Panel Data Research Center, Keio University

PDRC Discussion Paper Series

Japan Household Panel Survey (JHPS/KHPS) Sampling Weights

Mateus Silva Chang, Guillaume Osier, Kayoko Ishii

28 November, 2022

DP2022-003
https://www.pdrc.keio.ac.jp/en/publications/dp/8199/

Panel Data Research Center, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
info@pdrc.keio.ac.jp
28 November, 2022

Japan Household Panel Survey (JHPS/KHPS) Sampling Weights

Mateus Silva Chang, Guillaume Osier, Kayoko Ishii

PDRC Keio DP2022-003
28 November, 2022
JEL Classification: C8; C80; C83
Keywords: JHPS/KHPS; Sampling Weights; Weight calculation; Calibration; Survey

Abstract

Sampling weights are used to make inferences about the target population based on a specific sample. Given the importance of weighting survey observations when drawing inferences about the overall population, this paper documents the calculation of sampling weights for the Japan Household Panel Survey (JHPS/KHPS). First, we provide an overview of the sample design and structure of the JHPS/KHPS. This information is then used as a base to define the strategy adopted in the weight calculation process. Next, the integrated approach used to compute sampling weights for the initial fourteen waves (2004-2017) of the JHPS/KHPS and the different types of weights available are introduced. Finally, we provide advice on how the weights should be used and illustrate their effectiveness by comparing unweighted and weighted JHPS data with official statistics.

Mateus Silva Chang
Faculty of Economics, Keio University
2-15-45 Mita, Minato-ku, Tokyo
chang.mateus@keio.jp

Guillaume Osier
Statistics Luxembourg (STATEC)
13 Rue Erasme, 1468 Luxembourg

Kayoko Ishii
Faculty of Economics, Keio University
2-15-45 Mita, Minato-ku, Tokyo

Acknowledgement: The authors gratefully acknowledge the Japanese National Statistics Center (NSTAC) for providing the data from the Labor Force Survey from the Statistics Bureau of Japan that was necessary for the calculation of the Japanese population benchmark used in the weight calibration process. We also thank Akira Fukuda and Wenqing Chen for their support regarding the preparation of the figures for Section 8. This work was supported by Grant-in-Aid for Specially Promoted Research 17H06086 (2017/04/25 - 2022/03/31) and JSPS Program for Constructing Data Infrastructure for the Humanities and Social Sciences (R3-R4).

Japan Household Panel Survey (JHPS/KHPS) Sampling Weights[1]

Mateus Silva Chang
Keio University

Guillaume Osier
Statistics Luxembourg (STATEC)

Kayoko Ishii
Keio University

Summary

Sampling weights are used to make inferences about the target population based on a specific sample. Given the importance of weighting survey observations when drawing inferences about the overall population, this paper documents the calculation of sampling weights for the Japan Household Panel Survey (JHPS/KHPS). First, we provide an overview of the sample design and structure of the JHPS/KHPS. This information is then used as a base to define the strategy adopted in the weight calculation process. Next, the integrated approach used to compute sampling weights for the initial fourteen waves (2004-2017) of the JHPS/KHPS and the different types of weights available are introduced. Finally, we provide advice on how the weights should be used and illustrate their effectiveness by comparing unweighted and weighted JHPS data with official statistics.

[1] The authors gratefully acknowledge the Japanese National Statistics Center (NSTAC) for providing the data from the Labor Force Survey from the Statistics Bureau of Japan that was necessary for the calculation of the Japanese population benchmark used in the weight calibration process. We also thank Akira Fukuda and Wenqing Chen for their support regarding the preparation of the figures for Section 8. This work was supported by Grant-in-Aid for Specially Promoted Research 17H06086 (2017/04/25 - 2022/03/31) and JSPS Program for Constructing Data Infrastructure for the Humanities and Social Sciences (R3-R4).

1. Introduction

Ideally, the collection of data from all individuals in a target population would be the best option to study population characteristics, evaluate implemented policies or analyze the dynamic behavior of individuals and households. However, given time and resource constraints, the survey is a powerful alternative to the census as a tool to quickly gain information and insights from the target population. A survey is designed to guarantee that collecting data from a subgroup of the population, identified as a sample, will still enable us to draw inferences about the entire target population.

Despite the development of a careful sample design and the efforts made to implement the survey according to an established plan, unexpected problems and difficulties in collecting the data can lead to sample selection bias. In addition, in the case of panel data surveys, nonresponse and attrition are also problems that can affect the representativeness of the collected data. These distortions can be adjusted by a correction technique called weighting.

Sampling weights are values attributed to each sample observation to ensure that the metrics derived from the survey data are representative of the whole target population. The calculation of these weights must take into consideration the main features of the survey design, such as selection probabilities, the existence of regional clusters or unit nonresponse. Population benchmarks taken from external sources can also be used to adjust the sampling weights to generate more stable survey estimations. This adjustment process is known as calibration.

Given the importance of weighting survey observations when drawing inferences about the overall population, this paper documents the calculation of sampling weights to the Japan Household Panel Survey (JHPS/KHPS). In the first section, an overview of the sample design and structure of the JHPS/KHPS is provided. Next, the integrated approach used to compute sampling weights for the JHPS/KHPS and the different types of weights available are introduced. Finally, we provide advice on how the weights should be used and illustrate their effectiveness by comparing unweighted and weighted data with official statistics.

2. Overview and structure of the Japan Household Panel Survey (JHPS/KHPS)

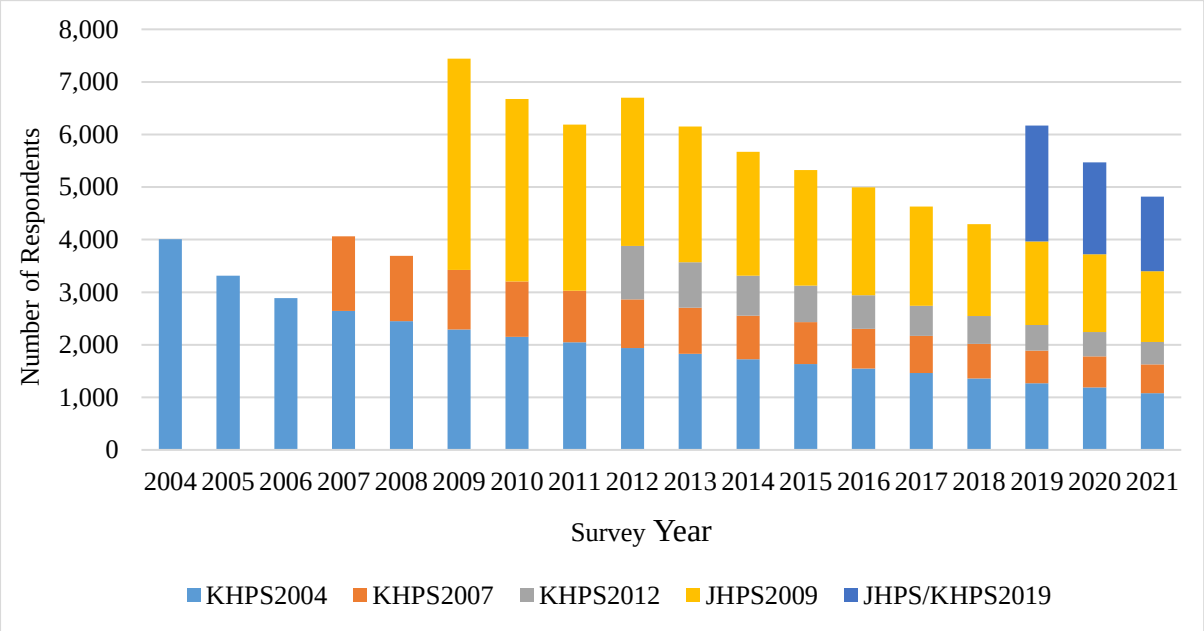
The Japan Household Panel Survey (JHPS/KHPS) was established in 2014 as a result of the integration of the Keio Household Panel Survey (KHPS), a survey that has been implemented since 2004, and the Japan Household Panel Survey (JHPS), a survey that was

introduced in 2009. The main objective of the JHPS/KHPS is to provide data that represent the Japanese population, thereby allowing the analyses of dynamic behavior by economic entities. The survey covers comprehensive topics such as household structure, individual attributes, academic background, employment status, time use, health condition, well-being, income, wealth and others.

2.1 Data Structure

The first wave of the KHPS was conducted in 2004 and collected information from a sample of 4,005 respondents aged 20 to 69 years. The KHPS sample was extended through the recruitment of an additional 1,419 individuals in 2007 and 1,012 more individuals in 2012. The first wave of the JHPS was conducted in 2009 and obtained data from 4,022 respondents aged 20 and over. Due to the similarity of these two surveys, the KHPS and the JHPS were combined in 2014 and named the "Japan Household Panel Survey (JHPS/KHPS)", thereby indicating the adoption of a common questionnaire. After the integration, the JHPS/KHPS received a top-up sample of 2,203 respondents aged 20 and over in 2019. Figure 1 shows the total sample sizes of the JHPS/KHPS for each survey year.

Figure 1. Sample size of the Japan Household Panel Survey (JHPS/KHPS)



Source: Authors, based on the JHPS/KHPS.

2.2 Sample Design

The survey subjects were restricted to individuals living in private dwellings in Japan and were selected according to a two-stage stratified random sampling method. For the case of

the KHPS, the samples were limited to individuals aged 20 to 69 years in the first wave, while for the JHPS, the samples were limited to individuals aged 20 and over in the first wave.[2]

To implement the two-stage stratified random sampling method, Japan was stratified into 24 strata following a regional and municipal classification. The number of subjects in each stratum was allocated in proportion to the registered population according to the Basic Resident Register of the previous year.[3] Next, the districts inside each stratum were selected following a systematic random sampling process, and an average of 10 subjects per district were selected until the predefined total number of subjects per stratum was achieved. Then, these subjects were also randomly sampled.

To guarantee the necessary number of respondents, reserve subjects were selected and used, when necessary, to replace original subjects who could not be contacted or declined to participate in the survey. For each original subject, between 3 and 5 reserve subjects were selected from the same surveyed district. Although these reserve subjects were also randomly selected, they had characteristics similar to those of the original subjects they were supposed to replace; i.e., they were from the same sex and age group (20s, 30s, 40s, 50s, and 60s for the KHPS sample and 20s, 30s, 40s, 50s, 60s and over for the JHPS sample).

2.2.1 Sample design of the top-up samples

In light of sample attrition problems, top-up samples were added to the KHPS in 2007 and 2012. The sample selection for the KHPS2007 sample and the KHPS2012 sample followed the sampling method adopted for the KHPS2004 sample. Two-stage stratified random sampling was adopted, with Japan being stratified into the same 24 strata and the number of subjects in each stratum being allocated in proportion to the registered population according to the Basic Resident Register of the previous year.

2.3 Survey methods and respondent tracking

The JHPS/KHPS uses the drop-off and pick-up (DOPU) method.[4] Regarding the drop-off and pick-up survey, an investigator visits the respondents, distributes the questionnaires and then once again visits the respondents on a later date to collect the completed questionnaires. If a subject is absent at the time of visitation, then the investigator attempts to

[2] The age range of the respondents were defined based on the age they completed in the year the survey was conducted and not on the age they were when they were selected.
[3] For example, on the KHPS first wave it was used the Basic Resident Register from March 31, 2003.
[4] In the first years of the JHPS several experiments were performed to see the effects of different survey methods on response rates. The experiments included, for instance, providing web survey option, conducting interview instead of paper-based questionnaire, and introducing incentive reward to investigators. Naoi, Yamamoto and Miyauchi (2010) examines the effects of each experiment on response rate of original subjects.

contact the subject via a different means of communication (such as leaving a note); all procedures are recorded. Usually, the survey is conducted at the beginning of February every year.

After the first wave, each selected subject was expected to be tracked. If in the following years a respondent has moved out, he or she is tracked to his or her new living location. In the event that a respondent has died or disappeared, his or her spouse can participate by serving as a substitute for the respondent in the survey; in such cases, the spouse is added as a new respondent and receives a new ID number. For cases where the respondent is not able to cooperate with the survey in a given year and has not clearly expressed his or her intention of not cooperating in following years, the investigator tries to contact him or her in subsequent years. For cases where such a respondent does cooperate in one of the following years, he or she is counted as a revival case.

2.4 The nonresponse issue

In general, all sampled subjects are requested to participate in a survey. However, it is common to have cases where the selected subject does not participate and is considered a nonrespondent. For the case of the KHPS and the JHPS, a different strategy was adopted, with both original and reserve subjects being prepared to guarantee the scheduled sample size. Therefore, it is not possible to calculate the response rate in a straightforward manner, as it is usually calculated in the first wave of other surveys.

An alternative way was to calculate the quasi-response rate by referring to the "Investigator Check Sheet" of each subject. First, the total number of original and reserve subjects contacted by the investigators is identified. Next, the quasi-response rate is calculated by using this number as the denominator and the number of total responses as the numerator. Table 1 displays the quasi-response rate for the first wave of each panel.

Table 1. The quasi-response rate for the first wave

		KHPS			JHPS	
		2004	2007	2012	2009	2019
(1)	Total number of subjects (original and reserve subjects)	13,430	5,868	3,183	12,549	9,465
(2)	Number of valid response[1]	4,005	1,419	1,012	4,022	2,203
(3)	Number of subjects contacted by the investigators[2]	9,665	4,223	2,425	10,075	6,585
(4)	In-touch rate [(3)/(1)*100]	72.0%	72.0%	76.2%	80.3%	69.6%
(5)	Quasi-response rate [(2)/(3)*100]	41.4%	33.6%	41.7%	39.9%	33.5%

Source: Authors, based on the JHPS/KHPS.

[1] Include reserve subjects.

[2] Given that the "Investigator Check Sheet" was not implemented in KHPS2004, the in-touch rate of KHPS2004 is calculated based on the in-touch rate of KHPS2007.

From the second wave onward, the panel attrition issue refers to the loss of sample subjects. The response rates after the second wave are calculated as the number of respondents divided by the number of subjects who responded to the survey in the previous year. Table 2 and Table 3 show the response rates for both the KHPS and the JHPS.

Table 2. Response rate of the KHPS

		2005 wave 2	2006 wave 3	2007 wave 4	2008[1] wave 5	2009 wave 6	2010 wave 7
(1)	Number of subjects	4,005	3,342	2,894	4,067	3,706	3,448
(2)	of which responded in last wave	4,005	3,314	2,887	4,062	3,691	3,422
(3)	of which revivals	0	28	7	5	15	26
(4)	Number of respondents	3,314	2,887	2,643	3,691	3,422	3,207
(5)	of which revivals	0	0	3	0	4	7
(6)	Response rate [((4)-(5))/(2)*100]	82.7%	87.1%	91.4%	90.9%	92.6%	93.55%
		2011 wave 8	2012 wave 9	2013[2] wave 10	2014 wave 11	2015 wave 12	2016 wave 13
(1)	Number of subjects	3,232	3,041	3,888	3,587	3,353	3,152
(2)	of which responded in last wave	3,207	3,030	3,877	3,568	3,312	3,124
(3)	of which revivals	25	11	11	19	41	28
(4)	Number of respondents	3,030	2,865	3,568	3,312	3,124	2,945
(5)	of which revivals	10	10	11	7	16	10
(6)	Response rate [((4)-(5))/(2)*100]	94.2%	94.2%	91.7%	92.6%	93.8%	94.0%
		2017 wave 14	2018 wave 15	2019 wave 16	2020 wave 17	2021 wave 18	
(1)	Number of subjects	2,970	2,769	2,572	2,409	2,281	
(2)	of which responded in last wave	2,945	2,741	2,549	2,378	2,244	
(3)	of which revivals	25	28	23	31	37	
(4)	Number of respondents	2,741	2,549	2,378	2,244	2,054	
(5)	of which revivals	11	14	8	14	19	
(6)	Response rate [((4)-(5))/(2)*100]	92.7%	93.0%	93.0%	93.8%	90.7%	

Source: Authors, based on the JHPS/KHPS.
[1] From 2008, this includes the 2007 top-up sample.
[2] From 2013, this includes the 2012 top-up sample.

Table 3. Response rate of the JHPS

		2010 wave 2	2011 wave 3	2012 wave 4	2013 wave 5	2014 wave 6	2015 wave 7
(1)	Number of subjects	4,022	3,490	3,170	2,839	2,596	2,384
(2)	of which responded in last wave	4,022	3,470	3,160	2,821	2,581	2,358
(3)	of which revivals	0	20	10	18	15	26
(4)	Number of respondents	3,470	3,160	2,821	2,581	2,358	2,198
(5)	of which revivals	0	6	4	8	6	6
(6)	Response rate [((4)-(5))/(2)*100]	86.3%	90.9%	89.1%	91.2%	91.1%	93.0%

		2016 wave 8	2017 wave 9	2018 wave 10	2019 wave 11	2020[1] wave 12	2021 wave 13
(1)	Number of subjects	2,211	2,060	1,897	1,759	3,824	3,328
(2)	of which responded in last wave	2,198	2,048	1,885	1,742	3,792	3,226
(3)	of which revivals	13	12	12	17	32	101
(4)	Number of respondents	2,048	1,885	1,742	1,589	3,226	2,763
(5)	of which revivals	9	3	4	7	14	25
(6)	Response rate [((4)-(5))/(2)*100]	92.8%	91.9%	92.2%	90.8%	84.7%	84.9%

Source: Authors, based on the JHPS/KHPS.
[1] From 2020, this includes the 2019 top-up sample.

3. Weight computation strategy

Hereafter, we describe how to develop weights for the first 14 waves (2004-2017) of the JHPS/KHPS. The computation of sampling weights for the JHPS/KHPS relies on "individual base weights", which are the backbone of the calculation procedure. Since the sample unit of the JHPS/KHPS is individual, the individual base weights are defined for the first wave of each sample: KHPS2004, KHPS2007, KHPS2012, and JHPS2009. For these waves, the base weights are calculated as the inverse of the selection probabilities of each individual in the sample. Next, these probabilities are adjusted for unit nonresponse and then calibrated to external population benchmarks. From the second wave onward, the base weights must be corrected to address panel attrition. These adjustments aim to maintain the representativeness of the data in the later years.

Next, the weights calculated for different waves and samples can be combined to produce different sets of JHPS/KHPS weights. The derived groups of weights are classified according to Table 4.

Table 4. Computed weights

Weight Type	Base	Working Sample
Longitudinal weights	KHPS 2004	Individuals initially selected for the KHPS2004 sample
	KHPS 2007	Individuals initially selected for the KHPS2004 sample and those selected for the KHPS2007 top-up sample
	KHPS 2012	Individuals initially selected for the KHPS2004 sample and those selected for the two KHPS top-up samples (KHPS2007 and KHPS2012)
	JHPS 2009	Individuals initially selected for the JHPS2009 sample
	JHPS/KHPS 2009	Individuals in the combined JHPS/KHPS sample, including individuals initially selected for the KHPS2004 sample, those selected for the KHPS2007 top-up sample, and those selected for the JHPS2009 sample
	JHPS/KHPS 2012	Individuals in the combined JHPS/KHPS sample, including individuals initially selected for the KHPS2004 sample, those selected for the two KHPS top-up samples (KHPS2007 and KHPS2012), and those selected for the JHPS2009 sample
Individual and household cross-sectional weights	KHPS	Individuals initially selected for the KHPS2004 sample and those selected for the two KHPS top-up samples (KHPS2007 and KHPS2012)
	JHPS	Individuals initially selected for the JHPS2009 sample
	JHPS/KHPS	Individuals initially selected for the KHPS2004 sample, those selected for the two KHPS top-up samples (KHPS2007 and KHPS2012), and those selected for the JHPS2009 sample

Source: Authors, based on the JHPS/KHPS data structure.

The aforementioned strategy basically follows the general steps mentioned in Watson (2012). In the next subsections, the procedure will be described in detail.

3.1 Computation of individual base weights for the first wave of each sample

The procedure described here applies not only to the KHPS2004 sample but also to the JHPS2009 sample and to the top-up samples (KHPS2007 and KHPS2012). The procedure for computing individual base weights for first-wave samples is basically a stepwise approach whereby weighting coefficients are calculated and adjusted at each step to take into account the main features of the sample selection procedure.

3.1.1 Step 1: Computation of individual design weights

By definition, design weights are defined for each individual in the initial year of each sample and are calculated as the inverse of the selection probability of each individual. The selection probability of each individual depends on the sampling design of the survey. For the JHPS/KHPS case, the survey subjects were selected by a two-stage stratified random sampling method, where Japan as a whole was stratified into 24 strata following a regional and municipal classification. The number of subjects drawn in each stratum was allocated in proportion to the size of the stratum in the registered population. Districts inside each stratum were selected following a systematic random sampling process until the predefined total number of subjects per stratum was achieved. These subjects were also randomly sampled.

Grounded in the aforementioned information, the design weights are calculated as the inverse of the selection probabilities of individual *i* in district *d* and stratum *s* as follows:

$$P_{ids} = Pr(d\ selected) \times Pr(i\ selected\ in\ d) = \frac{a_s}{A_s} \times \frac{b_{ds}}{B_{ds}}$$

where a_s is the total number of districts selected in stratum *s*; A_s is the total number of districts in stratum *s*; b_{ds} is the total number of individuals selected in district *d* localized in stratum *s*; and B_{ds} is the total number of individuals in district *d* localized in stratum *s*.

As the sample size of individuals in each stratum has been allocated to be proportional to the total number of individuals in the stratum, the probability of selection is the same for each individual.[5] For this reason, the design weight (dw_i) is constant for individual *i* in each initial sample:

[5] Given that the survey was employed using the drop-off/pick-up system (DOPU), the assumption of equal probability sampling is not entirely true. Nevertheless, given that the differences are marginal, we assume that the equal probability assumption is still valid.

$$dw_i = \frac{1}{P_{ids}} = \frac{N}{n}$$

where N is the size of the overall population, and n is the size of the total sample.

3.1.2 Step 2: Correction for unit nonresponse

Unit nonresponse refers to the failure to collect any survey information for a fraction of the individuals in the initial sample. There are several reasons for such a failure; for example, the dwelling may be difficult to locate or access or members of the household may be absent, refuse to respond to the questionnaire or suffer from health or more general incapacity problems. Unit nonresponse is detrimental to sample representativity when nonrespondents and respondents have different profiles with respect to the main characteristics of the survey.

To properly account for the bias caused by unit nonresponse, the traditional approach consists of estimating the probability of response θ_i of the respondents through logistic modeling and then adjusting the design weights accordingly. Let X_i designate a vector of response predictors, which must be available both on the respondents and the nonrespondents.[6] The estimated probability of response of individual i is given by the following logistic relationship:

$$\hat{\theta}_i = \frac{e^{\hat{A}X_i}}{1 + e^{\hat{A}X_i}}$$

where the vector $\hat{A}$ of the model coefficients is estimated from the sample units by maximum likelihood. Estimating $\hat{\theta}_i$, we obtain the individual design weights adjusted for unit nonresponse as follows:

$$\widetilde{dw}_i = \frac{dw_i}{\hat{\theta}_i}$$

If the response modeling is correct, the modified weights ($\tilde{d}w_i$) lead to unbiased estimators.

Unfortunately, for the JHPS/KHPS case, this traditional approach was not employed. The first reason for not employing this method was the lack of enough information received from the nonrespondents. For the first wave of each sample, it is only possible to obtain some minor information about the nonrespondents, such as dwelling type and the number of times

[6] There are several possible sources of auxiliary information for nonresponse correction, for instance the sampling frame itself, population censuses, administrative registers, *ad hoc* questions collected *de visu* by interviewers (e.g., dwelling type or neighborhood appearance) or information obtained from the data collection process (paradata) such as the mode of contact, the number of contact attempts or general information on the profile of the interviewer (Osier, 2016). Overall, in order to be powerful in reducing nonresponse bias, the X_i needs to be correlated with both the response propensity and the target variables of the survey.

the investigator visited the dwelling. Consequently, the necessary information for modeling response propensities is unavailable.

The second reason for not applying the traditional approach is the use of "reserve subjects" in the first wave of each sample of the JHPS/KHPS. To avoid the possible bias caused by nonresponsiveness in the first wave, for each original subject, between 3 and 5 reserve subjects were randomly selected (from the same district, sex and age category of their original subject).

3.1.3 Step 3: Calibration to external benchmarks

Although the aforementioned replacement by reserve subjects is expected to reduce nonresponse bias to a great extent, there is still a minor possibility of bias. To address this issue, the calibration approach can be used as an alternative to the traditional approach. Calibration (Deville, 1992) is a long-established technique that aims to incorporate external information into an estimator, thereby improving data accuracy. In other words, the objective of this technique is to calibrate the weights according to external population benchmarks, thereby reducing response bias and increasing the sample precision. In general, these external population benchmarks are constructed based on disaggregated population-level information obtained from auxiliary sources such as a census.

This technique also addresses the nonresponse issue (Lundström and Särndal, 1999; Särndal and Lundström, 2005) when the calibration variables are also significant response predictors; it has the advantage of not requiring the availability of data for both respondents and nonrespondents.[7] Given that in the case of the JHPS/KHPS, it is not possible to estimate the response propensity to adjust for unit nonresponse, a one-step calibration method was adopted to reduce nonresponse bias and increase data precision. The external population benchmarks used in the calibration process were obtained from the Labor Force Survey, Statistics Bureau of Japan.[8]

The weights obtained after the calibration step are the base weights (bw_i) for the individuals in the sample. From the second wave of each sample onward, these base weights are corrected yearly to address the panel attrition issue.

[7] Traditional nonresponse correction methods require data from both respondents and nonrespondents; thus, it is impracticable for cases where no data is available for nonrespondents.
[8] A description of the data used in the calculation of yearly population benchmarks is available in Appendix A.

3.2 Computation of individual base weights from the second wave onward

According to the JHPS/KHPS tracing rules, individuals selected and interviewed during the first wave become panel persons and are to be followed-up with in subsequent waves regardless of the place where they live. Thus, an individual who has moved to another dwelling must be recontacted at his or her new address. The follow-up of sample individuals leads to the collection of longitudinal data that in turn allow the analysis of socioeconomic trends over time.

From the second wave of each sample onward, the individual base weights calculated during the first wave have to be modified to compensate for panel attrition. In a panel setting, attrition refers to the loss of sample persons due to nonresponse. The main reasons for such loss are the refusal by a sample person to continue to participate in the study (survey fatigue), the inability to contact the person in cases of long-term absence, or being unable to locate a person when he or she has moved to another location. Attrition is different from being removed from the survey scope, for instance, when a person physically dies, leaves the country or joins a collective household or an institution.

Given the aforementioned issue, from the second wave of each sample onward, the base weights are updated based on two procedures. First, a logistic model is used to correct the problem of panel attrition in the JHPS/KHPS. Next, these adjusted weights are once again subjected to a calibration process based on external population benchmarks, aiming to increase the sample precision.

3.2.1 The attrition issue and the logistic model

Similar to unit nonresponse, attrition is a potential source of bias in estimates; thus, it needs to be dealt with to keep panel samples representative. The probability for an individual who has responded at wave *t* still to be responding at wave *t*+1 must be estimated through logistic modeling and then used to adjust the weights. Considering that $\tilde{b}_i^{(t)}$ is the base weight of *i* at *t*, in the first wave, $\tilde{b}_i^{(t)} = bw_i$ is the individual design weight adjusted for unit nonresponse and calibrated to external data sources. From the second wave onward, let $p_i^{t,t+1} = Pr(i \in \tilde{r}_{t+1} | i \in r_t)$ be the estimated probability for an individual *i* to respond at *t*+1 ($\tilde{r}_{t+1}$) given that it responded at *t* (r_t). Assuming $p_i^{t,t+1} > 0$ for all *i*, the base weight at *t*+1 is derived from the base weight at *t* as follows:

$$\tilde{b}_i^{(t+1)} = \frac{\tilde{b}_i^{(t)}}{p_i^{t,t+1}}$$

Contrary to the first wave, a great deal of auxiliary information collected at *t* is now available for correcting attrition at *t*+1. This information must be utilized as much as possible to build powerful response explanatory models. Selection methods exist in the literature (Schork, 2018) to detect the most powerful predictors among a list of possible candidates.

To estimate $p_i^{t,t+1}$, a traditional approach based on a logistic model is adopted. If a given individual *i* participated in survey wave *t*, then it is possible to use the information collected from this individual in wave *t* to estimate the probability that this individual *i* will participate again in wave *t+1*. First, a logistic regression model is employed to estimate the response propensity[9] as follows:

$$logit\,\{\Pr(resp_{it+1} = 1|X_{it})\} = \beta_0 + \beta_1 X_{it1} + \cdots + \beta_p X_{itp}$$

where $resp_{it+1}$ is the probability that individual *i* will participate in survey wave *t+1*, while X_{it} represents the auxiliary variables that are predictors of $resp_{it+1}$. The logistic regression model predictors include marital status, sex, age, income group, household size, etc.[10]

Next, the estimated response propensities $\widehat{Pr}(resp_{it+1} = 1|X_{it})$ obtained from the model are used to adjust the weights of each individual:

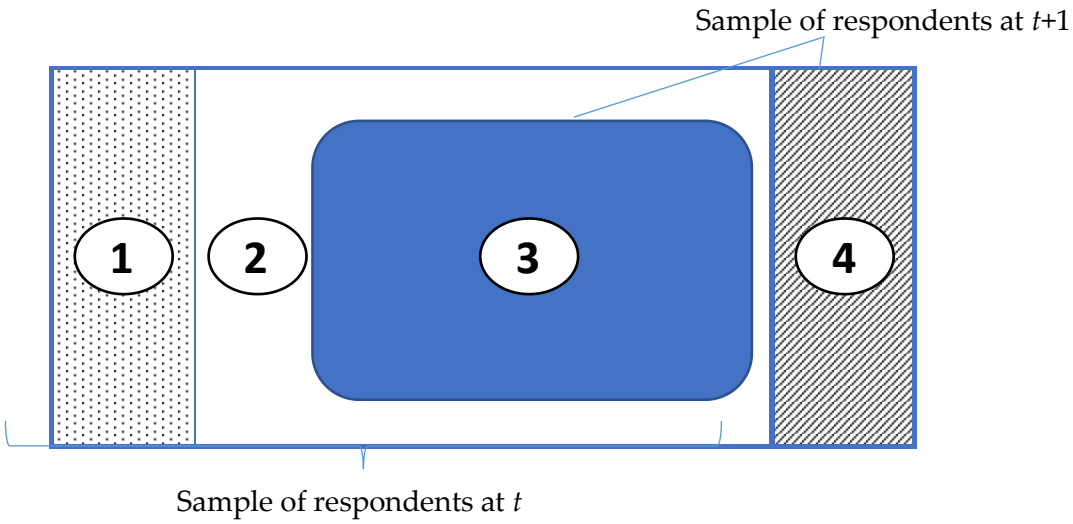
$$w_{it+1} = wb_{it} \times \frac{1}{\widehat{Pr}(resp_{it+1} = 1|X_{it})}$$

3.2.2 Re-entry, respondent substitution and calibration to external benchmarks

According to the JHPS/KHPS tracing rules, a respondent who does not participate in the survey at point *t* can return to the panel at *t+1* if he or she wishes to do so. These cases are known as re-entry, and they also demand specific treatment. The concept of re-entry is illustrated in Figure 2. First, consider that at time *t-1*, the respondents are identified by 1, 2, 3, and 4, as shown in Figure 2. Next, the respondents at time *t* are identified by 1, 2, and 3 (respondent 4 does not participate at *t*). At *t+1*, the respondent identified as 2 does not participate in the survey due to attrition, while the respondent identified by 1 does not participate because he or she is out of the survey scope. However, respondents identified as 3 and 4 do choose to participate in the survey. In this case, respondent 4 represents the re-entry cases, that is, cases that participated in *t-1* and *t+1* but did not participate in *t*.

[9] Individuals who turned "out-of-scope" (e.g., because of physical deaths) between t and t+1 must be removed from the calculations.
[10] A description of all predictors is available in Appendix B.

Figure 2. Sample dynamics between *t* and *t*+1



Source: Authors.

Given that the number of re-entry cases in the JHPS/KHPS is relatively small, individual re-entry is disallowed after the person exits the survey for the first time. In other words, re-entry cases receive a zero value weight after the respondent exits the survey for the first time.

Another specificity of the JHPS/KHPS tracing rule is the existence of cases where the spouse of the respondent serves as a substitute for the respondent after the respondent's death and continues participating in the survey. Given that the subjects of the JHPS/KHPS are the initially selected individuals, these cases are identified and automatically given a zero value weight for all waves occurring after the original respondent died.

The resultant weights are then calibrated according to external population benchmarks to increase the sample precision. After this process, the individual base weight at wave *t+1* is obtained.

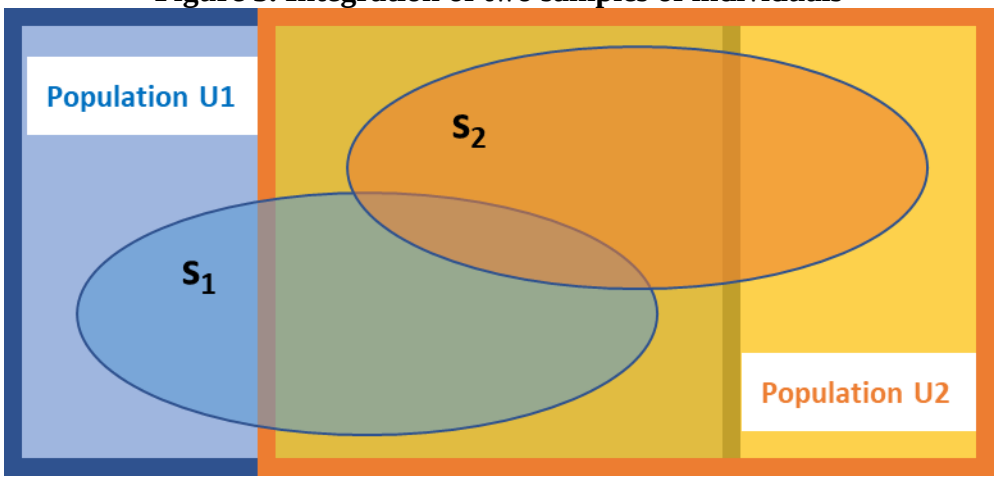
4. Integration of top-up samples

In the previous section, the strategy adopted in the computation of base weights for different samples and their update in the subsequent waves was explained. This strategy allows the computation of a group of sample weights for separate samples, e.g., the KHPS2004 sample, KHPS2007 top-up sample, KHPS2012 top-up sample, and JHPS2009 sample. This section explains how the weights from different samples can be combined to allow the possibility of having larger data that provide more stable results.

4.1 General formula

The problem of integrating two samples of individuals drawn from different populations at different points in time can be regarded as follows. Let s_1 and s_2 designate two samples of individuals who are representative of the populations U_1 and U_2, respectively. Assuming there is some overlap between those two groups, Figure 3 shows the image of the integration of two samples of individuals.

Figure 3. Integration of two samples of individuals



Source: Authors.

Each individual i in s_1 receives a sampling weight $w_i^{(1)}$, and each individual i in s_2 receives a sampling weight $w_i^{(2)}$. The weights for the integrated sample $\tilde{s} = s_1 \cup s_2 - s_1 \cap s_2$ are given by the following:

$$w_i^{INT} = \begin{cases} w_i^{(1)} & if\ i \in [s_1 \cap (U_1 - U_{inter})] \\ w_i^{(2)} & if\ i \in [s_2 \cap (U_2 - U_{inter})] \\ \theta w_i^{(1)} + (1-\theta) w_i^{(2)} & if\ i \in (s_1 \cap s_2 \cap U_{inter}) \\ \theta w_i^{(1)} & if\ i \in (s_1 \cap \bar{s}_2 \cap U_{inter}) \\ (1-\theta) w_i^{(2)} & if\ i \in (\bar{s}_1 \cap s_2 \cap U_{inter}) \end{cases}$$

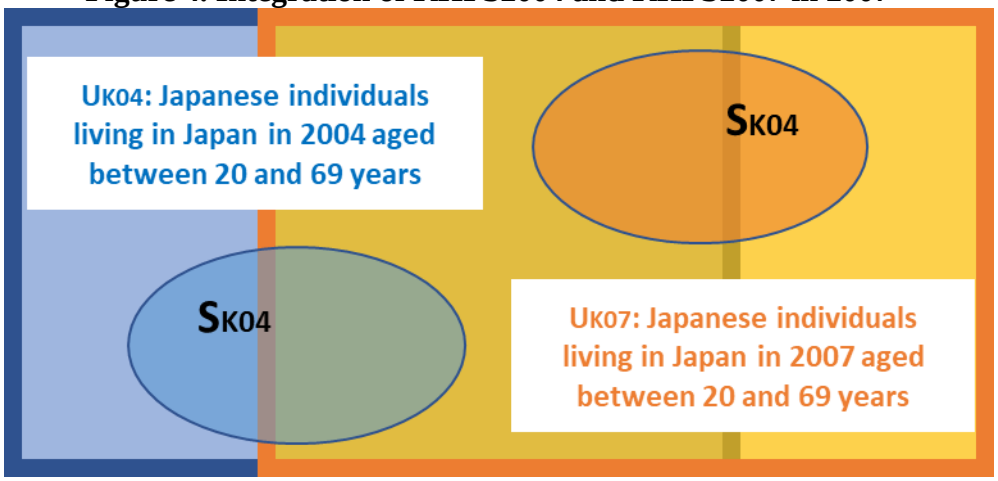
where $U_{inter} = U_1 \cap U_2$. The sharing parameter θ lies between 0 and 1. Generally, $\theta = 1/2$. An alternative option is to set $\theta = n_1/(n_1 + n_2)$, where n_1 is the size of s_1, and n_2 is the size of s_2.

4.2 Application to the KHPS2007 top-up sample

Regarding the computation of the weights for the combination of the KHPS samples from 2004 (S_{K04}) and 2007 (S_{K07}), it is first necessary to consider that each sample will represent the population from the period in which it was drawn (U_{K04} and U_{K07}). Based on set

theory, the calculation of the weights for the integrated sample will consider the union of the U_{K04} and U_{K07} populations. In this case, there will be an overlap in the population of Japanese individuals aged between 23 and 69 years who were living in Japan in 2007 and those who were also living in Japan in 2004.

Figure 4. Integration of KHPS2004 and KHPS2007 in 2007[11]



Source: Authors, based on the JHPS/KHPS data structure.

If each individual *i* in the KHPS2004 received a weight $w_{it}^{(K04)}$, and each individual *i* in the KHPS2007 received a weight $w_{it}^{(K07)}$, then the integrated weight w_{it}^{INT} can be calculated according to a general formula as follows:

$$w_{it}^{INT} = \begin{cases} w_{it}^{(K04)} & if \;\; i \in [S_{K04} \cap (U_{K04} - U_{inter})] \\ w_{it}^{(K07)} & if \;\; i \in [S_{K07} \cap (U_{K07} - U_{inter})] \\ \theta w_{it}^{(K04)} & if \;\; i \in (S_{K04} \cap \bar{S}_{K07}) \\ (1-\theta) w_{it}^{(K07)} & if \;\; i \in (\bar{S}_{K04} \cap S_{K07}) \end{cases}$$

where U_{inter} is the intersection of the population in KHPS2004 and KHPS2007 (U_{K04} and U_{K07}), and the parameter θ lies between 0 and 1.[12] To illustrate this process, consider the weights for 2007: respondents aged between 20 and 22 years maintain $w_{it}^{(K07)}$, respondents aged between 70 and 72 years maintain $w_{it}^{(K04)}$, respondents aged between 23 and 69 years from KHPS2004 have their weights adjusted to $\theta w_{it}^{(K04)}$, and respondents aged between 23 and 69 years from KHPS2007 have their weights adjusted to $(1-\theta) w_{it}^{(K07)}$.

[11] The sampling procedure does not allow repetition. Consequently, there is an intersection of the populations U_{K04} and U_{K07}, but no intersection in samples S_{K04} and S_{K07} is possible.
[12] The parameter θ can be calculated as the share of S_{K04} in the combined samples or it can simply be assumed to be 1/2. For simplicity, the second option was adopted.

This approach can be extended from 2008 onward to combine the base weights for the original KHPS2004 sample in year N ($N \geq 2008$) and those from the KHPS2007 top-up sample in year *N*, both of which are adjusted for attrition from 2008 onward. For the case of integrating the samples, the calibration process mentioned in the previous section is performed after the computation of the integrated weights.

4.3 Application to the KHPS2012 top-up sample

In 2012, another KHPS top-up sample was drawn from the resident population aged between 20 and 69 years. To identify the overlapping part U_{inter} of the reference population for the KHPS2004 sample with that of the KHPS2007 top-up sample, it is necessary to combine information about the age of the individuals in 2004 and 2007. Basically, an individual who is in the KHPS2012 top-up sample belongs to the overlapping set if at least one of these two conditions is met: the person was aged between 20 and 69 years in 2004, or the person was aged between 20 and 69 years in 2007.

4.4 Application to the JHPS2009

In 2009, the JHPS was launched based on a representative sample of the resident population aged 20 or older. Again, to combine the JHPS2009 sample with the KHPS sample (combining the KHPS2004 sample and the KHPS2007 top-up sample), we need to identify the individuals in the overlapping part. The rule is the same as that stated before. Basically, a sample individual who is in the JHPS2009 sample belongs to the overlapping sample if at least one of these two conditions is met: the person was aged between 20 and 69 years in 2004, or the person was aged between 20 and 69 years in 2007.

Finally, from 2012 onward, the JHPS2009 sample can also be merged with the combined KHPS sample (the KHPS2004 sample and the two top-up samples, KHPS2007 and KHPS2012). The intersection is identified as individuals who are in the JHPS2009 sample and meet at least one of these three conditions: the person was aged between 20 and 69 years in 2004, the person was aged between 20 and 69 years in 2007, or the person was aged between 20 and 69 years in 2012.

5. Computation of longitudinal weights

The combined base weights whose calculation was described in the previous section can be used for longitudinal estimation. As previously indicated in Table 6, six different groups of longitudinal weights were calculated. For the KHPS, longitudinal weights were calculated from 2004 onward (original KHPS2004 sample), from 2007 onward (original KHPS2004

sample + KHPS2007 top-up sample) and from 2012 onward (original KHPS2004 sample + KHPS2007 top-up sample + KHPS2012 top-up sample). In addition, longitudinal weights were calculated for the JHPS from 2009 onward (original JHPS2009 sample), for the combination of the JHPS/KHPS from 2009 onward (original KHPS2004 sample + KHPS2007 top-up sample + original JHPS2009 sample) and from 2012 onward (original KHPS2004 sample + KHPS2007 top-up sample + KHPS2012 top-up sample + original JHPS2009 sample).

These weights should be calibrated to reflect external population benchmarks for the population of reference. However, population benchmarks need to be calculated properly, taking into account the changes in population over time. The reference populations for the different JHPS/KHPS samples are listed in Table 5.

Table 5. JHPS/KHPS reference populations

Sample	Reference population[13]
Original KHPS sample (2004)	Resident population aged 20–69
Original KHPS sample (2005 onward)	Resident population aged 20+N – 69+N (N = CURRENT YEAR – 2004)
Original KHPS sample + KHPS2007 top-up sample	Resident population aged 20–72
Original KHPS sample + KHPS2007 top-up sample (2008 onward)	Resident population aged 20+M – 69+N (N = CURRENT YEAR – 2004) (M = CURRENT YEAR – 2007)
Original KHPS sample + KHPS2007 top-up sample + KHPS2012 top-up sample	Resident population aged 20–77 (N = CURRENT YEAR – 2004)
Original KHPS sample + KHPS2007 top-up sample + KHPS2012 top-up sample (2013 onward)	Resident population aged 20+M – 69+N (N = CURRENT YEAR – 2004) (M = CURRENT YEAR – 2012)
Original JHPS sample (2009)	Resident population aged 20 or more
Original JHPS sample (2010 onward)	Resident population aged 20+M or more (M = CURRENT YEAR – 2009)
Original KHPS sample + KHPS2007 top-up sample + original JHPS (2010 onward)	Resident population aged 20+M or more (M = CURRENT YEAR – 2009)
Original KHPS sample + KHPS2007 top-up sample + KHPS2012 top-up sample + original JHPS sample (2013 onward)	Resident population aged 20+M or more (M = CURRENT YEAR – 2012)

Source: Authors, based on the JHPS/KHPS data structure.

6. Computation of individual and household cross-sectional weights

In addition to the longitudinal weights, two other types of weights are calculated: individual cross-sectional weights and household weights.

[13] The resident population consists of the Japanese citizens living in Japan; it does not include foreigners living in Japan.

6.1 Individual cross-sectional weights

Three types of individual cross-sectional weights are calculated: weights for the KHPS that include the individuals selected for the KHPS2004 sample and the top-up samples collected in 2007 and 2012; weights for the JHPS that comprise the individuals from JHPS2009; and weights for the integrated JHPS/KHPS sample, including all the aforementioned samples.

Different from the longitudinal weights that apply to panel data, the individual cross-sectional weights allow inference for a specific year. For this reason, the cross-sectional weights are available for all years. They are available from 2004 to 2017 for the case of KHPS and JHPS/KHPS cross-sectional weights, while they are available from 2009 to 2017 for the JHPS cross-sectional weights.

As explained before, given that re-entry cases are disallowed, the organization of the individual cross-sectional weights is based on the calculated longitudinal weights.

6.2 Household cross-sectional weights

Based on the aforementioned three types of individual cross-sectional weights, three types of household cross-sectional weights are calculated. As the JHPS/KHPS subjects are individuals and not households, the household cross-sectional weights had to be calculated using the weight share method, as presented in Lavallée (2007). This method states that the household cross-sectional weight (W_{hit}) can be calculated according to the following formula:

$$W_{hit} = \frac{\sum_{i \epsilon r} w_{it} 1_{i \epsilon h}}{N_{hit}}$$

where the cross-sectional weight of individual *i*'s household *h* in year *t* (W_{hit}) is the result of the sum of the cross-sectional weight of all members from this household in year *t* (w_{it}) divided by the total number of members in household *h* in year *t* (N_{ht}).[14] In other words, the weight share method determines that the household weights are the average of the weights of members inside the survey scope. If the individual weights w_{it} are unbiased, then the household weights W_{hit} are unbiased as well.

[14] The first step in calculating the household cross-sectional weights is to obtain from the survey the data regarding the number of household members and their age in each wave. After that, it is possible to calculate the number of members in the household that are inside the scope of the survey.

7. Selection and use of the most appropriate weight

Given the complexity of the JHPS/KHPS structure, multiple types of weights are provided to meet the different needs of the users. The most appropriate weight for a given analysis must reflect the survey instrument, which is the source of the data being used in the analysis (KHPS, JHPS, or JHPS/KHPS), the combination of waves involved in the type of analysis and the indicators we seek to produce, and the level of analysis (household or individual).

The weight names and labels are given to help users choose the correct weight. The name and label of each weight reflect the combination of samples for which the weight is calculated, the level of analysis and its nature (cross-sectional analysis weight or longitudinal analysis weight).

Table 6. Available weights and corresponding definitions

Analysis level	Starts from year	Data source	Analysis Weight
Individual	2004	KHPS2004	Long_Weight_1
Individual	2009	JHPS2009	Long_Weight_2
Individual	2007	KHPS2004 and KHPS2007	Long_Weight_3
Individual	2012	KHPS2004, KHPS2007, and KHPS2012	Long_Weight_4
Individual	2009	KHPS2004, KHPS2007, and JHPS2009	Long_Weight_5
Individual	2012	KHPS2004, KHPS2007, JHPS2009, and KHPS2012	Long_Weight_6
Individual	2004	KHPS2004, KHPS2007, and KHPS2012	Cross_Weight_1
Individual	2009	JHPS2009	Cross_Weight_2
Individual	2004	KHPS2004, KHPS2007, JHPS2009, and KHPS2012	Cross_Weight_3
Household	2004	KHPS2004, KHPS2007, and KHPS2012	HH_1
Household	2009	JHPS2009	HH_2
Household	2004	KHPS2004, KHPS2007, JHPS2009, and KHPS2012	HH_3

Source: Authors, based on the JHPS/KHPS data structure.

If the analysis uses only data from a specific wave, the cross-sectional version of the weight would be the most appropriate option for this study. The cross-sectional weight is calculated and attributed to all sample members who answered the survey questionnaire in a specific wave, except for re-entry cases.[15] If the study uses data from multiple waves, then the appropriate version of the longitudinal weights should be selected.

[15] Another specificity of the JHPS/KHPS is the existence of cases where the spouse of the respondent served as a substitute for the respondent after his or her death and continued participating on the survey. Given that the subjects of the JHPS/KHPS are the initially selected individuals, these cases are identified and automatically received a zero value weight for all waves occurring after the original respondent died.

For example, the longitudinal weights for the integration of KHPS2004, KHPS2007, and KHPS2012 are available from 2012 to 2017. If a study requires the examination of panel data that starts in 2010, then the use of combined longitudinal weights for KHPS2004, KHPS2007, and KHPS2012 is not possible. In this case, the most appropriate would be the use of longitudinal weights for the combination of the KHPS2004 and KHPS2007 samples, since the weights would be available from 2007 to 2017.

7.1 Assumptions when not using weights[16]

As stated before, the use of weights aims to adjust possible distortions derived from sample selection bias, nonresponse, and attrition. These weights allow us to build population-level estimators in a statistically proper way that keeps bias and variance as small as possible. In general, an unweighted analysis does not correctly reflect the population structure unless some assumptions are true. In other words, the use of weights will not be necessary if we assume that population estimated parameters (means, measure of dispersion, etc.) do not differ between the following:

- Individuals from different Japanese regions;
- Individuals who participated in the first wave of each sample and those who did not;
- Individuals who continued to participate at later waves and those who did not; and
- Individuals from different samples.

For this reason, the theory suggests that researchers who publish or present unweighted estimates make these assumptions explicit. Otherwise, the use of weights to correct the possible distortions in estimations of population parameters is recommended. Furthermore, weights are generally not recommended in the case of model-based analyses, such as linear or logistic regression modeling, as they make estimates more volatile. That is why practitioners often prefer not to use them and keep data analysis unweighted, although there might be some concern about bias. However, there is no clear answer to this question, which actually depends on the data available. Thus, an empirical approach comparing weighted and unweighted results remains a sensible solution.

[16] This section was based on: Understanding Society (2019). The UK Household Longitudinal Study: waves 1-9 User Guide. University of Essex, Colchester, Essex.

7.2 Commands for using the weights in Stata[17]

The Japan Household Panel Survey (JHPS/KHPS) weights are available in the file named "Weights.dta". After merging this file with the JHPS/KHPS data, the "*svy*" command can be used to obtain estimates that correctly take into account the sample design.

First, it is necessary to specify the design of the survey using the svyset command. Suppose a study investigates the population features in 2013 using JHPS data. In this case, the appropriate weights are available in Column Cross_Weight_2 (JHPS Cross-sectional weights). The following command can be employed to specify the use of these weights:

svyset id [pweight = Cross_Weight_2]

Next, any compatible command needs to be prefixed with "*svy*". For example, a logistic regression for the data from 2013 using the weights can be performed in the following way:

svy: logistic depvar variable1 variable2 variable3 if year==2013

Suppose a study in which a longitudinal analysis using the JHPS/KHPS data for the period 2013-2017 needs to be performed. In this case, the appropriate weight would be found in Column Long_Weight_6 (JHPS/KHPS longitudinal weights). For the case of longitudinal analysis, it is necessary to copy the longitudinal weights of the last wave used in the analysis (in our example, the weights from 2017) and apply them to each respondent's previous waves (from 2013 to 2016 in our example).

In practice, after merging the weights file with the JHPS/KHPS data and keeping only the data for the period 2013-2017, the following procedure can be adopted:

1- Generate a new weighting variable for the longitudinal analysis of the period 2013-2017:

gen weight_lg17= Long_Weight_6

2- Copy the last wave weight value over the previous year:

sort id year
bysort id (year): replace weight_lg17=weight_lg17[_N]

[17] This section was based on: Understanding Society (2017). The UK Household Longitudinal Study: wave 1-7, 2009-2016 User Guide. Ed. Knies, Gundi. University of Essex, Colchester, Essex.

3- Specify the use of these weights:

svyset id [pweight = weight_lg17]

4- Perform the desired analysis:

svy: reg depvar variable1 variable2 variable3

Some commands in Stata are not compatible with the prefix "*svy*". For these cases, an alternative would be stating the use of the weight as follows:[18]

bysort year: table var1 [pw=weight_lg17]
xtset id year
xtreg depvar year i.var2 [pw=weight_lg17]

Table 7 provides more examples of weight choice based on the type of analysis employed by a given study.

Table 7. Examples of weight choice based on the analysis type

Data source	Longitudinal analysis of individual respondents	Cross-sectional analysis of individual respondents
KHPS2004	Long_Weight_1 weight from latest wave in the longitudinal analysis	
KHPS2004 + KHPS2007	Long_Weight_3 weight from latest wave in the longitudinal analysis	
KHPS2004 + KHPS2007 + KHPS2012	Long_Weight_4 weight from latest wave in the longitudinal analysis	Cross_Weight_1 weight from the analyzed wave
JHPS2009	Long_Weight_2 weight from latest wave in the longitudinal analysis	Cross_Weight_2 weight from the analyzed wave
KHPS2004 + KHPS2007 + JHPS2009	Long_Weight_5 weight from latest wave in the longitudinal analysis	
KHPS2004 + KHPS2007 + JHPS2009 + KHPS2012	Long_Weight_6 weight from latest wave in the longitudinal analysis	Cross_Weight_3 weight from the analyzed wave

Source: Authors, based on the JHPS/KHPS data structure.

8. Sampling weights and the representativeness of the data

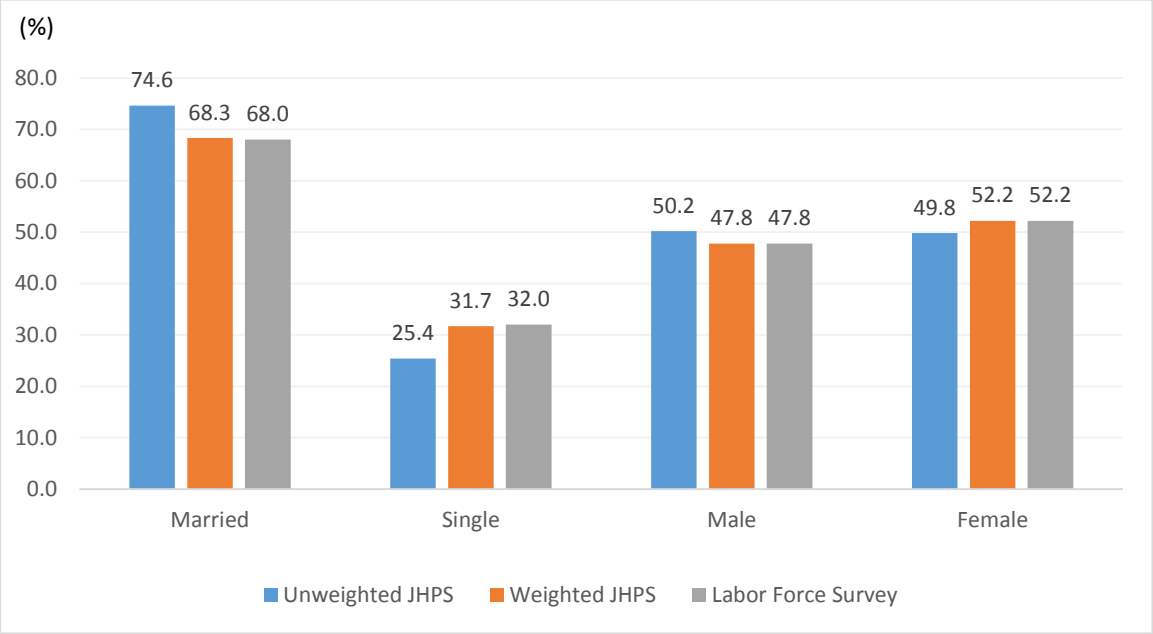
To illustrate and confirm the representativeness of the survey data after the adoption of the calculated sampling weights, this section presents a series of exercises in which weighted and unweighted values from the JHPS data are compared with the official statistics. At the first moment, individual-level characteristics are tested, while at the second moment, household-level information is observed. For the case of personal characteristics, the JHPS cross-sectional

[18] For more information on the commands and use of weights in Stata, please refer to the Stata user guide.

weight (Cross_Weight_2) is employed, while for household characteristics, the JHPS household weight (HH_2) is employed.

In the first exercise, the sex and marital status of the unweighted and weighted JHPS data in 2017 are compared with the data from the Labor Force Survey from the Statistics Bureau of Japan. Figure 5 illustrates the shares of the Japanese population according to sex and marital status. Interestingly, for the case of unweighted JHPS data, it is possible to observe a distortion in the data. However, after weighting the data, we verify that the shares are similar to those obtained from the Labor Force Survey.

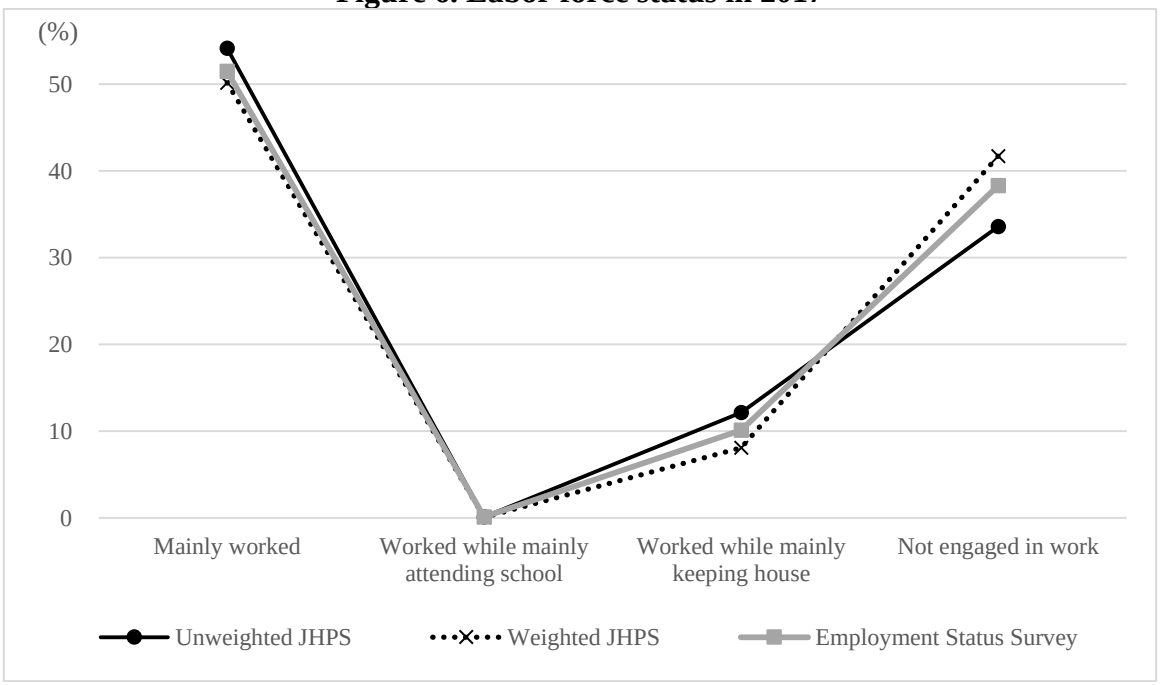
Figure 5. Share of sex and marital status in 2017



Source: Authors based on the JHPS data and the 2017 Employment Status Survey.

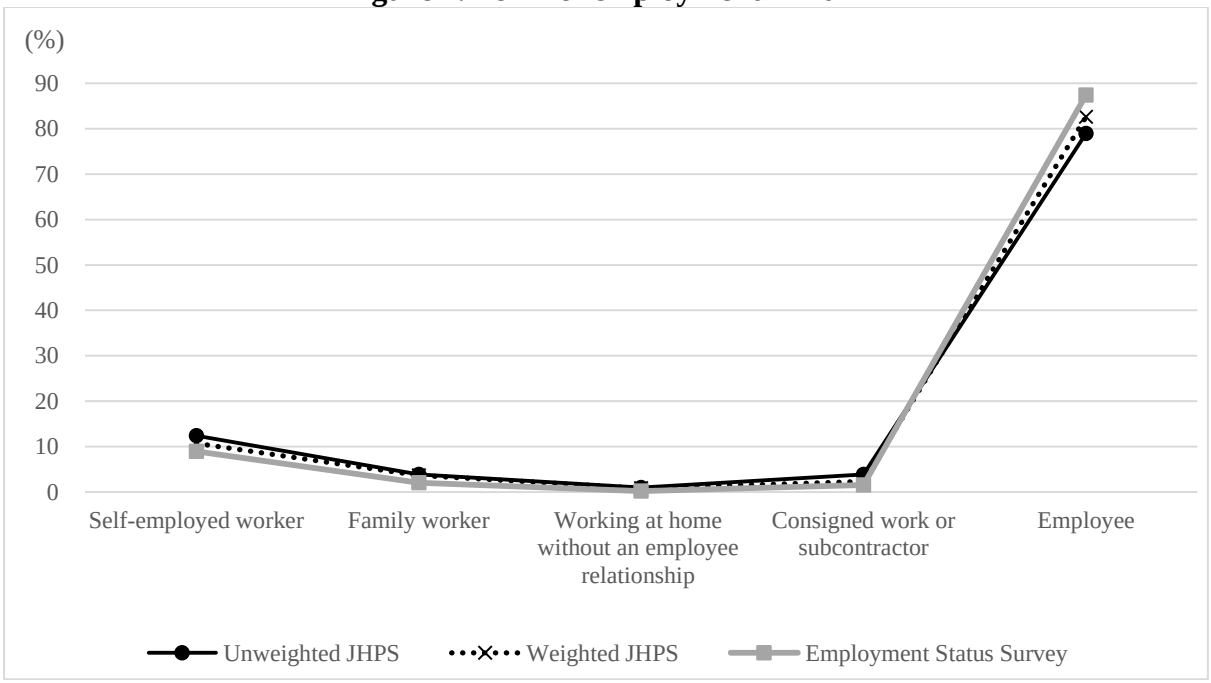
Next, a group of characteristics related to the job condition of the Japanese population is observed. Data on labor force status (Table 6), form of employment (Table 7), form of employment of employees only (Table 8), and firm size (Table 9) obtained from the 2017 Employment Status Survey are compared with weighted and unweighted JHPS data. In this case, the shares of the unweighted JHPS data are already very close to those found for the statistical data. Consequently, after the adoption of the weights, there is a slight change in which the shares with the weighted values get closer to the official statistics, thereby confirming that the use of sampling weights improves the representativeness of the JHPS data.

Figure 6. Labor force status in 2017



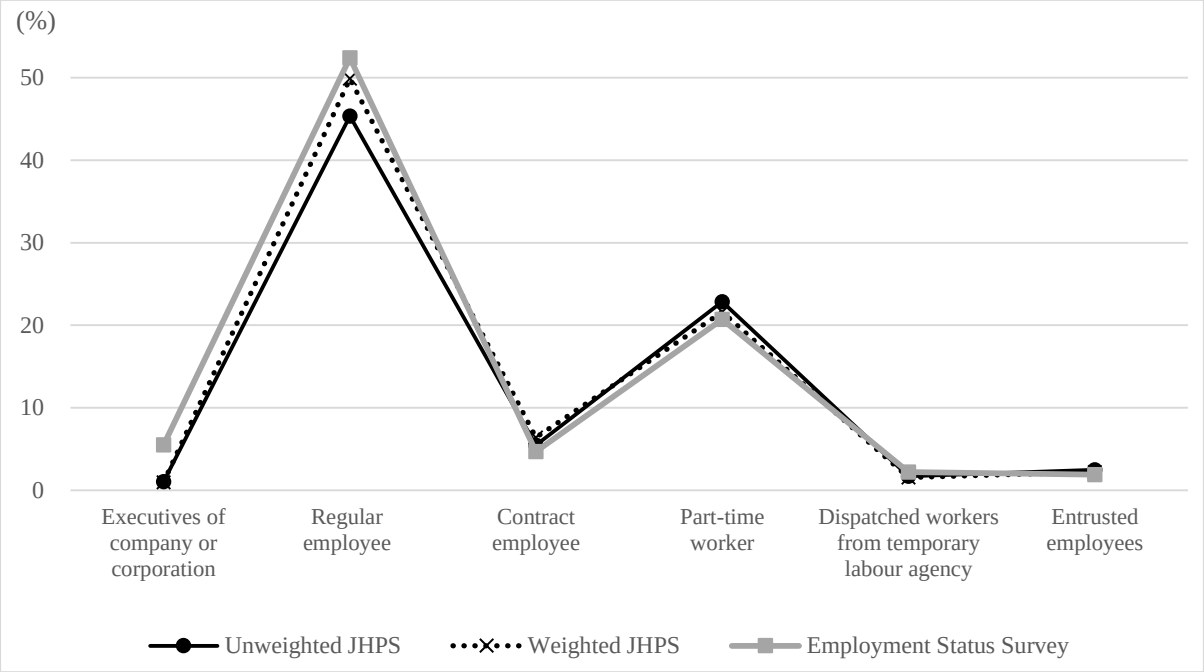
Source: Authors based on the JHPS data and the 2017 Employment Status Survey.

Figure 7. Form of employment in 2017



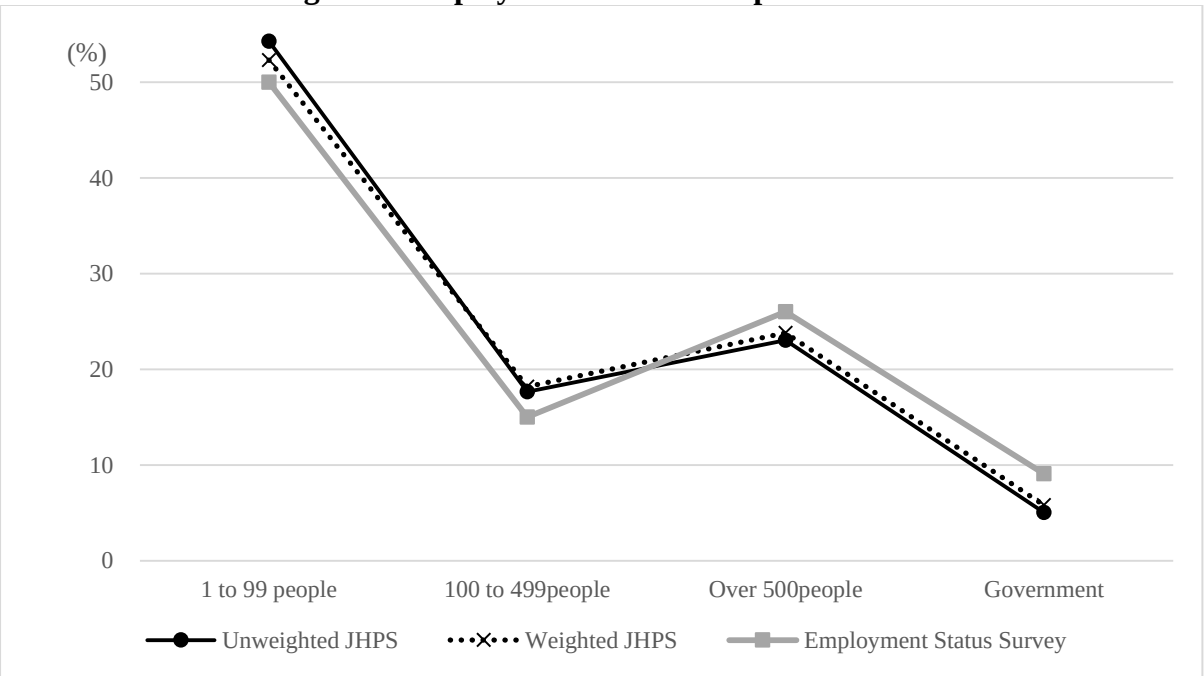
Source: Authors based on the JHPS data and the 2017 Employment Status Survey.

Figure 8. Form of employment in 2017 (only employees)



Source: Authors based on the JHPS data and the 2017 Employment Status Survey.

Figure 9. Employment size of enterprise in 2017

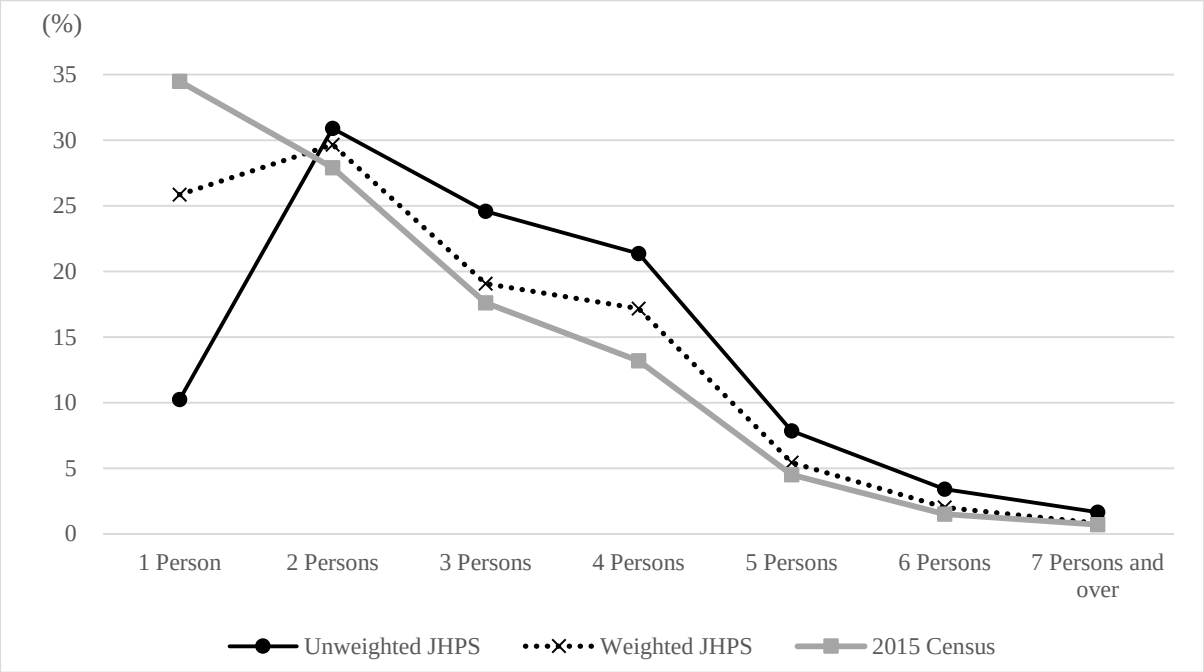


Source: Authors based on the JHPS data and the 2017 Employment Status Survey.

Regarding the household-level characteristics, data on household size (Figure 10), type of dwelling (Figure 11), and household income (Figure 12) are observed. Some of the variables in the JHPS, such as the household size and the type of dwelling, are originally biased. This bias is attributed to the fact that the JHPS is conducted using the drop-off and pick-up (DOPU) method, which tends to obtain more answers from detached-house respondents and fewer from

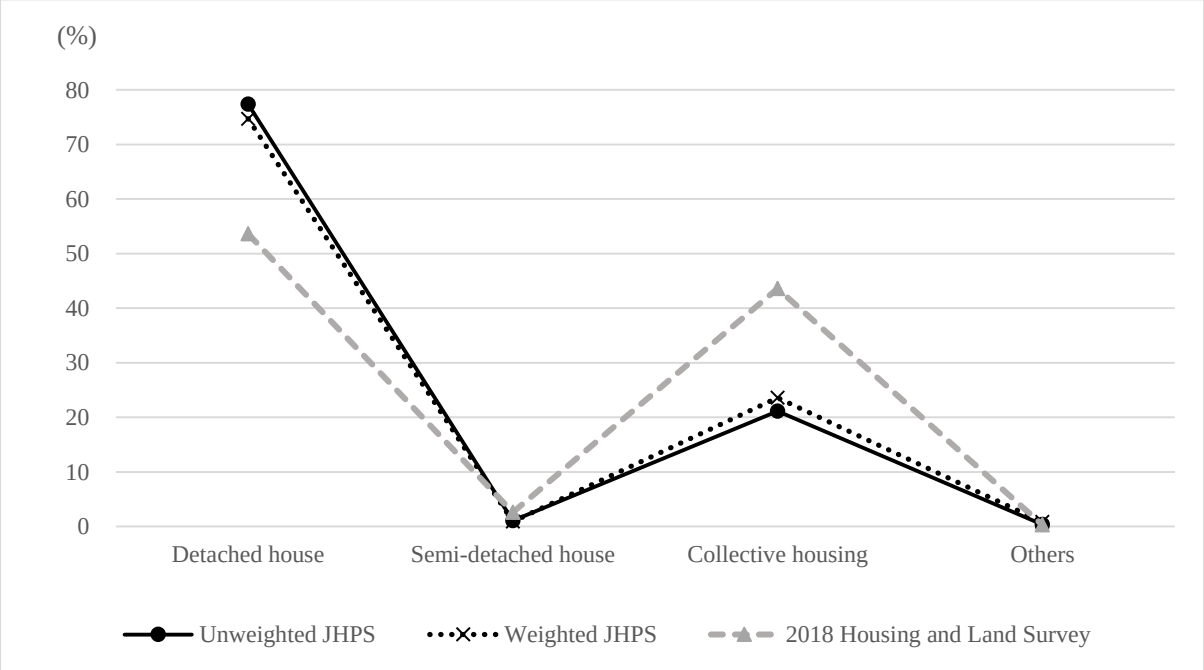
single-household respondents. Thus, it is possible to observe that the weights mitigate those biases but cannot completely correct them.

Figure 10. Household size in 2015



Source: Authors, based on the JHPS data and the 2015 Census.

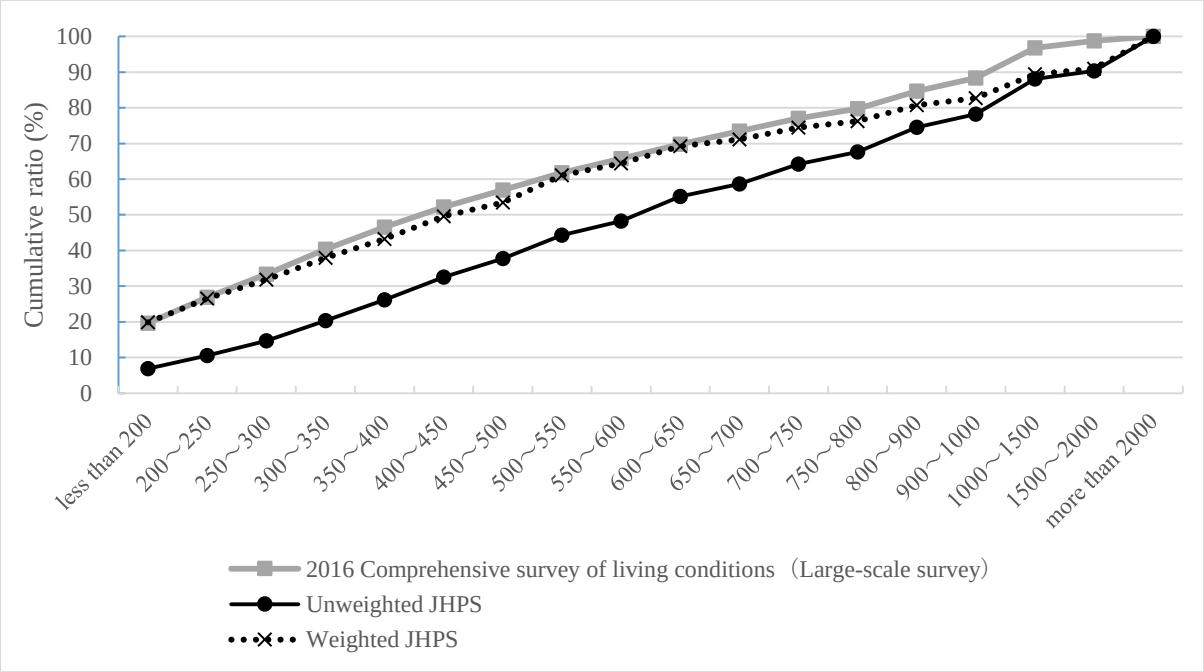
Figure 11. Type of dwelling in 2017



Source: Authors, based on the JHPS data and the 2018 Housing and Land Survey.

For household income, the JHPS collects less data from low-income households. However, the weighted data correct this problem, which suggests that the collection resembles what is observed in the official statistics.

Figure 12. Cumulative ratio of household income in 2016 (ten thousand yens)



Source: Authors, based on the JHPS data and the 2016 Comprehensive Survey of Living Condition (Large-scale Survey).

Last, Table 8 presents data for household deposits and savings, household securities, and household debts. Once again, we verify that the unweighted data slightly deviate from the official statistics; however, with the use of the sampling weights, it is possible to obtain values for the JHPS data that are more representative of the reality of the Japanese population.

Table 8. Deposits and saving, securities and debts in 2014

	Deposits and savings		Securities		Debts	
	Average amount (ten thousand yens)	Holding ratio (%)	Average amount (ten thousand yens)	Holding ratio (%)	Average amount (ten thousand yens)	Holding ratio (%)
Unweighted JHPS	912.2	79.8	229.9	27.2	613.1	43.9
Weighted JHPS	931.6	78.2	223.7	25.2	563.5	41.8
2014 Comprehensive Survey of Living Conditions	973.8	71.3	215.9	24.7	533.3	41.8

Source: Authors, based on the JHPS data and the 2014 Comprehensive Survey of Living Condition.

The figures and table presented in this section compare the official statistics for the Japanese population at the individual and household levels with the data collected for the JHPS. It is possible to conclude that the JHPS data provide a good representation of the Japanese population, with individual-level data being more representative than household-level data. The importance of the use of sampling weights is also illustrated, with weighted JHPS data achieving values closer to those of the official statistics than unweighted JHPS data.

References

Deville J.C. and Särndal C.E. (1992). *Calibration Estimators in Survey Sampling*. Journal of the American Statistical Association, pp. 376-382

Lavallée P. (2007). *Indirect Sampling*. Springer Series in Statistics.

Lundström S. and Särndal C.E. (1999). *Calibration as a Standard Method for Treatment of Nonresponse*. Journal of Official Statistics, Vol. 15, No. 2, pp. 305-327.

Naoi M., Yamamoto K., and Miyauchi T. (2010) JHPS Kaitoujokyo niokeru Chosajissihouhou no Performance (*in Japanese*), PDRC Discussion Paper DP-2009-005.

Osier G. (2016). *Unit non-response in household wealth surveys: Experience from the Eurosystem's Household Finance and Consumption Survey*. ECB Statistics Paper Series, No.15 / July 2016.

Särndal C.E. and Lundström S. (2005). *Estimation in Surveys with Nonresponse*. Wiley.

Schork J. (2018). *Automatic Variable Selection for Imputation Models: Common Methods Applied to EU-SILC*. Statec Économie et statistiques N° 98/2018. Available at : https://statistiques.public.lu/catalogue-publications/economie-statistiques/2018/98-2018.pdf

Understanding Society (2017). The UK Household Longitudinal Study: wave 1-7, 2009-2016 User Guide. Ed. Knies, Gundi. Colchester: University of Essex.

Understanding Society (2019). The UK Household Longitudinal Study: Wave 1-9, 2009-2018 User Guide. Colchester: University of Essex.

Watson N. (2012) Longitudinal and Cross-sectional Weighting Methodology for the HILDA Survey, *HILDA Project Technical Paper Series. No*.2/12.

Appendix A: Benchmarks used in weighting

The population benchmark data used in the weight calibration process were calculated based on made-to-order aggregated data provided by the Japanese National Statistics Center (NSTAC). A tabulation of the following data was requested for the NSTAC:

1- Japanese male population by 5-year-old category;
2- Japanese female population by 5-year-old category;
3- Japanese population by 5-year category and marital status;
4- Japanese population by 5-year category and region; and
5- Japanese population by 5-year category and employment status.

The tables for the aforementioned data were created based on the Labor Force Survey from the Statistics Bureau of Japan. The NSTAC elaborated the tables using quarterly data (January, February, and March) for 2004-2017. [19] Given that these are made-to-order tabulations and that the Statistics Bureau of Japan periodically updates its statistics, there is a possibility that the employed data can slightly differ from the data published by the Labor Force Survey, Statistics Bureau of Japan.

The population benchmark data are composed of the following calibration variables:
- Age group: yearly data of the Japanese population calculated by 5-year-old category (e.g., population from 20 to 24 years old, population from 25 to 29 years old, etc.).
- Sex: yearly data of the Japanese population by male and female group.
- Marital status: yearly data of the Japanese population by married and unmarried group.
- Region: yearly data of the Japanese population calculated per region (Hokkaido/Tohoku/Kanto/Chubu/Kinki/Chugoku/Shikoku/Kyushu and Okinawa).
- Employment status: yearly data of the Japanese population calculated by employment status (e.g., mainly working, working and studying, unemployed, etc.).

[19] Given the Great Disaster that occurred in 2011, the collection of data in March 2011 was not possible. Consequently, the obtained data for 2011 was calculated by the NSTAC based on the average of the data collected in January and February instead of quarterly data. In addition to this, data for the population of Tohoku is also not available for that year. In this case, Tohoku's population by 5-year-old category was calculated by employing the average of the available data for years 2010 and 2012.

Appendix B: Logistic regression model for response propensity

Based on the information available in each wave of the survey, the logistic regression model employed to control for attrition was defined as follows:

$$
\begin{aligned}
logit\,\{\Pr(resp_{it+1} = 1|X_{it})\} \\
&= \beta_0 + \beta_1 Marital_{it} + \beta_2 Sex_{it} + \beta_3 Age_{it} + \beta_4 HHsize_{it} + \beta_5 Income_{it} \\
&+ \beta_6 Kids_{it} + \beta_7 Change1_{it} + \beta_8 Change2_{it} + \beta_9 Change3_{it} \\
&+ \beta_{10} Change4_{it} + \beta_{11} Employment_{it} + \beta_{12} Health_{it} + \beta_{13} Region_{it} \\
&+ \beta_{14} City_{it} + \beta_{15} Wave_{it}
\end{aligned}
$$

where $resp_{it+1}$ assumes the value of 1 if individual *i* participates in the survey in year *t+1* and zero otherwise. The probability of individual *i* participating in the survey in year *t+1* is defined by individual *i*'s marital status, sex, age group, household size, income, the presence of kids in the household, changes in individual *i*'s household, employment status, health status, the region where individual *i* lives, the size of the city where individual *i* lives, and the number of survey waves individual *i* participated until year *t*.

Although the information for the auxiliary variables was collected from the survey in year *t* and all individuals *i* participated in this survey, it is not uncommon for the existence of situations where the respondent unintentionally forgets or intentionally does not answer a given question of the survey. The item nonresponse issue can lead to the deletion of the observations of individuals during the implementation of the logistic model. To avoid the deletion of cases with missing values, an imputation process based on multiple imputed chained equations (MICE) was performed before the implementation of the logistic model, replacing all the missing values with estimates.