Sampling Design and Weight

June 2025

1. Introduction

The research project entitled Health, Aging, and Retirement in Thailand (HART) is a longitudinal survey aimed at building a nationally representative data infrastructure to study health, aging, and retirement behaviors among the Thai population aged 45 and above. Utilizing a multi-stage probability sampling method, the project was designed to obtain a sample representative of two key demographic groups: the "pre-elderly" (ages 45–59) and the "elderly" (aged 60 and above).

The use of a nationally representative sample makes sampling design and weight essential components of the research. These statistical mechanisms serve to (1) control for sampling bias, and (2) enable estimation and generalization from the sample to the broader population of 22.7 million Thai individuals aged 45 and over.

This document aims to explain the sampling design and the calculation of survey weights used in the HART project. It covers key methodological aspects including the design of the sampling framework, the computation of weights, and the procedures for validating the accuracy and reliability of the weights. By understanding the rationale, methodology, and limitations of the sampling process, users of the HART dataset will be better equipped to apply the survey weights correctly. This ensures that research findings accurately reflect the demographic characteristics of the Thai population aged 45 and above and are comparable to internationally recognized aging surveys such as the Health and Retirement Study (HRS) in the United States and the Survey of Health, Ageing and Retirement in Europe (SHARE).

2. Sampling Design

The HART project commenced its first wave of data collection in 2015. The population of the HART study comprised **Thai individuals aged 45 years or older** as of 2015, or those born before 1970. According to data from the Ministry of Interior, the total number of individuals aged 45 and above nationwide was 22,714,654.

A total sample size of 5,600 units was determined for the study. The sampling design employed in the HART project was **multi-stages stratified random sampling**, as outlined below

2.1 Primary Stage: Selecting Provinces

In the first stage of sampling, the HART project divided the geographical area of Thailand into six regions as follows: (1) Bangkok and Vicinity, (2) Central Region (excluding Bangkok and Vicinity), (3) Eastern Region, (4) Northern Region, (5) Northeastern Region, and (6) Southern Region

Region	No. of provinces in region	Provinces selected
Bangkok and Vicinity	4	4
Central	16	2
Eastern	7	1
Northern	16	2
Northeastern	20	2
Southern	14	2
Total	**77**	**13**

All four provinces in Bangkok and Vicinity were taken with certainty because of their very large population. For the remaining regions, provinces were selected with probability proportional to size with a fixed sample size (PPS)-two provinces in Central, Northern, and Northeastern regions, and one province in the smaller Eastern region.

2.2 Secondary-Stage Selection: District Level

In the second stage of sampling, two districts were randomly selected from each province, which was sampled in Stage 1. An exception was made for Bangkok, where the population density is particularly high and the number of individuals aged 45 and older is substantial. Therefore, six districts (known as khet) were selected from Bangkok.

2.3 Third-Stage Selection: Individuals Aged 45 and Older

Within each sampled district, individuals aged 45 years and above were randomly selected from the district's population register to meet the required district sample size.

3. Calculation of Sample Weights

In HART project, the data obtained from the nationally representative survey are critically important for accurately reflecting the characteristics and behaviors of the population—Thai individuals aged 45 years and older. The HART study employed a multistage sampling design to ensure a diverse and comprehensive sample representing the population across all regions of the country.

However, due to differences in selection probabilities at each sampling stage, the raw data may not accurately reflect the true distribution of the national population. Therefore, the calculation of sampling weights is an essential process to adjust the sampled units. These weights are designed to compensate for unequal probabilities of selection and to correct for overrepresentation or underrepresentation of certain subgroups within the sample.

Weighting is thus a crucial tool for enhancing the accuracy of analytical results and for ensuring that survey data can be effectively used to inform health and social policy planning for the aging population in Thailand. In addition, the application of sampling weight strengthens the credibility of the study's findings, particularly when presenting nationally representative estimates on living conditions, health status, and retirement behaviors of the population.

For these reasons, the proper design and calculation of sample weights are important in the HART study to ensure that the research findings are both reflective of reality and useful for future policy development and planning.

3.1 Principles of Weight Calculation

The calculation of sampling weights is a crucial process in the HART study, intended to adjust sample data so that it more accurately reflects the characteristics of the target population. This is particularly important when individual sampling units are selected with unequal probabilities. The objective of the weighting procedure in HART is to correct for sample imbalance and align survey results with the national population of interest—Thai individuals aged 45 years and older.

A sample weight reflects the inverse of the probability of selection for a given sample unit. The basic formula is expressed as:

$$W_k = \frac{1}{\pi_k}$$

where W_k = the sample weight for unit k
π_k = the overall probability of selection for unit k

These weights help balance the data to ensure that it is representative of the population, particularly when certain subgroups are over- represented or under-represented in the sample relative to their true distribution in the population.

The HART study employed a multistage sampling design involving three stages:

1. **Province selection**: Provinces were selected using *Probability Proportional to Size (PPS) Sampling* with a fixed sample size.
2. **District selection**: Districts within each selected province were sampled using the same *PPS with fixed sample size* approach.
3. **Individual selection**: Individuals aged 45 and older were randomly selected from each selected district.

The overall probability of selection for an individual k in district j of province i, denoted as $\pi_{k,ij}$, is calculated as the product of the probabilities at each stage:

$$\pi_{k,ij} = \pi_i \times \pi_{j|i} \times \pi_{k|ij}$$

where: π_i = the probability of selecting province i
$\pi_{j|i}$ = the probability of selecting district j, province i
$\pi_{k|ij}$ = the probability of selecting individual k, district j, province i

3.2 Calculation of Selection Probabilities

3.2.1 Province-level probability (PPS, fixed n)

The sampling of provinces in each region will use **Probability Proportional to Size (PPS) Sampling with Fixed Sample Size method**. The calculation of the probability of sampling provinces in each region will use the same method, except for Bangkok and its vicinity, which selects every province in the region, resulting in the probability of each province being equal to 1. For other regions, the probability will be calculated from the following equation.

$$\pi_i = m\frac{Q_i}{\sum_{i\in U} Q_i}$$

where: π_i = the probability of selecting province i within each region
m = the number of provinces to be sampled in region
Q_i = the number of population persons in province i
$\sum_{i\in U} Q_i$ = the number of population in the region

3.2.2 District-level probability (PPS, fixed n)

The sampling of districts in each province will use the **Probability Proportional to Size (PPS) Sampling with Fixed Sample Size method**. The calculation of the probability of sampling districts in each province will use the same method. The probability is calculated from the following equation.

$$\pi_{j|i} = n_i\frac{Q_{ij}}{Q_i}$$

where: $\pi_{j|i}$ = the probability of selecting district j within province i
n_i = the number of districts sampled in province i
Q_{ij} = the number of targets-population persons in district j
Q_i = the number of target population size in province i

3.2.3 Individual-level probability (simple random sampling)

The sampling of individuals in each district uses **simple random sampling method** by calculating the probability of sampling individuals in each district from the following equation.

$$\pi_{k|ij} = \frac{q_{ij}}{Q_{ij}}$$

where: $\pi_{k|ij}$ = the probability of selecting individual k within district j, province i

q_{ij} = the number of samples required in district j , province i

Q_{ij} = target population size in the district j , province i

4. Verification and Normalization

Weights were checked so that their sum equals the total population of Thais aged 45 and over in 2015 (22,714,654 persons). Where necessary, weights were normalized to ensure this equality and to control for extreme values. The verification confirmed that the final weights sum precisely to the target population, supporting their use for national estimates.

Citation

If you use or refer to any of the documents, please cite them properly using the suggested citation formats below. This ensures academic integrity and helps others locate the original source.

Health, Aging, and Retirement in Thailand (HART). (2025). *Sampling Design and Weight*. Bangkok: National Institute of Development Administration (NIDA). Retrieved from https://hart.nida.ac.th/survey-design/