Recent Developments in Sample Survey Methodology

Yves Tillé
University of Neuchâtel

FORS
Lausanne, February 2011

Table of contents

1. Some history
2. Mutation of official statistics
3. Direct impact on the methodology
4. Innovations in the methods of sampling
5. Generalized calibration
6. Small area estimation or back to model-based estimation
7. Nonresponse
8. Back to the foundations of inference
9. Conclusions

The main steps

- Quetelet : Blunt refusal of the idea of sampling
- Controversial of Kiaer (1896) about the representative sample
- Neyman (1934) Random selection of the sample (as opposed to purposive selection, quota method)
- Royall developed *model-based framework*.

Quetelet, 1946

Quetelet

> *Ne pas se procurer la faculté de vérifier les documents que l'on réunit, c'est manquer à l'une des principales règles de la science. La statistique n'a de valeur que par son exactitude ; sans cette qualité essentielle, elle devient nulle, dangereuse même puisqu'elle conduit à l'erreur.*

Kiaer, 1896

Kiaer

- Presentation in Bern of the results of a work entitled *Observations et expériences concernant des dénombrements représentatifs* related to a sample selected in Norway.
- Kiaer (1896) first selects a sample of cities and municipalities.
- Next, in each municipality, he selects a subset of the individuals in function of the first letter of their name.
- Two-stage sampling, but the choice of the units is not random.

Strong Reactions against Kiaer, 1896

M. V. Mayr [...] C'est surtout dangereux de se déclarer pour ce système des investigations représentatives au sein d'une assemblée de statisticiens. On comprend que pour des buts législatifs ou administratifs un tel dénombrement restreint peut être utile - mais alors il ne faut pas oublier qu'il ne peut jamais remplacer l'observation statistique complète. Il est d'autant plus nécessaire d'appuyer là-dessus, qu'il y a parmi nous dans ces jours un courant au sein des mathématiciens qui, dans beaucoup de directions, voudraient plutôt calculer qu'observer. Mais il faut rester ferme et dire : pas de calcul là où l'observation peut être faite.

Reactions against Kiaer (continued), 1896

M. Milliet. Je crois qu'il n'est pas juste de donner par un vœu du congrès à la méthode représentative (qui enfin ne peut être qu'un expédient) une importance que la statistique sérieuse ne reconnaîtra jamais. Sans doute, la statistique faite avec cette méthode ou, comme je pourrais l'appeler, la statistique, pars pro toto, nous a donné ça et là des renseignements intéressants ; mais son principe est tellement en contradiction avec les exigences que doit avoir la méthode statistique, que, comme statisticiens, nous ne devons pas accorder aux choses imparfaites le même droit de bourgeoisie, pour ainsi dire, que nous accordons à l'idéal que scientifiquement nous nous proposons d'atteindre.

Acceptation of the idea of sampling

- In 1924, committee ISI (Artur Bowley, Corrado Gini, Adolphe Jensen, Lucien March, Verrijn Stuart, et Frantz Zizek) of the ISI in order to evaluate the pertinence of the representative method.
- Neyman advocates for the use of random sampling.

Design-based framework

- A sample S is selected with inclusion probabilities π_k. The total
$$Y = \sum_{k \in U} y_k$$
of the population U is estimated by the Horvitz-Thompson estimator
$$\widehat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$
- The units are selected with unequal inclusion probabilities according to the Neyman principle. The units with a larger variance are over-represented.

Design-based framework

- Calibration (Deville and Särndal, 1992) : the Horvitz-Thompson weights $d_k = 1/\pi_k$ are adjusted on population totals that are known at the population level by a register or a census.
- We search weights w_k close to the d_k that satisfy the calibration constraints:

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k.$$

- The choice of an appropriate distance enables us to obtain desired properties for the weights (positive weights, bounded weights).
- Calibration decreases the variances but adds a small bias.

Model-Based framework

- Royall (1977) : model $y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k$
- The parameter $\boldsymbol{\beta}$ is estimated by $\widehat{\boldsymbol{\beta}}$ by using the sample.
- The non-observed values $\hat{y}_k$ are predicted by $\hat{y}_k = \mathbf{x}_k' \widehat{\boldsymbol{\beta}}, k \in U \backslash S$.
- The total $Y = \sum_{k \in U} y_k$ of the population is estimated by

$$\widehat{Y}_{BLUE} = \sum_{k \in S} y_k + \sum_{k \in U \backslash S} \hat{y}_k .$$

- The sampling design and the inclusion probabilities are not taken into account.

The mainstream

- Impartiality principle in public statistics.
- Make the use of models a little bit suspicious.
- Modelling is however accepted for nonresponse of for small area estimation.

More sources

Multiple sources:

- Registers,
- Sample surveys,
- Censuses.

Major trend

The suppression of the censuses:

- Reticence of local authorities.
- Reticence of some users.
- The citizens have trouble understanding the interest of the censuses.

Trend to an increasing of nonresponse.

Increasing of the needs

More expectations.

- Request for desegregated statistics (small area estimation).
- New interests: inequality, wellness, health, sustainable development, migrations, gender, discrimination.

Direct impact on the methodology

- Before, we had a census, and the sample surveys were calibrated on the census.
- Now, we have several registers and several sample surveys and we must provide consistent statistics.
- Problem with the registers: they are not designed in order to produce statistics (missing data, errors, update problems).

Direct impact on the methodology

Example: calibration on wrong data.

- We search weights w_k that enable us to reproduce the totals known of the auxiliary variables.
- $$\sum_{k\in S} w_k \mathbf{x}_k = \sum_{k\in U} \mathbf{x}_k .$$
- Often, registers contain errors: measurement errors, missing values on the $\mathbf{x}_k$.
- The missing values can be imputed: mean imputation in post-strata for instance.
- Once the sample is selected, we wish to calibrate the totals on the totals of the register.

Direct impact on the methodology

Example : calibration on wrong data.

- What's the matter if there are errors in the register? NOTHING
- Does calibration on wrong data introduces a bias in the estimate? No, no more than if the register were correct.
- Obviously, under the condition to calibrate the false data on the false data.
- Why? Because the calibrated estimator is asymptotically unbiased, whatever be the variables $\mathbf{x}_k$.
- Calibration only reduces the variance in function of the correlation between the auxiliary variables $\mathbf{x}_k$ and the interest variable.

Direct impact on the methodology

One can thus draw some paradoxical rules:

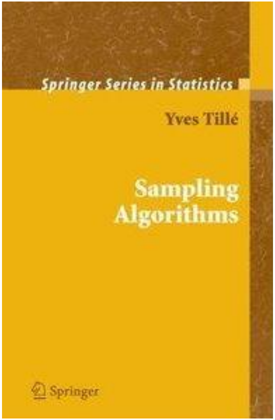
- We must calibrate the wrong data on the wrong data.
- We must definitely not update the registers with the sample surveys.
- If we update the registers, we obtain a random register and we introduce a bias.

Balanced sampling

It is possible to drawn balanced samples i.e. samples that have the same means in the sample as in the population Deville and Tillé (2004).

- Balanced sample $\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k .$
- Balanced sampling is an automatic calibration in the sampling design.
- Applications: French census, French master sample.

Balanced sampling



Co-ordination of samples

- The construction of dynamic registers poses the problem of the coordination of samples (Matei and Tillé, 2004, 2005; Nedyalkova et al., 2009).
- Rotation, management of the overlap, problems of births and deaths, splitting, fusions, etc.
- Not really completely solved. A perfect solution probably does not exist.
- Conflict between the quality of the coordination and the accuracy of the sampling design.

Generalized calibration Deville (2002, 2004); Chang and Kott (2008)

- In case of questionnaire nonresponse, one uses a model to estimate the probabilities of response of the respondents.
- One can directly estimate the parameter of this model by a calibration method.
- The error due to nonresponse and the sampling error are dealt with at the same time.
- The variables that explain the nonreponse do not need to be the same as the calibration variables.

Small area: Composite estimator

- Estimation based on the data

$$\bar{y}_S = \frac{1}{n_D} \sum_{S \cap D} y_k.$$

- Estimation based on a model and an auxiliary information

$$\bar{y}_M = \frac{1}{n_D} \sum_{S \cap D} \mathbf{x}_k \widehat{\boldsymbol{\beta}}.$$

- One used a weighted mean (coefficient α depends on the variances).

$$\alpha \bar{y}_S + (1 - \alpha) \bar{y}_M.$$

- More the number of observations is small, more the model-based estimation takes an important part of the estimator.

Small area

- Use of mixed models

$$y_k = \mathbf{x}_k \boldsymbol{\beta} + u_i + \varepsilon_k, k \in D_i,$$

- The random effect u_i is the effect of belonging to area i.
- With an approach based on a mixed model, it is possible to show that the optimal estimator is a composite estimator.
- Fuller and Battese (1973), Fay and Herriot (1979).
- Several hundreds of publications on small area estimation dealing with mixed models.
- Important need from the official statistical institutes.

Small area estimation

Questions:

- The random effects are generally used to take into account an uncontrolled effect in the experiments.
- The random effects are also used to reduce the number of degrees of freedom, because the estimation of all the effects is replaced by the estimation of the variance of the effects.
- By instance, we use the random effect for the individuals in panels. In fact, in panel data, we are not interested in the individuals.
- One generally uses the random effects for a nuisance variable that is not interesting.

Small area estimation

Problems:

- We are interested in small area and precisely the effect of the area is a random effect.
- The estimator obtained by using a mixed model is conditionally biased.
- The weighting system α depends on the area, which produces a problem of consistency.
- Do we sacrifice the principle of impartiality by using a model?
- Is it possible to make the small area estimates 'design-based' compatible?
- Finally, the most important is to have correlated auxiliary variables with the variable of interest.

Nonresponse

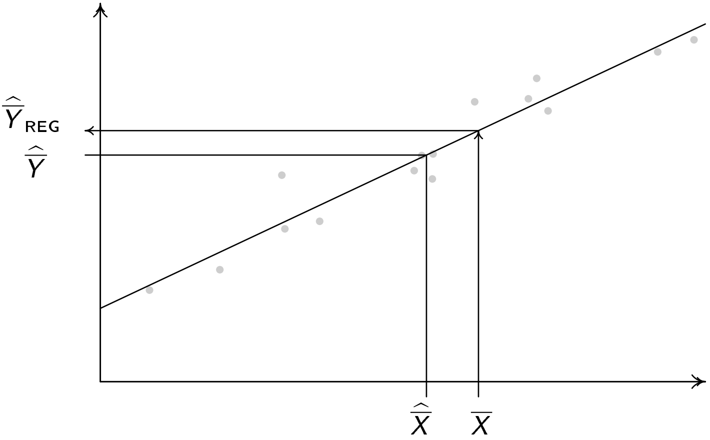
Questions:

- Item nonresponse: imputation
- Questionnaire nonresponse: reweighting by the inverse of the (estimated) probability of nonresponse.

Analogy with the two main sampling frameworks.

- Design based -> reweighting.
- Model-based -> prediction.

Nonresponse



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

Nonresponse

- Nonresponse and small domain send us back to the old dilemma: Must the inference be based on the sampling design or on a model?
- In some cases, both approaches are not necessarily incompatible.
- Under some technical conditions, the regression estimator can be presented under three forms (Nedyalkova and Tillé, 2009)

$$\widehat{Y}_{reg} = \widehat{Y} + \left(\mathbf{X} - \widehat{\mathbf{X}}_{\pi}\right)' \widehat{\mathbf{B}}.$$

$$\widehat{Y}_{reg} = \sum_{k \in S} w_k y_k.$$

$$\widehat{Y}_{reg} = \sum_{k \in S} y_k + \sum_{k \in \overline{S}} \widehat{y}_k.$$

Small area estimation

According to the presentation, the application of the different presentations to small area estimation gives different estimates:

$$\widehat{Y}_{reg} = \sum_{k \in S \cap D} w_k y_k.$$

$$\widehat{Y}_{reg} = \sum_{k \in S \cap D} y_k + \sum_{k \in \overline{S} \cap D} \widehat{y}_k.$$

Conclusion

- Do these new problems send us back to our old demons?
- Will nonresponse and small area estimation push us to a more model-based approach?
- Will we more calculate and less observe?
- Must we reconcile the design based approach and the model based approach? (Nedyalkova and Tillé, 2009)

Bibliography

Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95:555–571.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. In *Actes des Journées de Méthodologie, INSEE, Paris*.

Deville, J.-C. (2004). Calage, calage généralisé et hypercalage. Technical report, Internal document, INSEE, Paris.

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. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366):269–277.

Fuller, W. A. and Battese, G. E. (1973). Transformation for estimation of linear models with nested error structure. *Journal of the American Statistical Association*, 68:626–632.

Kiaer, A. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9, livre 2:176–183.

Matei, A. and Tillé, Y. (2004). On the maximal sample coordination. In Antoch, J., editor, *Proceedings in Computational Statistics, COMPSTAT'04*, pages 1471–1480. Physica-Verlag/Springer.

Matei, A. and Tillé, Y. (2005). Maximal and minimal sample co-ordination. *Sankhyā*, 67:590–612.

Nedyalkova, D., Qualité, L., and Tillé, Y. (2009). Tirages coordonnés d'échantillons à entropie maximale. Technical report, University of Neuchâtel.

Nedyalkova, D. and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521–537.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.