Noname manuscript No.
(will be inserted by the editor)

Estimation of a proportion in survey sampling using the logratio approach

Karel Hron · Matthias Templ · Peter Filzmoser

Received: date / Accepted: date

Abstract The estimation of a mean of a proportion is a frequent task in statistical survey analysis, and often such ratios are estimated from compositions such as income components, wage components, tax components, etc. In practice, the weighted arithmetic mean is regularly used to estimate the center of the data. However, this estimator is not appropriate if the ratios are estimated from compositions, because the sample space of compositional data is the simplex and not the usual Euclidean space. We demonstrate that the weighted geometric mean is useful for this purpose. Even for different sampling designs, the weighted geometric mean shows excellent behavior.

Keywords Survey sampling, Proportions, Compositional data, Logratio analysis

K. Hron
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, 771 46 Olomouc, Czech Republic
Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26, 771 46 Olomouc, Czech Republic
Tel.: +420585634605
Fax: +420585634002
E-mail: hronk@seznam.cz

M. Templ
Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 7, A-1040 Vienna, Austria
Tel.: +4315880110715
E-mail: templ@statistik.tuwien.ac.at

P. Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
Tel.: +4315880110733
E-mail: p.filzmoser@tuwien.ac.at

1 Introduction

Many surveys are concerned with the problem of estimating a mean of proportions. For example, one can be interested in the mean relative amount of time spent on working activities from the all-day activities, or the mean concentration of a pollutant in a study area. Usually, such examples are connected with a constant sum constraint, like 24 hours in the first case, or 1 (100) for proportions (percentages). However, this is not a necessary condition when considering proportional data. A counter-example are household expenditures of single-living persons: if the interest is on the relative contribution of foodstuffs on the overall expenditures, it is irrelevant if this part is expressed in euro or in percentages – both numbers lead to exactly the same information.

More formally, let o_i and O_i be the original values of two parts of a sample with $i = 1, \ldots, n$ observations/compositions, typically a part of interest for o_i and the remaining (amalgamized) parts for O_i. In our first example, o_i corresponds to the amount of time spent on working activities, and O_i on the other activities for the i-th person included in the sample. Then the mean proportion is usually estimated as $(\sum_{i=1}^{n} o_i)/(\sum_{i=1}^{n}(o_i + O_i))$, or it is directly estimated with the sample arithmetic mean $\frac{1}{n}\sum_{i=1}^{n} o_i$ when the data are already expressed in proportions or percentages. Nevertheless, this approach leads to some difficulties. In particular, these difficulties are connected with the concept of relative scale. Intuitively, the difference between one and two hours spent on working is not the same as between 10 and 11 hours. In the first case, two hours represent the double of one hour, while 11 hours is 1.1 times ten hours. The sample arithmetic mean estimator ignores the relative scale concept, with the consequence of misleading interpretations, for instance when comparing subpopulations. Note also that the standard mean estimators give in general two different answers if the data are provided with or without constant sum constraint. However, independent of the constraint we are interested in the same quantity, and thus a unique result should be expected.

Often we are not only interested in an estimation of the mean, but also in a conventionally calculated confidence interval around the mean. Using the above mean estimators for this purpose may lead to non-sense intervals, like negative values for percentage data. In this case, the interval can be simply cut to the non-negative part, but this violates the basic concept of a confidence interval as a tool for covering the true value of the population characteristic with a prescribed probability. The main reason for the undesired behavior of the confidence interval of percentage data is the improper underlying geometry. The percentages do not follow the usual Euclidean geometry in the real space, and consequently, also the Lebesgue measure as the underlying measure for any meaningful statistical inference is not appropriate there, see e.g. Mateu-Figueras and Pawlowsky-Glahn [2008] for details. Thus, although some technical adjustments of the normal confidence interval may seem to solve the problem of having negative values in interval estimators around the mean like in Kott and Liu [2009], the only concise solution can be to change the underlying geometry (and consequently also the measure) to a geometry that follows the properties of the sample space of proportional data [Pawlowsky-Glahn and Egozcue, 2001].

Another approach taking into account the relative scale is based on ratios $x_i = o_i/O_i$. Although in practice usually not only the ratios are available for the analysis, they are exclusively of interest when data carry only relative information, represented in proportions or percentages. Pawlowsky-Glahn and Buccianti [2011] demonstrate this for many examples

from natural and social sciences. For this reason, this case is of particular importance. The ratio does not change if the original data are expressed in percentages or in other units. This is also illustrated in Figure 1. The absolute values are shown with filled circles, their coordinate-wise arithmetic mean is indicated by the dashed lines and the open triangle. The grey lines from the origin represent the proportions. In fact, any data value on the grey line contains potentially the same information, namely the ratio of the original data. For reasons of comparability, the data can be normed such that the two parts forming the ratios sum up to 1. This constraint of sum 1 is indicated by the solid black line, and the symbols "+" are the projected data points. The solid triangle represents the projected mean. The center of the open circle is the mean of the projected data points, which differs from the projected mean. Formally,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} \frac{o_i}{O_i} \neq \frac{\frac{1}{n}\sum_{i=1}^{n} o_i}{\frac{1}{n}\sum_{i=1}^{n} O_i} = \frac{\bar{o}}{\bar{O}}. \tag{1}$$

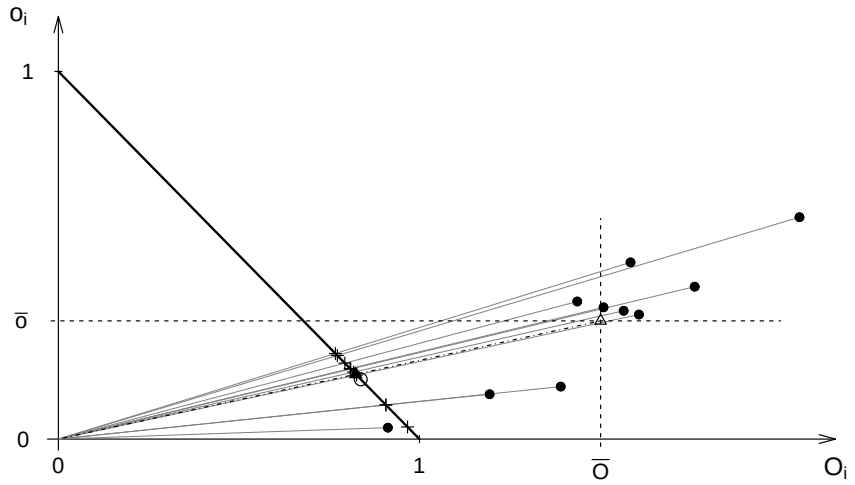


Fig. 1 Illustration of the geometry of proportional data (see text for detailed explanations).

For the geometric mean, on the other hand, we have

$$\bar{x}_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n} = \left(\prod_{i=1}^{n} \frac{o_i}{O_i}\right)^{1/n} = \frac{\left(\prod_{i=1}^{n} o_i\right)^{1/n}}{\left(\prod_{i=1}^{n} O_i\right)^{1/n}} = \frac{\bar{o}_G}{\bar{O}_G}. \tag{2}$$

Let us illustrate the above considerations (and extend them slightly) also with two numerical examples to show that the arithmetic mean is not *scale invariant* and does not have the property of *subcompositional coherence* [Egozcue, 2009], while the geometric mean does possess these properties.

Consider three sampled observations, compositions $\mathbf{x}_1 = (1, 2)$, $\mathbf{x}_2 = (1, 5)$, $\mathbf{x}_3 = (5, 5)$. They can easily be expressed in proportions that sum up to one – the result is $\mathbf{p}_1 = (0.333, 0.667)$, $\mathbf{p}_2 = (0.167, 0.833)$, $\mathbf{p}_3 = (0.500, 0.500)$. Both the arithmetic and geometric means are computed and we also rescale them to proportions. For the arithmetic mean we get $\overline{\mathbf{x}} = (0.368, 0.632)$ and $\overline{\mathbf{p}} = (0.333, 0.667)$, thus the ratios between their first and second parts equal 0.583 and 0.500, respectively. We conclude that the arithmetic mean is not *scale invariant* when dealing with relative information [Egozcue, 2009]. For the geometric mean we get in both cases $\mathbf{g} = (0.317, 0.683)$, with the corresponding ratio 0.464. The geometric mean preserves the ratios and therefore it is scale invariant. Note that these properties can be also generalized, see [Pawlowsky-Glahn and Egozcue, 2002] for details.

We go further and extend the previous example by adding one additional part: $\mathbf{x}_1^* = (1, 2, 5)$, $\mathbf{x}_2^* = (1, 5, 20)$, $\mathbf{x}_3^* = (5, 5, 5)$. Similar calculation as above lead to the arithmetic means $\overline{\mathbf{x}}^* = (0.143, 0.245, 0.612)$, the means of the proportions $\overline{\mathbf{p}}^* = (0.166, 0.259, 0.576)$ and the geometric means $\mathbf{g}^* = (0.128, 0.276, 0.595)$ (the values are again rescaled to sum one). For each estimator we compute again the ratio between the first and the second part, resulting in 0.583, 0.640 and 0.464, respectively. Obviously, the arithmetic mean from the original data and the geometric mean preserve the ratios from the previous case, but the result from the arithmetic mean after rescaling to proportions differs substantially (0.500 versus 0.640). Although the arithmetic mean of the proportions has the nice property that the constant sum constraint is preserved, it leads to inconsistent values, and it depends on the data scale.

Thus, for the geometric mean it is irrelevant whether ratios or the original data are used, the result is the same. This, however, only demonstrates the invariance of the geometric mean with respect to using original or the ratio data, and the lack of this invariance for the arithmetic mean, but it does not give an answer to the question which estimator or data information is appropriate for the problem.

In this paper we propose to use the logarithm of the ratio $x_i = o_i/O_i$ (logratio) in order to obtain an estimator that can be interpreted in the usual sense. (Note that the mean of the logratios can simply be transformed to the geometric mean.) In fact, data containing relative information are known under compositional data or compositions. The problem of statistical estimation with relative information is treated in many papers and books, e.g. in Aitchison [1986] and Pawlowsky-Glahn and Buccianti [2011]. The idea is to move the statistical analysis from the so-called Aitchison geometry on the simplex, that follows the nature of compositions, isometrically to the usual Euclidean space. Recently, the problem of sampling from finite populations of data carrying relative information was studied in Graf [2006a,b, 2011] for the Swiss Earnings Structure Survey. Here we focus not only on the case of estimating the mean, but we aim at going much deeper into the practical problems of survey sampling, like the treatment of different sampling schemes.

The paper is organized as follows: Section 2 provides details about the logratio approach to compositional data analysis. In Section 3 we discuss estimators for the population mean of data carrying relative information in case of finite populations. Numerical results with close-to-reality data sets are provided in Section 4. The final Section 5 concludes.

2 Elements of compositional data analysis

Generally, a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)$ consists of strictly positive parts $x_i > 0$, $i = 1, \ldots, D$, which carry only relative information, and the parts sum up to a constant, usually chosen as 1. The choice of the constant is not important, see Figure 1, and thus it is common to use the closure operation $\mathcal{C}$ that rescales the sum of the parts. Consequently, $\mathbf{x} \equiv \mathcal{C}(\mathbf{x})$ means that a composition $\mathbf{x}$ is expressed in terms of (an arbitrary) constant sum. Concretely, with the closure operation we identify the original composition $\mathbf{x} = (x_1, \ldots, x_D)$ with its representation, like in proportions,

$$\mathcal{C}(\mathbf{x}) = \left(\frac{x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i} \right).$$

This is a meaningful assumption. When e.g. relative contributions of single household expenditures on the total amount of expenditures are of interest rather than their absolute values in a specific (currency) unit, then the ratios between parts (that are exclusively of interest) obviously remain the same for $\mathbf{x}$ and $\mathcal{C}(\mathbf{x})$. In the following, we suppose that the compositions are already rescaled to unit constant sum; otherwise, the operator $\mathcal{C}$ is chosen to rescale the sum of parts to that constant.

2.1 Sample space and geometry of compositions

In this paper we are specifically interested in one compositional part x (e.g. proportion of food on all household expenditures). Accordingly, we deal with compositional data of the form $\mathbf{x} = (x, 1-x)$, $x > 0$, where possibly other compositional parts have been aggregated to the part $1-x$ (e.g. all different kinds of household expenditures except food, when exclusively the relative contribution of this part is of interest). Theoretical and practical aspects of the special case of analyzing "univariate" compositional data x are investigated in Filzmoser et al. [2009].

The set of all compositions $\mathbf{x}$, denoted as $\mathcal{S}$, forms a segment between the points $[1, 0]$ and $[0, 1]$ on the plane, see bold solid line in Figure 1. This one-dimensional subset of the first quadrant of the plane represents a special case of the simplex sample space, the sample space of D-part ($D \geq 2$) compositions. Since compositions are expressed in a relative scale, a special geometry on $\mathcal{S}$, different from the standard Euclidean one, needs to be used, when analyzing raw compositional data. This is nowadays known under the name Aitchison geometry [Egozcue and Pawlowsky-Glahn, 2006, Mateu-Figueras and Pawlowsky-Glahn, 2008] and results in operations of perturbation and power transformation as an analogon to the usual vector addition and scalar multiplication, as well as a new approach to norm and distance. Accordingly, for two 2-part compositions $\mathbf{x} = (x, 1-x)$ and $\mathbf{y} = (y, 1-y)$ and a real number α, the operations of perturbation and power transformation result in compositions

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[xy, (1-x)(1-y)] \quad \text{and} \quad \alpha \odot \mathbf{x} = \mathcal{C}[x^\alpha, (1-x)^\alpha],$$

respectively. Consequently, we set $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \odot \mathbf{y})$. The Aitchison norm and the Aitchison distance are real numbers

$$\|\mathbf{x}\|_A = \frac{1}{\sqrt{2}} \left| \ln \frac{x}{1-x} \right| \quad \text{and} \quad d_A(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{2}} \left| \ln \frac{x}{1-x} - \ln \frac{y}{1-y} \right|. \tag{3}$$

The Aitchison distance results in the definition of the Aitchison measure of the interval (a, b) of proportions, thus $\mathbf{a} = (a, 1-a)$ and $\mathbf{b} = (b, 1-b) \in \mathcal{S}$, as

$$\lambda_A(a, b) = \frac{1}{\sqrt{2}} \left| \ln \frac{b}{1-b} - \ln \frac{a}{1-a} \right|. \tag{4}$$

Together with the definition of the inner product (which is, however, trivial in the case of two-part compositions), the Aitchison geometry has all the well-known properties of the standard Euclidean geometry. The Aitchison geometry obviously follows the concept of relative scale. If we consider again the example of differences of time spent on working from the second paragraph of Section 1 (the values can be represented in proportions, or equivalently, taking the original constant sum constraint 24 into account), we get $\lambda_A(1, 2) = 0.74$ and $\lambda_A(10, 11) = 0.17$.

From Equations (3) and (4) it can be seen that the compositional information is processed by the logarithm of ratios (logratios). Setting

$$z = ilr(\mathbf{x}) = \frac{1}{\sqrt{2}} \ln \frac{x}{1-x} \tag{5}$$

(analogously for $\mathbf{y}$), the Aitchison norm and distance move to the usual Euclidean norm and distance on the real line and also $\lambda_A(a, b) = |ilr(\mathbf{b}) - ilr(\mathbf{a})|$. In fact, Equation (5) defines the one-to-one isometric logratio (ilr) transformation [Egozcue et al., 2003] that maps the compositions isometrically from the simplex with the Aitchison geometry to the (Euclidean) real line. This allows to use most of the standard statistical methods and estimators that rely on the Euclidean geometry for the transformed data, like the arithmetic mean or the empirical variance. Obviously, the interpretation of the ilr variable z is based on the description of the ratio between both compositional parts. Finally, note that the ilr transformation of two-part compositions reminds on the well known logit transformation used for the logistic regression model [Agresti, 2002]. However, for logistic regression the transformation is carried out mainly because of problems with the domain of the response variable, without any deeper geometrical background as presented above.

2.2 Distributional characteristics of compositions

In this section we want to characterize the distribution of compositions in order to define an appropriate estimator for the measure of central tendency. In particular, we justify theoretically why the geometric mean is appropriate as the corresponding estimator in the case of compositional data. This can be done directly on the simplex by taking advantage of the special properties of the Aitchison geometry [Pawlowsky-Glahn and Egozcue, 2002]. Consider a random two-part composition $\mathbf{X} = (X, 1-X)$. The center of the distribution of $\mathbf{X}$ is the $\boldsymbol{\xi} \in \mathcal{S}$ that minimizes the expectation $\mathsf{E}[d_A^2(\mathbf{X}, \boldsymbol{\xi})]$. Note that the random variable

$d_A^2(\mathbf{X}, \boldsymbol{\xi})$ for fixed $\boldsymbol{\xi}$ can be both discrete or continuous, depending on the number of realizations of the random composition $\mathbf{X}$. The minimum is reached at a value $\boldsymbol{\xi} = \mathrm{cen}(\mathbf{X})$, and it is called *center of* $\mathbf{X}$. Then, for two random compositions $\mathbf{X}$ and $\mathbf{Y}$, a real number a and a non-random composition $\mathbf{b} \in \mathcal{S}$, the following intuitive relations

$$\mathrm{cen}(\mathbf{X} \oplus \mathbf{Y}) = \mathrm{cen}(\mathbf{X}) \oplus \mathrm{cen}(\mathbf{Y}), \quad \mathrm{cen}(a \odot \mathbf{X} \oplus \mathbf{b}) = a \odot \mathrm{cen}(\mathbf{X}) \oplus \mathbf{b} \tag{6}$$

hold [Pawlowsky-Glahn and Egozcue, 2002]. In the following we will denote $\mathrm{cen}(\mathbf{X})$ simply by $\boldsymbol{\gamma}$. The mean value $\mathsf{E}[d_A^2(\mathbf{X}, \boldsymbol{\gamma})]$ is usually called *total variation of* $\mathbf{X}$ in this context, abbreviated by $\mathrm{totvar}(\mathbf{X})$ [Hron and Kubáček, 2011], and here it is equal to the variance of the *ilr* transformed variable Z of the composition $\mathbf{X}$,

$$\mathrm{totvar}(\mathbf{X}) = \mathrm{var}(Z) = \mathrm{var}\left(\frac{1}{\sqrt{2}} \ln \frac{X}{1-X}\right). \tag{7}$$

From its construction, the total variation represents a measure of total dispersion of $\mathbf{X}$ around $\boldsymbol{\gamma}$ and, also in the general case of a D-part composition, $D \geq 2$, its properties correspond to the standard case of a variance of a random variable. For a scalar $a \in \mathbf{R}$ and a non-random composition $\mathbf{b} \in \mathcal{S}$ it holds that

$$\mathrm{totvar}(a \odot \mathbf{X} \oplus \mathbf{b}) = a^2 \cdot \mathrm{totvar}(\mathbf{X}). \tag{8}$$

Let us consider a random sample of size n from an infinite population, $\mathbf{X}_1 = (X_1, 1 - X_1), \ldots, \mathbf{X}_n = (X_n, 1 - X_n)$, i.e. a set of independent random compositions, all of them with the same distribution as $\mathbf{X}$. It was proved in Pawlowsky-Glahn and Egozcue [2002] that the random composition

$$\mathbf{G} = \frac{1}{n} \odot (\mathbf{X}_1 \oplus \cdots \oplus \mathbf{X}_n) = \mathcal{C}\left(\sqrt[n]{\prod_{i=1}^{n} X_i},\ \sqrt[n]{\prod_{i=1}^{n} (1 - X_i)}\right), \tag{9}$$

called *sample center* and formed by geometric means of both parts, is the best linear unbiased estimator of $\boldsymbol{\gamma}$ in the sense of the Aitchison geometry. This means that this estimator is linear in the context of the Aitchison geometry. Moreover, it is unbiased, $\mathrm{cen}(\mathbf{G}) = \boldsymbol{\gamma}$, and for any other linear unbiased estimator $\mathbf{G}^*$ of $\boldsymbol{\gamma}$ we have that $\mathrm{totvar}(\boldsymbol{G}) < \mathrm{totvar}(\mathbf{G}^*)$. Both the theoretical and the sample center have an intuitive representation in the *ilr* coordinate as the mean value and the sample mean of Z, respectively. Practical computations of these characteristics are usually based on these properties. Note that the geometric mean was already used in some survey studies, e.g., for comparing price indices [McClelland and Reinsdorf, 1999], but without the theoretical background as mentioned above.

3 Finite populations of compositional data

Since we want to consider realistic scenarios in survey sampling, the case of finite populations of compositions needs to be considered. There are many practical examples of finite populations of compositional data in survey applications, see, e.g. Graf [2006a], or the introductory examples. The basic ideas of sampling are common for all types of populations, thus also for populations of compositional data. The sampling consists of selecting some

part of a population in order to make a statement for the whole population. In the basic sampling setup, the population U consists of a known, finite number N of units. Each unit is associated with a composition of a part of interest and the remaining part. This composition of each unit in the population is fixed, even if it is an unknown quantity, it is not a random one. The units in the population are identifiable and may be labeled with numbers $1, \ldots, N$.

The goal here is to determine proper population characteristics. In the previous section we came to the conclusion that each compositional data set may be characterized by center and total variation (i.e. variance of the corresponding *ilr* variable). This can be utilized also in the finite population case by defining the *finite population center* (finite population geometric mean),

$$\boldsymbol{\gamma}_N = \frac{1}{N} \odot (\mathbf{x}_1 \oplus \cdots \oplus \mathbf{x}_N) = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \mathbf{x}_i, \tag{10}$$

of all the compositions $\mathbf{x}_1 = (x_1, 1-x_1), \ldots, \mathbf{x}_N = (x_N, 1-x_N)$ in the whole population. Since the units are no longer random compositions as in the previous section, we denote them by lower-case letters in order to distinguish this case from the situation before. The population center can be equivalently computed as population mean of the *ilr* transformed units $z_1, \ldots, z_N$ by $\mu_N = 1/N \sum_{i=1}^{N} z_i$, and back-transformed using the inverse *ilr* (ilr^{-1}) transformation,

$$\boldsymbol{\gamma}_N = \mathcal{C}\left[\exp\left(\frac{\mu_N}{\sqrt{2}}\right), \exp\left(-\frac{\mu_N}{\sqrt{2}}\right)\right]. \tag{11}$$

Although here this formula was used just for back-transformation of the population mean, it holds also generally to move *ilr*-transformed data back to the simplex.

It is important to emphasize that computing the population total, as it is usual in sampling theory, would have no sense in this context, because there is no equivalent characteristic on the simplex. Moreover, because the population total variation coincides with the (usual) finite population variance,

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \mu_N)^2, \tag{12}$$

we can make all the following considerations concerning center and total variation in the *ilr* space. As a consequence, the center (as a result of the inverse *ilr* transformation of the population mean) can be used only to determine the population proportion.

Let us now briefly summarize the main approaches to the parameter estimation from finite populations, as they are described in Cochran [1977] and Thompson [2002]. We focus on the design-based inference, where designs are determined by assigning to each possible sample s the probability $\mathsf{P}(s)$ of selecting the sample. This is, in fact, the source of randomness. The probability $\mathsf{P}(s)$ is connected with the probability π_i that unit i is included in the sample, for $i = 1, \ldots, N$.

The first sampling design we want to consider is *simple random sampling*, also called simple random sampling (SRS) without replacement. The idea is to select n distinct units from the N units in the population in such a way that every possible combination of n units has the same chance to be the selected sample. This can be done by n selections, where at each step every unit of the population that has not already been selected has equal chance

of being selected. We have that $\mathsf{P}(s) = 1/\binom{N}{n}$ and $\pi_i = n/N$. Here the sample mean

$$\overline{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \tag{13}$$

is an unbiased estimator of the population mean. A back-transformation to the (unit) simplex, analogously as in (11), leads to an unbiased estimator of (10), the *sample center*. The variance of $\overline{z}$ is given by

$$\mathrm{var}(\overline{z}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}, \tag{14}$$

while its unbiased estimator is obtained by replacing N with n and μ_N by $\overline{z}$ in Equation (12).

A more general situation is proposed by *unequal probability sampling*, where different units in the population have different probabilities of being included in the sample [Cassel et al., 1977] (the above mentioned SRS represents a special case). The sampling procedure itself usually determines the different inclusion probabilities, which then have be taken into account for a reasonable estimation of population quantities. If we limit ourselves to designs without replacement, then the general Horwitz-Thompson estimator [Horvitz and Thompson, 1952] of the population mean is defined as

$$\hat{\mu}_\pi = \frac{1}{N}\sum_{i=1}^{n}\frac{z_i}{\pi_i}. \tag{15}$$

Obviously, direct computation on the simplex would lead to

$$\hat{\boldsymbol{\gamma}}_\pi = \frac{1}{N}\odot\bigoplus_{i=1}^{n}\left[\left(\frac{1}{\pi_i}\right)\odot\mathbf{x}_i\right] = \mathcal{C}\left(\sqrt[N]{\prod_{i=1}^{n}x_i^{1/\pi_i}},\ \sqrt[N]{\prod_{i=1}^{n}(1-x_i)^{1/\pi_i}}\right) \quad , \tag{16}$$

i.e. to the weighted geometric mean of both compositional parts, closed to unit constant sum constraint. It is an unbiased estimator of the finite population center $\boldsymbol{\gamma}_N$ in the sense of the Aitchison geometry, i.e. $\mathrm{cen}(\hat{\boldsymbol{\gamma}}_\pi \ominus \boldsymbol{\gamma}_N) = \mathbf{n}$, where $\mathbf{n} = \mathcal{C}(1,1)$ is neutral element on the simplex. This comes from the property

$$\mathrm{cen}(\hat{\boldsymbol{\gamma}}_\pi \ominus \boldsymbol{\gamma}_N) = ilr^{-1}(\mathsf{E}(ilr(\hat{\boldsymbol{\gamma}}_\pi \ominus \boldsymbol{\gamma}_N))) = ilr^{-1}(\mathsf{E}(\hat{\mu}_\pi - \mu_N)),$$

where the expectation is taken in sense of the sampling design [Pawlowsky-Glahn and Egozcue, 2002].

The variance of the (unbiased) estimator $\hat{\mu}_\pi$ equals (see Thompson [2002], p. 54)

$$\mathrm{var}(\hat{\mu}_\pi) = \frac{1}{N^2}\sum_{i=1}^{N}\left(\frac{1-\pi_i}{\pi_i}\right)z_i^2 + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)z_i z_j.$$

An unbiased estimator for the variance is given by

$$\widehat{\mathrm{var}}(\hat{\mu}_\pi) = \frac{1}{N^2}\sum_{i=1}^{n}\left(\frac{1-\pi_i}{\pi_i^2}\right)z_i^2 + \frac{1}{N^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\frac{z_i z_j}{\pi_{ij}},$$

where π_{ij} denotes the probability that both units i and j are included in the sample.

An approximate $(1-\alpha)100\%$ confidence interval for the parameter μ_N, based on the large-sample normal approximation for the estimator $\hat{\mu}_\pi$, is

$$(\hat{\mu}_\pi - u_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_\pi)}, \hat{\mu}_\pi + u_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_\pi)}),$$

where $u_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution. Note that for samples with small sample sizes (less than about 50) also the $(1-\alpha/2)$ quantile of the t-distribution with $n-1$ degrees of freedom can be applied instead of $u_{1-\alpha/2}$. Using the inverse ilr transformation, we arrive for the population proportion (coming from the unit-sum representation of the finite population center) at an interval (l, u), taking the first parts of the compositions $(l, 1-l) =$

$$\mathcal{C}\left(\exp\left[(\hat{\mu}_\pi - u_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_\pi)})/\sqrt{2}\right], \exp\left[-(\hat{\mu}_\pi - u_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_\pi)})/\sqrt{2}\right]\right) \tag{17}$$

and $(u, 1-u) =$

$$\mathcal{C}\left(\exp\left[(\hat{\mu}_\pi + u_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_\pi)})/\sqrt{2}\right], \exp\left[-(\hat{\mu}_\pi + u_{1-\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\mu}_\pi)})/\sqrt{2}\right]\right), \tag{18}$$

respectively, forming the upper and lower bounds of the corresponding confidence interval on the simplex. Applying the inverse ilr transformation, the confidence intervals are expressed in the Aitchison geometry on the simplex (with the corresponding Aitchison measure), thus they cannot produce difficulties even in extreme cases as described in Section 1 [Mateu-Figueras and Pawlowsky-Glahn, 2008].

A particular interesting case is the well-known *stratified sampling*. The main idea is to divide the population into subpopulations (regions or strata), and to select a sample by some design within each stratum. The principle of stratification is to partition the population in such a way that the units within each stratum are as similar as possible. Then, a stratified sample can provide greater precision than a simple random sample of the same size. The design is specially called *stratified random sampling* if the design within each stratum is simple random sampling. For example, a geographical region may be stratified into similar areas by means of some variable like habitat type, elevation or soil type. Human populations may be stratified on the basis of geographic region, city size, sex, or socioeconomic factors [Thompson, 2002]. Nevertheless, although stratified sampling is very popular in practice, it is most likely used for convenience reasons rather than for constructing homogeneous groups.

Let us assume that the sample of compositions is selected by simple random sampling from each of L strata in the population, with selections in different strata independent of each other. Suppose that within stratum U_h, $h = 1, \ldots, L$ any specified design is used to select the sample s_h of n_h units from N_h in the population; the total number of units in the population is $N = \sum_{h=1}^{L} N_h$ and the total sample size is $n = \sum_{h=1}^{L} n_h$. Further, suppose that one has an estimator $\hat{\mu}_h$ which is unbiased for the population mean μ_h in the corresponding stratum with respect to that design. Let $\mathrm{var}(\hat{\mu}_h)$ denote the variance of $\hat{\mu}_h$ and $\widehat{\mathrm{var}}(\hat{\mu}_h)$ its unbiased estimator. Then an unbiased estimator of the overall population mean μ_N is obtained by the weighted mean of the stratum estimators,

$$\hat{\mu}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_h \hat{\mu}_h. \tag{19}$$

Back-transformation of (19) to the simplex obviously results in the weighted geometric mean. Because of independence of the selections in different strata, the variance of the stratified estimator (and its unbiased estimator) is the sum of the individual stratum variances,

$$\mathrm{var}(\hat{\mu}_{st}) = \sum_{h=1}^{L} (N_h/N)^2 \mathrm{var}(\hat{\mu}_h), \quad \widehat{\mathrm{var}}(\hat{\mu}_{st}) = \sum_{h=1}^{L} (N_h/N)^2 \widehat{\mathrm{var}}(\hat{\mu}_h).$$

An important task is how to allocate the given total sample size $n = \sum_{h=1}^{L} n_h$ among the L strata. A reasonable allocation scheme estimates the population mean with the lowest variance for a fixed total sample under stratified random sampling [Neyman, 1934, so called *optimum allocation* or *Neyman allocation*],

$$n_h = \frac{n N_h \sigma_h}{\sum_{k=1}^{L} N_k \sigma_k}. \tag{20}$$

Note that the stratum population standard deviations σ_h, defined analogously as in (12) for all units in the stratum h, may in practice be estimated with sample standard deviations from past data. Applications of the Neyman basic idea to real-world problems were shown in many papers, see e.g. Sukhatme and Tang [1975], Kadane [2005]. Improvements to allow upper and lower bounds of the sample sizes within strata are given by Gabler et al. [2010].

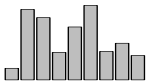
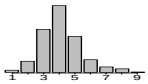
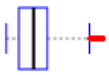
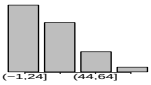
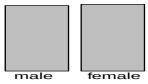
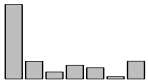
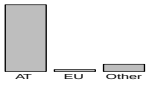
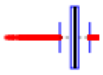
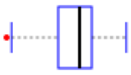
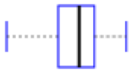
4 Numerical results

4.1 The Austrian close-to-reality EU-SILC population

This example and also the simulation study in Section 4.4 is motivated by the *European Union Statistics on Income and Living Conditions* (EU-SILC) 2006 of Austria, but is also applicable to many other survey data. EU-SILC is a complex panel survey conducted in EU member states and other European countries. It is mainly used for measuring risk-of-poverty and monitoring social cohesion in Europe [Atkinson et al., 2002] and to monitor the Europe 2020 goals. The EU-SILC data contain information about income, which is a composition of different income components. In addition, the original EU-SILC data contains more than 400 other categorical variables.

To carry out design-based simulations, population data are needed [for detailed discussion about design-based simulation studies we refer to Alfons et al., 2011a]. The procedure to generate such a close-to-reality population using the information of the sample is described in Kraft [2009] and Alfons et al. [2011c] and implemented in the R-package `simPopulation` [Kraft and Alfons, 2010]. The model-based methods therein can be considered as extensions of the data generation of Münnich et al. [2003] and Münnich and Schürle [2003]. Alfons et al. [2011c] have shown that the generated population is adequate and fulfills all the necessary requirements: the actual sizes of regions and strata are reflected; marginal distributions and interactions are considered; heterogeneities between subgroups are retained; the income components are simulated in a realistic manner; the household structure is reflected in the population.

Table 1 Variables of the synthetic EU-SILC population data from which samples are drawn with different sampling designs.

Variable	Type		Distribution	log-scale
Region	multinomial	9 levels		
Household size	ordinal scale	9 levels		
Age	continuous	from 1 – 97		
AgeCut	ordinal scale	4 levels		
Gender	binomial	2 levels		
Economic status	multinomial	7 levels		
Citizenship	multinomial	3 levels		
social	continuous/compositional			
other	continuous/compositional			
social/(social + other)	continuous/ratio			

4.2 Population characteristics

Table 1 lists the variables of the close-to-reality synthetic Austrian EU-SILC population data that are used in the following [for detailed information about the data generation, see Alfons et al., 2011c,b]. To get information about the distribution of the variables, small sparkboxes and sparkbars [Tufte, 2001] are included in this (graphical) table. It can be seen, that the variables *social* and *other*, and especially the ratio *social/(other+social)* are extremely right skewed. A detailed description of all these variables is provided in Eurostat [2004] and EU-SILC [2009].

The variable *social* (see Table 1) describes the household income from "social" transfers from family/children related allowances. The household income denoted by "other" includes

Table 2 95% confidence intervals for the arithmetic and the geometric mean.

	social/(social + other)		
estimator	CI left	mean	CI right
arithmetic mean (m())	0.444	0.518	0.591
arithmetic mean (m()/m())	0.314	0.378	0.442
geometric mean	0.493	0.626	0.742

income from rental of a property or land, housing allowances, regular inter-household cash transfer received, interest, dividends, profit from capital investments in unincorporated business, income received by people aged under 16, regular inter-household cash transfer paid and repayments/receipts for tax adjustment. For a detailed description of the variables we refer to European Commission et al. [2009]. The proportion of social income on the total property and transfer income or, equivalently, the ratio between the social income and the income from other sources (taking the geometric mean approach) may be of special interest to policy makers for comparative studies.

The problem of the presence of zero values is relevant when the geometric mean as measure of central tendency is considered. Nevertheless, when exclusively one proportion is of interest, it is rather a classification problem than a numerical one. If somebody has zero social transfers, then he/she does not belong to the group of socially supported citizens. Consequently, in practice such observations are usually excluded from the data set, see Martín-Fernández et al. [2011] for details.

4.3 Arithmetic mean versus geometric mean

We consider the subpopulation *female* from region *Burgenland*, living in single households. This subpopulation is one interesting domain defined by EU-SILC [2009]. Note that the sampling weights are the same for each observation in this domain. From this subpopulation a sample of size 50 is drawn (the population size equals 491) and confidence intervals are estimated from the ratio of the two compositional parts *social/other* and *social/(social + other)*. The ratio *social/other* stands for the amount of income from social transfers regarding to income in other areas, and *social/(social + other)* describes the income from social transfers on the whole income, often reported from statistical agencies and institutions [see, e.g., Leetmaa and Rennie, 2009].

Table 2 shows the three possible estimates of the population proportion and the corresponding conventionally calculated 95% confidence intervals. The differences between the arithmetic and geometric means are large and the confidence intervals even do not overlap. Note that the corresponding confidence interval for the population center, the geometric mean, was computed using formulas (17) and (18), i.e. considering the closure operation in order to represent the bounds in proportions (with respect to the remaining part of the composition).

Figure 2 shows the distribution (at population level) of the geometric and the arithmetic mean estimates in 246 domains (*gender* $\times$ *region* $\times$ *citizenship* $\times$ *four age classes*) for the ratio *social* to *social* + *other*. Again we observe large differences for the different estimation methods. The (closed) geometric mean estimates, given by back-transformation

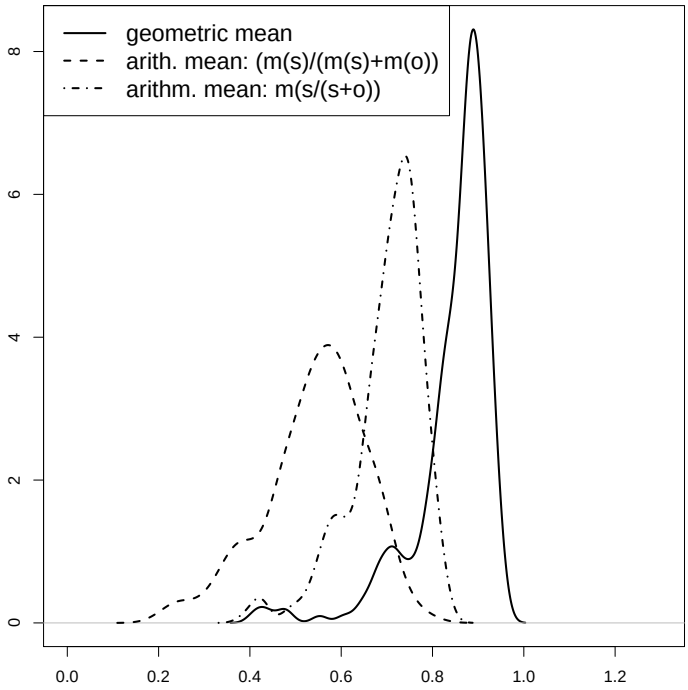


Fig. 2 Distribution of the geometric and arithmetic means in 246 domains at population level.

of (13) to the unit simplex, indicates the tightest distribution while, in comparison, especially the corresponding arithmetic mean estimates (m(s)/(m(s)+m(o))) is flat and broad. Domains that show outlying behavior (indicated by the two breaks in the left part of the distributions) are better reflected by the geometric mean estimates. Even more, it appears that m(s)/(m(s)+m(o)) is strongly affected by a few observations in the extreme right tail of the distributions for o. Also, observations on s/(s+o) that are far from 0.5 have too little influence on m(s/(s+o)). Figure 2 shows that using the geometric mean approach can give very different results from the two arithmetic mean options, so it is obvious that the choice of the estimator is an important issue. The differences in Table 2 thus did not only occur by chance, but these are systematic differences. On average, the geometric mean approach gives higher values.

As noted in the previous sections, the (closed weighted) geometric mean, or more precisely, its corresponding part should be chosen to estimate the center of the data, because it is coherent with the Aitchison geometry. The arithmetic mean can result in misleading interpretations.

4.4 Effect of the sampling design

In the context of survey sampling it is also interesting to look at the effects of sampling designs. Here we omit the standard solutions based on the arithmetic mean approach for

two reasons. First, the true values based on arithmetic and geometric means - calculated at population level - differs. It therefore makes no sense to compare arithmetic mean estimates with geometric mean estimates since the population truth is different. Secondly, we have already shown that the arithmetic mean approach is not suitable to estimate the center, and thirdly, we think that figures corresponding to arithmetic means, such as presented in Figures 3 and 4, would increase the length of the paper without much value added.

In this section, design-based simulations should show that the weighted geometric mean comes with no bias, independent of the chosen sampling design, similarly as in the case of the standard arithmetic mean approach. Therefore, the effect of the sampling design is evaluated by different settings. In Section 3 we described how to deal with sampling weights in compositional data from finite populations. In the following we compare the results with the unweighted version to see the importance of the results obtained in Section 3.

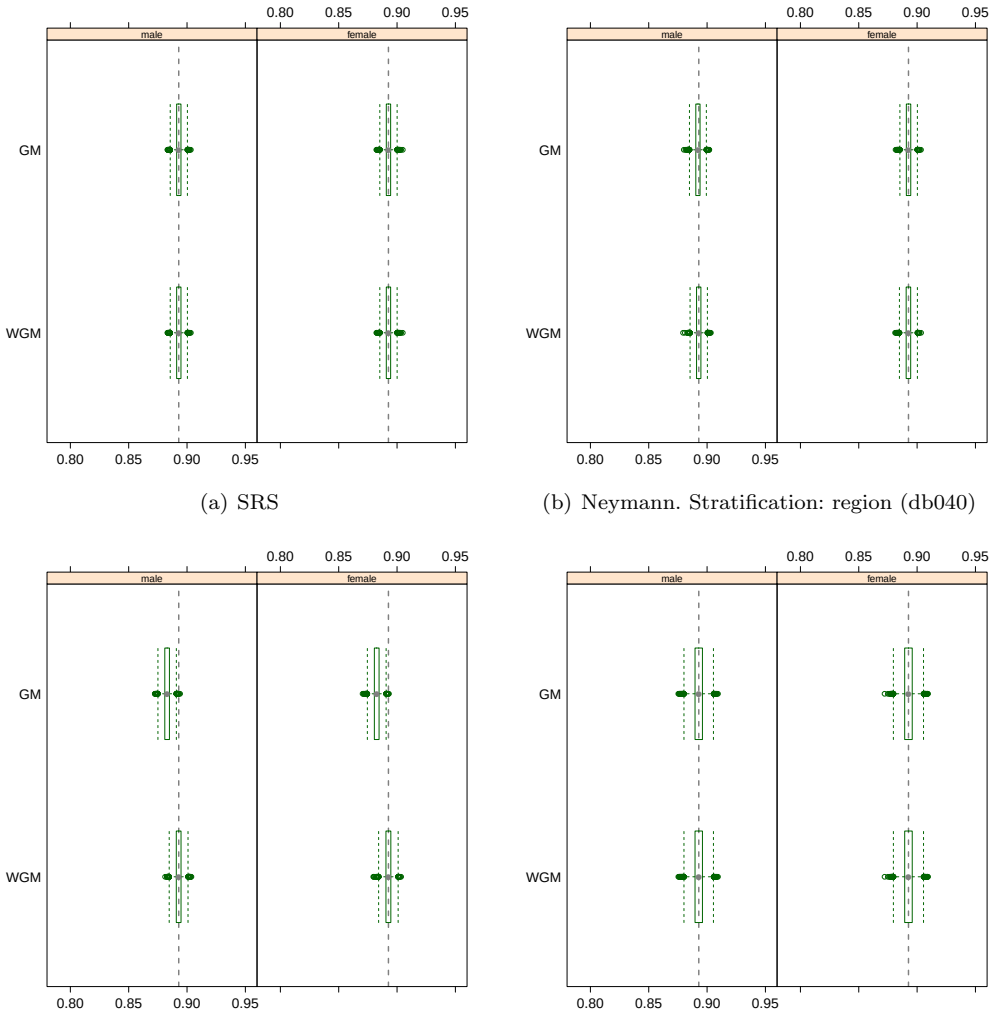
The investigation is performed in five different sampling designs which are indicated in Table 3. We also show the numbers of observations drawn from the Austrian synthetic population in each stratum in that table. All used sampling designs are very basic but often used in practice. Even oversampling is often used, for example, when the at-risk-at-poverty indicator is the most interesting indicator, then people who have income around the at-risk-at-poverty rate will be oversampled, i.e. people from large households or families with only one adult in the household. For our experiment we took an extreme case of $n_1 = 800$ and $n_2 = n_3 = \ldots = n_9 = 50$ to show an effect (since the inclusion probabilities are very different) of oversampling.

Table 3 Sample sizes in each stratum in the simulation. The strata correspond to the Austrian NUTS 2 classification and to gender × age class.

Design	B	NÖ	W	K	Stmk	OÖ	Sbg	T	V
SRS	15000								
Neyman	382	2460	2288	1323	2056	3933	840	960	757
equal size	1667	1667	1667	1667	1667	1667	1667	1667	1667
proportional	152	923	815	360	693	976	372	480	319
oversampling	800	50	50	50	50	50	50	50	50
	M 0-24	M 25-44	M 45-64	M 65+	F 0-24	F 25-44	F 45-64	F 65+	
equal size	1875	1875	1875	1875	1875	1875	1875	1875	

Figures 3 and 4 show boxplots of the estimation results obtained from estimating the (weighted) geometric mean on 10000 samples drawn from the synthetic EU-SILC population data. Various strategies are considered for drawing these samples. The estimations are carried out in the domain *gender*. In Figures 3 and 4, the true values - the geometric means calculated at population level - are plotted as grey vertical dashed lines. Naturally, using a simple SRS design, the geometric mean and the weighted geometric mean provide the same results without any bias (see Figure 3(a)). Using Neyman allocation to sample from a stratified population (Figure 3(b)), a very small bias is introduced by the unweighted version of the geometric mean. Larger bias may result with the unweighted version if other stratifications are used. Using equal size samples in each stratum (Figure 3(c)), $n_1 = n_2 = \ldots n_9 =$

1667, a large bias is introduced when estimating the geometric mean by not considering the sampling weights. This can also be seen in Figure 4(a), where other stratification variables are used. Using proportional sample sizes due to each strata (1/500), no bias is introduced (Figure 3(d)). However, if one region is oversampled ($n_1 = 800$, $n_2 = n_3, \ldots, n_9 = 50$) a large bias comes with the unweighted geometric mean estimates, shown in Figure 4(b).



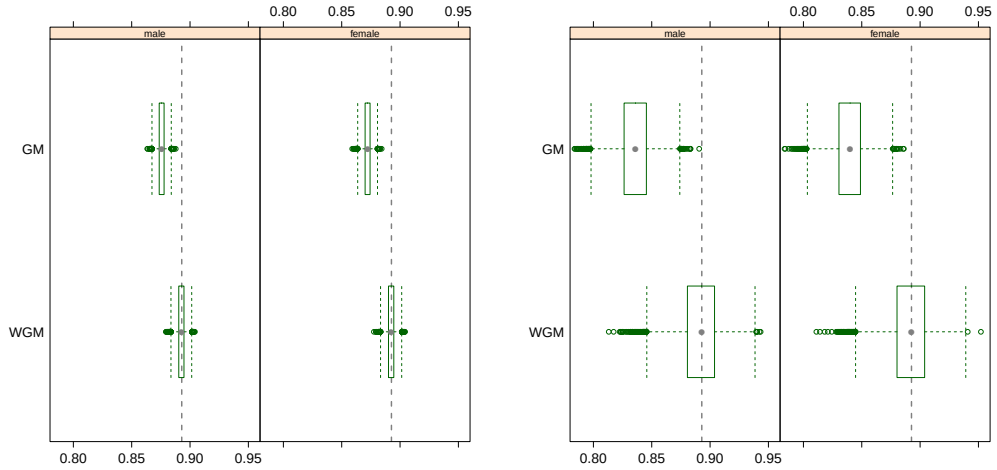
(a) SRS

(b) Neymann. Stratification: region (db040)

(c) Equal size. Stratification: region (db040)

(d) Proportional size. Stratification: region (db040)

Fig. 3 Simulation results for domains male/female from the (weighted) geometric mean concerning different sampling designs.



(a) Equal size. Stratification: age class × gender (rb090)

(b) Oversampling in one strata. Stratification: region (db040)

Fig. 4 Simulation results for domains male/female from the (weighted) geometric mean concerning equal size samples in each strata and oversampling in one region.

5 Conclusions

Data sets consisting of proportions or parts are compositional data which are represented in the so-called Aitchison geometry on the simplex and not in the usual Euclidean geometry. Thus, when mean proportions should be estimated, the classical mean is not useful because it is designed for the standard Euclidean geometry. As a consequence, the arithmetic mean does not fulfill requirements like scale invariance and subcompositional coherence, see the examples in Section 1. Instead, the geometrical mean has a solid theoretical basis in the compositional data framework. It has been demonstrated by numerical examples that the results for the geometric mean and the confidence interval around the mean are useful.

In design-based simulations we have shown that the weighted geometric mean approach comes with no bias. However, if the sampling weights are not considered, the bias may get large. Therefore, the sampling design should be considered in the estimation process.

An interesting question is how multivariate compositional data, consisting of several proportions on a whole, should be treated. The present approach in sampling theory is to consider only the single variables for univariate analyses, or to apply multivariate methods by ignoring the sampling design. An extension of the theory presented in this paper to the multivariate case of finite populations of proportions (compositions) is intended for our future research.

Acknowledgements We want to thank Prof. Anne Ruiz-Gazen (Toulouse School of Economics) for helpful suggestions. The authors would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the paper. The authors also gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

References

A. Agresti. *Categorial Data Analysis - 2nd edition.* John Wiley & Sons, New York, 2002.

J. Aitchison. *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press), 1986.

A. Alfons, J.P. Burgard, P. Filzmoser, B. Hulliger, J-P. Kolb, S. Kraft, R. Münnich, T. Schoch, and M. Templ. The AMELI simulation study. Research Project Report WP6 – D6.1, FP7-SSH-2007-217322 AMELI, 2011a. URL `http://ameli.surveystatistics.net`.

A. Alfons, P. Filzmoser, B. Hulliger, J-P. Kolb, S. Kraft, M. Münnich, and M. Templ. Synthetic data generation of SILC data. Research Project Report WP6 – D6.2, FP7-SSH-2007-217322 AMELI, 2011b. URL `http://ameli.surveystatistics.net`.

A. Alfons, S. Kraft, M. Templ, and P. Filzmoser. Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, pages 1–25, 2011c. 10.1007/s10260-011-0163-2.

T. Atkinson, B. Cantillon, E. Marlier, and B. Nolan. *Social Indicators: The EU and Social Inclusion.* Oxford University Press, New York, 2002.

C. Cassel, C. Sarndal, and H. H. Wretman. *Foundations of Inference in Survey Sampling.* Wiley, New York, 1977.

W.G. Cochran. *Sampling Techniques - 3rd edition.* John Wiley & Sons, New York, 1977.

J. J. Egozcue. Reply to "On the Harker variation diagrams..." by J. A. Cortés. *Mathematical Geosciences*, 41(7):829–834, 2009.

J.J. Egozcue and V. Pawlowsky-Glahn. Simplicial geometry for compositional data. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, editors, *Compositional Data Analysis in the Geosciences: From Theory to Practice*, volume 264 of *Special Publications*, pages 145–160. Geological Society, London, 2006.

J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003.

EU-SILC. Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC). EU-SILC LC-ILC/39/09/EN-rev.1, Directorate F: Social and information society statistics Unit F-3: Living conditions and social protection, EUROPEAN COMMISSION, EUROSTAT,, Luxembourg, 2009.

Directorate F: Social Statistics European Commission, Eurostat, Information Society Unit F-3: Living conditions, and social protection statistics. *Description of SILC user database variables: cross-sectional and longitudinal*, 2009. Version 2007.2.

Eurostat. Description of target variables: Cross-sectional and longitudinal. EU-SILC 065/04, Eurostat, Luxembourg, 2004.

P. Filzmoser, K. Hron, and C. Reimann. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment*, 407: 6100–6108, 2009.

S. Gabler, M. Ganninger, and R. Münnich. Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75(2):151–161, 2010.

M. Graf. Precision of compositional data in a stratified two-stage cluster sample: Comparison of the Swiss Earnings Structure Survey 2002 and 2004. 2006a. Joint Statistical Meeting

2006.

M. Graf. Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample: Analysis and precision assessment of wage components. Technical Report 338-0038, Swiss Federal Statistical Office, Neuchâtel, CH, 2006b.

M. Graf. Use of survey weights for the analysis of compositional data. In Pawlowsky-Glahn and Buccianti [2011], pages 114–127.

D.G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the Americal Statistical Association*, 47:663–685, 1952.

K. Hron and L. Kubáček. Statistical properties of the total variation estimator for compositional data. *Metrika*, 74(2):221–230, 2011.

J.B. Kadane. Optimal dynamic sample allocation among strata. *Journal of Official Statistics*, 21:531–541, 2005.

P. Kott and Y. Liu. One-sided coverage intervals for a proportion estimated from a stratified simple random sample. *International Statistical Review*, 77:251–265, 2009.

S. Kraft. Simulation of a population for the European Income and Living Conditions Survey. Master's thesis, Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria, 2009.

S. Kraft and A. Alfons. *`simPopulation`: Simulation of synthetic populations for surveys based on sample data*, 2010. R package version 0.1.1.

P. Leetmaa and H. Rennie. Household saving rate higher in the eu than in the usa despite lower income. household income, saving and investment, 1995-2007. Research report 29/2009, European Commission/EUROSTAT, 2009.

J. A. Martín-Fernández, J. Palarea-Albaladejo, and R. A. Olea. Dealing with zeros. In Pawlowsky-Glahn and Buccianti [2011], pages 43–58.

G. Mateu-Figueras and V. Pawlowsky-Glahn. A critical approach to probability laws in geochemistry. *Mathematical Geosciences*, 40(5):489–502, 2008.

R. McClelland and M. Reinsdorf. Small sample bias in geometric mean and seasoned CPI component indexes. Technical Report Working Paper 324, U.S. Department of Labor, Bureau of Labor Statistics, 1999. 31 p.

R. Münnich and J. Schürle. On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No. 4, University of Tübingen, 2003.

R. Münnich, J. Schürle, W. Bihler, H.-J. Boonstra, P. Knotterus, N. Nieuwenbroek, A. Haslinger, S. Laaksonen, D. Eckmair, A. Quatember, H. Wagner, J.-P. Renfer, U. Oetliker, and R. Wiegert. Monte Carlo simulation study of European surveys. DACSEIS Deliverables D3.1 and D3.2, University of Tübingen, 2003.

J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606, 1934.

V. Pawlowsky-Glahn and A. Buccianti, editors. *Compositional Data Analysis: Theory and Applications*, 2011. Wiley, Chichester.

V. Pawlowsky-Glahn and J.J. Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 15(5):384–398, 2001.

V. Pawlowsky-Glahn and J.J. Egozcue. BLU estimators and compositional data. *Mathematical Geology*, 34(3):259–274, 2002.

B.V. Sukhatme and V.K.T. Tang. Allocation in stratified sampling subsequent to preliminary test of significance. *Journal of the American Statistical Association*, 70:175–179, 1975.

S.K. Thompson. *Sampling - 2nd edition.* Wiley, New York, 2002.

E.R. Tufte. *The Visual Display of Quantitative Information - 2nd edition.* Graphics Press, Cheshire, CT, 2001.