On the maximum likelihood estimation of population and domain means

Janusz L. Wywiał

Department of Statistics Econometrics and mathematics
University of Economics in Katowice

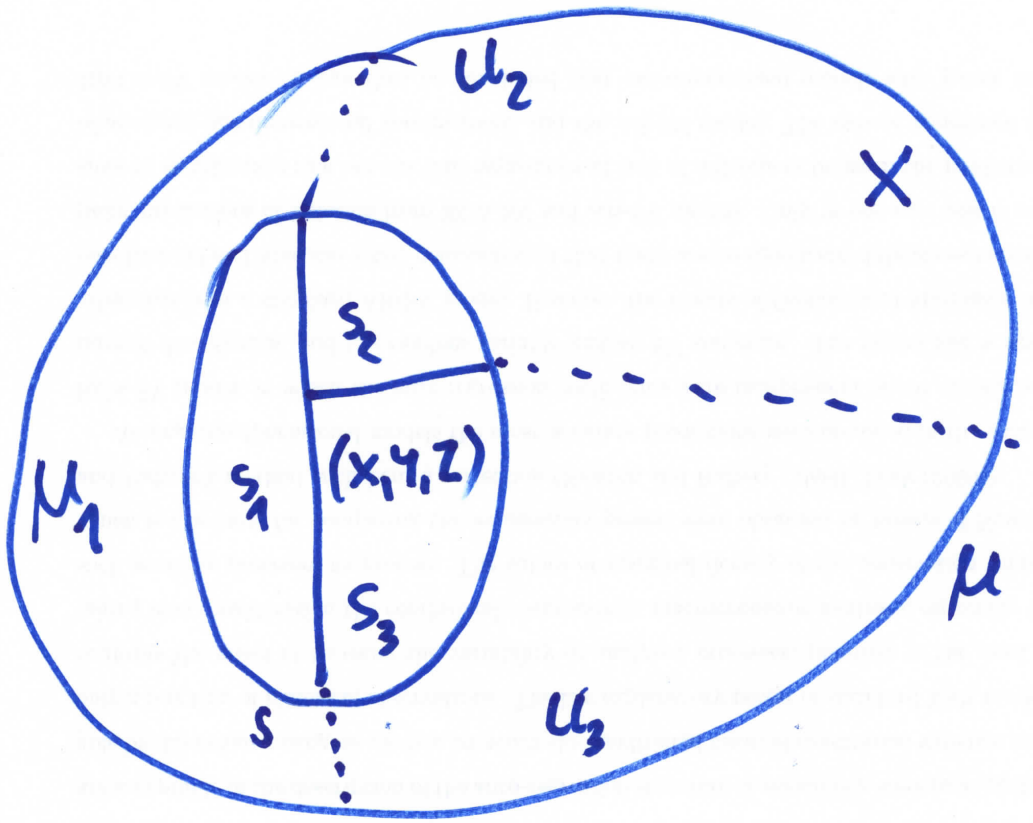
The 5th Congress of Polish Statistics
1-3 July 2025, Warsaw, Poland

Plan

- Formulation of the problem
- Model approach based on distribution mixture
- Maximum likelihood estimators
- Bivariate normal case
- Simulation analysis of estimation accuracy
- Conclussion.

Formulation of the problem

This problem was considered e.g. in [1], [2], [7]-[9], [12]-[14].



Model-design approach based on distribution mixture

Notation

$U = U_1 ... \cup U_h \cup ... \cup U_H$ - population as domains sum;

$d_k = [y_k, \mathbf{x}_k, \mathbf{z}_{k*}]$ - k-th observation of variable under study, auxiliary variable vector: $\mathbf{x}_k = [x_{k,1} ... x_{k,m}]$, vector identifying domains: $\mathbf{z}_{k*} = [z_{k,1} ... z_{k,H}]$ is zero vector except one element equal to 1; when $k \in U_h$, $\mathbf{z}_{k*} = \mathbf{z}_{k*}^{(h)}$;
$k = 1, ..., N$, $h = 1, .., H < N$, $1 \leq m < N$;

Let distribution of $[Y_k \mathbf{X}_k]$ be mixture of continuous densities:

$$f(y_k, \mathbf{x}_k) = \sum_{h=1}^{H} f(y_k, \mathbf{x}_k | \mathbf{Z}_{k*} = \mathbf{z}_{k*}^{(h)}) p_h, \quad p_h = P(\mathbf{Z}_{k*} = \mathbf{z}_{k*}^{(h)})$$

Model-design approach based on distribution mixture

Notation

$$f(y_k, \mathbf{x}_k, \theta) = \sum_{h=1}^{H} p_h f_h(y_k, \mathbf{x}_k, \theta_h), \quad k \in U$$

where: $\theta = [\theta_1 ... \theta_h ... \theta_H]$ and $\theta_h = [\theta_{h,1} ... \theta_{h,a}]$:

The marginal distribution of $\mathbf{X}_k$:

$$g(\mathbf{x}_k, \Theta_x) = \int_R f(y_k, \mathbf{x}_k, \Theta) dy_k = \sum_{h=1}^{H} p_h g_h(\mathbf{x}_k, \theta_{x,h}), \quad k \in U$$

where:
$g_h(\mathbf{x}_k, \theta_{x,h}) = \int_R f_h(y_k, \mathbf{x}_k, \theta_h) dy_k$,
$\theta_x = [\theta_{x,1} ... \theta_{x,H}]$, $\Theta_x = \{\theta_x, \mathbf{p}\}$.

Model-design approach based on distribution mixture

Sampling design

- A sample s of size $n \leq N$ is selected from population U according to a sampling design: $P(s) \geq 0$, $s \in \mathcal{S}$, $\sum_{s \in \mathcal{S}} P(s) = 1$ where $\mathcal{S}$ is sampling space;
- Inclusion probabilities of the sampling design: $\pi_k = \sum_{\{s: k \in s, s \in \mathcal{S}\}} P(s)$, $k = 1, ..., N$.
- $\underline{s} = U - s$ be the complement of s in U;
- $s = \bigcup_{h=1}^{H} s_h$, where $s_h \subseteq U_h$, n_h is the size of s_h, $n = \sum_{h=1}^{H} n_h$ is size of s, $1 < n_h \leq N_h$, $h = 1, ...H$.

Model-design approach based on distribution mixture

Estimated parameters

The main aim is to estimate:

- $\mu_h = E(Y_k|\mathbf{Z}_{k*} = \mathbf{z}_{k*}^{(h)})$ - domain mean for $k \in U_h$,
- p_h, $h = 1, ..., H$ - probabilities,
- $\mu = \sum_{h=1}^{H} p_h \mu_h$, population mean

Maximum likelihood estimation

Likelihood function

When the sample is selected according to preassigned inclusion probabilities, the pseudo-likelihood approach (see, [6], [10], [12]) leads to the following log-likelihood function:

$$l(\mathbf{d}_s, \mathbf{x}_{\underline{s}}) = l_1(\mathbf{d}_s) + l_2(\mathbf{x}_{\underline{s}}), \quad \mathbf{d}_s = \{d_k, k \in s\}$$

where the complete and incomplete functions are:

$$\begin{cases} l_1(\mathbf{d}_s) = \sum_{h=1}^{H} ln(p_h) \sum_{k \in s_h} \frac{1}{\pi_k} + \sum_{h=1}^{H} \sum_{k \in s_h} \frac{ln(f_h(y_k, \mathbf{x}_k, \theta_h))}{\pi_k}, \\ \\ l_2(\mathbf{x}_{\underline{s}}) = \sum_{k \in \underline{s}} \frac{ln(g(\mathbf{x}_k, \Theta_x))}{1 - \pi_k}. \end{cases}$$

EM-algorithm leads (see [3]-[5]) to replacing $l(\mathbf{d}_s, \mathbf{x}_{\underline{s}})$ with:

$$l^{(t)}(\mathbf{d}_s, \mathbf{x}_{\underline{s}}) = l_1(\mathbf{d}_s) + l_2^{(t)}(\mathbf{x}_{\underline{s}})$$

where $t = 0, 1, 2, \ldots$ - iterations,

$$l_2^{(t)}(\mathbf{x}_{\underline{s}}) = \sum_{h=1}^{H} \tau_h^{(t)} ln(p_h) + \sum_{h=1}^{H} \sum_{k \in \underline{s}} \frac{\tau_{h,k}^{(t)} ln(g_h(\mathbf{x}_k, \theta_{x,h}))}{1 - \pi_k},$$

$$\begin{cases} \hat{\tau}_h^{(t)} = \hat{\tau}_h(\hat{\Theta}_x^{(t)}) = \sum_{k \in \underline{s}} \frac{\tau_{h,k}^{(t)}}{1-\pi_k}, \\ \tau_{h,k}^{(t)} = \tau_h(\mathbf{x}_k, \hat{\Theta}_x^{(t)}) = \frac{p_h g_h(\mathbf{x}_k, \hat{\theta}_x^{(t)})}{g(\mathbf{x}_k, \hat{\Theta}_x^{(t)})}, \quad g(\ldots) = \sum_{h=1}^{H} p_h g_h(\ldots) \end{cases}$$

$\hat{\tau}_{h,k}^{(t)}$ is the posterior probability that the k-element ($k \in \underline{s}$) belongs to the h-th domain.

Maximum likelihood estimation

EM-algorithm

EM-algorithm provides approximated parameters $\hat{\Theta}^{(t+1)}$ and:

$$\hat{p}_h^{(t+1)} = \frac{\hat{N}_h + \hat{\tau}_h^{(t)}}{\hat{N} + \hat{\tau}^{(t)}}, \qquad h = 1, ..., H.$$

where

$$\hat{N}_h = \sum_{k \in s_h} \frac{1}{\pi_k}, \qquad \hat{N} = \sum_{h=1}^{H} \hat{N}_h = \sum_{k \in s} \frac{1}{\pi_k}, \qquad \hat{\tau}^{(t)} = \sum_{h=1}^{H} \hat{\tau}_h^{(t)}.$$

Statistics $\hat{N}$ and $\hat{\tau}^{(t)}$ are estimators of N.
$\tilde{N}_h^{(t)} = N\hat{p}_h^{(t)}$ estimates the expected values of the domain size;

In the case simple random sample drawn without replacement:

$$\hat{p}_h^{(t+1)} = \frac{1}{2}(\bar{p}_h + \bar{\tau}_h), \quad \bar{p}_h = \frac{n_h}{n}, \quad \hat{\tau}_h^{(t)} = \frac{1}{N-n} \sum_{k \in \underline{s}} \tau_{h,k}^{(t)}.$$

Maximum likelihood estimation

Bivariate normal model $N(\mu_{y,h}, \mu_{x,h}, \sigma^2_{y,h}, , \sigma^2_{x,h}, \rho_h)$, $h = 1, ..., H$

Regression type estimators of $\mu_{y,h}$:

$$\hat{y}_h^{(t+1)} = \bar{y}_{s_h} - \frac{\sigma_{xy,s_h}}{\hat{\sigma}^{2(t+1)}_{x,h}}(\bar{x}_{s_h} - \hat{x}_h^{(t+1)}),$$

$$\tilde{y}_h^{(t+1)} = \bar{y}_{s_h} - \frac{\sigma_{xy,s_h}}{\sigma^2_{x,s_h}}(\bar{x}_{s_h} - \hat{x}_h^{(t+1)}).$$

Ratio type estimator of $\mu_{y,h}$:

$$\check{y}_h^{(t+1)} = \bar{y}_{s_h}\frac{\hat{x}_h^{(t+1)}}{\bar{x}_{s_h}}$$

$t = 0, 1, 2, ...$

Maximum likelihood estimation

Bivariate normal model

Estimators based on the data observed in the sample s:

$$\bar{x}_{s_h} = \frac{1}{\hat{N}_h} \sum_{k \in s_h} \frac{x_k}{\pi_k}, \quad \bar{y}_{s_h} = \frac{1}{\hat{N}_h} \sum_{k \in s_h} \frac{y_k}{\pi_k}, \quad \hat{N}_h = \sum_{k \in s_h} \frac{1}{\pi_k},$$

$$\sigma^2_{x,s_h} = \frac{1}{\hat{N}_h} \sum_{k \in s_h} \frac{(x_k - \bar{x}_{s_h})^2}{\pi_k}, \quad \sigma^2_{y,s_h} = \frac{1}{\hat{N}_h} \sum_{k \in s_h} \frac{(x_k - \bar{y}_{s_h})^2}{\pi_k},$$

$$\sigma_{xy,s_h} = \frac{1}{\hat{N}_h} \sum_{k \in s_h} \frac{(x_k - \bar{x}_{s_h})(y_k - \bar{y}_{s_h})}{\pi_k}, \quad .$$

In the case of the simple random sample $\pi_k = \frac{n}{N}$ for all $k \in U$.

Maximum likelihood estimation

Bivariate normal model

Estimators based on $x_k \in U - s$:

$$\hat{x}_h^{(t+1)} = w^{(t)} \bar{x}_{s_h} + (1 - w_h^{(t)}) \bar{x}_{\underline{s},h}^{(t)},$$

$$\bar{x}_{\underline{s},h}^{(t)} = \frac{1}{\tau_h^{(t)}} \sum_{k \in \underline{s}} x_k \frac{\tau_{h,k}^{(t)}}{1 - \pi_k}, \qquad w^{(t)} = \frac{\hat{N}_h}{\hat{N}_h + \tau_h^{(t)}}, \qquad \bar{x}_h^{(0)} = \bar{x}_{s_h},$$

$$\hat{\sigma}_{x,h}^{2(t+1)} = w_h^{(t)} \sigma_{x,s_h}^2 + (1 - w_h^{(t)}) \sigma_{x,\underline{s},h}^{2(t)}, \qquad \hat{\sigma}_{x,h}^{2(0)} = \sigma_{x,s_h}^2,$$

$$\sigma_{x,\underline{s},h}^{2(t)} = \frac{1}{\tau_h^{(t)}} \sum_{k \in \underline{s}} \frac{(x_k - \bar{x}_{\underline{s},h}^{(t)})^2}{1 - \pi_k} \tau_{h,k}^{(t)},$$

In the case of the simple random sample $\pi_k = \frac{n}{N}$ for all $k \in U$.

Simulation analysis of the estimation accuracy

Description of the experiment

- Simple random samples $\{s_j, j = 1, ..., M\}$ are independently drawn without replacement from U,
- each s_j is partitioned among the domains in such a way that $s_j = s_{1,j} \cup ... \cup s_{h,j} \cup ... \cup s_{H,j}$ and $2 \leq n_h \leq n - 2(H-1)$, $h = 1, ..., H$.
- relative efficiency coefficient:

$$e(t_{s_h}) = \frac{mse(t_{s_h})}{v(\bar{y}_{s_h})} 100\%, \quad mse(t_{s_h}) = \frac{1}{M} \sum_{j=1}^{M} (t_{s_{h,j}} - \bar{y}_h)^2$$

$v(\bar{y}_{s_h}) = \frac{1}{M} \sum_{j=1}^{M} (\bar{y}_{s_{h,j}} - \bar{y}_h)^2$, $\bar{y}_h = \frac{1}{N_h} \sum_{j=1}^{M} y_{k,i}$,

- the relative bias:

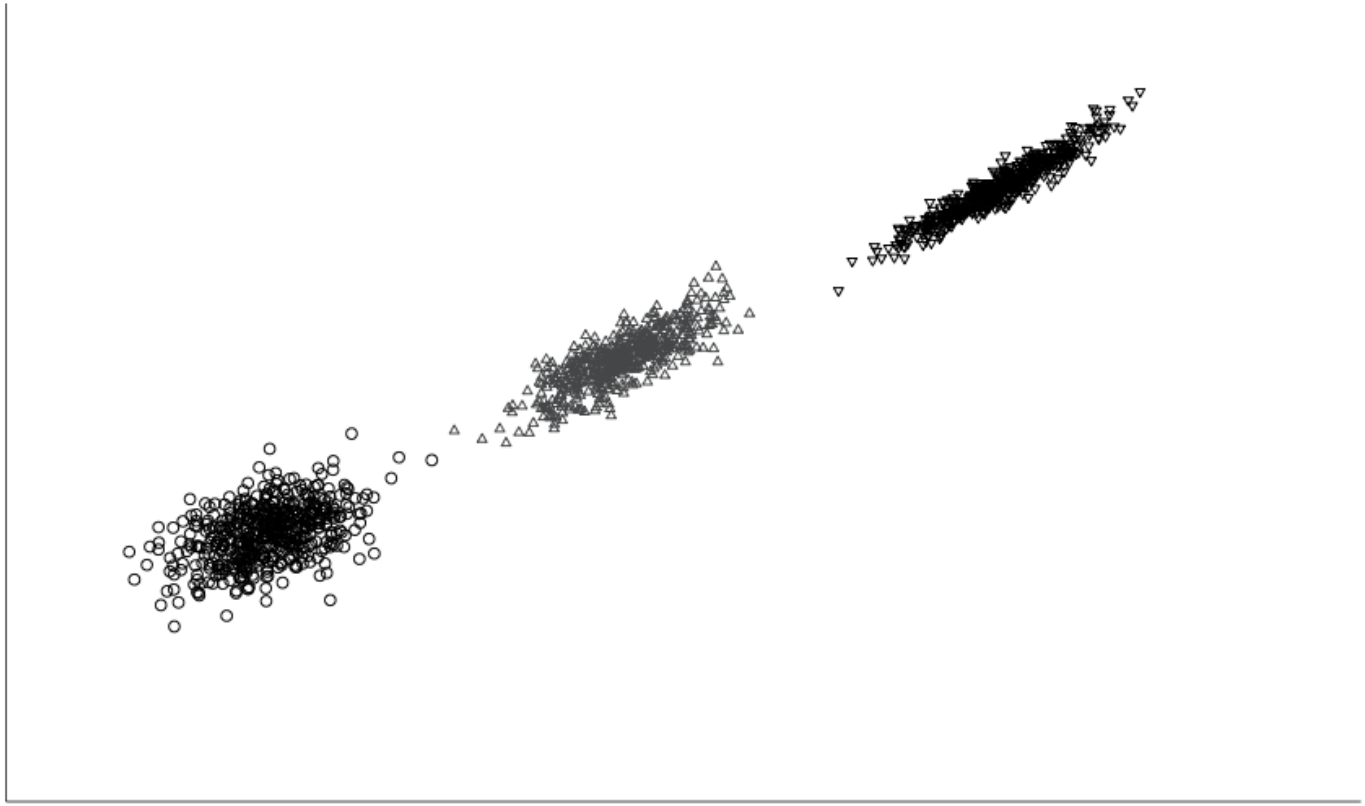
$$b(t_{s_h}) = \frac{|\bar{t}_{s_h} - \bar{y}_h|}{\sqrt{mse(t_{s_h})}} 100\%, \qquad \bar{t}_{s_h} = \frac{1}{M} \sum_{j=1}^{M} t_{s_{h,j}}, \quad h = 1, ..., H.$$

We assume that $M = 10000$.

Simulation analysis of the estimation accuracy

Spread of data generated from normal distribution mixture

$N_h = 500$, $p_h = 1/3$, $h = 1, 2, 3$;
$N(8, 4, 1, 1, 0.5)$, $N(14, 11.2, 1, 1, 0.8)$ and $N(20, 19, 1, 1, 0.95)$.



Simulation analysis of the estimation accuracy

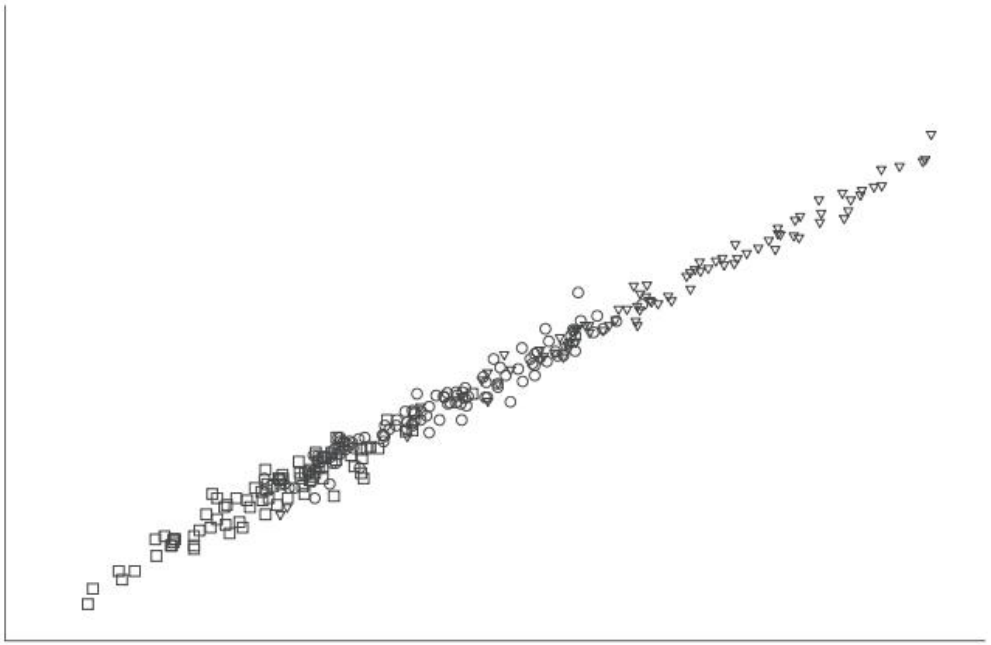
Table 1. Relative efficiency coefficients. Artificial data

t_{s_h}	n:	15	45	75	150
$\hat{y}_h^{(t)}$	1	94.6	88.9	86.2	86.2
	2	73.3	57.2	52.4	48.2
	3	64.2	44.5	35.9	29.5
$\tilde{y}_h^{(t)}$	1	286	202	131	98.9
	2	115	76.7	58.0	43.3
	3	599	405	159	40.6
$\check{y}_h^{(t)}$	1	241	205	135	97.4
	2	107	76.0	57.4	47.8
	3	569	406	158	40.7
$\hat{p}_h^{(t)}$	1	101	33.3	31.9	21.3
	2	127	59.0	38.3	21.5
	3	104	45.5	30.5	20.7
$\hat{y}^{(t)}$		83.3	37.3	27.8	20.9
$\tilde{y}^{(t)}$		45.3	28.4	24.8	20.7
$\check{y}^{(t)}$		41.7	26.7	24.6	20.6

Simulation analysis of the estimation accuracy

Spread of logarithmized data on Swedish municipalities [9]

Observations of y - municipal taxation revenues in 1985 and x - municipal employees in 1984 are divided by 30% and 70% quantiles of 1984 real estate value. $N_1 = N_3 = 86$, $N_2 = 109$.



Simulation analysis of the estimation accuracy of population mean

The population of 281 Swedish municipal units.

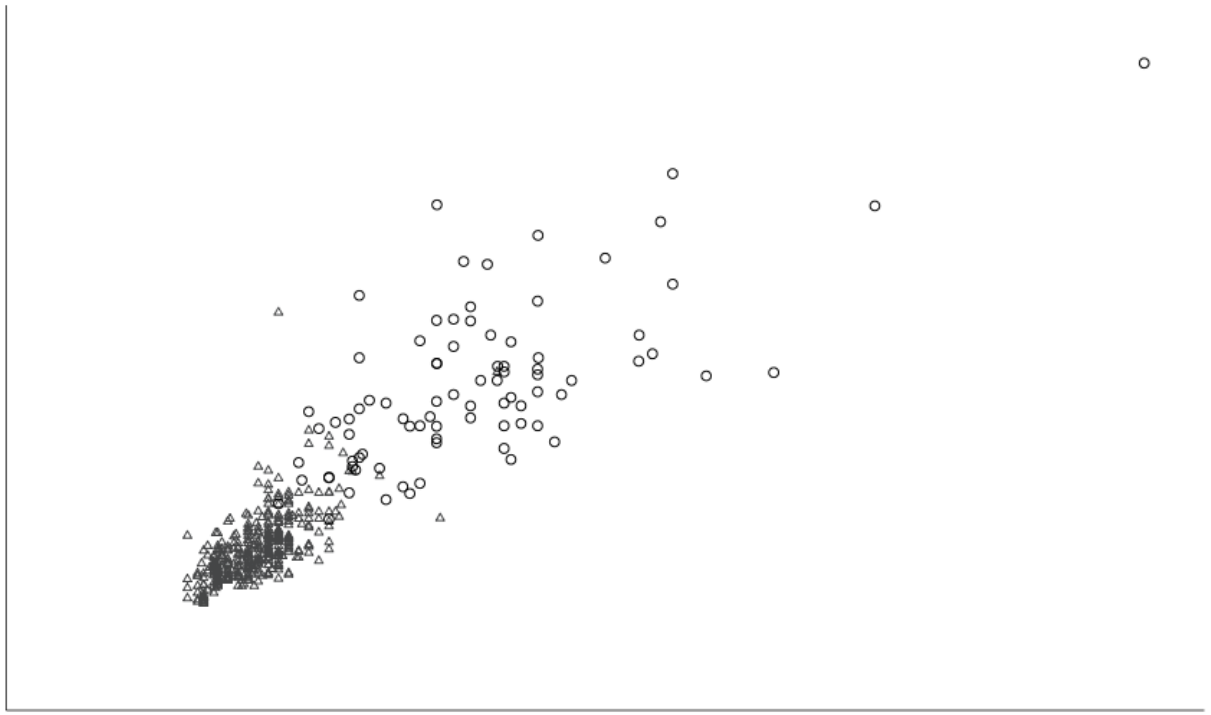
n:	8	14	28
1	2	3	4
e(.)			
$\hat{\bar{y}}^{(t)}$	85.1	58.4	46.2
$\tilde{\bar{y}}^{(t)}$	30.0	24.9	21.8
$\check{\bar{y}}^{(t)}$	49.0	43.0	43.9
b(.)			
$\hat{\bar{y}}^{(t)}$	8.0	7.9	5.2
$\tilde{\bar{y}}^{(t)}$	1.7	1.0	3.3
$\check{\bar{y}}^{(t)}$	31.6	18.0	9.4

Source: Own calculations.

Simulation analysis of the estimation accuracy

Data on current and starting salaries. Sourse: the SPSS dataset.

$N_1 = 390$ observations from officials and $N_2 = 84$ observations from managers.



Simulation analysis of the estimation accuracy

Data on current and starting salaries. Sourse: the SPSS dataset.

Table 3. Estimation accuracy of the population mean. Population of employees.

n:	15	24	48
1	2	3	4
e(.)			
$\hat{y}^{(t)}$	85.7	86.6	114
$\tilde{y}^{(t)}$	122	38.2	60.9
$\check{y}^{(t)}$	41.3	36.0	32.3
b(.)			
$\hat{y}^{(t)}$	47.0	51.7	78.9
$\tilde{y}^{(t)}$	50.9	15.1	4.0
$\check{y}^{(t)}$	20.0	10.3	10.5

Source: Own calculations.

Conclusions

- To estimate domain means in a finite population, the model-design approach was considered;
- The problem was considered as the estimation of the mean components of a mixture of probability distributions.
- In the case of a mixture of normal distributions, regression and ratio type estimators were derived.
- Example simulation analyses of the estimation accuracy showed that the proposed estimators of population means were more accurate than the simple sample mean. This was not always the case for the estimation of domain means.

Bibliography

[1] Chambers R.L., Skinner C.J. (2003). *Analysis of Survey Data*. Wiley, Chichester.
[2] Chen J., Sitter R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary informatio in complex surveys. *Statistica Sinica* 9, 385-406.

[3] Dempster A.P., Laird N.M., Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society* B, vol. 39.
[4] McLachlan G., Krishnan T. (1977). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc. New York.
[5] McLachlan G., Peel D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc. New York.
[6] Pfeffermann D. (1993). The role of sampling weights when modelling survey data. *International Statistical Review*, vol. 61.
[7] Skinner C. J., Holt D., Smith T. M. F. (1989). *Analysis of Complex Surveys.* New York Jahn Wiley.

Bibliography

[8] Rao J.N.K., Molina I. (2015). *Small Area Estimation*. John Wiley & Sons. Inc., Hoboken, New Jersey.
[9] Särndal C.E., Swensson B., Wretman J. (1992). *Model Assisted Survey Sampling* Springer-Verlag, New York.
[10] Thompson M.E. (1997). *Theory of Sample Surveys*. Chapman & Hall, London.
[11] Wywiał J. (2003). Some Contributions to Multivariate Methods in Survey Sampling. Katowice Univ. Economics *https* : *//www.sbc.org.pl/dlibra/publication/*706700.
[13] Wywiał J. (2023). **On the maximum likelihood estimation of population and domain means. *Journal of Statistical Theory and Prctice* 17, pp. 19**
https://link.springer.com/article/10.1007/s42519-023-00337-4
[14] Żądło T. (2006). On prediction of total value in incompletely specified domains. *Australian and New Zelland Journal of Statistics* vol. 48(3), 269-283.