Panel Data Research Center at Keio University

Longitudinal and Cross-sectional Weighting Methodology of Keio Household Panel Survey and Japan Household Panel Survey

Kayoko ISHII (Keio University)

Kayo NOZAKI (Hitotsubashi University)

Presentation at
the 20th International Panel Data Conference,
Hitotsubashi Hall in Tokyo, July 9, 2014

Topics

1. About DATA
   - Keio Household Panel Survey
   - Japan Household Panel Survey

2. How we develop the WEIGHTs

   2-1. Cross-sectional weight for Wave 1
   - Adjusting for unequal probabilities of selection and non-response.
   - Making inferences about the population from a sample.

   2-2. Longitudinal weight for Wave 2 and its followings
   - Adjusting for the sample attrition to make it close to wave 1 sample.

   2-3. Panel Allocation Factor for integration of panels
   - Integration of different samples.
     - KHPS main sample + top-up samples.
     - KHPS samples + JHPS sample

1. ABOUT DATA

1. DATA

- Keio Household Panel Survey (KHPS)
  - Since 2004
  - Approximately 4,000 individuals participated in Wave 1.
  - Top-up samples in 2007, and in 2012
  - Research object: persons aged 20-69 when selected.

- Japan Household Panel Survey (JHPS)
  - Since 2009
  - Research object: persons aged 20 and over when selected.
  - Approximately 4,000 individuals participated in Wave 1.
- The common features;
  - KHPS and JHPS have a lot of questions in common.
  - Sampling methods are almost the same among the two surveys.

Sampling methods of KHPS/JHPS

1. The unit of sampling is individual.
2. The sample is selected by two-stage stratified random sampling.
3. **Reserved subjects** are set who are in the same region, the same sex and the same age group as the original subjects.
4. For married respondents, the survey asks the same questions to their spouses.

More about “reserved subjects”

- The purpose
  - In case original subjects decline to participate in the survey, they are replaced by reserved subjects. This makes it possible to guarantee the scheduled sample size.
- Sampling methods of reserved subjects.
  - For one original subject, ten reserved subjects are allocated.
  - With the region, sex and age category controlled as the same as the original subjects, reserved subjects are selected randomly.

Two-stage stratified random sampling

1. Stratification.
   - The whole country was stratified into 24 strata by region and city class.
2. Allocate samples to each stratum.
   - The number of samples to each stratum was decided according to the population ratio.
3. Determine areas surveyed in each stratum.
   - The number of areas surveyed in each stratum was chosen randomly. In each area, approximately ten subjects are expected.

Do we really need weights ?

- The previous studies show that

  KHPS and JHPS are not seriously biased.

  (木村(2005), 直井・山本(2010))

  – Compared with the official statistics, there are small differences but most of them are not statistically significant.
    – Less single persons.
    – More self-employed. More house owners.

Still, it is better to have weights, especially for attrition problems.

2. DEVELOPING WEIGHTS

“General Steps in Weighting”
quoted from Watson (2012)

1. Determine which sample units are in-scope of the population.
2. Calculate the initial weights as the inverse of the probability of selection.
3. Adjust for non-response by developing response homogeneous groups or modeling response propensities.
4. Calibrate to known benchmarks to ensure the certain weighted estimates match (typically external) high quality totals.

(quoted from Watson 2012, p.8)

2-1. Cross-sectional weight

1. Determine which sample units are in-scope of the population.

   – All respondents are in-scope of the population.

   - The unit of sampling of KHPS/JHPS is INDIVIDUAL.
   - We create the cross-sectional weights only at the individual level.
     – For cross-sectional purposes, the units of interest are also household and enumerated persons, therefore we will think about providing weights for households and for enumerated persons in the future.

2-1. Cross-sectional weight

2. Calculate the initial weights as the inverse of the probability of selection.
   - We did not do the above, because:
     - For the original subjects, because of random sampling, the probabilities of selection are equal among them.
     - For the reserved subjects, their age, sex and location are the same as the original ones, so the distributions of these variables are fixed to the result of random sampling.
     - Random sampling is applied for the reserved subjects besides those variables.

2-1. Cross-sectional weight

3. Adjust for non-response.
4. Calibrate to known benchmarks

   – Adjusting for non-response with “raking”.
   - Unfortunately, we do not have enough information about subjects who declined to participate in the surveys.
   - We adjust for non-response with the a method called raking or iterative proportional fitting, and also calibrate to known benchmarks.

Raking

- Raking (or Iterative proportional fitting)
  - An alternative method of post-stratification method.
    - Specified by Deming and Stephan (1940).
    - Raking weights are computed using Bergmann's Stata IPFWEIGHT module.
      - Available at http://ideas.repec.org/c/boc/bocode/s457353.html or via STATA Ado-Search.
  - When you only have separate frequency tables for each benchmark variable (not cross-tabbed with each other), raking is useful. (Johnson 2008.)
- Benchmarks:
  - Education level, employment status, age group, marital status by sex, and region.

Example of Raking
(referring to Johnson 2008)

Example with education, marital status and age.

1. Compute education weight .
   Weight data by this weight.
   Generate the weighted frequency table for marital status.

2. Compute marital status weight
   Weight by education weight & marital status weight.
   Generate the weighted frequency table for age.

3. Compute age weight
   Weight by education weight & marital status weight & age weight.
   Generate the weighted frequency for education.

4. Compute a second education weight
   Weight by education weight & marital status weight & age weight & 2nd education weight.
   Generate the weighted frequency for marital status.

Continue process until the weighted frequencies and the population frequencies don't change.

Benchmarks used in raking

Variables	Benchmark statistics	KHPS2004	KHPS2007	KHPS2012	JHPS2009
educational level	Employment status survey, Ministry of Internal Affairs and Communications	2002	2007	2012	2007
employment status	Population census, Ministry of Internal Affairs and Communications	2005	2005	2010	2010
married by sex	Population census, Ministry of Internal Affairs and Communications	2005	2005	2010	2010
age groups	Population estimates, Ministry of Internal Affairs and Communications	Oct, 2003	Oct, 2006	Oct, 2011	Oct, 2008
region	KHPS/JHPS	KHPS2004	KHPS2007	KHPS2012	JHPS2009

The results of raking

	KHPS2004			JHPS2009		
	K04	benchmark	K04 weighted	J09	benchmark	J09 weighted
	(1)	(2)	(3)	(1)	(2)	(3)
education						
junior high school grad	13.6%	23.8%	23.8%	12.0%	19.6%	19.1%
high school grad	52.5%	43.6%	43.6%	50.4%	52.5%	51.1%
college grad	11.8%	13.7%	13.7%	11.9%	8.5%	8.3%
university grad	19.4%	16.2%	16.2%	22.8%	19.4%	18.9%
currently in school	2.7%	2.7%	2.7%	3.0%	2.8%	2.7%
marital status by sex						
man* married	36.8%	31.0%	31.0%	35.3%	30.6%	30.6%
woman* married	36.1%	33.0%	33.0%	36.6%	30.6%	30.6%
man* not married	13.2%	18.8%	18.8%	13.2%	17.5%	17.5%
woman * not married	14.0%	17.1%	17.1%	14.9%	21.4%	21.4%

note: (1) : un-weighted values
(2) : benchmark values
(3) : cross-sectional weighted values

2-2. Developing Longitudinal weight

Longitudinal weights are to adjust for attrition from the initial wave.

1. Determine which sample units are in-scope of the population.
   - All respondents who continue to respond from wave1 to wave t.

2. Calculate the initial weights as the inverse of the probability of selection.
   - The initial weight of longitudinal weight is the final cross-sectional weight.

3. Adjust for non-response.
   - To adjust for attrition, the initial weights of longitudinal weights are multiplied by the inverse of the probability of responding.
   - The probability of responding is predicted by constructing a logistic model. (If responding, Y=1, otherwise Y=0)

2-2. Developing Longitudinal weight

[Example: Longitudinal weight for wave5]

$$W_{adjust,wave5} = W_{adjust,wave1} \frac{1}{P_{response,wave2\ to\ wave5}}$$

$P_{response,wave2\ to\ wave5}$: the probability of responding from wave2 to wave5.

- The logistic regression for calculating the probability of continuously responding
  - Y: continuously responding=1, 0 otherwise.
  - X: Region, Number of household members, Sex, Original subject, Marital status, House owner, Educational level, Working condition, Age, Household income.
    - Xs are the information at the time of the initial wave.

Logistic model with continuously responding to the initial wave to wave t [Example: KHPS2004]

	KHPS 2004								
	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013
Number of household members									
2 people	0.331*	0.395**	0.392**	0.353**	0.375**	0.399***	0.326**	0.273*	0.216
	[0.186]	[0.162]	[0.156]	[0.153]	[0.153]	[0.154]	[0.155]	[0.155]	[0.156]
3 people	0.410**	0.452***	0.390**	0.353**	0.481***	0.484***	0.458***	0.359**	0.356**
	[0.185]	[0.160]	[0.154]	[0.151]	[0.151]	[0.151]	[0.152]	[0.152]	[0.153]
4 people	0.510***	0.601***	0.507***	0.418***	0.512***	0.504***	0.430***	0.347**	0.286*
	[0.196]	[0.168]	[0.161]	[0.158]	[0.157]	[0.157]	[0.157]	[0.158]	[0.158]
More than 5 people	0.424**	0.542***	0.564***	0.488***	0.481***	0.484***	0.428***	0.364**	0.323*
	[0.203]	[0.176]	[0.170]	[0.166]	[0.165]	[0.165]	[0.166]	[0.166]	[0.167]
Male	-0.0625	-0.0587	-0.132*	-0.158**	-0.167**	-0.180**	-0.151**	-0.182**	-0.146**
	[0.0917]	[0.0782]	[0.0744]	[0.0726]	[0.0716]	[0.0712]	[0.0711]	[0.0711]	[0.0712]
Original subject	-0.168	-0.134	-0.159*	-0.0986	-0.103	-0.118	-0.111	-0.111	-0.113
	[0.107]	[0.0912]	[0.0866]	[0.0847]	[0.0838]	[0.0834]	[0.0831]	[0.0833]	[0.0835]
Married	0.183	0.194*	0.193**	0.198**	0.185**	0.183*	0.228**	0.209**	0.164*
	[0.118]	[0.102]	[0.0971]	[0.0949]	[0.0944]	[0.0941]	[0.0945]	[0.0950]	[0.0955]
House owner	-0.171	-0.287***	-0.181**	-0.174**	-0.201**	-0.257***	-0.256***	-0.236***	-0.221***
	[0.111]	[0.0946]	[0.0879]	[0.0854]	[0.0845]	[0.0840]	[0.0836]	[0.0831]	[0.0832]
Education									
High School	0.161	0.176	0.318***	0.329***	0.406***	0.381***	0.397***	0.355***	0.321***
	[0.138]	[0.119]	[0.114]	[0.111]	[0.111]	[0.111]	[0.111]	[0.112]	[0.113]
Junior college or	0.565***	0.449***	0.481***	0.476***	0.590***	0.512***	0.553***	0.512***	0.460***
Higher professional school	[0.192]	[0.159]	[0.150]	[0.146]	[0.145]	[0.144]	[0.144]	[0.145]	[0.145]
University or	0.345**	0.315**	0.458***	0.453***	0.510***	0.433***	0.435***	0.405***	0.383***
Graduate school	[0.163]	[0.141]	[0.134]	[0.131]	[0.131]	[0.130]	[0.130]	[0.131]	[0.132]
Other or Nonresponse	0.332	0.329	0.359*	0.376**	0.429**	0.293	0.315*	0.338*	0.376**
	[0.243]	[0.206]	[0.192]	[0.188]	[0.187]	[0.185]	[0.185]	[0.185]	[0.184]

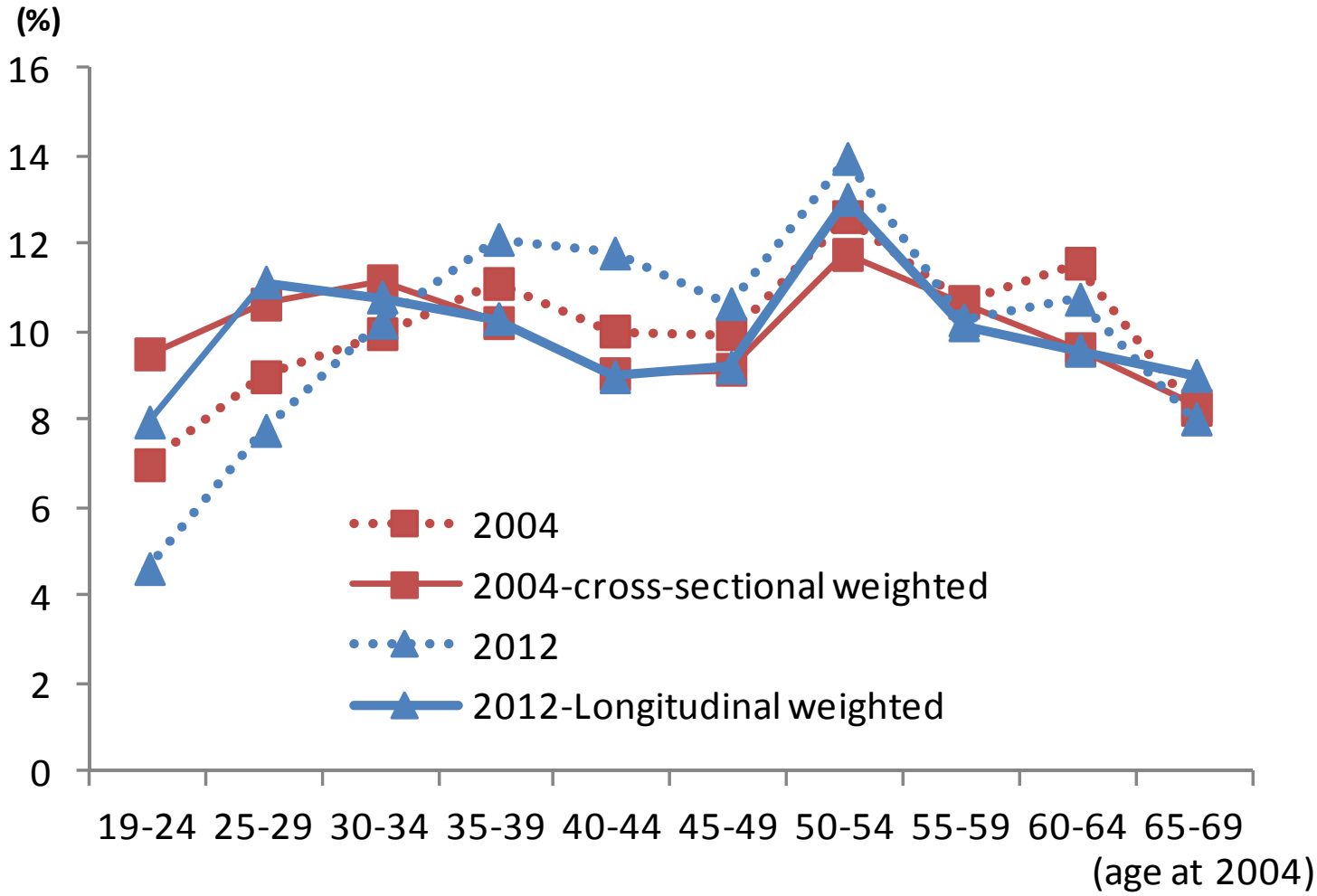
continued

	KHPS 2004								
	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013
Worker	-0.0829	-0.118	-0.0853	-0.0705	-0.148*	-0.129	-0.138*	-0.157*	-0.118
	[0.103]	[0.0885]	[0.0839]	[0.0817]	[0.0809]	[0.0805]	[0.0802]	[0.0805]	[0.0807]
Age									
30-39	0.121	0.257**	0.256**	0.345***	0.347***	0.456***	0.415***	0.449***	0.449***
	[0.144]	[0.124]	[0.118]	[0.115]	[0.114]	[0.114]	[0.114]	[0.115]	[0.116]
40-49	0.440***	0.444***	0.407***	0.522***	0.535***	0.615***	0.633***	0.614***	0.660***
	[0.156]	[0.131]	[0.124]	[0.121]	[0.120]	[0.120]	[0.120]	[0.121]	[0.121]
50-59	0.198	0.332**	0.406***	0.544***	0.605***	0.643***	0.583***	0.550***	0.526***
	[0.151]	[0.131]	[0.126]	[0.124]	[0.123]	[0.122]	[0.123]	[0.123]	[0.124]
60-69	0.392**	0.449***	0.349**	0.447***	0.520***	0.527***	0.538***	0.515***	0.460***
	[0.174]	[0.148]	[0.141]	[0.137]	[0.136]	[0.136]	[0.136]	[0.136]	[0.137]
Latter half in each age group	-0.0262	0.01	-0.0461	-0.018	-0.0431	-0.0176	0.00413	-0.00174	-0.0167
	[0.0859]	[0.0726]	[0.0689]	[0.0672]	[0.0662]	[0.0658]	[0.0656]	[0.0656]	[0.0657]
Household Income (ten thousand yen)									
360-559	0.161	0.148	0.0804	0.0263	0.001	0.0544	0.0535	0.0316	-0.0173
	[0.126]	[0.107]	[0.101]	[0.0972]	[0.0961]	[0.0952]	[0.0946]	[0.0944]	[0.0944]
560-849	0.16	0.0121	-0.0481	-0.0654	-0.153	-0.0835	-0.0822	-0.0953	-0.172
	[0.147]	[0.122]	[0.115]	[0.112]	[0.111]	[0.110]	[0.110]	[0.110]	[0.110]
More than 850	0.328**	0.190*	0.0685	0.0685	-0.0799	-0.0627	-0.0531	-0.0888	-0.115
	[0.136]	[0.113]	[0.106]	[0.103]	[0.101]	[0.0996]	[0.0993]	[0.0992]	[0.0992]
0 or Nonresponse	-0.327**	-0.313***	-0.304***	-0.247**	-0.319***	-0.291**	-0.328***	-0.395***	-0.384***
	[0.136]	[0.120]	[0.116]	[0.115]	[0.115]	[0.115]	[0.116]	[0.117]	[0.118]
Constant	0.712***	0.073	-0.248	-0.525***	-0.726***	-0.871***	-0.972***	-0.953***	-0.965***
	[0.237]	[0.203]	[0.195]	[0.191]	[0.190]	[0.189]	[0.191]	[0.191]	[0.192]
Number of observations	4,005	4,005	4,005	4,005	4,005	4,005	4,005	4,005	4,005
Log Likelihood	-1797	-2311	-2506	-2607	-2661	-2686	-2698	-2699	-2694

Robust standard errors in brackets

*** p<0.01, ** p<0.05, * p<0.1

Age distributions using longitudinal weights （wave1(2004) and wave9(2012) of KHPS2004data）



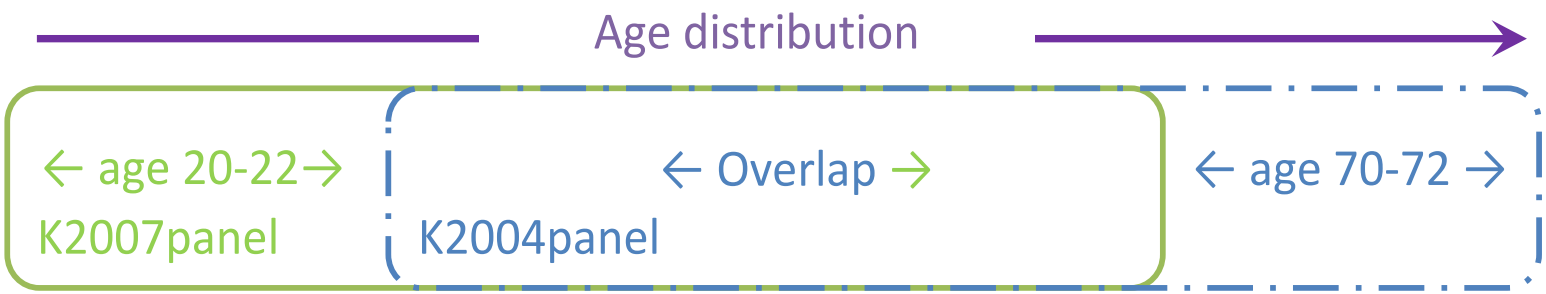
3. Panel Allocation Factor
integration of samples

- The purposes
  - To integrate KHPS main sample and the top-up samples.
  - To integrate KHPS sample and JHPS sample.
    - Fortunately, KHPS and JHPS have a lot of questions in common.
    - It allows you to analyze a bigger sample size.
- The concept of panel allocation factor
  - Identify the differences in survey population in each sample combined.
    - The part of the population of the main (or top-up) sample is not present in the population of the top-up (or main) sample.
  - Apply a weight less than 1 where the populations overlap.

Examples

Identifying the differences in survey population in each sample

- Integration of KHPS2004 sample and KHPS2007top-up sample in 2007.
  - The differences in survey population in each sample combined.
    - Persons aged 70-72 in K2004 are not present in the population of K2007topup .
    - Persons aged 20-22 in K2007topup are not present in the population of K2004.
  - Panel allocation factor
    - Apply a weight less than 1 to respondents aged 23-69 in K2004.
    - Apply a weight less than 1 to respondents aged 23-69 in K2007topup.

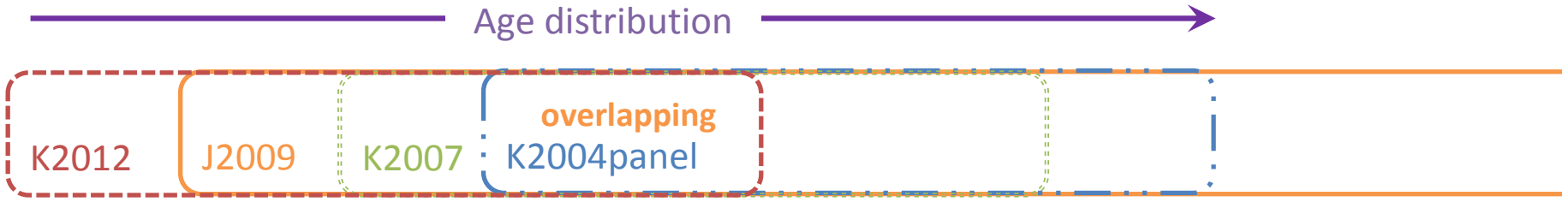


Another example

Identifying the differences in survey population in each sample

- Integration of K2004, K2007topup, J2009, and K2012topup in 2012.
  - The differences in survey population in each sample combined.
    - Persons aged 20-22 in K2012topup are not present in the population of other samples.
    - Persons aged 23-27 in K2007topup and K2012topup are not present in the population of other samples.

      And so on……
  - Panel allocation factor
    - Apply a weight less than 1 to respondents aged 28-69 in all samples.
    - Apply a weight less than 1 to respondents aged 25-27 in K07, J09, K12.
    - Apply a weight less than 1 to respondents aged 23-24 in J09, K12.
    - Apply a weight less than 1 to respondents aged 70-74 in K04, K07, J09.
    - Apply a weight less than 1 to respondents aged 75-77 in K04, J09.



How to calculate panel allocation factor

[Example] Integration of K04 and K07 top-up sample

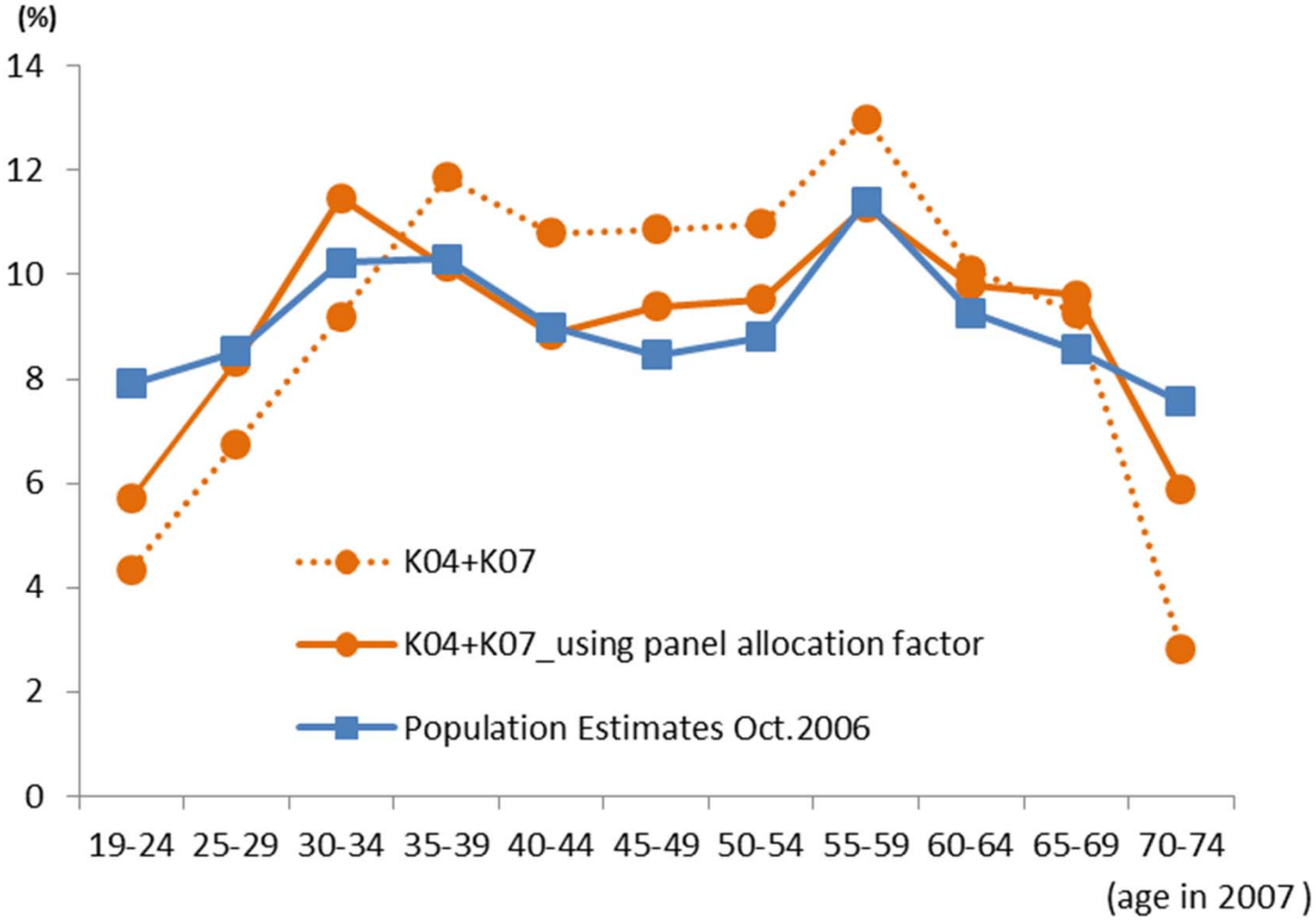
$$\hat{Y}_{combined} = \theta Y_{K04,wave4} + (1-\theta) Y_{K07,wave1}$$

$$\theta = \frac{n_{K04}}{n_{K04} + n_{K07}(\frac{d_{K04}}{d_{K07}})}$$

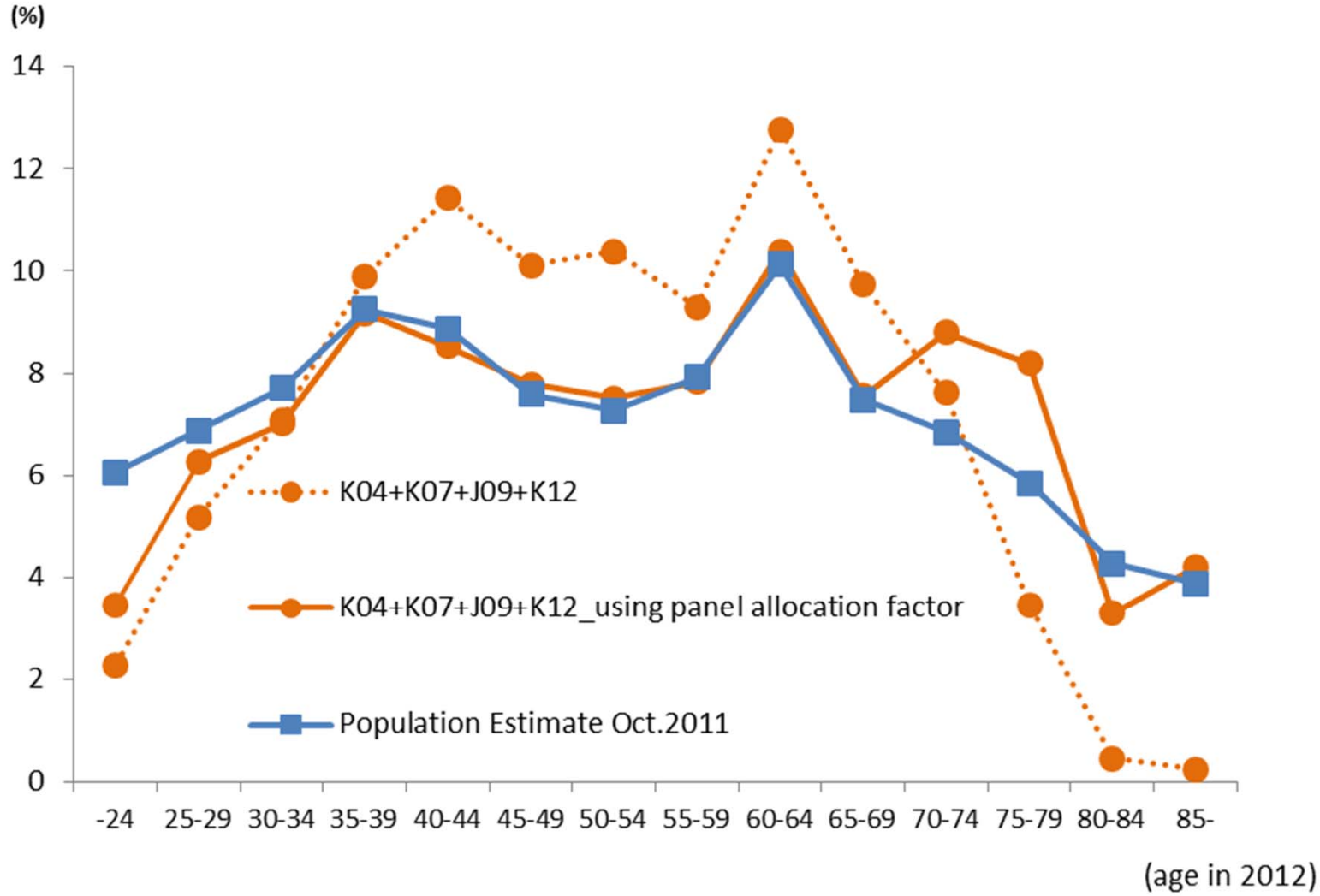
$$W_{combined} = \begin{cases} \theta \times w_{K04,adjust\ w6} : For\ the\ overlapping\ portion\ of\ K04 \\ (1-\theta) \times W_{K07,design} : For\ the\ overlapping\ portion\ of\ K07 \end{cases}$$

- θ is a panel allocation factor which becomes a weight for integration.
- n is the sample size of each panel at the time of integration.
- Suppose the sample designs of K04 and K07 are same, design effect(d_{K04}/d_{K07}) is equal to 1, therefore θ is decided by the ratio of those sample sizes.

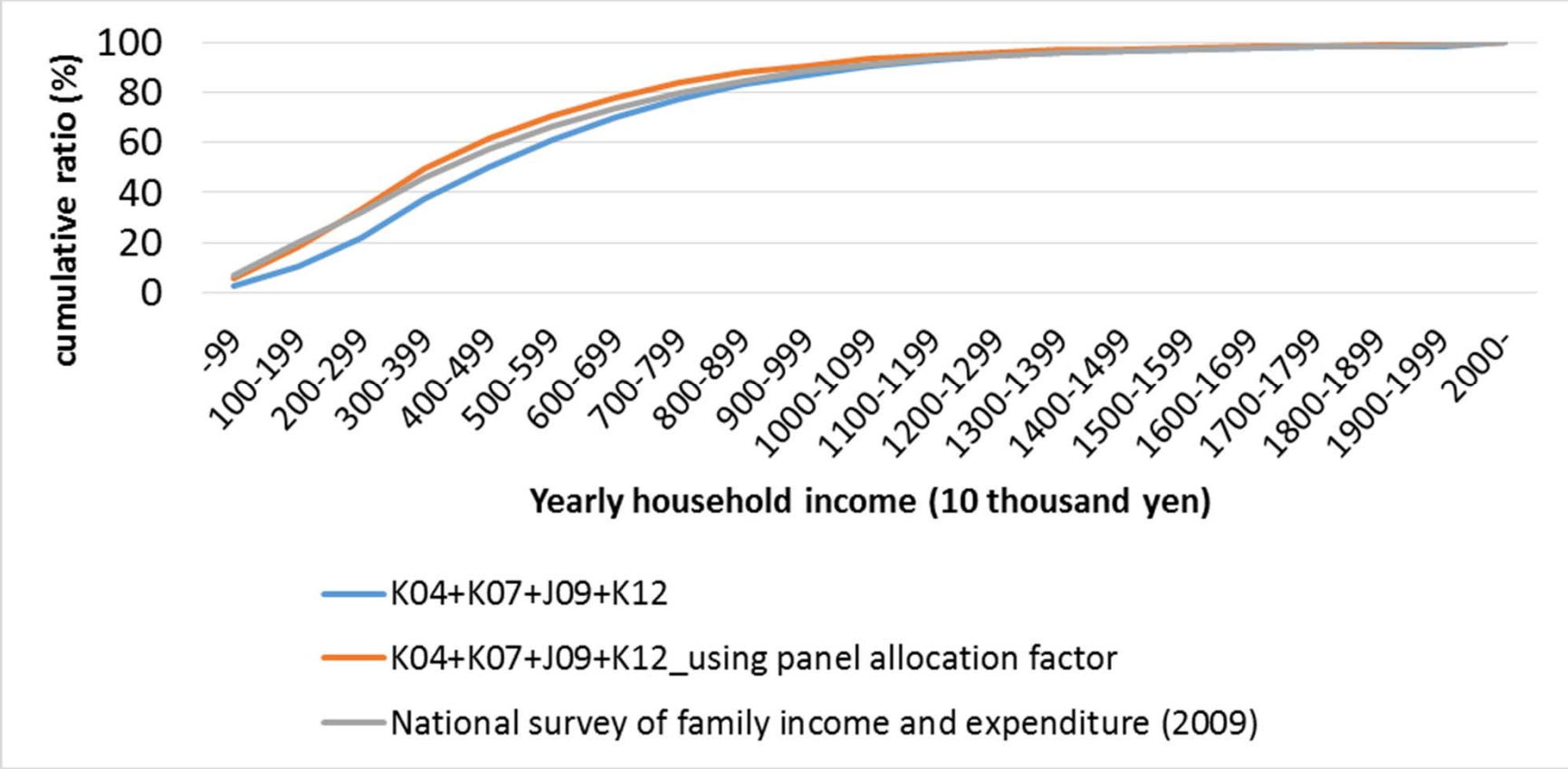
Age distribution after combining KHPS2004 and KHPS2007top-up using panel allocation factor



Age distribution after combining K04, K07top-up, J09 and K12top-up using panel allocation factor



Household income distribution compared with the official statistics



* We summarize KHPS and JHPS weighted by the inverse ratio of household size, because the sampling unit of National survey of family income and expenditure is household.

Conclusion and Limitation

- By using the weights, we could bring the age distribution in KHPS/JHPS close to the distribution of population estimates in the official statistics.

- We need to think about creating the cross-sectional weights for household and for enumerated persons.

- We need to make a regression analysis using weight, and check if the coefficient will change compared to one without weight.

Reference

- Deming, W. E., and F. F. Stephan. (1940), “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known”, *The Annals of Mathematical Statistics*, Vol. 11, No. 4, pp. 427-444.
- European Commission (2011), ‘Description of target variables: Cross-sectional and Longitudinal’, EU-SILC 065 (2011 operation).
- Johnson ,David R. (2008), “Using Weights in the Analysis of Survey Data” (PPT slide), Department of Sociology, Population Research Institute, The Pennsylvania State University.
- LaRoche, S. (2003), ‘Longitudinal and cross-sectional weighting of the Survey of Labour and Income Dynamics’, Income Research Paper Series, Cat.No. 75F0002MIE, No.007, Statistics Canada.
- Watson, N. (2012), “Longitudinal and cross-sectional weighting methodology for the HILDA Survey”, HILDA Project Technical Paper Series, No.2/12, Melbourne Institute of Applied Economic and Social Research.
- 木村正一 (2005)「2004年慶應義塾大学家計パネル調査の標本特性」,『日本の家計行動のダイナミズムⅠ』慶應義塾大学出版会.
- 直井道生・山本耕資 (2010)「日本家計パネル調査の標本設計と代表性」,『貧困のダイナミズム――日本の税社会保障・雇用政策と家計行動』慶應義塾大学出版会.

Acknowledgment

- The data for this analysis, Keio Household Panel Survey and Japan Household Panel Survey, were provided by the Keio University Panel Data Research Center.