WEIGHTING A SNOWBALL SAMPLE THROUGH A PSEUDO-CALIBRATION: THE CASE STUDY OF SAME-SEX CIVIL UNIONS IN ITALY

Marco D. Terribili

Abstract. This paper presents an approach to addressing the challenges of selection bias and non-probabilistic characteristics of snowball sampling design, particularly in the context of social research involving hidden or hard-to-reach populations. The primary aim is to refine the weighting methodology of snowball sampling by introducing the pseudo-calibration technique to adjust the direct weights, equal to one, according to some available auxiliary variables, with the ultimate goal of producing more reliable and unbiased estimates.
The data for this case study is derived from a snowball sample survey, called the "Over the Rainbow" project, carried out on Instagram users who tag their photos with popular LGBTQ+ community hashtags. The username list is collected using web-scraping tools that identify relevant users. To address the sampling design's limitations, the study employs calibration and post-stratification methods. Calibration involves adjusting the weights of the sample data to match known population totals, while post-stratification involves dividing the sample into subgroups that align with known demographic distributions.
The proposed sampling weights are benchmarked on the Istat survey on all civil unions between same-sex couples, celebrated by Italian municipalities since 2016; this known totals source represents a reliable external reference to adopt sampling weights. The expected results of this study are twofold. First, the application of calibration should yield sample weights that are more representative of the target population, by aligning the sample distribution with known population totals. Second, post-stratification is anticipated to define pseudo-weights, adjusting the direct weights equal to one, refining the sample by ensuring that the subgroups within the sample correspond proportionally to those in the broader population. The combination of these methods is expected to significantly reduce the biases associated with traditional snowball sampling and attribute sampling weights not just equal to 1, as usually happens in snowball sampling.
The paper contributes to the statistical and social research field by offering a methodologically sound approach to improving the accuracy of snowball sampling designs, with practical implications for studying hard-to-reach populations on social media platforms.

1. Introduction

Snowball sampling is a type of purposive sample, particularly useful for recruiting individuals who are difficult to identify and for whom a predefined

sampling framework is unavailable. One of the main advantages of snowball sampling is its ability to ease data collection. This method involves identifying one participant who meets the study's criteria and then asking them to recommend other individuals with similar characteristics, thereby expanding the sample. This approach can approximate a simple random sample, potentially leading to more robust and reliable results (Abdul-Quader et al., 2006).

On the other hand, snowballing is a sampling design that, due to its non-probabilistic nature, does not allow for classical and probabilistic statistical inference on the reference population: determining the appropriate sampling weights to make inferences on this kind of sample can be particularly challenging.

1.1. How to weight a snowball sample? The (few) literature suggestions.

When attempting to make inferences with a snowball sample, the choice of sampling weights depends directly on several design factors and data collection process characteristics.

If the linkage between initial and subsequent participants can be preserved, Respondent-Driven Sampling (RDS) can be used (Heckathorn, 1997). RDS allows for the adjustment of sampling weights based on the network structure of the participants, thus facilitating more reliable inferences. This sampling method entails multiple snowballing waves, repeating these waves until the condition of Markov equilibrium is verified. In this way, the probability of being included in the final wave is independent of the probability of being recruited by the first one, which gets "watered down" as more waves are added. The main advantage of this method is that it asymptotically approximates a simple random sample, but, on the other hand, it requires tracking the recruiting IDs of responding people, in contrast with the essential privacy and anonymity constraints (De Rosa et al., 2020).

However, when such linkages are not available, researchers face the challenge of weighting the sampling units appropriately. The literature about weighting issues in this data collection situation is relatively sparse, with only a few contributions offering guidance: Shafie (2010), Snijders (1992), and Vitalini (2010). These three papers contribute significantly to the understanding of snowball sampling, yet none fully bridge the knowledge gap regarding the weighting of snowball samples: Shafie's paper excels in proposing innovative weighting techniques to correct for selection bias in snowball samples, using simulations to demonstrate the effectiveness of these methods in providing more accurate estimates. Snijders' study provides a thorough analysis of the mathematical properties and challenges of snowball sampling, offering a strong theoretical foundation and highlighting the complexities involved in sample selection probabilities. Vitalini's PhD dissertation

offers a comprehensive review, underscoring the critical importance of known selection probabilities for accurate statistical inference. Despite these strengths, the issue of how to precisely weight snowball samples remains unresolved, leaving a critical gap in the methodology regarding snowball sampling.

In this context, alternative techniques such as calibration, post-stratification, and pseudo-calibration have been employed to mitigate the limitations of snowball sampling. These methods involve adjusting the sample to match known population characteristics, thereby improving the accuracy of the inferences drawn from the sample.

In this context, according to known totals availability, on which to constrain, data on same-sex people in civil unions, coming from a previous research project, has been pseudo-calibrated.

2. The case study: the collected civil union snowball sampling units.

The case study data, focusing on civil unions in Italy, are derived from the undersigned PhD research project (Terribili, 2022), titled "Over The Rainbow", OTR from now on. The data collection process involved a comprehensive survey of Instagram users who tagged their pictures with popular LGBTQ+ community hashtags. This initial list of 8290 users was compiled using a web-scraping procedure in R. Subsequently, a questionnaire was disseminated to these users through Instagram, leveraging the same social network they actively engage with. To further expand and diversify the sample, the snowball sampling method was employed, where participants were encouraged to share the survey with others in their network. This approach effectively broadened the scope and depth of the data collected for the research: 638 users, named by other respondents, actively participated in the survey, also without being initially listed by the web-scraping procedure.

Unfortunately, at least for this study, the population(s) involved in the OTR survey were extremely young[1] and non-representative of the whole population. In fact, being the OTR sample non-probabilistic, we cannot strictly talk about representativeness. Anyway, the characteristics of people caught can be considered a result too. For instance, the higher collaboration from women was an important feedback about the web-scraping phase: in fact, also if most of the collected posts by the web-scraping script (according to the hashtags contained) were male-oriented, most respondents were female.

[1] The OTR respondents' average age was 24.84 years old, a mean value which decreases observing the median (23), and even more the mode (20).

This design bias led to just 30 individuals in civil unions. This relatively exiguous sample size will be one of the most crucial issues of the weighting phase, as will be explained in the following paragraph.

2.1. Benchmarking at Istat same-sex civil unions survey

Despite the limited sample size available for our survey, we have been privileged to have a robust benchmark to refer to. Istat has conducted an extensive survey encompassing all civil unions between same-sex couples that have been registered at the civil registry offices of Italian municipalities since 2016[2]. This comprehensive data collection effort has resulted in a known total of 18962 individuals in same-sex civil unions, up to the point of our data collection. This substantial and reliable dataset serves as a crucial external reference, available at different levels of aggregation, as shown in the table below:

Table 1 – Individuals *in a civil union up to 31st January 2020 (end of OTR data collection phase).*

Level		Modalities	No. of people in Civil union
Sex		Male	6964
		Female	11998
Age class		18-39	7172
		40+	11770
Educational level		Lower High School	7848
		High School	6448
		Higher High School	4666
Sex-Geographic area of birth	Female	Abroad	610
		North-West	3546
		North-East	2093
		Centre	2575
		South & Islands	1513
	Male	Abroad	915
		North-West	2581
		North-East	1587
		Centre	1947
		South & Islands	1324
TOTAL			18692

Source: Istat civil unions survey

[2] On June 5, 2016, the law No. 76 (commonly named "Cirinnà's law", according the Senator who proposed and submitted the law proposal), introducing civil unions between same-sex couples, came into effect in Italy. During the second half of that year, 2336 civil unions were formed, a particularly significant number, reflecting couples who had long been waiting to formalize their emotional bond. After this initial boom, there was a gradual stabilization in terms of number of civil unions incurred (Istat, 2023).

By leveraging this benchmark, we can use calibration and post-stratification model to adopt sampling weights that respect these known totals making the OTR survey results more representative and statistically sound.

3. Pseudocalibrating to provide sampling weight to non-probabilistic sample units

Handling non-probabilistic samples through calibration and/or post-stratification methods entails some crucial initial choices. If the primary goal of these methods is to adjust the initial sampling weights of respondents to align with known population totals from an external survey, which initial weights to consider is not trivial, as it happens with the direct weights (given by the reciprocal of inclusion probabilities, as in the Horvitz-Thompson estimator) for probabilistic sample surveys.

The theory indeed suggests that, in the case of convenience samples, each statistical unit involved in the survey represents only itself, thus having an (initial) sample weight of 1. This does not preclude making inferences by referring to the universe, at least in cases where the population totals are known. However, such a small and constant sample weight certainly complicates the calibration procedure in terms of convergence to obtain a result.

In this context, a pseudo-calibration estimator can be adopted for integrating the non-probability sample with a probabilistic one, assuming both samples contain relevant information for estimating the population parameter. These proposed estimators employ pseudo-weights (Elliot, 2009; Baker et al., 2013), sharing a structural similarity with the adjusted projection estimators, but adopting a different inferential approach and informative setup (Golini and Righi, 2024).

The underlying idea of applied pseudo-calibration is to correct the direct weights (d_i), equal to 1 for all units belonging to the non-probabilistic sample, as suggested by the sampling literature, with an adjustment factor, to obtain a pseudo-weight (δ_i). In our case, we adopted a post-stratification to proceed with this adjustment.

The post-stratification domains represent an exhaustive and mutually exclusive partition of the population, which can be made more and more detailed and granular, for example, by geographic area, or by geographic area crossed with gender.

The easiest pseudo-weights (δ_i) set is intuitively given by the product of the direct weight d_i, as already said equal to 1, with the inverse of the sampling fraction (*n/N*) for all *n* sample units involved in the OTR survey (*s*):

$$\delta_i = d_i \cdot \frac{N}{n} = K \qquad \forall\, i \in s \tag{1}$$

Then, this sampling fraction, and its reciprocal, can also be computed for each mutually exclusive, and jointly exhaustive, subgroup - named *h* - of the population, for which the number of civil unions contracted in Italy between 2016 and January 2020 is known, providing different sets of pseudo-weights to use and to compare.

$$\delta_{i_h} = d_i \cdot \frac{N_h}{n_h} = K_h \qquad \forall\, i \in s_h \tag{2}$$

These different scenarios result in various sets of initial pseudo-weights, on which calibration was tested on the margins, given by the variables listed in Table 1.

Therefore, the final calibrated estimator is as follows:

$$\hat{t}_{Y_{CAL}} = \sum_{i \epsilon s} y_i \cdot w_i = \sum_{i \epsilon s} y_i \cdot \delta_i \cdot \gamma_i = \sum_{i \epsilon s} y_i \cdot \frac{N_h}{n_h} \cdot \gamma_i \qquad \forall\, i \in s, h \in H \tag{3}$$

$$\text{where} \begin{cases} min_{w_i}\{\sum_{k \in s} G_i(w_i - d_i)/q_i\} \\ \sum_{k \in s} x_i \cdot w_i = t_i \end{cases}$$

Basically, the final weights w_i are obtained by solving an optimization problem, in which $G_i(w_i - d_i)$ is, a non-negative, and strictly convex pseudo-distance function, continuously differentiable to w_i (Deville and Sarndal, 1992).

In this way, by progressively sharpening the pseudo weights system δ_i, a post-stratification system was first utilized, dividing the sample into exhaustive and not overlapped subgroups, according to the variables of sex and geographical area of birth, ensuring each subgroup reflects the population's known demographic distribution. Then the pseudo-weights δ_i have been calibrated to match the population's marginal distribution across variables such as sex, age classes, and geographic area, getting the final weights set w_i.

The pseudo-calibration can be, in other words, described as the process that leads to the calibrated final weight from a pseudo-weight, used to avoid the direct weight equal to 1. This technique highlights how each method strives to enhance the representativeness and reliability of survey data, when auxiliary data is available and when the minimum distance between weight sets is guaranteed and maintained.

3.1. The weights variability: something to consider and control

A calibration model was applied based on the availability of auxiliary data, with the input pseudo-weights being varied. The $1+CV^2$ (one plus Coefficient of Variation squared) formula of Kish (1992) was then calculated for the resulting final weights.

This calculation quantifies how the variability in weights approximatively affects the estimates, allowing us to assess and manage this variability to ensure that it does not lead to increased variability in the estimates. By carefully monitoring and adjusting the weights, we aim to maintain the precision and reliability of the survey estimates, thus enhancing the robustness of our inferential procedures.

Table 2 – *Weights systems Kish's variability indicator.*

Weight system label	Direct or Pseudo-weight	$1+CV^2$	Calibration model	$1+CV^2$
a.	$d_i = 1$	1	Sex + Age classes + Educational Status	3.135
b.	$\delta_i = N/n$	1		3.277
c.	$\delta_{ih} = N_h/n_h$, where h is a Geographical Area	1.107		3.344
d.	$\delta_{ij} = N_j/n_j$, where j is a Geographical Area * Sex	1.672		3.687
e.	$d_i = 1$	1	Geographical Area * Sex	1.672

Table 1 shows that, from the perspective of sample weights and their variability, using constant direct weights equal to 1 does not offer a substantial advantage once the weights are calibrated to meet the known marginal totals available from the external source Istat. The $1+CV^2$ of the final weights increases from 3.135 to 3.687, rising by only 0.552 points, making the initial weight system increasingly granular.

Moreover, the adoption of pseudo-weights, that are not constant and equal to 1 does not result in a significant increase in the variability of the weights (from 1 to 1.672), but it also ensures adherence to the reality of the non-probabilistic design, which, being convenience-based, deviates from simple randomness and constant weights.

3.2. Comparing the estimates: which calibration model does entail the lowest bias?

The Istat survey not only provides the previously mentioned valuable known totals but also estimates of survey variables common to both the Istat survey and the OTR survey, which were deliberately included in the questionnaire of the latter. These variables include experiencing discrimination from one of the parents, following the respondents coming out, and experiencing microaggressions in the workplace (simplified here as mobbing) due to their sexual orientation.

The table below shows the OTR estimates obtained with the different weighting systems illustrated in the previous paragraph, as well as the Istat benchmark

estimates. Additionally, the relative errors, in terms of the percentage coefficient of variation (CV%) of the obtained estimates, are reported, which are understandably high due to the small sample size of the OTR.

Table 3 – *Comparison OTR and Istat estimates, varying weighting system and relative estimation errors (CV%).*

		ESTIMATES (%)						COEFFICIENT OF VARIATION (%)[3]				
Proxy Variable in common	By Sex	a.	b.	c.	d.	e.	*Istat*	a.	b.	c.	d.	e.
Bad coming out (mother)	TOT	47,7	48,0	50,0	51,4	**37,8**	*21,8*	32,5	32,1	30,9	31,8	23,2
	F	36,3	41,7	**31,3**	43,4	23,1	*28,8*	54,3	53,9	59,5	57,3	53,5
	M	54,4	**51,7**	60,8	56,0	56,0	*18,1*	41,3	41,8	37,2	39,5	21,9
Bad coming out (father)	TOT	34,5	28,1	27,0	**22,9**	31,9	*19,8*	43,2	53,1	49,9	56,9	23,9
	F	15,5	7,9	5,5	2,7	**22,0**	*18,7*	110,0	259,9	259,1	531,4	43,2
	M	45,5	39,8	39,5	**34,7**	44,0	*20,4*	46,8	51,5	47,8	52,7	27,9
Mobbing	TOT	13,2	**15,7**	14,7	15,1	28,5	*20,0*	98,1	86,9	89,9	89,5	20,4
	F	16,4	**21,2**	25,3	29,7	12,5	*21,5*	84,5	83,3	73,5	60,9	71,9
	M	11,4	**12,5**	8,6	6,7	48,4	*20,4*	154,8	137,2	175,4	228,9	14,0

Table 3 shows the propensity estimates related to the three common survey variables between the OTR survey and the Istat survey (in italics), differentiated by the biological sex of the respondent. The estimates calculated from the OTR data are obtained using the five weighting systems mentioned above, differing in how the pseudo-sampling weights have been calculated.

The differences are significant, and there is no clear method of pseudo-calibration that is unequivocally better than the others. However, what emerges is that the use of direct weights that are constant and equal to 1 never results in the estimate closest to the Istat estimates benchmark. Weighting system b., the one obtained by calibrating the constant weights to be equal to the reciprocal of the overall sampling fraction, appears to be the "best" (in bold) more frequently, but still always with substantial differences. The variable "Mobbing", related to microaggressions in the workplace, seems to be the one estimated better by OTR, perhaps also because it is the most similar, from a definition and ontology point of view, between these two questionnaires.

The second part of the table, the one on the far right, shows the relative errors in terms of CV% of the estimates just seen. The errors are extremely high, making the estimates difficult to publish. This factor inevitably limits the generalizability of the observed results and confirms how crucial sample size is, in non-probability sampling, even more so than in probability sampling.

[3] Istat estimates' coefficient of variation were not published and are not known, but are extremely lower than the OTR ones, reported in Table 3.

4. Final remarks

The previous paragraphs comprehensively analyze the challenges and strategies associated with non-probabilistic samples, particularly those collected via snowball sampling. If non-probability samples are commonly accepted and used among survey statisticians, there must be a coherent framework, such as auxiliary information to employ, and an accompanying set of measures for evaluating their quality and reliability (Kim, 2024).

One of the key takeaways is that using direct weights equal to 1, a common practice, may result in slightly lower variability of the final weights but fails to consider the specific nuances of snowball sampling, making it insensitive to the unique characteristics of the sample units.

In this context, pseudo-calibration emerges as a crucial step, as it ensures that the sample aligns with known totals at their intersections, not just their margins. This alignment seems to effectively control the increase in weight variability, which is a significant concern when dealing with non-probabilistic samples. Calibration takes this a step further by ensuring that the estimates are more closely aligned with the Istat reference values, thereby enhancing the accuracy and reliability of the models.

Importantly, the choice of the best weighting system is highly dependent on the availability of auxiliary information. This dependency underscores the necessity of having comprehensive and accurate auxiliary data to inform the weighting process.

Concluding, it is important to say that making inferences from non-probabilistic samples is indeed possible. However, it is important to emphasize the critical importance of having a sufficiently large sample size and a robust sampling strategy. These elements are essential to obtaining reliable and valid estimates, thereby addressing the inherent challenges of non-probabilistic sampling methods.

Acknowledgements

I am deeply grateful to my superiors and colleagues at Istat for granting me leave to explore this innovative topic, showcasing the institute's commitment to cutting-edge research.

Lastly, I extend my sincere thanks to all the LGBTQ individuals I interviewed: their insights were invaluable to this research.

References

ABDUL-QUADER A. S., HECKATHORN D. D., SABIN K., SAIDEL T. 2006. Implementation and analysis of respondent-driven sampling: lessons learned from the field. *Journal of Urban Health*, No. 83, pp. 1-5

BAKER R., BRICK J. M., BATES N. A., BATTAGLIA M., COUPER M. P., DEVER J. A., TOURANGEAU R. 2013. Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, Vol. 1, No. 2, pp. 90-143

DE ROSA E., DE VITIIS C., INGLESE F., VITALINI A. 2020. Il web-Respondent Driven Sampling per lo studio della popolazione LGBT+, *RIEDS - The Italian Journal of Economic, Demographic and Statistical Studies*, Vol. 74, No. 1, pp. 73-84

DEVILLE J.C., SARNDAL C.E., 1992. Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, Vol. 87, No. 418, pp. 376-382

ELLIOTT M. R. 2009. Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights. *Survey Practice*, Vol. 2, No. 6

GOLINI N., RIGHI P. 2024. Integrating probability and big non-probability samples data to produce Official Statistics. *Statistical Methods & Applications*, Vol. 33, No. 2, pp. 555-580

HECKATHORN D. D. 1997. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*,Vol. 44, No. 2, pp. 174-199

KIM K.S. 2024. Methodology of Non-Probability Samples Through Data Integration. *American Journal of Biomedical Science & Research*. Vol. 21, No. 5

KISH L. 1992. Weighting for unequal Pi. *Journal of Official Statistics*, No. 8, pp. 183-200

ISTAT. 2023. Matrimoni e unioni civili in ripresa ma ancora non ai livelli pre-pandemia. *Statistiche Report*

SHAFIE T. 2010. Design-based estimators for snowball sampling. Available at SSRN 2471006

SNIJDERS T. A. 1992. Estimation on the basis of snowball samples: how to weight?. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, Vol. 36, No. 1, pp. 59-70

TERRIBILI, M. D. 2022. Surveying the LGBTQ population (s) through social media. *AG About Gender - International Journal of Gender Studies*, Vol. 11, No. 21

VITALINI, A. 2010. L'uso delle reti sociali per la costruzione di campioni probabilistici: possibilità e limiti per lo studio di popolazioni senza lista di campionamento. Milan: Università Cattolica del Sacro Cuore

---

Marco D. TERRIBILI, Istat, terribili@istat.it