Introduction to the Theory of Calibration

Yves Tillé

Institut de Statistique

University of Neuchâtel

Plan

- Reminder on the theory of sampling
- Regression estimator
- Calibration estimators, choice of the function of calibration
- General remarks
- What is a good calibration software?

Population and sample

- Finite population $U = \{1, 2, \ldots, k, \ldots, N\}$
- Variable of interest y.
- Values taken by the variable of interest on the population

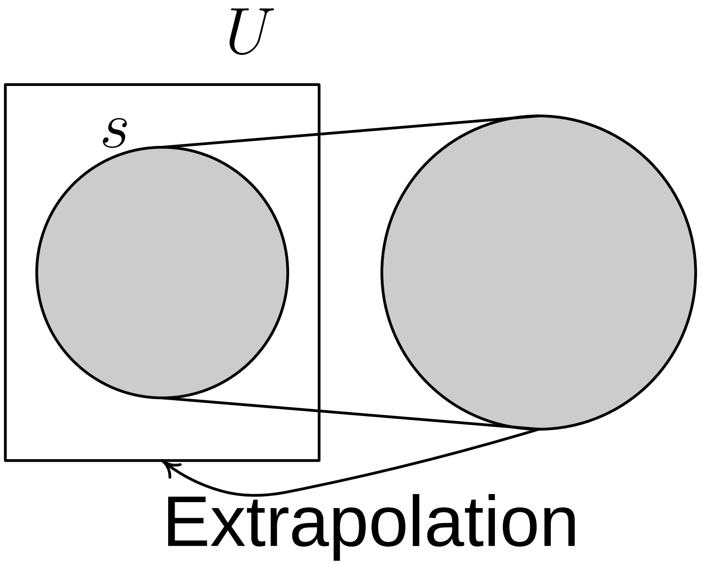
$$(y_1, \ldots, y_k, \ldots, y_N).$$

Functions of interest

- Total $Y = \sum_{k \in U} y_k$
- Mean $\overline{Y} = \frac{1}{N} \sum_{k \in U} y_k$
- Variance $S_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \overline{Y})^2$

Sample

- A sample s is a subset of the population U.



Sampling design

- A sampling design is a probability distribution on all the possible samples:

$$p(s) \geq 0, \text{ for all } s \subset U, \text{ and } \sum_{s \subset U} p(s) = 1.$$

- The random sample S is a random set such that $Pr(S = s) = p(s)$.

Basic estimation

- The inclusion probability $\pi_k, k \in U,$ can be derived from the sampling design

$$\pi_k = \sum_{s \ni k} p(s).$$

- Horvitz-Thompson estimator

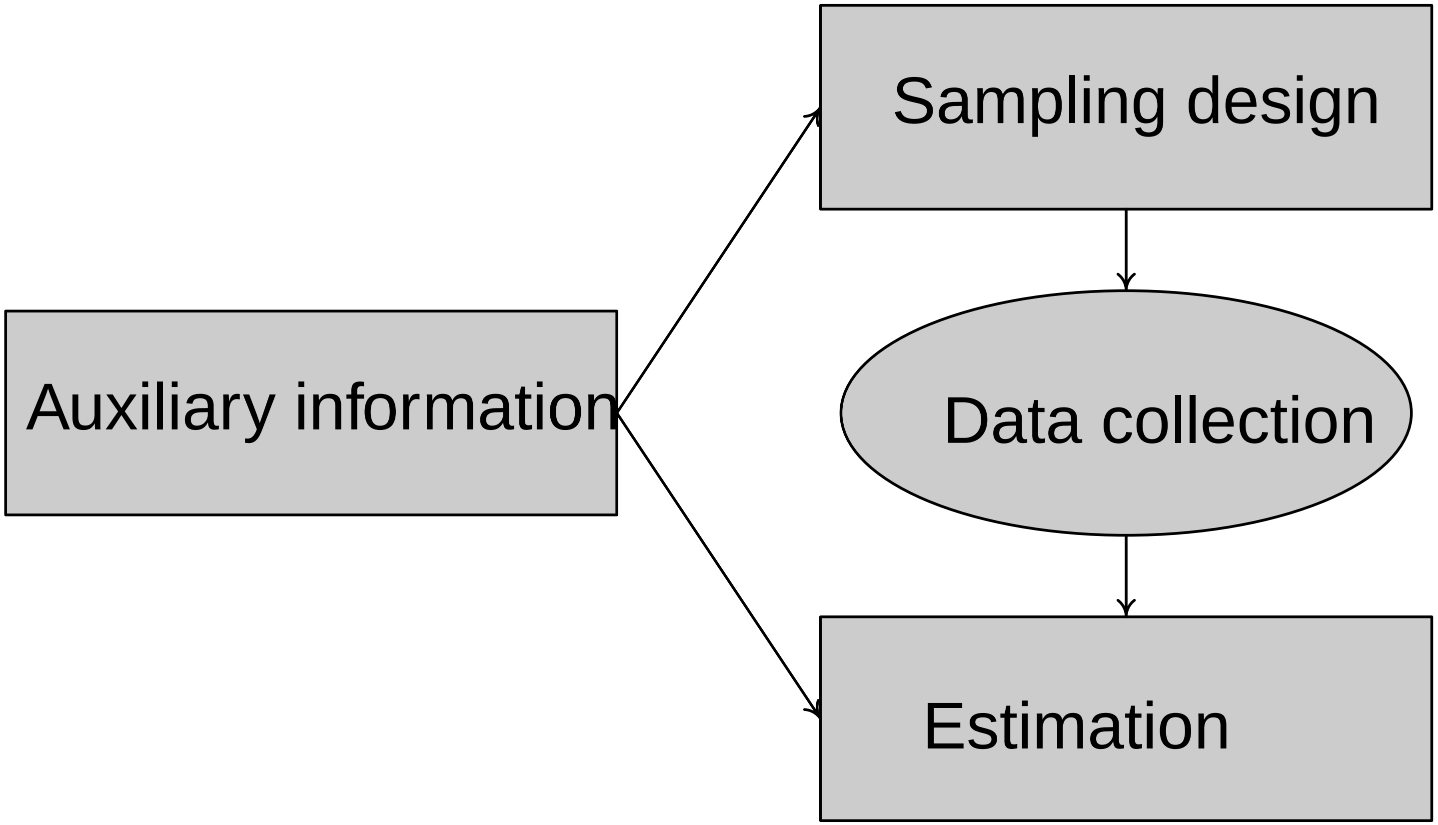
$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

- Unbiased if $\pi_k > 0$ for all $k \in U.$

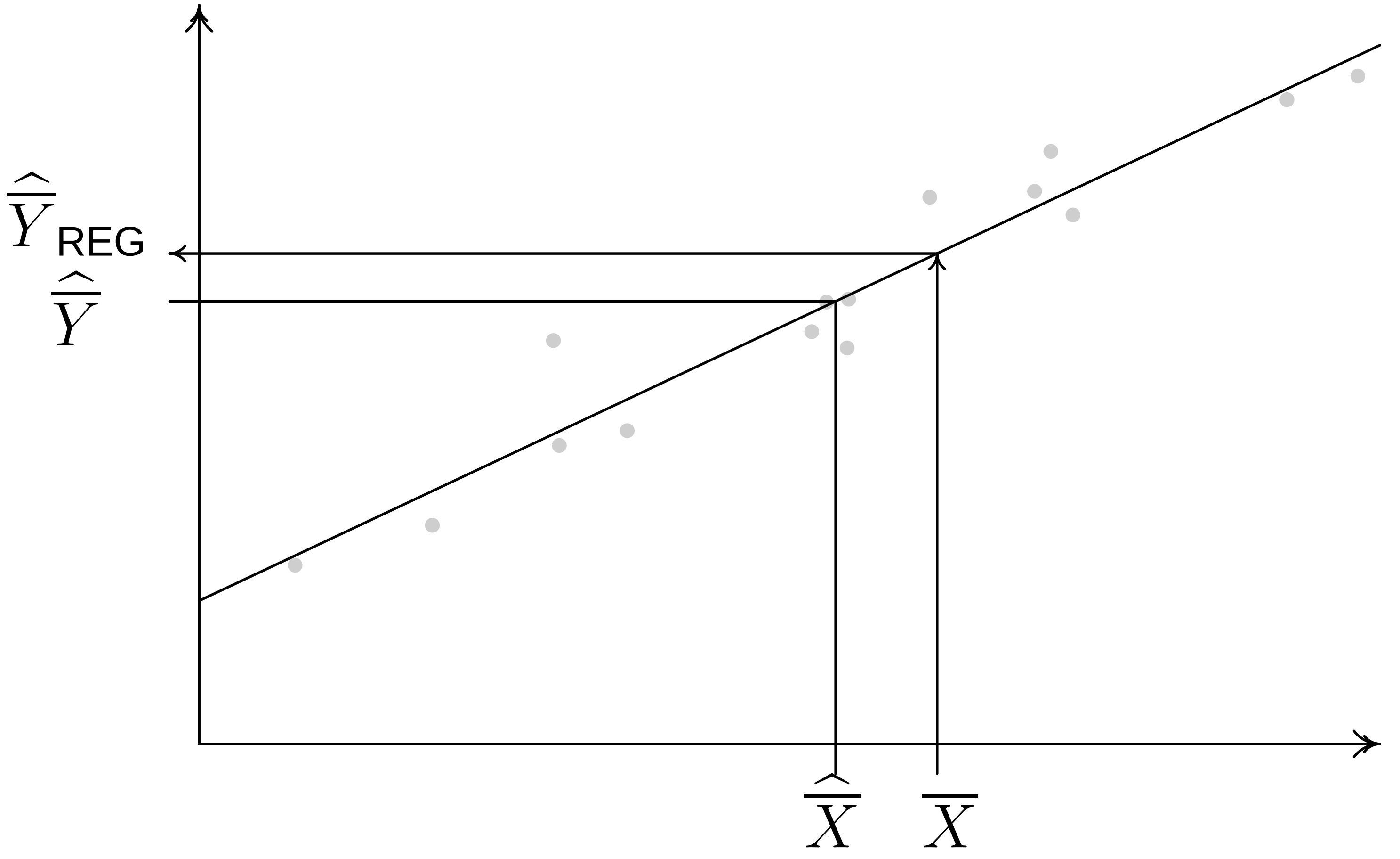
Auxiliary information

Auxiliary Information	**Interest Information**
Auxiliary variables	**Interest variables**
X	Y
known or partially known	unknown

Sampling and auxiliary information



Regression estimator



Regression estimator $\widehat{\overline{Y}}_{\text{REG}} = \widehat{\overline{Y}} + (\overline{X} - \widehat{\overline{X}})\widehat{b}$

Generalized regression estimator (GREG) (1)

- Multivariate auxiliary information given by the totals of p auxiliary variables $x_1, ..., x_p$.
- Vector $\mathbf{x_k} = (x_{k1}, ..., x_{kj}, ..., x_{kp})'$ of values taken by the p auxiliary variables on unit k.
- The total $\mathbf{X} = \sum_{k \in U} \mathbf{x_k}$, is assumed to be known.
- The aim is to estimate $Y = \sum_{k \in U} y_k$, using the information given by $\mathbf{X}$.

Generalized regression estimator (GREG) (2)

- GREG estimator: $\widehat{Y}_{\text{GREG}} = \widehat{Y} + (\mathbf{X} - \widehat{\mathbf{X}})'\widehat{\mathbf{b}}$
- $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$
- $\widehat{\mathbf{X}} = \sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k}$
- $\widehat{\mathbf{b}} = \left(\sum_{k \in S} \frac{q_k \mathbf{x}_k \mathbf{x}_k'}{\pi_k} \right)^{-1} \sum_{k \in S} \frac{q_k \mathbf{x}_k y_k}{\pi_k}$
- The q_k are weights.

Other presentation of the GREG

$$\widehat{Y}_{\mathsf{GREG}} = \sum_{k \in S} w_k y_k = \sum_{k \in S} \frac{g_k y_k}{\pi_k},$$

$$w_k = \frac{1}{\pi_k} \left\{ 1 + (\mathbf{X} - \widehat{\mathbf{X}})' \widehat{\mathbf{T}}^{-1} q_k \mathbf{x}_k \right\},$$

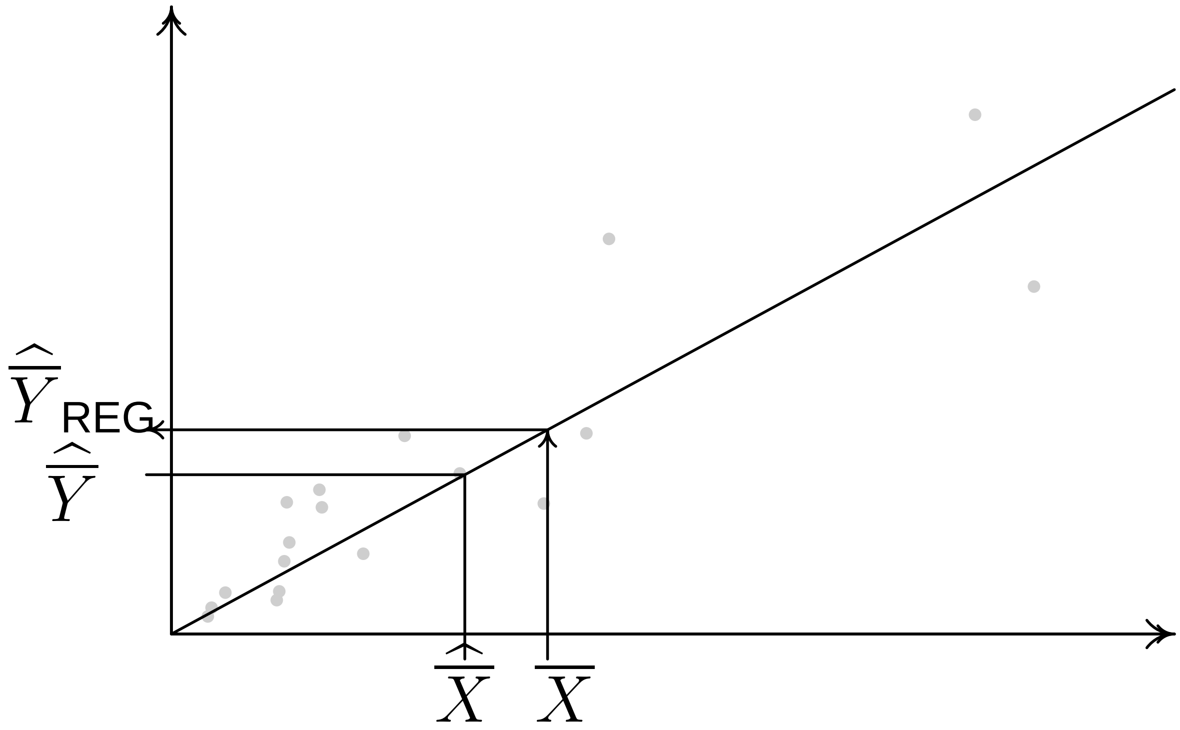
$$g_k = 1 + (\mathbf{X} - \widehat{\mathbf{X}})' \widehat{\mathbf{T}}^{-1} q_k \mathbf{x}_k \ ,$$

$$\widehat{\mathbf{T}} = \sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k' q_k}{\pi_k}.$$

PROBLEM: the weights can be negative.

Ratio estimator 1

If the regression line crosses the origin.



Ratio estimator 2

- Let $\mathbf{x}_k = x_k$ only one auxiliary variable, and $q_k = \frac{1}{x_k}$. Then
- $$\widehat{\mathbf{b}} = \left(\sum_{k \in S} \frac{q_k \mathbf{x}_k \mathbf{x}'_k}{\pi_k} \right)^{-1} \sum_{k \in S} \frac{q_k \mathbf{x}_k y_k}{\pi_k}$$
  $$= \left(\sum_{k \in S} \frac{x_k}{\pi_k} \right)^{-1} \sum_{k \in S} \frac{y_k}{\pi_k} = \frac{\widehat{Y}}{\widehat{X}}$$
- The regression estimator becomes the Ratio estimator
  $$\widehat{Y}_{\text{GREG}} = \widehat{Y} + (\mathbf{X} - \widehat{\mathbf{X}})'\widehat{\mathbf{b}} = \widehat{Y} + (X - \widehat{X})\frac{\widehat{Y}}{\widehat{X}} = \widehat{Y}\frac{X}{\widehat{X}}$$

Multivariate calibration

- Multivariate auxiliary information given by the totals of p auxiliary variables $x_1, ..., x_p$.
- Vector $\mathbf{x_k} = (x_{k1}, ..., x_{kj}, ..., x_{kp})'$ of values taken by the p auxiliary variables on unit k.
- The total

$$\mathbf{X} = \sum_{k \in U} \mathbf{x_k},$$

  is assumed to be known.
- The aim is to estimate $Y = \sum_{k \in U} y_k$, using the information given by $\mathbf{X}$.

Idea of calibration

- Horvitz-Thompson estimator $\sum_{k \in S} d_k y_k$,

  where $d_k = 1/\pi_k$.

- The idea consists of looking for new weights w_k as close as possible to d_k and such that

$$\sum_{k \in S} w_k \mathbf{x}_k = \mathbf{X} \text{ (calibration constraint).}$$

Pseudo-distance

- A pseudo-distance $G_k(.,.)$ between w_k and $d_k = 1/\pi_k$ is minimized,

$$\min_{w_k} \sum_{k \in S} \frac{G_k(w_k, d_k)}{q_k},$$

  under the constraints of calibration.
- $q_k, k \in S$, are strictly positive known coefficients.
- Function $G_k(.,.)$ is assumed to be strictly convex, positive and such that $G_k(d_k, d_k) = 0$.

Solution

- The weights w_k are then defined by

$$w_k = d_k F_k(\boldsymbol{\lambda}' \mathbf{x}_k),$$

where $d_k F_k(.)$ is the reciprocal of the function $G_k'(., d_k)/q_k$, with

$$G_k'(w_k, d_k) = \frac{\partial G_k(w_k, d_k)}{\partial w_k},$$

and $\boldsymbol{\lambda}$ is the Lagrange multiplier following from the constraints.

Identification of $\boldsymbol{\lambda}$

- The vector $\boldsymbol{\lambda}$ is obtained by solving the calibration equations:

$$\sum_{k\in S} w_k \mathbf{x}_k = \sum_{k\in S} d_k F_k(\boldsymbol{\lambda}' \mathbf{x}_k)\mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k.$$

  This system of equations can be non-linear (use of Newton method).

- Next the weights are computed $w_k = d_k F_k(\boldsymbol{\lambda}' \mathbf{x}_k)$.
- Finally, the calibrated estimator is $\widehat{Y}_{\text{CAL}} = \sum_{k\in S} w_k y_k$

Weights and g-weights

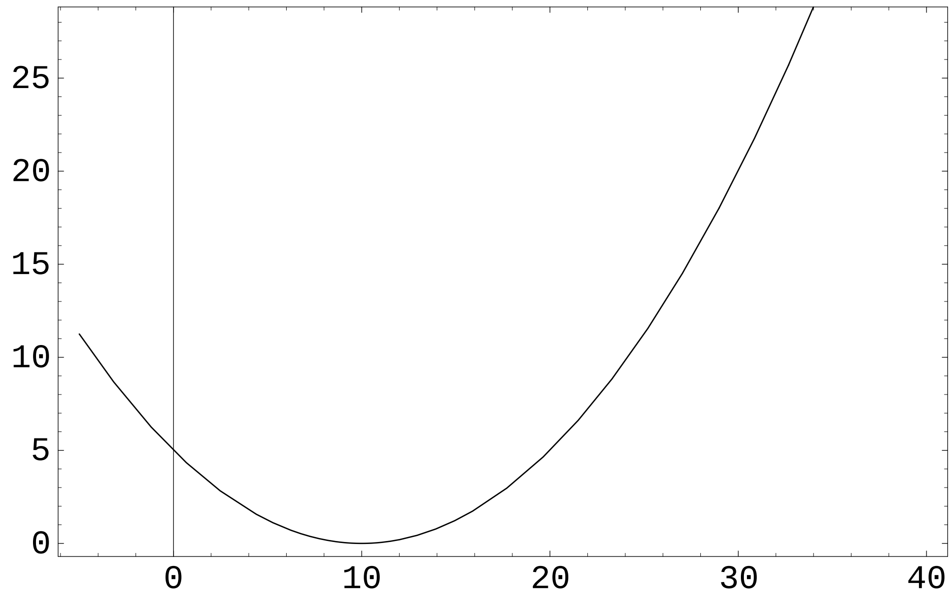
- The weights w_k are close to $d_k = 1/\pi_k$ and are such that $\sum_{k\in S} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k$
- g-weights $g_k = \pi_k w_k$ (close to 1).
- Horvitz-Thompson estimator $\widehat{Y} = \sum_{k\in S} \frac{y_k}{\pi_k}$
- Calibrated estimator $\widehat{Y}_C = \sum_{k\in S} w_k y_k = \sum_{k\in S} \frac{g_k y_k}{\pi_k}$.

The g_k are the distorsion of the weights with respect to the Horvitz-Thompson estimator.

Chi-square distance (1)

Suppose that $G_k(.,.)$ is chi-square function,

$$G_k(w_k, d_k) = \frac{(w_k - d_k)^2}{d_k},$$

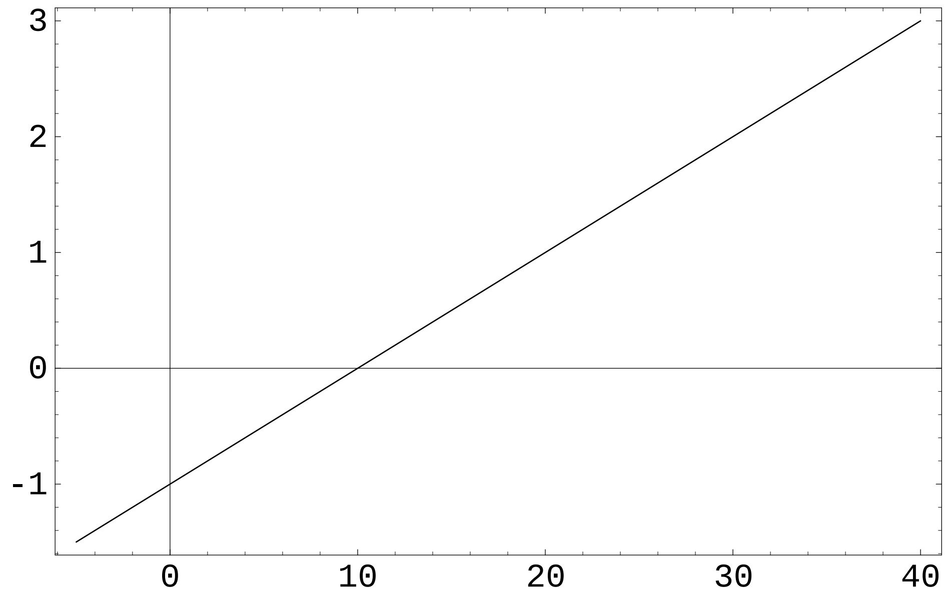


Linear method: function $G(w_k, d_k)$ with $q_k = 1$ and $d_k = 10$

Chi-square distance (2)

The derivative is

$$G'_k(w_k, d_k) = \frac{2(w_k - d_k)}{d_k},$$

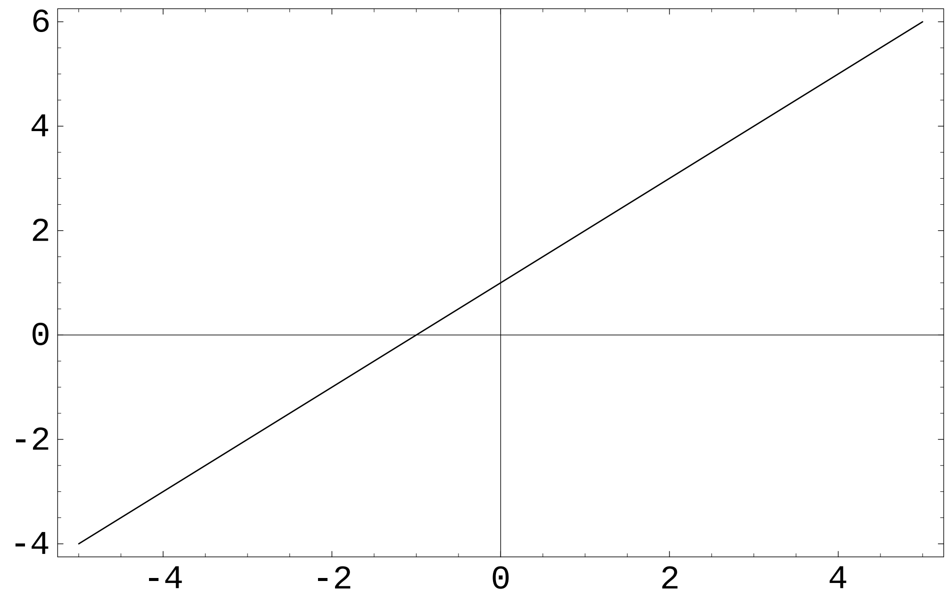


Linear method: function $G'(w_k, d_k)$ with $q_k = 1$ and $d_k = 10$

Chi-square distance (3)

- Calibration function

$$F_k(u) = 1 + q_k u.$$



Linear method: function $F_k(u)$ with $q_k = 1$

Chi-square distance (4)

- Weights $w_k = d_k F_k(u) = d_k(1 + q_k \boldsymbol{\lambda}' \mathbf{x}_k)$.
- The calibration equation is linear

$$\mathbf{X} = \widehat{\mathbf{X}} + \sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k' \boldsymbol{\lambda}$$

- Identification of $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda} = \left(\sum_{k \in S} d_k \mathbf{x}_k q_k \mathbf{x}_k' \right)^{-1} (\mathbf{X} - \widehat{\mathbf{X}})$$

Chi-square distance (5)

- Weights

$$
\begin{aligned}
w_k &= d_k F_k(u) = d_k(1 + q_k \boldsymbol{\lambda}' \mathbf{x}_k) \\
&= d_k \left[1 + q_k (\mathbf{X} - \widehat{\mathbf{X}})' \left(\sum_{k \in S} \frac{\mathbf{x}_k q_k \mathbf{x}_k'}{\pi_k} \right)^{-1} \mathbf{x}_k \right].
\end{aligned}
$$

Chi-square distance (6)

- The calibrated estimator is then equal to the generalized regression estimator which is

$$\widehat{Y}_{\mathrm{GREG}} = \widehat{Y} + (\mathbf{X} - \widehat{\mathbf{X}})'\widehat{\mathbf{b}},$$

where

$$\widehat{\mathbf{b}} = \widehat{\mathbf{T}}^{-1} \sum_{k \in S} \frac{\mathbf{x}_k y_k q_k}{\pi_k}.$$

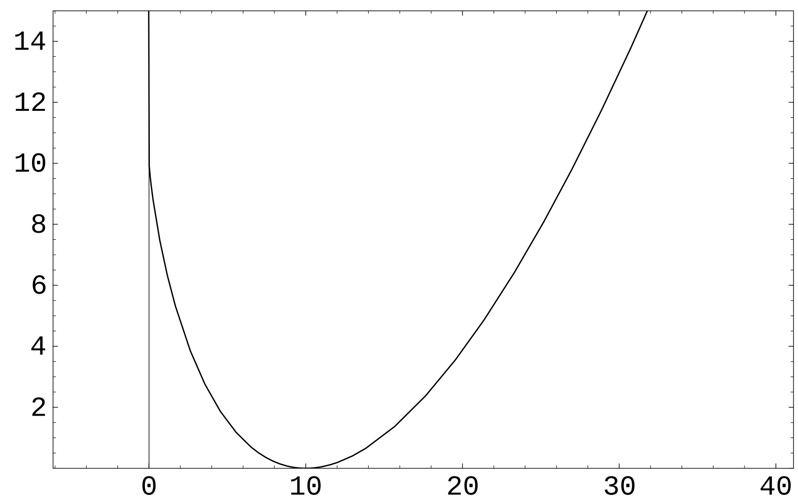
and

$$\widehat{\mathbf{T}} = \left(\sum_{k \in S} \frac{\mathbf{x}_k \mathbf{x}_k' q_k}{\pi_k} \right).$$

The raking ratio method: distance

Suppose that the distance is

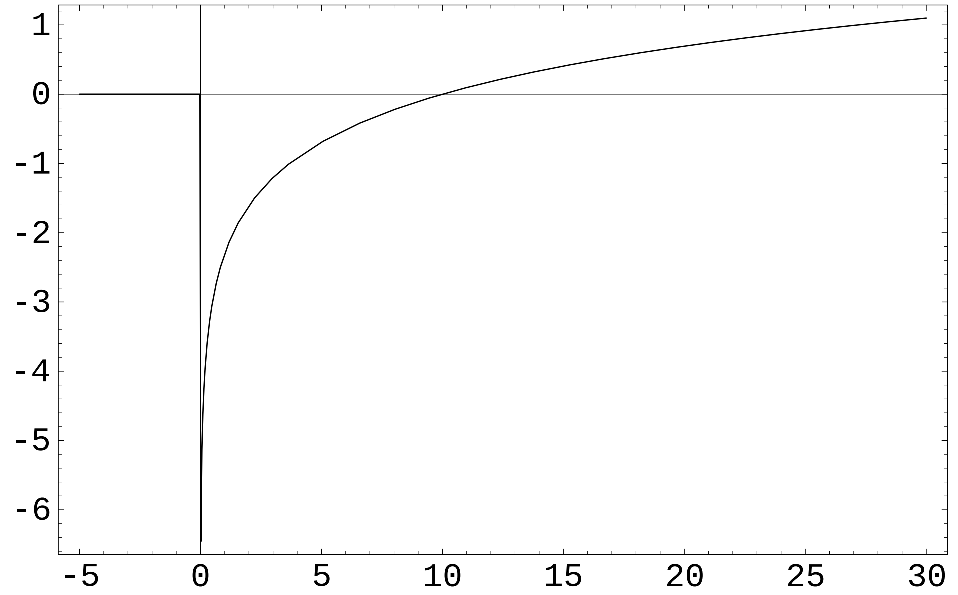
$$G(w_k, d_k) = w_k \log \frac{w_k}{d_k} + d_k - w_k.$$



“Raking ratio": function $G(w_k, = d_k)$ with $q_k = 1$ and $d_k = 10$

The derivative of the distance is

$$G'(w_k, d_k) = \log \frac{w_k}{d_k},$$

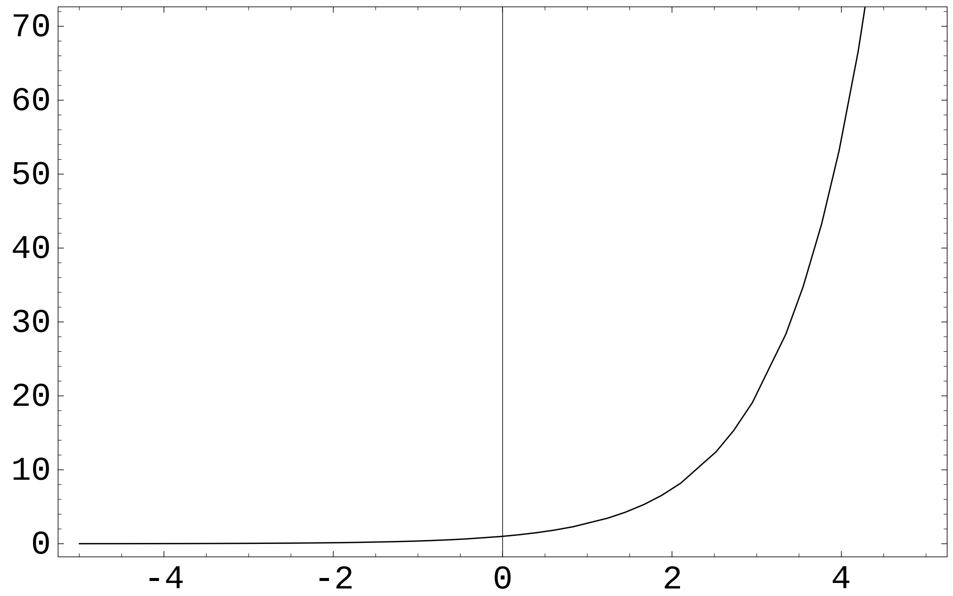


“Raking ratio": function $G'(w_k, d_k)$ with $q_k = 1$ and $d_k = 10$

The raking ratio method: calibration

The calibration function is

$$F_k(u) = \exp q_k u.$$



“Raking ratio": function $F_k(u)$ with $q_k = 1$

ADVANTAGE: The weights are always positive.

Marginal calibration(1)

Adjust the following table

80	170	150	400
90	80	210	380
10	80	130	220
180	330	490	1000

to the marginal column $(430, 360, 210)$,
and the marginal row $(150, 300, 550)$.

Calibration by row: iteration 1			
86.00	182.75	161.25	430.00
85.26	75.79	198.95	360.00
9.55	76.36	124.09	210.00
180.81	334.90	484.29	1000.00

Marginal calibration(2)

Calibration by column: iteration 2			
71.35	163.70	183.13	418.18
70.73	67.89	225.94	364.57
7.92	68.41	140.93	217.25
150.00	300.00	550.00	1000.00

Calibration by row: iteration 3			
73.36	168.33	188.31	430.00
69.85	67.04	223.11	360.00
7.65	66.12	136.22	210.00
150.87	301.49	547.64	1000.00

Marginal calibration(3)

Calibration by column: iteration 4			
72.94	167.50	189.12	429.56
69.45	66.71	224.07	360.23
7.61	65.79	136.81	210.22
150.00	300.00	550.00	1000.00

Calibration by row: iteration 5			
73.02	167.67	189.31	430.00
69.40	66.67	223.93	360.00
7.60	65.73	136.67	210.00
150.02	300.06	549.91	1000.00

Marginal calibration(4)

Calibration by column: iteration 6			
73.01	167.64	189.34	429.98
69.39	66.65	223.97	360.01
7.60	65.71	136.69	210.01
150.00	300.00	550.00	1000.00

Calibration by row: iteration 7			
73.01	167.64	189.35	430.00
69.39	66.65	223.96	360.00
7.60	65.71	136.69	210.00
150.00	300.00	550.00	1000.00

Marginal calibration(5)

Calibration by column: iteration 8			
73.01	167.64	189.35	430.00
69.39	66.65	223.96	360.00
7.60	65.71	136.69	210.00
150.00	300.00	550.00	1000.00

After 8 iterations, the adjustment is very accurate.

Pseudo-distances

α	$G^{\alpha}(w_k, d_k)$	$g^{\alpha}(w_k, d_k)$	$F_k^{\alpha}(u)$	$Type$
2	$\frac{(w_k - d_k)^2}{2d_k}$	$\frac{w_k}{d_k} - 1$	$1 + q_k u$	Chi-square
1	$w_k \log \frac{w_k}{d_k} + d_k - w_k$	$\log \frac{w_k}{d_k}$	$\exp(q_k u)$	Entropy
$1/2$	$2(\sqrt{w_k} - \sqrt{d_k})^2$	$2\left(1 - \sqrt{\frac{d_k}{w_k}}\right)$	$(1 - q_k u/2)^{-2}$	Hellinger Distance
0	$d_k \log \frac{d_k}{w_k} + w_k - d_k$	$1 - \frac{d_k}{w_k}$	$(1 - q_k u)^{-1}$	Inverse Entropy
-1	$\frac{(w_k - d_k)^2}{2w_k}$	$\left(1 - \frac{d_k^2}{w_k^2}\right)/2$	$(1 - 2q_k u)^{-1/2}$	Inverse Chi-square

Logistic function: distance

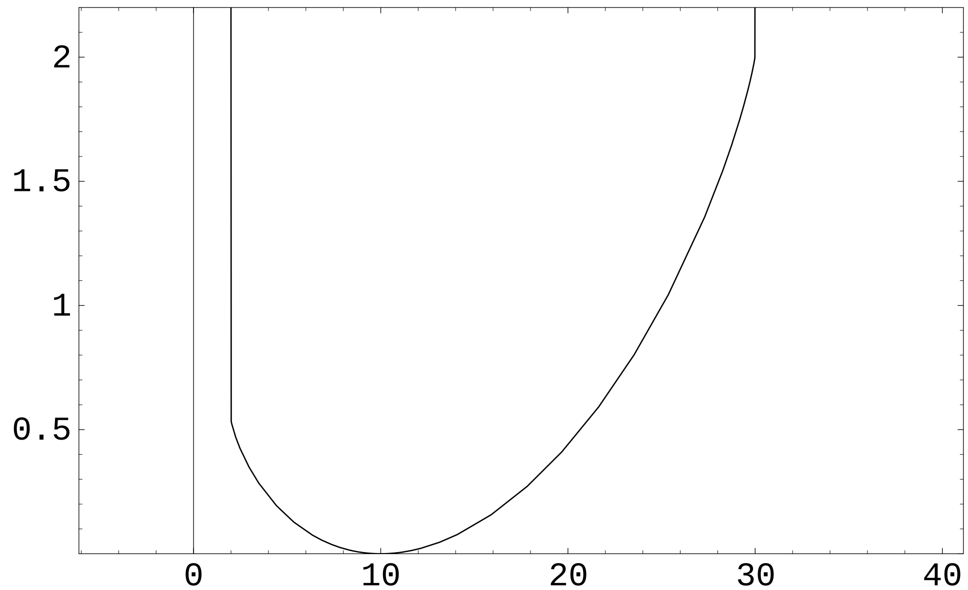
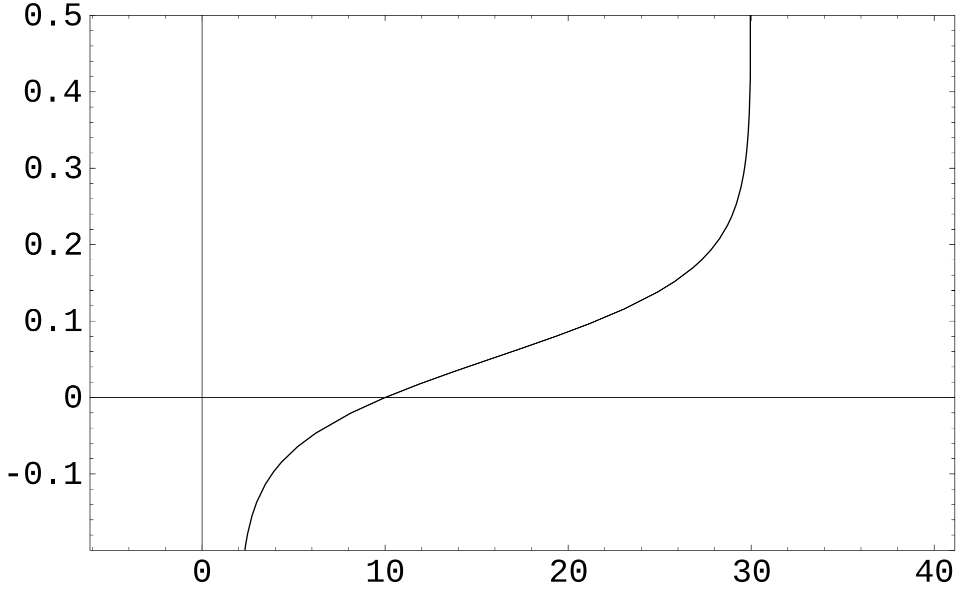


Figure 1: function $G(w_k, = d_k)$ with $q_k = 1$ and $d_k = 10$

Logistic function: derivative

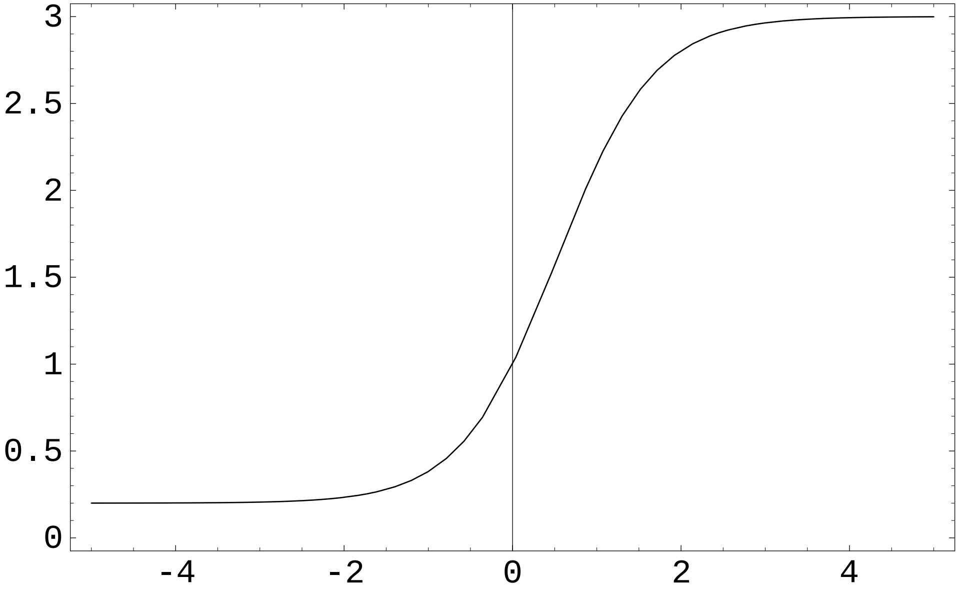
The derivative of the distance



function $G'(w_k, d_k)$ with $q_k = 1$ and $d_k = 10$

Logistic function

The calibration function is



function $F_k(u)$ with $q_k = 1$

ADVANTAGE: The weights are bounded.

Remarks on the calibration problem 1

- The weights can be bounded in such a way that

$$B^{-} \leq \frac{w_k}{d_k} \leq B^{+}.$$

  For instance $B^{-} = 0.4$ and $B^{+} = 3$.

- Other calibration functions can also be used.
- The variance of the regression estimator is a variance of residuals.

Remarks on the calibration problem 1

- Calibration to several stages (municipalities, households, individuals)
- If the calibration variables can explain the nonresponse, then a calibration can be used to correct at the same time the sampling error and the nonresponse error.

A good software of calibration?

- Easy to use?
- Which distance can be used?
- Possibility to impose bound?
- Special functionalities for non-responses?
- Computation of the estimator of variance, or at least of the residuals?
- Shareware?

References

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

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88:1013–1020.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies*. Dunod, Paris.