****

**The World Bank**

1818 H Street, NW

Washington, DC 20433

USA

Tajikistan

Household Budget Survey 2024

**SAMPLE DESIGN**

**Chiyu Niu**

**Economist, Poverty & Equity Global Practice, Europe and Central Asia**

*In collaboration with the*

**Tajikistan National Statistical Office (TajStat)**

8.15.2023[[1]](#footnote-1)

Contents

[Background 3](#_Toc144323464)

[Motivation 3](#_Toc144323465)

[Data Inputs 3](#_Toc144323466)

[Methodology 3](#_Toc144323467)

[Current Sample 4](#_Toc144323468)

[Refresh Sample 4](#_Toc144323469)

[Proposed Sample 5](#_Toc144323470)

[Selection of Clusters 6](#_Toc144323471)

[Conclusion 6](#_Toc144323472)

**Background** As part of an ongoing effort to quantify welfare and facilitate evidence based policies in Tajikistan, TajStat engages in the collection of consumption and expenditure data through a quarterly Household Budget Survey (HBS). This survey is intended to provide an impartial picture of the status and structure of income, expenditure, consumption and savings for the people of Tajikistan at the national and sub-national level, and to provide valid representativeness for the national level, for urban rural as-well-as for each primary administrative domain of interest; Dushanbe, Sogd, Khatlon, GBAO and DRS. Since 2009 the HBS has been the primary data source for welfare and poverty levels in Tajikistan.

**Motivation** Tajstat invited the World Bank team’s support to update the sampling design of the Household Budget Survey (HBS) for two reasons. First, the 2020 census data became available to update the sampling frame, which was based on the 2010 census. Second, Tajstat decided to increase the survey sample size from 3000 to 4000 households. As a result, this sample design improves 1) the HBS accuracy by realigning the sample with the latest census information on demographic shifts and changes in administrative areas, and 2) the HBS representativeness by increasing the sample size. The World Bank Listening-to-Tajikistan 2023 Baseline Survey will use the same sampling design as the HBS 2024.

**Data Inputs** This sampling design uses data from Census 2020 (“Census Frame\_2020\_cn.xlsx”), HBS 2021 (“TJK\_HBS2021\_Harmonization.dta”). The former provides information on census enumeration areas including their unique IDs, number of households, and population. HBS 2021 includes information on per capita household consumption.

**Methodology** The purpose of a sample design is to maximize the statistical efficiency given the survey budget or a fixed sample size. The statistical efficiency of a sample can be measured by the power defined as the probability of avoiding type-II error (wrongly accepting the null hypothesis), or equivalently, the standard error of the estimate of interests which in our case is the per capita expenditure.

Such equivalence can be understood in the following equation, where m is the sample size, Z\_1-a/2 is the Z-value for a two-sided type-I error (over rejecting a null hypothesis) of alpha percent, Z\_1-B/2 is the Z-value for a power of 1-B where B is probability of type-II error, s is the standard error of the outcome variable, delta is the clinically important difference in the outcome changes, (1 + (n-1)rho) is the design effects in a clustered sampling design which include both the cluster size n, and intra-class correlation rho. The ratio of sigma squared over delta squared is a function of the relative standard error

(RSE), so that when sample size m is fixed, the bigger the power, the smaller the RSE. The 2015 sample design utilized RSE as the measure for sample efficiency. This note employs the same measure. The RSE of a sample can be calculated as the following:

Where SE is the standard error of the estimate of the mean per capita expenditure, X bar is the mean per capita expenditure, s is the sample standard deviation of per capita expenditure, m is the sample size (or strata size when such calculation is done within a strata). A commonly accepted threshold of RSE for representativeness is 10 percent. The rest of this note explains the process of reducing RSE below the 10 percent threshold to achieve relevant representativeness given a fixed sample size.

**Current Sample** Table 1 shows the current HBS sample design and representativeness in terms of RSE. The sample size is 3000, stratified over 9 strata including Dushanbe, and the rural-urban areas in Sugd, Khatlon, RRP, and GBAO regions. The current sample design aimed to achieve representativeness on the national, 5 regions, and rural-urban levels, but not on the strata level. The clustered sampling follows a proportional-to-size (PPS) approach distributing 204 clusters across 9 strata based on total number of households per strata. The cluster size is determined by both the suggestions from the RSE analysis in 2015 and practical concerns on data collection difficulties in each stratum. Using the per capita expenditure estimates from HBS 2021, the RSEs aligns with the designed representativeness. However, GBAO region has the highest RSE of 9.8 percent which is on the edge of losing representativeness. Though the current HBS does not aim to achieve strata-level representativeness, many of the strata have satisfactory representativeness below 10 percent, with only RRP-Urban, GBAO-Urban and GBAO-Rural not passing the threshold.

Table 1. Current HBS Sample Design and RSE



**Refresh Sample** The refresh sample aims to add 1000 households to the current sample of 3000 households and to use census 2020 in the PPS sampling. Since the current sample has both attrition and replacements in some clusters, the cluster sizes differ slightly from the 2015 sampling design which took cluster design effects into account. The refresh sampling continues to use the cluster size as designed in the 2015 version which is indicated in the Table 2. The PPS procedure results in 59 additional clusters and 995 households compared to the current design. The resulting RSEs on the national, regional, and rural-urban levels are effectively reduced. However, GBAO region remains to have a high RSE at 9.13 percent while Dushanbe, Sugd, and Khatlon shows RSEs below 4 percent even before increasing the sample size. This means reallocation of sample size across these regions or strata can secure the representativeness of GBAO region while not sacrificing much of the rest.

Table 2. Refresh Sample Design and RSE



**Proposed Sample** Table 3 shows the proposed sample design that further reallocates sample observations across strata based on the PPS results in Table 2. Specifically, sample sizes of Sugd-Urban, Khatlon-Rural strata are reduced, while those of RRP-Urban, GBAO-Urban and GBAO-Rural strata increased. As a result, GBAO region reduced its RSE to 5.67 percent, and nearly all strata (except GBAO-urban) have RSEs below 10 percent. This means the proposed sampling design can make HBS 2024 representative on the strata level, too.



**Selection of Clusters** The selection of clusters in each stratum is based on the PPS approach. This approach assigns each cluster’s likelihood of selection relative to the percentage of stratum households residing in the cluster. By increasing the likelihood of selection for larger clusters, the sample will inherently capture more variation within the population, leading to lower design effects and higher precision. In the Tajikistan context, this method of selection further promotes representation in more densely populated and potentially more accessible districts. The dataset “xxx.dta” includes the selected clusters for each stratum.

**Conclusion** By using the most updated census information and increasing the sample size by 1000 households, the HBS 2024 sample design show significant improvements of survey representativeness and the level of confidence in resulting estimates. First, GBAO region shows a much higher level of representativeness in the proposed sample design compared to the current design. Second, by increasing the total sample size by 1000 households and reallocating observations from strata that already shows a good level of representativeness to those do not, the proposed sample design can achieve strata-level representativeness.

1. Major contributing inputs provided by Abduvali Kulov (Tajikistan Agency on Statistics, Demography, Employment, and Social Statistics, Department Head); Furkat Mirpochoev (Tajikistan Agency on Statistics , Census Department, Chief Specialist); Obert Pimhidzai (Lead Economist, World Bank), Alisher Rajabov (Senior Economist, World Bank), Avralt-Od Purevjav (Extended Term Consultant, World Bank). [↑](#footnote-ref-1)