The use of sampling weights in epidemiological research: An application to the KoCo19 study

Ronan Le Gleut [1], Christiane Fuchs [1,2] & the KoCo19 study group

[1] *Core Facility Statistical Consulting, Helmholtz Munich, Germany, ronan.legleut@helmholtz-munich.de*
[2] *Faculty of Business Administration and Economics, Bielefeld University, Germany, christiane.fuchs@uni-bielefeld.de*

Résumé. Les études épidémiologiques jouent un rôle central dans la compréhension des maladies, s'appuyant souvent sur des données d'enquêtes pour tirer des conclusions. Malgré les avantages potentiels de la recherche basée sur des enquêtes, l'intégration des poids de sondage, un aspect crucial pour garantir la représentativité des données, reste sous-utilisée dans ce domaine. Les poids de sondage jouent un rôle essentiel dans l'atténuation des biais de non-réponse et l'amélioration de l'exactitude des estimations de paramètres de population calculés à partir de statistiques descriptives. Dans les modèles de régression, les pondérations d'échantillonnage abordent des problèmes tels que l'hétéroscédasticité et le biais, contribuant à la précision et à la fiabilité des estimateurs. Cependant, des défis se posent dans les stratégies de pondération, des hypothèses devant parfois être formulées sur le plan d'échantionnage ou des décisions subjectives devant potentiellement être prises. Il devient alors crucial de trouver un équilibre entre la précision et la simplicité du modèle.

Cet article présente une approche pour calculer des poids dans les études observationnelles, i.e., sans information exhaustive sur le plan de sondage. Elle se rapproche de méthodes existantes utilisant des modèles de superpopulation ou de procédures doublement robuste. La méthode part de poids égaux attribués aux observations et les ajuste par le biais de méthodes de calage utilisant de l'information sociodémographique auxiliaire. Les caractéristiques de l'échantillon sont alors alignées sur les benchmarks de la population connus, corrigeant potentiellement les biais introduits par la non-réponse ou les erreurs de couverture. La méthode peut être particulièrement utile dans les enquêtes complexes où il est difficile d'obtenir un échantillon véritablement aléatoire et représentatif.

L'application des poids (d'échantillonnage) sera démontrée à l'aide de la cohorte représentative sur la COVID-19 à Munich (KoCo19). Cette étude longitudinale, avec un plan de sondage complexe à deux degrés en grappes, étudie la prévalence du virus SARS-CoV-2 et les facteurs de risque associés à une infection.

En conclusion, bien que l'utilisation de poids d'échantillonnage dans les études épidémiologiques soit complexe, elle constitue un élément important afin d'obtenir des statistiques précises et tirer des conclusions valables. En fin de compte, l'application judicieuse de poids (de calage), associée à une communication transparente et à des analyses de sensibilité, est importante afin d'améliorer la fiabilité des conclusions dans la recherche épidémiologique. Ceci est particulièrement important pour une planification et une intervention efficaces en matière de santé publique, où il est vital d'obtenir des taux de prévalence précis et des facteurs de risque fiables.

Mots-clés. Études épidémiologiques, Poids de sondage, Méthodes de calage, Représentativité.

Abstract. Epidemiological studies play a pivotal role in understanding disease patterns, often relying on survey data to draw meaningful insights. Despite the potential advantages of survey-based research, the incorporation of sampling weights, a crucial aspect of ensuring data representativeness, remains underutilized in this domain. In the realm of descriptive statistics, sampling weights play a vital role in mitigating nonresponse bias and improving the accuracy of population parameter estimates. In regression models, sampling weights address issues like heteroscedasticity and bias, contributing to the precision and reliability of parameter estimates. However, challenges arise in weighting strategies, with assumptions about accurate sampling design information and potential subjective decisions posing hurdles. Striking a balance between precision and model simplicity becomes crucial.

The article introduces an approach for calculating weights in observational studies without exhaustive sampling design information. It is somehow related to superpopulation model or doubly robust procedures. The method starts with equal weights assigned to observations, adjusting them based on auxiliary sociodemographic information through calibration methods. This aligns sample characteristics with known population benchmarks, potentially rectifying biases introduced by nonresponse or coverage errors. The approach might be particularly valuable in complex surveys where obtaining a truly random and representative sample is challenging.

The application of (sampling) weights will be demonstrated using the representative COVID-19 cohort in Munich (KoCo19). This longitudinal population-based study, with a complex two-stage cluster sampling design, investigates SARS-CoV-2 prevalence and risk factors for infection.

In conclusion, the article emphasizes that while the use of sampling weights in epidemiological studies is complex, it is an important component for deriving accurate statistics and making valid conclusions. Ultimately, the judicious application of (calibrated) weights, coupled with transparent reporting and sensitivity analyses, is important for advancing the reliability of epidemiological research findings. This is especially relevant for effective public health planning and intervention, where securing precise prevalence rates and distributions of risk factors is vital.

Keywords. Epidemiological studies, Sampling weights, Calibration methods, Representativeness.

1 Introduction

Epidemiological studies, pivotal in understanding disease patterns and risk factors, often rely on survey data. However, the use of sampling weights is not necessarily a common practice in such studies, even though they would offer advantages in achieving accurate descriptive statistics and enhancing the validity of regression models. Sampling weights are fundamental for ensuring the representativeness of survey samples, particularly in studies with complex

designs such as stratified or cluster sampling. This is crucial in providing an accurate portrayal of the target population. Using sampling weights may also help to mitigate bias introduced by differential response rates across different subgroups of the population. In regression models, sampling weights play an important role in addressing heteroscedasticity and enhancing the precision and robustness of effect estimates, particularly in epidemiological studies with varying dispersion across subgroups. Despite their advantages, challenges exist in accurately determining the sampling design details, requiring subjective decisions in selecting appropriate weighting strategies and balancing precision with model simplicity to avoid overfitting.

This article aims to present an approach to calculate weights without requiring detailed information about the sampling design, particularly in observational studies. The application of sampling weights will be demonstrated using the representative COVID-19 cohort in Munich (KoCo19).

2 Calibration weighting for observational studies

This section presents an approach for calculating weights without requiring detailed information about the sampling design. It is somehow related to the superpopulation model procedure detailed in Valliant, Dever & Kreuter (2018).

Initially, equal weights are assigned to all observations, assuming, e.g., a simple random sampling method. However, this assumption may not hold in many cases, as certain subpopulations may be over- or under-represented. If researchers can incorporate relevant information about the sampling design, it should be utilized to mitigate bias in the analysis, even though this may introduce an element of subjectivity into weight calculations.

To refine survey weights, auxiliary sociodemographic information about the target population is utilized. Calibration methods (Deville & Särndal, 1992) are then applied to align the sample's characteristics with known population totals or benchmarks for specific variables. This calibration technique helps rectify potential biases in survey data arising from nonresponse, coverage errors, or other sampling issues, and only requires to know population totals. Adjusting weights based on known population characteristics aims to enhance the accuracy and reliability of survey estimates, rendering them more representative of the target population. This is particularly valuable in complex surveys where obtaining a truly random and representative sample is challenging.

If it is possible to compute pseudo-inclusion probabilities prior to calibrating the weights, the method can be seen as a doubly robust procedure (Kang & Schafer, 2007). This means that it is approximately unbiased with respect to the quasi-randomization distribution (pseudo inclusion probabilities), to the distribution generated by the superpopulation model (calibration), or to both. However, without any available information on the sampling design, and if the superpopulation model is misspecified, the calibrated estimate may still exhibit bias. Nonetheless, the key questions is whether the bias and the variance have been reduced compared to unweighted estimate.

3 Example: The representative COVID-19 cohort Munich (KoCo19)

3.1 The KoCo19 cohort

The SARS-CoV-2 virus emerged as a global pandemic in mid-March 2020, just three months after the initial report on December 31, 2019, from the city of Wuhan, Hubei province, China. Seeking a deeper understanding of the actual case numbers, the prospective Munich COVID-19 cohort (KoCo19) was initiated in April 2020, during the initial wave of the pandemic. The cohort comprised 5313 participants aged 13 and above, residing in private households. Over the course of the pandemic, including different waves, variant occurrences, and the commencement of vaccination campaigns, four follow-ups were conducted at critical junctures. The response rates consistently exceeded 70%. In this population-based cohort study, assessment was made on the prevalence of SARS-CoV-2 antibodies (anti-S and anti-N), and participant responses to questionnaires covering sociodemographic information and potential risk factors for infection were collected. Starting from Follow-up 2, information on SARS-CoV-2 vaccination was incorporated.

3.2 Sampling design

The participants were selected through a two-stage cluster sampling design.

The first stage involved choosing 100 out of 755 constituencies using a rejective sampling design (Hajek, 1964). Initially, each constituency had an equal probability of being included in the sample ($\sim$13%). The sample of 100 constituencies underwent scrutiny to ensure its representativeness concerning Munich's population in terms of age structure, the percentage of the population with a migration background, households with children, and households with only one member. A sample was deemed representative if the mean fractions in the sample differed from the mean fractions across all 755 Munich constituencies by less than 10 percentage points. Only samples of 100 constituencies meeting these criteria had a non-zero probability of selection. A Monte Carlo simulation with 5000 iterations for random samples of 100 constituencies indicated inclusion probabilities at the constituency level ranging from 12% to 15% (Figure 1A), suggesting that the rejection step did not introduce significant bias.

The second stage involved selecting approximately 30 households for each of the 100 drawn constituencies (Figure 1B), totaling around 3000 households in the sample. These households were obtained through random routes starting in each selected constituency, which could be considered akin to systematic sampling with equal probabilities or simple random sampling. The random routes often crossed constituency borders, allowing a household to be included in the sample via its own constituency or a neighboring one (Figure 1C). To account for these multiple ways of inclusion, we considered first and second-order neighbors (neighbors of a neighbor) for each selected constituency and applied a generalized weight share method (Deville & Lavallée, 2006) for the household weights.

Lastly, all members aged 14 years and older were requested to provide blood samples.

If participants declined to provide blood samples, the sampling weights of the consenting participants within the same household were increased to represent the other members (unit nonresponse treatment within the household).

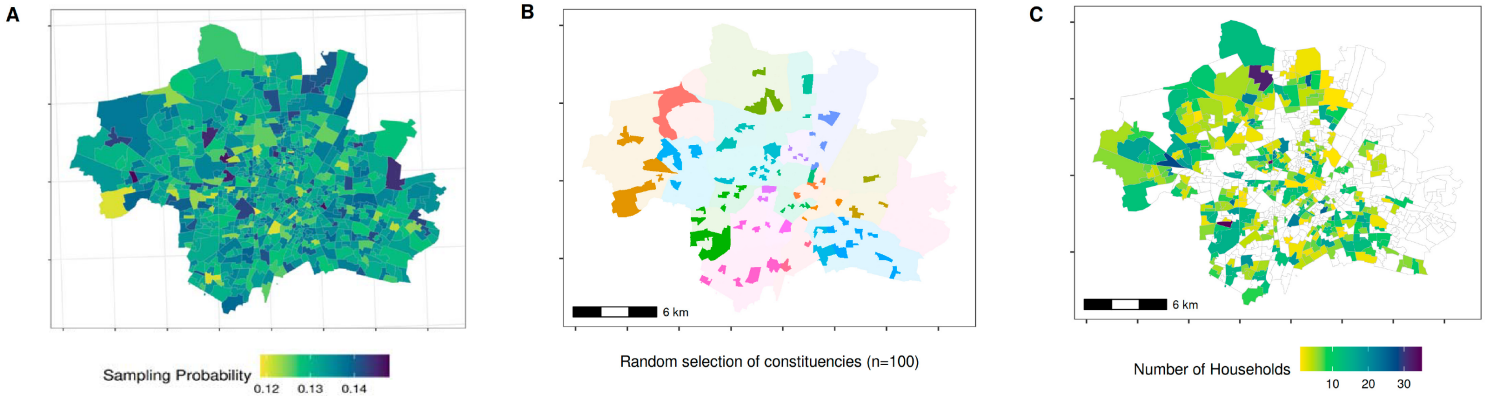

Figure 1: **A** Sampling probabilities for Munich constituencies (rejective sampling) using a Monte Carlo simulation. **B** Munich districts (distinguished by different colors) and the 100 selected start constituencies for the random walks (same color but in a darker shade). **C** All 2994 included households and their respective 368 constituencies.

3.3 Post-treatment and variance

After computing sampling weights for all participants, adjustments were made to account for attrition observed at each follow-up, modeling the underlying non-response mechanism (Särndal, Swensson & Wretman, 2003). The resulting weights underwent a final calibration based on the updated structure of the Munich population at each round, considering age, sex, country of birth, presence of children in the household, and single-member household distributions. Under some mild conditions on the sampling design, the variable of interest and the auxiliary variables used in the rejection rule, Fuller (2009) demonstrated that the regression estimator (equivalent to the calibrated estimator) for rejective sampling has similar properties to the regression estimator using the original selection procedure (simple random sampling with inclusion probabilities $\sim$13%), facilitating the calculation of variance and associated confidence intervals.

For the last three follow-ups, information on participants' vaccination status was obtained through questionnaires. Missing values (30% for Follow-up 2, 27% for Follow-up 3, and 8% for Follow-up 4) were imputed via multiple imputation ($m = 100$) using a Bernoulli distribution and crossing vaccination status with information on the immune response (anti-S and anti-N antibodies) for each round. In the last two follow-ups, approximately 93% and 97%, respectively, of participants were assumed vaccinated, in contrast to the city of Munich's reported vaccination rates of approximately 68% and 76% for the population older than 14 years. The calibration of cohort results is therefore of crucial importance.

The variance associated with the calibrated seroprevalence estimates was computed using

linearization and residual techniques (Deville, 1999). Essentially, the variance of the calibrated estimator is asymptotically equivalent to the variance of the total of the residuals of a linear regression using the linearized variable as a response and the auxiliary variables used in the calibration process as covariates. This variance (inference on finite population) accounts for uncertainty due to the different stages of the sampling design (selection of constituencies and households), the non-response mechanism (Juillard & Chauvet, 2018), and the calibration process.

Finally, the variability associated with the multiple imputation procedure was added to the variance of the seroprevalence estimates following the approach detailed in Honaker, King & Blackwell (2011). In short, the final variance estimate V is a combination of the average of the variance estimates $V_j,\ j = 1, \ldots, m$ (described above) over m replications and the variance of the m seroprevalence estimates $\theta_j,\ j = 1, \ldots, m$:

$$V = \frac{1}{m}\sum_{j=1}^{m} V_j + S^2\left(1 + \frac{1}{m}\right), \text{ with } S^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(\theta_j - \overline{\theta}\right)^2$$

The final seroprevalence estimates were obtained using the means of the m estimates, and 95% confidence intervals were computed assuming a normal distribution. Chauvet & Vallée (2020) have given general conditions for the asymptotic normality of the Horvitz-Thompson estimator under two-stage sampling designs.

In addition to accounting for the sampling design, the probabilities of the laboratory tests to yield false negatives or false positive results were considered. Following Sempos & Tian (2021), adjusted seroprevalence was calculated as $(\theta + sp - 1) / (sen + sp - 1)$, where sp represents the estimated specificity and sen represents the estimated sensitivity. While this adjustment is an exact formula for true seroprevalence, sensitivity and specificity, it becomes only approximate if the estimates are calculated and plugged in independently.

3.4 Unweighted estimates

As a sensitivity analysis, unweighted seroprevalence estimates were also computed along with their uncertainty. The variance was determined by a nonparametric cluster bootstrap procedure accounting for household clustering (Cameron, Gelbach & Miller, 2008). Seroprevalence estimates were calculated in each of the 5000 bootstrap samples (sampling households with replacement), and the variance of these estimates provided the uncertainty of the unweighted estimates. These seroprevalence estimates were also adjusted for the sensitivity and specificity of the laboratory tests.

3.5 Results

The weighted (calibrated) cumulative seroprevalence, adjusted for sensitivity and specificity, in private households for the Munich population aged 14 years and older increased from 1.6% (1.1-2.1%) in March 2020 to 12.4% (10.7-14.1%) in August 2021 and 14.5% (12.7-

16.2%) in November 2021 (Figure 2A). Without adjusting for vaccination status in Follow-ups 3 and 4, the seroprevalence would have been significantly lower: 8.5% (7.2-9.8%) for August 2021 and 10.5% (9.1-11.9%) for November 2021. Indeed, the proportion of vaccinated individuals is higher in the cohort compared to the general Munich population. Therefore, calibrating for vaccination status increases the weight of participants who are not vaccinated. As the seroprevalence is higher in the non-vaccinated population (Figure 2B), the overall seroprevalence, encompassing both vaccinated and non-vaccinated individuals, also increases with the calibration.

As a sensitivity analysis, unweighted estimates are presented in green (Figure 2A). Weighted estimates without calibration for vaccination status (in orange) and unweighted estimates are very close. Despite some weight sharing and nonresponse, the sample was already representative of the Munich population, or its characteristics do not influence the infection status (except for vaccination status).

The official number of positive cases is indicated in pink (Figure 2A) for the general population of Munich. In contrast to the KoCo19 cohort, this figure incorporates institutions such as nursing homes and encompasses potential cases of reinfection. Since the KoCo19 cohort is limited to private households and that the estimated seroprevalence does not account for multiple infections (neglected before the Omicron variant), comparing this estimate with the official number over time allows us to estimate a lower bound for the underreporting factor. The estimated underreporting factor changed over the rounds from 3.4 (2.4–4.4) at Baseline to 2.2 (2.0-2.5) at Follow-up 4.

Additional findings regarding sero-incidence, breakthrough infections, and infections among naïve subjects are available in Le Gleut et al. (2023).

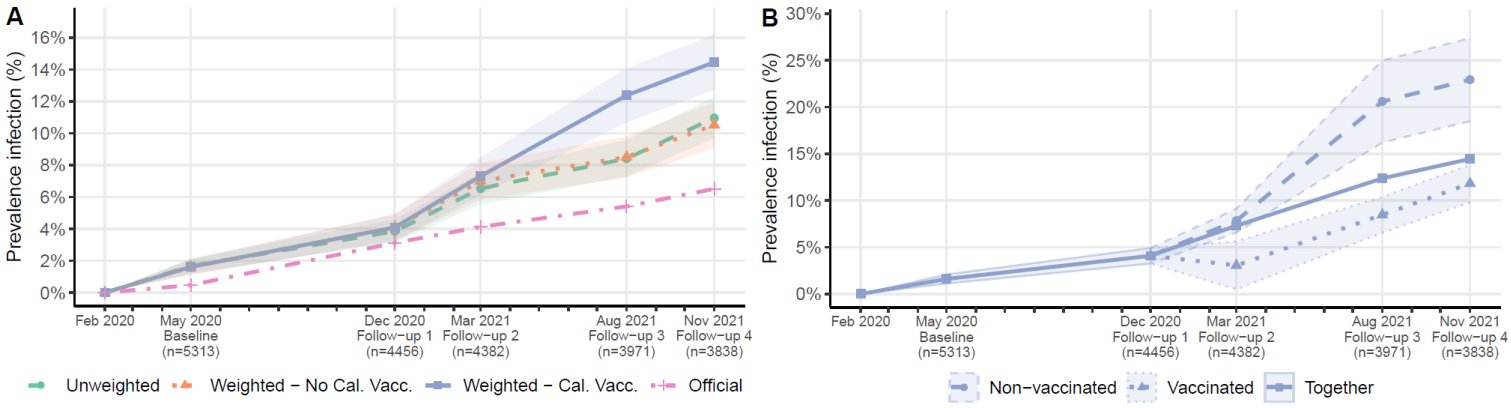

Figure 2: **A** Weighted and unweighted cumulative seroprevalence in private households and official numbers of cases reported by the authorities for the Munich population older than 13 years. **B** Seroprevalence estimates calibrated on the number of vaccinated people split according to the vaccination status of the same round.

Throughout the entire follow-up period, factors such as being born outside Germany, working in a high-risk job, and the area of residence per inhabitant were identified as infection risk factors, while other sociodemographic and health-related variables were not significant

(Le Gleut et al., 2023). This analysis utilized an extended Cox regression model (Anderson & Gill, 1982; Therneau & Grambsch, 2000) to account for intra-household clustering in the data and to obtain robust standard error estimates. During the examination of risk factors, the consideration of weights was omitted. However, in line with the concepts introduced by Solon, Haider, and Wooldridge (2015), one could have conducted appropriate diagnostics before determining the justification for using weights. A sensitivity analysis would have been pertinent as well, where the comparison between weighted and unweighted estimates (Wooldridge, 2001) could serve as a diagnostic tool for model misspecification or endogenous sampling.

Significantly more outcomes emerged from this cohort, encompassing a head-to-head evaluation of various seroassays (Olbrich et al., 2021), an integrative modeling approach to reported case numbers and seroprevalence utilizing a compartment model (Contento et al., 2023), and an evaluation of T cell reactivity following SARS-CoV-2 infection (Brand et al., 2021).

4 Discussion

In conclusion, the use of sampling weights in epidemiological studies is a complex endeavor with both advantages and challenges. While essential for achieving accurate descriptive statistics and enhancing the validity of statistical inferences, researchers must navigate the potential uncertainties associated with sampling design and subjective choices in weighting strategies. The judicious application of sampling weights, coupled with transparency in reporting, sensitivity analyses, and an understanding of study design, is essential to harness the full benefits of this methodology.

References

Brand, I., Gilberg, L., Bruger, J., Garí, M., Wieser, A., Eser, T. M., ... & Geldmacher, C. (2021). Broad T cell targeting of structural proteins after SARS-CoV-2 infection: High throughput assessment of T cell reactivity using an automated interferon gamma release assay. *Frontiers in Immunology*, 12, 688436.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics*, 90(3), 414-427.

Chauvet, G., & Vallée, A.A. (2020). Consistency of estimators and variance estimators in two-stage sampling. *J. Royal Stat. Soc.*, 82, 797–815.

Contento, L., Castelletti, N., Raimúndez, E., Le Gleut, R., Schälte, Y., Stapor, P., ... & KoCo19 study group. (2023). Integrative modelling of reported case numbers and seroprevalence reveals time-dependent test efficiency and infectious contacts. *Epidemics*, 43, 100681.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey methodology*, 25(2), 193-204.

Deville, J., & Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32(2), 165.

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418), 376-382.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4), 1491-1523.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software*, 45, 1-47.

Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, 22(4), 523 - 539.

Le Gleut, R., Plank, M., Pütz, P., Radon, K., Bakuli, A., Rubio-Acero, R., ... & Castelletti, N. (2023). The representative COVID-19 cohort Munich (KoCo19): from the beginning of the pandemic to the Delta virus variant. *BMC Infectious Diseases*, 23(1), 466.

Olbrich, L., Castelletti, N., Schälte, Y., Garí, M., Pütz, P., Bakuli, A., ... & KoCo19-Study Group. (2021). Head-to-head evaluation of seven different seroassays including direct viral neutralisation in a representative cohort for SARS-CoV-2. *Journal of General Virology*, 102(10), 001653.

Särndal, C. E., Swensson, B., & Wretman, J. (2003). Model assisted survey sampling. *Springer Science & Business Media*.

Sempos, C. T., & Tian, L. (2021). Adjusting coronavirus prevalence estimates for laboratory test kit error. *American journal of epidemiology*, 190(1), 109-115.

Solon, G., Haider, S. J., & Wooldridge, J. M. (2015). What Are We Weighting For? *Journal of Human Resources*, 50(2), 301-316.

Valliant, R., Dever, J., & Kreuter, F. (2018). Practical Tools for Designing and Weighting Survey Samples. *Springer*.

Wooldridge, J. M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric theory*, 17(2), 451-470.