New Sampling Design For The Swiss Job Statistics

Jean-Marc Nicoletti, Daniel Assoulin [1]

Abstract
In 2015 different aspects of the Swiss Job Statistics, JobStat, have been revised. One revision-point concerned the sampling design. In this paper we first briefly describe the different requirements for the new sampling design and the chosen strategy to satisfy them. Then we focus on the approach to construct the take-all strata. Due to skewed distributions of target variables, the construction of take-all strata is an important point for attending precision goals / efficiency at the sampling design stage for JobStat. At the other hand, experience shows that non-response in take-all strata can be a severe threat to precision goals, when ignored during the sampling design stage. In this paper we present the method developed for JobStat in order to construct a take-all strata with a 100% response requirement in a way its size is reasonable, the overall sample size is monitored and there is no need for further 100% response requirements to achieve the precision goals. The work relies heavily on the methods available in the R-package Stratification.

Key Words: Sampling design, stratification, take-all stratum, non-response

1. The Swiss job statistics (JobStat)

The JobStat (Swiss Job Statistics) survey provides quarterly estimates of employment variables like the number of employees for the establishments in Switzerland and at finer levels, like NACE 2 aggregates[2], economic sectors and NUTS 2 [3]. While the national plan only takes into account the estimation at NUTS 2-level, smaller regions have the options to increase their sample size to obtain regional estimates at a certain precision level. Till 2015, a stratified sampling was considered with the establishment as sampling and estimation unit. Since 2015, JobStat is included into FSO's Sample Coordination System for *enterprise* surveys. Therefore a cluster sampling is considered where the first level sampling unit is the enterprise and the second level is the establishment. If an enterprise is selected to be in the sample, all its establishments are selected too. Estimation is still performed at establishment level. Profiling information is integrated into JobStat and enterprises belonging to profiling groups are considered as a fully sampled stratum (profiling : partnership between enterprises and FSO that makes data transfer easier).

2. General aspects of the new sampling design

The sampling plan is composed of a national sampling plan and several regional sampling plans (only for the regions who asked for an increase of their sample size). The sampling probabilities for the national plan are constructed at enterprise level (primary sampling units) by using a stratification of FSO50 $\times$ size classes and a precision goal of CV $= 4\%$ at FSO50 level (FSO50 is an aggregation of NACE 2 into 50 activity classes, defined in table 2). For each region, requesting an increased sample size, a similar procedure with precision goals at economic sector level (CV=3%) has been applied (see table 3). The precision goal of the sampling plan is based on the estimation of the total of employees. The sampling

[1]Jean-Marc Nicoletti, Daniel Assoulin, Swiss Federal Statistical Office, Espace de l'Europe 10, 2010 Neuchtel, email: jean-marc.nicoletti@bfs.admin.ch, daniel.assoulin@bfs.admin.ch.

[2]NACE : General industrial classification of economic activities within the European communities.

[3]NUTS : Nomenclature of territorial units for statistics.

probabilities in each sampling plan are computed as for stratified sampling but used for Poisson sampling (sampling process considered in the FSO's Sample Coordination System). The sampling probabilities of the final (or integrated) sampling plan are calculated as the maximum between the sampling probabilities from national, respectively regional plans. In the next section we describe in more detail the construction of the national plan. The focus will be on the construction of the take-all strata and also the integration of profiling units into the sampling design.

3. National sampling plan

Let U the JobStat enterprises population and $Y = \sum_{i \in U} y_i$ the total employees. Let $\mathcal{H} = \{1, ..., 47\}$ denote the set of the considered FSO50 strata, $Y_h = \sum_{i \in U}$ the total of employees in U_h, $h \in \mathcal{H}$ and $\widehat{Y}_h$ its Horvitz-Thompson estimator with a coefficient of variation

$$\text{CV}_h = \frac{\sqrt{\text{Var}\left[\widehat{Y}_h\right]}}{Y_h}. \tag{1}$$

The national sampling plan is designed at the level of the primary sampling unit (enterprise) with the following approach.

1. Coefficient of variation $\text{CV}_h = 4\%$ for each $h \in \mathcal{H}$.
2. Profiling PROF is treated as a fully sampled stratum with the assumption of a 90% response rate.
3. Determination of a take-all stratum $\text{TA}_{RR=100\%}$ with a constraint on its size and with a response rate of 100%.
4. Calculate optimal size classes (two) and sample size allocation for enterprises outside $\text{TA}_{RR=100\%}$ and Profiling in line with our precision goals.
5. Verification of the expected precision for the extrapolation at establishment level.

3.1 Construction of the take-all strata

Take-all strata play an important role in sample design and estimation when the target variable is skewly distributed (see per example Hidiroglou, 1986). Considering a 100% response rate in the take-all stratum permits to reach precision goals with a reasonably sized sample. If at the other hand non-response in the take-all strata is just ignored during the sampling design stage this may jeopardize the precision level obtained at the estimation stage. The construction of the the national sampling plan first considers the determination of a take-all stratum $\text{TA}_{RR=100\%}$ with 100% response rate based on the following criteria : The size $|\text{TA}_{RR=100\%}|$ of the $\text{TA}_{RR=100\%}$ has to be big enough too allow us a reasonable sample size and in the same time it has to be not too big in order to be able to perform the time consuming operational duties in a given timespan (follow up of all the enterprises in $\text{TA}_{RR=100\%}$ in order to ensure a 100% response rate).

3.1.1 Take-all 100% response rate determination $TA_{RR=100\%}$

The determination of the $\text{TA}_{RR=100\%}$ is performed with an iteratively procedure using with the method LH (Lavallée-Hidiroglou) in the R-package stratification of L.-P. Rivest and S.

J	$N_{tot,J}$	$n_{s_{nat}}$
3	3924	10153
4	1989	11244
5	1158	13044
6	755	14877
7	494	16464

Table 1: Size of $\text{TA}_{RR=100\%}$, $N_{tot,J}$, in relation to J. The iterative procedure stops when $J = 7$.

Baillargeon (2009). Based on the precison target ($\text{CV}_h = 4\%$) the variable of interest Y (number of employees for each sampling unit), the request for a take-all stratum, the number of desired size classes and the response rate assumption inside these classes, the LH method will furnish boundaries for the size classes leading to a minimal sample size such that $\widehat{Y}_h$ will meet the targeted precision. For more explanation, see Rivest and Baillargeon (2009). In the following we describe the different steps of the iterative procedure used for $\text{TA}_{RR=100\%}$ determination.

1. Set : J=3 the number of desired size classes, targeted precision $\text{CV}_h = 4\%$, a response rate of 90%[4] for the size class $j = 1, ..., J-1$ and 100% in class J (flagged as the take-all stratum).
2. Apply the LH method within each $h \in \mathcal{H}$ (FSO50).
3. Let $N_{h,J}$ the size of the resulting J^{th} size class.
4. Compute $N_{tot,J} = \sum_{h\in\mathcal{H}} N_{h,J}$, the total number of enterprises allocated to size class J with 100% response rate (size of $\text{TA}_{RR=100\%}$).
5. If $N_{tot,J} > 600$, the maximal allowed size for the $\text{TA}_{RR=100\%}$ for JobStat[5], then repeat the procedure with $J = J + 1$. Stop otherwise, the $\text{TA}_{RR=100\%}$ has been defined with a size of $N_{tot,J}$.

Table 1 illustrates the obtained results. We note that $N_{tot,J}$ is decreasing in J. In order to have a reasonable size ($\leq$ 600) $\text{TA}_{RR=100\%}$ we have to put $J = 7$ which leads to a $\text{TA}_{RR=100\%}$ with 494 entreprises compared to 3924 with $J = 3$. After each iteration also the resulting total expected sample size ($n_{s_{nat}}$) is monitored. The calculation is explained in the following section.

3.1.2 Remaining enterprises (out of $\text{TA}_{RR=100\%}$ and PROF)

For the remaining enterprises, not in $\text{TA}_{RR=100\%,h}$ and not in Profiling (PROF_h), $V_h = U_h \setminus \left\{\text{TA}_{RR=100\%,h} \cup \text{PROF}_h\right\}$, $h \in \mathcal{H}$, we calculate an adjusted target CV. As $U_h = \text{TA}_{RR=100\%,h} \cup \left(\text{PROF}_h \setminus \text{TA}_{RR=100\%,h}\right) \cup V_h$ we have

$$Y_h = Y_{\text{TA}_{RR=100\%,h}} + Y_{\text{PROF}_h \setminus \text{TA}_{RR=100\%,h}} + Y_{V_h}, \quad (2)$$

and the variance of $\widehat{Y}_h$ can be written as

$$\text{Var}\left[\widehat{Y}_h\right] = \text{Var}\left[\widehat{Y}_{\text{TA}_{RR=100\%,h}}\right] + \text{Var}\left[\widehat{Y}_{\text{PROF}_h \setminus \text{TA}_{RR=100\%,h}}\right] + \text{Var}\left[\widehat{Y}_{V_h}\right]. \quad (3)$$

[4]Based in historic data, response rates in JobStat are around 90% or higher.

[5]600 was set as maximum size of $\text{TA}_{RR=100\%}$ for JobStat but this parameter can be adapted according the needs of a specific survey.

Note that Var $\left[\widehat{Y}_{\text{TA}_{RR=100\%,h}}\right] = 0$ ($\text{TA}_{RR=100\%}$ is fully sampled with 100% response rate). The variance from $\text{PROF}_h \setminus \text{TA}_{RR=100\%,h}$, the profiling part out of $\text{TA}_{RR=100\%,h}$, does not vanish with the assumption of a 90% response rate. Plugging (3) and (2) in equation (1) and solving for $\text{CV}_{V_h} = \frac{\text{Var}\left[\widehat{Y}_{V_h}\right]}{Y_{V_h}}$ leads to the adjusted target CV

$$\text{CV}_{V_h} = \frac{\sqrt{Y_h CV_h^2 - \text{Var}\left[\widehat{Y}_{\text{PROF}_h \setminus \text{TA}_{RR=100\%,h}}\right]}}{Y_{V_h}}, \tag{4}$$

After the adjusted CV (4) is computed, the LH stratification method is applied based on a Neyman allocation to each set V_h, $h \in \mathcal{H}$, with two size classes under a 90% response rate assumption and CV_{V_h} as target CV. This step provides an expected sample size $\widetilde{n}_{V_h} = \widetilde{n}_{V_{h,1}} + \widetilde{n}_{V_{h,2}}$ in V_h, where $\widetilde{n}_{V_{h,1}}$ and $\widetilde{n}_{V_{h,2}}$ are the sample sizes provided by the LH method in the size classes 1 and 2 respectively. The sampling probabilities in V_h are thus $\frac{\widetilde{n}_{V_h,1}}{N_{V_h,1}}$ and $\frac{\widetilde{n}_{V_{h,2}}}{N_{V_{h,2}}}$ if sampling unit i belongs to $V_{h,1}$ or $V_{h,2}$ respectively, with $N_{V_{h,1}}$ and $N_{V_{h,2}}$ the respective population sizes. The national sampling probabilities are thus

$$\pi_{nat,i} = \begin{cases} \frac{\widetilde{n}_{V_h,1}}{N_{V_h,1}} & \text{if sampling unit } i \in V_{h,1} \\ \frac{\widetilde{n}_{V_h,2}}{N_{V_h,2}} & \text{if sampling unit } i \in V_{h,2} \\ 1 & \text{if sampling unit } i \in \text{PROF} \cup \text{TA}_{RR=100\%} \end{cases}, h \in \mathcal{H}. \tag{5}$$

So, $\sum_{i \in U} \frac{1}{\pi_{nat,i}}$ leads to the resulting expected sample size for the national plan. Table 1 displays the expected sample sizes $n_{s_{nat}}$ for different sizes of the take-all strata ($N_{tot,J}$). In table 1 we can see that $J = 3$ would lead to an overall sample size of 10153, compared to 16464 for $J = 7$. This is the price to pay if we want a manageable $\text{TA}_{RR=100\%}$ size (ensuring a 100% response rate). Note that $N_{tot,J}$ is decreasing in J.

4. Regional sampling plans

The regional sampling plans are calculated for 5 regions (cantons or big cities) $\mathcal{K} = \{1, ..., 5\}$ who want an increase of their regional sample size for an improved precison level of the estimates in their own region. The construction of regionals sampling plans follows the same process as for the national sampling plan. Let $U^\star = \cup_{k \in \mathcal{K}} U_k^\star$, with $U_k^\star$ the set of enterprises belonging to the regions in $\mathcal{K}$. The considered domains where a certain level of precision has to be achivied are the regions crossed with the sectors $\mathcal{L} = \{2, 3\}$ (secondary and tertiary, see table 3). The regional sampling plans are designed at the first level sampling unit with a precision target of $\text{CV}_h = 3\%$ for each $(k, l) \in \mathcal{K} \times \mathcal{L}$. For each $i \in U^\star$ results a regional sampling probability $\pi_{reg,i}$.

5. Final sampling plan

The final sampling plan is the result of an integration of the regional sampling probabilities $\pi_{reg,i}$ and the national sampling probabilities (5) according

$$\pi_{\text{fin,i}} = \begin{cases} \max(\pi_{\text{nat,i}}, \pi_{\text{reg,i}}) & \text{if } i \in U^\star \\ \pi_{\text{nat,i}} & \text{elsewhere} \end{cases}. \tag{6}$$

These probabilities are then finally used to draw the sample in FSO's sample coordination system based on Poisson sampling. The expected sizes from the final sampling plan are presented in Table 4. To summarize, we first concentrated on the construction of the national sampling design, paying special attention to the construction of the take-all strata, as discussed in detail in this paper. Certain aspects of this design are summarized in Table 2 of the Appendix. Table 2 displays for each FSO50 activity level

- the size of the 100% response take-all strata $\text{TA}_{RR=100\%,h}$,
- the number of Profiling enterprises not contained in the 100% response take-all strata, $\text{PROF}_h \setminus \text{TA}_{RR=100\%,h}$,
- CV_{V_h}, the adapted target CV for entreprises outside $\text{TA}_{RR=100\%,h} \cup \text{PROF}_h$ according to (4),
- the expected sample size with, $\tilde{n}_{U_h}$, and without, $\tilde{n}_{V_h}$, entreprises in $\text{TA}_{RR=100\%,h} \cup \text{PROF}_h$,
- the expected CV, CV_{est}, when estimation is performed at the establishment level. Remember that while the sampling design was established with a target CV for the estimation on enterprise level, the effective estimation is performed on establishment level. Based on the results in Table 2 one can see that in general also the estimates on establishment level are well in line with the precision target of CV=4%. The existing deviations are due to multi-establishment enterprises with establishments in more than one FSO50 aggregate. The calculation is based on the assumption of a 100% response rate for enterprise in $\text{TA}_{RR=100\%}$ and 90% for enterprises outside the $\text{TA}_{RR=100\%}$. As enterprises are the reporting units, non-response is just considered in the first stage (enterprise).

In a similar way we produced the regional plans for which the corresponding figures are summarized in Table 3. Note that prior to establishing the regional plans we performed a partial recodification of multi-establishment enterprises with respect to their geographical location. This was needed in order to ensure that the sampling design, established on enterprise level, performed also well for estimating regional aggregates on an establishment level.

Finally, we calculated the sampling probabilities for each enterprise by taking the maximum sampling probability observed in the national and the regional design. The expected sample sizes, $\tilde{n}_{\text{U}_{\text{int},h}}$, based on these final sampling probabilities are reported in Table 4.

6. Conclusions

Due to skewed distributions of target variables, the construction of take-all strata is an important point for attending precision goals / efficiency at the sampling design stage for JobStat and other business surveys. The aim is to completely eliminate the contribution of very large enterprises to the variance of the estimator. However, non-response in take-all strata can jeopardize this effort and result in estimates that do not fulfill the precision targets.

In this paper we present the method developed for JobStat in order to construct a take-all strata taking into account non-response:

- The Lavallée-Hidiroglou method implemented in the Package Stratification is iteratively applied with an increasing size of strata (size classes) and the assumption of a 100% response rate in the take-all strata, and of 90% among other strata.

- After each iteration enterprises identified as belonging to the take-all strata are removed from the population and the LH-method is applied to the remaining enterprises requesting two strata with 90% response and an adjusted target CV, reflecting the existence of the 100% response take-all strata and Profiling. Profiling units outside the 100% take-all strata are considered as a separate take-all strata with 90% response rate.

- One observes that an increasing number of strata in the first point implies a decreasing number of enterprises in the take-all strata and at the other hand an increase of the total sample size resulting from the sample size calculated in point 2, added to the size of the take-all strata.

- Therefore, the iteration was stopped as soon the size of the take-all strata was down to a manageable size, meaning the 100% response can be achieved by the production unit. Further reduction would lead to an increased sample size with constant precision which is not desirable.

The method heavily relies on methods already implemented in the package stratification and is easy to implement. While it proved to work well for JobStat, we think it could also be useful in other business statistics with skewed distributions of target variables.

REFERENCES

Louis-Paul Rivest, Sophie Baillargeon (2009), A General Algorithm for Univariate Stratification, International Statistical Review.

Renaud, A., Panchard, C. and Potterat, J. (2008), "Statistique de l'emploi. Révison 2007: méthodes d'estimation,", Numéro de commande: 338-0055, FSO.

Renaud, A. (2008), "Statistique de l'emploi. Révison 2007: cadre de sondage et échantillonage," Numéro de commande: 338-0052, FSO.

Särndal, C.E., Swensson, B. et Wretman, J. (1992), Model Assisted Survey Sampling, Springer.

M. A. Hidiroglou (1986), "The Construction of a Self-Representing Stratum of Large Units in Survey Design." The American Statistician 40, no. 1: 27-31.

A. Tables

FSO50 ($h \in \mathcal{H}$)	$\lvert \text{TA}_{RR=100\%,h} \rvert$	$\lvert \text{PROF}_h \setminus \text{TA}_{RR=100\%,h} \rvert$	$CV_{V_h}(\%)$	$\tilde{n}_{V_h}$	$\tilde{n}_{U_h}$	$CV_{\text{est}}(\%)$
Sector 2						
5-9	8	12	5.34	67	87	4.57
10-12	21	22	5.83	245	288	4.09
13-15	10	2	5.21	231	243	4.17
16-18	10	9	4.29	537	556	3.95
19-20	13	7	6.98	98	118	4.16
21	5	4	10.71	45	54	3.82
22-23	12	5	4.88	271	288	4.07
24-25	14	3	4	436	453	3.94
26	15	8	6.37	177	200	4.17
27	7	5	6.5	126	138	3.93
28	10	6	4.82	274	290	3.95
29-30	6	1	1	69	76	4.48
31-33	9	2	2	387	398	3.84
35	7	52	7.28	112	171	4.13
36-39	5	20	4.89	176	201	3.99
41-42	15	10	4.75	407	432	3.97
43	12	61	4.28	522	595	3.94
Sector 3						
45	8	24	4.63	460	492	3.96
46	13	87	4.56	669	769	3.90
47	13	148	7.4	339	500	4.33
49	12	33	6.5	295	340	3.93
50-51	4	2	9.32	57	63	3.88
52	11	21	7.66	138	170	3.33
53	2	5	25.18	17	24	4.10
55	12	33	4.41	395	440	3.93
56	13	22	4.61	425	460	3.77
58-60	6	18	6.23	273	297	4.05
61	3	7	17.81	34	44	3.84
62-63	8	31	4.53	608	647	3.87
64	9	48	10.4	181	238	4.00
65	8	63	13.03	54	125	3.94
66	11	76	5.05	403	490	3.93
68	7	27	4.57	471	505	3.97
69	9	10	4.62	475	494	4.00
70	13	58	5.19	398	469	3.45
71	5	26	4.17	653	684	3.95
72	15	8	7.94	118	141	3.88
73-75	10	12	4.7	453	475	4.00
78	14	2	6.04	253	269	4.04
79-82	16	33	5.47	468	517	3.96
84	15	71	19.93	35	121	3.94
85	18	20	5.61	411	449	3.80
86	31	42	6.18	377	450	3.98
87	5	42	4.42	242	289	3.85
88	12	19	4.75	440	471	4.03
90-93	5	15	4.23	682	702	3.88
94-96	7	104	4.19	630	741	3.95
U	494	1336		14634	16464	

Table 2: National plan with expected sizes and CV in each FSO50 (aggregates of NACE 2) at enterprises level and establishment level (expected precision CV_{est}).

Region ($U_k^\star$, ($k \in \mathcal{K}$)	Sector ($l \in \mathcal{L}$)	$\lvert \text{TA}_{RR=100\%,k,l} \rvert$	$\lvert \text{PROF}_{k,l} \setminus \text{TA}_{RR=100\%,k,l} \rvert$	$CV_{V_{k,l}}(\%)$	$\tilde{n}_{V_{k,l}}$	$\tilde{n}_{U_{k,l}}$
Geneva	2	10	7	4.20	614	631
Geneva	3	16	87	4.29	865	968
Neuchâtel	2	9	2	4.56	621	632
Neuchâtel	3	14	20	5.20	705	739
Saint-Gall	2	12	28	5.51	360	400
Saint-Gall	3	9	142	5.16	461	612
Vaud	2	13	35	5.28	341	389
Vaud	3	19	195	4.74	824	1038
Zürich City	2	18	12	6.15	258	288
Zürich City	3	31	234	5.95	597	862
$U^\star$		151	762		5646	6559

Table 3: List of regional sampling plans (region crossed with sectors $\mathcal{K} \times \mathcal{L}$).

FSO50 ($h \in \mathcal{H}$)	$\tilde{n}_{\text{U}_h}$	$\tilde{n}_{\text{U}_{\text{int},h}}$	$\tilde{n}_{\text{U}_{\text{int},h}}$ - $\tilde{n}_{\text{U}_h}$
Sector 2			
5-9	87	88.7	1.7
10-12	288	347.9	59.9
13-15	243	258.1	15.1
16-18	556	635.7	79.7
19-20	118	126.9	8.9
21	54	59.7	5.7
22-23	288	299.7	11.7
24-25	453	562.3	109.3
26	200	330.9	130.9
27	138	148.8	10.8
28	290	331.8	41.8
29-30	76	77.2	1.2
31-33	398	462.6	64.6
35	171	179.0	8.0
36-39	201	209.4	8.4
41-42	432	560.5	128.5
43	595	1191.9	596.9
Sector 3			
45	492	517.3	25.3
46	769	849.1	80.1
47	500	664.8	164.8
49	340	363.6	23.6
50-51	63	64.3	1.3
52	170	175.3	5.3
53	24	25.9	1.9
55	440	456.3	16.3
56	460	525.0	65.0
58-60	297	297.5	0.5
61	44	45.3	1.3
62-63	647	662.4	15.4
64	238	259.4	21.4
65	125	132.4	7.4
66	490	497.5	7.5
68	505	517.8	12.8
69	494	525.9	31.9
70	469	502.3	33.3
71	684	733.3	49.3
72	141	141.1	0.1
73-75	475	506.6	31.6
78	269	291.6	22.6
79-82	517	557.3	40.3
84	121	176.0	55.0
85	449	513.0	64.0
86	450	606.6	156.6
87	289	386.6	97.6
88	471	487.1	16.1
90-93	702	716.1	14.1
94-96	741	883.3	142.3
Total	16464	18952.0	2488.0

Table 4: Final and national plan expected sizes $\tilde{n}_{\text{U}_{\text{int},h}}$ and $\tilde{n}_{\text{U}_h}$, respectively, in each FSO50 (aggregates of NACE 2).