**Description of the sampling plan for Building the Barricades' quantitative survey**

**Presentation**

This document is a description of the sampling plan for Building the Barricades' quantitative survey carried out between July and December 2019, in the favelas of Maré, in Rio de Janeiro, Brazil.

According to the Maré Population Census (Redes da Maré, 2019), Maré's 16 favelas were home to 139,073 residents in 2013, including 97,878 aged 18 or over. In the present study, 1,211 people were interviewed, randomly selected from residents aged over 18.

**Research population**

The research population corresponds to the group of residents aged over 18 living in households in Maré's 16 favelas. The Maré Population Census address register used for selecting the sample has 47,776 addresses.

**Geographical stratification of the research population**

The research population was stratified into three geographical strata, composed of groups of Maré's favelas, determined according to the favelas' location, housing characteristics and social dynamics, in addition to road access and shared public facilities.

Geographical strata 1, 2 and 3 include, respectively, four, nine and three Maré favelas, as shown in Table 1, below.

| Table 1 – Maré's favelas included in each geographical stratum | |
| --- | --- |
| Geographical strata | The favelas which form each stratum |
| Area 1 | Nova Holanda, Parque Maré, Parque Rubens Vaz and Parque União |
| Area 2 | Baixa do Sapateiro, Conjunto Bento Ribeiro Dantas, Conjunto Esperança,  Conjunto Pinheiros, Morro do Timbau, Nova Maré, Salsa and Merengue, Vila do João, Vila dos Pinheiros |
| Area 3 | Marcílio Dias, Parque Roquete Pinto, Praia de Ramos |

The distribution of the research population according to age group and sex within each of the sample's geographical stratum is shown in Table 2 below.

| Table 2 – Distribution of the population aged over 18, according to age group and sex, in each geographical stratum | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Geographical stratum and sex | Total | Age group | | | | |
| 18 to 29 | 30 to 49 | 50 to 65 | 66 or above | Age unknown |
| Area 1 | 38,313 | 12,749 | 17,048 | 5,845 | 2,287 | 384 |
| Male | 18,648 | 6,270 | 8,517 | 2,786 | 891 | 184 |
| Female | 19,588 | 6,472 | 8,521 | 3,057 | 1,395 | 143 |
| Unknown | 77 | 7 | 10 | 2 | 1 | 57 |
| Area 2 | 47,415 | 14,744 | 21,412 | 7,890 | 2,795 | 574 |
| Male | 22,843 | 7,211 | 10,578 | 3,660 | 1,138 | 256 |
| Female | 24,470 | 7,518 | 10,825 | 4,221 | 1,657 | 249 |
| Unknown | 102 | 15 | 9 | 9 | – | 69 |
| Area 3 | 12,150 | 3,433 | 5,743 | 2,154 | 746 | 74 |
| Male | 5,833 | 1,683 | 2,803 | 1,032 | 288 | 27 |
| Female | 6,295 | 1,746 | 2,939 | 1,122 | 458 | 30 |
| Unknown | 22 | 4 | 1 | – | – | 17 |
| Source REDES DA MARÉ. Population Census (2013). Rio de Janeiro: Redes da Maré, 2019. | | | | | | |

**Stages of sample selection in the sampling plan**

The sampling plan used in the survey has two stages of stratification and selection: private households and adult resident (aged 18 or older).

In the first stage, households were selected by inverse sampling (Haldane, 1945; Vasconcellos *et al*., 2005; Vasconcellos *et al*., 2013).

In the second stage, an adult resident was selected with equiprobability among the adult residents of the selected household.

Inverse sampling is a sequential sampling procedure, widely used in systems that carry out blood counts, and has the advantage of eliminating, or at least greatly mitigating, the non responses observed in classic household surveys.

In classic household sampling, a fixed number of households to be interviewed by stratum or sample unit is determined, and they are selected using a random sampling method. Then, the selected households are visited and the responses and non responses (refusals, households which are vacant, used occasionally, demolished etc.) are observed; non responses cause a loss in the sample, which is different in each stratum or sample unit. To deal with this problem, three alternatives can be employed:

(1) substituting non responding households, which always generates selection biases, regardless of the substitution rule used;

(2) increasing the sample size by applying an expected non response rate, which is not the most ideal because the non response rate is different according to stratum and causes waste when the sample ends up larger than expected, or it does not solve the problem, when the non response rate is higher than expected; or

(3) the use of oversampling, which is difficult to apply and usually increases the cost of data collection, since compensating for losses usually occurs after the data collection work has finished, making it necessary to return to the field to complete the sample.

Inverse sampling alters this paradigm by changing the sample size to a fixed number of successes (completed interviews), regardless of the number of households visited to obtain the interview. Its form of randomization is done by defining a random order of addresses to visit.

Thus, in inverse sampling, the interviewer receives the list of addresses and visits them sequentially until they reach the number determined or they exhaust the survey area (stratum or sample unit).

The method was originally described by Haldane (1945) for counting blood elements and transferred to use in household sampling by Vasconcellos and colleagues in 2005.

**Sample size**

Given that there is no information about the prevalence of the main variables studied in the research population, the sampling plan started from another strategy. A minimum proportion of 3% (Pmin = 0.03) was chosen, for which the relative standard error of the estimate should be no greater than 60% (dR = 60%), which corresponds to a relative margin of error of 1.8%, with a 95% confidence level (1-α = 0.95).

According to Cochran (1977) and assuming simple random sampling without replacement (AAS), the sample size necessary to estimate proportions equal to or greater than Pmin with relative standard error not exceeding dR at the 1-α confidence level is given by:

(1)

With fixed parameters, using the equation in (1) leads to a sample size of 346 adults (or households, since, in the present study, only one adult was selected in the second sample stage) in each geographical stratum.

As each geographical stratum constitutes a controlled estimate domain, the same sample size was maintained for each one, ensuring the estimate of proportions of 3% with a relative standard error of 60% at a 95% confidence level in each stratum.

However, as a set of questions in the survey referred to people up to 65 years old and it was not considered operational to set a maximum age for interviewees, the sample size was increased to 400 adults in each geographical stratum, leading to a final sample size of 1,200 households (or adults), adding together the samples from the three strata.

This small increase in sample size leads to a small reduction in the margin of error. It should also be noted that the results for Maré as a whole have a lower margin of error than for each particular stratum. However, what is most relevant is to determine the sample's calculation parameters, because, strictly speaking, the sampling errors should be estimated for each variable from the data collected.

**The probabilistic sampling scheme**

The probability of household i from geographical stratum h being included in the sample, represented by, is given by:

, where, (2)

th is the total number of households in stratum h;

vh is the total number of households visited in stratum h;

ah is the total number of households with adults (eligible households) in stratum h; and

eh is the total number of households interviewed in stratum h.

Note that the three fractions above indicate, respectively, the probability that the household will be visited, the probability that the household will be eligible and the probability that the household will be interviewed. The “- 1” in the third fraction corresponds to the stopping rule of the sequential procedure, in other words, the loss of a degree of freedom because the last eligible household is also the last one interviewed.

The probability that the adult Ahij will be included in the sample, given that the household dhi was selected, represented by, is given by:

, where, (3)

nahi is the number of adults in household hi from stratum h.

Thus the probability of inclusion of any adult, represented by , is given in the equation below:

(4)

The design's natural weight corresponds to the inverse of the equation shown in (4), with the observation that the second fraction must be eliminated to avoid expanding the number of eligible households for the total population. Thus, the natural weight of the design for any adult, represented by is given in (5)

(5)

**Calibration of sampling weights**

In household surveys, the sample represents the population of households and this always causes bias in population estimates for sex and age groups, due to the diversity of resident numbers in each household, their ages and sex.

In addition, all the probabilities of inclusion in the sample depend on an outdated register. In this research, the probabilities of inclusion reflect the population of Maré in 2013, therefore six years before the time of the survey.

Thus, international literature recommends calibrating the sample weights so that they can be adjusted to independent estimates of the population by sex and age, whose date of reference is close to the research period (Silva, 2004).

In this case, a post-stratification estimator was used to calibrate sampling weights, such that the calibrated weight of adult j from household i of stratum h, represented by, is equal to their design weight multiplied by a calibration factor, defined by the post-stratum p, represented by. Thus, the calibrated weight is given by:

(6)

, where, (7)

is the value of the independent estimate of the number of people in the post-stratum p; and

is the total estimated number of people in post-stratum p.

The eight post-strata are defined by crossing sex with age group (18 to 29; 30 to 49; 50 to 65; 66 and above). The data for the independent estimate Maré's population is presented in Table 3.

As the most recent census data available is from 2013 and the sample was collected in 2019, the population size of each stratum was estimated by sex and age on July 1, 2019.

Based on the population estimates (1) for Brazil's Federal Units and simple age, produced by the Brazilian Institute of Geography and Statistics, IBGE, for 2010 (IBGE 2010 Demographic Census), 2013 (the year of Maré's Census) and 2019 (time of research), the linear trend method was used (Madeira and Simões, 1972) to estimate the population of Maré for July 1, 2019, shown in Table 3.

(1) Available at: [ftp://ftp.ibge.gov.br/Projecao\_da\_Populacao/Projecao\_da\_Populacao\_2018 /projecoes\_2018\_populacao\_idade\_simples\_2010\_2060.xls](ftp://ftp.ibge.gov.br/Projecao_da_Populacao/Projecao_da_Populacao_2018/projecoes_2018_populacao_idade_simples_2010_2060.xls)

| Table 3 – Estimate of Maré's population aged over 18 for July 1st 2019, by sex and age group | | | |
| --- | --- | --- | --- |
| Age Group | Total | Men | Women |
| Total | 101,549 | 49,435 | 52,114 |
| 18 to 29 | 30,603 | 15,186 | 15,417 |
| 30 to 49 | 45,911 | 22,911 | 23,000 |
| 50 to 65 | 17,750 | 8,418 | 9,332 |
| 66 or above | 7,285 | 2,920 | 4,365 |

**Operational procedures and actual sample data**

Two operational procedures should be detailed in this section: (1) the procedure for randomizing the order of visiting the households; and (2) the procedure for selecting the adult to be interviewed in the household.

The procedure for randomizing the addresses was done by first placing all the addresses of each stratum in alphabetical order by favela, street, street number and complement. For each address, a uniform random number was generated in the interval (0;1). Then, the addresses of each stratum were ordered in ascending order of these randomly generated numbers. In a more scientific way, a random permutation of the addresses of each stratum was generated.

Each interviewer received a small set of addresses to visit and, when completing work on them, received a new batch of addresses. Using this procedure, it is possible to ensure that the interviewers respect the random order of the list.

Table 4 shows the number of addresses in Maré, how many addresses were given to the interviewers and how many were actually interviewed.

| Table 4 – Number of addresses available in the Maré Census, number of addresses supplied to the interviewers and the number of households actually interviewed | | | |
| --- | --- | --- | --- |
| Geographical strata | Total number of addresses | Addresses given to the interviewers | Households interviewed |
| Total | 47,776 | 3,136 | 1,211 |
| Area 1 | 19,231 | 1,130 | 406 |
| Area 2 | 22,336 | 1,054 | 400 |
| Area 3 | 6,209 | 952 | 405 |

It can be seen in Table 4, that the addresses for all buildings entered the sampling procedure, in other words, they had a non-zero probability of being selected. Indeed, there was no prior separation of addresses due to their purpose in 2013, considering that their use could have changed between 2013 and 2019, when the survey was conducted.

The procedure for selecting the adult to be interviewed in the selected household was based on the selection of simple random samples for each possible household size. If the household had a single adult resident, they were selected. If there were two adults, a sample of size 1 was selected from numbers 1 and 2, that is, sometimes the number 1 was selected, sometimes the number 2. If the household had three residents, a sample of size 1 was selected from numbers 1, 2, 3. This procedure was repeated for up to 15 residents. Thus, a selection table with 48,000 rows and 16 columns was designed, as shown in Table 5.

In order to select the household resident, the first step is to order the ages of residents over 18 years old (adults), from the oldest to the youngest. After recording this, the line corresponding to the household is consulted. If the interviewer is in the first household, the first line is consulted; if in the second household, the second line, and so on. On the consulted line, the column corresponding to the household's number of adult residents is looked at - if, for example, there are five adult residents, column 5. Thus, the intersection of line x column indicates the selected resident's number in the order recorded. If the number indicated is number 2, the selected resident is the second oldest.

| Table 5 – Table for selecting the adult to be interviewed | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Selected household's nº in the order list | Number of adult residents in the household | | | | | | | | | | | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 1 | 1 | 1 | 2 | 1 | 2 | 3 | 4 | 2 | 3 | 8 | 11 | 8 | 8 | 10 | 9 |
| 2 | 1 | 1 | 1 | 3 | 3 | 5 | 7 | 2 | 1 | 8 | 1 | 3 | 5 | 4 | 15 |
| 3 | 1 | 2 | 1 | 2 | 1 | 2 | 7 | 2 | 2 | 3 | 4 | 3 | 2 | 3 | 5 |
| 4 | 1 | 2 | 1 | 3 | 3 | 5 | 1 | 5 | 2 | 3 | 10 | 10 | 6 | 6 | 10 |
| 5 | 1 | 2 | 3 | 2 | 5 | 4 | 4 | 7 | 5 | 2 | 4 | 12 | 11 | 3 | 2 |
| 6 | 1 | 1 | 2 | 4 | 4 | 5 | 2 | 5 | 5 | 9 | 2 | 12 | 10 | 6 | 2 |
| 7 | 1 | 1 | 2 | 4 | 3 | 2 | 6 | 6 | 8 | 10 | 1 | 9 | 9 | 6 | 2 |
| 8 | 1 | 2 | 3 | 1 | 4 | 3 | 6 | 1 | 6 | 2 | 3 | 6 | 4 | 11 | 1 |
| 9 | 1 | 1 | 2 | 3 | 2 | 6 | 1 | 2 | 2 | 3 | 3 | 10 | 10 | 1 | 14 |
| 10 | 1 | 1 | 2 | 1 | 4 | 1 | 3 | 5 | 2 | 9 | 6 | 3 | 6 | 10 | 11 |
| 11 | 1 | 1 | 3 | 2 | 2 | 1 | 3 | 8 | 8 | 3 | 8 | 10 | 3 | 4 | 2 |
| 12 | 1 | 1 | 2 | 2 | 5 | 1 | 5 | 5 | 6 | 1 | 3 | 2 | 1 | 9 | 5 |
| 13 | 1 | 2 | 3 | 1 | 4 | 3 | 1 | 1 | 2 | 3 | 7 | 4 | 13 | 2 | 13 |
| 14 | 1 | 1 | 2 | 3 | 2 | 3 | 1 | 8 | 4 | 10 | 8 | 4 | 6 | 8 | 1 |
| **. . .** | | | | | | | | | | | | | | | |

The search for households to be interviewed, through the sequential procedure of inverse sampling, gave the results shown in Table 6.

| Table 6 – Results from the household visits | | | | |
| --- | --- | --- | --- | --- |
| Results from the household visits | Total | Area 1 | Area 2 | Area 3 |
| Total | 3,136 | 1,130 | 1,054 | 952 |
| Interviews carried out | 1,211 | 406 | 400 | 405 |
| Selected person absent on all visits | 201 | 52 | 84 | 65 |
| Selected person absent (traveling, in hospital etc.) | 12 | 4 | 3 | 5 |
| Selected person declined interview | 32 | 9 | 16 | 7 |
| No adult in a condition to be interviewed | 14 | 5 | 5 | 4 |
| Household refused to participate | 169 | 66 | 60 | 43 |
| Household closed or inaccessible | 1.048 | 414 | 330 | 304 |
| Household not located or non-existent (\*) | 352 | 130 | 124 | 98 |
| Household vacant (unoccupied, travel, renovation etc.) | 72 | 33 | 29 | 10 |
| Non residential unit | 25 | 11 | 3 | 11 |

(\*) These cases stem from the unpredictable nature of addresses in favelas, with many unnamed alleys and unnumbered households. The instruction for collection was to classify the address as non-existent or not found when in doubt about accessing the household for the selected address.

It can be seen that, adding together the three strata, 3,136 addresses needed to be visited to have a sample of 1,211 interviewed households. This means that 61% of the addresses visited led to non responses, which would be a serious problem if classic household sampling had been used.

At the end of the data collection operation in the 1,211 interviewed households, a sample of 1,211 adults was obtained, whose distribution by age group is shown in Table 7.

| Table 7 – The number of adults by age group, according to geographical strata | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Geographical strata | Total | Age group and sex | | | | | | | |
| 18 to 29 | | 30 to 49 | | 50 to 65 | | 66 and above | |
| Men | Women | Men | Women | Men | Women | Men | Women |
| Total | 1,211 | 113 | 168 | 190 | 303 | 116 | 166 | 49 | 106 |
| Area 1 | 406 | 32 | 66 | 68 | 101 | 36 | 53 | 14 | 36 |
| Area 2 | 400 | 38 | 50 | 63 | 107 | 34 | 63 | 18 | 27 |
| Area 3 | 405 | 43 | 52 | 59 | 95 | 46 | 50 | 17 | 43 |

Table 7 demonstrates the common selection biases in household surveys. The total number of women interviewed was 743 (61.3%) and the number of adults aged 66 years and over was 155 (12.8%). However, Table 3 indicates that women make up 51.3% of the population and adults aged 66 and over, 7.2%, which proves the availability bias (the sample has more women and more older people because more are available for interview).

It is precisely this availability bias associated with the increase in the total population that justify the calibration of the sample weights.

**References**

COCHRAN, WG. Sampling techniques. 3rd Ed. New York: John Wiley & Sons, 1977.

HALDANE, JBS. On a method of estimating frequencies. Biometrika 1945; 33:222-225.

MADEIRA, J. L.; SIMÕES, C. C. da S. Estimativas preliminares da população urbana e rural, segundo as unidades da federação, de 1960/1980, por uma nova metodologia. Revista Brasileira de Estatística, 1972, 33(129): 3-11.

Redes da Maré. Censo Populacional da Maré (2013). Rio de Janeiro: Redes da Maré, 2019. ISBN: 978-85-61382-08-7. CDD 304.66. CDU 351.778.57. Available at: <https://www.redesdamare.org.br/media/downloads/arquivos/CensoMare_WEB_04MAI.pdf>

SILVA, PLN. Calibration Estimation: When and Why, How Much and How. Rio de Janeiro: IBGE (Textos para discussão da Diretoria de Pesquisa, nº 15).

VASCONCELLOS, MTL; SILVA, PLN; SZWARCWALD CL. Sampling design for the World Health Survey in Brazil. Cadernos de Saúde Pública 2005; 21(S):S89-S99.

VASCONCELLOS MTL; SILVA PLN; ANJOS LA. Sample design for the Nutrition, Physical Activity and Health Survey (PNAFS), Niterói, Rio de Janeiro, Brazil. Estadística 2013; 65(184):47-61.