A new sampling design for the Swiss Earnings Structure Survey

Lionel Qualité

Swiss Federal Statistical Office

International Conference on Establishment Statistics (ICES VI) | June 15, 2021

Earnings Structure Survey

- Collects information on wages, qualifications, employment rate, job category, etc. employment category, etc.
- Sample survey conducted every two years.
- Produces results on labour costs, wage distributions.
- Data used by the *Salarium* wage calculator, and to study wage discriminations.

Sampling design, until 2016 - 1

- Business sample stratified by size class, kind of activity and geographical area.
- Poisson selection since 2012.
- Two stage survey (businesses provide data on a part only of their workforce).
- Sample sizes derived from target precisions for estimators of mean wages.

Sampling design, until 2016 - 2

- Neyman optimal allocation for the mean wage estimator.
- Parameters estimated on previous surveys.
- Minimum sample sizes imposed within strata (but a posteriori groupings for nonresponse correction).
- Interdependencies between SFSO sample and cantonal extensions.

New for 2018

- Data on incomes delivered by the compensation offices (OASI incomes).
- Comprehensive (almost) but
  1. available with a two-year delay,
  2. no information on employees activity rates,
  3. sometimes available only at the head of group level.

Use of OASI incomes

- As a proxy for wages.
- To calculate sample allocation according to chosen accuracy targets,
- *(and to choose/adapt these targets depending on the resulting size)*.
- To calculate sample extensions that supplement the SFO sample.
- Without limitations or apprehensions related to the use of previous survey data.

Tasks

- Validation of the similarity between OASI incomes and ESS wages using the 2016 survey data.
- Allocation with the purpose of estimating medians.
- Allocation for several precision objectives.
- Write a report (Qualité and Potterat 2020).

Validation-1

- An opportunity to review the ESS variance estimators (Poisson design, calibration, two stages sampling).
- Two computations:
  1. What would be the estimated variance for the 2016 survey if we had collected OASI incomes rather than wages? (if close to variances with survey data then we may use OASI incomes as proxy)
  2. What is the estimated variance using the 2016 survey data if different first and second stage sampling rates are used? (helpful to check that the new design has the expected precision).

Validation-2

- Note: if we change the second degree design, the variance estimator is the sum of 3 terms instead of the usual two terms.
- Relatively simple (and estimable) for a Poisson sampling design.
- Results: compatible with the confidence we have in the variance estimates, i.e. not identical but relatively similar.

Variance of median estimators - linearisation

- Solution: derive a linearised variable z_k and reduce to the variance of an estimator of total.
- For q_α an estimator of the α-quantile of a variable y_k, Deville (1999) then Osier (2009) suggest:

$$z_k = -\frac{1}{Nf(q_\alpha)}\left[I(y_k \leq q_\alpha) - \alpha\right].$$

where $f(u) = \frac{1}{hN}\sum_{k\in U}\phi\left(\frac{u-y_k}{h}\right)$, ϕ is a kernel function, h a bandwidth parameter, I is the indicator function and N is the size of population U.

Variance of median estimators - linearisation

- Tillé and Vallée (2019) following Graf (2015):

$$z_k = -\frac{1}{Nf(q_\alpha)}\left[\Phi\left(\frac{q_\alpha - y_k}{h}\right) - F(q_\alpha)\right],$$

  where $\Phi' = \phi$ and $F' = f$.
- Typical choice: f the Gaussian density function and $h = (4/3)^{1/5}\sigma N^{-1/5}$, where σ is the standard deviation of y.
- z_k needs to be estimated (e.g. by plug-in).

Simulations

- The ESS sample size is too large for simulations $\rightarrow$ simulation on *small* SRS samples, with or without calibration.
- Choice of h: adaptation to extreme values (e.g. replace σ with $min\{\sigma, (q_{.75} - q_{.25})/1.34\}$ in the usual formulas.)
- Choice of linearised variable has to match the choice of quantile estimator! (otherwise: variance estimation bias, incorrect coverage rate, etc.).
- Median computed by the SAS-surveymeans procedure shows larger bias and variance than those deriving from kernel density estimators (at least for small samples).

Example

Table: 10'000 simulations, SRS samples, size 5'000, calibrated weights.

Method	Median	$\mathrm{E}(\widehat{q}_{.5})$	Approx. S	$\mathrm{E}(\widehat{S})$	Simul. S.D.	Cov. rate
m0	55'948	55'974	350.92	351.08	347.80	0.9473
m1	55'958	55'963	314.88	315.60	317.72	0.9470
m2	55'956	55'959	315.86	316.50	315.40	0.9483

m0: SAS-surveymeans and Osier approximation, m1: Gaussian kernel, m2: Exponential kernel.

Sample allocation - 1

- Uniform first stage units selection rate π_h within *strata* h (combination of size class, kind of activity and geographical area, similar as 2016).
- Approximated variance as a function of strata sample sizes $n_h = N_h \pi_h$:

$$V = \sum_{h \in H} \frac{d_h^2}{n_h} - b_h.$$

- n_h need not be integer and may be arbitrarily close to 0.
- d_h and b_h depend on data, expected response rates, second stage sampling rates, calibration, etc.

Sample allocation - 2

- Optimal for the total of one variable, under cost constraint, or minimum cost for a given variance.
- with n_h between lower bound a_h and upper bound b_h: there are numerous references (Neyman 1934, Aeberhardt and Marcus 2006, Koubi and Mathern 2009, but also Gabler et al. 2012, etc.)
- While $a_h = 0$ is admissible for a Poisson design, ability to set $a_h > 0$ is useful to compute sample extensions.
- Partial solution in Gabler et al.

Sample allocation - 3

- Remark that the order of elements of $A = \{a_h\sqrt{c_h}/d_h, b_h\sqrt{c_h}/d_h; h = 1, \ldots, H\}$ gives an order of (de-)activation of constraints, where c_h is the average sampling cost in stratum h.
- I *(I think)* found the same solution as Aeberhardt and Marcus 2006.
- *Prooved* that it was the exact solution to the problem.
- Programming difficulties: tied values in A ; $d_h = 0$; comparisons of real numbers ; infinite variances $(n_h = 0)$; etc.

Sample allocation - 4

- Variation Coefficient under 5% required for publication of ESS results. Target VC: 3%.
- (main) Statistics of interest: total labour costs by activity section (level 1 of the classification), by size class ; median wages for 39 fields of activity (compositions based on level 3 of the classification), by european NUTS2 regions, and activity fields combined with NUTS2 regions (target VC= 5%). Extensions: same targets as for NUTS2 regions.
- 342 SFO targets, 345 extension targets.
- 2016 sample: 50,600 businesses, including all businesses with 50 employees or more.
- Expectations for 2018: cautiously reduce sampe size.

Sample allocation - 5

- Convex optimisation problem, but solution usually not in the interior of the feasible domain.
- Sub-optimal solution for 2018: objectives considered one at a time and sample augmented where needed.
- Careful choice of the order in which objectives are considered.

Sample allocation - 6

- Calculation of objectives: if the minimum VC exceeds 3% (resp. 5%), aiming for the best possible VC is far too expensive.
- If the minimum VC is close to 3% (resp. 5%), aiming for it can also be expensive.
- Implemented solution: add 0.5 or 1 pt to the minimum VC if it is close to or exceeds 3% (resp. 5%).
- Calibration causes a dependency between domains, even if they are non overlapping.
- Sample extensions:
  - Strictly as a supplement of the SFSO sample in order to obtain desired local precisions,
  - Additional businesses selected only within the canton that pays for the extension.

Sample allocation - 7

- Result for the 2018 sample: 46'000 businesses.
- *Validation* using a small simulation (approx. 2 days computation time for 1'000 repetitions).
- VCs compatible with expectations, except in areas where one or a few businesses:
  - make up for a large share of jobs,
  - have different incomes from the rest of the domain.
- Recommendation: obtain responses from these units.

Improvements for the 2020 survey

- Replace target VC for total labour costs with VC for mean wages as it is a better match for published results,
- Force a positive lower bound on sampling rates,
- Replace the sequential allocation procedure by a call to a convex optimizer in the CVXR package of the R statistical software (turns out it works mostly, apart from a few rounding errors).

Bibliography - 1

- Aeberhardt, R. & Marcus, V. (2006). *Mesure et Contrôle de la Précision dans un Plan de Sondage Complexe. Cas de l'Enquête sur la Structure des Salaires de 2006.* Presented at an internal methods seminar, INSEE.
- Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: linearisation and residual techniques. *Survey Methodology*, 25, pp. 193-204.
- Gabler, S., Ganninger, M. & Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75, pp. 151-161.

Bibliography - 2

- Graf, M. (2017). *A simplified approach to linearisation variance for surveys*. Technical report, University of Neuchâtel.

- Koubi, M. & Mathern, S. (2009). *La nouvelle méthode d'échantillonnage de l'enquête trimestrielle ACEMO depuis 2006. Amélioration de l'allocation de Neyman.* Working document of the Direction de l'animation de la recherche, des études et des statistiques, 146.

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

Bibliography - 3

- Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearisation techniques. *Survey Research Methods*, 3, pp. 167-195.

- Qualité, L. & Potterat, J. (2020). *Enquête suisse sur la structure des salaires 2018: Révision du plan de sondage.* Methodological report, SFSO.

- Vallée, A. A., & Tillé, Y. (2019). Linearisation for Variance Estimation by Means of Sampling Indicators: Application to Non-response. *International Statistical Review*, 87(2), 347-367.