Inter-American Development Bank

Dealing with Hard-to-Reach Populations in Panel Data: Respondent-Driven Survey (RDS) and Attrition

Country Department
Andean Group

Wladimir Zanoni
Jimena Romero
Nicolás Chuquimarca
Emmanuel Abuelafia

TECHNICAL NOTE Nº
IDB-TN-02800

September 2023

Dealing with Hard-to-Reach Populations in Panel Data: Respondent-Driven Survey (RDS) and Attrition

Wladimir Zanoni
Jimena Romero
Nicolás Chuquimarca
Emmanuel Abuelafia

Inter-American Development Bank
Country Department Andean Group
September 2023

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Dealing with hard-to-reach populations in panel data: respondent-driven survey (RDS) and attrition / Wladimir Zanoni, Jimena Romero, Nicolás Chuquimarca, Emmanuel Abuelafia.
p. cm. — (IDB Technical Note ; 2800)
Includes bibliographical references.
1. Demographic surveys-Ecuador. 2. Demographic surveys-Peru. 3. Immigrants-Ecuador. 4. Immigrants-Peru. I. Zanoni, Wladimir. II. Romero, Maria Jimena. III. Chuquimarca, Nicolás. IV. Abuelafia, Emmanuel. V. Inter-American Development Bank. Country Department Andean Group. VI. Series.
IDB-TN-2800

http://www.iadb.org

Dealing with Hard-to-Reach Populations in Panel Data: Respondent-Driven Survey (RDS) and Attrition*

Wladimir Zanoni[1] Jimena Romero [2] Nicolás Chuquimarca [3]
Emmanuel Abuelafia [1]

[1]Inter-American Development Bank, Washington DC
[2]Stockholm University
[3]Universidad San Francisco de Quito

September, 2023

Abstract

Hidden populations, such as irregular migrants, often elude traditional probabilistic sampling methods. In situations like these, chain-referral sampling techniques like Respondent-Driven Surveys (RDS) offer an effective solution. RDS, a variant of network sampling sometimes referred to as "snowball" sampling, estimates weights based on the network structures of friends and acquaintances formed during the sampling process. This ensures the samples are representative of the larger population. However, one significant limitation of these methods is the rigidity of the weights. When faced with participant attrition, recalibrating these weights to ensure continued representation poses a challenge. This technical note introduces a straightforward methodology to account for such attrition. Its applicability is demonstrated through a survey targeting Venezuelan migrants in Ecuador and Peru.

Keywords: Respondent driven survey, migration, hidden population, attrition
JEL Codes: C83, J11, J15

*The authors gratefully acknowledge Osmel Manzano, Tomás Bermúdez, and Felipe Muñoz for their valuable support. We are indebted to the Equilibrium team for their data collection and survey design work. We greatly appreciate Lucina Rodriguez, Cynthia Van Der Werf, Diana Rivera, and Emily Díaz aid on data handling and manuscript amendments.

1. Introduction

Hard-to-reach populations refer to specific sub-groups that are challenging to access due to factors such as physical and geographical constraints, as well as their unique social and economic situations (Shaghaghi, Bhopal, & Sheikh, 2011). Traditional probabilistic sampling methods may struggle to yield a representative sample of these populations, especially when they exhibit significant heterogeneity or are restricted by legal and cultural barriers (Shaghaghi et al., 2011; Rozo, 2021).

For instance, Rozo (2021) underscores the significance of obtaining referrals when engaging with hard-to-reach demographics. In the study of Venezuelan migrants in Colombia— a group known for its trust-related challenges— Rozo (2021) utilize a variation of snowball sampling. This approach facilitated the selection of a representative sample to analyze the impact of a comprehensive amnesty program for migrants.

Several chain-referral sampling techniques have been devised to engage with populations that are more diverse and expansive than those reachable through conventional probabilistic sampling strategies (Shaghaghi et al., 2011). A notable example is snowball sampling, which is contingent upon referrals from initially identified individuals (seeds) to induct new participants. Another method is the Respondent-Driven Survey (RDS), which calculates weights by considering the attributes of social networks during the sampling process, ensuring that the resulting sample accurately mirrors the broader population. Consequently, these methodologies enable direct sampling of the target group, proving to be both time-efficient and cost-effective compared to traditional strategies, which necessitate extensive sampling for marginal accessibility.

Nevertheless, these techniques possess inherent limitations. The weighting mechanisms are largely fixed to the initial sample composition. Thus, when subsequent surveys are conducted and participants drop out, there remains an ambiguity in adjusting these weights to maintain consistency and representation across longitudinal datasets.

The methodology introduced in this document aims to advance the understanding and application of RDS in longitudinal studies, particularly when contending with the challenge of participant attrition. The unique relevance of this approach lies in its capability to adjust and maintain the accuracy of sampling weights, even in the presence of disruptions such as attrition. As a case in point, we apply our methodology to data collected from Venezuelan migrants in Ecuador and Peru—a population that's challenging to access. The importance of this dataset is multifold: not only does it provide insight into the specific vulnerabilities and lived experiences of these migrants amidst a global pandemic, but it also serves as a tangible example of how sampling weights can be corrected and made robust using the proposed method. By doing so, our study illuminates the complexities faced by this migrant community, all the while demonstrating the utility and significance of refining sampling techniques in such contexts.

2. RDS data and weights construction

2.1 Sampling

Respondent-Driven Survey is a form of network sampling (or "snowball" sampling) that is adjusted from an analysis of network structure in order to draw unbiased and representative estimations. It is a response to a dilemma typically applied to hard-to-reach populations, in which traditional probabilistic sampling methods fail to cover the objective population, and "snowball"-like methods are prone to biases of the sampling method.

The sampling process starts with a set of initial respondents, who refer their peers, who in turn refer their peers on different waves. Because of the "small world" literature or the principle of "six-degrees of separation" (Watts, 2004); this approach could potentially reach any member of the population in only six waves theoretically. The process of RDS begins with a set of initial respondents, or seeds. The selection of the seed is typically nonrandom and directed where the 'hidden' population could be easier to reach (public venues or health centers). The seeds should be diverse and well-networked, but they do not need to be chosen randomly. The seed then refers friends and acquaintances to be interviewed, accompanied by two financial incentives: to refer peers and to complete the survey (Heckathorn, 2007).

Heckathorn (2002) compares this sampling process as a Markov chain, as it is a stochastic process with two characteristics: i) it can assume a limited number of states (e.g. classification of ethnic groups, worker industry, etc.), and ii) the process is state-dependent, where the probability of moving from one state to another depends on the transition probability matrix. If this is complied with, then the sampling process is expected to reach an equilibrium mix of recruits independent of the characteristics of the initial mix. Moreover, this convergence towards equilibrium occurs at a fast, geometric rate (Heckathorn, 1997; Kemeny & Snell, 1960).

2.2 Weights

RDS employs a specialized mathematical formula that takes into consideration the intrinsic network structure of the sampling process to derive unbiased estimators. Essentially, this approach factors in both the multiplicity and reciprocity of the sampled respondents.

Multiplicity comes into play when each respondent enumerates the eligible acquaintances known to them. This count acts as a proxy for their respective network sizes. Given that respondents with larger networks have a higher propensity to be surveyed, a weightage is applied to respondents based on the reciprocal of their declared network sizes.

The tie between the underlying network structure and the actual population estimation is anchored on the reciprocity model. This model integrates the intricate networks of acquaintances and friends within which the peer recruitment is conducted. A distinctive aspect of each participant in this model is the identification based on their recruiting peer or those they have subsequently recruited. The ensuing reciprocal linkages offer insights into the potential recruitment

pool's size, facilitating the adjustment of estimators to ensure they are asymptotically unbiased (Heckathorn, 2007).

For RDS to be effective, it is pivotal that the study population remains interconnected. Additionally, the size of this population should be sufficiently large to accommodate extensive referral chains without looping back to previous participants[1].

Heckathorn (2002) demystifies the principal concept of population proportion estimation, which can be harnessed as weights. Broadly, the process initiates by defining the ties between two individuals within the same demographic:

$$T_{a,b} = P_a N_a S_{a,b}$$

In where $T_{a,b}$ Is the number of ties between person A and B, P_a is the proportion of the population of person A based on a characteristic, N_a is the proportion of crosscutting ties, and $S_{a,b}$ is the probability that a member of group A will form a tie with a member of group B. Assuming this is the same for B ($T_{b,a} = P_b N_b S_{a,b}$) and that the number of ties between A and B is the same reciprocally:

$$P_a = \frac{S_{b,a} N_b}{S_{b,a} N_b + S_{a,b} N_a}$$

This is the estimate of the population size of A based on the reciprocity model based on two sources of data: the transition probabilities, and self-reported personal network size (Heckathorn, 2002). This is then adjusted assuming different network sizes.

3. Attrition in RDS and proposed methodology

The RDS approach, as illustrated, places considerable emphasis on the interconnectedness of sampled peers and the network size of each individual respondent. However, in constructing longitudinal series, attrition can disrupt these intricate network characteristics. Consequently, the RDS weights, designed to be representative and unbiased, may no longer hold these properties due to these disruptions.

To rectify the RDS weights in light of attrition, a two-pronged methodology is proposed:

i) First, calculate the probability of attrition for each ensuing period.

ii) Subsequently, modify the original RDS weights by taking the inverse of the retention probability.

By implementing this adjustment process, the RDS weights can be recalibrated to maintain their intended properties, even when faced with the challenges of attrition in longitudinal studies.

[1] To check further technical details on the method Mail Man School of Public Health (2016) offers a comprenhensive guide.

Step 1: Likelihood of attrition for each period

Estimate the probability of attrition using the most predictive variables from the baseline sample. The dependent variable is attrition in each of the subsequent periods and is predicted with variables from the baseline sample. This regression is weighted by the RDS initial weights.

Using a simple logistic regression, let $P(y_j)$ be the probability of attrition of period t in log-odds $\ln\left(\frac{p(y_{j,t})}{1-p(y_{j,t})}\right)$, $x_{i,j,0}$ covariates i at baseline ($t = 0$) for the individual j that determine the likelihood of attrition, and w_j the weights for individual j:

$$P(y_{j,t}) = \sum_{i=1}^{n} \beta_i x_{i,j,0} w_{j,0} \tag{1}$$

We can get a vector of estimated probabilities of attrition for each period t and individual j: $\hat{P}(y_{j,t})$

Step 2: Adjust weights on subsequent periods

Once the probabilities are estimated, they are taken to the corresponding period to adjust for the RDS weights. Each RDS weight is multiplied by the inverse of the probability of retention, or not attrition. Let $\hat{P}(y_{j,t})$ be the probability of attrition:

$$\hat{w}_{j,t} = \frac{1}{1-\hat{P}(y_{j,t})} w_{j,0} \tag{2}$$

In other words, each RDS weight is multiplied by the inverse of the estimated probability of retention, or not attrition.

4. Application: Adjusting a Longitudinal RDS Panel of Migrants

The case in point is the scenario involving irregular Venezuelan migrants in Ecuador and Peru. Characterized as a hard-to-reach population (Rozo, 2021); our efforts were concentrated on capturing an empirical snapshot of these migrants' vulnerabilities and characteristics To this end, a meticulous survey was designed and administered.

4.1 Panel RDS in Ecuador and Peru

The sampling process was initiated independently in both Ecuador and Peru, using distinct seeds for each country.

In Ecuador, the sampling process commenced with an initial set of 15 seeds. This expanded across 17 waves, ultimately culminating in 64 distinct networks and encompassing a total of 1,203 respondents.

Conversely, in Peru, the sampling began with a smaller set of 5 seeds. However, it saw expansion through 14 waves, leading to the establishment of 80 networks and capturing the responses of 1,231 individuals.

For both countries, the baseline survey was rolled out in December 2020. Three subsequent rounds of data collection were executed on a monthly basis (spanning January to March) using the Whatsapp platform. A notable observation from this exercise was the pronounced attrition observed over these rounds.

Table 1: Attrition in RDS panel: Percentage of absent sample

Ecuador	Baseline	Round 1	Round 2	Round 3	Round 4	Round 5
Respondent	1203	987	949	1000	797	763
Absent	0	216	254	203	406	440
Attrition (%)	0	22	27	20	51	58
Peru						
Respondent	1231	1004	953	958	747	680
Absent	0	227	278	273	484	551
Attrition (%)	0	23	29	28	65	81

4.2 Weights at baseline

At baseline, the weights reflect the population size for each network. The sum of the weights in Ecuador and Peru (416k and 1MM respectively as shown in table 2) are close to the estimated size of the population in each country. Most of the networks have a smaller population size showing a right-skewed distribution (see figures 1, 2). Using weights we can estimate socioeconomic characteristics as shown in tables 3 and 4, and estimate the difference between these characteristics between gender of the migrant.

Table 2: Sum of unadjusted weights

	Ecuador	Perú
Baseline	416,841	1,000,000
Round 1	326,913	785,078
Round 2	320,062	742,466
Round 3	347,176	776,422
Round 4	325,306	609,348
Round 5	277,888	521,841

Table 3: Balance table at baseline - Ecuador

	Mean	Obs	Male	Female	Diff
Age	31.757	1,203	31.739	31.768	0.03
	(10.904)		(13.162)	(12.203)	
Married	0.422	1,203	0.462	0.397	-0.06
	(0.494)		(0.003)	(0.006)	
From Caracas	0.087	1,203	0.104	0.077	-0.03
	(0.282)		(0.023)	(0.016)	
In Pichincha	0.467	1,203	0.526	0.43	-0.10**
	(0.499)		(0.002)	(0.004)	
Has children in Ecuador	0.509	1,195	0.419	0.564	0.15**
	(0.5)		(0.006)	(0.004)	
Has children in Venezuela	0.173	1,194	0.229	0.139	-0.09**
	(0.378)		(0.015)	(0.014)	
Years in Ecuador	2.003	1,203	2.079	1.956	-0.12
	(2.767)		(3.884)	(3.763)	
Children in school	0.712	585	0.721	0.707	-0.01
	(0.453)		(0.021)	(0.017)	
Has health insurance	0.109	1,176	0.121	0.102	-0.02
	(0.312)		(0.016)	(0.017)	
Education (years)	13.4	1,192	13.04	13.622	0.58*
	(3.228)		(1.281)	(0.987)	
Valid ID	0.442	1,203	0.454	0.435	-0.02
	(0.497)		(0.003)	(0.004)	
Sends remittances	0.337	1,174	0.418	0.287	-0.13**
	(0.473)		(0.006)	(0.012)	
Wage (dollars)	210.119	799	229.64	192.582	-37.06
	(180.117)		(2629.06)	(12857.807)	
Weekly laboral hours	28.177	759	34.535	22.572	-11.96***
	(21.851)		(44.312)	(39.246)	
Employed	0.24	1,144	0.347	0.172	-0.17***
	(0.427)		(0.01)	(0.015)	
Has laboral contract	0.179	305	0.148	0.218	0.07
	(0.384)		(0.026)	(0.031)	
Had COVID-19	0.053	1,149	0.062	0.047	-0.01
	(0.224)		(0.013)	(0.009)	
Received help for COVID-19	0.384	1,191	0.312	0.428	0.12**
	(0.486)		(0.012)	(0.004)	
Lost employment due to COVID-19	0.56	1,039	0.476	0.619	0.14**
	(0.497)		(0.002)	(0.008)	
Sends less remittances due to COVID-19	0.045	751	0.036	0.052	0.02
	(0.208)		(0.012)	(0.013)	

Table 4: Balance table at baseline - Perú

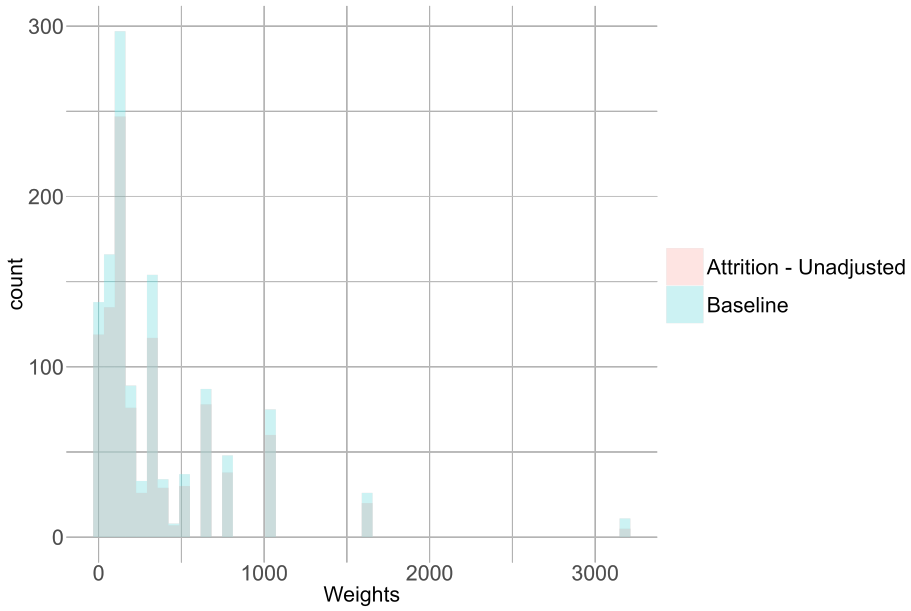
	Mean	Obs	Male	Female	Diff
Age	31.13	1,201	31.451	30.939	-0.51
	(9.953)		(10.572)	(8.173)	
Married	0.408	1,231	0.369	0.431	0.06
	(0.492)		(0.009)	(0.004)	
From Caracas	0.091	1,231	0.121	0.074	-0.05
	(0.288)		(0.021)	(0.01)	
In Lima	0.938	1,231	0.936	0.938	0.00
	(0.242)		(0.015)	(0.01)	
Has children in Perú	0.56	1,224	0.451	0.624	0.17***
	(0.497)		(0.004)	(0.007)	
Has children in Venezuela	0.214	1,226	0.255	0.19	-0.06*
	(0.41)		(0.015)	(0.014)	
Years in Perú	2.273	1,231	2.429	2.181	-0.25**
	(1.239)		(2.315)	(0.251)	
Children in school	0.517	591	0.554	0.501	-0.05
	(0.5)		(0.006)	(0)	
Has health insurance	0.15	1,202	0.132	0.16	0.03
	(0.357)		(0.015)	(0.013)	
Education (years)	12.976	1,140	12.936	12.999	0.06
	(3.277)		(1.15)	(0.959)	
Valid ID	0.514	1,231	0.566	0.483	-0.08*
	(0.5)		(0.005)	(0.001)	
Sends remittances	0.426	1,201	0.473	0.397	-0.08
	(0.495)		(0.002)	(0.006)	
Wage (soles)	622.236	832	726.46	530.469	-195.99***
	(370.275)		(13370.611)	(12772.696)	
Weekly laboral hours	31.14	784	38.964	25.028	-13.94***
	(24.751)		(44.317)	(45.916)	
Employed	0.517	1,185	0.669	0.425	-0.24***
	(0.5)		(0.012)	(0.004)	
Has laboral contract	0.101	566	0.141	0.058	-0.08**
	(0.301)		(0.02)	(0.015)	
Had COVID-19	0.138	1,130	0.128	0.145	0.02
	(0.346)		(0.019)	(0.013)	
Received help for COVID-19	0.274	1,199	0.166	0.339	0.17***
	(0.446)		(0.02)	(0.009)	
Lost employment due to COVID-19	0.514	1,032	0.437	0.569	0.13**
	(0.5)		(0.005)	(0.005)	
Sends less remittances due to COVID-19	0.891	867	0.881	0.897	0.02
	(0.312)		(0.018)	(0.019)	

4.3 Weights with attrition

For the next rounds, the presence of attrition affects the representativeness of the sample. With the attrited sample, the sum of the weights is reduced significantly in the following periods (see the rounds 1-5 in table 2).

The histograms in figures 1 and 2 show the count of the weights and baseline and how many are lost due to attrition.

Figure 1: Histogram or RDS Weights Unadjusted - Ecuador



In the subsequent analysis, we employed the attrited sample in tandem with the baseline weights to re-evaluate the socioeconomic profiles, placing a specific emphasis on gender disparities. The ramifications of weight loss and diminished representativeness become glaringly evident when we examine the balance tables derived from the attrited sample. These tables, specifically Tables 5 and 6, present the balance outcomes for both Ecuador and Peru, respectively. It is crucial to note that these tables account for the unadjusted weights in the face of attrition. Two interesting patterns emerge: (1) Some previously significant differences vanish, and (2) some differences widen. The opposite direction in both patterns demonstrates that the new estimates are affected by subject attrition in a non-clear manner.

These shifts not only underline the necessity of accurate weight adjustment but also caution against the potential pitfalls of drawing conclusions from unadjusted longitudinal datasets, especially in the context of pronounced attrition.

Figure 2: Histogram or RDS Weights Unadjusted - Peru

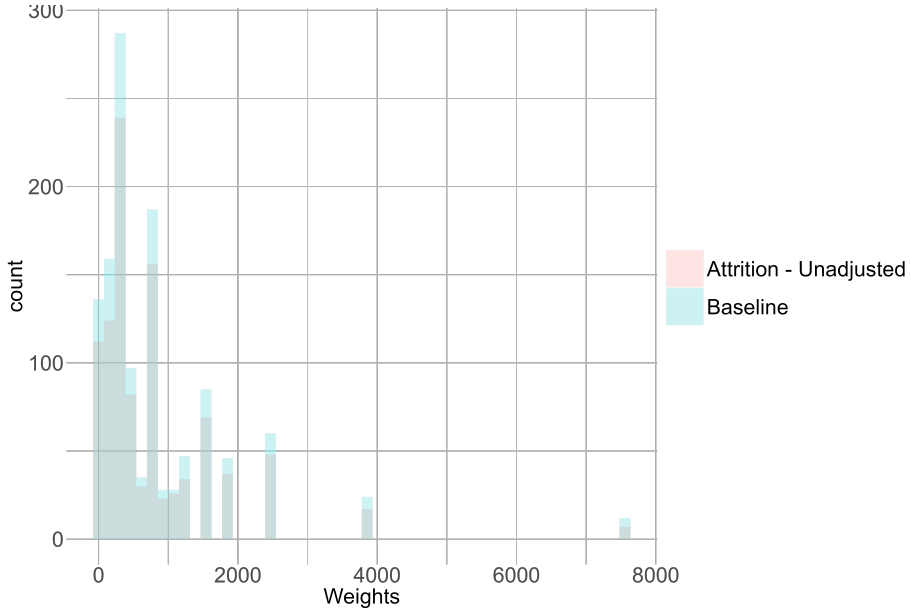


Table 5: Balance table with attrition - Ecuador

	Mean	Obs	Male	Female	Diff
Age	32.012	987	31.407	32.388	0.98
	(10.777)		(14.348)	(13.979)	
Married	0.417	987	0.452	0.394	-0.06
	(0.493)		(0.003)	(0.007)	
From Caracas	0.073	987	0.064	0.078	0.01
	(0.26)		(0.012)	(0.02)	
In Pichincha	0.437	987	0.492	0.403	-0.09*
	(0.496)		(0.001)	(0.006)	
Has children in Ecuador	0.523	980	0.451	0.567	0.12**
	(0.5)		(0.004)	(0.004)	
Has children in Venezuela	0.183	978	0.242	0.146	-0.10**
	(0.387)		(0.016)	(0.015)	
Years in Ecuador	2.006	987	2.093	1.951	-0.14
	(2.825)		(4.927)	(4.466)	
Children in school	0.726	476	0.726	0.726	0.00
	(0.447)		(0.024)	(0.019)	
Has health insurance	0.102	968	0.132	0.083	-0.05
	(0.303)		(0.019)	(0.018)	
Education (years)	13.534	978	13.234	13.72	0.49
	(3.157)		(1.059)	(1.132)	
Valid ID	0.452	987	0.467	0.443	-0.02
	(0.498)		(0.002)	(0.004)	
Sends remittances	0.366	971	0.466	0.303	-0.16***
	(0.482)		(0.002)	(0.012)	
Wage (dollars)	207.833	659	220.648	196.816	-23.83
	(183.019)		(2716.911)	(15099.727)	
Weekly laboral hours	27.213	621	33.06	21.982	-11.08***
	(21.273)		(35.411)	(38.575)	
Employed	0.229	941	0.301	0.182	-0.12**
	(0.42)		(0.013)	(0.017)	
Has laboral contract	0.173	244	0.137	0.211	0.07
	(0.379)		(0.025)	(0.036)	
Had COVID-19	0.058	945	0.071	0.049	-0.02
	(0.233)		(0.016)	(0.01)	
Received help for COVID-19	0.398	978	0.329	0.441	0.11**
	(0.49)		(0.012)	(0.004)	
Lost employment due to COVID-19	0.551	852	0.513	0.578	0.07
	(0.498)		(0.001)	(0.006)	
Sends less remittances due to COVID-19	0.046	618	0.043	0.048	0.00
	(0.209)		(0.015)	(0.015)	

Table 6: Balance table with attrition - Perú

	Mean	Obs	Male	Female	Diff
Age	31.553	977	32.119	31.219	-0.90
	(9.963)		(12.306)	(9.246)	
Married	0.413	1,004	0.379	0.432	0.05
	(0.493)		(0.01)	(0.004)	
From Caracas	0.091	1,004	0.101	0.084	-0.02
	(0.287)		(0.019)	(0.012)	
In Lima	0.932	1,004	0.928	0.934	0.01
	(0.252)		(0.018)	(0.011)	
Has children in Perú	0.595	998	0.474	0.666	0.19***
	(0.491)		(0.002)	(0.01)	
Has children in Venezuela	0.221	999	0.27	0.192	-0.08*
	(0.415)		(0.017)	(0.015)	
Years in Perú	2.296	1,004	2.484	2.186	-0.30**
	(1.333)		(2.958)	(0.305)	
Children at school	0.51	511	0.553	0.492	-0.06
	(0.5)		(0.007)	(0.001)	
Has health insurance	0.15	981	0.14	0.156	0.02
	(0.357)		(0.017)	(0.013)	
Education (years)	13.011	933	12.899	13.072	0.17
	(3.403)		(1.392)	(1.13)	
Valid ID	0.538	1,004	0.565	0.521	-0.04
	(0.499)		(0.005)	(0.001)	
Sends remittances	0.429	981	0.477	0.4	-0.08
	(0.495)		(0.002)	(0.006)	
Wage (soles)	638.341	671	740.954	544.327	-196.63***
	(376.321)		(15926.201)	(15832.137)	
Weekly laboral hours	32.175	629	40.569	25.234	-15.34***
	(24.983)		(49.409)	(48.56)	
Employed	0.504	970	0.672	0.405	-0.27***
	(0.5)		(0.014)	(0.006)	
Has laboral contract	0.116	457	0.154	0.075	-0.08**
	(0.321)		(0.022)	(0.019)	
Had COVID-19	0.153	931	0.142	0.159	0.02
	(0.36)		(0.023)	(0.014)	
Received help for COVID-19	0.286	982	0.146	0.369	0.22***
	(0.452)		(0.022)	(0.008)	
Lost employment due to COVID-19	0.534	844	0.427	0.608	0.18***
	(0.499)		(0.006)	(0.008)	
Sends less remittances due to COVID-19	0.886	707	0.868	0.898	0.03
	(0.319)		(0.022)	(0.021)	

4.4 Adjusted weights

In order to correct these problems, we implemented the two-step method proposed in this paper: i) estimate the likelihood of attrition for each round and missing individual, ii) adjust the weights on subsequent periods with the inverse of the probability of retention. After the adjustment, the weights are back to being representative of the population. The sum of the weights is similar to the population.

Table 7: Sum of adjusted weights

	Ecuador	Perú
Baseline	416,841	1,000,000
Round 1	346,455	838,764
Round 2	353,707	834,950
Round 3	356,458	847,366
Round 4	449,162	1,008,523
Round 5	450,923	1,005,575

In addition, the distribution of the weights is smoother (see figures 3 and 4). This is more evident when looking at the density distributions in figures 5 and 6. As a result, the predicted means and differences using the adjusted weights approach better the original baseline information.

Figure 3: Histogram or RDS Weights Adjusted - Ecuador

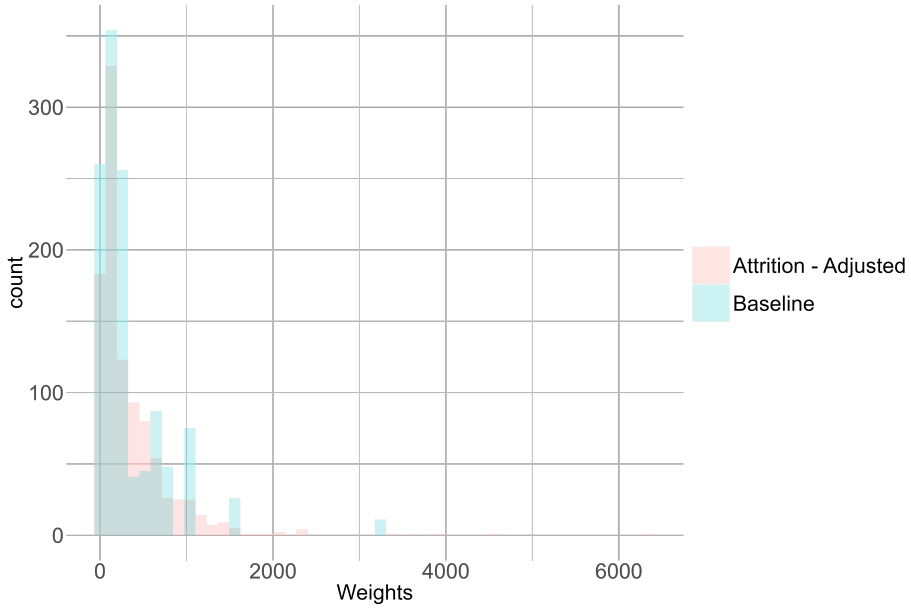


Figure 4: Histogram or RDS Weights Adjusted - Peru

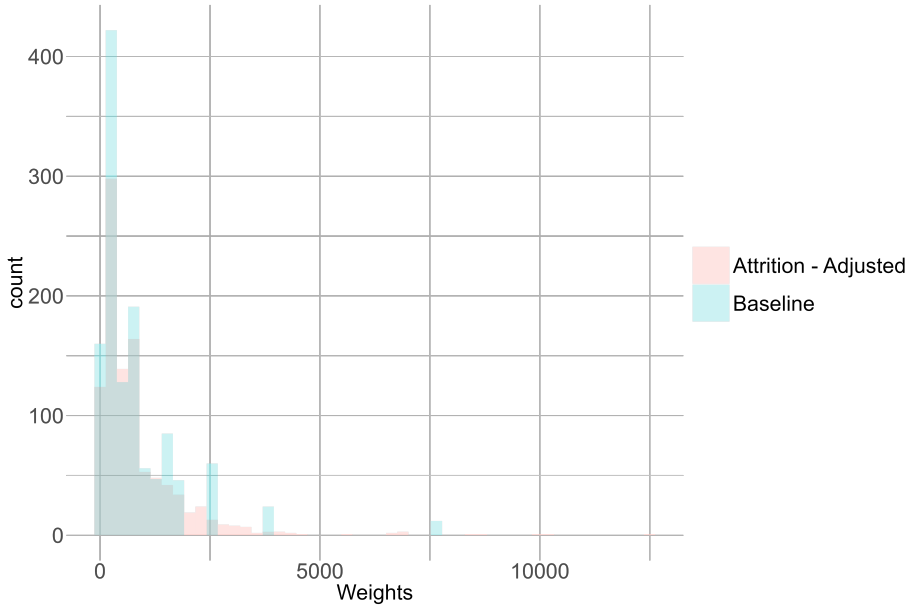


Figure 5: RDS Weights Adjusted density distribution - Ecuador

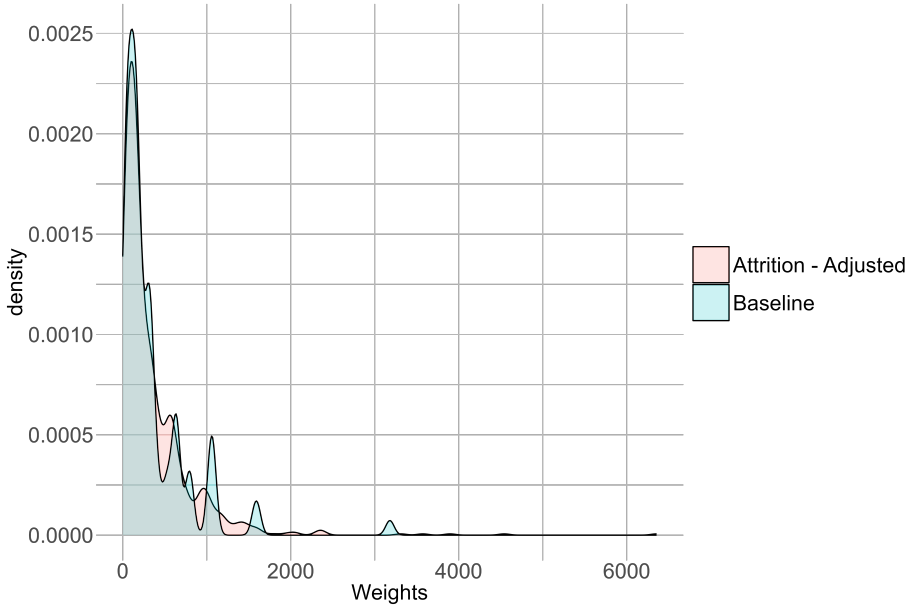


Figure 6: RDS Weights Adjusted density distribution - Peru

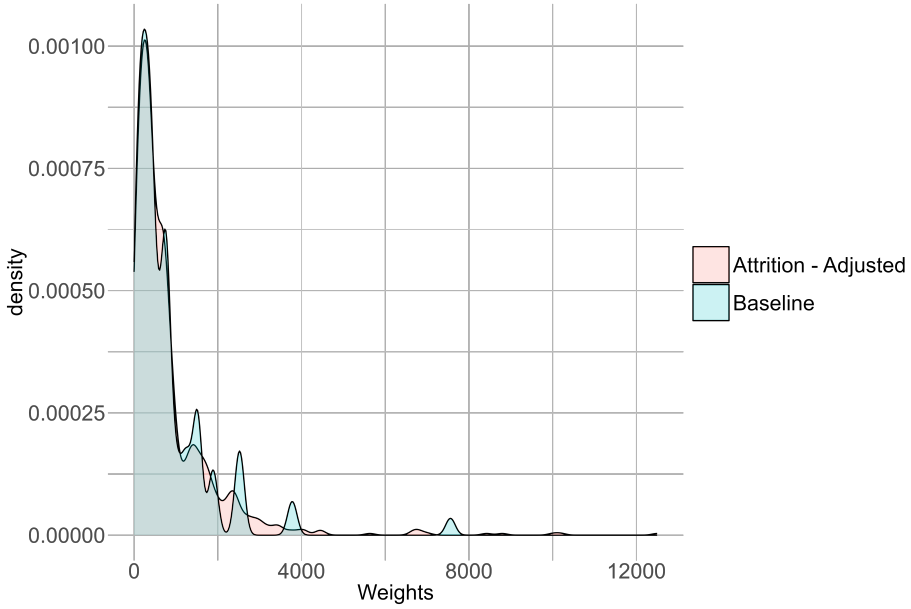


Table 8: Balance table with attrition and adjusted - Ecuador

	Mean	Obs	Male	Female	Diff
Age	32.077	987	31.495	32.426	0.93
	(10.675)		(13.478)	(13.648)	
Married	0.416	987	0.464	0.387	-0.08
	(0.493)		(0.003)	(0.008)	
From Caracas	0.082	987	0.061	0.095	0.03
	(0.275)		(0.011)	(0.027)	
In Pichincha	0.447	987	0.508	0.41	-0.10*
	(0.497)		(0.001)	(0.006)	
Has children in Ecuador	0.515	980	0.455	0.551	0.10*
	(0.5)		(0.003)	(0.004)	
Has children in Venezuela	0.173	978	0.237	0.135	-0.10**
	(0.379)		(0.017)	(0.015)	
Years in Ecuador	2.005	987	2.095	1.951	-0.14
	(2.772)		(5.159)	(3.951)	
Children in school	0.731	476	0.728	0.733	0.00
	(0.444)		(0.025)	(0.019)	
Has health insurance	0.106	968	0.144	0.082	-0.06*
	(0.307)		(0.02)	(0.018)	
Education (years)	13.62	978	13.287	13.819	0.53
	(3.183)		(1.055)	(1.124)	
Valid ID	0.468	987	0.482	0.46	-0.02
	(0.499)		(0.001)	(0.003)	
Sends remittances	0.359	971	0.467	0.294	-0.17***
	(0.48)		(0.003)	(0.013)	
Wage (dollars)	208.552	659	225.888	194.066	-31.82
	(180.928)		(2727.616)	(15127.332)	
Weekly laboral hours	27.34	621	33.317	22.148	-11.17***
	(21.226)		(37.57)	(39.219)	
Employed	0.232	941	0.297	0.191	-0.11**
	(0.422)		(0.013)	(0.021)	
Has laboral contract	0.173	244	0.152	0.194	0.04
	(0.379)		(0.025)	(0.038)	
Had COVID-19	0.052	945	0.065	0.045	-0.02
	(0.223)		(0.014)	(0.01)	
Received help for COVID-19	0.399	978	0.319	0.447	0.13**
	(0.49)		(0.013)	(0.004)	
Lost employment due to COVID-19	0.542	852	0.507	0.567	0.06
	(0.499)		(0.001)	(0.006)	
Sends less remittances due to COVID-19	0.041	618	0.043	0.04	0.00
	(0.199)		(0.015)	(0.013)	

Table 9: Balance table with attrition and adjusted - Perú

	Mean	Obs	Male	Female	Diff
Age	31.607	977	32.018	31.359	-0.66
	(10.047)		(13.514)	(9.672)	
Married	0.403	1,004	0.368	0.425	0.06
	(0.491)		(0.011)	(0.005)	
From Caracas	0.088	1,004	0.1	0.08	-0.02
	(0.283)		(0.02)	(0.012)	
In Lima	0.929	1,004	0.926	0.932	0.01
	(0.256)		(0.018)	(0.012)	
Has children in Perú	0.559	998	0.442	0.629	0.19***
	(0.497)		(0.005)	(0.009)	
Has children in Venezuela	0.215	999	0.264	0.185	-0.08*
	(0.411)		(0.018)	(0.015)	
Years in Perú	2.286	1,004	2.452	2.187	-0.27**
	(1.254)		(2.338)	(0.35)	
Children in school	0.514	511	0.57	0.491	-0.08
	(0.5)		(0.01)	(0.001)	
Has health insurance	0.145	981	0.135	0.151	0.02
	(0.352)		(0.017)	(0.013)	
Education (years)	12.939	933	12.785	13.024	0.24
	(3.446)		(1.586)	(1.16)	
Valid ID	0.536	1,004	0.557	0.523	-0.03
	(0.499)		(0.005)	(0.001)	
Sends remittances	0.432	981	0.474	0.406	-0.07
	(0.496)		(0.002)	(0.006)	
Wage (soles)	635.705	671	739.492	540.553	-198.94***
	(382.114)		(16162.816)	(17070.231)	
Weekly laboral hours	32.142	629	40.514	25.142	-15.37***
	(24.997)		(52.986)	(52.14)	
Employed	0.497	970	0.66	0.4	-0.26***
	(0.5)		(0.015)	(0.007)	
Has laboral contract	0.116	457	0.155	0.073	-0.08**
	(0.321)		(0.022)	(0.018)	
Had COVID-19	0.14	931	0.135	0.144	0.01
	(0.348)		(0.026)	(0.014)	
Received help for COVID-19	0.276	982	0.143	0.357	0.21***
	(0.447)		(0.025)	(0.009)	
Lost employment due to COVID-19	0.53	844	0.429	0.599	0.17**
	(0.499)		(0.006)	(0.007)	
Sends less remittances due to COVID-19	0.878	707	0.868	0.884	0.02
	(0.328)		(0.022)	(0.028)	

4.5 How much better?

In the upcoming table, we juxtapose gender-based socioeconomic differences derived from three distinct data sets: the baseline sample, the attrited sample with unadjusted weights, and the attrited sample adjusted using our proposed weighting methodology.

For Ecuador, the sum of squared errors for the attrited sample totals 176.730 across all variables. However, with the adjusted weights, this figure dramatically decreases to 28.915. It's important to emphasize that the bulk of this discrepancy is attributed to key variables, such as age, wage, and work hours. Conversely, for Peru, the results are more varied.

Table 10: Errors of gender difference - Ecuador

Variable	Baseline Difference	Attrition Difference Unadjusted	Attrition Difference Adjusted	Sum of Squared Errors Unadjusted	Sum of Squared Errors Adjusted
Age	0.03	0.98	0.93	0.902	0.810
Married	-0.06	-0.06	-0.08	0.000	0.000
From Caracas	-0.03	0.01	0.03	0.002	0.004
In Pichincha	-0.10**	-0.09*	-0.10*	0.000	0.000
Has children in Ecuador	0.15**	0.12**	0.10*	0.001	0.002
Has children in Venezuela	-0.09**	-0.10**	-0.10**	0.000	0.000
Years in Ecuador	-0.12	-0.14	-0.14	0.000	0.000
Children in school	-0.01	0.00	0.00	0.000	0.000
Has health insurance	-0.02	-0.05	-0.06*	0.001	0.002
Education (years)	0.58*	0.49	0.53	0.008	0.002
Valid ID	-0.02	-0.02	-0.02	0.000	0.000
Sends remittances	-0.13**	-0.16***	-0.17***	0.001	0.002
Wage (dollars)	-37.06	-23.83	-31.82	175.033	27.458
Weekly laboral hours	-11.96***	-11.08***	-11.17***	0.774	0.624
Employed	-0.17***	-0.12**	-0.11**	0.003	0.004
Has laboral contract	0.07	0.07	0.04	0.000	0.001
Had COVID-19	-0.01	-0.02	-0.02	0.000	0.000
Received help for COVID-19	0.12**	0.11**	0.13**	0.000	0.000
Lost employment due to COVID-19	0.14**	0.07	0.06	0.005	0.006
Sends less remittances due to COVID-19	0.02	0.00	0.00	0.000	0.000
Sum of squared errors				176.73	28.915

5. Concluding Remarks

Since its seminal introduction by Heckathorn (1997), the Respondent Driven Sampling (RDS) methodology has witnessed considerable advancements and found varied applications across a range of research domains. Innovations like Wejnert and Heckathorn (2008)'s leveraging of the internet for RDS harnessed the web's expansive reach for rapid information dissemination. Schonlau and Liebau (2012) further eased the application by integrating a user-friendly RDS

Table 11: Errors of Gender Difference - Perú

Variable	Baseline Difference	Attrition Difference Unadjusted	Attrition Difference Adjusted	Sum of Squared Errors Unadjusted	Sum of Squared Errors Adjusted
Age	-0.51	-0.90	-0.66	0.152	0.023
Married	0.06	0.05	0.06	0.000	0.000
From Caracas	-0.05	-0.02	-0.02	0.001	0.001
In Lima	0.00	0.01	0.01	0.000	0.000
Has children in Perú	0.17***	0.19***	0.19***	0.000	0.000
Has children in Venezuela	-0.06*	-0.08*	-0.08*	0.000	0.000
Years in Perú	-0.25**	-0.30**	-0.27**	0.002	0.000
Children in school	-0.05	-0.06	-0.08	0.000	0.001
Has health insurance	0.03	0.02	0.02	0.000	0.000
Education (years)	0.06	0.17	0.24	0.012	0.032
Valid ID	-0.08*	-0.04	-0.03	0.002	0.003
Sends remittances	-0.08	-0.08	-0.07	0.000	0.000
Wage (soles)	-195.99***	-196.63***	-198.94***	0.410	8.702
Weekly laboral hours	-13.94***	-15.34***	-15.37***	1.960	2.045
Employed	-0.24***	-0.27***	-0.26***	0.001	0.000
Has laboral contract	-0.08**	-0.08**	-0.08**	0.000	0.000
Had COVID-19	0.02	0.02	0.01	0.000	0.000
Received help for COVID-19	0.17***	0.22***	0.21***	0.002	0.002
Lost employment due to COVID-19	0.13**	0.18***	0.17**	0.002	0.002
Sends less remittances due to COVID-19	0.02	0.03	0.02	0.000	0.000
Sum of squared errors				2.544	10.811

module within the Stata software, streamlining weight estimations.

Over the years, RDS has been instrumental in mapping elusive populations: from HIV patients and migrants to sex workers, LGBTQ+ individuals, and specific health-afflicted communities. Distinctive applications such as those by Tyldum (2021) on Central/Eastern European migrants and Michaels, Pineau, Reimer, Ganesh, and Dennis (2019) on the LGBTQ+ populace in the U.S. underscore RDS's adaptability and scope.

However, it's pivotal to recognize RDS's constraints. Often a fallback in many research undertakings, RDS becomes the method of choice primarily when traditional probability sampling proves impractical Heckathorn (2002); Shaghaghi et al. (2011); Tyldum (2021). The methodology grapples with challenges like ascertaining respondents' truthful reporting of their personal network size and achieving randomness through extensive referral chains. Such nuances emphasize the judicious deployment of RDS in academic pursuits.

Our research contributes by highlighting the nuances of studying elusive populations using panel data. We present a novel method to recalibrate population weights amidst participant attrition. By unveiling a weight adjustment approach grounded in observable data, we aim to offer more accurate population projections. This method hinges on the assumption that observed longitudinal attrition is non-random, driven by specific attributes influencing response rates in

subsequent phases.

The insights in this paper lay the groundwork for a broader research trajectory, aimed at circumventing inherent procedural limitations. For one, our weight recalibration approach operates with a 'memory-free' premise, determining attrition probabilities solely from the extant period's data. While efficient, it may overlook potential serial correlations in these probabilities. Another area warranting refinement is our baseline assumption of RDS reaching equilibrium. To fortify our analyses, an inaugural convergence assessment for pertinent variables is advisable. Such avenues signal further research opportunities to refine and fortify our methodology.

References

Heckathorn, D. D. (1997, May). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, *44*(2), 174–199. Retrieved 2021-08-15, from https://academic.oup.com/socpro/article-lookup/doi/10.2307/3096941 doi: 10.2307/3096941

Heckathorn, D. D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, *49*(1), 11–34. Retrieved 2021-09-09, from https://www.jstor.org/stable/10.1525/sp.2002.49.1.11 (Publisher: [Oxford University Press, Society for the Study of Social Problems]) doi: 10.1525/sp.2002.49.1.11

Heckathorn, D. D. (2007, August). 6. Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, *37*(1), 151–208. Retrieved 2021-07-15, from http://journals.sagepub.com/doi/10.1111/j.1467-9531.2007.00188.x doi: 10.1111/j.1467-9531.2007.00188.x

Kemeny, J. G., & Snell, J. L. (1960). *Finite markov chains*. (Van Nostrand.

Mail Man School of Public Health, C. (2016, Aug). *Respondent-driven sampling*. Retrieved from https://www.publichealth.columbia.edu/research/population-health-methods/respondent-driven-sampling

Michaels, S., Pineau, V., Reimer, B., Ganesh, N., & Dennis, J. M. (2019). Test of a hybrid method of sampling the lgbt population: Web respondent driven sampling with seeds from a probability sample. *Journal of Official Statistics*, *35*(4), 731–752.

Rozo, S. (2021). *Tips for collecting surveys of hard-to-reach populations*. Retrieved 2021-07-26, from https://blogs.worldbank.org/impactevaluations/tips-collecting-surveys-hard-reach-populations

Schonlau, M., & Liebau, E. (2012). Respondent-driven sampling. *The Stata Journal*, *12*(1), 72–93. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/1536867X1201200106

Shaghaghi, A., Bhopal, R. S., & Sheikh, A. (2011, December). Approaches to recruiting 'hard-to-reach' populations into research: A review of the literature. *Health Promotion Perspectives*, *1*(2), 86–94. Retrieved 2021-07-26, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3963617/ doi: 10.5681/hpp.2011.009

Tyldum, G. (2021). Surveying migrant populations with respondent-driven sampling. Experiences from surveys of east-west migration in Europe. *International Journal of Social Research Methodology*, *24*(3), 341–353.

Watts, D. (2004). *Six degrees: The science of a connected age*. W. W. Norton. Retrieved from https://books.google.se/books?id=1gueFWR7qjoC

Wejnert, C., & Heckathorn, D. D. (2008). Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods & Research*, *37*(1), 105–134. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0049124108318333?casa_token=cHU3VgCcKzAAAAAA:wj7w67jW34yBnf6xXFrXiGrqf3qwe4Po4HSDlnYTnOHPyYMetEy4rHTs55tmMncr8hdcRbhulirbopw doi: 10.1177/0049124108318333