DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 1/97

Simultaneous inference for proportions in arbitrary sampling designs

by

Andreas Stich*

June 1997

Abstract

In this paper simultaneous confidence intervals for proportions in arbitrary sampling schemes are constructed. To accomplish this the asymptotic joint distribution of proportions is derived in arbitrary sampling designs. An application to household proportions in Germany in 1993 and Monte Carlo simulations are given. It turns out that conventional estimators fail in some sampling designs while the confidence intervals taking into account the sampling design behave well in all instances.

*Seminar für Wirtschafts– und Sozialstatistik, Universität zu Köln, Albertus–Magnus–Platz, 50923 Köln, Deutschland; Tel: +49/221/470 2809, Fax: +49/221/470 5074, e–mail: stich@wiso.uni-koeln.de

1 Introduction

A widely used tool for displaying statistical data is the frequency distribution. Studies on income, poverty or wealth almost always define income (or wealth) classes and report how large a proportion of the population falls into these classes. Usually, estimates of the population proportions are derived under the assumption that samples are drawn by simple random sampling. Mukhopadhyay and Chattopadhyay (1993) give sampling designs under which spherical confidence regions for the estimated proportions can be determined. However, this approach is only suited for special sampling designs. If different sampling designs are used, this method does not work any longer. Latorre (1993, 1995) derives estimates when the sampling design is stratified random sampling or two-stage sampling. In this paper Latorre's approach is generalized to arbitrary sampling designs. The goal of this paper is to estimate population proportions fully exploiting the information which is given in the sampling design.

Let the income variable be denoted by X with continuous cumulative distribution function F. Define K income classes

$$[0, a_1[, [a_1, a_2[, \ldots, [a_{i-1}, a_i[, \ldots, [a_{K-1}, \infty[,$$

The vector $p(F) = (p_1(F), \ldots, p_K(F))$ with $p_i(F) = \int_{a_{i-1}}^{a_i} dF(x)$ gives the proportion of individuals in each class. Let $p(F_n)$ be the corresponding vector of sample proportions of the observations $p(F_n) = (p_1(F_n), \ldots, p_K(F_n))$ where

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$$

is the empirical distribution function. In infinite populations the vector of proportions $p(F)$ is to be estimated.

In finite populations let N be the number of elements belonging to the population, and let $N_1, \ldots, N_K$ be the sizes of the income classes. In this case the vector of proportions is given by

$$p(F_N) = \left(\frac{N_1}{N}, \ldots, \frac{N_K}{N} \right) = (p_1(F_N), \ldots, p_K(F_N)).$$

This paper establishes simultaneous confidence intervals for $p(F_N)$ in this setup. The paper is organized as follows. The next section presents an auxiliary model. Section 3 derives the asymptotic distribution of the estimators. Section 4 deals with simultaneous confidence intervals, section 5 gives an empirical application, and section 6 presents Monte Carlo simulations. The last section concludes.

2 The auxiliary model

Before introducing the auxiliary model, some definitions are stated. A finite population is denoted by a set U and the elements are identified by labels from 1 to N, i.e., a finite population is a set $U = \{1, \ldots, N\}$. A sample ω is a subset of U.

A sampling experiment will create a sample ω according to a probability distribution P. This probability distribution $\{P(\omega)|\omega \subset U\}$ is called sampling design.

The inclusion probability of first order of unit i is defined as

$$\pi_i = P(i \in \omega) = \sum_{\{\omega \subset U | i \in \omega\}} P(\omega).$$

Usually in sampling theory the sample ω is obtained according to a specified sampling design from a finite population U and the stochastic element in this procedure is the randomization of the sample $\omega \in U$. In this paper a superpopulation model will be used. The finite population vector $x_N = (x_{1N}, \ldots, x_{NN})$ is assumed to be the realized outcome of a vector random variable $X_N = (X_{1N}, \ldots, X_{NN})$ with cumulative distribution function $\mathfrak{F}_N$. If $X_{1N}, \ldots, X_{NN}$ are i.i.d. with continuous cumulative distribution function F_X it yields $\mathfrak{F}_N = F_X^N$. More details on superpopulation models can be found in Cassel et al. (1977). In superpopulation models the sample ω is assumed to be fix, i.e., the sample ω from a finite population U is interpreted as follows: The subset ω of labels from U and the corresponding units in the finite population are fixed. The stochastic element in this model is the creation of the finite population vector x_N.

In the following an extension of the superpopulation model, the auxiliary model, introduced by Nygård and Sandström is used (see Sandström (1983), Nygård and Sandström (1985a, 1985b, 1989)). In addition to the above assumptions a vector of weights associated with the sample and regarded as deterministic, is considered. Statistical inference in this model is in coincidence with usual statistical inference. X_N is a sequence of r.v. for which statements on distribution properties are made.

Let $T(.)$ be a stochastic functional. Notice that $T(F)$ is a parameter and $T(F_N)$ a stochastic variate. Below, asymptotic results for a statistic of the form $\sqrt{n}(T(F_n) - T(F_N))$ will be computed. Because both F_n and F_N are random variables confidence statements are of the form considered in Royall (1971).

Consider a sequence of populations $U_t = \{1, 2, \ldots, N_t\}$ such that $N_t \to \infty$ with $t \to \infty$. For fixed t denote the sample by ω_t with sample size n_t and assume that $n_t \to \infty$ so that the sample fraction $f_t = n_t/N_t \to f$, $0 < f < 1$ with $t \to \infty$. When t increases, new subsets of U_t are chosen such that ω_t is not necessarily a subset of ω_{t+1}.

Definition 2.1
Let w_{it} be bounded deterministic weights, $\forall t, \forall i \in U_t$ and

$$\bar{w}_t = \frac{1}{n_t} \sum_{i \in \omega_t} w_{it} \neq 0$$

the mean of the weights. The weighted empirical distribution function (w.e.d.f.) is given by

$$F_{n_t}^*(x) = \frac{1}{n_t} \sum_{i \in \omega_t} \frac{w_{it}}{\bar{w}_t} \mathbb{1}_{X_i \leq x}. \tag{2.1}$$

Assumption 2.1
Let w_{it} be defined as above. Assume that

$$\max_{i\in\omega_t}\left(\frac{w_{it}}{\bar{w}_t}\right)^2 \leq d^2 < \infty \quad \forall t.$$

If the weights are equal to some positive constant, i.e., $w_{it} = c > 0$, $\forall i \in \omega_t$, then $F^*_{n_t}$ is identical to the "ordinary" empirical distribution function F_{n_t}. If $w_{it} = \pi_{it}^{-1}$ with π_{it} being the known inclusion probabilities, $F^*_{n_t}$ coincides with the Horvitz–Thompson estimator of the finite population c.d.f. F_{N_t}. Furthermore, choosing w_{it} as some positive constant and substituting n_t and ω_t by N_t and U_t, respectively, the w.e.d.f. is the finite population c.d.f. F_{N_t}.

Notice that the empirical distribution function can be written as:

$$F_{n_t}(x) \quad = \quad \frac{1}{n_t}\sum_{i\in\omega_t} \mathbb{1}_{X_i\leq x} = \frac{1}{n_t}\sum_{i\in\omega_t}\delta_{X_i}$$

with

$$\delta_{X_i}(x) = \begin{cases} 0 & X_i > x \\ \\ 1 & X_i \leq x \end{cases}$$

the Dirac-function, i.e., the one point distribution with mass 1 at point X_i.

3 The asymptotic distribution

Let

$$p(F^*_{n_t}) = (p_1(F^*_{n_t}), \ldots, p_K(F^*_{n_t}))$$

with

$$p_i(F^*_{n_t}) = F^*_{n_t}(a_i) - F^*_{n_t}(a_{i-1})$$

be the estimator for the vector of proportions.

Lemma 3.1
(Asymptotic distribution of w.e.d.f.) Under assumption 2.1

(i) $\sqrt{\frac{n_t}{1+v_t^2}}(F^*_{n_t} - F) \overset{asy}{\sim} \mathbb{B}$

(ii) $\sqrt{\frac{n_t}{1+v_t^2-f_t}}(F^*_{n_t} - F_{N_t}) \overset{asy}{\sim} \mathbb{B}$

with v_t^2 the squared coefficient of variation of the weights

$$v_t^2 = \frac{n_t^{-1}\sum\limits_{i\in\omega_t}(w_{it} - \bar{w}_t)^2}{\bar{w}_t^2}$$

and $\mathbb{B}$ a Brownian bridge, i. e., the distribution of $\mathbb{B}(F(x))$ is normal with $E(\mathbb{B}(F(x))) = 0$, $E(\mathbb{B}(F(x)), \mathbb{B}(F(y))) = \min(F(x), F(y)) - F(x)F(y)$ and $\mathbb{B}(0) = \mathbb{B}(1) \equiv 1$.

Proof:

(i):
The proof of (i) can be found in the proof of lemma 5.2 in Sandström (1983). Koul (1970) proofs the result for independent but not necessarily identical r.v.'s.

(ii):
Because of the uniform distribution of $F(X)$ over the interval $[0,1]$ for every r.v. the proof is done for a uniformly distributed r.v.'s. The result can be used for arbitrary, continuous r.v.'s by inserting $F(x)$ in the formulae (see Shorack and Wellner (1986), p. 99). Using chapter 3.3 of Shorack and Wellner (1986) shows the result of the lemma. At first the interesting expression can be rewritten as

$$\begin{aligned}
&F^*_{n_t} - F_{N_t} \\
&\quad = \quad F^*_{n_t} - F - (F_{N_t} - F) \\
&\quad = \quad \frac{1}{n_t}\sum_{i\in\omega_t}\frac{w_{it}}{\bar{w}_t}(\delta_{X_i} - F) - \frac{1}{N_t}\sum_{i\in U_t}(\delta_{X_i} - F) \\
&\quad = \quad \frac{1}{n_t}\sum_{i\in\omega_t}\frac{w_{it}}{\bar{w}_t}(\delta_{X_i} - F) + \frac{1}{n_t}\sum_{i\in\omega_t}(-f_t)(\delta_{X_i} - F) - \frac{1}{N_t}\sum_{i\in U_t\setminus\omega_t}(\delta_{X_i} - F) \\
&\quad = \quad \frac{1}{n_t}\sum_{i\in\omega_t}\left[\frac{w_{it}}{\bar{w}_t} - f_t\right](\delta_{X_i} - F) + \frac{1}{n_t}\sum_{i\in U_t\setminus\omega_t}(-f_t)(\delta_{X_i} - F) \\
&\quad = \quad \frac{1}{n_t}\sum_{i\in U_t} r_{it}(\delta_{X_i} - F)
\end{aligned}$$

with

$$r_{it} = \begin{cases} \frac{w_{it}}{\bar{w}_t} - f_t & i \in \omega_t \\ \\ -f_t & i \in U_t\setminus\omega_t \end{cases} . \tag{3.1}$$

This means

$$\sqrt{\frac{n_t}{1+v_t^2-f_t}}(F^*_{n_t} - F_{N_t}) = \frac{1}{\sqrt{\sum_{i\in U_t} r_{it}^2}}\sum_{i\in U_t} r_{it}(\delta_{X_i} - F), \tag{3.2}$$

because of

$$\begin{aligned}
\sum_{i\in U_t} r_{it}^2 &= \sum_{i\in\omega_t}\left(\frac{w_{it}}{\bar{w}_t} - f_t\right)^2 + \sum_{i\in U_t\setminus\omega_t} f_t^2 \\
&= n_t(1+v_t^2) + n_t f_t - 2f_t\sum_{i\in\omega_t}\frac{w_{it}}{\bar{w}_t}
\end{aligned}$$

$$
\begin{aligned}
&= n_t(1+v_t^2) + n_t f_t - 2n_t f_t \frac{n_t^{-1}\sum_{i\in\omega_t} w_{it}}{\bar{w}_t} \\
&= n_t(1+v_t^2) + n_t f_t - 2n_t f_t \\
&= n_t(1+v_t^2 - f_t).
\end{aligned}
$$

Furthermore

$$
\max_{i\in U_t} \frac{r_{it}^2}{n_t(1+v_t^2-f_t)} = \frac{\max_{i\in U_t}\left[\max\left[\left(\frac{w_{it}}{\bar{w}_t} - f_t\right)^2, f_t^2\right]\right]}{n_t(1+v_t^2-f_t)} \overset{t\to\infty}{\longrightarrow} 0.
$$

This is true because of assumption 2.1 and $f_t \to f$.

Now look at the empirical process

$$
\mathbb{B}_t(s) = \frac{1}{\sqrt{\sum_{i\in U_t} r_{it}^2}} \sum_{i\in U_t} r_{it}(\mathbb{1}_{\xi_i \le G_{it}(s)} - G_{it}(s)) \quad 0 \le s \le 1 \tag{3.3}
$$

for arbitrary cumulative distribution functions G_{it} in $[0,1]$ and uniformly distributed r.v.'s ξ_i If $\xi_i = F(X)$ then G_{it} are cumulative distribution functions of a uniform distribution over $[0,1]$ Thus $\max_{i\in U_t} ||G_{it} - I|| \to 0$ for $t \to \infty$ with I the identity in $[0,1]$. From this follows that all conditions of corollary 1 in Shorack and Wellner (1986) p. 109 are fulfilled and hence $\mathbb{B}_t$ converges to a Brownian Bridge. Because (3.3) with $s = F(x)$ equals (3.2) the statement is proved. □

Lemma 3.2
Under assumption 2.2.1

(i)

$$
\sqrt{\frac{n_t}{1+v_t^2}}[p(F_{n_t}^*) - p(F)] \overset{asy}{\sim} N(0, \Sigma_p)
$$

with

$$
\Sigma_p = \begin{pmatrix}
p_1(F)(1-p_1(F)) & -p_1(F)p_2(F) & \cdots & -p_1(F)p_K(F) \\
-p_2(F)p_1(F) & \ddots & & \vdots \\
\vdots & & \ddots & \vdots \\
-p_K(F)p_1(F) & \cdots & \cdots & p_K(F)(1-p_K(F))
\end{pmatrix}. \tag{3.4}
$$

(ii) converges additionally $f_t \to f$ for $t \to \infty$

$$
\sqrt{\frac{n_t}{1+v_t^2-f_t}}[p(F_{n_t}^*) - p(F_{N_t})] \overset{asy}{\sim} N(0, \Sigma_p) \tag{3.5}
$$

with Σ_p as in (3.4).

Proof:

A finite vector of points from a Brownian Bridge is multivariate normal. Application of the δ–method (see Rao (1973), p. 387) yields the examined random vector. The elements of the covariance matrix are

$$E((\mathbb{B}(F(a_j)) - \mathbb{B}(F(a_{j-1})))^2) = p_j(F)(1 - p_j(F))$$

and for $j < k$:

$$E((\mathbb{B}(F(a_j)) - \mathbb{B}(F(a_{j-1})))(\mathbb{B}(F(a_k)) - \mathbb{B}(F(a_{k-1})))) = -p_j(F)p_k(F).$$

□

Using lemma 5.1 in Sandström (1983) a consistent variance estimator for σ_p^2 and Σ_p is given by changing F to $F_{n_t}^*$ because $F_{n_t}^*$ converges in probability to F. The estimator for Σ_p is

$$\hat{\Sigma}_p^* = \begin{pmatrix} p_1(F_{n_t}^*)(1 - p_1(F_{n_t}^*)) & -p_1(F_{n_t}^*)p_2(F_{n_t}^*) & \cdots & -p_1(F_{n_t}^*)p_K(F_{n_t}^*) \\ -p_2(F_{n_t}^*)p_1(F_{n_t}^*) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ -p_K(F_{n_t}^*)p_1(F_{n_t}^*) & \cdots & \cdots & p_K(F_{n_t}^*)(1 - p_K(F_{n_t}^*)) \end{pmatrix}. \quad (3.6)$$

For fixed sample size n define $n^* = (1 + v^2 - f)/n$ and $\Psi = n^*\Sigma_p$. From this definition it follows that the variance of $p(F_{n_t}^*)$ is Ψ, and a consistent estimator of Ψ is given by $\hat{\Psi}^* = n^*\hat{\Sigma}_p^*$. Furthermore, for large samples equation (3.5) is equivalent to $p(F_{n_t}^*) \sim N(p(F_N), \Psi)$.

Notice that the distribution of $p(F_{n_t}^*)$ is K–variate singular multinormal since the components of the random vector $p(F_{n_t}^*)$ add to unity, because of

$$\begin{aligned} \sum_{i=1}^{K} p_i(F_{n_t}^*) &= \sum_{i=1}^{K} \left(F_{n_t}^*(a_i) - F_{n_t}^*(a_{i-1})\right) = \sum_{i=1}^{K} F_{n_t}^*(a_i) - \sum_{i=1}^{K} F_{n_t}^*(a_{i-1}) \\ &= F_{n_t}^*(\infty) - F_{n_t}^*(0) = 1 - 0 = 1. \end{aligned}$$

4 Simultaneous inference

With the result of the above section confidence intervals can be built for the unknown population proportions $p_i(F_N)$. This section follows the work of Latorre (1995). The results can be applied directly to the estimation problem dealt with in this paper because of the asymptotic normality of the proportion estimator. Latorre (1995) gives three types of simultaneous confidence intervals: the Scheffé-type, the Bonferroni-type and the Sidàk-type. Thomas (1989) carries out a Monte Carlo simulation for various simultaneous confidence intervals using data from a two–stage cluster sample. He found that Bonferroni intervals based on transformations of the estimated proportions behave best. Because

the development of the intervals can be found in the articles only short sketches of the derivation and the intervals are given here.

Let $\mathbf{1}_{K:i}$ be a vector of length K whose elements are zero except the i–th which is equal to one. Furthermore, let p_i^- and p_i^+ be the lower and upper limits of the confidence interval for $p_i(F_N)$ and $p_0(F_{n_t}^*)$ the vector consisting of the first $K-1$ elements of $p(F_{n_t}^*)$ with a nonsingular multinormal distribution and mean vector $p_0(F_N)$ whose elements are the first $K-1$ elements of $p(F_N)$. The dispersion matrix is Ψ_0 which is consistently estimated by the nonsingular matrix $\hat{\Psi}_0^*$, being the $(K-1)\times(K-1)$ upper-left sub-matrix of $\hat{\Psi}^*$. Following Latorre (1995) the $(1-\alpha)$–simultaneous confidence intervals examined by the classical Scheffé projection method are.

$$\mathbf{1}'_{K-1:i}p_0(F_N) \in \mathbf{1}'_{K-1:i}p_0(F_{n_t}^*) \pm \sqrt{\chi^2_{K-1,1-\alpha}\mathbf{1}'_{K-1:i}\hat{\Psi}_0^*\mathbf{1}_{K-1:i}} \tag{4.1}$$

$i = 1,\ldots,K-1$ and

$$p_K(F_N) \in p(F_{n_t}^*) \pm \sqrt{\chi^2_{K-1,1-\alpha}\hat{\Psi}^*_{KK}} \tag{4.2}$$

where $\chi^2_{l,\beta}$ is the β-quantile of a χ^2-distribution with l degrees of freedom and $\hat{\Psi}^*_{KK}$ the K-th diagonal element of $\hat{\Psi}^*$. The intervals obtained in (4.1) and (4.2) are conservative. Shorter intervals can be computed using the Bonferroni inequality. The simultaneous confidence intervals of Bonferroni-type are

$$\mathbf{1}'_{K:i}p(F_N) \in \mathbf{1}'_{K:i}p(F_{n_t}^*) \mp z_{\alpha/2K}\sqrt{\mathbf{1}'_{K:i}\hat{\Psi}^*\mathbf{1}_{K:i}} \quad i = 1,\ldots,K \tag{4.3}$$

where z_β is the β-quantile of the standard normal distribution.

The Sidàk-type simultaneous confidence intervals are given by

$$\mathbf{1}'_{K:i}p(F_N) \in \mathbf{1}'_{K:i}p(F_{n_t}^*) \mp z_{(1-(1-\alpha)^{1/K})/2}\sqrt{\mathbf{1}'_{K:i}\hat{\Psi}^*\mathbf{1}_{K:i}} \quad i = 1,\ldots,K. \tag{4.4}$$

These intervals are shorter than the Bonferroni-type intervals. The different lengths can be explained by the different critical values. At usual levels of α the following inequality holds

$$\sqrt{\chi^2_{K-1,\alpha}} > |z_{\alpha/2K}| > |z_{(1-(1-\alpha)^{1/K})/2}|.$$

For a suitably smooth function g, $g(p_i(F_{n_t}^*))$ will be asymptotically $N(g(p_i(F_N)), (g'(p_i(F)))^2\hat{\Psi}^*_{ii})$. Bonferroni intervals can be obtained by inverting the corresponding intervals on the $g(p_i(F_N))$'s:

$$\mathbf{1}'_{K:i}p(F_N) \in g^{-1}\left(g(\mathbf{1}'_{K:i}p(F_{n_t}^*)) \mp z_{\alpha/2k}g'(\mathbf{1}'_{K:i}p(F_{n_t}^*)\sqrt{\mathbf{1}'_{K:i}\hat{\Psi}^*\mathbf{1}_{K:i}}\right) \tag{4.5}$$

(see Thomas (1989), chapter 3.4). Suitable choices for g are, e.g., $g_1(y) = \ln(y)$ or the logit $g_2(y) = ln(y/(1-y))$. For two stage cluster sampling Thomas (1989) shows that these intervals with $t_{r-1,1-\alpha/2k}$, the $1-\alpha/2k$–Quantile of the Student–t–distribution with $r-1$ degrees of freedom and r the number of clusters, instead of $z_{\alpha/2k}$ are the best simultaneous confidence intervals. Because arbitrary sampling designs are considered in this note this transformation will not taken into account. Further investigations for special sampling designs are necessary.

The limits of the simultaneous confidence intervals can be found in the following table

Table 1: Simultaneous confidence intervals for population proportions

Scheffé	p_i^-	$p_i(F_{n_t}^*) - \sqrt{\chi^2_{K-1,1-\alpha}\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}$
	p_i^+	$p_i(F_{n_t}^*) + \sqrt{\chi^2_{K-1,1-\alpha}\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}$
Bonferroni	p_i^-	$p_i(F_{n_t}^*) + z_{\alpha/2K}\sqrt{\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}$
	p_i^+	$p_i(F_{n_t}^*) - z_{\alpha/2K}\sqrt{\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}$
Sidàk	p_i^-	$p_i(F_{n_t}^*) + z_{(1-(1-\alpha)^{1/K})/2}\sqrt{\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}$
	p_i^+	$p_i(F_{n_t}^*) - z_{(1-(1-\alpha)^{1/K})/2}\sqrt{\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}$
transf.	p_i^-	$g^{-1}\left(g(p_i(F_{n_t}^*)) + z_{\alpha/2K}g'(p_i(F_{n_t}^*))\sqrt{\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}\right)$
Bonferroni	p_i^+	$g^{-1}\left(g(p_i(F_{n_t}^*)) - z_{\alpha/2K}g'(p_i(F_{n_t}^*))\sqrt{\frac{1+v^2-f}{n}p_i(F_{n_t}^*)(1-p_i(F_{n_t}^*))}\right)$

5 Application to the Socio–Economic–Panel data

The German Statistical Yearbook 1995 gives the number of households in 1993 classified by their monthly household net income (see Statistisches Bundesamt (1995), p. 554, table 21.7.2). From this table the household proportions (i.e., p_i) in the different income classes can be computed. To show the different behaviour of the estimates based on the sampling design and the "usual" proportion estimates, this paper examines these estimates and simultaneous confidence intervals of the household proportions which fall into the different net income-classes in Germany in 1993 using the German Socio–Economic–Panel (SOEP) data. The panel participants are interviewed annually; the data are recorded on the household level as well as on the individual level.

The panel includes a large section on income and earnings in the preceding year. The main respondent of a household is asked in each wave of the SOEP: "If you add all: What is the total monthly amount of the household net income of all household members today? Please report the monthly net amount, i.e. the amount after tax and social insurance. Regular payments like aid for dwelling and children, student loans, alimony payments, etc. add to the amount! In the case of "I don't know", please estimate the amount".

Notice that the "usual" proportion estimates are

$$p_i(F_n) = \frac{n_i}{n} \quad i = 1, \ldots, k$$

with n_i the number of observations which fall in class i. Simultaneous confidence intervals are given in Latorre (1995). They are

$$p_i(F) \in p_i(F_n) \pm \sqrt{\chi^2_{K-1,1-\alpha} \frac{1}{n} p_i(F_n)(1 - p_i(F_n))} \quad i = 1, \ldots, K \tag{5.1}$$

and

$$p_i(F) \in p_i(F_n) \mp z_{\alpha/2K} \sqrt{\frac{1}{n} p_i(F_n)(1 - p_i(F_n))} \quad i = 1, \ldots, K. \tag{5.2}$$

The sample of the SOEP consists of three subsamples with slightly different sampling designs. The basic design in all subsamples is a two-stage sampling with systematic sampling and sampling probability proportional to size. More information can be found in Rendtel (1995), pp. 23–25. The inclusion probabilities of the different households are given in the SOEP data. With this information the sampling design can be taken into account by using Horvitz-Thompson-estimates (HT-estimates) for the household proportions by setting the weights in (2.1) to π_i^{-1}. For the transformed Bonferroni intervals $g_1(y) = \ln(y)$ is used.

Table 2: Proportions of household by income classes, "usual" point estimates and 95% confidence intervals 1993

income classes	p_i	$p_i(F_n)$	95% confidence intervals	
in DM	in %	in %	with (5.1)	with (5.2)
<1200	8.68	6.68	[5.45,7.91]	[5.81,7.55]
1200 - 1800	13.32	10.11	[8.62,11.60]	[9.06,11.15]
1800 - 2500	18.64	18.55	**[16.63,20.47]**	**[17.20,19.90]**
2500 - 3000	11.15	13.25	[11.58,14.92]	[12.07,14.43]
3000 - 4000	17.94	22.10	[20.05,24.15]	[20.66,23.54]
4000 - 5000	12.33	14.52	[12.78,16.26]	[13.30,15.75]
5000 - 6000	7.29	7.10	**[5.84,8.37]**	**[6.21,8.00]**
6000 - 10000	8.57	7.06	[5.79,8.32]	[6.17,7.95]
>10000	2.08	0.63	[0.24,1.02]	[0.35,0.90]

Table 3: Proportions of household by income classes, HT–point estimates and 95% confidence intervals 1993

income classes	p_i	$p_i(F^*_{n_t})$	95% confidence intervals			
in DM	in %	in %	Scheffè	Bonferroni	Sidák	tr. Bonferroni
<1200	8.68	9.35	**[7.27,11.44]**	**[7.88,10.82]**	**[7.89,10.82]**	**[7.99,10.94]**
1200 - 1800	13.32	12.52	**[10.15,14.89]**	**[10.85,14.19]**	**[10.86,14.19]**	**[10.96,14.31]**
1800 - 2500	18.64	19.45	**[16.62,22.29]**	**[17.46,21.45]**	**[17.46,21.44]**	**[17.56,21.55]**
2500 - 3000	11.15	11.91	**[9.59,14.23]**	**[10.28,13.54]**	**[10.28,13.54]**	**[10.38,13.66]**
3000 - 4000	17.94	18.90	**[16.10,21.71]**	**[16.93,20.88]**	**[16.93,20.87]**	**[17.03,20.98]**
4000 - 5000	12.33	13.33	**[10.90,15.77]**	**[11.62,15.05]**	**[11.62,15.04]**	**[11.72,15.16]**

income classes	p_i	$p_i(F^*_{n_t})$	95% confidence intervals			
in DM	in %	in %	Scheffè	Bonferroni	Sidák	tr. Bonferroni
5000 - 6000	7.29	6.95	**[5.13,8.77]**	**[5.67,8.23]**	**[5.67,8.23]**	**[5.78,8.36]**
6000 - 10000	8.57	6.90	**[5.09,8.72]**	[5.62,8.18]	[5.63,8.18]	[5.74,8.31]
>10000	2.08	0.68	[0.09,1.26]	[0.26,1.09]	[0.26,1.09]	[0.37,1.25]

Intervals in bold print cover the value given by the official statistics. It can easily be seen that the confidence intervals of the HT-estimates cover most of the official proportions while the "usual" estimates only cover the true values in two out of nine income classes.

6 Monte Carlo simulations

In this section Monte Carlo simulations are carried out in order to test the behaviour of the design based estimators. HT-estimators are used for the simulations.

There are many possible candidates as population income distribution functions. Functional forms which capture empirically observed income distributions more or less well are, e.g., the Lognormal, Pareto, Gamma, Singh-Maddala distribution, Generalized Beta of first and second kind etc. (see e.g. Brachmann et al., 1996, for a discussion). Monte Carlo simulations for the Lognormal (LN) and the Singh-Maddala (SM) distribution are performed. The former has two parameters, $\mathrm{LN}(\mu, \sigma^2)$, and distribution function

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right) dt.$$

A nice property of the Lognormal is that the shape of the distribution (and thus the inequality) is entirely driven by the parameter σ^2. The simulations were carried out with parameters $\mu = 1$ and $\sigma^2 = 0.25$ (fitting a Lognormal distribution to German data would bring about a parameter estimate for σ^2 of roughly 0.28).

While the advantage of the Lognormal distribution is its dependency on just one parameter the advantage of the Singh-Maddala distribution (SM) is its good fit to real world data. It is a three parameter distribution, $\mathrm{SM}(a, b, c)$, but only b and c are shape parameters. The distribution function is

$$F(x) = 1 - \frac{1}{(1 + ax^b)^c}$$

Two sets of parameters are used for the SM. The first one (SM1), namely $a = 100$, $b = 2.8$ and $c = 1.7$, roughly mirrors the German household income distribution. The second one (SM2) is an extremely unequal distribution with arbitrarily chosen parameters, $a = 100$, $b = 2$ and $c = 0.7$.

From each distribution a finite population of 10000 observations is drawn. 10 classes are created and the true proportions for the populations are computed. The values can be found in table 4. For the Monte Carlo simulations 4 sampling designs are considered:

- Simple random sampling without replacement (SRS). The inclusion probabilities are: $\pi_i = n/N \ \forall i \in U$.

- Poisson sampling: For each element of the population a Bernoulli-experiment with parameter α_i is carried out. If the realization is 1 the i-th element enters the sample. This sampling design produces samples with random sample size. For this sampling design the inclusion probabilities are $\pi_i = \alpha_i$
- Stratified sampling with 2 classes (strat. 1): The first class contains the 25% smallest observations. The sampling design within the classes is SRS. The sample size for each class is $n/2$.
- Stratified sampling with 5 classes (strat. 2): The first class contains the 50% smallest observations, the second class the next 25%, the third class 12.5% the fourth class 6.25% and the last class 6.25%. The sampling design within the classes is SRS. The sample size for each class is $n/5$.

For each sampling design 10000 samples with sample size $n = 500$ and $n = 1000$ are drawn from the three finite populations. The coverage probabilities for the simultaneous confidence intervals for $\alpha = 0.05$ are given in tables 5 – 7.

Table 4: Classes and true proportions for the Monte-Carlo simulations

LN		SM1		SM2	
$[a_{i-1}, a_i[$	p_i (in %)	$[a_{i-1}, a_i[$	p_i (in %)	$[a_{i-1}, a_i[$	p_i (in %)
$[0, 1[$	2.41	$[0, 0.05[$	3.50	$[0, 0.1[$	38.78
$[1, 2[$	23.54	$[0.05, 0.10[$	18.26	$[0.1, 0.2[$	28.07
$[2, 3[$	31.12	$[0.10, 0.15[$	27.09	$[0.2, 0.3[$	12.55
$[3, 4[$	20.18	$[0.15, 0.20[$	22.55	$[0.3, 0.4[$	6.28
$[4, 5[$	11.40	$[0.20, 0.25[$	13.05	$[0.4, 0.5[$	3.56
$[5, 6[$	5.62	$[0.25, 0.30[$	6.97	$[0.5, 0.6[$	2.36
$[6, 7[$	2.77	$[0.30, 0.35[$	3.73	$[0.6, 0.7[$	1.46
$[7, 8[$	1.46	$[0.35, 0.40[$	1.84	$[0.7, 0.8[$	1.15
$[8, 9[$	0.63	$[0.40, 0.45[$	0.96	$[0.8, 0.9[$	0.79
$[9, \infty[$	0.78	$[0.45, \infty[$	1.60	$[0.9, \infty[$	5.00

Table 5: Coverage probabilities (in %) of the simultaneous 95%–confidence intervals for LN

CI-type	sampling design							
	$n = 500$				$n = 1000$			
	SRS	Poisson	strat. 1	strat. 2	SRS	Poisson	strat. 1	strat. 2
Scheffè	93.03	92.13	77.19	96.05	97.09	97.86	94.41	98.93
Bonferroni	79.42	77.57	71.34	94.72	85.78	87.30	81.96	95.51
Sidák	79.42	77.39	71.23	94.72	85.78	87.15	81.96	95.51
trans. Bonf.	89.47	87.99	74.25	96.58	96.10	94.08	90.81	97.29
(5.1)	99.08	98.84	00.00	00.00	98.27	98.12	00.00	00.00
(5.2)	83.85	86.13	00.00	00.00	89.85	90.23	00.00	00.00

Table 6: Coverage probabilities (in %) of the simultaneous 95%–confidence intervals for SM1

CI-type	sampling design							
	$n = 500$				$n = 1000$			
	SRS	Poisson	strat. 1	strat. 2	SRS	Poisson	strat. 1	strat. 2
Scheffè	97.10	95.31	91.74	97.47	99.28	98.58	97.78	98.83
Bonferroni	84.23	84.30	81.80	90.95	89.73	90.01	84.45	95.73
Sidák	84.23	84.10	81.80	90.95	89.69	89.82	84.41	95.73
trans. Bonf.	94.27	93.07	89.05	96.72	94.91	94.25	93.03	97.15
(5.1)	97.82	98.28	00.00	00.00	99.55	99.54	00.00	00.00
(5.2)	86.02	86.68	00.00	00.00	94.27	94.00	00.00	00.00

Table 7: Coverage probabilities (in %) of the simultaneous 95%–confidence intervals for SM2

CI-type	sampling design							
	$n = 500$				$n = 1000$			
	SRS	Poisson	strat. 1	strat. 2	SRS	Poisson	strat. 1	strat. 2
Scheffè	95.23	95.04	84.91	100.00	98.17	98.42	93.92	100.00
Bonferroni	80.79	79.00	70.72	99.62	88.34	87.42	83.18	99.77
Sidák	80.79	78.89	70.72	99.56	87.83	87.23	82.57	99.77
trans. Bonf.	93.39	91.96	82.36	99.58	95.55	94.38	91.65	99.73
(5.1)	97.10	97.27	00.00	00.00	98.40	98.54	00.00	00.00
(5.2)	81.37	83.99	00.00	00.00	89.48	91.99	00.00	00.00

It can be seen that the Scheffè-intervals are conservative and exceed the nominal coverage level in 16 out of 24 cases, while the coverage probabilities of the Bonferroni and Sidák-intervals are almost always too small. The transformed Bonferroni-intervals behave better than the former two. Their coverage probabilities are much closer to 95% than the probabilities of the Bonferroni and Sidák-intervals. Furthermore they are less conservative than the Scheffè-intervals.

The "usual" confidence intervals only give good results for the SRS and Poisson sampling designs. In these cases they show the same behaviour as the Scheffè and Bonferroni-intervals, but have always higher coverage rates. In the SRS design the similar behaviour is due to the fact that the only difference of the estimators is the lack of the factor $-f$ in (5.1) and (5.2). In stratified sampling both intervals fail. The coverage rate is 0 in all Monte Carlo simulations.

In small sample sizes the Scheffè and transformed Bonferroni-intervals give good results, while for $n = 1000$ the trans. Bonferroni-intervals are the better choice as they are less conservative. This is true for the three populations considered here. The Bonferroni and Sidák-intervals have coverage probabilities which are too small in most cases (except strat. 2).

The "usual" intervals should only be used for the sampling designs SRS and Poisson-sampling. However, in these designs the design based intervals can be used, as well, because they are slight transformations of the "usual" intervals.

The Monte Carlo simulations show that the design based confidence intervals behave reasonably well in a range of sampling designs of practical interest. Thus they should preferred to the "usual" intervals.

7 Conclusion

As can be seen in this paper the joint distribution of proportion estimates is asymptotically multinormal in a quite general framework. With this result simultaneous confidence intervals can be given for arbitrary sampling designs. Now it is possible to take the additional information into account which is given by knowing the sampling design. So better estimates and confidence intervals can be computed. Surely, this can only be done if there is some information about the sampling design. The example in section 5 shows that the sampling design based estimators behave much better than the "usual" estimators. Hence the former estimators should be used for the SOEP data. This result is supported by Monte Carlo simulations, which show that the sampling design based estimators behave better in stratified sampling where the "usual" estimators fail completely. Further research has to be done whether the design based estimates are better in other situations as well.

References

Brachmann, K.; Stich, A.; Trede, M. (1996), Evaluating parametric income distribution models, *Allgemeines Statistisches Archiv* 80: 285–298.

Cassel, C.M.; Särndal, C.E.; Wretman, J.H. (1977): *Foundations of inference in survey sampling.* John Wiley & Sons, New York.

Koul, H.L. (1970): Some convergence theorems for ranks and weighted empirical cumulatives. *The Annals of Mathematical Statistics,* 41, 1768-1773.

Latorre, G. (1993): Simultaneous inference for proportions in two-stage sampling on income data. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali*, 13, 33-51.

Latorre, G. (1995): Simultaneous inference for proportions in stratified random sampling. *Research on Economic Inequality,* Vol. 6, 255-268.

Mukhopadhyay, N.; Chattopadhyay, S. (1993): Simultaneous estimation of proportions in a finite population. *Calcutta Statistical Association Bulletin*, 43, 65-73.

Nygård, F.; Sandström, A. (1985a): Estimating Gini and Entropy inequality parameters. *Promemorior Fran P/STM Nr. 13,* 1985-01-09.

Nygård, F.; Sandström, A. (1985b): The estimation of the Gini and the entropy inequality parameters in finite populations. *Journal of Official Statistics,* 1, 399-412.

Nygård, F.; Sandström, A. (1989): Income inequality measures based on sample surveys. *Journal of Econometrics,* 42, 81-95.

Rao, C.R. (1973): *Linear statistical inference and its applications.* New York. John Wiley.

Rendtel, U. (1995): *Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität.* Campus, Fankfurt/ New York.

Royall, R.M. (1971): Linear regression models in finite population sampling theory. in: *Godambe, V.R.; Sprott, D.A. (eds.), Foundations of statistical inference.* Holt, Rinehart and Winston of Canada, Toronto.

Sandström, A. (1983): Estimating income inequality. Large sample inference in finite populations. *Research Report 1983:5, Department of Statistics, University of Stockholm.*

Shorack, G.R.; Wellner, J.A. (1986): *Empirical processes with applications to statistics.* New York, Wiley & Sons.

Statistisches Bundesamt (1995): Statistisches Jahrbuch 1995 für die Bundesrepublik Deutschland, Metzler-Poeschel, Stuttgart.

Thomas, D.R. (1989): Simultaneous confidence intervals for proportions under cluster sampling. *Survey Methodology,* 15, 187-201.