Methods of Standardisation

13

Abby Israëls

Statistics Methods (201302)

Statistics Netherlands

The Hague/Heerlen, 2013

Explanation of symbols

.	data not available
*	provisional figure
**	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2012–2013	2012 to 2013 inclusive
2012/2013	average for 2012 up to and including 2013
2012/'13	crop year, financial year, school year etc. beginning in 2012 and ending in 2013
2010/'11– 2012/'13	crop year, financial year, etc. 2010/'11 to 2012/'13 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands
Grafimedia

Cover
Teldesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form:
www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1876-0333

60165201302 X-37

Table of Contents

1. Introduction to the theme

1.1 General description and reading guide

Demographers and health statisticians frequently encounter a problem that involves comparing the results of populations that have different structures with respect to background characteristics. But, of course, it also occurs in many other disciplines. An example of this is comparing mortality figures from cardiovascular diseases for populations with a different age distribution. Given a similar health care system, countries with a young population will usually have lower mortality rates than countries with a much older population. In this case, a country's gross (crude) mortality rate is therefore not a good indicator of the health of its citizens. Only when the data are examined for age effects, by only comparing individuals in the same age class, is it possible to make a fair comparison. To this end, we can determine age-specific mortality rates for each population. We can also determine averages that are adjusted (corrected) for age: *standardised mortality rates*[1] or, more generally, *standardised averages*. Standardisation methods have been developed for this, for which we distinguish:

1. a target variable (*Y*);
2. populations, i.e.
   a) the populations to be compared, and
   b) a 'standard population' (reference population);
3. variables by which we standardise, the so-called 'distorting characteristics' (or 'confounding characteristics');
4. a target function (average, $\bar{Y}$) or target parameter (expectation, $\mu = E(\bar{Y})$).

In the example mentioned, *Y* is dying/not dying from a cardiovascular disease in a certain period and population, and μ is the underlying mortality probability. Further, the countries distinguished between are the populations, and one specific country, for example, can be chosen as the standard. Finally, we perform standardisation by the variable of Age (in classes).

Target variable *Y* can be a binary (0/1) variable, such as death/no death, but it can also be a quantitative variable, such as annual wages or the number of pregnancies. The type of variable has an impact on the determination of confidence margins and on the modelling. It is also possible to standardise a complete frequency distribution or all the scores from a frequency distribution.

[1] Statistics Netherlands defines mortality rate as the number of deaths in a certain period per 1000 or 10,000 residents. In this report, we define the mortality rate as the average number of death per resident.

The standard population can be one of the populations to be distinguished. Oftentimes, however, the union of the populations studied is taken as the standard ('sum population'). A hypothetical reference group can also be used as the standard. If we want to monitor the differences between populations over time, we can introduce a standard year. For example, people have constructed a hypothetical (i.e. simplified) population for Europe in 1950 and Europe in 2000.

We standardise by 'distorting variables', which would otherwise prevent a fair comparison of the target function for different populations. Mortality rates and morbidity rates (prevalence, or the average number of care contacts) are virtually always standardised at Statistics Netherlands by age and gender, or by age per gender. However, to compare absentee percentages between different groups of employees, the scale level of the employees may also be understood as a distortion, for which adjustment is required. Distorting variables are variables for which the effect on the target variable is well-known. Our goal is to 'calculate away' the effects of these distorting variables to make the remaining effects or changes visible.

Because describing the standardisation in general terms involves tedious formulations, this will be done as far as possible using the example of mortality rates that must be standardised by age. The use of the formulas will provide the general explanation needed.

Standardisation has been used for actuarial calculations since the mid-18th century (see Keiding, 1987), a time when neither the pocket calculator nor mechanical calculation tools were available. Other applications of standardisation are mortality figures by cause of death (as stated previously), the number of hospital admissions, fertility rates, disposable income for different target groups (e.g. adjusted for differences in the size and composition of the household), etc.

Traditionally, there are two methods of standardisation: direct and indirect standardisation. In *direct standardisation* (chapter 2), for each population, the distribution of the distorting characteristics in the standard population is used. In *indirect standardisation*, for each population, the mortality rate is compared with the mortality rate that would be obtained if the age-specific mortality rates were equal to those of the standard population (chapter 3). Linear regression can also be used to adjust mortality rates for distorting characteristics. For mortality rates, the obvious choice is therefore to use logistic regression, because Mortality is a binary variable. This and other forms of *regression analysis* are discussed in chapter 4. We also discuss the link between these forms and direct and indirect standardisation. For the sake of simplicity and due to the fact that less data are needed for standardisation, mainly direct and indirect standardisation were used originally. Finally, chapter 5 takes a look at some relationships between the different methods. Naturally, the report also describes advantages and disadvantages of the methods in various situations.

1.2 Scope and relationship with other themes

As stated above, standardisation methods are frequently used to compare mortality rates of different populations, where adjustments are made for differences in the age structure. Life tables (see the Methods Series document 'Life Tables' by Van der Meulen, 2009) can be used as the basic material for direct and indirect standardisation.

Standardisation methods have a strong similarity to composite index numbers; see the theme "Index numbers" from the Methods Series (Van der Grient and De Haan, 2011). For both topics, this concerns the presentation of summary measures, where weighting is performed on the categories of the 'distorting' characteristics. We discuss this similarity in the subsections 2.5.1 and 3.5.1.

1.3 Place in the statistical process

Standardisation can be viewed as a further analysis of the data. At Statistics Netherlands, however, many health statistics and some population statistics are published in both a standardised and unstandardised form, because the presentation of only the unstandardised figures can easily lead to incorrect interpretations. Calculating standardised figures is therefore often a standard component of the output.

1.4 Definitions

Concept	Description
Mortality rate (Gross mortality rate)	Number of deaths in a certain period per number x of the population. Often, x is given the value of 1, 1000 or 10,000. In this report, we use x=1.
Mortality ratio	Quotient of the gross mortality rate of the population studied and the standard population.
Mortality figure	Any figure that involves mortality. It may be used for the absolute number of died persons.
Standardisation	Adjusting aggregate figures for the influence of distorting (confounding) characteristics.
Standardised average	Average after correction (adjustment) for the effect of distorting characteristics
Standardised mortality rate	Adjusted gross mortality rate, by correcting for the effect of distorting characteristics (example of a standardised average).
Direct standardisation	Standardisation method in which mortality figures (especially rates) of one or several populations are weighted by a characteristic of one particular 'standard population'.
Indirect standardisation	Standardisation method in which an observed mortality figure (especially a rate) is compared with the corresponding figure (rate) that is obtained by adopting the age-specific mortality rates of an external population.
CMF	Comparative Mortality Figure . It is a direct standardised mortality ratio; see formula (2.2).
SMR	Standard Mortality Ratio; It is based on indirect standardisation; see formula (3.1).

1.5 General notation

We use the subscript i for the classes of the distorting characteristic (i=1,…,I), j for the populations considered (j=1,…,J) and s for the standard or reference population. Furthermore, in a certain period,

N_{ij} = number of people in age class i, population j,

D_{ij} = mortality (number of deaths) in age class i, population j.

The above data are the basic data from which the following can be derived:

$N_{+j} = \sum_i N_{ij}$ = size of population j,

$q_{i|j} = N_{ij}/N_{+j}$ = age distribution for population j,

$D_{+j} = \sum_i D_{ij}$ = mortality (number of deaths) in population j,

$Y_{ij} = D_{ij}/N_{ij} = \overline{D}_{ij}$ = *age-specific mortality rate* in (i,j) = average mortality per resident in (i,j),

$Y_{\cdot j} = D_{+j}/N_{+j} = \overline{D}_{\cdot j}$ = *(gross) mortality rate* = average mortality per resident in j.

A mortality rate Y_{ij}, can be seen as a realization of a *mortality probability* μ_{ij}, such that $\hat{\mu}_{ij} = Y_{ij}$.

The gross mortality rate is a weighted sum of the age-specific mortality rates:

$$Y_{\cdot j} = \sum_i q_{i|j} Y_{ij} \quad . \tag{1.1}$$

For the standard population, the subscript 'j' in these formulas is replaced by 's'. Often, the union of all populations j (=1,..,J) is used for the standard population. In that case,

$$D_{is} = \sum_j D_{ij}, \qquad D_{+s} = \sum_j D_{+j}, \qquad N_{is} = \sum_j N_{ij}, \qquad N_{+s} = \sum_j N_{+j} \quad . \tag{1.2}$$

If the mortality rates relate to a certain period, we can use the average over a number of dates from that period as the population size N_{ij} (for example, the average of the population at the start and at the end of the period), or the population size on the median date. In practical terms, it is rarely necessary to work more precisely, but we can also define N_{ij} as the number of person years in a certain period, i.e. the sum of the individual risk periods for all people, expressed in years.

D_{ij} is an aggregate statistic. We can define the underlying variable D (Mortality) with person as the object type. D is then a binary variable with values 1 (deceased) and 0 (not deceased), and D_{ijk} is the score on variable D of individual k from age class i and population j (k=1,…,N_{ij}).

More generally, we can assume a quantitative variable Y with individual scores Y_{ijk} where k = 1,…,N_{ij}. Y_{ij} is thus defined as the *average* of these individual scores. We can also see Y_{ij} as an aggregate of individual scores Y_{ijk}. If Y is binary, then Y_{ijk} = D_{ijk}, because $N_{ijk} \equiv 1$. Standardisation methods are derived in terms of Y and N, but

because they are often applied to binary variables, we will represent most of the formulas in terms of *D* and *N*.

Finally, we define *ratios*. The *age-specific mortality ratio* for population *j* compared to the standard population *s* is

$$R_i = \frac{D_{ij} / N_{ij}}{D_{is} / N_{is}} = \frac{Y_{ij}}{Y_{is}} \quad (i=1,\ldots,I) \; . \tag{1.3}$$

The *gross mortality ratio* for population *j* compared to the standard population *s* is

$$R = \frac{D_{+j} / N_{+j}}{D_{+s} / N_{+s}} = \frac{Y_{.j}}{Y_{.s}} \; . \tag{1.4}$$

For the sake of convenience, we have omitted the subscripts *j* and *s* from these rates.

As a result of different age distributions, the gross mortality ratio *R* does not necessarily fall between the maximum and minimum R_i. Standardised mortality ratios (CMF and SMR), however, do satisfy this requirement.

To help the reader become more familiar with the notation, we have provided a specific example in table A and B of appendix 2; see columns (1)-(10). A number of standardisation methods will be applied to this table in the following chapters. Section 2.4 sets out the original goal of the analysis. The table covers the period from 1979 to 1986. The numbers of deaths (D_{ij} and D_{is}) therefore relate to an eight-year period. The population sizes (N_{ij} , etc.) in this table are sums of the year totals for these eight years, as the approximation of the number of person years; we could also have used the averages over the years instead. The year total in year *t* is calculated as the average of the population size on 1 January of year *t* and 1 January of year *t+1*.

2. Direct standardisation

2.1 Short description

In direct standardisation, for each population *j*, the age-specific mortality rates Y_{ij} are weighted using a standard age distribution (age distribution in the standard population) instead of using the individual age distribution as in the gross mortality rate in formula (1.1). This results in the *direct standardised mortality rate* for population *j*:

$$Y_j^{DIR} = \sum_i q_{i|s} Y_{ij} \quad . \tag{2.1}$$

Dividing by the gross mortality rate in the standard population results in the *Comparative Mortality Figure*:

$$CMF = \frac{\sum_i q_{i|s} Y_{ij}}{\sum_i q_{i|s} Y_{is}} = \frac{\sum_i q_{i|s} Y_{ij}}{Y_{\cdot s}} \quad . \tag{2.2}$$

So CMF is a measure for the *ratio* of the mortality in populations *j* and *s*, adjusted for age. The calculations of (2.1) and (2.2) are both called *direct standardisation*.

2.2 Applicability

1. The CMF makes it possible to compare the mortality in a population *j* with the mortality in the standard population. Because a fixed standard is used, the CMFs also enable the comparison of the mortality rates in multiple populations *j*, as the denominators are the same. In indirect standardisation, comparison of mortality rates from different populations is problematic (chapter 3). For this reason, direct standardisation is generally preferred if we want to compare the mortality, for example, in multiple countries or regions, for multiple years or for various ethnic groups.
2. Formula (2.4) in section 2.3.1 will demonstrate that the CMF can be written as a weighted average of the age-specific mortality ratios R_i from (1.3), where weighting is performed using the fraction of deaths in the standard population. A discussion has arisen about the application of direct standardisation if the R_i differ strongly from one another. Some authors think that, in this case, only the age-specific mortality rates and/or ratios should be published, and that the CMF is only a useful summary measure of the R_i if these are reasonably homogeneous. On the other hand, we do publish simple average scores (for example, a gross mortality rate), without requiring everyone (or every age class) to score that average. For other authors this is a justification for calculating a CMF, also if the R_i differ strongly from each other. However, in this case, one must be aware of the effect that the weights of R_i have on the outcome, and it is

therefore a good idea to also present the R_i. More information about the advantages and disadvantages of determining standardised figures can be found in Fleiss (1973, chapter 13).

3. Selecting a fixed (non-stochastic) or very large standard population increases the accuracy of the CMF and simplifies the calculation of standard errors (section 2.3.2). Often, the selected standard is the union of all populations j=1,…,J, i.e. all the countries, regions or periods to be observed. Sometimes, international agreement must be obtained about the choice of the appropriate standard population.

   At Statistics Netherlands, direct standardisation is also used for long time series of mortality and morbidity rates, for which not just different years, but also different populations (for example, ethnic groups) are compared to each other, by standardising by age (separately per gender, or together); j is then a combination of population and year. In this case, there are more choices for the standard population. Sometimes, standardisation is performed by selecting the sum population (for example, all ethnic groups) in a certain base year as the standard. But we can also choose to only standardise within the year (with, each year, the sum population of the ethnic groups of that year as the standard), as is done in the standardisation of general practitioner contacts, where there is not yet a long time series.

4. Applying direct standardisation (or, in general, applying 'propensity score weighting methods') is not recommended if, for one or more age classes, $q_{i|j}$ is very small and the associated $q_{i|s}$ is much larger. $Y_{ij} = D_{ij}/N_{ij}$ is then based on few observations and still counts heavily in (2.1). The variances of Y_j^{DIR} and CMF, which are presented in section (2.3.2), are consequently very large. In table B of appendix 2, we would already have difficulty with this if we were to include the age class 65+, but certainly if we were to split this class. Here, the ratio $q_{i|s}/q_{i|j}$ is equal to 9.95/0.16 = 63, which means that the contribution for this age class to the standard error in direct standardisation is 63 times as large as for the gross mortality rate $Y_{.s}$. This is only minimally compensated for by the other age classes. Obviously, besides the q ratio, the number of observations, N_{ij}, also has an effect on the standard error.

   The problem can also arise for countries with fewer older people in an international comparison of mortality rates. Due to the risk of large variances in the case of direct standardisation, we should avoid to split the higher age categories if this causes a strong increase in the q ratio while the number of observations (N_{ij}) is small. For the same reason, we must restrict the number of distorting characteristics used for standardisation purposes. Notice that all interactions between these characteristics are included; see section 4.5.1 for more information.

   A practical example at Statistics Netherlands is the Dutch National Medical Registration (Landelijke Medische Registratie). In this register, the number of hospital admissions by patient's country of origin is directly standardised by the

age distribution of the total Dutch population (per gender and in total). To obtain reliable standardised figures, the population was initially limited to people aged 0 to 50 years, and later included people up to 60 years of age (when the ethnic minority population in older age groups had grown).

It also occurs that Y_{ij} is unknown as mortality rates are not available in each age class for some of the populations. In that case, direct standardisation is impossible, and indirect standardisation is often performed instead.

5. Standardisation is not only applicable to dummy variables such as death/no death, but also to quantitative *Y*-variables. For example, in Israëls and De Ree (1981), standardisation was applied for a comparison of wages between different economic business sectors, for which standardisation is performed by employee age and education.

2.3 Detailed description

2.3.1 Determining the CMF

By multiplying the numerator and denominator of formula (2.2) by N_{+s} , we can write the CMF as

$$CMF = \frac{\sum_i N_{is} Y_{ij}}{\sum_i N_{is} Y_{is}} = \frac{\sum_i (N_{is} / N_{ij}) D_{ij}}{D_{+s}} \ . \qquad (2.3)$$

The denominator is now the number of deaths in the standard population, and the numerator is the *direct standardised number of deaths* in population *j*, i.e. the number of people that would have died if population *j* had the age distribution of the standard population.

The CMF can also be written as a weighted sum of the age-specific mortality ratios R_i with weights $w_{is} = D_{is} / D_{+s}$:

$$CMF = \frac{\sum_i N_{is} Y_{ij}}{\sum_i N_{is} Y_{is}} = \frac{\sum_i N_{is} Y_{is} R_i}{\sum_i N_{is} Y_{is}} = \frac{\sum_i D_{is} R_i}{\sum_i D_{is}} = \sum_i w_{is} R_i \ . \qquad (2.4)$$

We can therefore see the CMF as a summary measure for the age-specific mortality ratios R_i, with weights w_{is} proportional to D_{is}. In section 2.2, item 2, we already questioned the presentation of the CMF when the R_i are too heterogeneous. In section 3.5, we show the similarity of formula (2.4) to the Laspeyres price index.

2.3.2 Standard error of Y_j^{DIR} and CMF

When determining the standard error of Y_j^{DIR} or CMF, N_{ij} and N_{is} are usually known population sizes and therefore have zero variance. Also when these are estimated population sizes or sample sizes, it is justifiable to work conditionally on these

numbers, as we are comparing mortality probabilities. The numbers of deaths D_{ij} and D_{+s} (i=1,…,I) are also population figures. However, the D_{ij} are generally treated as stochastic. In this situation, dying or not dying is seen as the result of a probability mechanism that could also have had a different outcome. For example, it is assumed that the number of deaths $\underline{D}_{ij}$ is binomially distributed with parameters N_{ij} and mortality probability p_{ij}, which means that the variance of $\underline{D}_{ij}$ is equal to[2]

$$Var\ (\underline{D}_{ij}) = N_{ij}\, p_{ij}\,(1 - p_{ij}) \tag{2.5}$$

and the estimated variance is equal to

$$\text{var}(\ \underline{D}_{ij}) = N_{ij}\ \frac{D_{ij}}{N_{ij}}(1 - \frac{D_{ij}}{N_{ij}}) = D_{ij}\,(1 - \frac{D_{ij}}{N_{ij}})\ . \tag{2.6}$$

If the mortality probability is small, i.e. if $D_{ij} << N_{ij}$ (or $Y_{ij} << 1$), then $\underline{D}_{ij}$ is Poisson-distributed by approximation, as a result of which $Var\ (\underline{D}_{ij}) = N_{ij}\, p_{ij}$ and $\text{var}(\ \underline{D}_{ij}) = D_{ij}$.

Based on the binomial distribution of $\underline{D}_{ij}$, the estimated variance of the direct standardised mortality rate is

$$\text{var}(\ Y_j^{dir}) = \text{var}(\sum_i q_{i|s}\ \frac{\underline{D}_{ij}}{N_{ij}}) = \sum_i q_{i|s}^2\ \frac{1}{N_{ij}^2}\ \text{var}(\ \underline{D}_{ij}) = \sum_i q_{i|s}^2\ \frac{1}{N_{ij}^2}\ D_{ij}\,(1 - \frac{D_{ij}}{N_{ij}})\ . \tag{2.7}$$

Here, it is assumed that the estimated mortality rates for different age classes are independent. Traffic accidents or epidemics disrupt this assumption, but this disruption will usually be relatively small. For the variance of the standardised *number* of deaths, we must multiply the variance from (2.7) by N_{+s}^2. The 95% confidence margin of the direct standardised mortality rate is 1.96 times the square root of (2.7), assuming the normal distribution.

Strictly speaking, for the variance of the CMF, we also deal with the stochastic of D_{+s}; see formula (2.3). This stochastic is neglected in the literature, because the standard population (usually the sum population) is almost always extremely large. Chiang (1984) even states that only D_{ij} should be considered as stochastic. The variance estimator of the CMF according to formula (2.2) is therefore

$$\text{var}(\ CMF\ ) = \frac{1}{Y_{\cdot s}^2}\ \text{var}(\ Y_j^{dir}) = \frac{1}{D_{+s}^2}\sum_i \frac{N_{is}^2}{N_{ij}^2} D_{ij}\,(1 - \frac{D_{ij}}{N_{ij}})\ . \tag{2.8}$$

Chiang (1961, 1984) bases the variance calculations on a slightly different situation, namely that of life tables (Van der Meulen, 2009), which does not involve annual mortality, but death in a certain age class.

---

[2] We underline the stochastic parameter $\underline{D}_{ij}$ in the variance formulas, to distinguish this from the realisations D_{ij}.

If Y is a quantitative variable, the variance formulas must be adapted. In this situation, either a theoretical distribution is assumed for $\underline{Y}_{ij}$, or its variance estimation is based on the observed distribution.

To determine the 95% confidence interval of the CMF, we can base ourselves on the normality of the CMF and use $1.96\,SE\,(CMF\,) \equiv 1.96\sqrt{\text{var}(\,CMF\,)}$ as the margin. Because rates are asymmetrical, Breslow and Day (1987) recommend a log transformation. This gives $1.96\,SE\,\{\ln(\,CMF\,)\} = 1.96\,\{SE\,(CMF\,)\}\,/\,CMF$ as 95% margin for the natural logarithm of CMF, after which the interval can be back transformed using the exponential transformation. The same transformation can be used to test 'CMF = 1'.

The test of whether the CMFs of two different populations *j* and *j'* compared to the same standard population are equal can be easily derived from this (Breslow and Day, 1987), as this test boils down to the fact that the quotient of the two direct standardised mortality rates (or of the two CMFs) is equal to 1.

If population *j* is a part of the standard population, as is the case if the standard population is the union of all considered populations *j*, then we can test slightly more accurately by comparing the mortality in population *j* with that in the union of the other populations, *s\j*; see Yule (1934).

2.4 Example

Example 1. *Mortality figures of Turkish and Dutch men, 0-44 years of age: direct standardisation*

In Hoogenboezem and Israëls (1990), analyses were performed of the differences in mortality rates between Turkish, Moroccan and Dutch residents of the Netherlands by various causes of death in the years 1979-1988. The reason behind this was the fact that questions had been asked in the Lower House of the Dutch Parliament about high death rates among Turkish and Moroccan children in the Netherlands, compared to Dutch children of the same age. In Hoogenboezem and Israëls (1990), indirect standardisation was used. In this example, for comparative purposes, we present the results of direct standardisation on the data of table A in appendix 2, while we will discuss the results of indirect standardisation in section 3.4. Please note that the data from table A deviates slightly from the data in Hoogenboezem and Israëls (1990); we limit ourselves here to the years 1979-1986 and to 'men < 45 years'.

The direct standardised mortality rate according to formula (2.1) is equal to 0.00137, i.e. 13.7 per 10,000 people; see column (11) in table A in appendix 2. The CMF according to formula (2.2) is therefore equal to 13.71/8.71 = 1.575; see column (12). Multiplying the numerator and denominator by N_{is} / 10,000 = 38,287,704 / 10,000 shows that the direct standardised *number* of Turkish deaths in the period 1979-1986, the numerator of formula (2.3), is equal to 52,501, which is 1.575 times the number of deceased Dutch men of 33,336. The conclusion is that the standardised mortality among Turkish men up to 45 years of age is somewhat more than 1½ times

as high as the mortality for the Dutch nationals. The fact that the higher death rate is not constant over the age classes is demonstrated by the values of R_i in table A. For children, the mortality ratio is much larger than 1½.

If we had applied 'reverse standardisation', i.e. a comparison of the mortality of Dutch nationals (j) with that of Turkish immigrants as standard, then this would have produced a CMF of 0.619. This differs only minimally from the reciprocal of 1.575, but this is not generally true, because different standards are used.

We could have also included the age classes 45-65 and 65+ (table B). In that case, we would have obtained a CMF of 0.611 instead of 1.575! Not only is the mortality in the higher age classes among Turkish immigrants lower than among those of Dutch origin, these classes, by far, have the largest weight, because the most Dutch nationals die in them. The number of Turkish immigrants aged 65+ is even so small that a further split by cause of death is not possible, because the variance of the CMF by cause of death would increase too much. For this reason, indirect standardisation was used in Hoogenboezem and Israëls (1990); see section 3.4.

Assuming normality of the CMF, the 95% confidence interval for CMF is (1.462; 1.687), symmetrical around 1.575. If we assume normality of ℓn(CMF), which is a better option, then we obtain the asymmetrical confidence interval (1.466; 1.692). The difference is small. Due to the low mortality probabilities, we assumed that the Mortality variable is Poisson-distributed.

2.5 Characteristics

2.5.1 Relationship with the Laspeyres price index

The Methods Series report 'Index numbers' (Van der Grient and De Haan, 2011) presents the following formula:

$$P_L^{t,0} = \frac{\sum_i q_i^0 p_i^t}{\sum_i q_i^0 p_i^0} = \sum_i w_i^0 \frac{p_i^t}{p_i^0} = \sum_i w_i^0 I_i^{t,0} \quad . \tag{2.9}$$

Here, $P_L^{t,0}$ is the Laspeyres price index in reporting period t compared to base period 0, p_i^t is the average price of article i in reporting period t, q_i^0 is the consumed 'quantity' of article i in base period 0, and $w_i^0 = q_i^0 p_i^0 / \sum_i q_i^0 p_i^0$ is the weight of the single price index number $I_i^{t,0} = p_i^t / p_i^0$ of article i in the Laspeyres price index.

The same as for us, the q_i are relative contributions (consumption patterns for articles instead of age distributions), and the w_i are weights. Average prices p_i take the place of age-specific mortality rates Y_i, and index $I_i^{t,0}$ takes the place of ratio $R_i^{j,s} \equiv R_i$.

However, the interpretation of formulas (2.9) and (2.4) is somewhat different. For index numbers, we are always comparing average prices in two periods with one another, weighting the prices with quantities *q*. The populations are articles in two periods, between which an average price increase is defined. In demographic and health statistics, but also in other fields, the standardisation usually involves differences between populations at the same point in time. However, for long time series, Statistics Netherlands does standardise over time, with the population of a base year as the standard. If more than two populations are involved in the analysis, for price indices, there is always a time *ordered* series of price index numbers; for the standardisation of populations at the same point in time, the union of all populations is often used as the standard population.

For price index numbers, both the average prices and the quantities are stochastic, unless there is complete observation of prices and/or transactions. Individually measured prices are realisations of a quantitative variable. In standardisation for mortality or morbidity, the sizes (*N*) are usually fixed and only mortality is a stochastic variable, which, moreover, is binary. This simplifies the calculation of confidence intervals.

Prices can also be compared spatially/geographically (between countries) instead of over time; see the last paragraph of section 3.5.1 for more information.

2.5.2 Standardisation of nominal variables

Up to this point, variable D was the binary variable Mortality. We can calculate standardised averages more generally for each category of a nominal variable using a multinomial distribution; see De Ree and Israëls (1982). Per category, the same formulas are applicable as for ‘death/no death’. For example, in Hoogenboezem and Israëls (1990), the formulas are also applied to mortality by cause of death, even if indirect standardisation was ultimately selected in that situation (see section 3.5.2 for indirect standardisation for nominal variables). It is easy to show that the direct standardised mortality rates per cause of death add up to the direct standardised mortality rate for all causes of death together, and that the CMFs per cause of death add up in a weighted manner, with the number of deaths in the standard population as weights. Indeed, the denominators of the CMFs per cause of death also add up to the denominator of the CMF for the total mortality.

2.6 Quality indicators

- Besides calculating standard errors and performing tests (section 2.3.2), we can also study the stability of the solution by conducting a sensitivity analysis. For example, we can examine how the standardised figures react to combining age classes. If this leads to large differences, we have a problem: the solution is then apparently instable. A solution with more classes will have a greater variance, but less bias. The bias is only measurable if we assume a certain model, for example a linear relationship between age and the target variable. It therefore cannot always be determined whether the mean square error (mean quadratic

deviation) increases or decreases due to the combination of classes. As a rule, the choice will be made to combine classes if it gives a large reduction in variance.

- It is a good idea to not only compare the standardised mortality in population *j* with that in the standard population at an aggregated level, by determining the CMF and associated margin, but also to test whether age-specific mortality probabilities are equal; in other words, whether $\mu_{ij} = \mu_{is}$ for i=1,…,I. See appendix 1 for this test.

3. Indirect standardisation

3.1 Short description

In *indirect standardisation,* we calculate the Standard Mortality Ratio (SMR), which is also called the Standard Morbidity Ratio,

$$SMR = \frac{\sum_i q_{i|j} Y_{ij}}{\sum_i q_{i|j} Y_{is}} = \frac{Y_{\cdot j}}{\sum_i q_{i|j} Y_{is}} \quad . \tag{3.1}$$

The difference between this and the CMF is that the weights $q_{i|s}$ are replaced by $q_{i|j}$ in both the numerator and the denominator. The numerator is the gross mortality rate of population *j*; the denominator is the mortality rate in population *j* if the age-specific mortality rates were the same as those of the standard population. The SMR thus indicates proportionally how many more or fewer deaths there are in population *j* than in the standard population, if this had the age distribution of population *j*.

3.2 Applicability

We have seen in section 2.2, item 4 that direct standardisation leads to large standard errors if one or more age-specific mortality rates are based on small numbers and, despite this, still weigh heavily in the calculation. Indirect standardisation is not sensitive to this and is therefore preferred in this situation.

A second reason to use indirect standardisation is when direct standardisation is not possible because the necessary data are missing. Often, the age-specific mortality rates Y_{ij} (or the number of deaths D_{ij}) are not known for all populations *j*, and this means that the Y_j^{DIR} cannot be calculated. In indirect standardisation, the age-specific mortality rates are only needed for the standard population, and these are often still provided.

In summary, unreliable or missing age-specific mortality rates Y_{ij} are a reason for not using direct standardisation.

In indirect standardisation, the mortality per population *j* is compared with that of the standard population. However, comparing multiple populations *j with each other* can lead to interpretation problems. Direct standardisation is more suitable for that, because fixed weights are used, namely that of the age distribution of the standard population. Section 3.5 contains more information about this subject.

3.3 Detailed description

3.3.1 Determining the SMR

We can also write the SMR as

$$SMR = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{is}} = \frac{D_{+j}}{\sum_i N_{ij} D_{is} / N_{is}} = \frac{D_{+j}}{\sum_i (N_{ij} / N_{is}) D_{is}} \ . \qquad (3.2a)$$

Here, the numerator is the *observed* number of deaths in population *j*, and the denominator the *expected* number of deaths in population *j* if the mortality probability per age class would be equal to that of the standard population. The SMR is thus also represented as

$$SMR = O_j \big/ E_j \ , \qquad (3.2b)$$

where $O_j = D_{+j}$ stands for *observed count* and $E_j = \sum_i E_{ij} = \sum_i N_{ij} Y_{is}$ for *expected count*, a more frequently used notation in statistics. Actually, this is an elaboration of a model-based approach, in which a Poisson model is assumed for the number of deaths $\underline{D}_{ij}$ with expectation

$$E_{ij} \equiv E(\underline{D}_{ij}) = N_{ij} E(\underline{Y}_{ij}) = N_{ij} \mu_{ij} = N_{ij} Y_{is} \ . \qquad (3.3a)$$

The $\mu_{ij} = E(\underline{Y}_{ij})$ are parameters for the age-specific; in other words, μ_{ij} is the mortality probability for people from cell (*i*,*j*), with Y_{ij} as the realisation. Breslow and Day (1975) assume a multiplicative model for these mortality probabilities,

$$E(\underline{Y}_{ij}) = \mu_{ij} = \varphi_i \theta_j \ , \qquad (3.3b)$$

where φ_i is the effect of age on the mortality probability, and θ_j the effect of the population. This multiplicative model thus assumes that the mortality probabilities do not depend on the interaction 'Age x Population', which means that the observed age-specific mortality ratios R_i are reasonably homogeneous; see comment 2 in section 2.2.[3]

Substituting (3.3b) in (3.3a) means that it is assumed that the random variable $\underline{D}_{ij}$ has a Poisson distribution with parameter $(N_{ij}\varphi_i\theta_j)$. Model (3.3) can be seen as a Poisson regression (McCullagh and Nelder, 1989). A specific estimation of the parameters leads to the SMR as estimator for θ_j; see Breslow and Day (1975). We will come back to this in subsection 4.5.2.

An alternative form for formula (3.2a), comparable to formula (2.4) for the CMF, is

[3] The parameters φ_i are the effect of Age on Mortality, but may also be the effect of several distorting characteristics, such as Age x Income. One may also exclude interactions between such characteristics or use other kind of regression models for the estimation of the E_{ij}. For example, a logistic regression model without interactions has been used for the Hospital Standardised Mortality Ratio (HSMR) in Israëls et al. (2012). One still speaks about SMR and indirect standardisation.

$$SMR = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{is}} = \frac{\sum_i N_{ij} Y_{is} R_i}{\sum_i N_{ij} Y_{is}} = \sum_i w_i^* R_i \ , \qquad (3.4)$$

where $w_i^* = N_{ij} Y_{is} \Big/ \sum_i N_{ij} Y_{is}$. Therefore, like the CMF, the SMR is a weighted average of the age-specific ratios R_i, although the weights are more difficult to interpret. It is therefore not surprising that a number of convenient characteristics that apply to direct standardisation do not apply to indirect standardisation. We will discuss this further in section 3.5.

A third representation of the SMR is:

$$SMR = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{is}} = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{ij} / R_i} = \left(\frac{\sum_i N_{ij} Y_{ij} R_i^{-1}}{\sum_i N_{ij} Y_{ij}} \right)^{-1} = \left(\sum_i w_{ij} R_i^{-1} \right)^{-1} \qquad (3.5)$$

where $w_{ij} = D_{ij} / D_{+j}$. In section 3.5, we demonstrate the similarity of this to the Paasche price index.

Like for direct standardisation, the ratio (SMR instead of CMF) can be brought to the level of (standardised) mortality rates by multiplying by $Y_{.s}$, the gross mortality rate in the standard population:

$$Y_j^{INDIR} = SMR \times Y_{.s} = \frac{Y_{.j}}{\sum_i q_{i|j} Y_{is}} Y_{.s} \ . \qquad (3.6)$$

Y_j^{INDIR} may be called the *indirect standardised mortality <u>rate</u>*. Likewise, multiplying the SMR by D_{+s} leads to the *indirect standardised number of deaths* in population *j*. Usually, however, we limit ourselves to the SMR. The other indirect standardised figures do not provide any additional interpretation, and are not recommended. Fleiss (1973, p. 169) demonstrates that Y_j^{INDIR} can be larger (or smaller) than all Y_{ij}, which is undesirable.

3.3.2 Standard error of SMR

We could assume that $\underline{D}_{+j}$ is binomially distributed with parameters N_{+j} and mortality probability $p_{.j}$, where $\hat{p}_{.j} = Y_{.j}$. However, mortality probabilities differ strongly between age classes. For this reason, analogous to section 2.3.2, we assume a binomial distribution of $\underline{D}_{ij}$ with parameters N_{ij} and p_{ij} (i=1,…,I) . In this situation, in each age class, everyone has the same mortality probability. Making use of formula (2.5), the variance of $\underline{D}_{+j}$ is

$$Var\,(\underline{D}_{+j}) = Var\,(\sum_i \underline{D}_{ij}) = \sum_i Var\,(\underline{D}_{ij}) = \sum_i N_{ij}\, p_{ij} (1 - p_{ij}) \ . \qquad (3.7)$$

The estimator of this is

$$\text{var}(\underline{D}_{+j}) = \sum_i N_{ij} Y_{ij} (1 - Y_{ij}) = \sum_i D_{ij} (1 - \frac{D_{ij}}{N_{ij}}) \ , \qquad (3.8)$$

According to formula (3.2a), this leads to

$$\text{var}(SMR) = \text{var}(\frac{\underline{D}_{+j}}{E_j}) = \frac{1}{E_j^2} \sum_i D_{ij} (1 - \frac{D_{ij}}{N_{ij}}) \ , \qquad (3.9)$$

where E_j is the expected number of deaths. We assume here that E_j is not stochastic, which means that D_{is} is understood to be non-stochastic. This can be justified by looking at the issue in a model-based way, as Breslow and Day do (see formula (3.3)), or when the D_{is} are so large that their effect on the variance of the SMR is negligible. Without the model assumption, the variance of the SMR will be larger, because D_{is} is then considered to be stochastic.

Just as in direct standardisation, the formulas become simpler if we assume that $\underline{D}_{ij}$ is Poisson distributed. From formula (3.9), it then follows that the variance of SMR is equal to SMR^2/D_{+j}. For confidence intervals and for the test of whether SMR = 1, a log transformation must first be performed on the SMR, like for the CMF in section 2.4 (Breslow and Day, 1987).

3.4 Example

Mortality figures for Turkish and Dutch men, 0-44 years of age: indirect standardisation

We will now apply indirect standardisation to table A in appendix 2. The results can be found in columns (13) and (14).

The number of deaths per 10,000 Turkish men, Y_j, in the period 1979-1986, was equal to 14.6 (column 8). This is the numerator in formula (3.1). The denominator, the mortality figure per 10,000 Turkish men if the age-specific mortality rates were the same as those for Dutch men, is equal to 9.05 (column 13). The SMR for Turkish men compared to Dutch men is therefore equal to 14.6/9.05 = 1.616, which is slightly higher than the CMF of 1.575 from section 2.4.[4] This is the case because there are relatively more Turkish people than Dutch people in the age classes with high mortality ratios R_i, namely the 1 to 14-year-olds; compare formulas (2.2) and (3.1).[5] The difference with direct standardisation for the population up to 45 years of age is therefore not very large. If we had limited the case to this age, direct standardisation could also have been used. Hoogenboezem and Israëls (1990) used indirect standardisation, because direct standardisation would have led to large

[4] Note that both standardised mortality figures are smaller than the gross mortality rate of 1.68 (column 10). This means that age explains part of the higher mortality among Turkish men.

[5] Y_j^{INDIR} , the indirect standardised mortality rate according to formula (3.6), is equal to 1.616 x 8.7 x 10^{-4} = 0.00141, i.e. 14.1 deaths per 10,000.

standard errors for other ethnic groups, especially for Turkish and Moroccan women. Furthermore, the study was comparing immigrants with native Dutch people, and not the different immigrant groups to each other, as was explained in section 2.4.

If we consider all the ages, according to the distribution of table B in appendix 2, then direct standardisation does deviate strongly from indirect standardisation: CMF = 0.611 and SMR = 0.920. This difference is caused by the much smaller share of the age categories 45-64 and 65+ for Turkish men compared to Dutch men. Relatively few Turkish men died in these categories.

The 'reverse standardisation', i.e. the mortality of Dutch men (*j*) compared to the Turkish men as standard, generates an SMR of 0.635. This differs only very little from the reciprocal of 1.616, but this is not generally true, because different standards are used.

Assuming normality of the SMR, the 95% confidence interval for SMR (1.507; 1.726), is symmetrical around 1.616. If we assume normality of ℓn(SMR), which is a better option, then we obtain the asymmetrical confidence interval (1.510; 1.730). The difference is small. Due to the low mortality probabilities, we have assumed here that the Mortality variable is Poisson distributed.

3.5 Characteristics

3.5.1 Relationship with the Paasche price index

The Methods Series report 'Index numbers' (Van der Grient and De Haan, 2011) presents the following formula:

$$P_P^{t,0} = \frac{\sum_i q_i^t p_i^t}{\sum_i q_i^t p_i^0} = \left[\sum_i w_i^t \left(\frac{p_i^t}{p_i^0} \right)^{-1} \right]^{-1} = \left[\sum_i w_i^t \left(I_i^{t,0} \right)^{-1} \right]^{-1}, \qquad (3.10)$$

where $P_P^{t,0}$ is the Paasche price index in reporting period *t* compared to base period 0, p_i^t is the average price of article *i* in reporting period *t*, q_i^t the consumed quantity of article *i* in reporting period *t* and $w_i^t = q_i^t p_i^t / \sum_i q_i^t p_i^t$ the weight of the single price index number $I_i^{t,0} = p_i^t / p_i^0$ of article *i* in the Paasche price index. We see here the similarity between formulas (3.10) and (3.5), analogous to the similarity between the CMF and the Laspeyres price index number which is described in section 2.5.

Prices can also be compared spatially/geographically (between countries) instead of over time. Such international 'purchasing power parities' (Van der Grient and De Haan, 2011) are determined using both the formulae of direct and indirect standardisation, after which an average of the two is taken. Such a procedure is not often used in standardisation methods, as it complicates the interpretation.

3.5.2 Standardisation of nominal variables

In subsection 2.5.2, we discussed the situation in which the total deaths are split by cause of death. The same formulas apply for each cause of death as for 'total number of deaths'. Because for the SMR, like for the CMF, both the numerators and the denominators for the causes of death add up to the numerator and denominator for all causes of death together respectively, the total SMR is also a weighted sum of the SMRs per cause of death. The expected numbers of deaths per cause of death form the weights.

3.5.3 Comparing the SMRs of two populations with the standard

In section 3.2 we already stated that a comparison between two populations using indirect standardisation is problematic. Fleiss (1973, p. 161) gives an example of two populations j and j' with exactly the same age-specific mortality rates, i.e. $Y_{ij}=Y_{ij'}$, but with different age distributions. This leads to SMR$\{j{:}s\} \neq$ SMR$\{j'{:}s\}$; see, for example, formula (3.4). Hence, in indirect standardisation there is no complete adjustment for age differences between the populations (Rothman, 1986). However, in practice, the differences are usually small. The two SMRs are the same if s is the union of the two populations j and j'. In that case, $Y_{ij}=Y_{ij'}=Y_{is}$ and both SMRs are equal to 1. More generally, if $Y_{ij}/Y_{ij'}$ is constant, the ratio of the SMRs is equal to that constant. Formula (2.4) shows that if $Y_{ij}=Y_{ij'}$, then it is always true that CMF$\{j{:}s\}$ = CMF$\{j{:}s\}$, because a fixed standard is used.

3.6 Quality indicators

See section 2.6.

4. Regression analysis

4.1 Short description

In regression analysis, a dependent variable Y is written as a function of one or more explanatory variables. We usually use the linear (additive) regression model $Y=X\beta+\varepsilon$. For each X-variable included in the model, the associated β-parameter indicates its effect on Y, after adjustment for the effect of the other X-variables on Y. Qualitative explanatory variables can be included by creating dummy variables. With regression analysis, we can adjust effects for one another; therefore, in our case, we can also adjust the effect of a population j on the mortality rate for age effects. We can also calculate averages for the populations, adjusted for such age effects. If Age and/or other distorting characteristics are categorized, we may call this 'regression standardisation'. Because the explanatory variables are qualitative, we could also call this method analysis of variance instead of regression analysis. In our case, it is an analysis of variance of Mortality on Age (classes i) and Population (classes j).

4.2 Applicability

In the case of multiple distorting characteristics, all interactions between these characteristics are automatically included in direct and indirect standardisation (however, see footnote 3). Combinations of these characteristics can be considered as a single 'product variable'. Regression analysis can also deal with models without interactions. These models are much more economical in the number of parameters. Quantitative X-variables (covariates) can also be included in the regression equation. In this sense, regression analysis can do more than standardisation. But standardisation methods are conceptually clearer for our objective (section 1.1). This is due mainly to the fact that, in direct and indirect standardisation, each population j is directly compared (in pairs) with population s, and because only one age distribution, N_{ij} or N_{is}, is used for this purpose. In 'regression standardisation', in the case of a sum population, all (two or more) studied populations are jointly analysed, and the age distributions of all populations have an effect on the end result. That this is less simple already follows from the fact that an inverse must be calculated to obtain results. Intuitively, regression analysis seems to be similar to standardisation, as demonstrated in section 4.1. In section 4.5, we will further address the comparison between regression analysis and standardisation. We will then see that direct standardisation is equivalent to a weighted form of regression analysis (section 4.5.1)

It does not matter whether the target variable Y with scores Y_{ijk} is quantitative or binary (dummy variable). In the latter case, logistic regression is an alternative; see section 4.5.2. The regression analyses can often take place in an aggregated manner, just as for direct and indirect standardisation.

4.3 Detailed description

For the regression analysis, we can assume, on the one hand, a situation in which the standard population *s* is the union of all studied populations *j*, such that formula (1.2) applies. This means that all populations *j* are compared to one other, but are also implicitly compared with their complement *s**j*. On the other hand, population *s* can be an external population of which population *j* is not a part. In this case, a regression analysis can be performed in which only populations *j* and *s* are included, with $j \cup s$ as the sum population. In this section, we assume the first situation with respect to notation, which means that, for example, N_{is} is the population size in age class *i* for all populations *j* together. In section 4.4, we will present an example with the second situation.

Because regression analysis is an individual model, we here use the notation at the individual level, as presented in section 1.5. D_{ij} is thus a dummy variable (Mortality) with the score D_{ijk}=1 if individual *k* from age class *i* and population *j* has died in the study period, and otherwise D_{ijk}=0 (*i*=1,…,*I*; *j*=1,…,*J*). We can use the notation Y_{ijk} instead of D_{ijk} to generalise the theory to quantitative variables. As stated in section 1.5, $Y_{ijk} = D_{ijk}$ for dummy variables, because N_{ijk}=1.

Analysis of variance of *Y* on the qualitative variables Age and Population can be represented as

$$Y_{ijk} = D_{ijk} = \beta_0 + \sum_i \beta_i^A X_{ik}^A + \sum_j \beta_j^P X_{jk}^P + \varepsilon_{ijk} \quad . \tag{4.1}$$

Here, X_i^A is the dummy variable for age class *i* (or more generally for the *i*th class of the distorting characteristic) and X_j^P for the *j*th population. Thus, $X_{ik}^A = 1$ if individual *k* is in the *i*th age class (and otherwise $X_{ik}^A = 0$), and $X_{jk}^P = 1$ if *k* is part of the *j*th population (and otherwise $X_{jk}^P = 0$). In addition, the ß's are regression coefficients and the ε_{ijk} are disturbances.

To identify the parameters, we impose constraints. For this purpose, we utilise the constraints used in a multiple classification analysis[6] (MCA):

$$\sum_i \sum_j \sum_k \beta_i^A = \sum_i \sum_j N_{ij} \beta_i^A = \sum_i N_{is} \beta_i^A = 0 \tag{4.2}$$

and

$$\sum_i \sum_j \sum_k \beta_j^P = \sum_i \sum_j N_{ij} \beta_j^P = \sum_j N_{+j} \beta_j^P = 0 \quad . \tag{4.3}$$

---

[6] Multiple classification analysis (MCA) is a procedure of SPSS that can only be performed in the syntax mode, using the ANOVA command. However, the MCA parameters pertaining to constraints (4.2) and (4.3) can also easily be derived afterwards from an analysis of variance in which other constraints are used. Section 4.4 explains this using an example.

This means that there is no specific reference category, but the parameters for each variable (distorting characteristic and population) have a weighted average of zero over the categories, using the category frequencies as weights.

Due to these constraints, $ß_0$ becomes equal to μ, the expectation of the number of deaths, and the *ß*-parameters are deviations from this. We can now see $\mu + \beta_j^P$ as the expected (average) score for population *j* after adjustment for the distorting characteristics. We can consider its estimator as the mortality rate standardised by means of regression analysis.

Minimisation of $\sum_i \sum_j \sum_k e_{ijk}^2 = \sum_i \sum_j N_{ij} e_{ij}^2$, the sum of squares of the residuals, leads to constraints for the parameter estimators b_i^A and b_j^P , which are analogous to (4.2) and (4.3). From this, it follows that the general average is equal to

$$\hat{\mu} = \overline{Y}_{\cdot s} = \overline{D}_s = \frac{1}{N_s} \sum_i \sum_j \sum_k D_{ijk} \quad . \tag{4.4}$$

The regression-standardised average mortality rate can now be defined as

$$Y_j^{REGR} = \hat{\mu} + b_j^P = \overline{Y}_{\cdot s} + b_j^P \quad . \tag{4.5}$$

Instead of using formula (4.1), which is based on individual observations, we can also estimate the parameters at aggregated level,

$$Y_{ij} = \overline{D}_{ij} = \mu + \sum_i \beta_i^A X_i^A + \sum_j \beta_j^P X_j^P + \varepsilon_{ij} \quad . \tag{4.6}$$

where $\varepsilon_{ij} = \frac{1}{N_{ij}} \sum_k \varepsilon_{ijk}$.

Minimisation of $\sum_i \sum_j N_{ij} e_{ij}^2$ by the parameters leads to the same parameter estimators, when using the same constraints.

In regression analysis, it is rather uncommon to add the average $\hat{\mu}$ to an estimated parameter, whereas this is usual for standardised averages. In SPSS – General Linear Model, there is an option comparable to (4.5), namely the Estimated Marginal Means procedure, but this option is not particularly useful.

4.4 Example

Mortality figures for Turkish and Dutch men, 0-44 years of age: regression standardisation

We now apply regression analysis to two populations: Turkish men (*j*) and Dutch men (*s*). The sum population is therefore $j \cup s$, and the total size in age class *i* is $N_{ij}+N_{is}$. For this, we use the numbers from columns (2) and (3) of table A; the columns (6) and (7) from this table indicate how many of them have a score of 1 on the target variable of Mortality; the others have the score of 0.

The estimation of formula (4.1) under the MCA constraints (4.2) and (4.3) leads to $\hat{\mu} = 0.000879$ and $\beta^{P}_{Turk} = 0.000548$, as shown in column 1 of table 1. The mortality figure standardised by regression for Turkish men is therefore equal to $\hat{\mu} + \beta^{P}_{Turk} = 0.00143$. In direct and indirect standardisation, our figures were $Y_j^{DIR} = 0.00137$ and $Y_j^{INDIR} = 0.00141$ respectively; the difference is minimal. Note that, in indirect and regression standardisation, the same weights N_{ij} are used: this means that all individuals from population *j* count the same (see section 4.5.2). However, in the regression standardisation in this example, we are forced to use slightly different overall weights, because we are working with the union of Turkish and Dutch men N_{ij}+N_{is}. Because N_{is} is large compared to N_{ij} , that makes little difference in this example.

Table 1. Parameter estimates for regression standardisation with Turkish men (j) and Dutch men (s) as populations

Variable	Parameter estimates	
	$\sum_{i=1}^{6} N_{ij\cup s}\hat{\beta}_i^{A} = \sum_{j=1}^{2} N_{+j}\hat{\beta}_j^{P} = 0$ (MCA constraints)	$\hat{\beta}_6^{A} = \hat{\beta}_2^{P} = 0$ (reference categories)
Constant term	0.000879	0.00162
Age		
0	0.003464	0.002720
1-4	-0.000358	-0.001103
5-14	-0.000610	-0.001355
15-24	-0.000151	-0.000896
25-34	-0.000052	-0.000797
35-44	0.000745	0
Population		
Turkish men	0.000548	0.000556
Dutch men	-0.000008	0

The last column of table 1 shows the parameter estimates of an analysis of variance in which the last category of Age and Population is omitted. As stated in footnote 6, we can easily calculate the parameter estimates of the MCA solution from this. This can be done by subtracting the corresponding weighting average (weighted with the numbers of people) from each parameter, for each variable. The subtracted averages are then added to the constant term. Note that the differences between estimated parameters associated with the same variables are the same in column 1 and 2.

4.5 Characteristics

4.5.1 Direct standardisation and regression

Israëls and De Ree (1981) demonstrate that the direct standardised mortality rate Y_j^{DIR} can be obtained as the result of a *weighted* linear regression, when comparing a number of populations *j* with each other and with their sum population *s*. Direct standardisation uses the same additive model (4.6) as regression analysis, but with a different loss function for the estimation of the ß-parameters: $\sum_i \sum_j N_{is} e_{ij}^2$ is minimised instead of $\sum_i \sum_j N_{ij} e_{ij}^2$. This means that the squared deviations e_{ij}^2 are not weighted with their actual cell frequencies N_{ij}, but with the frequencies N_{is} from the standard population. The constraints from (4.2) and (4.3) are adapted as a result. Constraint (4.3) now becomes $\sum_j \beta_j^P = 0$, which is easy to see when replacing N_{ij} by N_{is}; the weights are no longer dependent on *j*. Using this loss function and these constraints, the estimator for $\mu + \beta_j^P$ will now be equal to the direct standardised average.

Because, in unweighted regression analysis in the case of a sum population *s*, all individuals count the same, 'regression standardisation' has smaller standard errors than direct standardisation, and therefore leads to more efficient estimators for ß, if the model applies. Whereas in regression standardisation the parameters for all populations *j* must be simultaneously estimated, in direct standardisation this can also be done separately for each population *j*. As a result, direct standardisation is simpler and more transparent.

4.5.2 Indirect standardisation and regression

In contrast to direct standardisation, indirect standardisation uses N_{ij} for weighting the mortality rates Y_{ij}, as can be seen in the formulas from chapter 3. All individuals therefore count the same. In this perspective, indirect standardisation is more similar to regression standardisation than direct standardisation is. On the other hand, indirect standardisation is not based on an additive model for the mortality probability, but on a multiplicative model for the denominator of the SMR, as we saw in formula (3.3b).

There, we demonstrated that the model behind the SMR can be considered as a multiplicative regression of Y_{ij} on the variables of Age and Population. We can write formula (3.3b), $E(Y_{ij}) = \mu_{ij} = \varphi_i \theta_j$, analogously to formula (4.1) and (4.6) with parameters μ, β_i^A and β_j^P instead of φ_i and θ_j. If we take the standard population as reference group *j'* (= *s*\j), β_j^P becomes the ratio of the rates between population *j* and the reference group, adjusted for age. In general (also for quantitative *Y*-variables), we could estimate the parameters by taking the logarithms to the left and

the right of the equals sign and applying the least squares method. However, if the numbers of deaths $\underline{D}_{ij}$ are Poisson distributed with expectation parameters $N_{ij}\mu_{ij} = N_{ij}\varphi_i\theta_j$, then Poisson regression with maximum likelihood (ML) estimation is possible. Please note that this ML estimator deviates slightly from SMR{*j*:*s*}, as Breslow and Day (1975) demonstrate. The same thing happens when logistic regression is applied.

If *Y* is binary, as for the variable of Mortality (*D*), logistic regression on Age and Population is an alternative, making a distinction between the populations *j* and *s*\j. We have performed that for the example from section 3.4. The logistic regression model is $\ell n\{p/(1 - p)\} = X\beta$ or $p/(1-p) = \exp(X\beta)$, where $\exp(\beta_j^P)$ represents the effect on the odds ratio "death/no death" for population *j* compared to the reference population *j'*, after adjusting for age. If (1-*p*) is almost 1, this is a reasonable approximation for the SMR. ML estimation gives $\hat{\beta}_j^P = 1.616$, which, when rounded, is identical to the SMR in example 3.4, and it also leads to nearly the same confidence interval. If the standard population is large and *p* is very small, this is a good approximation. If the standard population is much smaller, then the value of $\hat{\beta}_j^P$ changes, which is undesirable if we want to really standardise. In that case, the SMR obviously remains equal to 1.616.

5. Comparison of the methods

In this theme report, we discussed the two best known standardisation methods, direct and indirect standardisation. The related indexes for the ratio of the mortality rates between the population under study and the standard population are the CMF and SRM respectively. From a theoretical perspective, the indices are equal if the ratio of the age-specific mortality rates in studied population *j* and the standard population *s* is constant; in other words, $Y_{ij}=c.Y_{is}$ for $i=1,\ldots,I$, or $R_i=c$. The weighting of these age-specific mortality ratios R_i is irrelevant in this case. Large differences between the CMF and SRM only arise if the R_i are strongly heterogeneous. However, in that case, the more obvious choice is to publish age-specific mortality figures than of standardised figures. Standardised rates will still be published if there is an obligation to do so, e.g. if standardised figures must be delivered for a lot of different populations (for example, for all 20 causes of death from a certain classification or for all European countries).

The CMF and SMR are obviously also equal if *j* and *s* have the same age distribution ($q_{i|j}=q_{i|s}$). In that case, no adjustment for age is necessary. Finally, they are equal in expectation if the differences in mortality rate and age distribution are uncorrelated.

Kilpatrick (1962) sees the CMF and SMR as estimators for the same parameter, assuming a perfect homogeneity of mortality rates. In that case, it is possible to indicate in which situation the CMF is more efficient than the SMR, and vice versa; see also Van der Maas and Habbema (1981). It can be efficient to take a weighted or unweighted average of the two. However, in practice, the homogeneity will never apply exactly, and this type of combined index for descriptive statistics is less transparent.

In that sense, regression analysis is somewhat more problematic in that the standardisation is less transparent if there are multiple populations with the sum population as standard. Direct and indirect standardisation can always be determined in pairs (*j*,*s*), whereas the regression result also depends on the other populations, *j'*. If only one population *j* is compared with the standard *s*, a sum population must be created in order to perform an unweighted regression analysis.

6. References

Breslow, N.E. and N.E. Day (1975), Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases* 28, 289-303.

Breslow, N.E. and N.E. Day (1987), *Statistical methods in cancer research, Volume II – The design and analysis of cohort studies*. International Agency for Research on Cancer, Scientific Publications no. 82, Lyon.

Chiang, C.L. (1961), Standard errors of the age-adjusted death rate. *USGPO, Vital statistics* 47, 275-285.

Chiang, C.L. (1984), *The life table and its applications*. Robert Krieger Publishing Company, Malabar, Florida.

De Ree, S.J.M. and A.Z. Israëls (1982), *Een noot over het gebruik van standaardisatie bij nominale variabelen*. Internal report, Statistics Netherlands, Voorburg.

Fleiss, J.L. (1973), *Statistical methods for rates and proportions*. Wiley, New York.

Hoogenboezem, J. and A.Z. Israëls (1990), Sterfte naar doodsoorzaak onder Turkse en Marokkaanse ingezetenen in Nederland, 1979-1988. *Maandbericht Gezondheidsstatistiek* 9 (8), 5-20.

Israëls, A.Z. and S.J.M. de Ree (1981), *Standaardisatie, met toepassing op het Loonstructuuronderzoek 1972*. Internal report, Statistics Netherlands, Voorburg.

Israëls, A., J. van der Laan, J. van den Akker-Ploemacher and A. de Bruin (2012), *HSMR 2011: Methodological report*. Statistics Netherlands, The Hague. http://www.cbs.nl/en-GB/menu/themas/gezondheid-welzijn/publicaties/artikelen/archief/2012/2012-hsmr2011-method-report.htm.

Keiding, N. (1987), The method of expected number of deaths, 1786–1886–1986. *International Statistical Review* 55, 1-20.

Kilpatrick, S.J. (1962), Occupational mortality indices. *Population studies* 16, 175-187.

McCullagh, P. and J.A. Nelder (1989), *Generalized linear models*. Chapman and Hall, London.

Molenaar, W. (1973), *Simple approximations to the Poisson, binomial and hypergeometric distributions*. Mathematisch Centrum, Report SW 9/73, Amsterdam.

Rothman, K.J. (1986), *Modern epidemiology*. Little Brown and Co, Boston.

Van der Grient, H.A. en J. de Haan (2008), *Index numbers*. Methods Series document, Statistics Netherlands, The Hague [English translation from Dutch in 2011].

Van der Maas, P.J. and J.D.F. Habbema (1981), Standaardiseren van ziekte- en sterftecijfers: mogelijkheden en beperkingen. *Tijdschrift voor Sociale Geneeskunde* 59 (8), 259-270.

Van der Meulen, A. (2009), *Theme: Life tables and Survival analysis, Subtheme: Life tables*. Methods Series document, Statistics Netherlands, The Hague [English translation from Dutch in 2012].

Yule, G.U. (1934), On some points relating to vital statistics, more especially statistics on occupational mortality. *Journal of the Royal Statistical Society* 97, 1-84.

Appendix 1. Equality test of age-specific mortality probabilities

We want to test whether two age-specific mortality probabilities, μ_{ij} and μ_{is} (or μ_{ij} and $\mu_{ij'}$) are the same for a certain *i*. For this hypothesis, their estimators $Y_{ij} = D_{ij} / N_{ij}$ and $Y_{is} = D_{is} / N_{is}$ do not differ significantly. We can also formulate the hypothesis as $\mu_{ij} / \mu_{is} = 1$, for which $R_i = Y_{ij} / Y_{is}$ is the test statistic.

For the sake of generality, we assume that D_{ij} and D_{is} are parameters subject to chance and therefore presuppose a superpopulation. We assume that D_{ij} and D_{is} are binomially distributed, with the mortality probabilities p_{ij} and p_{is} respectively. We can now test our hypothesis using the 'Fisher exact test'. The required statistics, or actually the realisations thereof, are included in table A; the number of survivors is indicated by '*A'*. Conditional to the total row and total column of this table, the number of deaths D_{ij} is then hypergeometrically distributed with parameters N_{ij}, D_{i+} and N_{i+}. If D_{ij} deviates too strongly from $N_{ij} D_{i+} / N_{i+}$, this leads to the rejection of the hypothesis $\mu_{ij} = \mu_{is}$.

Table A. Frequency table for the Fisher exact test

	Number of deaths	Number of survivors	Total
j (Turkish people)	D_{ij}	$A_{ij} = N_{ij}$-D_{ij}	N_{ij}
s (Dutch people)	D_{is}	$A_{is} = N_{is}$-D_{is}	N_{is}
Total	D_{i+}	$A_{i+} = N_{i+}$-D_{i+}	N_{i+}

The probability that the hypergeometrically distributed D (unknown number of deaths in population *j*) is smaller than or equal to the realisation D_{ij}, i.e. $P[D \le D_{ij}]$, can be approximated by a Poisson distribution

$$P[D \le D_{ij}] \approx P[v \le D_{ij} \mid \mu = \frac{(2N_{ij} - D_{ij})(2D_{i+} - D_{ij})}{2(N_{is} + A_{i+} + 1)}]$$

where *v* is a Poisson-distributed statistic with parameter μ. Another approximation for $P[D \le D_{ij}]$ is

$$P[D \le D_{ij}] \approx P[u \le 2\frac{\sqrt{(D_{ij} + 1)(A_{is} + 1)} - \sqrt{A_{ij} D_{is}}}{\sqrt{N_{i+} - 1}}] \ .$$

where *u* is a standard-normal distributed statistic. The best way is to use the Poisson approximation if μ is smaller than 30. See Molenaar (1973) for more information about these approximations. If the sample is large enough, a Chi-squared test can be used for table A instead of the hypergeometric test.

For the simultaneous test of $\mu_{ij} = \mu_{is}$ for all age classes (i=1,…,I), we can use the Chi-squared test for an Ix2 table, with the number of deaths and the number of survivors as columns. For quantitative dependent variables, the t-test and F-test can be used.

Appendix 2. Use of standardisation on mortality figures of Turkish men (j) compared to Dutch men (s), 1979-1986

Table A. Age up to 45 years

(1) i	(2) N_{ij}	(3) N_{is}	(4) $q_{i\|j}$	(5) $q_{i\|s}$	(6) D_{ij}	(7) D_{is}	(8) $Y_{ij} = D_{ij}/N_{ij}$	(9) $Y_{is} = D_{is}/N_{is}$	(10) $R_i = Y_{ij}/Y_{is}$	(11) $q_{i\|s}Y_{ij}$	(12) CMF formula (2.2)	(13) $q_{i\|j}Y_{is}$	(14) SMR formula (3.1)
			%	%			per 10,000	per 10,000		per 10,000		per 10,000	
0	18 298	671 538	3.20	1.75	138	2 863	75.4	42.6	1.77	1.32		1.37	
1-4	68 183	2 700 772	11.94	7.05	87	1 372	12.8	5.1	2.51	0.90		0.61	
5-14	138 922	8 270 650	24.33	21.60	103	2 168	7.4	2.6	2.83	1.60		0.64	
15-24	135 924	9 777 094	23.80	25.54	158	7 054	11.6	7.2	1.61	2.97		1.72	
25-34	87 608	9 299 700	15.34	24.29	110	7 625	12.6	8.2	1.53	3.05		1.26	
35-44	122 078	7 567 950	21.38	19.77	239	12 254	19.6	16.2	1.21	3.87		3.46	
Total	571 013	38287 704	100	100	835	33 336	14.6	8.7	1.68	13.71	**1.575**	9.05	**1.616**

Key: (1) Age class; (2) Size of population *j* by age; (3) Size of population *s* by age; (4) Age distribution *j*; (5) Age distribution *s*; (6) Number of deaths *j*; (7) Number of deaths *s*; (8) Age-specific mortality rate *j*; (9) Age-specific mortality rate *s*; (10) Mortality ratio of population *j* compared to that of population *s*; (11) Calculation of numerator of CMF; (12) CMF; (13) Calculation of denominator of SMR; (14) SMR

Table B. All ages

(1) I	(2) N_{ij}	(3) N_{is}	(4) $q_{i\|j}$	(5) $q_{i\|s}$	(6) D_{ij}	(7) D_{is}	(8) $Y_{ij} = D_{ij}/N_{ij}$	(9) $Y_{is} = D_{is}/N_{is}$	(10) $R_i = Y_{ij}/Y_{is}$	(11) $q_{i\|s}Y_{ij}$	(12) CMF formula (2.2)	(13) $q_{i\|j}Y_{is}$	(14) SMR formula (3.1)
			%	%			per 10,000	per 10,000		per 10,000		per 10,000	
0	18 298	671 538	2.82	1.23	138	2 863	75.4	42.6	1.77	0.93		1.20	
1-4	68 183	2 700 772	10.51	4.94	87	1 372	12.8	5.1	2.51	0.63		0.53	
5-14	138 922	8 270 650	21.41	15.12	103	2 168	7.4	2.6	2.83	1.12		0.56	
15-24	135 924	9 777 094	20.95	17.87	158	7 054	11.6	7.2	1.61	2.08		1.51	
25-34	87 608	9 299 700	13.50	17.00	110	7 625	12.6	8.2	1.53	2.13		1.11	
35-44	122 078	7 567 950	18.81	13.83	239	12 254	19.6	16.2	1.21	2.71		3.05	
45-64	76 838	10 982 309	11.84	20.07	345	105 593	44.9	96.1	0.47	9.01		11.39	
65+	1 031	5 443 711	0.16	9.95	39	364 872	378.3	670.3	0.56	37.64		1.06	
Tot.	648 882	54 713 724	100	100	1 219	503 801	18.8	92.1	0.20	56.24	**0.611**	20.41	**0.920**

Key: (1) Age class; (2) Size of population *j* by age; (3) Size of population *s* by age; (4) Age distribution *j*; (5) Age distribution *s*; (6) Number of deaths *j*; (7) Number of deaths *s*; (8) Age-specific mortality rate *j*; (9) Age-specific mortality rate *s*; (10) Mortality ratio of population *j* compared to that of population *s*; (11) Calculation of numerator of CMF; (12) CMF; (13) Calculation of denominator of SMR; (14) SMR

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Standaardisatiemethoden				
1.0	21-01-2010	First Dutch version	Abby Israëls	Agnes de Bruin Heymerik van der Grient Sander Scholtus
1.1	01-05-2013	Minor corrections	Abby Israëls	
English version: Methods of Standardisation				
1.1E	01-05-2013	First English version	Abby Israëls	