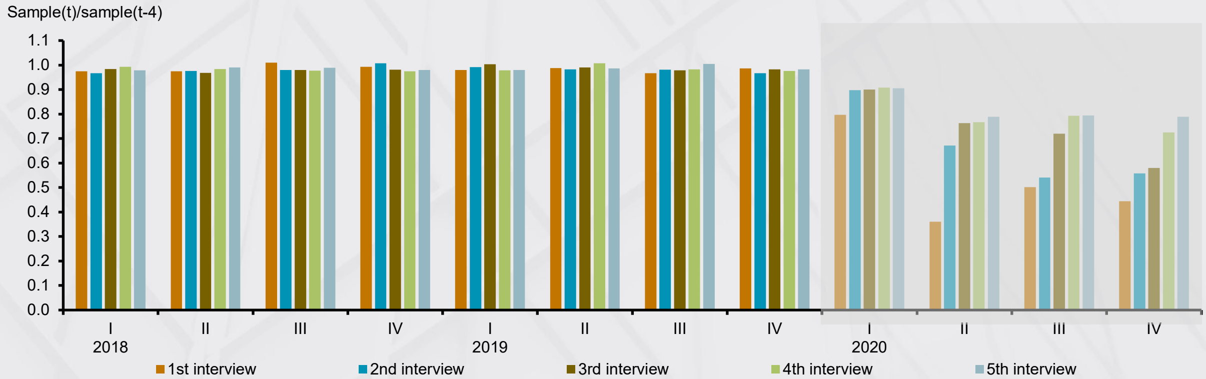
Continuous National Household Sample Survey estimates adjusted by the response rate reduction during the pandemic

The Continuous National Household Sample Survey (Continuous PNAD) has been facing challenges inherent to the change in data collection, from face-to-face to by telephone, due to the pandemic.[1] The response rate reduction is one of the greatest concerns, especially in the groups of individuals first interviewed[2] as of 2020Q2. This fact, mentioned by labor market analysts and by the Brazilian Institute of Geography and Statistics (IBGE)[3] itself, is illustrated in Figure 1.

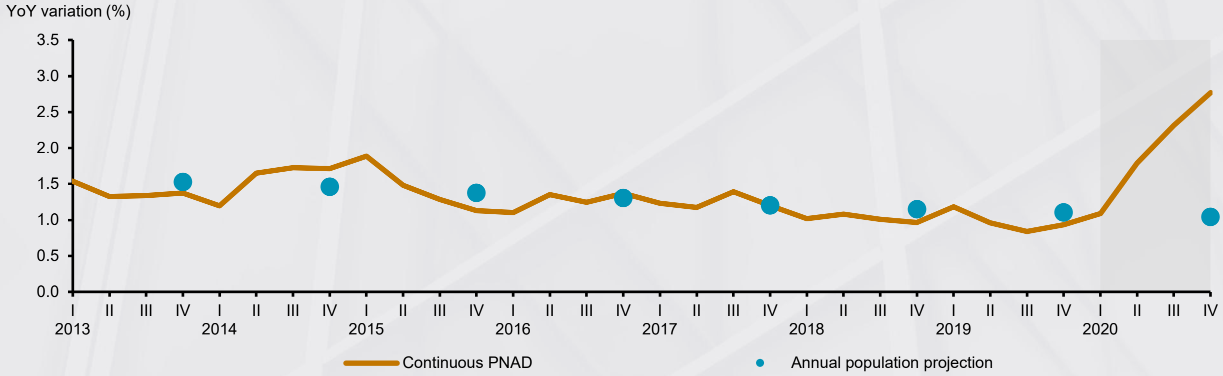
Figure 1 – Sample size variation by interview group



Source: IBGE

This box analyses the effects of the Continuous PNAD response rate reduction in the survey's leading economic indicators. A key motivation is the acceleration of the working age population[4] (WAP) during the pandemic, detached from IBGE population projections (Figure 2). In addition, one observes an increase of the population aged 40 years or older and a sharp decrease in the population aged less than 40 years (Figure 3).

Figure 2 – Working age population growth



Source: IBGE

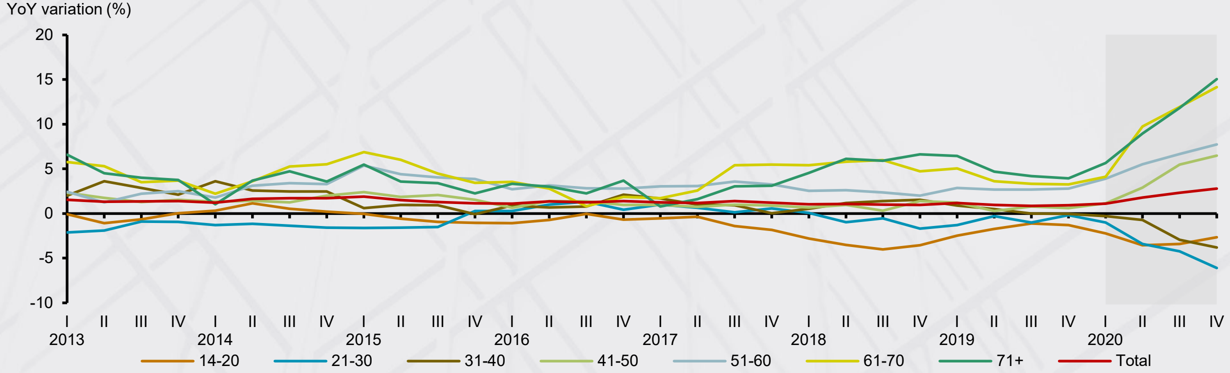
[1] Similar challenges are being faced worldwide. Further details on this subject may be seen in the International Labor Organization Report (ILO, 2020).

[2] The survey's sample rotation scheme consists in monitoring respondents during five consecutive quarters by means of interview groups (1 to 5).

[3] Response rate reduction is also referred to as a reduction in the number of interviews in each sample. See, for example, the Technical Notes 8/2020 ("Informações referentes à divulgação dos dados do 2º trimestre de 2020") and 2/2021 ("Sobre o processo de ponderação da PNAD Contínua") from Continuous PNAD/IBGE and the *Carta de Conjuntura do Instituto de Pesquisa Econômica Aplicada (IPEA) Number 50 – Nota de Conjuntura 22 – 1º trimestre de 2021* ("A redução no número de entrevistas na PNAD Contínua durante a pandemia e sua influência para a evolução do emprego formal").

[4] Population aged 14 years or older.

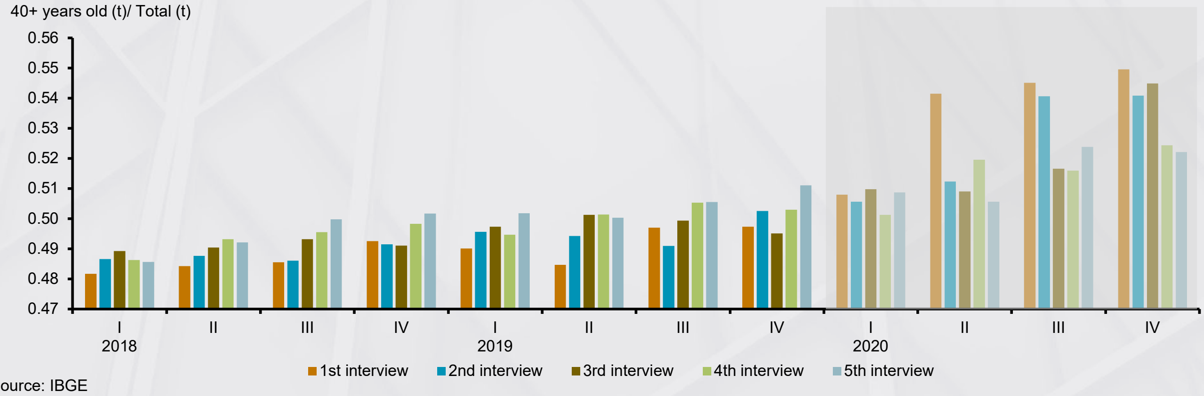
Figure 3 – Working age population growth by age group



Source:IBGE

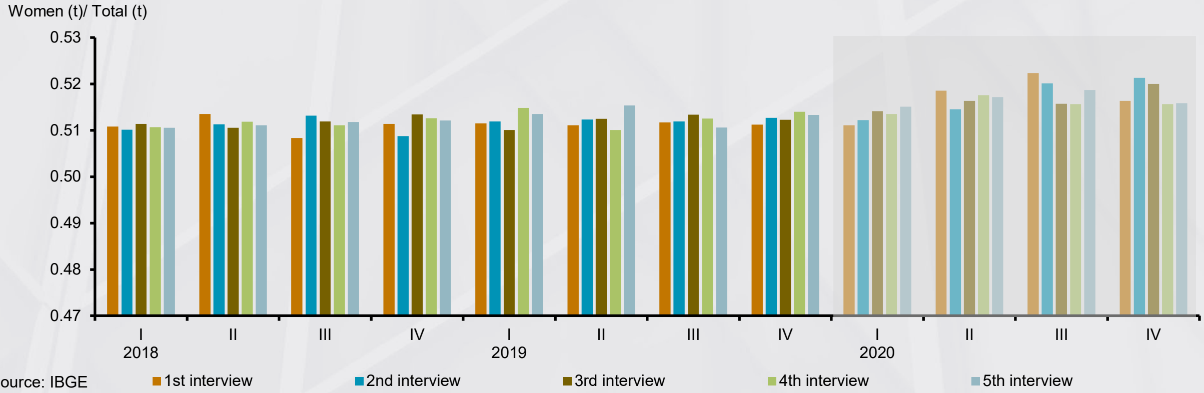
The structure of sex and age of respondents in household samples is usually different from that of population, a phenomenon known in the literature as "availability bias". Particularly regarding the Continuous PNAD, from one quarter to another, greater availability is observed for women and older individuals –they are more frequently found. Thus, in the same quarter, in general, the samples with already interviewed individuals tend to include a greater number of women and, especially, older individuals. According to Figures 4 and 5, the availability bias –especially that of older individuals (aged 40 years or older)– clearly increased as of 2020Q2 in all interview groups, but particularly in those groups started when the pandemic had already begun, for which the first contact occurred by telephone[5]. The sudden change in the proportion of women and older individuals may reflect a selection bias related, for example, to the access to telephone numbers or the propensity to answer the IBGE's contact attempts.

Figure 4 – Sample share of people older than 40 years old by group of interview



Source: IBGE

Figure 5 – Sample share of women by group of interview



Source: IBGE

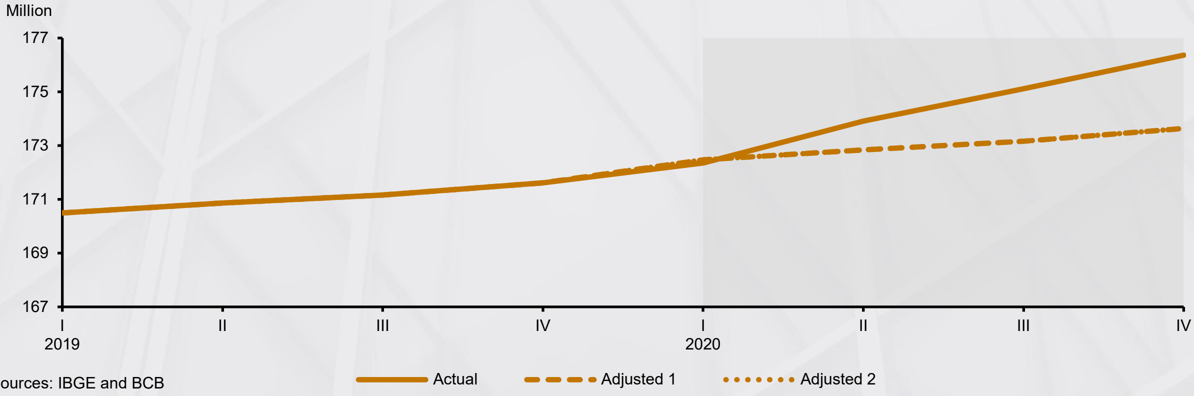
5/ That is, in the interviewed group 1 in 2020Q2, in the interviewed groups 1 and 2 in 2020Q3, and the interviewed groups 1 to 3 in 2020Q4.

The Continuous PNAD assumes a random sample loss and redistributes sample weights uniformly among individuals of the same primary sampling unit. The employment of this assumption in the context of an increased availability bias may explain the acceleration of WAP growth, since the increase in the proportion of older individuals would be interpreted as a demographic change inherent to the population itself.

The clear impact of the change in the way of collecting information on demographic statistics raises interest in evaluating the extent by which the same impact may occur in the estimates of labor market variables, such as employed population, unemployment rate, and participation rate. Two related empirical strategies are applied with the goal of evaluating this issue, using quarterly Continuous PNAD microdata. Both strategies aim at mitigating the availability bias associated with the response rate reduction by using the 2019 samples and the IBGE's population projection for 2020 as benchmarks for adjusting the survey's estimates.

In the first approach, all interview groups started in 2020 are excluded. Next, the remaining observations are classified by sex, age bracket[6] and federation unit (UF). The reason for this exclusion is based on two assumptions: i) interview groups started in 2020 increase the survey's availability bias; and ii) individuals in the interview groups started in 2019 are more similar to those of the 2020 quarterly samples should the response rate had not been reduced. Next, the working age population is calibrated by means of the ratio between the WAP of the IBGE population projection and the WAP obtained by means of the aforementioned exclusion, so as to keep the total population by groups of sex x age bracket x UF in accordance with the IBGE's projection.[7]

Figure 6 – Working age population



Sources: IBGE and BCB

[6] Five age brackets were defined: 14-24 years, 25-39 years, 40-54 years, 55-69 years and 70 years or older. The result of estimates is robust to alternative definitions of age brackets.

[7] The simulated estimate of some variable s_2 of the labor market has the following formal definition:

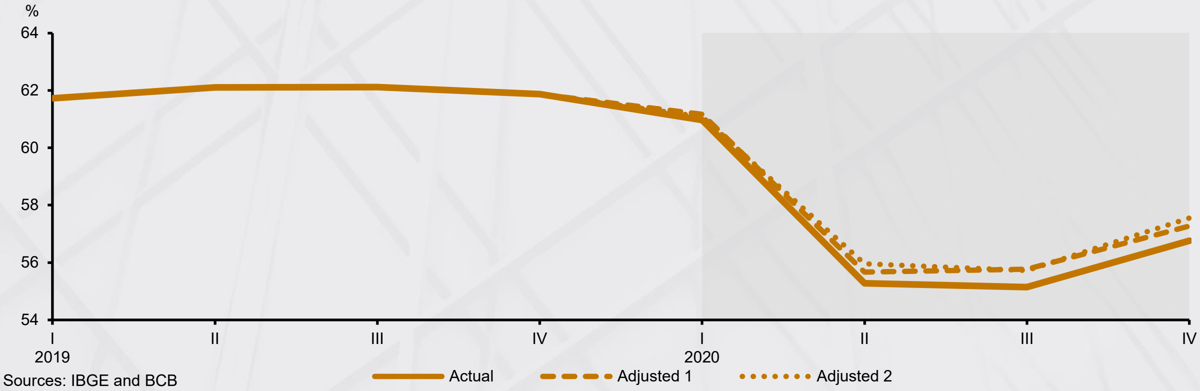
$$X_t^{s_1} = \sum_{G(g)} \sum_{A(a)} \sum_{U(u)} X_{g,a,u,t}^{s_1} \left(\frac{\mathrm{PIT}_{g,a,u,t}^{IBGE}}{\mathrm{PIT}_{g,a,u,t}^{s_1}} \right), \qquad (1)$$

where $X_{g,a,u,t}^{s_1} = \sum_{i \in F(g,a,u,E(t))} w_{i,t} \mathbb{I}(X)$.

That is, the total of the simulated variable *X* (simulation 1: s_1) of sex *g*, age bracket *a*, UF *u*, in the quarter *t* is equal to the sum of the sample weighting (*w*) of all individuals *i* in the condition of variable *X* (identified by the indicator function $\mathbb{I}(X)$) and included in the set *F*, which depends on the sex, age bracket, UF, and interview groups $E(t)$, which, in turn, will depend on the 2020 quarter: $E(t) = \sum_{s=t+1}^{5} e_s$, in which *e* refers to the interview group. Therefore, in I 2020 (*t* = 1) only the groups of individuals in the second and fifth interviews are considered, in II 2020, the groups of the third and fifth interviews and so on until IV 2020, which considers only the group of the fifth interview (begun –was the group of the first interview– in IV 2019). Thus, all interview groups of 2020 begun in 2019. It is also noteworthy that the definition of $\mathrm{PIT}_{g,a,u,t}^{s_1}$ is the same of $X_{g,au,t}^{s_1}$ and that $\mathrm{PIT}_{g,a,u,t}^{IBGE}$ is the WAP of the quarter *t* of 2020 that uses the annual growth rates of the IBGE's population projection by sex, age bracket, and UF uniformly applied to the WAPs of the previous four quarters in 2019. The total aggregate of the simulated variable *X*, $X_t^{s_1}$, sums $X_{g,au,t}^{s_1}$ multiplied by the ratio of PITs and UF ($U(u)$), age bracket ($A(a)$) and sex ($G(g)$).

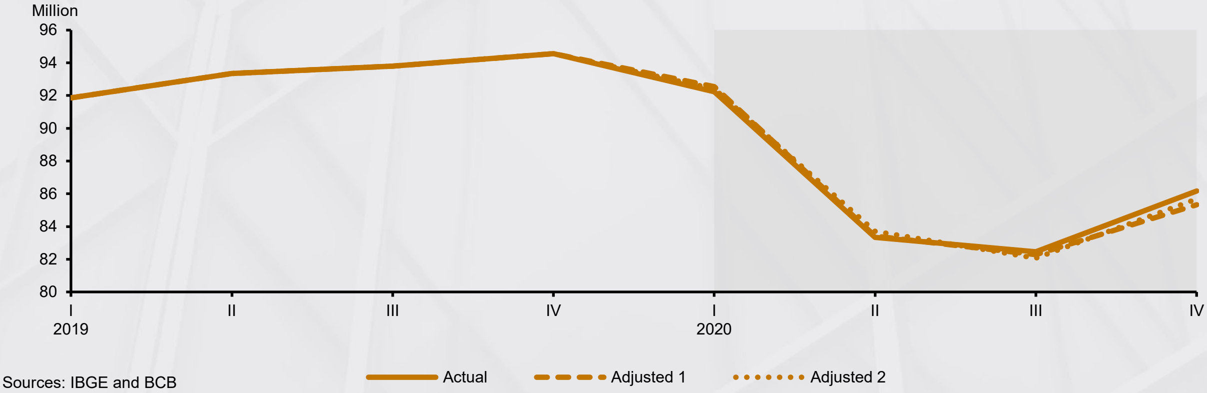
The second approach directly adjusts the sample weights –similarly to the one suggested and still being implemented by IBGE in the Technical Note 2/2021 of the Continuous PNAD– as the procedure for correcting survey's potential biases. First, it applies a logistic regression with several demographic and geographic covariates for each quarter of 2020 and for each interview group, with regard to the same interview groups in 2019 quarters, to estimate propensity scores for individual responses to interviews of the Continuous PNAD. Then, the propensity score adjusts the individual sample weighs in each quarter of 2020. Intuitively, for each interview group, the aim of this adjustment by propensity score is to give more (less) weight to individuals in each quarter of 2020 that are more (less) akin to individuals in the 2019 quarterly samples according to the characteristics defined by the covariates in the logistic regression. Next, the total population is calibrated as in the first approach.[8]

Figure 7 – Participation rate



Sources: IBGE and BCB

Figure 8 – Employed population



Sources: IBGE and BCB

As expected, the result of both empirical implementations are similar in general. In the first approach the samples of interview groups started in 2020 are excluded, remaining only samples of interview groups started in 2019. In the second approach, the demographic structure of 2020 quarterly samples is approximated by the demographic structure observed in 2019. Thus, both are attempts to mitigate the selection bias in observable (demographic) variables in 2020.

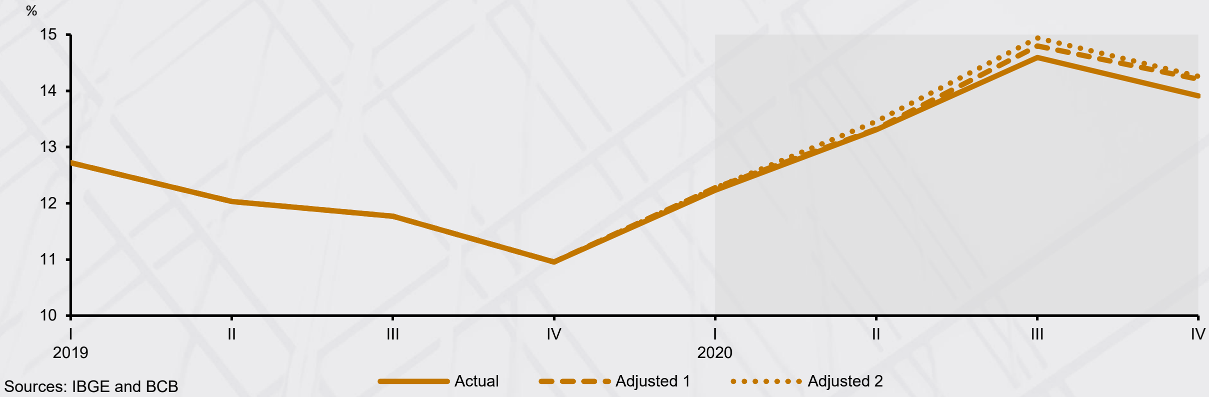
8/ / The formal definition of simulation 2 (^{S_2}) of some variable of the labor market follows:

$$X_t^{S_2} = \sum_{G(g)} \sum_{A(a)} \sum_{U(u)} X_{g,au,t}^{S_2} \left(\frac{\mathrm{PIT}_{g,a,u,t}^{IBGE}}{\mathrm{PIT}_{g,a,u,t}^{S_2}} \right), \qquad (2)$$

where $X_{g,au,t}^{S_2} = \sum_{i \in F(g,a,u)} \hat{\theta}_{i,t|2019} w_{i,t} \mathbb{I}(X);$

such as $\hat{\theta}_{i,t|2019}$ is the estimate of the response propensity of an individual *i* in the quarter *t* of 2020, given the demographic structure of 2019, according to a set of covariates of a logistic regression. It is defined as $logit\left(\theta_{i,e,t|2019}\left(y_{i,e,t} = 1 \middle| \boldsymbol{x}_{i,e,t}\right)\right) = \boldsymbol{x}_{i,e,t}^T \boldsymbol{\beta}_{e,t}$, so that in the implementation $y_{i,e,t} = 1$ if the individual is in some quarter of 2019 in the interviewed group *e*, and $y_{i,e,t} = 0$ if the individual is in the quarter *t* of 2020 and in the interviewed group *e*. *x* is a vector of covariates (dummies for metropolitan region, attend school, educational ranges, domiciliary condition, age bracket, great region, sex, rural, kind of domicile, groups by number of individuals aged at least 14 years and UFs) and β is the vector of coefficients. The choice of covariates is based on Almeida et al. (2019).

Figure 9 – Unemployment rate



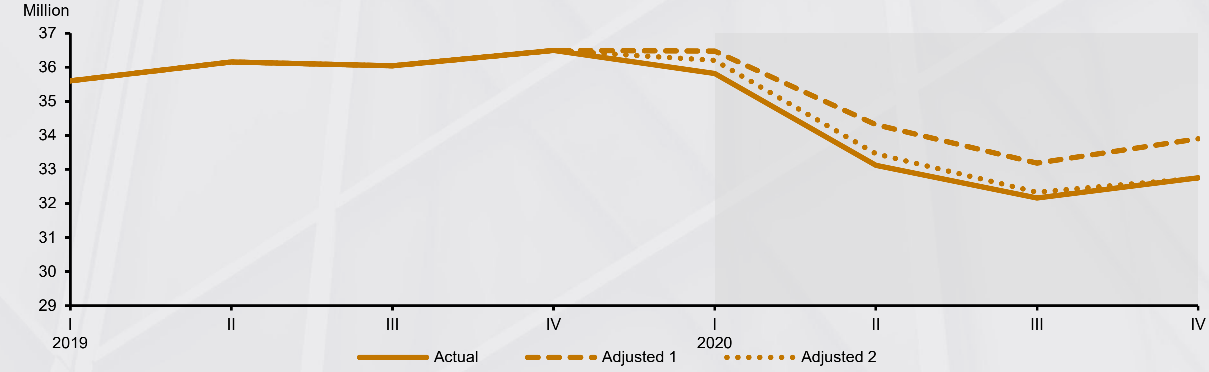
Sources: IBGE and BCB

Results shown in Figures 6 to 9 might seem surprising, at first glance, since an adjustment in the WAP of nearly 3.8 million people throughout 2020 implies small changes in labor market economic variables. The new estimates are likely within the confidence interval for the Continuous PNAD domain estimates. For example, the adjustments suggest that the estimate reported by the IBGE overestimated the employed population by nearly 0.5 million people in 2020Q4. The unemployment rate would have been nearly 0.4 p.p. higher than that reported (13.9%) and the participation rate would have reduced less, nearly 0.8 p.p. (57,6% against 56,8%), both in 2020Q3.

The adjustment results suggest that a possible selection on observable demographic variables –which is associated with the increased availability bias due to the sample reduction caused by the change to phone call data collection during the pandemic– does not seem to have led to significant changes in key aggregate estimates linked to employment and labor market participation in the Continuous PNAD. This does not exclude, however, the possibility that these data collection problems have implied biases associated with selection on *ex-ante* unobservable variables by the survey.[9]

One example that might suggest this type of bias is illustrated in Figures 10 and 11. Figure 10 shows the formal employed population[10] and indicates that –according to the first empirical approach of the adjustment, which excludes the interview groups started in 2020 from the sample– the decline in formal employment in 2020 would have been 0.7 and 1.2 million lower than the officially reported. By excluding the interview groups started in 2020 from the sample, the first approach is able to mitigate the selection bias in both observables (demographic) and unobservables. The second empirical adjustment strategy, which is strictly demographic and adjusts sample weighs by propensity scores, indicates a fall of formal

Figure 10 – Population in formal occupation



Sources: IBGE and BCB

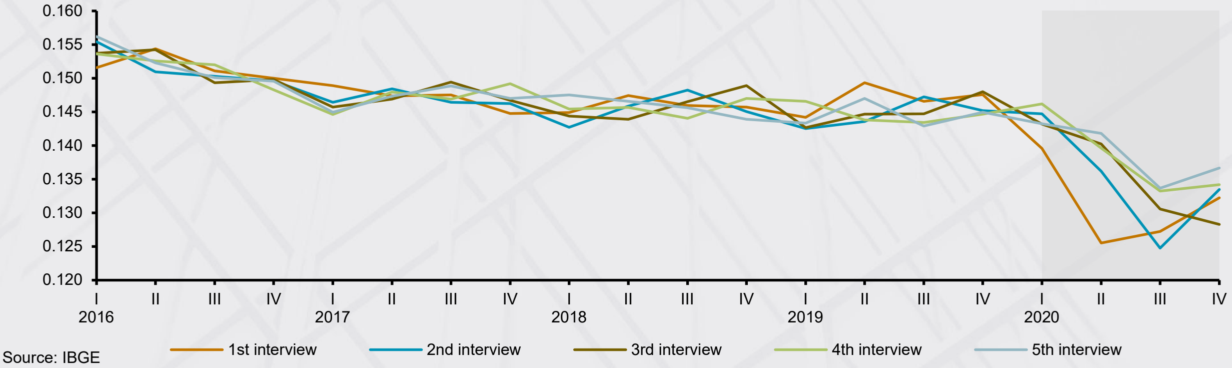
[9] Selection on ex-ante unobservable variables indicates that the survey may be involuntarily selecting individuals in specific conditions within or outside the workforce, for example. On the other hand, selection on observable variables means that the survey selects individuals according to known characteristics (observables) in the sample design, as, for example, age bracket and sex.

[10] Formal jobs are defined as private and public sector registered employees and non-registered public sector employees that contribute to the social security. Excludes household workers, statutory and military.

employment closer to the originally reported. Figure 11 illustrates how the proportion of individuals with formal occupation varied differently among interview groups after the pandemic.[11]

Figure 11 – Sample share of people in formal occupation by group of interview

People in formal occupation (t)/Total (t), (%)



Source: IBGE

Summing up, this box investigates the impacts of the drop of Continuous PNAD's response rate on labor market indicators. The results of two distinct adjustment methods indicate that the likely selection bias in observable demographic variables does not seem to imply relevant changes in aggregate estimates of employment and participation in the labor market. However, these methods are not able to fully correct selection on unobservable variables bias.[12] If this type of bias indeed exists, monitoring the labor market outlook during the pandemic is even more challenging, not only to analysts, but also to institutions that carry out household sample surveys around the world.

References:

Almeida, P. A. *et al*. Pesos longitudinais para a Pesquisa Nacional por Amostra de Domicílios contínua (PNAD contínua). **Boletim Mercado de Trabalho**: Conjuntura e Análise, n. 67, out. 2019 Available at: <https://www.ipea.gov.br/portal/images/stories/PDFs/mercadodetrabalho/191101_bmt_67_nt_pesos_longitudinais.pdf>.

Corseuil, C. H, Russo, F. A redução no número de entrevistas na PNAD Contínua durante a pandemia e sua influência para a evolução do emprego formal. **Carta de Conjuntura**: Mercado de Trabalho, n. 50, Nota de conjuntura 22, 1º trimestre de 2021. Available at: <https://www.ipea.gov.br/portal/images/stories/PDFs/conjuntura/210318_cc_50_nota_22_amostra_da_pnad_continua.pdf>.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua)**: informações referentes à coleta do mês de abril de 2020. Rio de Janeiro: IBGE, 2020. (Nota Técnica). Available at: <https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Nota_Tecnica/Nota_Tecnica_Divulgacao_2Tri2020_Agosto_2020.pdf >.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua)**: Sobre o processo de ponderação da PNAD Contínua. Rio de Janeiro: IBGE, 2021. (Nota Técnica). Available at: <https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Nota_Tecnica/Nota_Tecnica_02_2021_Sobre_o_processo_de_ponderacao.pdf >.

[11] It is noteworthy that this aggregate behavior of the formal employment in the PNAD Continua sample in Figure 11 is the same for age brackets and sex. Thus, this does not represent the composition effect of these demographic groups.

[12] The first adjustment method minimizes the selection bias for unobservable variables. However, it cannot be employed as of 2021Q1, since all individuals in the sample will have responded the survey only by telephone.

ILO – INTERNATIONAL LABOUR ORGANIZATION. Ilostat: COVID-19 impact on the collection of labour market statistics. Geneva: ILO, 2020. Available at: < https://ilostat.ilo.org/topics/covid-19/covid-19-impact-on-labour-market-statistics/>