PAA Extended Abstract - Please do not cite

How do populations aggregate?

Dennis M. Feehan*

March 22, 2019

Abstract

I show that when a population can be partitioned into subgroups, the death rate for the entire population can be written as the weighted harmonic mean of the death rates in each subgroup, where the weights are given by the numbers of deaths in each subgroup. In the full paper, I will provide some insight for why this phenomenon arises, and show that it generalizes to any type of occurrence-exposure rate. I use these relationships as the starting point for investigating how populations aggregate.

*UC Berkeley, feehan@berkeley.edu

1 Relationship

Most populations can be described at several different scales. For example, people living in the United States can be considered one large population or, at a finer scale, a collection of fifty different state populations.

Given a population and a partition of the population into subgroups, what is the relationship between the death rate in the subgroups and the death rate for the aggregate population? We show that occurrence-exposure rates aggregate across scales according to an elegant and well-understood mathematical relationship: the weighted harmonic mean.

Definition. Let $\vec{x}, \vec{w} \in \mathbb{R}^n$ and let $x_i > 0$ and $w_i > 0$ for all i. Then the **Weighted Harmonic Mean** of the $\vec{x}$ values, with weights given by the $\vec{w}$ values, is

$$H[\vec{x}; \vec{w}] = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}}.$$

Result. *Suppose that a population has been partitioned into K subgroups and let the death count, exposure, and death rate in subgroup i be ${}_nD_x^i$, ${}_nL_x^i$, and ${}_nM_x^i = \frac{{}_nD_x^i}{{}_nL_x^i}$. Then the death rate for the aggregate population, M, is the weighted harmonic mean of the death rates in the subgroups, with weights given by the number of deaths in each subgroup:*

$$M = H[\vec{M}; \vec{D}] = \frac{\sum_i {}_nD_x^i}{\sum_i \frac{{}_nD_x^i}{{}_nM_x^i}},$$

where $\vec{M}$ and $\vec{D}$ are vectors of the K group-specific death rates and death counts.

The weighted harmonic mean is a natural average to use for quantities whose reciprocals aggregate according to the usual (arithmetic) mean. Thus, the result suggests that the reciprocal of death rates – i.e., exposure per unit of death – is a quantity that aggregates in the usual way.

2 Proof

Proof. Since the subgroups form a partition of the aggregate population, each death falls into exactly one subgroup, and each unit of exposure falls into exactly one subgroup. Thus, the total exposure in the aggregate population can be written as ${}_nL_x = \sum_i {}_nL_x^i$ and the total number

of deaths can be written ${}_nD_x = \sum_i {}_nD_x^i$. The death rate in the aggregate population is thus ${}_nM_x = \frac{{}_nD_x}{{}_nL_x} = \frac{\sum_i {}_nD_x^i}{\sum_i {}_nL_x^i}$.

It remains to show that $\frac{\sum_i {}_nD_x^i}{\sum_i {}_nL_x^i}$ is equal to the weighted harmonic mean $H[\vec{M};\vec{D}]$. For each subgroup i, we have $\frac{{}_nD_x^i}{{}_nM_x^i} = {}_nD_x^i \times \frac{{}_nL_x^i}{{}_nD_x^i} = {}_nL_x^i$. Applying the definition of the weighted harmonic mean, we obtain

$$\begin{aligned} H[\vec{M};\vec{D}] &= \frac{\sum_i {}_nD_x^i}{\sum_i \frac{{}_nD_x^i}{{}_nM_x^i}} \\ &= \frac{\sum_i {}_nD_x^i}{\sum_i {}_nL_x^i} = {}_nM_x. \end{aligned}$$

□

3 Related relationships

The natural (arithmetic) mean can also be used to aggregate across subgroups, now using the exposure as the weights, rather than the deaths. So, using $AM[\vec{x},\vec{w}]$ to refer to the weighted arithmetic mean, we have:

$$HM[\vec{M};\vec{D}] = AM[\vec{M};\vec{L}],$$

or

$$\frac{\sum_i D_i}{\sum_i {}_nL_x} = \underbrace{\frac{\sum_i {}_nD_x}{\sum_i \frac{1}{{}_nM_x} {}_nD_x}}_{\text{Harmonic mean}} = \underbrace{\frac{\sum_i {}_nM_x \ {}_nL_x}{\sum_i {}_nL_x}}_{\text{Arithmetic mean}}$$

Thus, viewing aggregation from the perspective of the harmonic mean puts emphasis on the numbers of deaths, while the arithmetic mean focuses on the amount of exposure.

The derivation of the result does not depend on any particular feature of death rates, and would goes through for any type of occurrence-exposure rate; for example, it could also describe the aggregation of fertility rates across mutually exclusive subpopulations.

While aggregated occurrence-exposure rates can be decomposed using either the harmonic or the arithmetic mean, life table probabilities are not so flexible: they can be aggregated only according to the arithmetic mean (see the Appendix).

We can obtain interesting demographic relationships by consider a stationary population and taking

the limit as the width of the age interval used to calculate the rate, n, goes to 0 or goes to the entire age range.

The limit as the age range shrinks: hazards

As the width of the age interval goes to 0, the life table death rate becomes the instantaneous risk of death, i.e. the hazard:

$$\lim_{n \to 0} {}_n m_x = \mu_x.$$

We have

$$\begin{aligned}
\mu_x &= \lim_{n \to 0} {}_n m_x \\
&= \lim_{n \to 0} H[\vec{m}, \vec{d}] \\
&= \lim_{n \to 0} \frac{\sum_i {}_n d_x^i}{\sum_i \frac{{}_n d_x^i}{{}_n m_x^i}} \\
&= \frac{\sum_i \lim_{n \to 0} {}_n d_x^i}{\sum_i \lim_{n \to 0} \frac{{}_n d_x^i}{{}_n m_x^i}} \\
&= \frac{\sum_i l_x^i \mu_x^i}{\sum_i l_x^i} = \frac{\sum_i l_x^i \mu_x^i}{\sum_i \frac{l_x^i \mu_x^i}{\mu_x^i}}
\end{aligned}$$

provided that $\lim_{n \to 0} \frac{{}_n d_x^i}{{}_n m_x^i}$ exists and is nonzero.

Thus, the hazard of death in the aggregate population at exact age x aggregates in the same way as the death rates: it can be understood as a harmonic mean of the hazards, weighted by the numbers of deaths, or as the arithmetic mean of the hazards, weighted by the number of survivors to exact age x.

The limit as the age range gets wide: life expectancy

In a stationary population, life expectancy at birth is equal to the reciprocal of the life table crude death rate. Since the crude death rate can be considered the life table death rate over the entire age range, ${}_\infty m_0$, we have

$$\frac{1}{e_0} = {}_\infty m_0 = H[\vec{m}; \vec{d}],$$

where $\vec{m}$ and $\vec{d}$ are the K subgroup life table death rates and life table deaths. This relationship enables us to relate the life expectancy at birth in the aggregate group to the life expectancies at birth in the subgroups:

$$e_0 = \frac{1}{H[\vec{m}; \vec{d}]}.$$

Since the number of life table deaths is equal to the radix, i.e., ${}_\infty d_0^i = l_0^i$, and that the life table exposure over the entire age range is equal to the cohort years of life lived, i.e., ${}_\infty L_0^i = T_0^i$, we have

$$\begin{aligned} e_0 &= \frac{1}{H[\vec{m}; \vec{d}]} \\ &= \frac{\sum_i T_0^i}{\sum_i l_0^i} \\ &= \frac{\sum_i l_0^i e_0^i}{\sum_i l_0^i}. \end{aligned}$$

Thus, the aggregate life expectancy at birth is equal to the weighted arithmetic average of the subgroup life expectancies at birth, where the weights are given by the proportion of births in each subgroup. This result is consistent with the intuition that the harmonic mean makes sense for quantities that are additive on a reciprocal scale; thus, when considering the reciprocal of death rates – such as e_0 in the life table stationary population – arithmetic mean aggregation is appropriate.

The harmonic mean also yields a second decomposition of the aggregate e_0:

$$H[\vec{e_0}; \vec{T_0}] = \frac{\sum_i T_0^{(i)}}{\sum_i \frac{T_0^{(i)}}{e_0^{(i)}}} = \frac{T_0}{\sum_i l_0} = \frac{T_0}{l_0} = e_0.$$

In this second interpretation, e_0 plays the role of a rate, weighted by an amount of exposure.

4 History

To my knowledge, these harmonic mean decompositions have not been previously been studied in the context of aggregating demographic rates. However, there are several classic studies that are related, in that they study aggregation or use the harmonic mean in different ways. Rogers (1975) used formal demography to study different kinds of aggregation across regions, with particular attention to inter-regional migration flows. Keyfitz et al. (2005) (Sec. 1.5) analyzed the problem

of aggregation in the context of population projections and growth rates; in that analysis, Keyfitz finds that independently projecting subregions and adding the projections will, in general, produce estimated population size that is larger than the one obtained from projecting the aggregate directly. Schoen (2013) investigated the harmonic mean as it relates to two-sex population models, which incorporate marriage markets. More generally, there is statistical literature on the properties and uses of the harmonic mean (e.g., Carvalho 2016).

5 Applications

Understanding how populations aggregate has practical and theoretical applications. To illustrate a practical application, I present the results of an analysis that illustrates how choosing the wrong mean when aggregating can cause appreciable errors in realistic situations. I then turn to a discussion of how understanding aggregation via the harmonic mean may help us to conceptually understand the mortality process through a new lens.

5.1 Magnitude of potential errors when aggregating incorrectly

In this section, I show that appreciable errors can arise when death rates are not aggregated correctly. To do so, I create a pseudo-population based on the lifetables found in the US Mortality database[1]. The US Mortality database has life tables for each US State and for Washington, D.C. I treat these life tables as cohorts, each of which is initially the same size. I then compare three strategies for aggregating death rates for each single year of age, for each sex, and for both sexes combined.

The first strategy for aggregating death rates uses the definition of the aggregate death rate

$$ {}_nM_x = \frac{\sum_i {}_nd_x^{(i)}}{\sum_i {}_nL_x^{(i)}}. $$

We have seen elsewhere that this true aggregate is equal to $HM[\vec{M}, \vec{D}]$ or, alternatively, $AM[\vec{M}, \vec{L}]$.

Figure 1 shows the aggregated age-specific death rates from the pseudo-population created from the 51 state lifetables. Figure 2 shows the relative standard deviation in death rates across the 51 states at each age. I define the relative standard deviation to be

$$ \text{rel-sd} = \frac{\text{sd}({}_nM_x^{(i)})}{{}_nM_x}, $$

[1] https://usa.mortality.org/

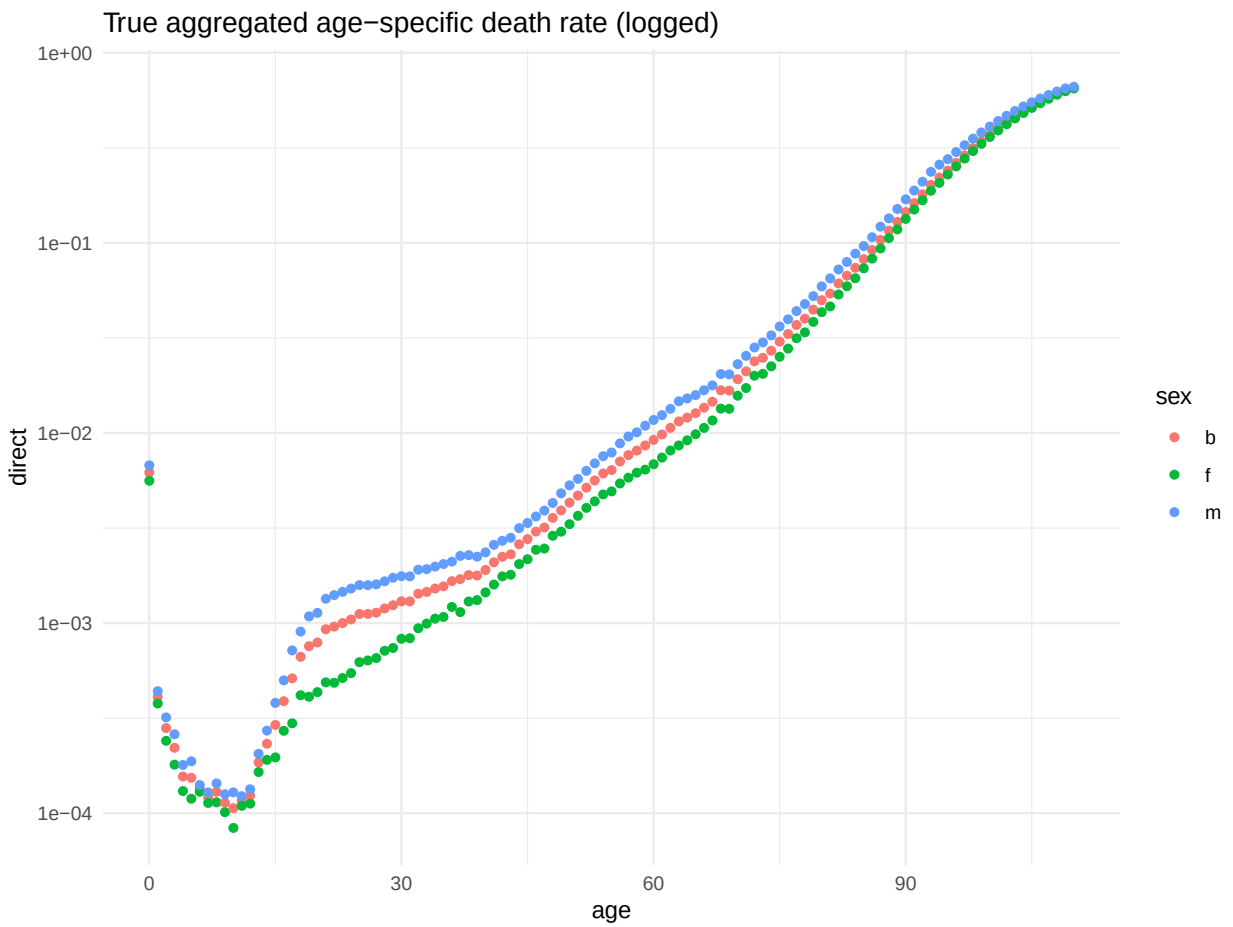

Figure 1: Aggregate death rates across 51 sub-national psuedo-units.

where $\text{sd}({}_nM_x^{(i)})$ is the standard deviation across the 51 state life table death rates. The relative standard deviation thus quantifies the amount of spread in death rates across the 51 subnational units, when compared to the true aggregate death rate.

For each age-sex group, I compare the true aggregate death rate to two alternate strategies for aggregating death rates. The first alternate strategy is the simple (unweighted) arithmetic mean

$$ {}_nM_x^I = \frac{1}{51}\sum_i {}_nM_x^{(i)}. $$

The second alternate strategy is the arithmetic mean weighted by the number of deaths, rather than the exposure:

$$ {}_nM_x^{II} = \frac{\sum_i {}_nM_x^{(i)}\,{}_nD_x^{(i)}}{\sum_i {}_nD_x^{(i)}}. $$

Figure 2: Relative standard deviation death rates by age and sex, across 51 sub-national units.

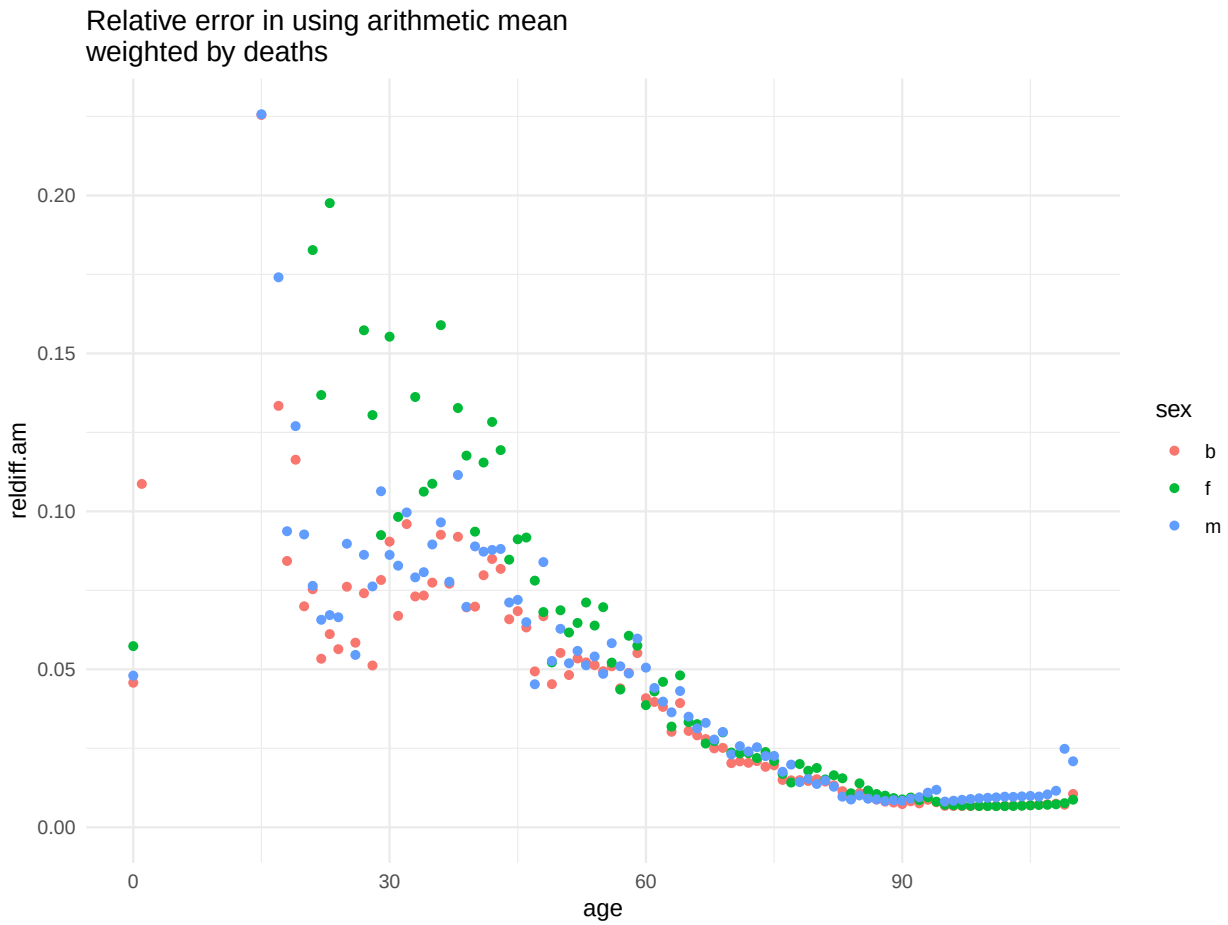

Figure 3: Relative errors in aggregating using the incorrectly weighted arithmetic mean

To summarize the error for each strategy, I plot the relative error, $\frac{{}_nM_x - {}_nM_x^{(1)}}{{}_nM_x}$.

5.2 A different lens on mortality

Focusing on aggregation via the harmonic mean may offer an alternative way to conceptualize the mortality process.

While the arithmetic mean is appropriate for aggregating quantities that are observed on a natural scale, the harmonic mean is usually appropriate for quantities whose natural scale is their reciprocal. The classic example is *length-biased sampling*, in which units are likely to be observed in a given state in proportion to the time spent in that state. Viewed from this perspective, the harmonic mean places emphasis on $\frac{1}{{}_nM_x}$. While the conventional death rate quantifies a number of deaths per unit of exposure, $\frac{1}{{}_nM_x}$ focuses on the amount of exposure accrued for each death. In a sense, the arithmetic mean takes the perspective that the we have observed in each subunit a certain amount of exposure, along with the associated deaths; it is thus appropriate to weight by the amount of

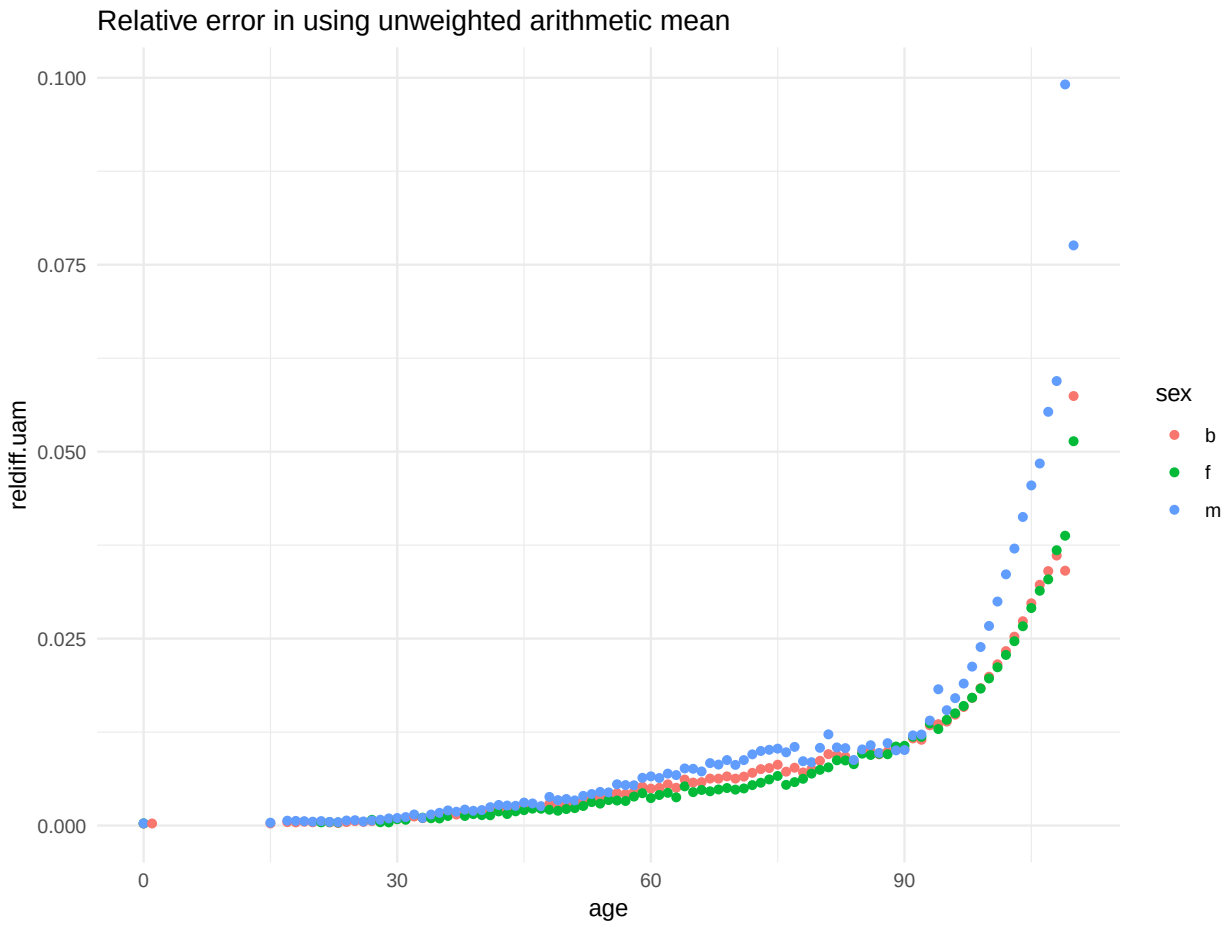

Figure 4: Relative errors in aggregating using the unweighted arithmetic mean

exposure. The harmonic mean relationship, on the other hand, takes the perspective that we have observed a number of deaths from each subunit, along with their associated exposure.

6 Next steps

The results developed in this extended abstract can be expanded in several ways. Two main areas of focus seem particularly important: first, I hope to expand and deepen the intuition behind the harmonic mean relationship, paying particular attention to the relationship between the harmonic mean and length-biased sampling (e.g. Carvalho 2016). Second, I hope to find further practical applications for the harmonic mean decompositions introduced in this study.

7 References

Carvalho, M. de. (2016). Mean, What do You Mean? *The American Statistician*, *70*(3), 270–274.

Keyfitz, N., Caswell, H., Caswell, H., & Keyfitz, N. (2005). *Applied mathematical demography* (Vol. 47). Springer.

Rogers, A. (1975). *Introduction to multiregional mathematical demography*. John Wiley & Sons.

Schoen, R. (2013). *Modeling multigroup populations*. Springer Science & Business Media.

8 Appendix

8.1 Aggregating probabilities

Here, I demonstrate that life table probabilities aggregate according to the usual, arithmetic mean, weighted by the number of synthetic cohort members who are alive at the start of the age interval.

Result. *Suppose that a population has been partitioned into K subgroups and let the probability of death in group i by ${}_nq_x^i$ and let the number of cohort members surviving to the start of the age interval in subgroup i be ${}_nl_x^i$. Then the life table probability of death for the aggregate population, ${}_nq_x$, is the weighted arithmetic mean of the corresponding life table probability in the subgroups, with weights given by the number of people at the start of the time period in each subgroup:*

$$M = A[\vec{q}; \vec{l}] = \frac{\sum_i l_x^i \ {}_nq_x^i}{\sum_i l_x^i},$$

where $\vec{q}$ and $\vec{l}$ are vectors of the K group-specific probabilities of death and surviving cohort members at the start of the age interval.

Proof. This follows from the fact that ${}_nq_x = \frac{{}_nd_x}{{}_nl_x}$. Applying the definition of the weighted harmonic mean, we obtain

$$\begin{aligned} A[\vec{q}; \vec{l}] &= \frac{\sum_i {}_nq_x^i \ l_x^i}{\sum_i l_x^i} \\ &= \frac{\sum_i \frac{{}_nd_x^i}{l_x^i} \ l^i}{\sum_i l_x^i} \\ &= \frac{\sum_i {}_nd_x^i \ l^i}{\sum_i l_x^i} \\ &= \frac{{}_nd_x}{l_x}. \end{aligned}$$

□