A note on quantile balanced sampling

Maciej Beręsewicz*

Abstract

This note presents an extension of balanced sampling to accommodate quantiles. The concept of quantile calibration, as proposed by Harms and Duchesne (2006), is applied to balanced sampling, thereby enabling the preservation of specific quantiles (e.g. quartiles, deciles) in the sample. This results in improvements in the estimation of quantiles and their variance. The paper presents a limited simulation study that demonstrates the performance of the proposed method.

Keywords: cube method, survey design, quantile estimation, calibration.

*Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Statistics, Al. Niepodległości 10, 61-875 Poznań, Poland, E-mail: maciej.beresewicz@ue.poznan.pl; Statistical Office in Poznań, ul. Wojska Polskiego 27/29 60-624 Poznań, Poland.

1 Introduction

The concept of balanced sampling via the cube method was originally proposed by Deville and Tillé (2004) and has subsequently been further developed for general sampling designs (Chauvet, 2009; Tillé, 2011), variance estimation (cf. Chauvet, Haziza, & Lesage, 2017; Deville & Tillé, 2005) and computational efficiency (Chauvet & Tillé, 2006; Hasler & Tillé, 2014). The concept was subsequently employed to achieve spatially balanced sampling (cf. Grafström, Lundström, & Schelin, 2012). In the existing literature on balanced sampling, the focus is on methods where the sampling procedure (i.e. sampling weights) should reproduce known totals of auxiliary variables. However, researchers or official statisticians may be interested in balancing other quantities, such as the median or deciles. This is because National Statistical Institutions rely on access to administrative data that contains rich unit-level data information, which may be crucial for business and mixed-mode surveys.

This paper extends the concept of balanced sampling to encompass quantiles by employing the methodology proposed by Harms and Duchesne (2006) for calibration estimators. As a result of this modification, quantile estimation becomes more efficient than it is with standard balanced sampling. To highlight the importance of quantiles, we use the term *quantile balanced sampling*. Furthermore, the proposed approach can be applied to stratified balanced sampling or data matching (Jauslin & Tillé, 2023).

The paper is structured as follows. Section 2 introduces the basic notation and presents balanced sampling. Section 3 defines quantile balanced sampling and discuss estimation of variance. Section 4 and Appendix contains limited simulation studies of the performance of the proposed method. The paper concludes with a summary and discussion. The proposed approach can be easily implemented in existing open-source software, for instance in R (R Core Team, 2023) via the `BalancedSampling` (Grafström, Lisic, & Prentius, 2024) and `jointCalib` (Beręsewicz, 2024) packages.

2 Basic setup

2.1 Notation

Let $U = \{1, ..., N\}$ denote the target population consisting of N labelled units. Each unit k has an associated vector of J auxiliary variables $\boldsymbol{x}$ and the target variable y, with their corresponding values $\boldsymbol{x}_k$ and y_k, respectively. Let $\{(y_k, \boldsymbol{x}_k, d_k), k \in S\}$ be a dataset of a probability sample of size n_A selected according to some sampling design $p(s)$.

The goal is to estimate a finite population total $\tau_y = \sum_{k \in U} y_k$ or the mean $\bar{\tau}_y = \tau_y / N$ of the variable of interest y. The Horvitz-Thompson is the well-known estimator of a finite population total, which is expressed as $\hat{\tau}_{y\pi} = \sum_{k \in s} d_k y_k$, where s denotes a probability sample of size n, $d_k = 1/\pi_k$ is a design weight and π_k is the first-order inclusion probability of the k-th element of the population U. This estimator is unbiased for τ_y i.e. $E\left(\hat{\tau}_{y\pi}\right) = \tau_y$ however may have high variance and thus are interested in selecting samples according to some efficient sampling design, for instance balanced sampling presented in the next section.

2.2 Balanced sampling

Definition 1. *(Balanced sampling) A sampling design $p(s)$ is said to be balanced with respect to the auxiliary variables $x_1, \ldots, x_J$, if and only if it satisfies the balancing equations given by*

$$\forall_{j=1,\ldots,J} \sum_{k \in U} \frac{x_{kj} s_k}{\pi_k} = \tau_{\boldsymbol{x}_j},$$

for all $s \in \mathcal{S}$ such that $p(\boldsymbol{s}) > 0$ where $\tau_{\boldsymbol{x}} = \sum_{k \in U} \boldsymbol{x}_k = \left(\sum_{k \in U} x_{k1}, \ldots, \sum_{k \in U} x_{kJ}\right)^T$ is vector of known population totals.

Balanced sampling can thus be viewed as a restriction on the support. Indeed, only the samples that satisfy the balancing equations have a strictly positive probability, that is, the support is

$$\mathcal{Q} = \left\{ \boldsymbol{s} \in \mathcal{S} \,\middle|\, \sum_{k \in U} \frac{\boldsymbol{x}_k s_k}{\pi_k} = \tau_{\boldsymbol{x}} \right\},$$

Special cases of the balanced sampling designs are unequal probability, cluster or stratified designs (cf. Tillé, 2011). Deville and Tillé (2004) proposed the cube method that efficiently

selects samples to meet the balancing constraints. Balanced sampling resembles calibration in such way that the sampling weights $d_k = 1/\pi_k$ should reproduce known population totals.

2.3 Calibration estimator for a total and a mean

Let $\boldsymbol{x}_k^\circ$ be a J_1-dimensional vector of auxiliary variables (benchmark variables) for which $\tau_{\boldsymbol{x}^\circ}$ is assumed to be known. Note that we distinguish auxiliary variables with known totals by $^\circ$ and quantiles by *.

The main idea of calibration for probability samples is to look for new calibration weights w_k that are as close as possible to original weights d_k and reproduce known population totals $\tau_{\boldsymbol{x}^\circ}$ exactly. In other words, in order to find new calibration weights w_k we have to minimise a distance function $D(\boldsymbol{d}, \boldsymbol{v}) = \sum_{k \in s} d_k G\left(\frac{v_k}{d_k}\right) \rightarrow \min$ to fulfil calibration equations $\sum_{k \in s} v_k \boldsymbol{x}_k^\circ = \tau_{\boldsymbol{x}^\circ}$, where $\boldsymbol{d} = (d_1, \ldots, d_n)^T$, $\boldsymbol{v} = (v_1, \ldots, v_n)^T$ and $G(\cdot)$ is a function that must satisfy some regularity conditions: $G(\cdot)$ is strictly convex and twice continuously differentiable, $G(\cdot) \geq 0$, $G(1) = 0$, $G'(1) = 0$ and $G''(1) = 1$. Examples of $G(\cdot)$ functions are given by Deville and Särndal (1992). For instance, if $G(x) = \frac{(x-1)^2}{2}$, then using the method of Lagrange multipliers the final calibration weights w_k can be expressed as $w_k = d_k + d_k \left(\tau_{\boldsymbol{x}^\circ} - \hat{\tau}_{\boldsymbol{x}^{\circ}\pi}\right)^T \left(\sum_{j \in s} d_j \boldsymbol{x}_j^\circ \boldsymbol{x}_j^{\circ T}\right)^{-1} \boldsymbol{x}_k^\circ$ where $\hat{\tau}_{\boldsymbol{x}^{\circ}\pi}$ is the vector of estimated totals using π_k.

The final calibration estimator of a population total τ_y can be expressed as $\hat{\tau}_{y\boldsymbol{x}^\circ} = \sum_{k \in s} w_k y_k$, where w_k are calibration weights obtained using a specific $G(\cdot)$ function and notation $y\boldsymbol{x}^\circ$ is used to denote that calibration using $\boldsymbol{x}^\circ$ was applied. In the next section we show how calibration is extended for estimation of quantiles.

2.4 Calibration estimator for a quantile

Harms and Duchesne (2006) considered the estimation of quantiles using the calibration approach in a way very similar to what Deville and Särndal (1992) proposed for a finite population total τ_y (see Berger and Munoz (2015) for further information on additional estimation techniques and their comparative analysis). By analogy, in their approach it is not necessary to know values of all auxiliary variables for all units in the population. Below we briefly discuss the problem of finding calibration weights under this setup.

We want to estimate a quantile $Q_{y,\alpha}$ of order $\alpha \in (0, 1)$ of the variable of interest y, which

can be expressed as $Q_{y,\alpha} = \inf\{t \,|F_y(t) \geq \alpha\}$, where $F_y(t) = N^{-1}\sum_{k\in U} H(t-y_k)$ and the Heavyside function is given by

$$H(t-y_k) = \begin{cases} 1, & t \geq y_k, \\ 0, & t < y_k. \end{cases} \tag{1}$$

We assume that $\boldsymbol{Q}_{\boldsymbol{x}^*,\alpha} = \left(Q_{x_1,\alpha},\ldots,Q_{x_{J_2},\alpha}\right)^T$ is a vector of known population quantiles of order α for a vector of auxiliary variables $\boldsymbol{x}_k^*$, where $\alpha \in (0,1)$ and $\boldsymbol{x}_k^*$ is a J_2-dimensional vector of auxiliary variables. It is worth noting that, in general, the numbers J_1 and J_2 of auxiliary variables may be different. It may happen that for a specific auxiliary variable its population total and the corresponding quantile of order α will be known. However, in most cases quantiles will be known for continuous auxiliary variables, unlike totals, which will generally be known for categorical variables.

In order to find new calibration weights w_k that reproduce known population quantiles in a vector $Q_{\boldsymbol{x}^*,\alpha}$, an interpolated distribution function estimator of $F_y(t)$ is defined as $\hat{F}_{y,cal}(t) = \frac{\sum_{k\in s} w_k H_{y,s}(t,y_k)}{\sum_{k\in s} w_k}$, where the Heavyside function in formula (1) is replaced by the modified function $H_{y,s}(t,y_k)$ given by

$$H_{y,s}(t,y_k) = \begin{cases} 1, & y_k \leqslant L_{y,s}(t), \\ \vartheta_{y,s}(t), & y_k = U_{y,s}(t), \\ 0, & y_k > U_{y,s}(t), \end{cases} \tag{2}$$

where appropriate parameters are defined as $L_{y,s}(t) = \max\{\{y_k, k\in s \mid y_k \leqslant t\}\cup\{-\infty\}\}$, $U_{y,s}(t) = \min\{\{y_k, k\in s \mid y_k > t\}\cup\{\infty\}\}$ and $\vartheta_{y,s}(t) = \frac{t-L_{y,s}(t)}{U_{y,s}(t)-L_{y,s}(t)}$ for $k = 1,\ldots,n$, $t\in\mathbb{R}$. From a practical point of view a smooth approximation to the step function, based on the logistic function can be used i.e. $H(x) \approx \frac{1}{2} + \frac{1}{2}\tanh kx = (1+e^{-2kx})^{-1}$, where a larger value of k corresponds to a sharper transition at $x = 0$.

A calibration estimator of quantile $Q_{y,\alpha}$ of order α for variable y is defined as $\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha)$, where a vector $\boldsymbol{w} = (w_1,\ldots,w_n)^T$ is a solution of an optimization problem $D(\boldsymbol{d},\boldsymbol{v}) = \sum_{k\in s} d_k G\left(\frac{v_k}{d_k}\right) \to \min$ subject to the calibration constraints $\sum_{k\in s} v_k = N$ and $\hat{\boldsymbol{Q}}_{\boldsymbol{x}^*,cal,\alpha} = \left(\hat{Q}_{x_1,cal,\alpha},\ldots,\hat{Q}_{x_{J_2},cal,\alpha}\right)^T = \boldsymbol{Q}_{\boldsymbol{x}^*,\alpha}$ or equivalently $\hat{F}_{x_j,cal}\left(Q_{x_j,\alpha}\right) = \alpha$, where $j = 1,\ldots,J_2$.

As in the previous case, if $G(x) = \frac{(x-1)^2}{2}$, then, using the method of Lagrange multipliers,

the final weights w_k can be expressed as $w_k = d_k + d_k \left(\mathbf{T_a} - \sum_{k\in s} d_k \boldsymbol{a}_k\right)^T \left(\sum_{j\in s} d_j \boldsymbol{a}_j \boldsymbol{a}_j^T\right)^{-1} \boldsymbol{a}_k$, where $\mathbf{T_a} = (N, \alpha, \ldots, \alpha)^T$ and the elements of $\boldsymbol{a}_k = (1, a_{k1}, \ldots, a_{kJ_2})^T$ are given by

$$a_{kj} = \begin{cases} N^{-1}, & x_{kj} \leq L_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ N^{-1}\vartheta_{x_j,s}\left(Q_{x_j,\alpha}\right), & x_{kj} = U_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ 0, & x_{kj} > U_{x_j,s}\left(Q_{x_j,\alpha}\right), \end{cases} \tag{3}$$

with $j = 1, \ldots, J_2$ and $\vartheta_{x_j,s}$ is defined similarly as $\vartheta_{y_j,s}$.

In the method described above it is assumed that a known population quantile is reproduced for a set of auxiliary variables, i.e. that the process of calibration is based on a particular quantile (of order α). For instance, it could be the median $\alpha = 0.5$ or a vector of $\boldsymbol{\alpha} = (0.1, \ldots, 0.9)$ for a specific variable.

3 Quantile balanced sampling

3.1 Definitions

In this section we provide the definition of quantile balanced sampling i.e. method that allow to select a sample taking into account not only totals but also a specific quantile of auxiliary variables.

Definition 2 (Quantile balanced sampling)**.** *A sampling design $p(s)$ is said to be quantile balanced if and only if it satisfies the balancing equations with respect to $\boldsymbol{x}^\circ$ variables*

$$\forall_{j=1,\ldots,J_1} \sum_{k\in U} \frac{x_{kj}^\circ s_k}{\pi_k} = \tau_{\boldsymbol{x}_j^\circ},$$

and balancing equations with respect to $\boldsymbol{x}^*$ through $\boldsymbol{a}$ variables

$$\forall_{j=1,\ldots,J_2} \sum_{k\in U} \frac{a_{kj} s_k}{\pi_k} = \alpha_j,$$

for all $s \in \mathcal{S}$ such that $p(\boldsymbol{s}) > 0$ and where $\tau_{\boldsymbol{x}^\circ}$ is defined as previously, α_j is a scalar (e.g. 0.5; median) and a_{kj} is defined either using Heavyside function given by (3) or approximation using logistic function.

Remark 1. *Definition 2 can be further extended to balance multiple quantiles i.e. a vector of* $\boldsymbol{\alpha}_j$ *for a given* x_j^* *variable. For instance, a researcher may be interested in balancing quartiles (i.e.* $\boldsymbol{\alpha} = (0.25, 0.50, 0.75)$*) or deciles (i.e.* $\boldsymbol{\alpha} = (0.1, \ldots, 0.9)$*). Another situation may be that one may balance different quantiles for different variables. For instance, median for say* x_1^* *but deciles for* x_2^**. This could be written down as follows*

$$\forall_{\substack{j=1,\ldots,J_2 \\ \alpha \in \mathcal{A}_j}} \sum_{k \in U} \frac{a_{kj}^{\alpha} s_k}{\pi_k} = \alpha_j,$$

where $\mathcal{A}_j$ *is a set of quantiles for a given variable* j*.*

Remark 2. *We do not define quantile balanced sampling with respect to only quantiles because the null-space (kernel) of such matrix may not exist. Thus, it may not be possible to select sample based only on balancing quantiles using the cube method.*

Deville and Tillé (2004) showed that the cube method used for selection of balanced sampling accurately reproduce known population totals i.e.

$$\left| \frac{\hat{\tau}_{\boldsymbol{x}_j^\circ} - \tau_{\boldsymbol{x}_j^\circ}}{\tau_{\boldsymbol{x}_j^\circ}} \right| < O\left(\frac{p}{n}\right),$$

were $\hat{\tau}_{\boldsymbol{x}_j^\circ} = \sum_{k \in S} (\pi_k)^{-1} \boldsymbol{x}_j^\circ$, p is the number of variables and n is the sample size. Therefore, the same procedure applied for quantiles via the $\boldsymbol{a}$ variables

$$\left| \frac{\hat{\alpha}_j - \alpha_j}{\alpha_j} \right| < O\left(\frac{p}{n}\right),$$

where $\hat{\alpha}_j = \sum_{k \in S} (\pi_k)^{-1} a_{kj}$ is a given α quantile for a given $\boldsymbol{x}_j$ variable.

This, the drawback of this procedure is that inclusion of quantiles increases number of variables p and thus the accuracy of the cube method may be lower than the standard balanced sampling.

3.2 Extensions

Chauvet (2009) proposed an extension of the balanced sampling for stratified populations and Hasler and Tillé (2014) proposed an algorithm for highly stratified populations. This sampling design can be easily accounted for by calculating quantiles for specific stratas. In such setting

the following definition can be proposed.

Definition 3 (Stratified quantile balanced sampling)**.** *A sampling design $p(s)$ is said to be stratified quantile balanced if and only if it satisfies the balancing equations with respect to $\boldsymbol{x}^\circ$ variables*

$$\forall_{\substack{j=1,\ldots,J_1\\h=1,\ldots,H}} \sum_{k\in U_h} \frac{x^\circ_{kj} s_k}{\pi_k} = \tau_{\boldsymbol{x}^\circ_j},$$

and balancing equations with respect to $\boldsymbol{x}^*$ through $\boldsymbol{a}$ variables

$$\forall_{\substack{j=1,\ldots,J_2\\h=1,\ldots,H}} \sum_{k\in U_h} \frac{a_{kj} s_k}{\pi_k} = \alpha_j,$$

where $h = 1, \ldots, H$ denotes stratum in the population U.

This approach may be useful in particular for business surveys where samples are selected proportionally to the size within stratas defined by NACE groups.

3.3 Variance estimation

To estimate the variance of the quantile balanced sampling we can use the procedure introduced by Deville and Tillé (2005). Let $\boldsymbol{x}_k = \left(\left(\boldsymbol{x}^\circ_k\right)^T, \boldsymbol{a}_k^T\right)^T$ be a vector of auxiliary variables, then a family of variance estimators for balanced sampling proposed by Deville and Tillé (2005) is given by

$$\widehat{\text{var}}\left(\hat{\tau}_{y\pi}\right) = \sum_{k\in s} c_k \frac{\left(y_k - \boldsymbol{x}_k^T \hat{\boldsymbol{b}}\right)^2}{\pi_k^2}, \tag{4}$$

where

$$\hat{\boldsymbol{b}} = \left(\sum_{\ell\in s} c_\ell \frac{\boldsymbol{x}_\ell \boldsymbol{x}_\ell^T}{\pi_\ell^2}\right)^{-1} \sum_{\ell\in s} c_\ell \frac{\boldsymbol{x}_\ell y_\ell}{\pi_\ell^2}$$

and the c_k are the solutions of the nonlinear system

$$1 - \pi_k = c_k - \frac{c_k \boldsymbol{x}_k^T}{\pi_k} \left(\sum_{\ell\in S} c_\ell \frac{\boldsymbol{x}_\ell \boldsymbol{x}_\ell^T}{\pi_\ell^2}\right)^{-1} \frac{c_k \boldsymbol{x}_k}{\pi_k}.$$

Theorem 1. *Quantile balanced sampling is at least as efficient as balanced sampling in estimating the population total (or mean) of the variable of interest y, i.e.,*

$$Var(\hat{\tau}_{y\pi}^{QBS}) \leq Var(\hat{\tau}_{y\pi}^{BS}),$$

where $\hat{\tau}_{y\pi}^{QBS}$ and $\hat{\tau}_{y\pi}^{BS}$ are the Horvitz-Thompson estimators of the population total τ_y under quantile balanced sampling and balanced sampling, respectively.

Proof. Using the proposed approach we simply concatenate vectors $\boldsymbol{x}_k^\circ$ and $\boldsymbol{a}_k$ and the linear regression, that is basis of this approach, consist of two set of model parameters:

$$\hat{y}_k = \left(\boldsymbol{x}_k^\circ\right)^T \hat{\boldsymbol{b}} + \boldsymbol{a}_k^T \hat{\boldsymbol{\gamma}}.$$

Adding $\boldsymbol{a}_k^T \hat{\boldsymbol{\gamma}}$ may improve the fit and thus the sum of residuals given in the nominator of (4) will be smaller than with just $\boldsymbol{x}_k^\circ$. □

In the next section we study the performance of the proposed method using a limited simulation study.

4 Simulation study

In this simulation study we use data from Särndal, Swensson, and Wretman (1992, Appendix B) on municipalities in Sweden. Data contain 284 observations and 11 variables. In the simulation study we are interested in estimating population total, mean, median and 90$^{\text{th}}$ of revenues from 1985 municipal taxation (in millions of kronor; denoted as `RMT85`). Additional simulation study based on artificial data covering balanced and quantile balanced sampling followed by calibration is presented in Appendix A.

We used the following 6 variables as auxiliary information: the 1975 population (in thousands; denoted as `P75`), the number of Conservative seats in municipal council (denoted as `CS82`), the number of Social-Democratic seats in municipal council (denoted as `SS82`), total number of seats in municipal council (denoted as `S82`), the number of municipal employees in 1984 (denoted as `ME84`) and real estate values according to 1984 assessment (in millions of kronor; denoted as `REV84`).

We use unequal probability sample design where the inclusion is proportional to the `P75` variable with sample size equal to 50. We consider two methods:

- balanced sampling with constraints on totals of all six auxiliary variables (denoted as `Balanced`),
- quantile balanced sampling with constraints on totals and
  - medians for all six auxiliary variables (denoted as `QB medians`). In this approach constraints consisted of 12 equations.
  - quartiles of all six auxiliary variables (denoted as `QB quartiles`). In this approach constraints consisted of 24 equations.

In the results, we report Monte Carlo bias (Bias), variance (Variance) and relative mean square error (RMSE) of total, mean, median and 90^{th} percentile based on $R = 10,000$ simulations for the `RMT85` variable. The simulation was conducted in R (R Core Team, 2023) using the `BalancedSampling` (Grafström, Lisic, & Prentius, 2024, the `cube` function) and the `jointCalib` (Beręsewicz, 2024).

Table 4.1: Monte Carlo simulation results for mean, median, 90^{th} percentile and total of RMT85 variable

Characteristic	Method	Bias	Variance	RMSE
Mean	Balanced	1.73	448.46	21.25
	QB medians	1.25	351.33	18.79
	QB quartiles	0.81	136.10	11.69
Median	Balanced	1.39	457.56	21.44
	QB medians	0.70	392.03	19.81
	QB quartiles	-0.05	266.26	16.32
90^{th} percentile	Balanced	5.31	3,267.42	57.41
	QB medians	4.86	2,913.60	54.20
	QB quartiles	6.32	2,135.73	46.64
Total	Balanced	5.32	866,550.56	930.90
	QB medians	-3.36	808,737.37	899.30
	QB quartiles	18.70	727,652.85	853.23

Monte Carlo simulation results for the mean, median, 90^{th} percentile and total are presented in the Table 4.1. For all characteristics the proposed quantile balanced sampling performed better, in particular the RMSE significantly decreased for the mean and median when quartiles are used (`QB quartiles`). When only medians are used the results are not significantly different from the one obtained from standard balanced sampling.

Figure 1 presents Monte Carlo coefficient of variation (CV) for the estimated quantiles of the auxiliary variables. The solid line refer to standard balanced sampling, dashed to QB sampling with medians and dotted to QB with quartiles. For all variables we observe improvement in terms of precision but it varies across the variables. For instance, for the variables regarding the number of seats (SS82 and S82) the improvement is slight (less than 1 p.p.) while for the rest of variables improvement ranges between 2-5 p.p.

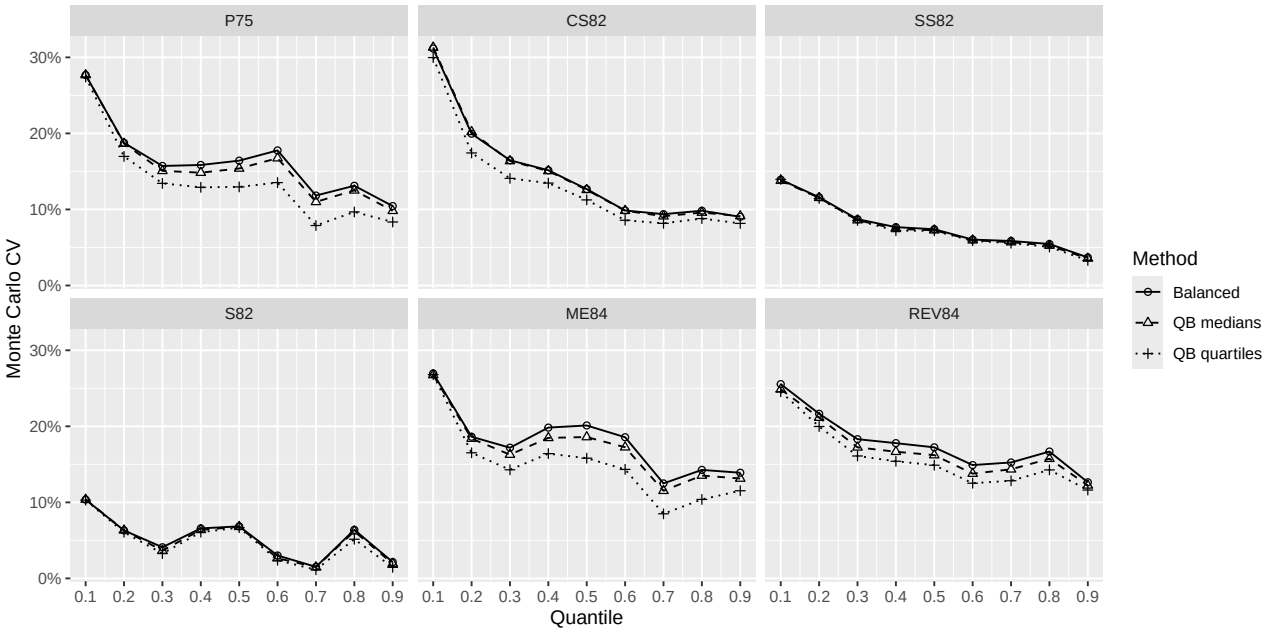

Figure 1: Monte Carlo coefficient of variation of the estimated quantiles of the auxiliary variables

This limited simulation study shows that the proposed methods lead to more efficient sampling procedure which is observed through improvements in RMSE. In this simulation study we did not verify properties of estimators based on combination of balanced sampling and calibration as suggested by Tillé (2011) due to small sample size. Results regarding joint application of these methods are presented in Appendix A.

5 Summary

In this note we proposed an extension of the balanced sampling to account for quantiles. This method is based on adding new balancing constraints so the selected sample will match specified quantiles of auxiliary variables (e.g. median, deciles). As shown in a limited simulation study inclusion of deciles significantly improves not only quantiles but also means and totals.

The proposed method may be of interest to National Statistical Institutes wishing to make more extensive use of administrative data in the process of sample selection. The use of information not only on totals but also on distributions can improve estimates, especially for business surveys.

Acknowledgements

The authors' work has been financed by the National Science Centre in Poland, OPUS 20, grant no. 2020/39/B/HS4/00941. Codes to reproduce simulations from the paper are available at https://github.com/ncn-foreigners/paper-q-sampling. I would like to thank Tomasz Żądło for discussion on this note.

References

Beręsewicz, M. (2024). *jointCalib: A Joint Calibration of Totals and Quantiles* [R package version 0.1.2, https://ncn-foreigners.github.io/jointCalib/]. https://doi.org/10.32614/CRAN.package.jointCalib

Berger, Y. G., & Munoz, J. F. (2015). On estimating quantiles using auxiliary information. *Journal of Official Statistics*, *31*(1), 101–119.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, *35*(1), 115–119.

Chauvet, G., Haziza, D., & Lesage, É. (2017). Examining some aspects of balanced sampling in surveys. *Statistica Sinica*, 313–334.

Chauvet, G., & Tillé, Y. (2006). A fast algorithm for balanced sampling. *Computational Statistics*, *21*(1), 53–62.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382.

Deville, J.-C., & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, *91*(4), 893–912.

Deville, J.-C., & Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of statistical planning and Inference*, *128*(2), 569–591.

Grafström, A., Lisic, J., & Prentius, W. (2024). *BalancedSampling: Balanced and Spatially Balanced Sampling* [R package version 2.0.6]. https://doi.org/10.32614/CRAN.package.BalancedSampling

Grafström, A., Lundström, N. L., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, *68*(2), 514–520.

Harms, T., & Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, *32*(1), 37–52.

Hasler, C., & Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational statistics & data analysis*, *74*, 81–94.

Jauslin, R., & Tillé, Y. (2023). An efficient approach for statistical matching of survey data through calibration, optimal transport and balanced sampling. *Journal of Statistical Planning and Inference*, *225*, 121–131.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer Science & Business Media.

Tillé, Y. (2011). Ten years of balanced sampling with the cube method: An appraisal. *Survey methodology*, *37*(2), 215–226.

Appendix for the paper

A note on quantile balanced sampling

A Additional simulation study

We generate a finite population $\mathcal{F}_N = \{\boldsymbol{X}_k = (X_{k1}, X_{k2}), \boldsymbol{Y}_k = (Y_{k1}, Y_{k2}) : k = 1, \ldots, N\}$ with size $N = 100,000$, where Y_{k1} is the continuous outcome, Y_{k2} is the binary outcome, $X_{k1} \sim N(1,1)$ and $X_{k2} \sim \text{Exp}(1)$. The finite population $\mathcal{F}_N$ were generated using the outcome variables presented in Table A.1, where $\alpha_k \sim N(0,1)$ and $\epsilon_k \sim N(0,1)$ and X_{k1}, X_{k2}, α_k and ϵ_k are mutually independent. The variable α_k induces dependence of Y_{k1} and Y_{k2} even after adjusting for X_{k1} and X_{k2}. In the simulation study we consider two sample sizes: 500 and 1,000.

Table A.1: Outcome models used in the simulation study

Type	Form	Formulae
Continuous	linear (Y_{11})	$Y_{k1} = 1 + X_{k1} + X_{k2} + \alpha_k + \epsilon_k$
	non-linear (Y_{12})	$Y_{k2} = 0.5(X_{k1} - 1.5)^2 + X_{k2}^2 + \alpha_k + \epsilon_k$
Binary	linear (Y_{21})	$P(Y_{k1} = 1 \mid X_{k1}, X_{k2}; \alpha_k) = \text{logit}(-1 + X_{k1} + X_{k2} + \alpha_k)$
	non-linear (Y_{22})	$P(Y_{k2} = 1 \mid X_{k1}, X_{k2}; \alpha_k) = \text{logit}\{0.5(X_{k1} - 1.5)^2 + X_{k2}^2 + \alpha_k\}$

In the results, we report Monte Carlo bias (Bias), variance (Var) and relative mean square error (RMSE) based on $R = 10,000$ simulations for each Y variables and for each target quantity i.e. total, mean, median and 90^{th} percentile[1]. We consider the following methods: balanced sampling (denoted as BS) without and with calibration to totals and quantile balanced sampling (denoted as QBS) without and with calibration to totals and quantiles jointly.

Table A.2 contains results for the totals, while table A.3 contains results for the means, medians and 90th percentiles. For linear continuous and binary target variables, balanced sampling with calibration and quantile balanced sampling with calibration are comparable in terms of bias, variance and RMSE. However, a slight improvement in terms of variance is observed when comparing balanced sampling and quantile balanced sampling only (for all characteristics). The main difference is observed when we consider non-linear continuous variables (Y_{12}). For both sample sizes, the variance and RMSE of the total are lower than for balanced sampling.

[1]Median and percentile were calculated for Y_{11} and Y_{12} only.

The RMSE for balanced sampling and quantile balanced sampling without calibration is larger by approximately 5% and with calibration by approximately 16%. The same interpretation can be applied to other characteristics, including the mean, median, and 90th percentile. For the binary non-linear variable (Y_{22}), the observed differences are negligible.

Table A.2: Monte Carlo simulation results for totals

Variable	Method	Calibration	Bias	Variance	RMSE
		$n = 500$			
Y_{11}	BS	No	6.77	86,214.99	9,285.21
		Yes	93.49	39,885.16	6,316.16
	QBS	No	-54.47	63,230.22	7,951.93
		Yes	-64.42	41,643.62	6,453.51
Y_{12}	BS	No	0.22	88,944.48	9,431.04
		Yes	12.87	60,363.18	7,769.39
	QBS	No	-26.25	79,910.95	8,939.33
		Yes	-292.96	44,046.85	6,643.24
Y_{21}	BS	No	-22.18	6,068.50	2,463.53
		Yes	2.66	3,972.66	1,993.15
	QBS	No	1.58	4,358.97	2,087.82
		Yes	10.15	3,919.81	1,979.88
Y_{22}	BS	No	12.52	7,422.63	2,724.48
		Yes	31.72	3,569.58	1,889.60
	QBS	No	12.24	4,221.55	2,054.68
		Yes	12.71	3,565.99	1,888.43
		$n = 1,000$			
Y_{11}	BS	No	66.28	43,328.87	6,582.80
		Yes	2.42	19,639.60	4,431.66
	QBS	No	33.13	31,609.11	5,622.30
		Yes	41.99	20,262.34	4,501.57
Y_{12}	BS	No	47.62	44,416.02	6,664.71
		Yes	-26.02	29,894.58	5,467.66
	QBS	No	54.36	39,819.42	6,310.50
		Yes	-77.00	21,628.80	4,651.31
Y_{21}	BS	No	25.16	2,995.54	1,730.95
		Yes	15.09	1,941.90	1,393.60
	QBS	No	0.86	2,277.80	1,509.24
		Yes	3.39	1,972.86	1,404.59
Y_{22}	BS	No	28.45	3,665.87	1,914.86
		Yes	9.44	1,703.88	1,305.36
	QBS	No	-2.21	2,073.23	1,439.87
		Yes	-4.97	1,676.92	1,294.97

Table A.3: Monte Carlo simulation results for mean, median and 90^{th} quantile

			Y_{11}			Y_{12}			Y_{21}			Y_{22}		
Parameter	Method	Calib	Bias	Var	RMSE	Bias	Var	RMSE	Bias	Var	RMSE	Bias	Var	RMSE
						$n = 500$								
Mean	BS	No	0.11	0.66	8.14	0.05	0.82	9.05	-0.01	0.04	2.09	0.02	0.04	1.94
		Yes	0.09	0.40	6.32	0.01	0.60	7.77	0.00	0.04	1.99	0.03	0.04	1.89
	QBS	No	-0.05	0.61	7.78	-0.03	0.79	8.90	0.00	0.04	2.03	0.01	0.04	1.92
		Yes	-0.06	0.42	6.45	-0.29	0.44	6.64	0.01	0.04	1.98	0.01	0.04	1.89
Median	BS	No	0.24	1.01	10.07	0.10	0.98	9.89						
		Yes	0.02	0.83	9.09	-0.02	0.84	9.16						
	QBS	No	0.10	0.97	9.85	-0.09	0.95	9.75						
		Yes	-0.08	0.84	9.17	-0.26	0.79	8.89						
90^{th} quantile	BS	No	1.07	3.19	17.90	0.24	2.97	17.24						
		Yes	0.70	2.41	15.53	0.01	2.34	15.30						
	QBS	No	0.72	3.09	17.60	0.35	2.96	17.20						
		Yes	0.71	2.48	15.77	0.40	2.12	14.57						
						$n = 1,000$								
Mean	BS	No	0.02	0.33	5.74	0.02	0.41	6.39	0.01	0.02	1.46	0.01	0.02	1.34
		Yes	0.00	0.20	4.43	-0.03	0.30	5.47	0.02	0.02	1.39	0.01	0.02	1.31
	QBS	No	0.01	0.30	5.50	0.04	0.39	6.27	-0.00	0.02	1.46	-0.01	0.02	1.34
		Yes	0.04	0.20	4.50	-0.08	0.22	4.65	0.00	0.02	1.40	-0.00	0.02	1.29
Median	BS	No	0.09	0.49	7.02	-0.06	0.48	6.90						
		Yes	-0.02	0.40	6.31	-0.18	0.41	6.39						
	QBS	No	0.05	0.47	6.86	0.06	0.47	6.87						
		Yes	-0.00	0.40	6.35	0.01	0.39	6.22						
90^{th} quantile	BS	No	0.50	1.61	12.69	0.08	1.52	12.34						
		Yes	0.33	1.23	11.09	-0.05	1.18	10.86						
	QBS	No	0.71	1.52	12.36	0.17	1.46	12.07						
		Yes	0.70	1.17	10.85	0.21	1.03	10.14						