Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Uffizi federal da statistica UST

Coordinated sampling: Theory, method and application at SFSO

Lionel Qualité

Swiss Federal Statistical Office

ENBES workshop on Coordinated Sampling for Business Surveys | March 1st 2019

Business surveys at SFSO

- Business register with $\approx 600'000$ active units.
- 5-6 coordinated business surveys each year.
- Typically “stratified” (size and industry), cut-off of smallest units.
- Collection of uncoordinated surveys of local units (e.g. price index statistics), or selected on behalf of other offices or partners,
- Other non-random surveys (e.g. profiling) or surveys on different populations (e.g. non hotel accomodations).

Coordinated surveys

- Rotating panels (e.g. Value-added statistic - WS and Job statistic - Besta. . .): 5 rotation blocs, updated annually (WS) or irregularly (Besta),
- Repeated - every other year - surveys (e.g. Earnings structure survey, Continuing training survey . . .) with renewed samples,
- Possibly one occasion surveys.

Coordination needs - 1

- Accomodate repeated surveys, panels, updated panels, rotating panels, one-occasion surveys,
- compatible with updating of the sampling frame,
- spread response burden over the population.

Coordination needs - 2

- Allow different “stratifications” for different surveys or different sampling occasions,
- make it possible to use new sampling frame information (wages, turnover) for future sampling designs,
- $\Rightarrow$ Exactly respects freely chosen inclusion probabilities.

Notations

- π_k^t probability that unit k is selected at time t, π_k^{ts} at times t and s,
- independent surveys: $\pi_k^{ts} = \pi_k^t \pi_k^s$,
- positive coordination for unit k if $\pi_k^{ts} > \pi_k^t \pi_k^s$, negative otherwise,
- "optimal" coordination at bounds

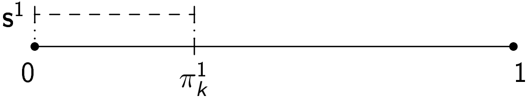
$$\underbrace{\max(0, \pi_k^t + \pi_k^s - 1)}_{\text{optimal negative coordination}} \leq \pi_k^{ts} \leq \underbrace{\min(\pi_k^t, \pi_k^s)}_{\text{optimal positive coordination}} .$$

Brewer's method

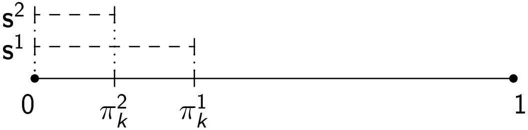
- Poisson Transversal designs.
- For each $k \in U$, generate a permanent random number $u_k \sim \text{Unif}[0,1]$ (only one for all the sampling occasions),
- First occasion: select k if $u_k < \pi_k^1$
- Second occasion:
  - Positive coordination. select k if $u_k < \pi_k^2$
  - Negative coordination. select k if $\pi_k^1 < u_k < \pi_k^1 + \pi_k^2$ (when $\pi_k^1 + \pi_k^2 \leq 1$)

Brewer's two samples selection method - 1

- First sampling occasion

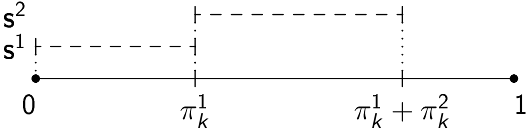


- Positive coordination when $\pi_k^2 \leq \pi_k^1$

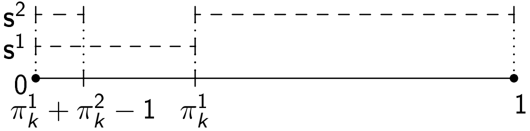


Brewer's two samples selection method - 2

- Negative coordination when $\pi_k^1 + \pi_k^2 \leq 1$

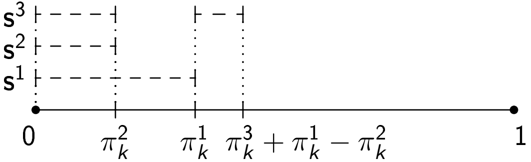


- Negative coordination when $\pi_k^1 + \pi_k^2 \geq 1$



Generalization to 3 or more surveys

1. Put an order on sub-intervals of $[0, 1]$ according to desired coordination rules,
2. construct selection zone for new survey,
3. example: third survey positively coordinated with second then negatively coordinated with first.



After t surveys

1. For each unit: record $t+1$ selection intervals and corresponding longitudinal samples,
2. To select a new sample: for each unit, rank all intervals in function of coordination priorities,
3. add intervals to selection set until their total length exceeds π_k^{t+1},
4. split last interval into selection and no-selection intervals so that selection probability is π_k^{t+1}.

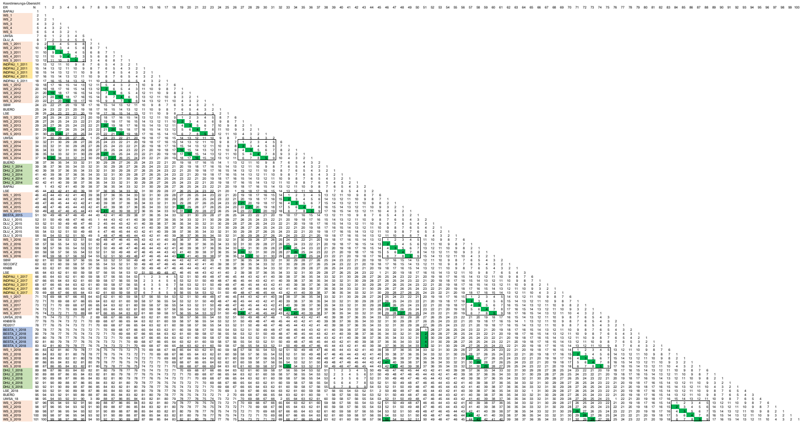
Coordinated Poisson Sampling

- Extends Brewer et al. (1972)'s method of two samples selection with permanent random numbers,
- allows to select coordinated one-occasion surveys, panels or rotating panels,
- accommodates dynamic populations with births, deaths, as well as mergers, split-offs, take-overs, break-ups,
- has transversal Poisson sampling designs (independent unequal-probabilities unit selections),
- has some optimality properties for sample coordination.

Coordinated Poisson Sampling

- Coordination with respect to the survey with highest priority is optimal (π_k^{ts} is at its bound),
- if negative coordinations with priorities in chronological order then longitudinal design is systematic,
- always strictly respects inclusion probabilities (if the random number generator. . .)

Business surveys 2009-2019



Koordinierungs-Übersicht

Coordination priorities

- Open question: priority given to coordination between different occasions for a same survey or to coordination with other recent surveys ?
- E.g. is it better to reselect a business into a rotating panel (contrary to what was initially advertised) or to select it for another rotating panel (implying new training, costs, etc.) ?
- Currently, we do the former.

Programming difficulties

- Comparing "real" numbers (interval endpoints) when they are 'equal', e.g.: if $\pi_k^t = 1$ for some t then there is no *new* interval.
- efficiently storing longitudinal samples (e.g. no proper boolean type in SAS),
- large number of independent sortings: could profit from parallel/distributed computing.

Random sample size

- Not new: unfortunately there is non-response in all the surveys selected with this program,
- variance of calibrated (Hájek) estimator with Poisson sampling is close to that of Horvitz-Thompson estimator with fixed size sampling (*when inclusion probabilities are equal*), and we always calibrate,
- effect on sample sizes and budget is negligible.

Unit selection independance

- No choice of the transversal sampling designs, only of the inclusion probabilities,
- ⇒ no cluster, multi-phase or balanced sampling,
- ⇒ usually not suitable for face-to-face surveys,
- ⇒ no multi-level coordination (e.g. businesses/local units, households/persons).
- Currently: coordination at business/household level.
- N.b.: could look at using coordinated sampling at the lower level working with conditional inclusion probabilities.

Side Benefits

- Simple and correct procedure for repeated surveys with varying populations and inclusion probabilities,
- standardization of sampling procedures, files, designs and samples storage, *weighting and variance estimation methods*, etc.
- *Poisson sampling* ⇒ simplified variance computations and estimation, no strata collapsing (replaced with calibration variables selection). . .

Effect on burden spreading/repeated selections

- Difficult to evaluate for business surveys: mix of positive and negative coordinations due to panels and rotating panels,
- Modest: for large or very small businesses it makes no difference,
- not all surveys are selected within this system.
- In population surveys, hundreds of thousands of multiple selections avoided.

Updating the sampling frame: Ideally

- Based on the typology of demographic events in Eurostat business registers recommandations manual,
- reflects continued or discontinued existence of businesses,
- Takeover and Split-off: one unit retains its history and others are deleted or created with virgin history,
- Merger and Break-up: new units are created with virgin history,
- Simultaneous with changes in the business register.

But. . .

- Using a table of events recorded in our business register (BR),
- missing some important information on takeovers and split-offs: identification number may change so that we do not know which business continues,
- recorded events include backs-and-forths, erroneous mutations, fictitious units, etc.

Really

- Retain independance between units selections: either the sampling history of one disappearing unit is inherited by a new unit or a new virgin history and random number is created,
- if identifier persits then unit sampling history follows,
- Link units or groups of units in BR at time a to units or group of units in BR at time $b > a$ using events and their timestamp,
- automatically transfer history when there is only one possibility.

Really - continued

- other cases (one to many, many to one, many to many) are forwarded to BR administrators for a decision on the history transfer.
- ⇒ no automatic procedure ⇒ no continuous frame updating,
- frame is updated every semester, ≈ 200 cases forwarded to BR administrators each time.

Sustainability: data and computation growth

- Stored data: longitudinal samples (support of longitudinal designs),
- seems necessary if one wants to choose coordination type with all past samples,
- for each unit, after t survey occasions: $t + 1$ samples,
- it is the minimum number of samples if inclusion probabilities are freely chosen (Wynn 1977),
- $\Rightarrow$ data $\geq O(N \cdot t^2)$ (plus interval endpoints, random numbers, etc. - N is the population size).

Sustainability: data and computation growth

- Computations probably $> O(N \cdot t^2)$,
- current implementation tested to $t = 238$ for business surveys (works but annoyingly slow in the end),
- failed at $t = 210$ for household surveys ($N \approx 3.5m$),
- limit: matrix size in SAS IML, but computation times are also problematic,
- better implementations are possible, but a growth rate of t^2 is too fast.

Sustainability: data and computation growth

- Groups of units share common designs/supports, at least in the beginning,
- consecutive negatively coordinated surveys that always receive successive coordination priorities may be grouped,
- possible to reinitialize the system retaining only part of the information on previous surveys, e.g. selections and selection probabilities of units during a few selected periods or in a few selected surveys or groups of surveys,
- → used a couple of times for our population and household surveys.

Sustainability: data and computation growth

- Free choice of inclusion probabilities and of coordinations may be too much to ask for,
- constrained inclusion probabilities ('strata') help reduce "effective" population size N,
- using only negative coordination and chronological priorities is equivalent to random number shifting,
- that is what we ended-up doing for household surveys as, after 8 years, positive coordination was never used.

Sustainability: unanticipated needs/requests

- Possibility to meet some unanticipated requests, e.g. move from a panel survey to a rotating panel,
- unit independance greatly helps finding solutions: only need to consider relatively small longitudinal designs when reinitializing the system, everything is computable,
- also helps with computations in other cases, e.g. introducing some dependence between units selection by using coordinated sampling as a part of a multilevel sampling design.

Conclusion and assessment after 10 years

- Does not answer the needs of all NSIs: no *simple and efficient* coordination of surveys at different levels, not a huge lifespan - or not at its full capacity,
- but strongly contributed to standardize our operations,
- and to confidently select samples for our repeated surveys,
- lived up to our expectations at SFSO,
- especially since all requirements were not known in advance.

Brewer, K., Early, L., and Joyce, S. (1972).
Selecting several samples from a single population.
Australian Journal of Statistics, 3:231–239.

Eurostat (2010).
Business registers - recommandations manual.
Tech. Rep. , Publications Office of the European Union, Luxembourg.
ISBN 978-92-79-14659-6.

Qualité, L. (2009).
Unequal probability sampling and repeated surveys.
Thèse de doctorat, Université de Neuchâtel, Neuchâtel, Suisse.

Wynn, H. (1977).
Convex sets of finite population plans.
Annals of Statistics **5**, 414–418.