14es Journées de méthodologique statistique

29 mars 2022

Session spéciale en hommage à Jean-Claude Deville

Olivier Sautory

I. Mes premiers contacts avec Jean-Claude Deville

Ensae

- stage élève administrateur

- cours Processus aléatoires du second ordre

- GT : analyse harmonique qualitative sur les femmes mariées 3 fois

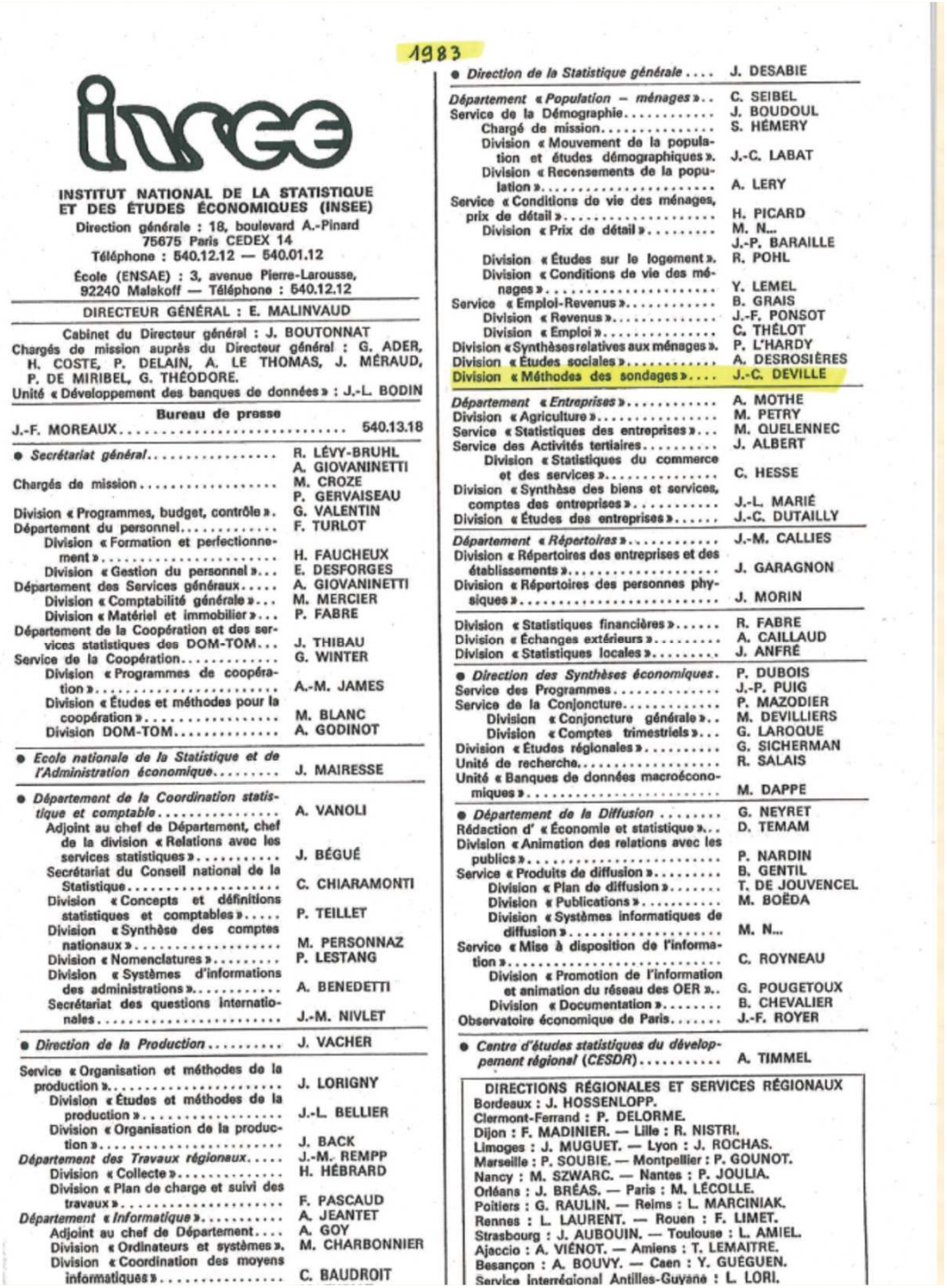
Insee

- assistant en statistique à l'Ensae

- département de la démographie
  (échantillon démographique permanent)

II. La division Méthodes de sondages

Jean-Claude remplace Fernand Chartier en 1983



1983

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES (INSEE)
Direction générale : 18, boulevard A.-Pinard
75675 Paris CEDEX 14
Téléphone : 540.12.12 — 540.01.12
École (ENSAE) : 3, avenue Pierre-Larousse,
92240 Malakoff — Téléphone : 540.12.12

DIRECTEUR GÉNÉRAL : E. MALINVAUD

Cabinet du Directeur général : J. BOUTONNAT
Chargés de mission auprès du Directeur général : G. ADER, H. COSTE, P. DELAIN, A. LE THOMAS, J. MÉRAUD, P. DE MIRIBEL, G. THÉODORE.
Unité « Développement des banques de données » : J.-L. BODIN

Bureau de presse
J.-F. MOREAUX 540.13.18

• *Secrétariat général*	R. LÉVY-BRUHL A. GIOVANINETTI
Chargés de mission	M. CROZE P. GERVAISEAU
Division « Programmes, budget, contrôle »	G. VALENTIN
Département du personnel	F. TURLOT
Division « Formation et perfectionnement »	H. FAUCHEUX
Division « Gestion du personnel »	E. DESFORGES
Département des Services généraux	A. GIOVANINETTI
Division « Comptabilité générale »	M. MERCIER
Division « Matériel et immobilier »	P. FABRE
Département de la Coopération et des services statistiques des DOM-TOM	J. THIBAU
Service de la Coopération	G. WINTER
Division « Programmes de coopération »	A.-M. JAMES
Division « Études et méthodes pour la coopération »	M. BLANC
Division DOM-TOM	A. GODINOT
• *École nationale de la Statistique et de l'Administration économique*	J. MAIRESSE
• *Département de la Coordination statistique et comptable*	A. VANOLI
Adjoint au chef de Département, chef de la division « Relations avec les services statistiques »	J. BÉGUÉ
Secrétariat du Conseil national de la Statistique	C. CHIARAMONTI
Division « Concepts et définitions statistiques et comptables »	P. TEILLET
Division « Synthèse des comptes nationaux »	M. PERSONNAZ
Division « Nomenclatures »	P. LESTANG
Division « Systèmes d'informations des administrations »	A. BENEDETTI
Secrétariat des questions internationales	J.-M. NIVLET
• *Direction de la Production*	J. VACHER
Service « Organisation et méthodes de la production »	J. LORIGNY
Division « Études et méthodes de la production »	J.-L. BELLIER
Division « Organisation de la production »	J. BACK
Département des Travaux régionaux	J.-M. REMPP
Division « Collecte »	H. HÉBRARD
Division « Plan de charge et suivi des travaux »	F. PASCAUD
Département « Informatique »	A. JEANTET
Adjoint au chef de Département	A. GOY
Division « Ordinateurs et systèmes »	M. CHARBONNIER
Division « Coordination des moyens informatiques »	C. BAUDROIT
• *Direction de la Statistique générale*	J. DESABIE
Département « Population — ménages »	C. SEIBEL
Service de la Démographie	J. BOUDOUL
Chargé de mission	S. HÉMERY
Division « Mouvement de la population et études démographiques »	J.-C. LABAT
Division « Recensements de la population »	A. LERY
Service « Conditions de vie des ménages, prix de détail »	H. PICARD
Division « Prix de détail »	M. N... J.-P. BARAILLE
Division « Études sur le logement »	R. POHL
Division « Conditions de vie des ménages »	Y. LEMEL
Service « Emploi-Revenus »	B. GRAIS
Division « Revenus »	J.-F. PONSOT
Division « Emploi »	C. THÉLOT
Division « Synthèses relatives aux ménages »	P. L'HARDY
Division « Études sociales »	A. DESROSIÈRES
Division « Méthodes des sondages »	J.-C. DEVILLE
Département « Entreprises »	A. MOTHE
Division « Agriculture »	M. PETRY
Service « Statistiques des entreprises »	M. QUELENNEC
Service des Activités tertiaires	J. ALBERT
Division « Statistiques du commerce et des services »	C. HESSE
Division « Synthèse des biens et services, comptes des entreprises »	J.-L. MARIÉ
Division « Études des entreprises »	J.-C. DUTAILLY
Département « Répertoires »	J.-M. CALLIES
Division « Répertoires des entreprises et des établissements »	J. GARAGNON
Division « Répertoires des personnes physiques »	J. MORIN
Division « Statistiques financières »	R. FABRE
Division « Échanges extérieurs »	A. CAILLAUD
Division « Statistiques locales »	J. ANFRÉ
• *Direction des Synthèses économiques*	P. DUBOIS
Service des Programmes	J.-P. PUIG
Service de la Conjoncture	P. MAZODIER
Division « Conjoncture générale »	M. DEVILLIERS
Division « Comptes trimestriels »	G. LAROQUE
Division « Études régionales »	G. SICHERMAN
Unité de recherche	R. SALAIS
Unité « Banques de données macroéconomiques »	M. DAPPE
• *Département de la Diffusion*	G. NEYRET
Rédaction d' « Économie et statistique »	D. TEMAM
Division « Animation des relations avec les publics »	P. NARDIN
Service « Produits de diffusion »	B. GENTIL
Division « Plan de diffusion »	T. DE JOUVENCEL
Division « Publications »	M. BOÉDA
Division « Systèmes informatiques de diffusion »	M. N...
Service « Mise à disposition de l'information »	C. ROYNEAU
Division « Promotion de l'information et animation du réseau des OER »	G. POUGETOUX
Division « Documentation »	B. CHEVALIER
Observatoire économique de Paris	J.-F. ROYER
• *Centre d'études statistiques du développement régional (CESDR)*	A. TIMMEL

DIRECTIONS RÉGIONALES ET SERVICES RÉGIONAUX
Bordeaux : J. HOSSENLOPP.
Clermont-Ferrand : P. DELORME.
Dijon : F. MADINIER. — Lille : R. NISTRI.
Limoges : J. MUGUET. — Lyon : J. ROCHAS.
Marseille : P. SOUBIE. — Montpellier : P. GOUNOT.
Nancy : M. SZWARC. — Nantes : P. JOULIA.
Orléans : J. BRÉAS. — Paris : M. LÉCOLLE.
Poitiers : G. RAULIN. — Reims : L. MARCINIAK.
Rennes : L. LAURENT. — Rouen : F. LIMET.
Strasbourg : J. AUBOUIN. — Toulouse : L. AMIEL.
Ajaccio : A. VIÉNOT. — Amiens : T. LEMAITRE.
Besançon : A. BOUVY. — Caen : Y. GUÉGUEN.
Service interrégional Antilles-Guyane : L. LORI.

L'histoire du calage

Le 1er texte de Jean-Claude sur le calage

Deville J.-C. (1988). Estimation linéaire et redressement sur information auxiliaire d'enquête par sondage, *Essais en l'honneur d'Edmond Malinvaud*, 915-927.

- construction de nouveaux poids "proches" des poids de sondage (de façon que les estimateurs linéaires utilisant ces poids soient "presque" sans biais) et qui "calent" l'échantillon sur un ensemble de totaux.

- minimisation d'une distance quadratique → généralisation de la procédure REDRE de Y. Lemel (poids de sondage quelconques et pondérations pour les individus).

- lien avec l'estimateur par régression généralisée de C.-E. Särndal (1982)

- explicitation de la méthode dans le cas de calage sur deux variables qualitatives

$$\Sigma = \sum d_k G\left(\frac{w_k}{d_k}\right) - \lambda \left(\sum w_k x_k - x \right)$$

$$\frac{\partial \Sigma}{\partial w_k} = G'\left(\frac{w_k}{d_k}\right) - \lambda x_k = 0$$

$$G'\left(\frac{w_k}{d_k}\right) = \lambda x_k$$

$$\frac{w_k}{d_k} = F(\lambda x_k)$$

$G(x) = \frac{1}{2}(x-1)^2$

$G' = x - 1 = u$

$F = u + 1$

$$\frac{w_k}{d_k} = 1 + \lambda x_k$$

$$w_k = d_k + \lambda x_k d_k$$

$$\sum w_k x_k = \sum d_k x_k + \lambda \sum d_k x_k^2 = X$$

$$= N \hat{x} + \lambda N \hat{x^2} = X$$

$$\lambda = \frac{X - \hat{x}}{\hat{x^2}}$$

$$w_k = d_k + d_k x_k \frac{X - \hat{x}}{\hat{x^2}}$$

$$\sum w_k y_k = \sum d_k \hat{y} + \frac{X - \hat{x}}{\hat{x^2}} \sum d_k x_k y_k$$

Aspects théoriques du calage sur marges

1. Le problème

Population : U = {1 ... k ... N}

s : échantillon de taille n tiré de U, avec les probabilités d'inclusion

$$\pi_k = \text{Prob}\ (k \in s)$$

Y = variable d'intérêt, dont on veut estimer le total sur la population :

$$T_Y = \sum_{k \in U} y_k$$

Estimateur de Horvitz-Thompson : $\hat{T}_{Y\pi} = \sum_{k \in s} \frac{1}{\pi_k} y_k = \sum_{k \in s} d_k\, y_k$

où d_k = ***poids de sondage***, ou ***poids d'échantillonnage***, de k.

Information auxiliaire

J variables auxiliaires $X_1 \ldots X_j \ldots X_J$ connues sur s, et dont on connaît **les totaux sur la population** $T_{X_j} = \sum_{k \in U} x_{jk}$.

Si l'information auxiliaire est relative à des variables catégorielles, cela signifie que l'on connaît les **effectifs** des modalités de ces variables, i.e. les totaux des variables indicatrices[1] associées à ces modalités.

---

[1] L'indicatrice associée à une modalité d'une variable catégorielle est la variable qui vaut 1 pour les unités qui prennent cette modalité, 0 sinon.

Problème

Tenir compte de cette information pour améliorer l'estimateur de Horvitz-Thompson. On va chercher un nouvel estimateur de T_Y :

$$\hat{T}_{Yw} = \sum_{k \in s} w_k \, y_k$$

où les nouveaux poids w_k :

- sont "proches" des poids initiaux d_k
- vérifient les **équations de calage** : $\forall \, j = 1 \ldots J \quad \sum_{k \in s} w_k \, x_{jk} = T_{X_j}$

2. Résolution théorique

On choisit une fonction G telle que $G(w_k/d_k)$ mesure la "distance" entre le poids initial d_k et le poids final w_k :

- $G(1) = 0$
- G est positive et convexe : $G(w_k/d_k)$ est d'autant plus élevée que le rapport w_k/d_k est éloigné de 1.

Les poids w_k cherchés sont solutions du problème d'optimisation :

$$\min_{w_k} \sum_{k \in s} d_k \, G(w_k/d_k) \quad \text{avec} \quad \sum_{k \in s} w_k \, x_k = T_X$$

en posant $x_k = (x_{1k} \ldots x_{jk} \ldots x_{Jk})'$ et $T_X = (T_{X_1} \ldots T_{X_j} \ldots T_{X_J})'$

Résolution

Lagrangien : $$L = \sum_{k \in s} d_k \, G(w_k / d_k) - \lambda' \left(\sum_{k \in s} w_k \, x_k - T_X \right)$$

où $\lambda = (\lambda_1 \ldots \lambda_j \ldots \lambda_J)'$ est un vecteur de multiplicateurs de Lagrange.

Les conditions du 1er ordre donnent : $$\boxed{w_k = d_k \, F(x_k' \, \lambda)}$$

où F est la fonction réciproque de la dérivée de la fonction G.

Le vecteur λ est déterminé par la résolution du système non linéaire des **équations de calage** :

$$\boxed{\sum_{k \in s} d_k \, F(x_k' \, \lambda) \, x_k = T_X}$$

Ce système d'équations peut être résolu par la méthode itérative de Newton.

La convergence est obtenue lorsque $$\max_{k \in s} \left| \frac{w_k^{(i+1)}}{d_k} - \frac{w_k^{(i)}}{d_k} \right| < \varepsilon$$

3. Les fonctions G usuelles

On indique les fonctions G(r) (où $r = w_k/d_k$), et F(u) (où $u = x_k' \lambda$)

La méthode linéaire

$$G(r) = \frac{1}{2}(r-1)^2 \qquad F(u) = 1+u$$

On montre qu'il y a convergence de l'algorithme de Newton dès la 2ème itération, et que :

$$\hat{T}_{Yw} = \sum_{k \in s} w_k y_k = \hat{T}_{Y\pi} + \hat{b}_1\left(T_{X_1} - \hat{T}_{X_1\pi}\right) + \hat{b}_2\left(T_{X_2} - \hat{T}_{X_2\pi}\right) + \ldots + \hat{b}_J\left(T_{X_J} - \hat{T}_{X_J\pi}\right)$$

où $\hat{b}_1, \hat{b}_2 \ldots \hat{b}_J$ sont les coefficients de la régression (pondérée) de ***Y*** sur

$\boldsymbol{X_1}$... $\boldsymbol{X_J}$ dans l'échantillon.

⇨ $\hat{T}_{Yw}$ est donc l'**estimateur par régression généralisée**[2] du total T_Y.

[2] C.-E. Särndal (1980, 1982)

La méthode exponentielle, ou méthode du raking ratio

$$G(r) = r\log r - r + 1 \qquad F(u) = \exp u \,(> 0)$$

Dans le cas d'un calage sur des variables catégorielles uniquement, on montre que cette méthode se confond avec la méthode du *raking ratio*[3], appelée aussi *Iterative Proportional Fitting* (ou encore méthode R.A.S.)

[3] W.E. Deming, F.F. Stephan (1940)

La méthode "logit"

L et U sont deux constantes telles que $L < 1 < U$.

$$G(r) = \left[(r - L) \log \frac{r - L}{1 - L} + (U - r) \log \frac{U - r}{U - 1} \right] \frac{1}{A} \qquad \text{si } L < r < U$$

$$= \infty \quad \text{sinon}$$

$$\text{où } A = \frac{U - L}{(1 - L)(U - 1)}$$

$$F(u) = \frac{L(U - 1) + U(1 - L)\exp(A u)}{U - 1 + (1 - L)\exp(A u)} \quad \text{et} \quad L < F(u) < U$$

i.e. méthode raking ratio "bornée" : les rapports de poids sont compris entre L (< 1) et U (> 1) .

La méthode linéaire tronquée

$$G(r) = \frac{1}{2}(r-1)^2 \quad \text{si } L \le r \le U$$

$$= \infty \quad \text{sinon}$$

$$F(u) = 1 + u \in [L, U]$$

Article soumis à JASA

(novembre 1989)

Publié dans les actes des JMS 1991

Calibration Estimators and Generalized Raking Techniques in Survey Sampling

Jean-Claude Deville
Institut national de la statistique
et des études économiques
Paris, France

Carl-Erik Särndal
Université de Montréal
Montréal, Canada

Abstract. This paper is about estimation of the finite population total in the presence of univariate or multivariate auxiliary information. Attention is focused on alternative weighting systems that reflect a given auxiliary information. There are two parts to the paper: (i) derivation of a weighting system with the aid of a distance measure and a set of calibration equations; (ii) an application to the case where the information consists of known marginal counts in a two- or multi-way table; this is called generalized raking.

The general regression estimator (GREG) was conceived with multivariate auxiliary information in mind. It is ordinarily justified by a regression relationship between the study variable y and the auxiliary vector x. But the GREG can be derived by a different route by focusing instead on the weights. The sampling weight of the k:th observation is π_k^{-1}, where π_k is the inclusion probability of k. We show that the weights implied by the GREG are as close as possible, according to a given distance measure, to the π_k^{-1} while respecting side conditions called calibration equations. These state that the sample sum of the weighted auxiliary variable values must equal the known population total for that auxiliary variable. That is, the calibrated weights must give perfect estimates when applied to each auxiliary variable. This consistency check on a weighting system is required by many practitioners. The GREG uses the auxiliary information efficiently, but the weights are not always without reproach. Negative weights can occur; in some applications such weights make no sense. It is natural to seek the root of the dissatisfaction in the

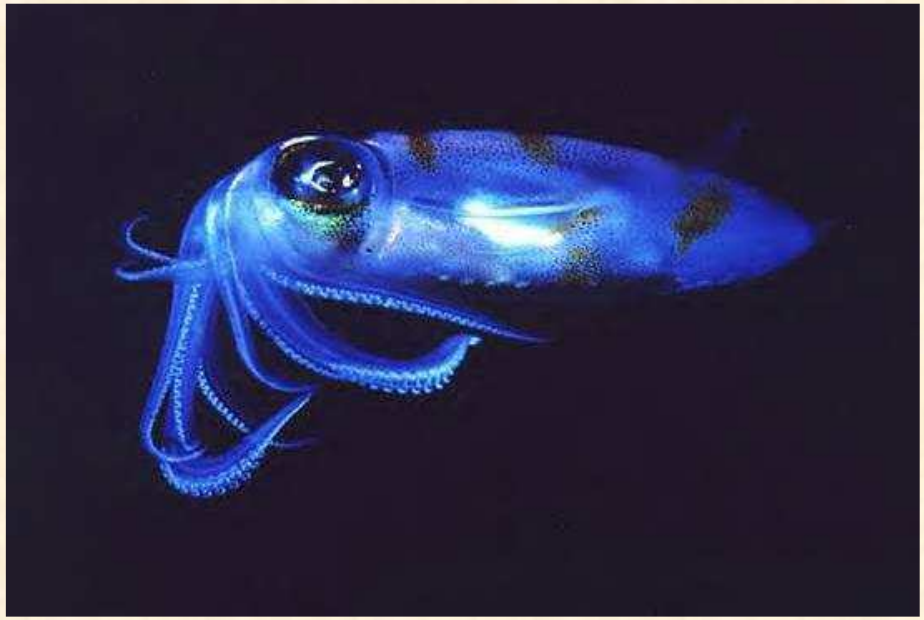
```
Random number seed (-32768 to 32767)?
? 9
? 16
? 7
? 32
? 18
? 38

 1          9.         16.          7.         32.         18.         38.
      0.00000     0.00000     0.00000     0.00000     0.00000     0.00000
 0  0

Ajustement:1lineaire,2raking,3logit,4lin.tronq,5:1/1+x,6chi2,7Hellinger? 1

 -0.296 -0.022 -0.014  0.267
 -0.296 -0.022 -0.014  0.267
meth 1 =
  0.970937  0.703848  1.244791  0.977702  1.252668  0.985578 1
   8.73844  11.26156   8.71354  31.28646  22.54802  37.45198
   0.00000     0.00000     0.00000     0.00000     0.00000     0.00000
Total=   0
 0

Ajustement:1lineaire,2raking,3logit,4lin.tronq,5:1/1+x,6chi2,7Hellinger?
1LIST   2RUN  3LOAD"  4SAVE"  5CONT 6,"LPT1 7TRON 8TROFF9KEY    0SCREEN


 -0.296 -0.022 -0.014  0.267
 -0.296 -0.022 -0.014  0.267
meth 1 =
  0.970937  0.703848  1.244791  0.977702  1.252668  0.985578 1
   8.73844  11.26156   8.71354  31.28646  22.54802  37.45198
   0.00000     0.00000     0.00000     0.00000     0.00000     0.00000
Total=   0
 0

Ajustement:1lineaire,2raking,3logit,4lin.tronq,5:1/1+x,6chi2,7Hellinger? 2

 -0.296 -0.022 -0.014  0.267
 -0.322 -0.026 -0.021  0.256
 -0.323 -0.026 -0.021  0.256
 -0.323 -0.026 -0.021  0.256
meth 2 =
  0.935060  0.724029  1.258732  0.974652  1.265185  0.979649 3
   8.41554  11.58446   8.81113  31.18888  22.77334  37.22666
   0.00000     0.00000     0.00000     0.00000     0.00000     0.00000
Total=   0
 0

Ajustement:1lineaire,2raking,3logit,4lin.tronq,5:1/1+x,6chi2,7Hellinger?
1LIST   2RUN  3LOAD"  4SAVE"  5CONT 6,"LPT1 7TRON 8TROFF9KEY    0SCREEN

%-33593.652 -0.425 -0.489%33580.145
Overflow
Overflow
Overflow
Overflow
 15 Overflow
```

9	16	25
7	32	40
18	38	60
40	80	

0,8 / 1,15

La macro Calmar

Calmar (CALage sur MARges) = macro SAS mettant en oeuvre 4 des méthodes de calage de l'article (1989-1990).



Exemple

1	2	2
2	1	4
4	2	6

→

1	1	2
3	1	4
4	2	6

Solution de la méthode linéaire

Exemple

1	2	2
2	1	4
4	2	6

→

0.91	1.09	2
3.09	0.91	4
4	2	6

Solution de la méthode raking ratio

La macro Calmar

Utilisée pour la 1ère fois sur l'enquête sur la consommation alimentaire de 1987, puis sur l'enquete "Modes de vie" de 1989.

Atelier de méthodologie statistique 1990-1991.

Document de travail et présentation aux JMS 1991 : Redressements d'échantillons d'enquêtes auprès des ménages par calage sur marges.

Nombreux échanges avec Statistique Canada 1990-1991 :
Michel Hidiroglou, Pierre Lavallée, Diane Stukel.

→ Macro Caljack (1993)

Calibration Estimators in Survey Sampling

JEAN-CLAUDE DEVILLE and CARL-ERIK SÄRNDAL

Reprinted from the Journal of the American Statistical Association
Volume 87, Number 418, June 1992

Generalized Raking Procedures in Survey Sampling

JEAN-CLAUDE DEVILLE, CARL-ERIK SÄRNDAL, and OLIVIER SAUTORY

Reprinted from the Journal of the American Statistical Association
Volume 88, Number 423, September 1993

SCB STATISTICS SWEDEN

UNIVERSITY OF ÖREBRO · SWEDEN

PROGRAM FOR WORKSHOP ON USES OF AUXILIARY INFORMATION IN SURVEYS.

October 5-7, 1992.

MONDAY, October 5.

9.00 - 9.45 Words of welcome (B. Dernulf / N. Lind)

Opening remarks on the theme of the workshop (C.E. Särndal)

9.45 -10.10 Coffee break

A. Auxiliary information for improved sampling and estimation.

I. *General theoretical approaches.*

Chair: C. E. Särndal.

10.10-10.40 J.N.K. Rao: Estimating distribution functions from sample survey data using auxiliary information at the estimation stage.

10.40-11.10 J. C. Deville: Constrained Sampling, Conditional Inference, Weighting: Three aspects of the utilisation of auxiliary information.

11.10-11.30 Discussion (J. H. Wretman)

11.30-12.00 R. L. Wright: Choosing Sample Sizes for Sample Designs with Auxiliary Information.

12.00-13.30 LUNCH

13.30-14.00 Y. Tillé: Estimation with auxiliary information using conditional inclusion probability.

14.00-14.20 Discussion (B. Swensson)

WEDNESDAY, October 7.

C. Analysis of survey data and computerized estimation systems.

Chair: D. Thorburn

9.00 - 9.30	R. Chambers: Nonparametric regression with complex survey data.
9.30 -10.00	O. Sautory: Calibration on known marginal counts for sample surveys: practical experiences at I.N.S.E.E.
10.00-10.20	Discussion (C. Cassel)
10.20-10.40	Coffee break
10.40-11.10	V. Estevao, M.A. Hidiroglou and C.E. Särndal : Requirements on a Generalized Estimation System at Statistics Canada.
11.10-11.40	C. Andersson and L. Nordberg: CLAN - a program for the computation of standard errors in survey sampling.
11.40-12.00	Discussion (J. C. Deville)
12.00-14.00	Closing Remarks and LUNCH.

La macro Calmar (suite)

Diffusion à l'intérieur de l'Insee : le plus souvent sans correction préalable de la non-réponse totale

Procédure « légitimée » par :
Dupont F. (1996). Calage et redressement de la non-réponse totale. *Actes des Journées de Méthodologie Statistique, 15 et 16 décembre 1993, INSEE-Méthodes n°56-57-58.*

Diffusion en France : Services statistiques ministériels, Banque de France, Ined, …

Et dans les INS étrangers 1993-1994 : Westat (G. Kalton), Australie, Suisse, Finlande, Suède, Norvège, Etats-Unis (Bureau of Labor Statistics), Royaume-Uni, Eurostat, Argentine,Nouvelle-Zélande, …

Le calage généralisé

Deville J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Article présenté au congrès de l'AFCAS, Sherbrooke, Québec.*

Méthode programmée par J. Le Guennec : Calmar 2 (2000-2002).

Présentations par J.-C. Deville (théorie) et J. Le Guennec (deux applications) aux JMS 2002…

… et par O. Sautory au symposium de Statistique Canada de 2003 : *Calmar2: a new version of the Calmar calibration adjustment program*

Création de l'Unité de Méthodologie Statistique (UMS) en 1994, à partir de la division « Méthodes statistiques et sondages » :
- division méthodologie d'élaboration et d'analyse des données
- cellule méthodologie de recueil des données

Unité de 6 personnes, plus des méthodologues de passage (thésards, Statistique Canada)

POUVEZ-VOUS NOUS EXPLIQUER SIMPLEMENT POURQUOI C'EST COMPLIQUÉ ?
PÉTILLON

IV. D’autres aventures partagées avec Jean-Claude

Elaboration d’un logiciel de calcul de précision adapté aux enquêtes-ménages de l’Insee : POULPE

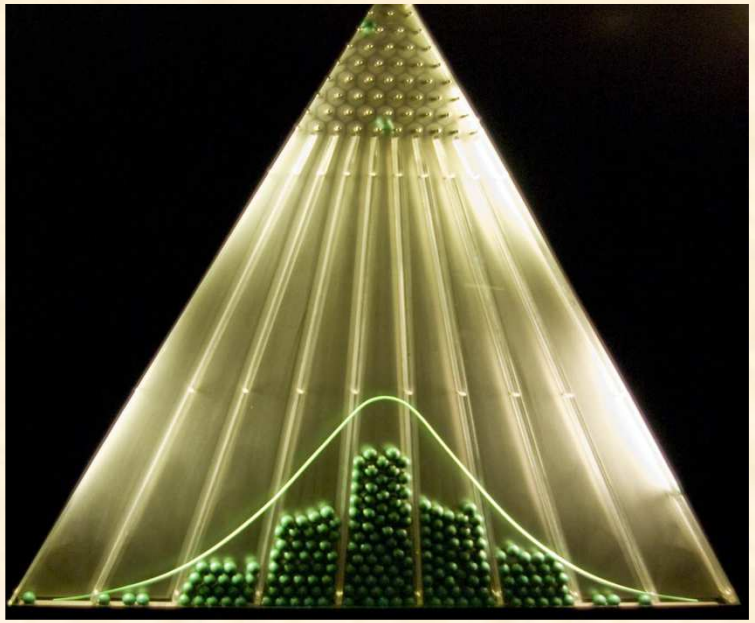
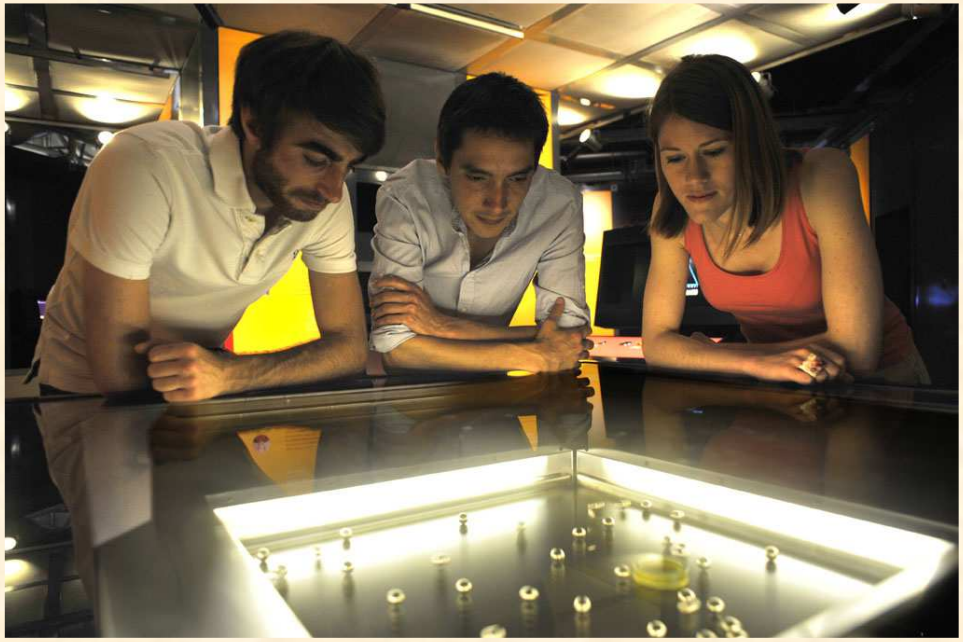
Cours de formation continue sous l’égide d’Eurostat (TES) :

• Techniques d’échantillonnage (1991-1998)

• Utilisation de l’information auxiliaire (1997-2001)

Participation à la création de l’espace *Probabilités et Statistiques* du nouvel *Îlot mathématiques* de la *cité des Sciences et de l’Industrie* à la Villette (1994-1995).

Planche de Galton



Mouvement brownien