For *4th Congress of Polish Statistics*, Warsaw, Poland; 2–4 July 2024.

Relaxed calibration of sampling weights

Nicholas T. Longford

SNTL Statistics Research and Consulting, London, UK

(Mohn Centre for Children's Health and Wellbeing, Imperial College London)

sntlnick@sntl.co.uk

Keywords:

Adjustment of sampling weights; auxiliary information; calibration (benchmarking); population (national) surveys.

To appear in *Survey Methodology* **50**, Dec. 2024.

Literature

Deville and Särndal (1992) — Särndal (2007)

Lohr and Raghunathan (2017), Devaud and Tillé (2019)

Monograph by Tillé (2020)

Haziza and Beaumont (2017) — model-based approaches

Chambers (1996); Guggemos and Tillé (2010) — ridge regression

Some adaptations:

Cardot, Goga and Shehzad (2017); Dagdoug, Goga and Haziza (2023)

!! A similar problem in causal analysis (matching/balancing)

Calibration

A long-standing problem in population surveys & official statistics

Auxiliary information: known population totals of some variables

Calibration: adjusting the sampling weights so that:

- weighted sample totals agree with the known population totals
- total of the sample weights stays the same
- the weights are changed as little as possible
- the weights are not too dispersed (efficiency)

Hard calibration: no leeway for any discrepancies

Soft calibration: thresholds for the discrepancies

Solutions; old and new

Raking for discrete variables:

— Iteratively adjust the weights for one variable at a time

Quadratic programming: optimisation constrained by the thresholds

Problem: solution may not exist or may be unsatisfactory

Solution: discard some variables, change some thresholds

— *improvise* with a black box

Proposal: replace *tresholds/constraints* with *penalties*

— *simplicity:* noniterative solution/algorithm

— *transparency:* properties are easy to study/explore

— *optimality* in a well-defined sense

Notation and formalities

A realised population survey:

— focal variable y (values $\mathbf{y}$), weights $\mathbf{w}$ (estimator $\hat{\theta} = \mathbf{w}^\top \mathbf{y}$)

— other variables, vector $\mathbf{x} = (x_0 = 1, x_1, \ldots, x_K)$;

— data matrix $\mathbf{X}$ $[n \times (K+1)]$

— population totals, vector $\mathbf{t} = (t_0 = N, t_1, \ldots, t_K)$;

Calibration: adjusted weights $\mathbf{u} = C(\mathbf{w}; \mathbf{X}, \mathbf{t})$,

$$[\boldsymbol{\delta} =] \quad \mathbf{X}^\top \mathbf{u} - \mathbf{t} \doteq \mathbf{0}, \quad \text{i.e.,} \quad \delta_k = \mathbf{X}_k^\top \mathbf{u} - t_k \doteq 0, \quad 0 \le k \le K$$

subject to small $||\mathbf{u} - \mathbf{w}||$

small $\mathrm{var}(u)$

E.g., thresholds $D_k \ge 0$ on the discrepancies; $|\,\delta_k\,| \le D_k$

Motivation. Thresholds $\to$ penalties

Replace the constraints $\delta_k^2 \le D_k^2$

with a single constraint $\delta_0^2 + \delta_1^2 + \cdots + \delta_K^2 \le D$.

Minimise

$$\sum_{k=0}^{K} p_k \, \delta_k^2 \qquad \left(= \boldsymbol{\delta}^\top \mathbf{P} \boldsymbol{\delta} \right)$$

subject to constraints on efficiency and small change $(\mathbf{u} - \mathbf{w})^\top (\mathbf{u} - \mathbf{w})$

Priorities p_k, $0 \le k \le K$ to be set.

Next: minimise

$$F(\mathbf{u}; \mathbf{w}) = \sum_{k=0}^{K} p_k \, \delta_k^2 + R \, (\mathbf{u} - \mathbf{w})^\top (\mathbf{u} - \mathbf{w}) + S \left(\mathbf{u}^\top \mathbf{u} - \frac{1}{n} \mathbf{u}^\top \mathbf{1} \mathbf{1}^\top \mathbf{u} \right)$$

Unconstrained optimisation

Invariance ... we can assume that $R + S = 1$.

Quadratic objective function

$$F(\mathbf{u};\mathbf{w}) \;=\; \mathbf{u}^\top \mathbf{H}\mathbf{u} - 2\mathbf{u}^\top \mathbf{s} + E\,,$$

where

$$\begin{aligned} \mathbf{H} \;&=\; \mathbf{I}_n + \mathbf{X}\mathbf{P}\mathbf{X}^\top \\ \mathbf{s} \;&=\; R\mathbf{w} + (1-R)\,\frac{t_0}{n}\,\mathbf{1}_n + \mathbf{X}\mathbf{P}\mathbf{t}\,. \end{aligned}$$

Minimum: $\quad \mathbf{u}^* = \mathbf{H}^{-1}\mathbf{s}; \quad F(\mathbf{u}^*;\mathbf{w}) \;=\; E - \mathbf{s}^\top \mathbf{H}^{-1}\mathbf{s}.$

Minimum *always* exists and has a closed form.

... setting $\mathbf{P} = \mathrm{diag}(p_k)$ and R (*tuning* parameters)

$\mathbf{H}^{-1}$ and a link to ridge regression

$\mathbf{H} = \mathbf{I} + \mathbf{L}$, where $\mathbf{I}$ is easy to invert and $\operatorname{rank}(\mathbf{L}) \leq K+1$

$$\left(\mathbf{I} + \mathbf{X}\mathbf{P}\mathbf{X}^\top\right)^{-1} = \mathbf{I} - \mathbf{X}\left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$$

$$\left[\hat{\boldsymbol{\beta}} =\right] \left(\mathbf{P}^{-1} + \mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$

— affinity with (generalised) *ridge regression*

Alternative: A recursive algorithm for evaluating $\mathbf{H}^{-1}\mathbf{s}$,

— operating only with vectors of length n

Estimator of the population total:

$$\hat{\theta}(\mathbf{u}; R) = (1-R)\left\{\frac{t_0\,\bar{y}}{1+np_0} + \mathbf{t}^\top\hat{\boldsymbol{\beta}}\right\} + R\mathbf{w}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

Properties of $\delta_k(\mathbf{P}, R)$

$$\frac{\partial \delta_k^2}{\partial p_k} = -2\delta_k^2 \mathbf{X}_k^\top \mathbf{H}^{-1} \mathbf{X}_k \; < 0$$

— large δ_k^2 is easy to reduce; small δ_k^2 is difficult to reduce

— do not try to wipe out the last bit of discrepancy

$$\frac{\partial \delta_k}{\partial R} = \mathbf{X}_k^\top \mathbf{H}^{-1} \left(\mathbf{w} - \frac{t_0}{n} \mathbf{1}_n \right)$$

— linear dependence on R

Simple *micro-management* of individual δ_k

Examples and simulations:

— easy control of the discrepancies δ_k

Summary

Calibration as a routine operation

— with a unique closed-form solution

— that reflects the perspectives, judgements and priorities

— easy control of the discrepancies

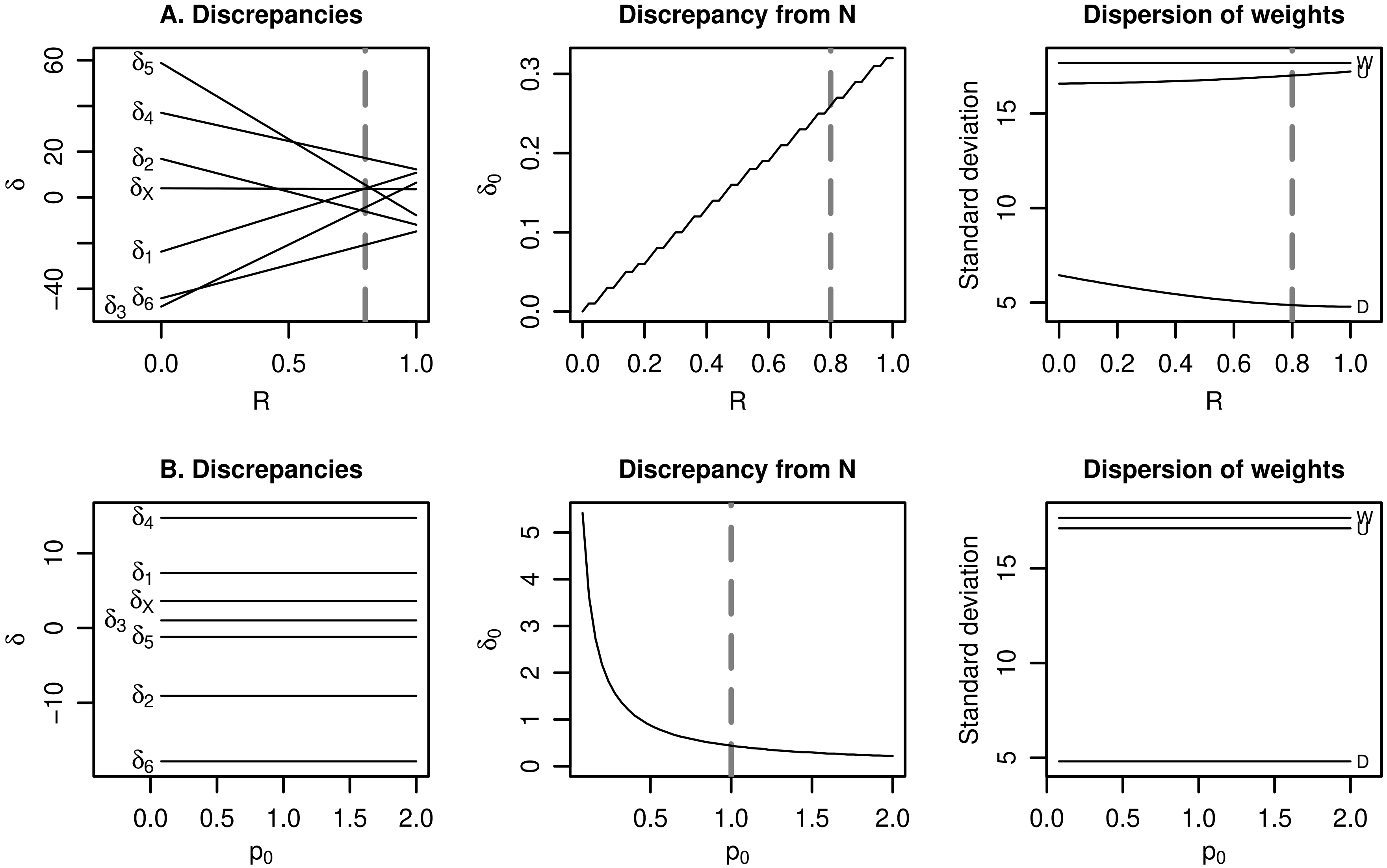
Old: Include/exclude in calibration

New: Set priorities for calibration; use *all* available information

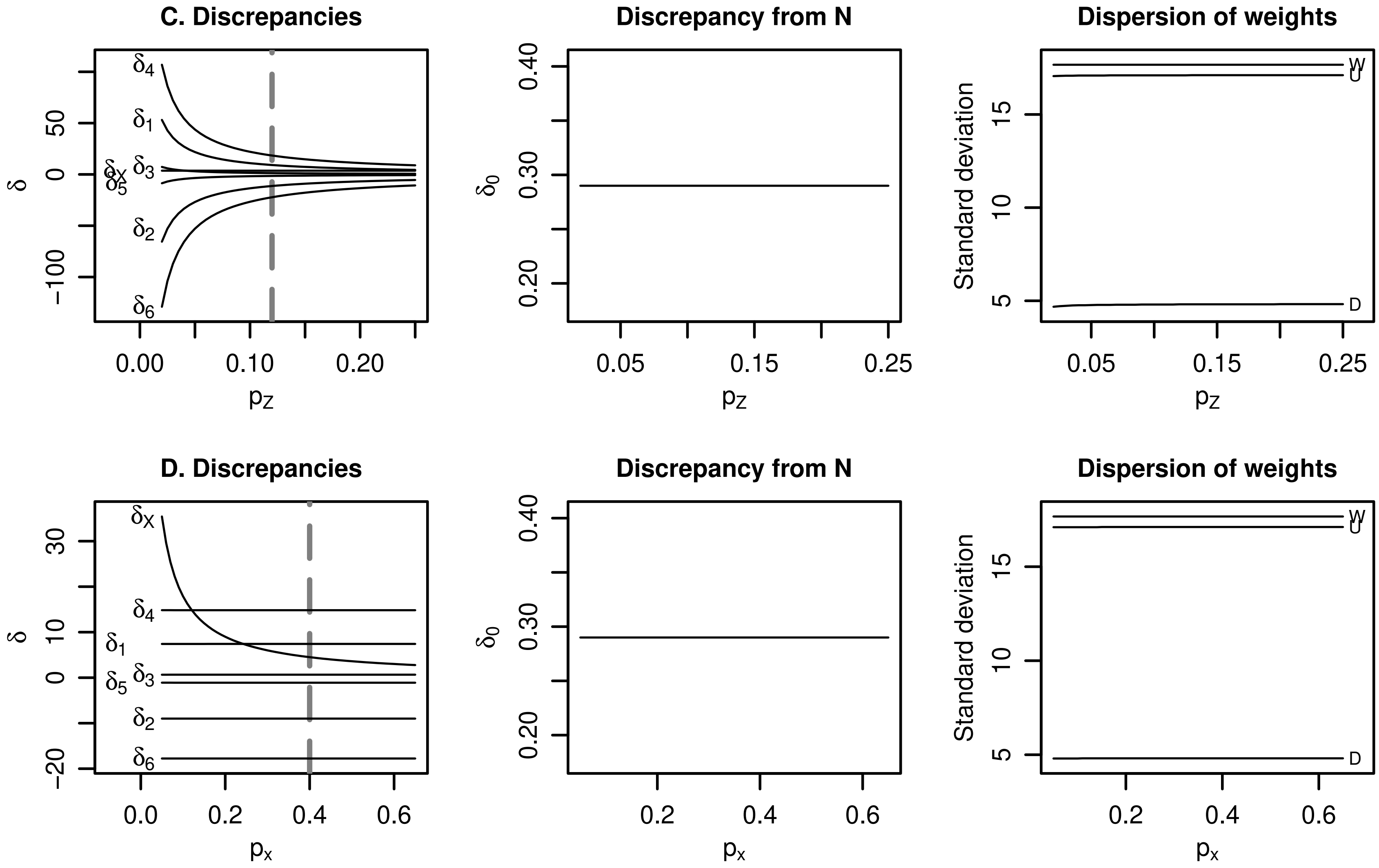
Analytical connection with balancing in causal analysis:

— matching the means of group A with *weighted* means of group B

Thank you — Dziękuję bardzo. *Questions? — Pytania?*



Dependence of $\boldsymbol{\delta}$, $\mathrm{var}(u)$, $\mathrm{var}(u-w)$ and R and p_0.



Dependence of $\boldsymbol{\delta}$, var(u), var$(u-w)$ and p_k , $k = 1, \ldots, K-1$ (for categories) and p_K (for a cont. variable).