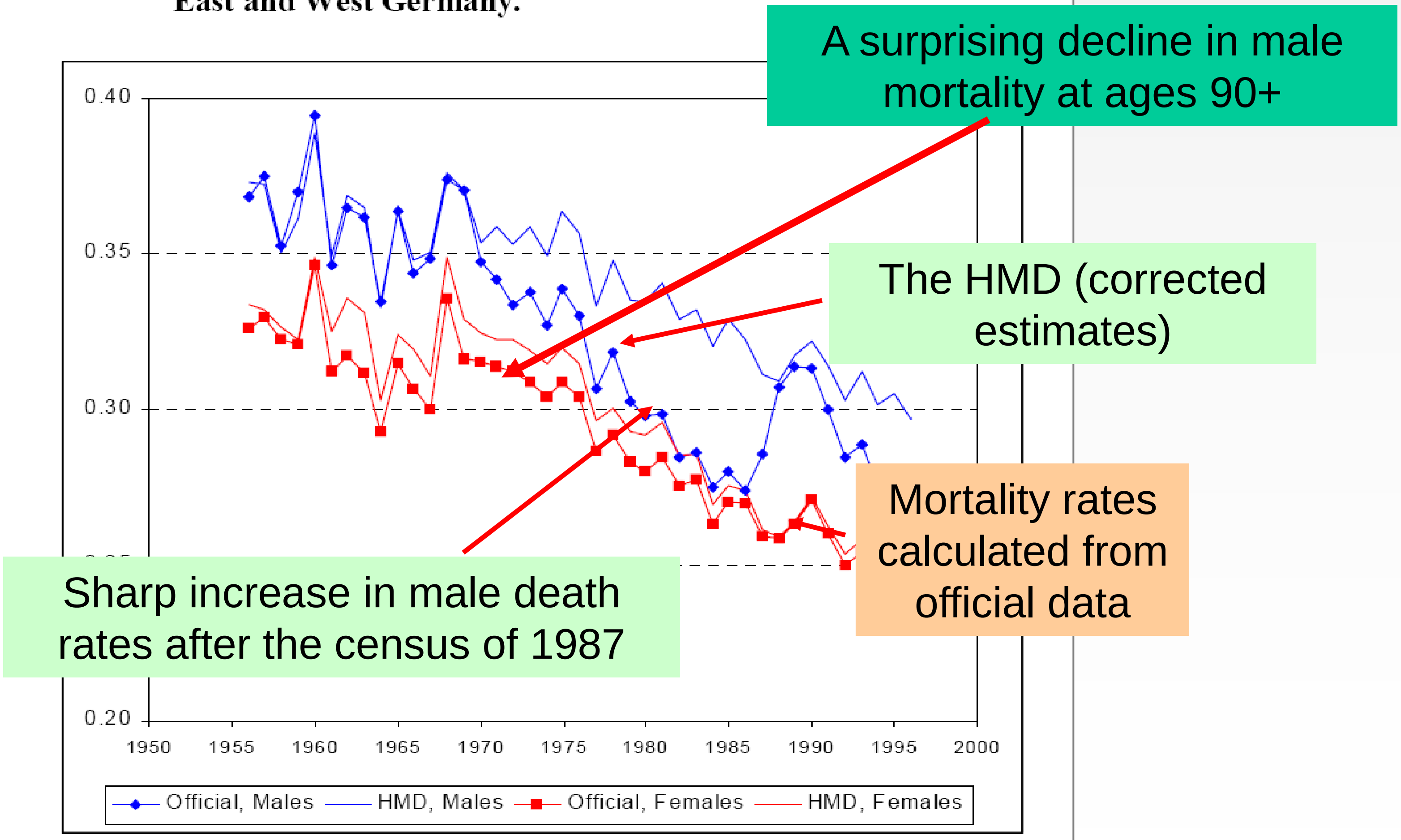
Change in definition of Maori in the census of 1991 from the one based on ethnicity of parents to the one based on self-identification. The new definition caused a jump in Maori population, but the death counts were not corrected simultaneously. Observation of continuous time series and additional information help to identify the problem.

Source: The HMD project, 2012.

Data quality: inter-censal population overstatement due to unregistered out-migration (Germany)

Implausibly low death rates of men aged 90+ (almost equal) to death rates of females and unusual jumps in male death rates help to see the problem. Correction is made on the basis of pension funds' data (DRV)

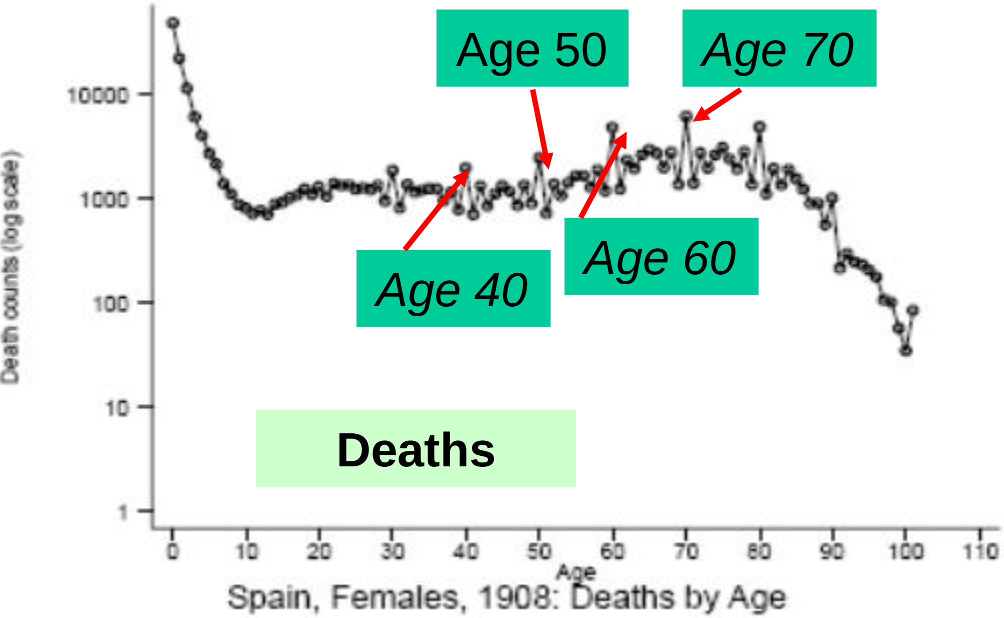
Figure 9: Trends in death rates at age 90+, calculated from the official and the HMD population estimates, for the combined population of East and West Germany.

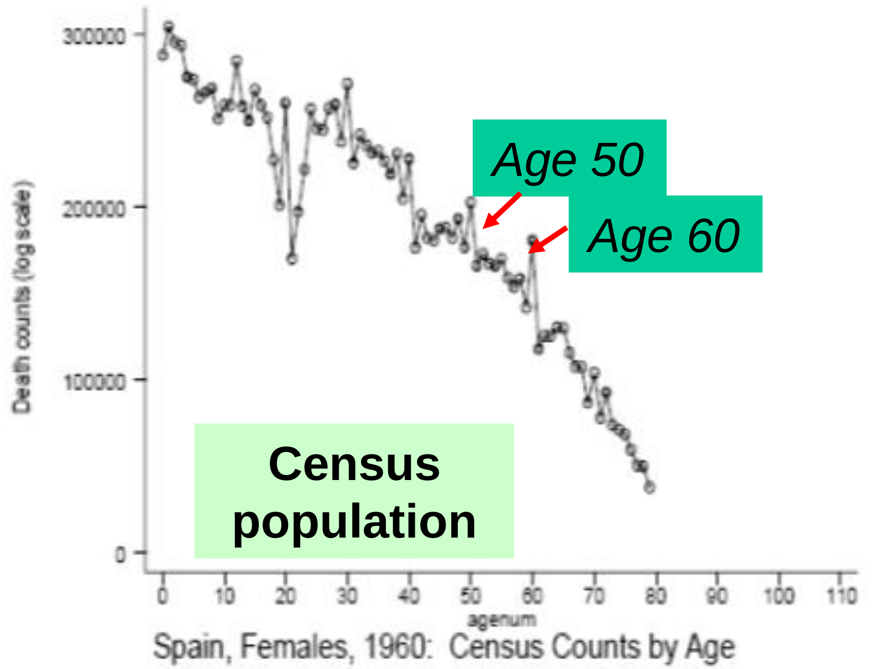


Source: Jdanov, Scholz, Shkolnikov, 2005.

Data quality: age heaping at ages ended by 0 and by 5 due to imprecise registration of age (1)

Digit preference or age heaping



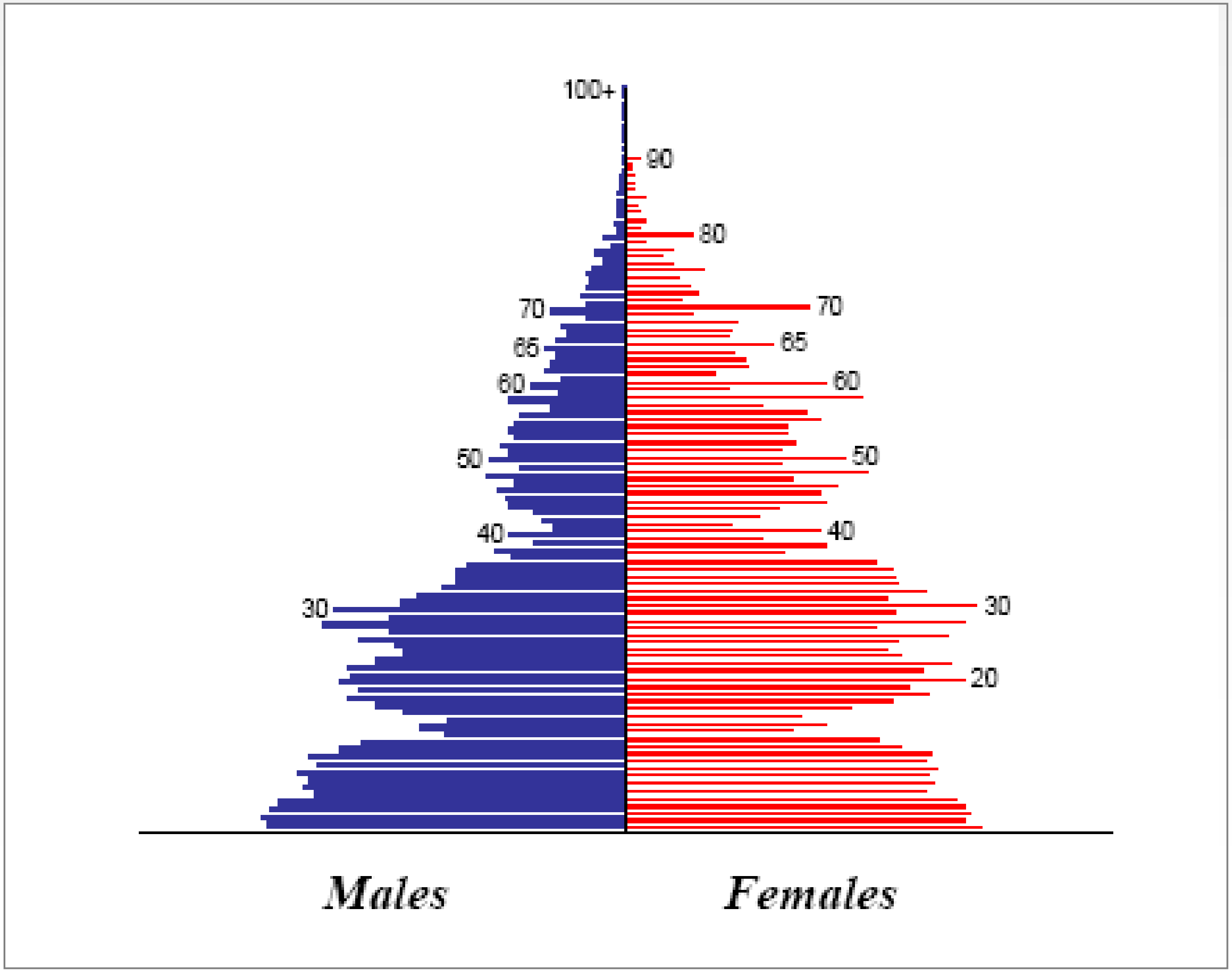


Spanish female deaths by age, 1908
Source: Glei et al., 2007.

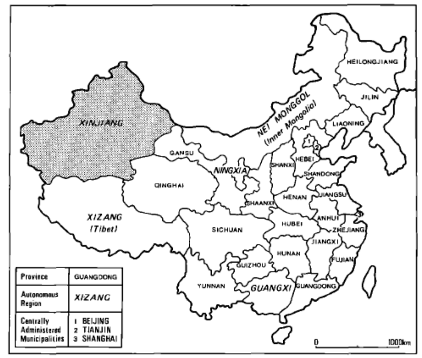
Spanish female census counts by age, 1960
Source: Glei et al., 2007.

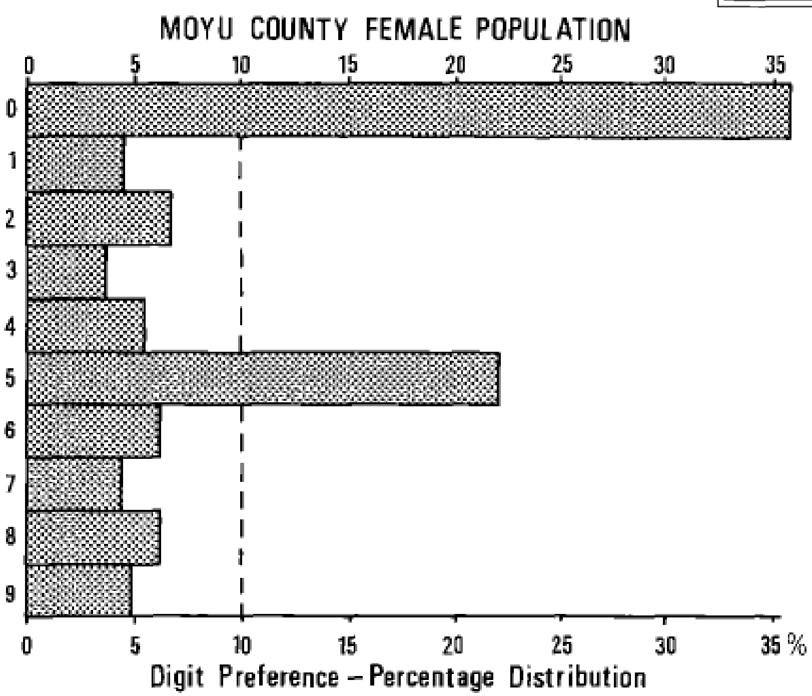
Data quality: age heaping at ages ended by 0 and by 5 due to imprecise registration of age at census (2)

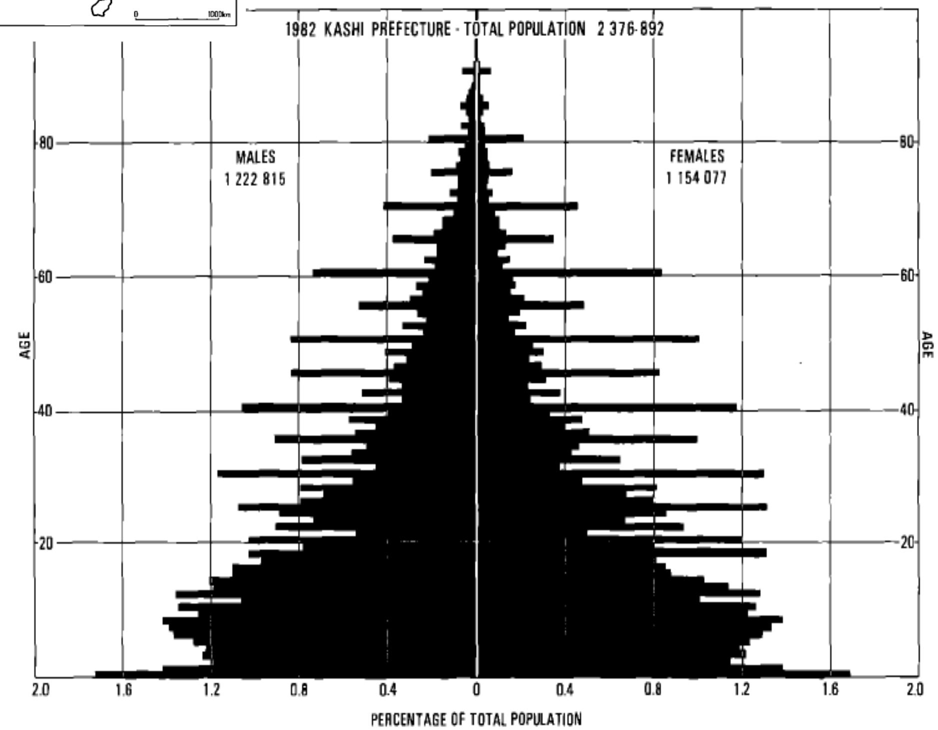
Digit preference or age heaping in census counts
Belarus, Census as of January 15, 1959



Data quality issues: severe age heaping in a Muslim region of China (Uygur population in Xinjiang Autonomous region, census of 1982)



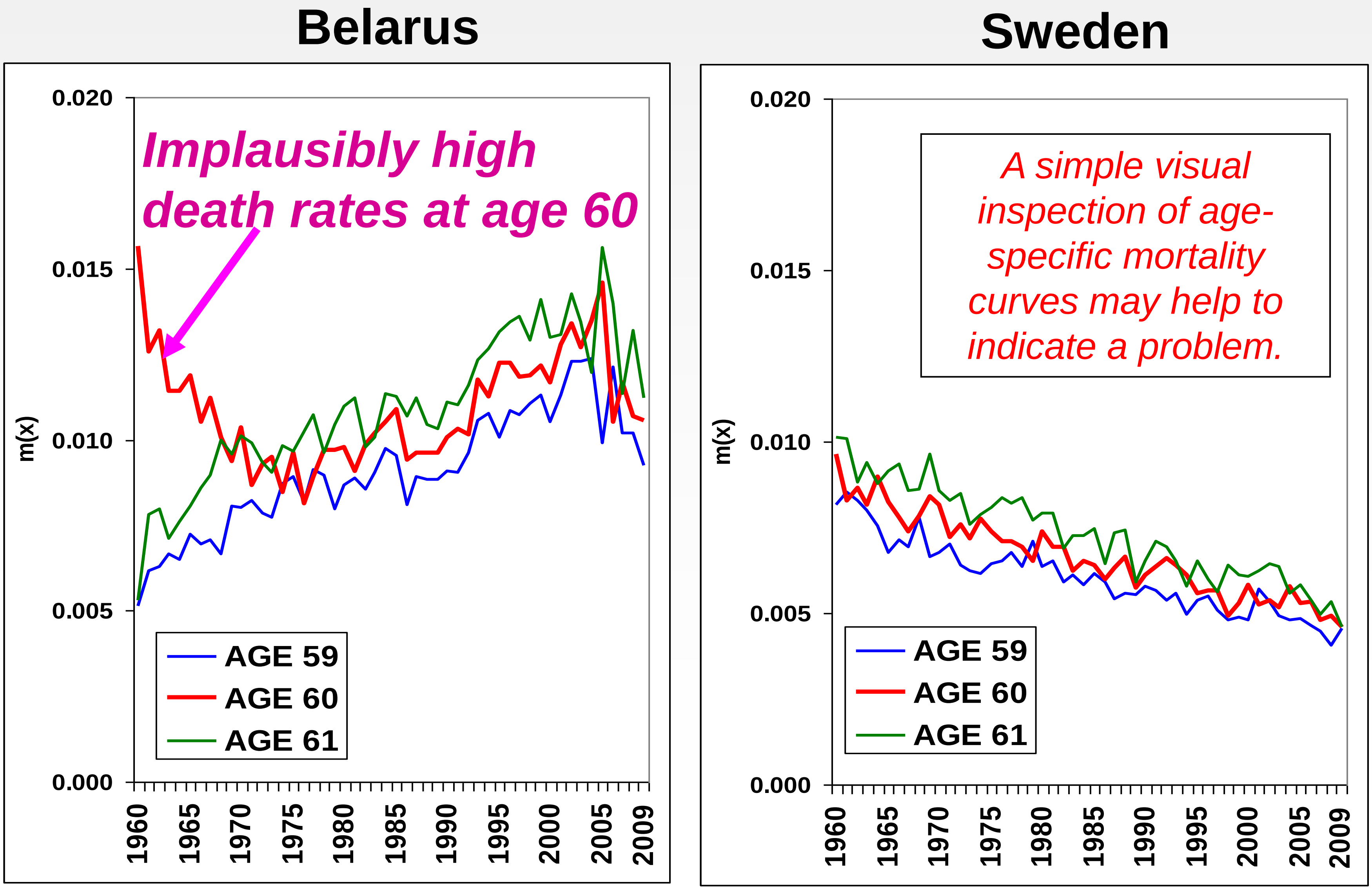




Source: Jowett, Li, 1992

Data quality: age heaping at ages ended by 0 and by 5 due to imprecise registration of age (2)

Consequences of digit preference (age heaping) for age-specific mortality rates



Data quality: measures of age heaping (1)

Whipple's Index	Quality of Data	Deviation from Perfect
<105	very accurate	< 5%
105-110	relatively accurate	5-9.99%
110-125	OK	10-24.99%
125-175	bad	25-74.99%
>175	very bad	>= 75%

"Perfect" = Whipple's index for country with high quality population statistics, e.g. Sweden.

(1) The Whipple's Index =

$$\frac{\text{(sum of numbers at ages 25, 30, 35, ..., 60)} \times 100 \times 5}{\text{total number between ages 23 and 62}}$$

The Whipple's Index for ages over 90=

$$\frac{\text{(sum of numbers at ages 95, 100, 105)} \times 100 \times 5}{\text{total number between ages 93 and 107}}$$

A comparison of the Whipple's Index for centenarians, China 1990 and Sweden 1985-1994

	Male			Female			Both Sexes		
	China	Sweden	Dif. %	China	Sweden	Dif. %	China	Sweden	Dif. %
Survivors Ages 95+	82.7	88	-6	87.2	90.8	-4	86.2	90.1	-4.3
Deaths Ages 95+	91	91.1	-6	88.5	96	-4	89.2	94.7	-5.8

Source: Wang, Zeng, Jeune, Vaupel,1999.

The Whipple's index does not show any age heaping in China (Han population)

Data quality: measures of age heaping (2)

2. Kannisto's Age heaping index (for ages (70, 80 or 90)

$$AHI_i = \frac{D_i}{\exp\left(\frac{1}{5}\sum_{y=i-2}^{i+2}\ln(D_y)\right)}$$

where D_i is number of deaths at age *i.*
Note: age heaping is present if AHI>1.10

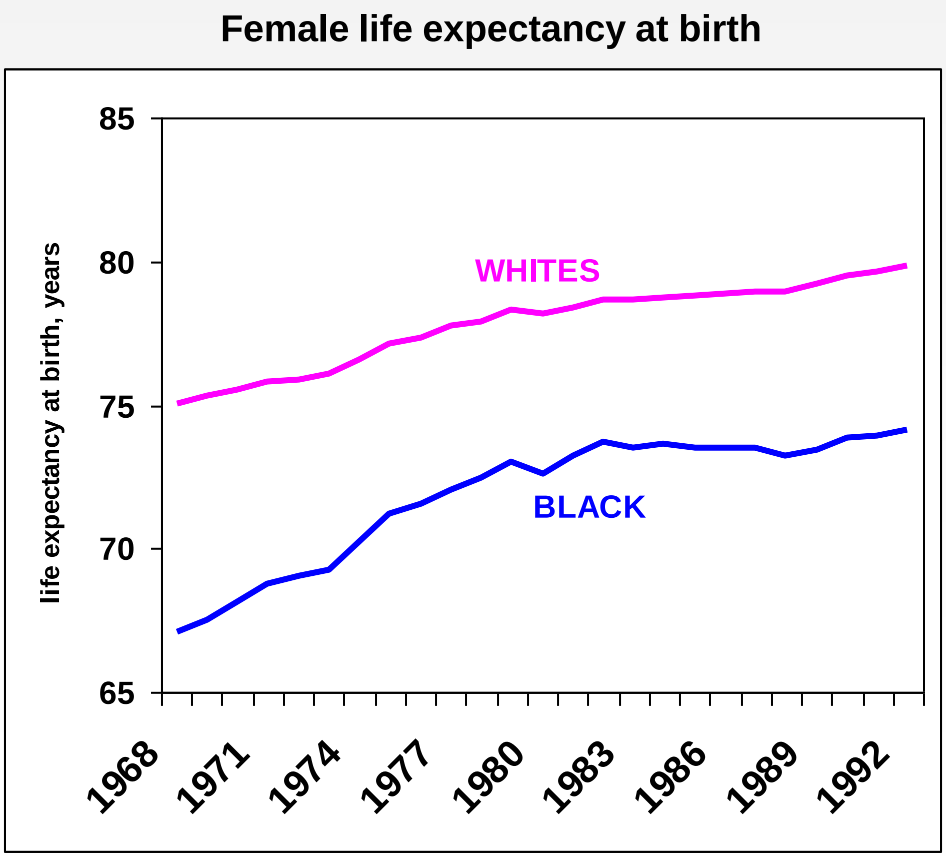
3. Ratios of probabilities of dying

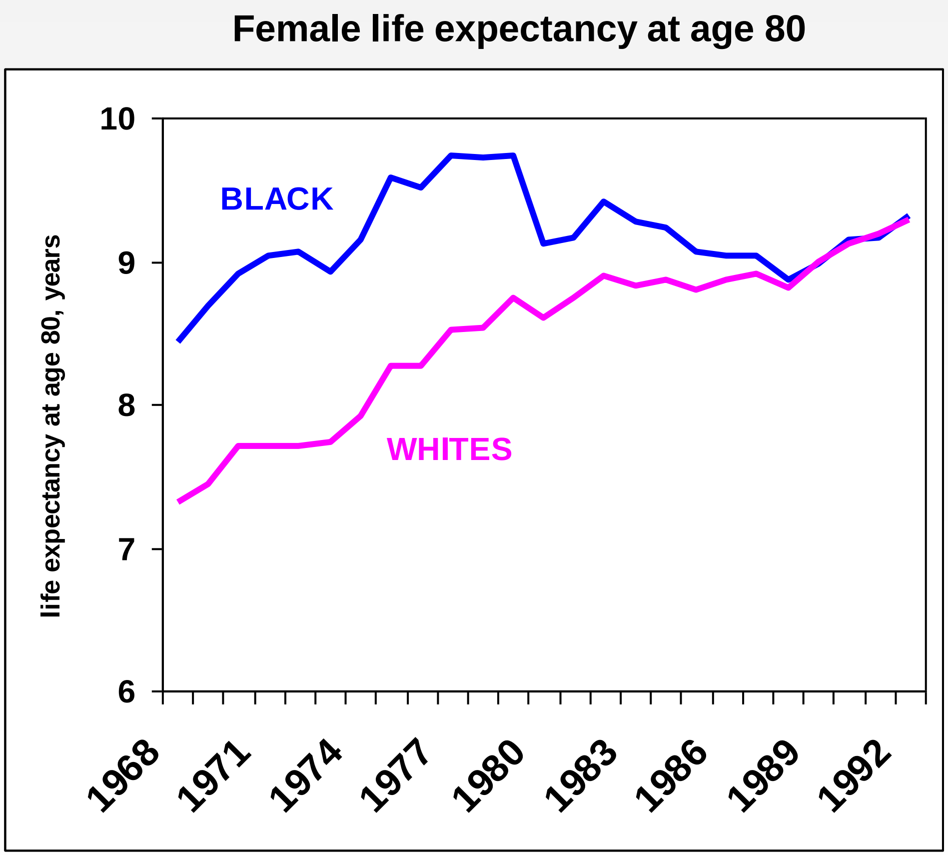
q(80)/q(81) ; q(90)/q91

Sources:
Wang, Z., Zeng, Y., Jeune, B., & Vaupel, J. (1999). Age Validation of Han Chinese Centenarians. In: B. Jeune & J. Vaupel (Eds)*. Monographs on Population Aging.Vol. 6* (pp. 195-214). Odense: Odense University Press.
Kannisto, V. (1999). Assessing the Information on Age at Death of Old Persons in National Vital Statistics. *Validation of Exceptional Longevity. Monographs on Population Aging. Vol. 6.* (pp. 240-249). Odense: Odense University Press.
Jdanov, D., Jasilionis, D., Soroko, E.L., Rau,R., Vaupel, J.W. (2008).
Beyond the Kannisto-Thatcher Database on Old Age Mortality: An Assessment of Data Quality at Advanced Ages. MPIDR Working Paper WP 2008-013.

Data quality: age overstatement (evidence from black-white mortality differences in the US)

Life expectancies at ages 0 and 80 in the US





Source: Berkeley Mortality Database, 2007.

Checking for age overstatement

- **Check the plausibility of e(65) and e(80) estimates**

- **T(100) / T (70) or T(100) / T(80)** *simultaneously checking for mortality cross-overs (e.g. mortality is very high at young and adult ages, but it is unreasonably low at old ages)* (Coale&Kisker, 1986).

Note: T(100) / T(70) is not applicable to low mortality countries showing very rapid progress in decreasing mortality (such as Japan or France).

2. Deaths at age 105+ / Deaths at age 100+

The consequence of age overstatement is unreasonably high life expectancies at old age.

Source: Jdanov, D., Jasilionis, D., Soroko, E.L., Rau,R., Vaupel, J.W. (2008).
Beyond the Kannisto-Thatcher Database on Old Age Mortality: An Assessment of Data Quality at Advanced Ages.
MPIDR Working Paper WP 2008-013.

Assessment of age overstatement: Coale & Kisker (1986)

T(100) / T(70) in various countries relative to T(100) / T(70) for Sweden (1980)

Costa Rica	1950	16.57
Dominican Republic	1950	82.14
El Salvador	1950	51.87
Guatemala	1950	47.06
Haiti	1950	27.61
Panama	1950	40.90
Argentina	1947	9.78
Bolivia	1950	59.59
Brazil	1950	44.13
Venezuela	1950	29.04

Denmark	1950	0.42
Hungary	1960	0.65
Ireland	1970	0.60
Netherlands	1960	0.61
Norway	1960	0.68
Switzerland	1960	0.45

Assessment of the quality of age registration at old age: Evidence from data quality studies

Vaino Kannisto (1994) introduced a set of data quality checks and grouped countries into four data quality groups

Best data quality group:
Belgium, France, the Netherlands, and Sweden. The Czech, Danish, Finnish (from 1951), Italian, Japan (since1971), Scottish, Swiss, Polish, and Western German data are also assigned to this category as they show best data quality throughout the whole period covered with exceptions of one or two periods with acceptable data quality.

Acceptable data quality group:
Australia, Austria, the Czech Republic (from 1981), England & Wales, Estonia, East Germany (except 1961-1970), Ireland (from 1981), Japan (before 1970), Latvia (from 1971), Luxemburg (from 1981), New Zealand (from 1961), New Zealand Non-Maori (from 1961), Norway (from 1961), Portugal (from 1961), Scotland, Slovakia (except 1961-1970), Spain (from 1961), and Slovenia.

Conditionally acceptable group:
Canada (from 1951), Finland (before 1951), Iceland, Latvia (before 1971), Lithuania (from 1981), Luxembourg (before 1981), New Zealand Non Maori (before 1961), Norway (before 1961), and the USA.

Weak data quality group:
Canada (before 1951), Chile, Ireland (before 1981), Lithuania (before 1981), New Zealand (1951-1960), Portugal (before 1961), and Spain (before 1961).

Source: Jdanov, D., Jasilionis, D., Soroko, E.L., Rau,R., Vaupel, J.W. (2008). Beyond the Kannisto-Thatcher Database on Old Age Mortality: An Assessment of Data Quality at Advanced Ages. MPIDR Working Paper WP 2008-013.

Once a poor quality of data is detected what can one do about it?

❖ Be aware of the problem. Address it in discussion. Try to understand causes of the problem.

❖ Avoid using problematic parts of the data. For example, if mortality at infant age is problematic, use e_5 instead of e_0. If mortality at ages 80+ is problematic, use the interval life expectancy $_{80}e_0=(T_0-T_{80})/l_0$.

❖ Correct populations and/or deaths using additional or alternative sources providing reliable data.

❖ If needed, age heaping can be treated by smoothing or by redistribution of excess deaths at 0- and 5-ages.

❖ Use model mortality age curves instead of original age-specific death rates. Gomperz or logistic curves can be used to fit mortality at old ages.

❖ Use mortality from the most relevant (for your case) model life tables instead of defective or missing data.

Once a poor quality of data is identified what can one do about it?

Model life tables represent methods to accumulate the past mortality patterns of countries with good-quality data in a transparent way. They provide some „standard“ mortality age curves corresponding to approximately known overall or age-specific levels of mortality.

Steps for construction of the MLTs:

1. Gather a lot of good quality life tables

2. Find simple relationships for modeling a major part of variation in the data. For example, linear relationship between age-specific qx and e10 in Coale-Demeny MLTs, linear relationship between logit-transformed mortality curves (Brass).

3. Classify the data into series of MLTs that can be used as a standard.

UN MLT (1958, 1982); Coale-Demeny MLT (1966, 1983), Ledermann MLT (1969), Coale-Guo MLTs (1989), INDEPTH MLTs (2004)

Read more in *J.Duchêne, 2006*

Once a poor quality of data is identified what can one do about it?

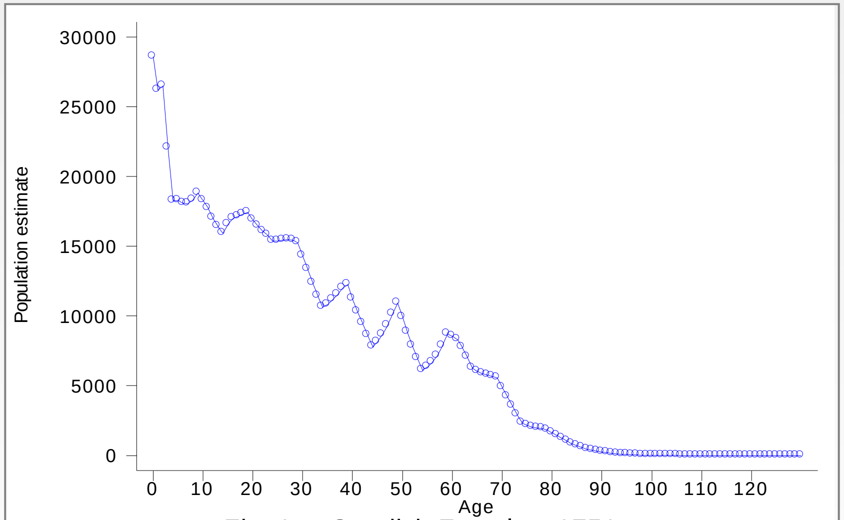


Fig. 1a: Swedish Females, 1751

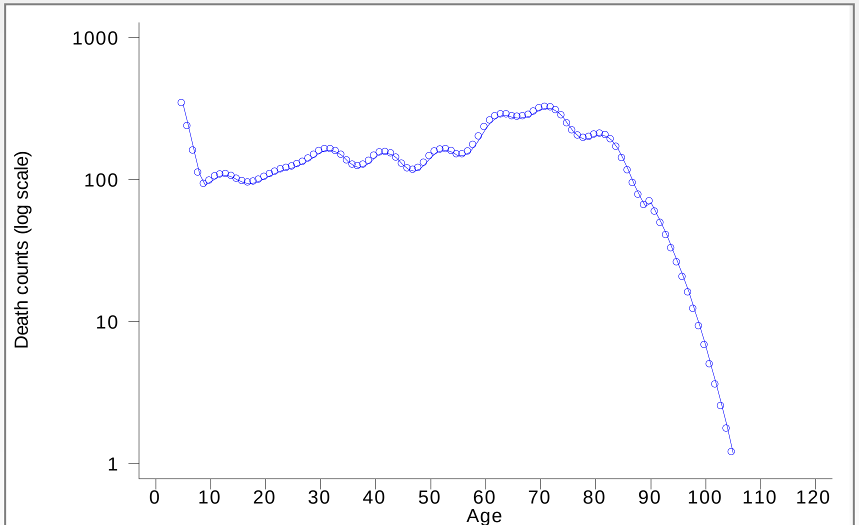


Fig. 1b: Swedish Females, 1751

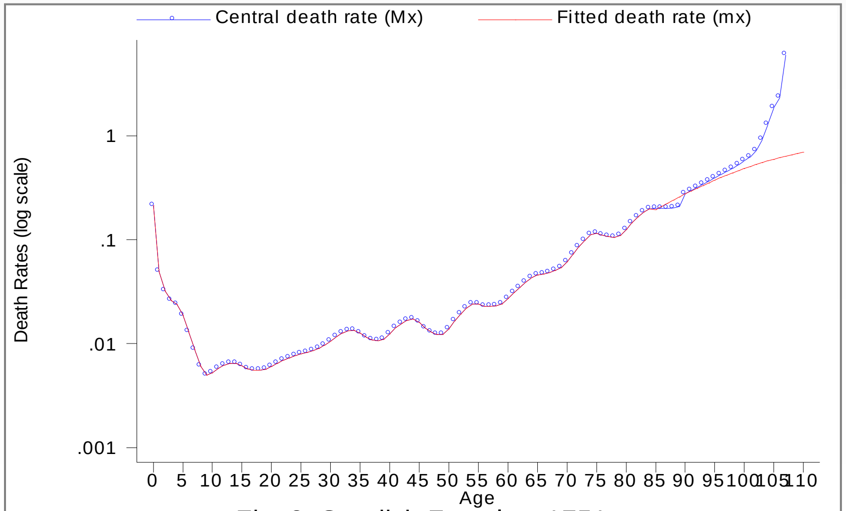


Fig. 2: Swedish Females, 1751

We would advise using the life tables by 5-year (or 10-year) age groups (rather than the 1x1 data)

Source: ABOUT MORTALITY DATA FOR SWEDEN , HMD 2008

Homework task

1. Используя данные по смертности для Ирландии (женщин) рассчитайте Kannisto's age heaping index для возраста 80 лет для всех годов. Данные можно агрегировать по 5-летним периодам. Постройте график и коротко сформулируйте свое заключение.

PART 2. STATISTICAL ERRORS

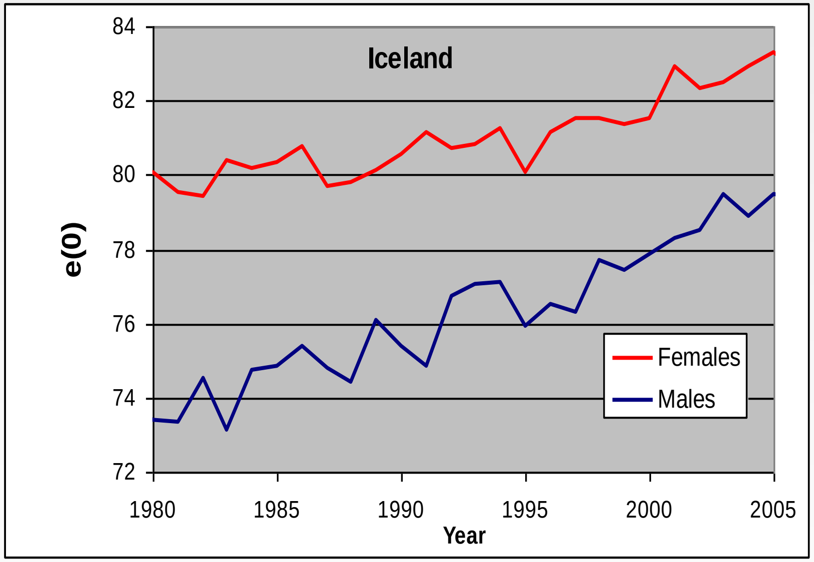
Why to care about randomness and confidence limits?

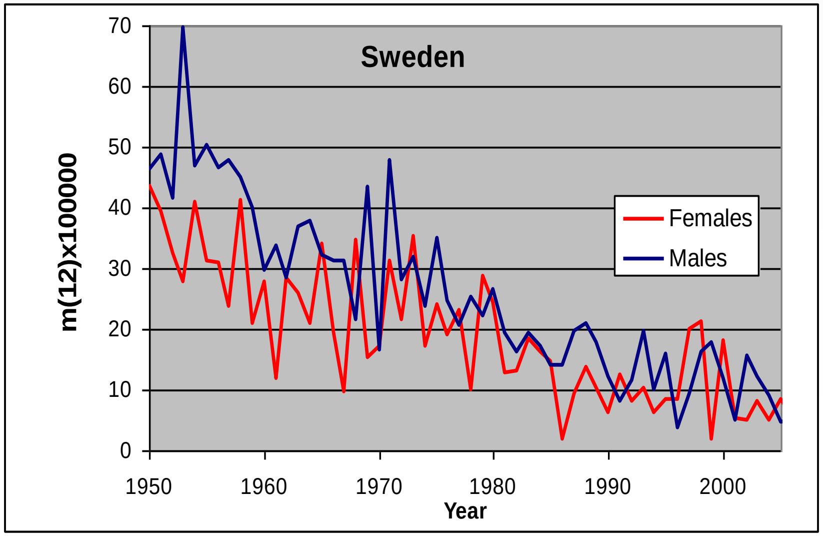
There are two approaches to measurements.

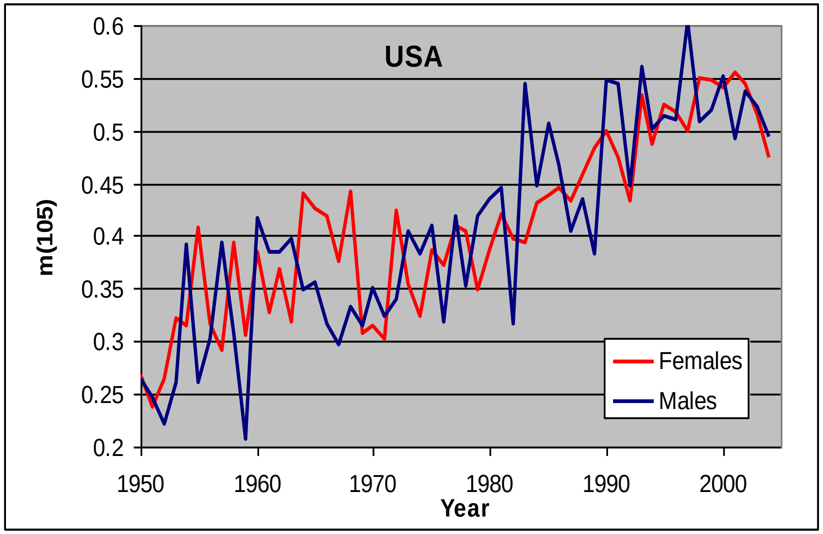
Deterministic approach: „if 3 deaths were registered among 1000 people in 2005 then the death rate is exactly 3 per 1000“.

Stochastic approach: „if 3 deaths were registered among 1000 people in 2005 then the death rate can be estimated with confidence of 0.95 as 3 plus-minus 0.4 per 1000 “.

In most cases demographers operate with large populations and the deterministic approach is just fine. However ...







Standard error of the elementary death rate

To understand well this and following slides, you need to know what is written in sections "Expectation of sample proportion and sample mean", "Variance of sample proportion and sample mean", "The binomial distribution", "The normal distribution", "The central limit theorem", and "Confidence intervals of means and proportions" of the document "Elements of Probability and Statistics".

$$M_x = \frac{D_x}{P_x}$$

The rule of variance yields the estimate of variance:

$$S^2_{M_x} = \frac{1}{P_x^2} S^2_{D_x} \qquad \textbf{(1)}$$

***D_x* is a binomial random variable in *N_x* trials with the probability of of dying *q_x*. The expectation and the variance for this variable are:**

$$E(D_x) = N_x q_x$$

$$\sigma^2_{D_x} = N_x q_x (1 - q_x)$$

The estimate of variance

$$S^2_{D_x} = D_x(1 - \hat{q}_x) = P_x M_x (1 - \hat{q}_x) \qquad \textbf{(2)}$$

Formulae (1) and (2) together yield

$$S^2_{M_x} = \frac{1}{P_x} M_x (1 - \hat{q}_x)$$

The standard error of Mx

$$SE_{M_x} = \sqrt{\frac{1}{P_x} M_x (1 - \hat{q}_x)} \approx \sqrt{\frac{1}{P_x} M_x}$$

Standard error of the elementary death probability

As we know the variance of the estimate of probability q_x is

$$\sigma_{q_x}^2 = \frac{1}{N_x} q_x (1 - q_x) \qquad (1)$$

The unknown number of trials (population size) is estimated as:

$$N_x = \frac{D_x}{\hat{q}_x} \qquad (2)$$

(1) and (2) together yield the estimate of variance

$$S_{q_x}^2 = \frac{1}{D_x} \hat{q}_x^2 (1 - \hat{q}_x)$$

The standard error of q_x is

$$SE_{q_x} = \hat{q}_x \sqrt{\frac{1}{D_x}(1 - \hat{q}_x)} \approx \hat{q}_x \sqrt{\frac{1}{D_x}}$$

Standard error of linear aggregates: SDRs, SMRs etc.

For any linear combination of elementary death rates R, the rule of variance yields:

$$R = \sum_x w_x M_x \qquad S_R^2 = \sum_x (w_x)^2 S_{M_x}^2$$

For the crude death rate:

$$CDR = \frac{D}{P} = \frac{1}{P}\sum_x P_x M_x \qquad w_x = \frac{P_x}{P}$$

For the directly standardized death rate:

$$SDR = \sum_x \theta_x^s M_x \qquad w_x = \theta_x^s$$

For the (indirectly) standardized mortality ratio:

$$SMR = \frac{D}{D_{\exp}} = \frac{D}{\sum_x P_x M_x^s} = \frac{\sum_x P_x M_x}{\sum_x P_x M_x^s} = \sum_x \left(\frac{P_x}{\sum_x P_x M_x^s} \right) M_x \qquad w_x = \frac{P_x}{\sum_x P_x M_x^s}$$

Standard error of the life and the healthy life expectancies

For development of the formulae see: Chiang (1984) and the EHEMU Technical Report 2006_3.

Conventional LE:

$$S_{e_x}^2 = \frac{1}{l_x^2} \sum_{y=x}^{\omega-1} l_y^2 [(1-a_y)n + e_{y+n}]^2 S_{p_y}^2$$

Healthy LE:

$$S_{eH_x}^2 = \frac{1}{l_x^2} \sum_{y=x}^{\omega-1} l_y^2 [(1-a_y)n(1-\pi_y) + eH_{y+n}]^2 S_{p_y}^2 + \frac{1}{l_x^2} \sum_{y=x}^{\omega-1} L_y^2 S_{(1-\pi_y)}^2$$

CI-LE_plus_simulations

CI-HLE_plus_simulations

Homework task

2. Посчитайте доверительный интервал для e0 и e80 для Исландии и России для последнего доступного в HMD года.