Use of Survey Weights for the Analysis of Compositional Data: Some Simulation Results

MONIQUE GRAF

Statistical Methods Unit, Swiss Federal Statistical Office, Espace de l´Europe 10, 2010 Neuchâtel, Switzerland,
monique.graf@bfs.admin.ch

The compositional space can be seen as a vector space, where the vector addition corresponds to perturbation and the multiplication by a scalar corresponds to powering (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001). Whereas perturbation is a widely used operation in applications of compositional analysis, powering is somewhat neglected. Survey data analysis on the other hand is a domain of applied statistics where the use of weights is predominant. The reason for introducing weights in survey data analysis is threefold: 1. the use of complex survey designs with unequal inclusion probabilities, 2. the correction of non-response, and 3. calibration procedures. We shall introduce briefly the rationale for weights in survey analysis and then discuss the connection between survey weights and the powering operation. Several examples will be given.

Surveys are essentially built to optimize the estimation of totals in population subgroups for a number of variables. Practically, a key variable is chosen and the design is optimized for this variable, the trade-off being between cost and precision. Totals are estimated by weighted sums of the sampled values. The weights are extrapolation factors that depend on the survey design. It is an important aspect of the data quality to inform the user on the measurement error of the published figures. Survey design and estimation are described e.g. in Särndal, Swensson and Wretman (1992).

In a survey context, the interest is taken in totals or means across cases, but in a compositional context, totals have no meaning. So if we want to average cases, we have to go back to the original measurement scale and then make the closure operation. For the geometric mean composition on the contrary, the result is the same, whether the amounts are averaged first and then a average composition is computed, or whether the geometric mean of the compositions is computed directly and then closed.

The design-based approach does not make any assumptions on the distribution of compositions. This opens the way to parametrization by general partitions (Aitchison, 1986, section 2.7) without the drawback of ad hoc assumptions on multivariate normality (Aitchison, 1986, definition 6.7). In household expenditure surveys for instance, a hierarchy of commodities with broad categories are subdivided into more detailed goods. A general partition can follow this organization and may be a more convenient way to convey the information on the surveyed units. The joint probability distribution of transforms of this general partition is derived from the distribution of the sample inclusion indicator.

After a brief review of survey methodology, we apply the design-based principles to the estimation of compositions, of compositional transforms and of their covariance matrix on a small population. The properties of the estimators will be investigated by simulation. The talk will end with a discussion.

1 Introduction

The aim of this paper is to provide some illustrations on the variance-covariance matrix estimators proposed in Graf (2011) in the context of compositions. Only the material that is necessary to understand the simulation results will be recalled here without justification.

2 Elements of Survey Design

An authoritative reference on the design and analysis of surveys based on register information is provided by Särndal, Swensson and Wretman (1992). A succinct review is given below.

The peculiarity of data collection in finite population surveys is the existence of registers, that is of lists of statistical units (persons, households, companies), called sampling frames. A sample is selected from the frame. Sound scientific rules require that the selection should include a chance mechanism.

Probability sample Consider a finite population whose elements are listed in a frame. A probability sample S is an element taken randomly out of a set $\mathbf{S}$ of possible samples with a known probability of selection $p(.)$. The procedure defining

$$\Pr(S = s) = p(s), s \in \mathbf{S}$$

must give every element in the frame a non-zero probability of selection.

Thus the probability π_k of selection of unit k in the population is given by the probability that one of the samples containing k ($s \ni k$) is selected. (The symbol $\ni$ is used in order to emphasize the fact that s is variable and k is fixed.)

$$\pi_k = \Pr(S \ni k) = \sum_{\{s \in \mathbf{S}\,:\, s \ni k\}} p(s). \tag{1}$$

In the same way, the joint probability of selection π_{ik} of units i and k is given by the probability that one of the samples containing both i and k is selected.

$$\pi_{ik} = \sum_{\{s \in \mathbf{S}\,:\, s \ni i \text{ and } k\}} p(s) \qquad\qquad \pi_{kk} = \pi_k. \tag{2}$$

2.1 Examples of survey designs

We give below two examples of survey design.

Bernoulli design In the Bernoulli design, each unit is sampled with a fixed probability π and independently. We have

$$\pi_k = \pi_{kk} = \pi \text{ and } \pi_{ik} = \pi^2\,,\; i \neq k.$$

The sample size is random in this case with expectation $N\pi$, where N is the population size.

Fixed size designs In a fixed size design, n units out a population of size N are selected without replacement according to prescribed probabilities π_k. The first case is is the simplest one:

- If π_k is constant, we have the simple random sampling design without replacement. One easily sees that in this case $\pi_k = n/N$ and $\pi_{ik} = n(n-1)/(N(N-1))$. It is easy to set up a selection algorithm that generates a random sample according to this design.
- In the more general case of unequal selection probabilities with the constraint of a fixed size design, the selection algorithm and the computation of the second order selection probabilities are more complicated. Different sampling algorithms exist, Tillé (2006).

2.2 Design-based estimation

The purpose of the survey is the estimation of some statistic $Q(S)$, viewed as a function of the random sample S. Its design-based distribution is derived from $p(.)$.

$$\begin{aligned}
\mathrm{E}(Q(S)) &= \sum_{s \in \mathbf{S}} p(s)Q(s), \\
\mathrm{Var}(Q(S)) &= \sum_{s \in \mathbf{S}} p(s)Q^2(s) - [\mathrm{E}(Q(s))]^2, \\
\mathrm{Cov}(Q_1(S), Q_2(S)) &= \sum_{s \in \mathbf{S}} p(s)Q_1(s)Q_2(s) - \mathrm{E}(Q_1(s))\ \mathrm{E}(Q_2(s)).
\end{aligned}$$

The sample membership estimator is a fundamental example of a statistic $Q(S)$.

Sample membership indicator Let us denote as before by S a random sample with distribution $p(.)$. To each unit in the population an indicator variable of sample inclusion is attached:

$$z_k(S) = \begin{cases} 1 & \text{if } k \in S \ , \\ 0 & \text{otherwise.} \end{cases}$$

The (randomization) distribution of the binary random variables $z_k(S)$, $k = 1, ..., N$ gives the inclusion probability of element k and the joint inclusion probability of elements i and k:

$$\Pr(z_k(S) = 1) = \pi_k, \qquad \Pr(z_i(S)z_k(S) = 1) = \pi_{ik}.$$

The first and second moments of $z_k(S)$ are easily obtained. Using equations (1) and (2):

$$\begin{aligned} \mathrm{E}(z_k(S)) &= \sum_{s\in\mathbf{S}} p(s)z_k(s) = \sum_{\{s\in\mathbf{S}\,:\,s\ni k\}} p(s) = \pi_k, \\ \mathrm{Var}(z_k(S)) &= \mathrm{E}(z_k^2(S)) - \mathrm{E}(z_k(S))^2 = \pi_k - \pi_k^2, & (3) \\ \mathrm{Cov}(z_i(S), z_k(S)) &= \pi_{ik} - \pi_i\pi_k. & (4) \end{aligned}$$

Note that the sample size is given by $n_S = \sum_{k\in U} z_k(S)$ and, depending on the sampling design, it can be fixed or random.

Linear statistic Let y_k be some quantity of interest attached to individual k, for instance an age class indicator or the income. The quantity y_k is observed if k is in the sample and unknown otherwise. The aim is to estimate the total of y_k, $k \in U$, for instance the total number of persons in a given age class or the total income in the population. The parameter of interest is the sum

$$\theta = \sum_{k\in U} y_k.$$

- Horvitz-Thompson or π estimator
  The estimator
  $$\hat\theta_1(S) = \sum_{k\in U} \frac{y_k z_k(S)}{\mathrm{E}\left(z_k(S)\right)} = \sum_{k\in U} \frac{y_k z_k(S)}{\pi_k} = \sum_{k\in S} \frac{y_k}{\pi_k} \tag{5}$$
  is called the Horvitz-Thompson or π estimator.

  It is unbiased for θ:
  $$\mathrm{E}\left(\hat\theta_1(S)\right) = \sum_{k\in U} \frac{y_k \,\mathrm{E}\left(z_k(S)\right)}{\mathrm{E}\left(z_k(S)\right)} = \sum_{k\in U} y_k = \theta,$$

- Variance of $\hat\theta_1(S)$
  It is given by:
  $$\mathrm{Var}\left(\hat\theta_1(S)\right) = \sum_{i,k\in U} \frac{y_i y_k}{\pi_i\pi_k} \mathrm{Cov}\left(z_i(S), z_k(S)\right) = \sum_{i,k\in U} \frac{y_i y_k}{\pi_i\pi_k} \left(\pi_{ik} - \pi_i\pi_k\right), \tag{6}$$
  where $\mathrm{Cov}\left(z_i(S), z_k(S)\right)$ is given by equation (3) if $i = k$ and by equation (4) if $i \neq k$.

  Notice that y_k is not considered as a random variable, but rather as a multiplicative constant applied to the random sample inclusion indicator.

- Variance estimator of $\hat\theta_1(S)$
  The variance of $\hat\theta_1(S)$ in equation (6) is given by a sum over the whole population, but y_i and y_k are only known on the sample, so we need an estimator. Its Horvitz-Thompson estimator is given by
  $$\widehat{\mathrm{Var}}\left(\hat\theta_1(S)\right) = \sum_{i,k\in U} \frac{y_i y_k}{\pi_i\pi_k} \left(\pi_{ik} - \pi_i\pi_k\right) \frac{z_i(S)z_k(S)}{\pi_{ik}} = \sum_{i,k\in S} \frac{y_i y_k}{\pi_i\pi_k} \frac{\pi_{ik} - \pi_i\pi_k}{\pi_{ik}}. \tag{7}$$
  This estimator is unbiased for $\mathrm{Var}\left(\hat\theta_1(S)\right)$ in equation (6), because $\mathrm{E}\left(z_i(S)z_k(S)\right) = \pi_{ik}$.

2.3 Variance of a nonlinear statistics by the linearization method

Suppose that the statistic $Q(S)$ is twice differentiable with respect to $(z_k(S), k = 1, ..., N)$. Let us write for simplicity $z_k(S) = z_k$ and $Q(S) = Q(z_1, z_2, ..., z_N)$.
Recall that $\mathrm{E}(z_k) = \pi_k$. Let us denote the first partial derivative of Q with respect to z_k evaluated at $(\pi_1, \pi_2, ..., \pi_N)$ by $Q^{'}_k$ and expand Q up to the order 1:

$$Q(z_1, z_2, ..., z_N) \approx Q(\pi_1, \pi_2, ..., \pi_N) + \sum_{k\in U} Q^{'}_k(z_k - \pi_k).$$

Replacing $\mathrm{E}\left(Q(z_1, z_2, ..., z_N)\right)$ by its first order approximation $Q(\pi_1, \pi_2, ..., \pi_N)$, we have that the variance is evaluated by the variance of the linear statistic $\sum_{k\in U} Q^{'}_k z_k$:

$$\mathrm{Var}\left(Q(z_1, z_2, ..., z_N)\right) \approx \sum_{i,k\in U} Q^{'}_i Q^{'}_k \,\mathrm{Cov}\left(z_i(S), z_k(S)\right) = \sum_{i,k\in U} Q^{'}_i Q^{'}_k \left(\pi_{ik} - \pi_i\pi_k\right). \tag{8}$$

.

Variance estimator Let us denote by $\hat{z}_1, \hat{z}_2, \hat{z}_N$ the indicators of the actual sample and let

$$\widehat{Q^{'}}_k = \frac{\partial Q(S)}{\partial z_k}|_{\hat{z}_1,\hat{z}_2,\hat{z}_N}.$$

The linearized statistics is then

$$Q(S) - \mathrm{E}(Q(S)) \approx \sum_{k\in U} \widehat{Q^{'}}_k(z_k - \pi_k).$$

By this expansion, the estimator of the linearized variance is

$$\widehat{\mathrm{Var}}\left(Q(S)\right) \approx \sum_{i,k\in U} \widehat{Q^{'}}_i \widehat{Q^{'}}_k \left(\pi_{ik} - \pi_i\pi_k\right) \frac{z_i(S)z_k(S)}{\pi_{ik}} = \sum_{i,k\in S} \widehat{Q^{'}}_i \widehat{Q^{'}}_k \frac{\pi_{ik} - \pi_i\pi_k}{\pi_{ik}}. \tag{9}$$

Vector statistics The design-based covariance matrix estimator of a vector statistic $\mathbf{Q}(S)$ follows along the same lines:

$$\begin{aligned} \widehat{\mathbf{Q}^{'}}_k &= \frac{\partial \mathbf{Q}(S)}{\partial z_k}|_{\hat{z}_1,\hat{z}_2,\hat{z}_N} \\ \widehat{\mathrm{Var}}(\mathbf{Q}(S)) &\approx \sum_{i,k\in S} \widehat{\mathbf{Q}^{'}_i}(\widehat{\mathbf{Q}^{'}_k})^t \frac{\pi_{ik} - \pi_i\pi_k}{\pi_{ik}}. \end{aligned} \tag{10}$$

2.4 Four variance estimators for compositions

Let t_k be some total amount for unit $k \in U$ (U is a population of size N) and let $w_k(= 1/\pi_k)$ be the corresponding survey weight. The total amount t_k is distributed among D components $t_k\mathbf{p}_k, k = 1, ..., N$, where $\mathbf{p}_k$ is the corresponding D-parts composition.

The D-part composition giving the cases closed arithmetic mean of amounts and closed geometric mean estimators are respectively

$$\bar{\mathbf{p}}_{am}(S) = \frac{1}{\sum_{i\in U} w_i t_i z_i(S)} \sum_{k\in U} w_k t_k z_k(S)\,\mathbf{p}_k, \tag{11}$$

$$\bar{\mathbf{p}}_{gm}(S) = \frac{1}{\sum_{i\in U} w_i z_i(S)} \odot \prod_{k\in U} \left[w_k z_k(S)\right] \odot \mathbf{p}_k. \tag{12}$$

In the above expression, the product is computed componentwise.

Notice that only the closed geometric mean is compatible with Aitchison's geometry, see Pawlowsky-Glahn and Egozcue (2002).

Closed arithmetic mean of amounts From the design-based viewpoint, it is a ratio of sums. The design-based covariance matrix estimator $\mathbf{V}_{am}$ of $\bar{\mathbf{p}}_{am}(S)$ in equation (11) is given by equation (10), with:

$$\widehat{\mathbf{Q}'_k} = \widehat{\mathbf{Q}'_k}^{(am)} = \frac{w_k t_k \left(\mathbf{p}_k - \bar{\mathbf{p}}_{am}(S)\right)}{\sum_{i\in S} w_i t_i},\ k \in S. \tag{13}$$

Additive logratio of the geometric mean composition The transform $\mathrm{alr}(\bar{\mathbf{p}}_{gm}(S))$ is obtained from equation (11) with $\mathrm{alr}(\mathbf{p}_k)$ replacing $\mathbf{p}_k$, and w_k replacing $w_k t_k$. Thus the corresponding covariance matrix $\mathbf{V}_{\mathrm{alr}\,.gm}$ is given by equation (10) with

$$\widehat{\mathbf{Q}'_k} = \widehat{\mathbf{Q}'_k}^{(\mathrm{alr}\,.gm)} = \frac{w_k \left[\mathrm{alr}(\mathbf{p}_k) - \mathrm{alr}(\bar{\mathbf{p}}_{gm}(S))\right]}{\sum_{i\in S} w_i}. \tag{14}$$

The covariance matrix of other transforms are obtained along the same lines.

Closed geometric mean composition The covariance matrix of $\bar{\mathbf{p}}_{gm}$, denoted by $\mathbf{V}_{pgm}$, is given by equation (10) with:

$$\widehat{\mathbf{Q}'_k} = \widehat{\mathbf{Q}'_k}^{(gm)} = \frac{\partial}{\partial z_k}\bar{\mathbf{p}}_{gm}(S) = \left(\mathrm{diag}(\bar{\mathbf{p}}_{gm}(S)) - \bar{\mathbf{p}}_{gm}(S)\bar{\mathbf{p}}_{gm}(S)^t\right) \begin{bmatrix} \mathbf{I}_{D-1} \\ \mathbf{0}^t_{D-1} \end{bmatrix} \widehat{\mathbf{Q}'_k}^{(\mathrm{alr}\,.gm)}. \tag{15}$$

Additive logratio of the arithmetic mean composition Linearizing $\mathrm{alr}(\bar{\mathbf{p}}_{am}(S))$, we obtain :

$$\widehat{\mathbf{Q}'_k} = \widehat{\mathbf{Q}'_k}^{(\mathrm{alr}\,.am)} = \mathrm{diag}\left(\frac{\bar{\mathbf{p}}_{am,-D}}{\bar{p}_{am,D}}\right)^{-1} \frac{w_k t_{kD}(\mathbf{p}_{\mathbf{k},-\mathbf{D}}/p_{kD} - \bar{\mathbf{p}}_{am,-D}/\bar{p}_{am,D})}{\sum_i w_i t_{iD}}. \tag{16}$$

These variance-covariance matrix estimators will be used to derive confidence domains for the mean compositions $\bar{\mathbf{p}}_{am}$ and $\bar{\mathbf{p}}_{gm}$.

Confidence domains Let us denote by $\bar{\mathbf{p}}$ either $\bar{\mathbf{p}}_{am}$ or $\bar{\mathbf{p}}_{gm}$, by $\mathbf{V}$ the corresponding covariance matrix in Euclidian geometry, and by $\mathbf{V}_{\mathrm{alr}}$ the covariance matrix of $\mathrm{alr}(\bar{\mathbf{p}})$. We obtain confidence domains in Aitchison's geometry by back-transformation of the confidence domains computed for the alr transforms. These domains can be compared with those obtained in Euclidian geometry. Let $d = D-1$.

If the sample size is large enough, the critical value at the $(1-\alpha)$ confidence level is $\chi^2_{d;1-\alpha}$, the $(1-\alpha)$ quantile of the chi-square distribution with d degrees of freedom. Then we have:

1. the confidence domain for $\mathbf{p}$ in Euclidian geometry is limited by a d dimensional ellipsoid, because $\mathbf{V}$ is of rank d. Let $\mathbf{U}$ be a $D\times d$ matrix, such that $\mathbf{U^t}\,\mathbf{V}\,\mathbf{U}$ is of full rank. The confidence domain for $\mathbf{p}$ is given by

$$\left\{\mathbf{p}\in\mathbb{R}^D \mid (\mathbf{p}-\bar{\mathbf{p}})^t\,\mathbf{U}^t\left(\mathbf{U}^t\mathbf{V}\mathbf{U}\right)^{-1}\mathbf{U}\,(\mathbf{p}-\bar{\mathbf{p}}) \le \chi^2_{d;1-\alpha}\right\} \tag{17}$$

2. the confidence domain for $\mathrm{alr}(\mathbf{p})$ is

$$\left\{\mathrm{alr}(\mathbf{p})\in\mathbb{R}^d \mid (\mathrm{alr}(\mathbf{p})-\mathrm{alr}(\bar{\mathbf{p}}))'\,\mathbf{V}^{-1}_{\mathrm{alr}}\,(\mathrm{alr}(\mathbf{p})-\mathrm{alr}(\bar{\mathbf{p}})) \le \chi^2_{d;1-\alpha}\right\} \tag{18}$$

3. the corresponding domain for $\mathbf{p}$ in Aitchison's geometry is a subset of the simplex S^D:

$$\left\{\mathbf{p}\in S^D \mid (\mathrm{alr}(\mathbf{p})-\mathrm{alr}(\bar{\mathbf{p}}))'\,\mathbf{V}^{-1}_{\mathrm{alr}}\,(\mathrm{alr}(\mathbf{p})-\mathrm{alr}(\bar{\mathbf{p}})) \le \chi^2_{d;1-\alpha}\right\} \tag{19}$$

Domains 2. and 3. were used in Graf (2006).

For small sample size n, different options are available for the critical value, see Rao, Scott and Skinner (1998). If we restrict to one-stage designs without stratification, as in the examples below, the critical value $\chi^2_{d;1-\alpha}$ is replaced by Hotelling's T^2 namely

$$\frac{(n-1)d}{n-d}\,F_{d,n-d;1-\alpha},$$

where $F_{d,n-d;1-\alpha}$ is the upper α quantile of the F distribution, according to the classical theory (because in this case the nominal and actual degrees of freedom are equal). Equivalently, one can compute the actual coverage $1-\tilde{\alpha}$ of the domains in equations (17) to (19). With $F_{d,n-d}$ denoting the cdf, it is given by

$$1-\tilde{\alpha} = F_{d,n-d}\left(\frac{n-d}{(n-1)d}\,\chi^2_{d;1-\alpha}\right). \tag{20}$$

3 Simulation example

In order to see how a complex survey design enters into play, we consider a small population. The following example utilizes the dataset ”election” provided by Lumley (2010). This dataset contains voting data from the US 2004 presidential election. It provides the number of votes per ”precinct” (state or county, depending on data availability) for the 3 candidates: George W. Bush, John Kerry and Ralf Nader. The total number of precints is $N = 4600$.

Lumley's sampling design is of fixed size (number of sampled precincts is $n = 40$). Precinct k, $k = 1, ..., N$, is sampled with probability π_k proportional to the total number of votes T_k. Thus we have:

$$\pi_k = 40\,T_k \,/ \sum_{i=1,...,4600} T_i$$

The idea behind that choice of inclusion probabilities is to give the same weight to each vote at the individual voter's level. In Lumley's example a sample is drawn using Tillé's splitting method, implemented in the ”sampling” package by Matei and Tillé (2009). This sampling design provides an exact algorithm for the computation of the second order selection probabilities π_{ik} in equation (2). Thus variance-covariance matrices of the type given in equation (10) can be computed without difficulty. Note that the knowledge of the first order selection probabilities does not determine the sampling design uniquely. Other selection algorithms will provide other second order inclusion probabilities with the same π_k, see Tillé (2006) for details.

3.1 Datasets

Whereas the minimum number of votes is 2 for Bush, and 1 for Kerry, Nader had no vote at all in 1778 precincts.

Transformed data To avoid problems with zero parts, the dataset is slightly transformed: if, in a specific precinct, the votes for Nader are less than 1% of the total votes, he is given that 1%. The votes of the best elected candidate in the precinct is changed accordingly, so that the number of votes per precinct remains constant. Figure 1 shows two boxplots per candidates: the number of votes per precinct in the original data (left) and in the transformed data (right). One can see that the parts for Bush and Kerry are hardly modified. As can be observed from the boxplot for the total of the votes, the size of the precinct varies widely, and consequently the inclusion probabilities.

Artificial data In order to begin with a less extreme case for compositional data analysis, an artificial dataset was created:

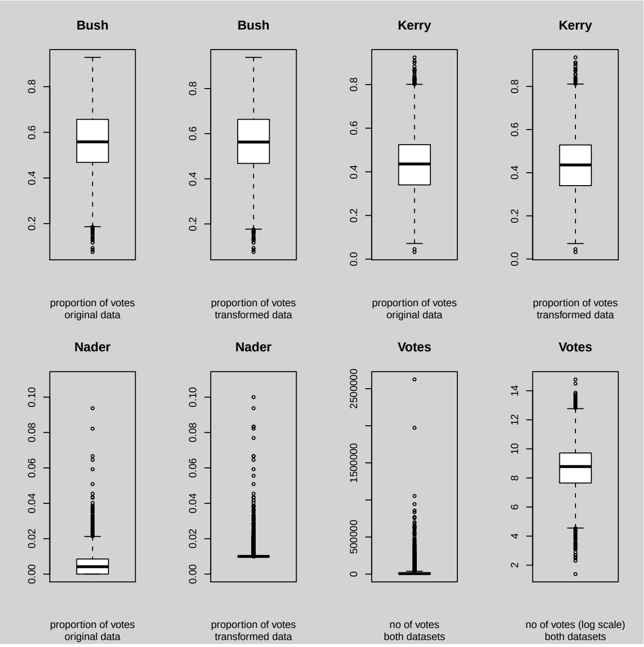

Figure 1: 2004 US presidential elections: original and transformed data

Let us denote by `Bush` the number of votes for Bush in the original dataset (similarly for the other candidates). Set

$$\begin{aligned} B &= \log(\texttt{Bush}+10) \\ K &= \log(\texttt{Kerry}+10) \\ N &= \log(\texttt{Nader}+10) \\ T &= B+K+N \end{aligned}$$

The sampling design for the artificial dataset is Tillé's splitting method again, but with π_k proportional to the new T_k. Figure 2 shows the boxlots of B, K, N and T (above), and $B/T, K/T$ and N/T (below).

3.2 Simulation setup

Tillé's sampling design is applied in 3 cases

1. Artificial data, sample size $n = 40$;
2. Artificial data, sample size $n = 10$;
3. Transformed data, sample size $n = 40$.

1000 simulations were performed in each case. In each simulation,

- the arithmetic mean, the geometric mean, their alr transforms and the four corresponding covariance matrices derived in section 2.4 are computed.
- confidence domains of the linearized estimators are computed in all four cases, using equation (17) for "pam" and "pgm" and equation (18) for "alrpam" and "alrpgm".
- a binary variable recording whether the true value (obtained from the complete dataset) is within the confidence domain is computed.

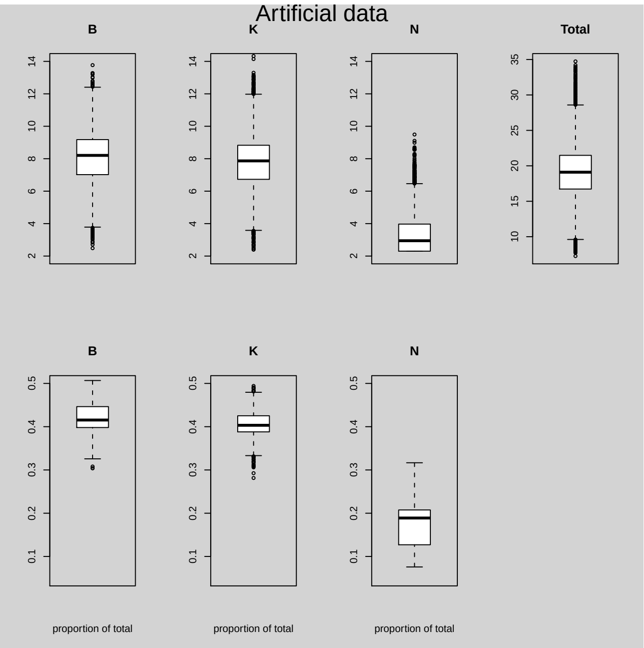

Figure 2: Artificial data:
B=log(Bush+10), K=log(Kerry+10), N=log(Nader+10), Total=B+K+N.

3.3 Results and discussion

The confidence domains in the alr space are back-transformed to the simplex, using equation (19), to ease the interpretation. Note that exactly the same domains would have been obtained from the linearization based on another transform.

Examples of the confidence domains in the simplex For case 1 (artificial data, sample size 40), the confidence domains based on one random sample are plotted in figure 3. We see that

- the arithmetic and geometric mean estimates are near;
- the confidence domains are similar, whether they are computed in the Euclidian or Aitchison's geometry.

On the contrary, in case 3 (transformed data, sample size 40), where the estimates are near the boundary of the simplex, the confidence domains may vary widely , see figure 4. These confidence domains have different properties, namely:

- We cannot be certain that the confidence domains under the heading "direct linearization", that is in Euclidian geometry, are entirely within the simplex.
- The back-transformed alr domains under the heading "linearization in Aitchison's geometry" are interior regions of the simplex by definition.
- In the centred representation in figure 5, domains in Euclidian geometry are simply translated; but domains in Aitchison's geometry change their shape.

Coverage The above procedure, where the critical value for the confidence domain is determined from the chi-square distribution with 2 degrees of freedom: $\chi^2_2(0.95) = 5.99$, is strictly valid when the 3 following conditions are met:

- the covariance matrix is exact, not estimated;

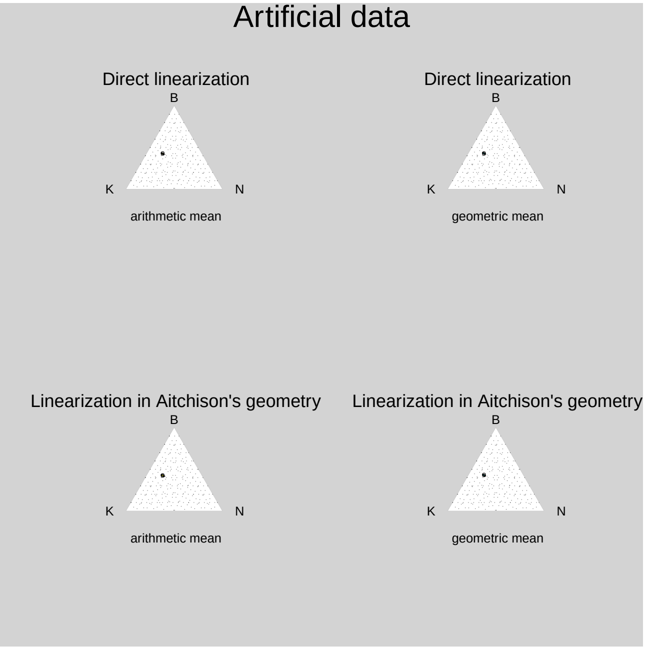

Figure 3: Artificial data:the four confidence domains for a simulated sample of size 40.

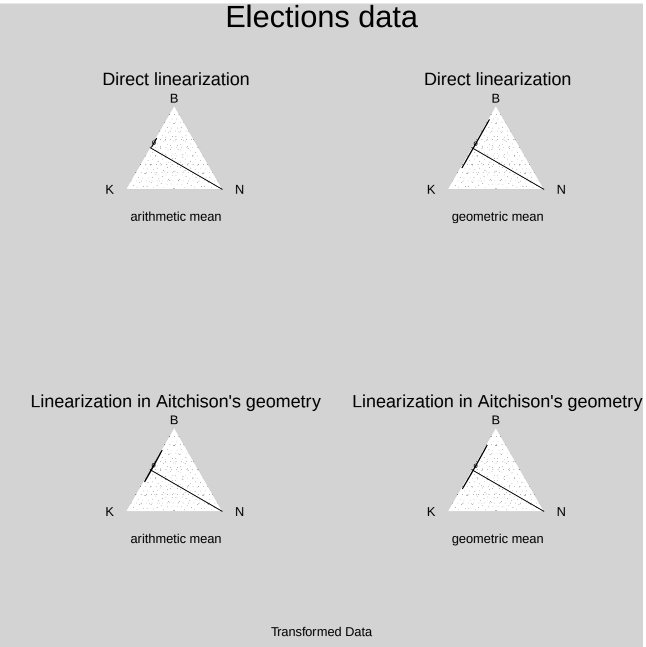

Figure 4: Transformed data: the four confidence domains for a simulated sample of size 40.

datasets	cov.matrix	n	pam	pgm	alrpam	alrpgm	nominal
artificial	estimate	40	0.941	0.937	0.972	0.936	0.934
artificial	mean	40	0.960	0.962	0.986	0.964	0.950
artificial	estimate	10	0.867	0.865	0.892	0.867	0.870
artificial	mean	10	0.956	0.952	0.981	0.955	0.950
transformed	estimate	40	0.793	0.290	0.709	0.279	0.934
transformed	mean	40	0.952	0.856	1.000	0.843	0.950

Table 1: Coverage results (1000 simulations). Nominal confidence level is 0.95 for ”mean covariance matrix”, and is given by equation (20) for ”covariance matrix estimate”.

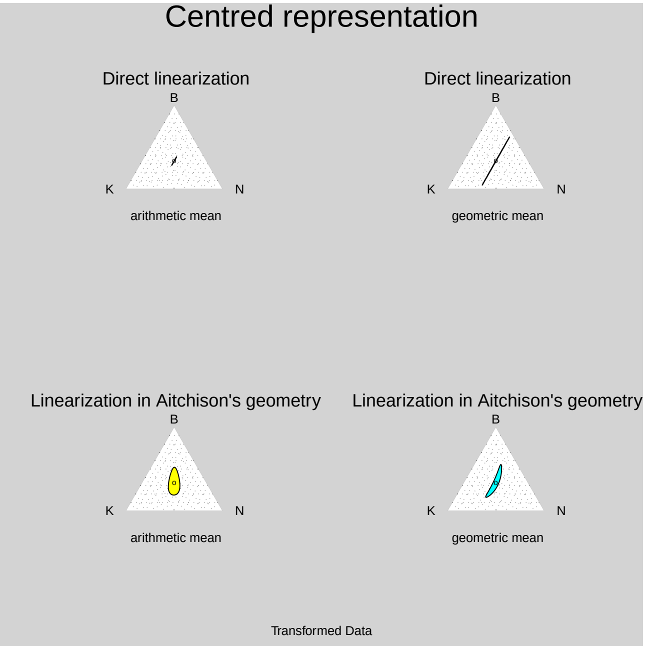

Figure 5: The four centred confidence domains corresponding to figure 4.

- the sample size is large enough for the normal approximation of the linearized estimators to be valid;

- for the direct linearization, the confidence domain is entirely within the simplex.

Let us consider the binary variable indicating whether the confidence domain centred on the estimate contains the true value. In order to determine the importance of the first condition, for each iteration the binary variable is computed using two different covariance matrices: 1. the covariance matrix estimate at each iteration, 2. the average covariance matrix over the 1000 iterations. Table 1 shows the coverage results for the 1000 simulations of the confidence domains defined by the critical value of the χ^2 distribution with 2 degrees of freedom (equations 17 and 18). First column shows the dataset; column 2 "cov.matrix" indicates the type of covariance matrix; the sample size n is in column 3; the average value of the binary variable over the 1000 simulations is given in column 4 (pam) for the closed arithmetic mean of amount (equation 11), in column 5 (pgm) for the closed geometric mean (equation 12), in column 6 (alrpam) and 7 (alrpgm) for the corresponding estimates in the alr space. In the last column, the nominal coverage is 0.95, when the estimated covariance matrix is replaced by the average over the 1000 simulation, and is given by equation (20) when the estimated covariance matrix at each run is used. We see that

- Artificial data, sample size 40: the coverage rates are reasonably close to the nominal 0.95 (under the chi-square distribution). We observe a slight undercoverage that is explained by the nominal coverage based from the F distribution, except for "alrpam", where the procedure is slightly conservative. When the covariance matrix is fixed at the mean over the 1000 simulations, the coverage rates increase slightly.

- Artificial data, sample size 10: when the covariance matrix is estimated at each iteration, results are close to the nominal F-based coverage. When this matrix is fixed at the mean over the 1000 simulations, the coverage is near to 0.95, as expected.

- Transformed data, sample size 40: undercoverage is present in all four domains, but is much more severe for the geometric mean whether in Euclidian or in Aitchison's geometry. This may be due to the lack of normality in this case. When the mean covariance matrix is used instead of the estimated one, the coverage improves dramatically, but only the "pam" coverage is correct. The overcoverage of "alrpam" contrasts with the undercoverage of "alrpgm".

Conclusions The above simulation results show that when the composition approaches the boundary of the simplex, the linear approximation of the logarithm becomes poor and explains partly the undercoverage of the confidence domains of the closed geometric mean composition. The fact that in Euclidian geometry the domains can lie partly outside the simplex does not seem to be predominant. In all cases, the best results in terms of coverage were obtained for the closed arithmetic mean of amounts in Euclidian geometry. In view of this study, the linearization method to obtain confidence domains for compositions can be recommended, when the compositions are not too close to the boundary of the simplex, especially for small sample size.

References

Aitchison, J. (1986). *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). 416 p.

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

Graf, M. (2011) Use of Survey Weights for the Analysis of Compositional Data. In Pawlowsky-Glahn, V. and Buccianti, A. (eds) (2011) *Compositional Data Analysis: Theory and Applications*, Chapter 9, Wiley.

Lumley, T., survey: Analysis of complex survey samples, R package version 3.21.

Matei, A. and Y. Tillé, sampling: Survey sampling, R package version 2.3.

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

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

Rao, J.N.K, Scott, A. J. and C. J. Skinner (1998). Quasi-Score Tests with Survey Data *Statistica Sinica* 8,1059-1070.

Särndal, C. E. and Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer Series in Statistics.

Tillé, Y. (2006), *Sampling Algorithms.* Springer Series in Statistics.