Sampling Method the Indonesia High-frequency Phone-based Monitoring of COVID-19 Impacts on Households (HiFy)

(May 22, 2023)

Sampling Frame

The sampling frame of the Indonesia high-frequency phone-based monitoring of socio-economic impacts of COVID-19 on households was the list of households enumerated in three recent World Bank surveys, namely Urban Survey (US), Rural Poverty Survey (RPS), and Digital Economy Household Survey (DEHS). The US was conducted in 2018 with 3,527 sampled households living in the urban areas of 10 cities and 2 districts in 6 provinces. The RPS was conducted in 2019 with the sample size of 2,404 households living in rural areas of 12 districts in 6 provinces. The DEHS was conducted in 2020 with 3,107 sampled households, of which 2,079 households lived in urban areas and 1,028 households lived in rural areas in 26 districts and 31 cities within 27 provinces. Overall, the sampled households drawn from the three surveys across 40 districts and 35 cities in 27 provinces (out of 34 provinces).

For the final sampling frame, six survey areas of the DEHS which were overlapped with the survey areas in the UPS were dropped from the sampling frame. This was done in order to avoid potential bias later on when calculating the weights (detailed below). The UPS was chosen to be kept since it had much larger samples (2,016 households) than that of the DEHS (265 households).

Sample size

The HiFy survey was initially designed as a 5-round panel survey. By end of the fifth round, it is expected that the survey can maintain around 3,000 panel households. Based on the experience of phone-based, panel survey conducted previously in other study in Indonesia, the response rates were expected to be around 60 percent to 80 percent. However, learned from other similar surveys globally, response rates of phone-based survey, moreover phone-based panel survey, are generally below 50 percent. Meanwhile, in the case of the HiFy, information on some of households' phone numbers was from about 2 years prior the survey with a potential risk that the targeted respondents might not be contactable through that provided numbers (already inactive or the targeted respondents had changed their phone numbers). With these considerations, the estimated response rate of the first survey was set at 60 percent, while the response rates of the following rounds were expected to be 80 percent. Having these assumptions and target, the first round of the survey was expected to target 5,100 households, with 8,500 households in the lists. The actual sample of households in the first round was 4,338 households or 85 percent of the 5,100 target households. However, the response rates in the following rounds are higher than expected, making the sampled households successfully interviewed in Round 2 were 4,119 (95% of Round 1 samples), and in Rounds 3, 4, 5, 6, 7, and 8 were 4,067 (94%), 3,953 (91%), 3,686 (85%), 3,471 (80%), 3,435 (79%), 3,383 (78%) respectively. The number of balanced panel households up to Rounds 3, 4, 5, 6, 7, and 8 are 3,981 (92%), 3,794 (87%), 3,601 (83%), 3,320 (77%), 3,116 (72%), and 2,856 (66%) respectively.

Stratification

Two stratifications were applied in selecting sampled households, namely the explicit and implicit stratifications. In Indonesia, during the preparation of the survey, the number of positive cases of COVID-19 was mostly in DKI Jakarta province (about 50%) and in other cities in Java (about 30%). Also, the reported positive cases were mostly in urban areas. Considering the spread of the COVID-19 cases back then (about one month after the outbreak), the samples were first explicitly stratified into 5 strata (regions), namely DKI Jakarta, Java Non-DKI Jakarta Urban, Java Non-DKI Jakarta Rural, Outside Java Urban, and Outside Java Rural. Besides DKI Jakarta stratum which was the epicenter of COVID-19 epidemic in Indonesia and thus oversampled, samples in other strata were allocated proportionally to population of the respective strata/region. Besides the regional stratification as the explicit stratification, an implicit stratification in terms of gender and education level (i.e., junior secondary or lower, senior secondary, and tertiary) of the

head of households were applied in selecting sample of households, with the assumptions that the impacts of the pandemic might be different across those different households' demographic characteristics.

Sampling method

Three stages of sampling strategies were applied. For the first stage, districts (as primary sampling unit (PSU)) were selected based on probability proportional to size (PPS) systematic sampling in each stratum, with the probability of selection was proportional to the estimated number of households based on the National Household Survey of Socio-economic (SUSENAS) 2019 data. Prior to the selection, districts were sorted by provincial code. In the second stage, villages (as secondary sampling unit (SSU)) were selected systematically in each district, with probability of selection was proportional to the estimated number of households based on the Village Potential Census (PODES) 2018 data. Prior to the selection, villages were sorted by sub-district code. In the third stage, the number of households was selected systematically in each selected village. Prior to the selection, all households were sorted by implicit stratification, that is gender and education level of the head of households. If the primary selected households could not be contacted or refused to participate in the survey, these households were replaced by households from the same area where the non-response households were located and with the same gender and level of education of households' head, in order to maintain the same distribution and representativeness of sampled households as in the initial design.

In the Round 8 survey where we focused on early nutrition knowledge and early child development, we introduced an additional respondent who is the primary caregiver of under 5 years old in the household. We prioritized the mother as the target of caregiver respondents. In households with multiple caregivers, one is randomly selected. Furthermore, only the under 5 children who were taken care of by the selected respondent will be listed in the early child development module.

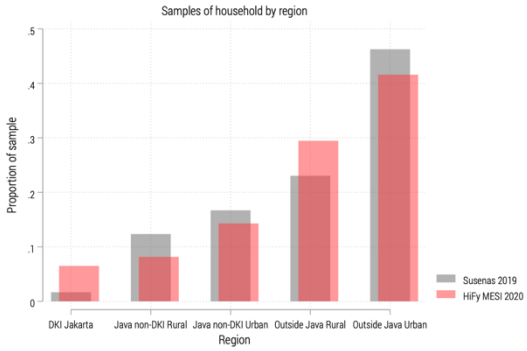
Weighting

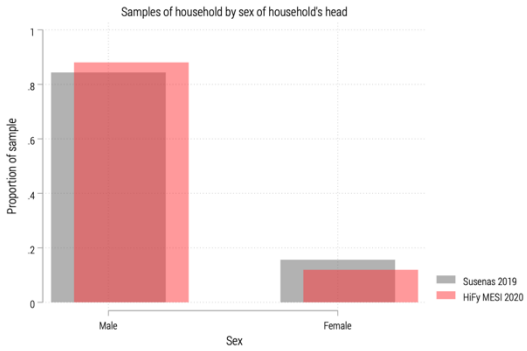
Since the sampling design of the three surveys was not the same, calculating the weights for all households combined is complicated. As a practical alternative, household weights were first calculated independently by each initial survey and then combined them all together afterward. For this approach to be properly applied without potential bias, there should not be overlapped survey areas across different surveys.

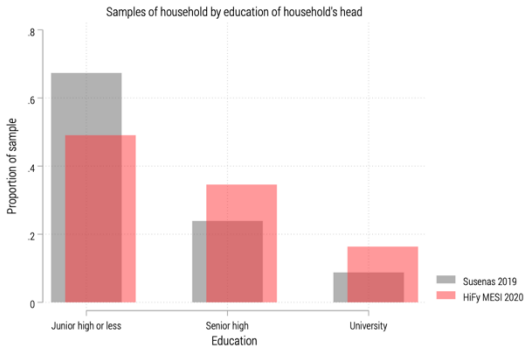
The household weights were calculated for both cross-section for each round and panel for all rounds of the survey. In each round of the survey, the initial sampling weight was calculated following the original sampling method of the survey from which the sampled households were drawn. A sampling weight trimming using the mean and standard deviation of the weights was then conducted to reduce weight variability. In particular, the weight trimming was applied to some outlier weights (only on small proportion of the samples), while keeping the total of the weights remain the same. Afterward, the weights were calibrated using a raking method to ensure the total estimates of the households with respect to designated variables were comparable with the population estimates of those variables from the SUSENAS 2019. The designated variables included region (DKI Jakarta, Java Non-DKI Jakarta Urban/Rural, Outside Java Urban/Rural), gender of household's head, and level of education of household's head (junior secondary and lower, senior secondary, and tertiary).

The comparisons of unweighted and weighted distributions between the HiFy and SUSENAS 2019 for the designated variables were presented below.

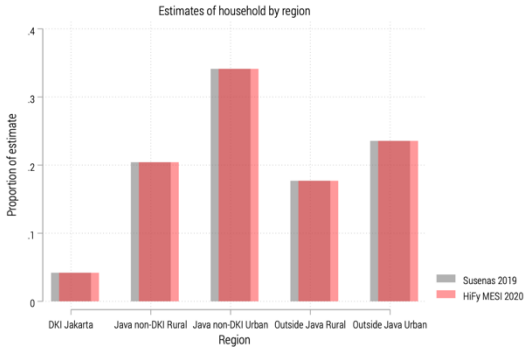
UNWEIGHTED

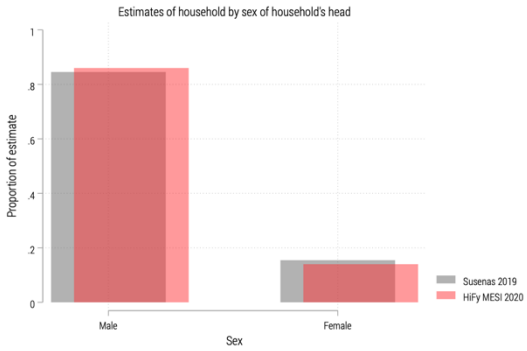


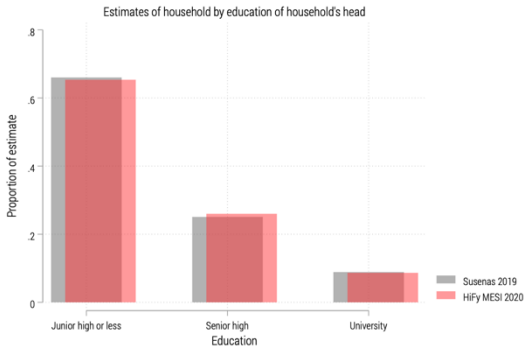




WEIGHTED







Meanwhile, the primary caregiver weights of round 8 were calculated by multiplying the household weights and the number of eligible caregivers in the selected households.

Attrition

The attrition occurred when respondents were not able to be interviewed, which was mostly because their phones were unreachable or unanswered. A test for whether attrition was random showed that the dropped households over rounds were generally random. The survey weight calculation has taken into account any

small associations between participation rate and certain key characteristics. Therefore, for analysis requiring panel households, attrition bias is not a concern when interpreting changes between rounds.

NOTE: Setting-up sampling parameter

Prior to using the HiFy data for an analysis, a sampling set-up needs to be done using the following commands to declare the survey design in STATA:

```
svyset psu [pweight=weight], strata(strata) fpc(N1h) vce(linearized)  singleunit(certainty) || ssu, fpc(N2hi) || tsu, fpc(Zhij)
```

where,
psu is the primary sampling unit (district)
ssu is the secondary sampling unit (village)
tsu is the tertiary sampling unit (household)
weight is the designated weight
strata is strata for PSU (region)
N1h is the finite population correction for PSU (number of districts in stratum-h)
N2hi is the finite population correction for SSU (number of villages in district-i)
Zhij is the finite population correction for TSU (number of households in village-j)

Meanwhile for the Round 8 module 11. Knowledge about Early Nutrition and 12. Early Child Development, a designated weight is prepared in hify_covid19_weight_R8_caregiver.dta. To declare the survey design in STATA, we apply the following command:

```
svyset psu [pweight=mot_wgt], strata(strata) fpc(N1h) vce(linearized) singleunit(certainty) || ssu, fpc(N2hi) || tsu, fpc(Zhij) || mot_id, fpc(n_elig)
```