Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Amostragem Por Área: Sistematização e Aplicação

Pedro Reis Simões **13/0047902**

Brasília

2017

Pedro Reis Simões **13/0047902**

Amostragem Por Área: Sistematização e Aplicação

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2017

Dedico este trabalho à todos os amigos,

colegas e professores que me ajudaram tanto

durante a graduação e que estiveram disponíveis

nos momentos em que mais precisei.

Agradecimentos

Agradeço primeiramente ao meu pai Humberto pelo apoio que me deu durante toda a minha vida, sendo ele um modelo de sucesso que venho tentando e tentarei seguir sempre. Eu aprendo algo novo quase todas as vezes em que conversamos, como se ele fosse uma fonte inesgotável de conhecimento e sabedoria.

Ao meu irmão Guilherme pela disponibilidade e por estar ao meu lado em alguns dos momentos mais difíceis de minha vida, principalmente durante a transição do segundo grau para a universidade.

Ao avô de minha irmã, o Alcir Fleith. Se não fosse por ele eu não teria sequer começado os estudos na Universidade de Brasília. Enquanto eu mesmo e todos os outros já não tinham mais fé de que eu pudesse ser aprovado, foi o Alcir quem viu o meu nome nos jornais e me comunicou a ótima notícia.

Ao grande amigo de universidade Rodrigo Marques dos Santos, que esteve ao meu lado em momentos de grande dificuldade e grande alegria. Compartilhamos muitos aprendizados e excelentes momentos. Com certeza é um amigo que possui alma de ouro. Espero que eu tenha feito por ele ao menos metade do que ele fez por mim.

À amiga Marina Barros de Oliveira, pela carisma, inteligência e sabedoria. Além dos bons e agradáveis momentos juntos, pude ter o privilégio de aprender algumas lições de vida com ela, bem como compartilhar momentos de estudo.

Finalmente, agradeço ao professor Alan Ricardo da Silva, pela excelente orientação. Ele esteve sempre disponível e jamais me esquecerei das nossas ótimas e longas reuniões. Sou muito grato pelo seu apoio, confiança e paciência.

"Se soubermos de onde veio, então poderemos saber para onde enviar."

Karrye

Resumo

O presente trabalho tem como objetivo sistematizar e aplicar a técnica de amostragem por área. Para isto, foram utilizadas as definições apresentadas por Cochran (1977) e Kish (1965), um banco de dados gentilmente cedido por Jim Lepkowski da Universidade de Michigan nos Estados Unidos como parte do *Sampling for Survey Statisticians* e um estudo de caso na Universidade de Brasília realizado com o objetivo de ilustrar o método. Para a realização de uma amostra por área, primeiro são obtidas imagens aéreas do local que será estudado e, a partir delas, são divididos os conglomerados de acordo com características geográficas facilmente observáveis, como ruas ou rios. Em seguida, são sorteados os blocos que serão visitados e, apenas nos selecionados, são construídos mapas internos para a realização do próximo sorteio entre os elementos pertencentes à cada conglomerado. Os estudos de caso confirmaram os principais propósitos da técnica, que são a obtenção de registros populacionais para o uso futuro, a economia de recursos para a realização da amostra, como tempo e dinheiro, e a atribuição de probabilidades não nulas de seleção à todos os elementos.

Palavras-chaves: Amostragem por Área. Conglomerado. Bloco. Registros Populacionais.

Lista de Tabelas

5.1 Distribuição de salas de aula por localidade, com respectivas probabilidades de seleção e pesos amostrais (valores aproximados) 50
5.2 Distribuição de salas de aula por localidade, com respectivas probabilidades de seleção e pesos amostrais (valores aproximados) para a nova divisão dos blocos 52
5.3 Distribuição de salas de aula por localidade, organizadas pela ordem do sorteio 53
5.4 Probabilidades de seleção dos blocos da amostra e suas respectivas salas de aula 54
5.5 Quantidades observadas de alunos do sexo feminino e masculino para as salas da amostra 64

Lista de Figuras

3.1 Hierarquia geográfica das unidades definida pelo Departamento do Censo dos Estados Unidos 24
3.2 Exemplo de enumeração de domicílios em um local hipotético 26

5.1 Mapa do distrito de Madurai 36
5.2 Mapa de Melur Taluk 37
5.3 Mapa da vila Pathinettangudy 41
5.4 Mapa da aldeia Pallapatti 42
5.5 Mapa da aldeia Melapathinettangudui 43
5.6 Mapa da aldeia Keelapathinetanguddi 44
5.7 Dados relativos às trinta primeiras vilas listadas de Melur Taluk . . . 45
5.8 Foto de satélite da Asa Norte 48
5.9 Foto de satélite da Universidade de Brasília 48
5.10 Blocos considerados no estudo em ênfase 49
5.11 Desenho à Mão do Interior do BSAN 56
5.12 Desenho à Mão do Interior do PJC e do PAT 57
5.13 Enumeração das salas do BSAN 59
5.14 Enumeração das salas do PJC e do PAT 60
5.15 Tabela de números aleatórios com os RN's obtidos para cada um dos blocos selecionados 62

Sumário

Resumo **v**

1 INTRODUÇÃO **1**

1.1 OBJETIVOS 2

2 TEORIA DE AMOSTRAGEM **3**

2.1 Introdução 3

2.2 AMOSTRAGEM ALEATÓRIA SIMPLES 4

2.3 AMOSTRAGEM ESTRATIFICADA 9

2.4 AMOSTRAGEM SISTEMÁTICA 14

2.5 AMOSTRAGEM DE CONGLOMERADOS 16

2.5.1 1 Estágio 17

2.5.2 2 Estágios 20

3 AMOSTRAGEM POR ÁREA **23**

3.1 INTRODUÇÃO 23

3.2 PREPARAÇÃO DOS MAPAS 24

3.3 SELEÇÃO DE BLOCOS E DOMICíLIOS 27

3.4 BLOCOS MUITO PEQUENOS OU MUITO GRANDES 28

3.5 MODIFICANDO PARA TRÊS OU MAIS ESTÁGIOS 30

4 MATERIAL E MÉTODOS **31**

4.1 MATERIAL 31

4.2 MÉTODOS 31

5 ANÁLISE DOS RESULTADOS **34**

5.1 INTRODUÇÃO 34

5.2 ESTUDO DE CASO: DADOS DA ÍNDIA 34

5.2.1 Apresentação dos Mapas 34

5.2.2 Tamanho da Amostra 38
5.2.3 Fração Amostral 38
5.2.4 Análise da Amostra do Primeiro Estágio 39
5.2.5 Preparação das Unidades Amostrais Primárias 40
5.2.6 Seleção dos Conglomerados 40
5.2.7 Cálculo das Taxas Amostrais 44
5.2.8 Criação dos Segmentos 46
5.2.9 Listagem e Início das Entrevistas 46
5.3 ESTUDO DE CASO: UNIVERSIDADE DE BRASÍLIA 47
5.3.1 Apresentação dos Mapas 47
5.3.2 Tamanho da Amostra 49
5.3.3 Fração Amostral 50
5.3.4 Análise da Amostra do Primeiro Estágio 51
5.3.5 Preparação das Unidades Amostrais Primárias 51
5.3.6 Seleção dos Blocos 52
5.3.7 Seleção das Salas de Aula 54
5.3.8 Criação dos Segmentos 58
5.3.9 Listagem e Início das Visitas 63
5.3.10 Resultados Para o Estudo na UnB 65

6 CONCLUSÕES **67**

REFERÊNCIAS **69**

Capítulo 1

INTRODUÇÃO

Nos campos da pesquisa e do conhecimento é comum a necessidade da observação de um fenômeno e/ou característica diretamente na população alvo. No entanto, o trabalho de coleta destas informações pode se tornar uma tarefa difícil. A delimitação da área de estudo, a existência de uma base com os elementos da população, bem como a garantia de que todo elemento dessa população alvo tenha probabilidade de pertencer à amostra maior do que zero são alguns dos impecílios que o pesquisador pode encontrar. Neste sentido, a amostragem por área bem sistematizada se torna uma opção viável capaz de fornecer as respostas desejadas e resolver os problemas de seleção da amostra.

Segundo Kish (1965), a amostragem por área é um procedimento prático de listagem que possibilita a criação de bons registros populacionais para seleção de domicílios. Esse tipo de amostragem fornece um efetivo e conveniente cadastro de domicílios e pessoas por diversas razões: (1) com um procedimento de mapeamento oficial toda a população de domicílios (elementos) pode, claramente, ser identificada com uma lista definida de blocos e segmentos (conglomerados). (2) Essas identificações não se alteram do momento da listagem até o período de pesquisa. (3)

O pesquisador de campo pode prontamente e claramente identificar os limites dos blocos e segmentos, bem como os domicílios pertencentes a eles. (4) Os domicílios servem de convenientes intermédios para amostrar pessoas, porque são prontamente identificáveis, relativamente estáveis, usualmente contém poucas pessoas e cada pessoa pode ser identificada com um, e apenas um, domicílio.

Além disso, segundo Lepkowski e Heeringa (2016), muitas das populações estão distribuídas ao longo de uma larga área geográfica e seria inviável realizar um procedimento completamente aleatório no local, dadas as limitações de tempo e de custo, principalmente. A amostragem por área tem como objetivo também reduzir esse custo por meio de conglomerados, que são divididos e selecionados de acordo com com as definições da técnica que serão apresentadas posteriormente neste trabalho.

1.1 OBJETIVOS

O objetivo geral do trabalho é apresentar as etapas para seleção de unidades de análise utilizando a técnica de amostragem por área. Os objetivos específicos são:

- Caracterizar a amostragm por área;
- Fazer uma aplicação com dados reais.

Capítulo 2

TEORIA DE AMOSTRAGEM

2.1 Introdução

Bussab e Morettin (2002) definem população e amostra da seguinte forma: *População é o conjunto de todos os elementos ou resultados sob investigação. Amostra é qualquer subconjunto da população.*

Desta forma, existem diversas técnicas de amostragem para as mais diversas situações. Cada uma destas técnicas apresentará vantagens e desvantagens, cabendo então ao pesquisador escolher a mais adequada.

Como dificilmente é possível observar toda a população (ou seja, realizar um censo), pode-se listar algumas das principais vantagens em utilizar a amostragem, comparada com o censo, de acordo com a realidade e a perspectiva do século anterior (Cochran, 1977): se os dados forem obtidos apenas de uma fração da população, então os gastos financeiros serão menores; coletar informações apenas da amostra é um processo mais rápido; em alguns casos o pesquisador deve ser bem treinado ou o equipamento muito específico para a pesquisa, sendo assim inviável realizar a pesquisa em toda a população por falta de insumos; e, por conta dos itens anteriores,

a acurácia da pesquisa é maior quando se trabalha apenas com uma amostra. Entretanto, deve-se ressaltar que, com o aprimoramento da tecnologia, algumas amostras apresentam uma quantidade tão grande de dados e informações que são tão complexas de serem analisadas - e obtidas - quanto alguns dados censitários.

Para a realização de uma amostra representativa e de qualidade, Cochran (1997) recomenda seguir os seguintes passos de como deve ser o processo de planejamento e execução de uma pesquisa por amostragem: o objetivo da pesquisa deve ser claro; a população da qual a amostra será retirada muito bem especificada; os dados que serão coletados dos elementos bem definidos e relevantes para atingir o objetivo; o grau de incerteza (ou seja, o erro máximo tolerado para os estimadores que serão obtidos) desejado na pesquisa ser preciso, evidenciando o tempo e o custo que serão investidos nela; definir os métodos de mensuração e coleta dos dados (questionários, entrevista, observação, etc.); escolher o método de seleção da amostra; aplicar um pré-teste para revelar possíveis erros antes de realizar a pesquisa em uma escala maior; realizar a pesquisa de campo organizadamente; e obter um registro que poderá ser utilizado em pesquisas futuras.

Dessa forma, algumas técnicas de amostragem serão brevemente especificadas nas seções seguintes.

2.2 AMOSTRAGEM ALEATÓRIA SIMPLES

Segundo Cochran (1997), Amostragem Aleatória Simples (AAS) é um método para selecionar uma amostra de n elementos de um total de N, na qual todas as

$\binom{N}{n}$ possíveis amostras distintas possuem a mesma probabilidade de serem selecionadas. Para o sorteio ser aleatório, enumeram-se os elementos da população de 1 até N e atribui-se a mesma probabilidade de seleção a todos os números. Para o processo ser aleatório, geralmente sçao utilizados programas computacionais ou, como era comumente feito antes da existência dos recursos tecnológicos atuais, utiliza-se uma tabela de números aleatórios, que pode ser encontrada por exemplo em RAND (2001).

A notação e as definições utilizadas serão as mesmas sugeridas por Cochran (1977). Os valores pertencentes aos elementos da população serão denotados por $y_1, y_2, \ldots, y_N$, e os valores observados na amostra por $y_1, y_2, \ldots, y_n$. Seguem algumas definições para a AAS sem reposição.

Média populacional

$$\overline{Y} = \sum_{i=1}^{N} \frac{y_i}{N}. \tag{2.1}$$

Média amostral, estimador da média populacional

$$\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}. \tag{2.2}$$

Total populacional

$$T = \sum_{i=1}^{N} y_i. \tag{2.3}$$

Estimador do total populacional

$$\hat{T} = N\overline{y}. \tag{2.4}$$

Variância de y em uma população finita

$$S^2 = \sum_{i=1}^{N} \frac{(y_i - \overline{Y})^2}{N-1}. \tag{2.5}$$

Estimador da variância de y em uma população finita

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n-1}. \tag{2.6}$$

Variância de $\overline{y}$

$$V(\overline{y}) = (1-f)\frac{S^2}{n}, \tag{2.7}$$

na qual $f = \frac{n}{N}$ é a fração amostral. A expressão $(1-f)$ é conhecida como correção para populações finitas ou do inglês fpc (finite population correction).

Estimador da variância de $\overline{y}$

$$\widehat{V}(\overline{y}) = (1-f)\frac{s^2}{n}. \tag{2.8}$$

Erro padrão de $\overline{y}$

$$EP(\overline{y}) = \sqrt{V(\overline{y})}. \tag{2.9}$$

Estimador do erro padrão de $\overline{y}$

$$\widehat{EP}(\overline{y}) = \sqrt{\widehat{V}(\overline{y})}. \tag{2.10}$$

Variância de $\hat{T}$

$$V(\hat{T}) = N^2(1-f)\frac{S^2}{n}. \tag{2.11}$$

Estimador da variância de $\hat{T}$

$$\widehat{V}(\hat{T}) = N^2(1-f)\frac{s^2}{n}. \tag{2.12}$$

Estimador do erro padrão de $\hat{T}$

$$\widehat{EP}(\hat{T}) = \sqrt{\widehat{V}(\hat{T})}. \tag{2.13}$$

Entretanto, o pesquisador pode estar interessado em saber qual a proporção ou porcentagem de elementos possuem determinada característica, fazendo com que a variável observada seja qualitativa. Nestes casos, Cochran (1977) apresenta também algumas definições. Sejam A e a os números de unidades da população e da amostra, respectivamente, que pertencem à classe C. Portanto, a proporção de unidades que pertencem à classe C na população e na amostra são $P = \frac{A}{N}$ e $p = \frac{a}{n}$, respectivamente. Seja ainda y_i o valor observado no i-ésimo elemento da população ou da amostra e $y_i = 0, 1$, assumindo 1 ou 0 se pertence ou não, respectivamente, à classe C. Seguem algumas definições para o caso de proporções em AAS sem reposição.

Média populacional

$$\overline{Y} = \frac{A}{N} = P. \tag{2.14}$$

Média amostral, estimador da média populacional

$$\overline{y} = \frac{a}{n} = p. \tag{2.15}$$

Variância de y

$$S^2 = \frac{N}{N-1} PQ, \tag{2.16}$$

em que $Q = 1 - P$.

Estimador da variância de y

$$s^2 = \frac{n}{n-1} pq. \tag{2.17}$$

em que $q = 1 - p$.

Variância de p

$$V(p) = \frac{(N-n)}{N-1} \frac{PQ}{n}. \tag{2.18}$$

Estimador da variância de p

$$\widehat{V}(p) = (1-f)\frac{pq}{n-1}. \tag{2.19}$$

Intervalo de confiança para a média amostral

$$IC(\overline{Y}, (1-\alpha)) = \left[\overline{y} - z_{\alpha/2}\sqrt{\frac{\widehat{V}(p)}{n}}; \overline{y} + z_{\alpha/2}\sqrt{\frac{\widehat{V}(p)}{n}}\right], \tag{2.20}$$

em que $(1-\alpha) \times 100\%$ é o nível de confiança e $z_{\alpha/2}$ é o quantil da distribuição normal.

Entretanto, existe ainda uma decisão a ser tomada pelo pesquisador: qual o tamanho da amostra? Segundo Cochran (1977) e para o caso de AAS, pode-se decidir o tamanho da amostra n com base em conhecimentos de estudos anteriores e do maior erro tolerado para a pesquisa. Portanto, suponha que deseja-se estimar o valor da média de uma variável em uma população utilizando AAS. Sob a hipótese de normalidade para a distribuição da média amostral, o tamanho da amostra neste caso seria:

Tamanho mínimo da amostra para grandes populações ($f \approx 0$)

$$n_0 = \frac{z^2_{\alpha/2}\sigma^2}{\epsilon^2}, \tag{2.21}$$

na qual σ^2 é a variância da variável em análise e ϵ é o erro tolerado pelo pesquisador.

A estimação do tamanho de amostra também pode ser utilizada para o caso de proporções. Para isso, utiliza-se o seguinte estimador:

Tamanho mínimo da amostra para o caso de proporções (para $f \approx 0$)

$$n_{0prop} = \frac{z^2_{\alpha/2}p(1-p)}{\epsilon^2}. \tag{2.22}$$

Mas uma informação precisa sobre o valor de p nem sempre está disponível e, portanto, o cálculo de n_{0prop} pode ser feito considerando a variância máxima, ou seja, o valor de p que maximiza $p(1-p)$. No caso, $p = 0,5$. Então temos que:

Tamanho da amostra para o caso de proporções (caso conservador com $f \approx 0$)

$$n_{0propmax} = \frac{z^2_{\alpha/2} 0,25}{\epsilon^2}. \tag{2.23}$$

Entretanto, o cálculo de n_0 tem como pressuposto que o valor de N tende ao infinito, ou seja, que a população é "muito grande". O mesmo cálculo do tamanho de amostra pode ser feito utilizando uma correção para populações finitas. Desta forma, utiliza-se o seguinte estimador:

Tamanho da amostra com correção para população finita

$$n = \frac{n_0}{1 + \frac{n_0}{N}}. \tag{2.24}$$

A fórmula geral para o tamanho da amostra é apresentada em (2.21). Independente da técnica utilizada, os valores de $z^2_{\alpha/2}$ e ϵ^2 são constantes e apenas o valor de σ^2 é alterado conforme a técnica utilizada. A Equação (2.24) é a fórmula geral para o tamanho da amostra com correção para população finita.

2.3 AMOSTRAGEM ESTRATIFICADA

Segundo Cochran (1977), em amostragem estratificada a população de N indivíduos é dividida em subpopulações com $N_1, N_2, \ldots, N_L$ unidades, respectivamente. Um elemento não pode pertencer a duas ou mais subpopulações diferentes,

fazendo com que

$$N_1 + N_2 + \ldots + N_L = N. \tag{2.25}$$

Cada uma dessas subpopulações recebe o nome de estrato. Se uma amostra aleatória simples for realizada dentro de cada estrato, todo o processo é chamado de amostra aleatória estratificada.

Usando a notação proposta por Cochran (1977), em que h se refere ao estrato e i ao i-ésimo elemento dentro do estrato, temos:

- N_h é o número total de elementos;
- n_h é o número de elementos pertencentes à amostra;
- y_{ih} é o valor obtido do i-ésimo elemento;
- $W_h = \frac{N_h}{N}$ é o peso do estrato;
- $f_h = \frac{n_h}{N_h}$ é a fração amostral do estrato;
- $\overline{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$ é a verdadeira média do estrato;
- $\overline{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$ é a média amostral do estrato;
- $S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \overline{Y}_h)^2}{N_h - 1}$ é a verdadeira variância do estrato.

Com os valores acima e ainda segundo Cochran (1977), seguem algumas definições para a amostragem aleatória estratificada sem reposição e no caso das amostras retiradas de cada estrato serem independentes entre si:

Estimador da média populacional

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h, \tag{2.26}$$

na qual L é o número de estratos.

Variância do estimador da média populacional

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h}. \tag{2.27}$$

Note a semelhança entre (2.27) e (2.7). A estrutura é exatamente a mesma, mas no caso de amostragem estratificada é necessário realizar os cálculos para cada estrato e multiplicá-los pelos seus respectivos pesos ao quadrado. Em seguida, somam-se todos os valores obtidos.

Estimador de S_h^2

$$s_h^2 = \sum_{i=1}^{n_h} \frac{(y_{hi} - \overline{y}_h)^2}{n_h - 1}. \tag{2.28}$$

Note a semelhança dessa equação com a (2.6).

O estimador do total populacional em amostragem estratificada é obtido de forma análoga ao caso da AAS, apresentado em (2.4), na qual simplesmente encontra-se o produto entre N e o estimador da média populacional, no caso $\overline{y}_{st}$. A variância do estimador do total populacional também é obtido de forma análoga ao caso da AAS, apresentado em (2.11), mas, ao invés de realizar um cálculo geral, encontra-se primeiro os valores em cada estrato para então somá-los.

Entretanto, assim como na AAS, o pesquisador pode estar interessado em analisar a proporção ou a porcentagem de elementos que pertencem a uma determinada classe C. Utilizando as mesmas notações para o caso de proporções em AAS, apenas com a adição do subíndice h em alguns símbolos para denotar alguma característica do h-ésimo estrato, e ainda segundo Cochran (1977), seguem algumas definições para o estudo de proporções em amostra aleatória estratificada sem reposição no caso das

amostras retiradas de cada estrato serem independentes entre si:

Variância do estimador da proporção populacional

$$V(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2(N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}. \tag{2.29}$$

Note que o estimador da proporção populacional é $p_{st} = \overline{y}_{st}$.

Intervalo de confiança para a média

$$IC(\overline{Y}_{st}, (1-\alpha)) = \left[\overline{y}_{st} - t_{n-1,\alpha/2}\sqrt{V(\overline{y}_{st})}; \overline{y}_{st} + t_{n-1,\alpha/2}\sqrt{V(\overline{y}_{st})}\right], \tag{2.30}$$

na qual $n-1$ representa os graus de liberdade da distribuição t de Student e $t_{\alpha/2}$ é o quantil da distribuição t de Student.

Quanto à determinação dos tamanhos das amostras, Cochran (1977) sugere que eles sejam escolhidos de modo a minimizar a variância do estimador da média dado um custo específico ou de minimizar o custo dada uma variância do estimador da média específica. Sendo assim, a partir da função de custo linear, chega-se a seguinte definição:

Função de custo linear

$$C = c_0 + \sum_{h=1}^{L} c_h n_h, \tag{2.31}$$

na qual C é o custo total do trabalho de amostragem, c_0 é o custo fixo e c_h é o custo por unidade.

A partir da Equação (2.31), Cochran (1977) mostra como distribuir o tamanho dos estratos:

Alocação ótima

$$n_h = n \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} (N_h S_h / \sqrt{c_h})}, \tag{2.32}$$

em que o valor de n dependerá da situação desejada:

- se o custo for fixado e deseja-se obter os n_h's que minimizam a variância, então $n = \frac{(C-c_0)\sum_{h=1}^{L}(N_h S_h/\sqrt{c_h})}{\sum_{h=1}^{L}(N_h S_h \sqrt{c_h})}$;

- se a variância for fixada e deseja-se obter os n_i's que minimizam o custo, então $n = \frac{\left(\sum_{h=1}^{L} W_h S_h \sqrt{c_h}\right)\sum_{h=1}^{L} W_h S_h/\sqrt{c_h}}{V+(1/N)\sum_{h=1}^{L} W_h S_h^2}$.

Alternativamente, Cochran (1977) apresenta ainda mais dois tipos de alocação: a alocação de Neyman e a alocação proporcional. Em ambas a problemática é análoga à alocação ótima, mas a equação de n_h sofre algumas alterações. Na alocação de Neyman os custos são fixos, ou seja, iguais para todos os estratos. Na alocação proporcional, o valor de n_h é proporcional ao tamanho do estrato. Ambos casos são definidos a seguir:

Alocação de Neyman

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}. \tag{2.33}$$

Alocação proporcional

$$n_h = nW_h. \tag{2.34}$$

Com o objetivo de mensurar a eficiência de um método de amostragem, ou seja, obter a menor variância dos estimadores possível, Kish (1965) propôs uma medida chamada *deff* (*design effect*). O *deff* é uma razão entre as variâncias para determinada estatística. Para fins de comparação, a variância colocada no denominador é de um método padrão e mais simples, no caso, AAS com reposição. Dessa forma é

possível observar o quão mais eficiente é um método de amostragem mais complexo em relação a AAS em termos da variância do estimador desejado.

Tem-se então a seguinte fórmula geral do *deff*:

***deff* geral**

$$deff = \frac{\widehat{V}(\phi_{Delin})}{\widehat{V}(\phi_{AAS})}, \tag{2.35}$$

em que ϕ_{Delin} é a estatística obtida através do método mais complexo de amostragem e ϕ_{AAS} é a mesma estatística, porém obtida a partir de uma AAS.

Para a amostragem estratificada, podemos obter os seguintes *deff*'s:

***deff* para o caso de amostragem aleatória estratificada com alocação proporcional**

$$deff(st_{pr}) = \frac{\widehat{V}(\overline{y}_{st_{pr}})}{\widehat{V}(\overline{y})} = 1 - \left(\frac{\sigma_e^2}{\widehat{V}(\overline{y})}\right), \tag{2.36}$$

na qual $\widehat{V}(\overline{y})$ é o estimador da variância da média em AAS, obtido na equação (2.8) e $\sigma_e^2 = \sum_{h=1}^{L} W_h(\overline{Y}_h - \overline{Y})^2$ é a variância entre os estratos.

***deff* para o caso de amostragem aleatória estratificada com alocação ótima**

$$deff(st_{ot}) = \frac{\widehat{V}(\overline{y}_{st_{ot}})}{\widehat{V}(\overline{y})} = 1 - \left(\frac{\sigma_e^2}{\widehat{V}(\overline{y})}\right) - \left(\frac{\sigma_{dp}^2}{\widehat{V}(\overline{y})}\right), \tag{2.37}$$

na qual $\sigma_{dp}^2 = \sum_{h=1}^{L} W_h(\sigma_h - \overline{\sigma})^2$ é a variabilidade entre os desvios padrões dos estratos, $\sigma_h = Y_{hi} - \overline{Y}_h$ é o desvio padrão dos valores do estrato h e $\overline{\sigma}$ é a média entre os desvios padrões de cada estrato.

2.4 AMOSTRAGEM SISTEMÁTICA

Segundo Cochran (1977), uma população com N elementos é enumerada de 1 a N

obedecendo algum sistema de ordenação. Para a seleção de n elementos, seleciona-se aleatoriamente uma unidade dentre as k primeiras, em seguida dentre as $2k$ primeiras, e assim sucessivamente. Por exemplo, se $k = 20$ e o primeiro número escolhido é $RN = 8$, então retira-se da população os elementos de número 8, 28, 48, 68, e assim por diante. Na amostragem sistemática, a seleção do primeiro elemento determina toda a amostra e as possíveis amostras são mutualmente excludentes, ou seja, um mesmo elemento pode pertencer a apenas a uma delas. Temos então que $k = fracNn$, ou seja, quanto maior o tamanho da amostra n menor será o valor de k.

Cochran (1977) descreve também algumas das vantagens da amostragem sistemática sobre o método de AAS:

- é mais fácil e mais rápido realizar o desenho amostral em campo;
- na prática, a amostragem sistemática divide a população em estratos e retira-se um elemento de cada um, fazendo com que a amostra tenda a ser mais heterogênea e representativa.

Seja $\overline{y}_s$ a média de uma amostra sistemática e $N = nk$. Note que, ao assumir este valor para N, existirão k possíveis amostras sistemáticas possíveis de serem retiradas, dado que a população foi dividida em k estratos e que apenas um deles será selecionado. Segundo Cochran (1977), temos então algumas definições:

Estimador da média populacional

$$\overline{y}_s = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{2.38}$$

Variância do estimador da média populacional

$$V(\overline{y}_s) = \frac{S^2}{n}\left(\frac{N-1}{N}\right)[1+(n-1)\rho_w], \tag{2.39}$$

na qual $\rho_w = \frac{2}{(n-1)(N-1)S^2}\left[\sum_{i=1}^{k}\sum_{j<u}(y_{ij}-\overline{Y})(y_{iu}-\overline{Y})\right]$ é o coeficiente de correlação entre os pares de unidades que estão na mesma amostra sistemática, y_{ij} e y_{iu} é o valor observado no j-ésimo elemento do i-ésimo estrato e o valor observado no u-ésimo elemento do i-ésimo estrato, respectivamente, e $\overline{Y}$ é a média populacional. Note que, quanto menor for o valor de ρ_w, menor será o valor de $V(\overline{y}_s)$, ou seja, quanto maior for a variabilidade entre os elementos da amostra menor será a variância do estimador da média.

2.5 AMOSTRAGEM DE CONGLOMERADOS

Em diversas situações dentro do problema de amostragem não é possível ou não é viável realizar um cadastro de todos os elementos da população para se realizar um sorteio. Por isso, Cochran (1977) define a amostragem de conglomerados como uma técnica em que as unidades sorteadas são os blocos a que os elementos pertencem. Em um típico problema de amostrar pessoas de uma cidade, por exemplo, como é inviável obter um registro com todas as pessoas de uma determinada região, faz-se primeiro uma amostra dos domicílios do local para depois se obter os dados de todos os residentes nos domicílios amostrados.

Desta forma, Cochran (1977) apresenta algumas definições de amostragem de conglomerados (também muito conhecida pelo seu nome em inglês *cluster*), abordadas nas duas subseções seguintes.

2.5.1 1 Estágio

Para o caso em que cada conglomerado possuir o mesmo número de elementos, e se o método de AAS sem reposição for utilizado para sortear os conglomerados e a correção para populações finitas seja desconsiderada, Cochran (1977) traz a seguinte definição. Com a letra u representando o u-ésimo conglomerado, sejam M_u o tamanho do conglomerado u, S_u^2 a variância dentro do conglomerado u e C_u o custo relativo para obter registros do conglomerado u. Então, o custo relativo dada uma variância especificada ou uma variância relativa dado um custo especificado é $\propto \frac{C_u S_u^2}{M_u^2}$ (em que o símbolo $\propto$ significa "proporcional a").

Segundo Cochran (1977), em amostragem de conglomerados é necessário fazer uma distinção entre a média por conglomerado $\overline{Y}$, que iguala-se a (2.1) da AAS, e a média por elemento $\overline{\overline{Y}} = \sum_{i=1}^{A} \frac{y_i}{AM} = \frac{\overline{Y}}{M}$, na qual A é o número de conglomerados na população e M é a quantidade de elementos em cada conglomerado. A partir destes resultados, segue que:

Média amostral por elemento

$$\overline{\overline{y}} = \sum_{i=1}^{a} \frac{y_i}{aM}, \tag{2.40}$$

em que a é o número de conglomerados na amostra.

Variância entre elementos

$$S^2 = \frac{\sum_{i,j}(y_{ij} - \overline{\overline{Y}})^2}{AM - 1}, \tag{2.41}$$

em que y_{ij} representa o valor observado no j-ésimo elemento do i-ésimo conglomerado.

Estimador da variância entre elementos

$$s^2 = \frac{\sum_{i,j}(y_{ij} - \overline{\overline{y}})^2}{aM - 1}. \tag{2.42}$$

Variância da média amostral por elemento

$$V(\overline{\overline{y}}) = \frac{1-f}{aM} S^2[1 + (M-1)\rho], \tag{2.43}$$

na qual $\rho = \frac{2\sum_{i=1}^{A}\sum_{j<k}(y_{ij}-\overline{\overline{Y}})(y_{ik}-\overline{\overline{Y}})}{(M-1)(AM-1)S^2}$ é o coeficiente de correlação intraconglomerado, semelhante ao ρ apresentado na equação (2.39).

Estimador da variância da média amostral por elemento

$$\widehat{V}(\overline{\overline{y}}) = \frac{1-f}{aM} s^2[1 + (M-1)\rho]. \tag{2.44}$$

Assim como nas outras técnicas de amostragem, o pesquisador pode estar interessado em observar proporções. Neste caso, utilizando as mesmas notações de proporção utilizadas na seção de AAS, Cochran (1977) define a variância da proporção amostral da seguinte forma:

Variância da proporção amostral

$$V(p) = \frac{A-a}{Aa} \sum_{i=1}^{A} \frac{(p_i - P)^2}{A-1}. \tag{2.45}$$

Estimador da variância da proporção amostral

$$\widehat{V}(p) = (1-f)M \sum_{i=1}^{a} \frac{(p_i - p)^2}{a-1}. \tag{2.46}$$

Um outro caso neste tipo de amostragem é o de conglomerados com tamanhos diferentes. Cochran (1977) também apresenta algumas definições para esta situação:

Média populacional por elemento, conglomerados de tamanhos diferentes

$$\overline{\overline{Y}}_{cong} = \frac{1}{M_0} \sum_{i=1}^{A} \sum_{j=1}^{M_i} y_{ij}, \tag{2.47}$$

na qual M_i é o tamanho do conglomerado i e $M_0 = \sum_{i=1}^{A} M_i$.

Média amostral por elemento, conglomerados de tamanhos diferentes

$$\overline{\overline{y}}_{cong} = \frac{1}{m_0} \sum_{i=1}^{a} \sum_{j=1}^{M_i} y_{ij}, \tag{2.48}$$

na qual $m_0 = \sum_{i=1}^{n} M_i$.

Variância entre elementos

$$S^2_{cong} = \frac{\sum_{i,j} (y_{ij} - \overline{\overline{Y}}_{cong})^2}{M_0 - 1}. \tag{2.49}$$

Estimador da variância entre elementos

$$s^2_{cong} = \frac{\sum_{i,j} (y_{ij} - \overline{\overline{y}}_{cong})^2}{m_0 - 1}. \tag{2.50}$$

Variância da média amostral por elemento, conglomerados de tamanhos diferentes

$$V(\overline{\overline{y}}_{cong}) = \frac{1-f}{m_0} S^2 \left[1 + \left(\frac{M_0}{A} - 1 \right) \rho \right]. \tag{2.51}$$

Estimador da variância da média amostral por elemento, conglomerados de tamanhos diferentes

$$\widehat{V}(\overline{\overline{y}}_{cong}) = \frac{1-f}{m_0} s^2 \left[1 + \left(\frac{m_0}{a} - 1 \right) \rho \right]. \tag{2.52}$$

Variância da proporção amostral, conglomerados de tamanhos diferentes

$$V(p) = \frac{A-a}{Aa\overline{M}^2} \frac{\sum_{i=1}^{A} M_i^2 (p_i - P)^2}{A-1}, \tag{2.53}$$

em que $\overline{M}$ é o tamanho médio dos conglomerados na população.

Estimador da variância da proporção amostral, conglomerados de tamanhos diferentes

$$\widehat{V}(p) = \frac{1-f}{a} \sum_{i=1}^{a} \left(\frac{m_i^2}{\overline{m}} \right) \frac{(p_i - p)^2}{a-1} \tag{2.54}$$

em que $\overline{m}$ é o tamanho médio dos conglomerados da amostra.

Intervalo de confiança para a média

$$IC(\overline{\overline{Y}}, (1-\alpha)) = \left[\overline{\overline{y}} - t_{n-1,\alpha/2} \sqrt{\widehat{V}(p)}; \overline{\overline{y}} + t_{n-1,\alpha/2} \sqrt{\widehat{V}(p)} \right]. \tag{2.55}$$

2.5.2 2 Estágios

Existem ainda as situações em que a amostra deve ser feita em mais de um estágio. Nesses casos, os elementos da população pertencem a subgrupos de conglomerados. Se não for possível ou viável considerar todos os elementos dentro dos conglomerados sorteados tal que uma amostra seja realizada dentro de cada um deles, então esse é um caso de amostragem de conglomerados em dois estágios.

Para apresentar as definições serão utilizadas as notações e os resultados obtidos por Cochran (1977). Sejam a a quantidade de conglomerados na amostra, M o tamanho de cada conglomerado e m a quantidade de subunidades retiradas de cada conglomerado na amostra. Algumas notações seguem abaixo:

- y_{ij} é o valor observado na j-ésima subunidade do i-ésimo conglomerado;
- $\overline{y}_i = \sum_{j=1}^{m} \frac{y_{ij}}{m}$ é a média por subunidade dos valores do i-ésimo conglomerado;
- $\overline{\overline{y}} = \sum_{i=1}^{a} \frac{\overline{y}_i}{a}$ é a média total por subunidade;

- $S_1^2 = \frac{\sum_{i=1}^{A}(\overline{Y}_i - \overline{\overline{Y}})^2}{A-1}$ é a variância entre os elementos dos conglomerados;

- $S_2^2 = \frac{\sum_{i=1}^{A}\sum_{j=1}^{M}(y_i j - \overline{Y}_i)^2}{A(M-1)}$ é a variância entre os elementos das subunidades.

Note que quando $m = M$ temos uma amostragem de conglomerados em um estágio.

Com o pressuposto de que a amostra dos conglomerados e das subunidades foram realizadas segundo o método de AAS, tem-se (Cochran, 1977):

Variância da média total por subunidade

$$V(\overline{\overline{y}}) = \frac{1-f_1}{a}S_1^2 + \frac{1-f_2}{ma}S_2^2, \tag{2.56}$$

na qual $f_1 = \frac{a}{A}$ e $f_2 = \frac{m}{M}$ são as frações amostrais no primeiro e no segundo estágio respectivamente.

Variância amostral entre os elementos dos conglomerados

$$s_1^2 = \frac{\sum_{i=1}^{a}(\overline{y}_i - \overline{\overline{y}})^2}{a-1}. \tag{2.57}$$

Variância amostral entre os elementos das subunidades

$$s_2^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{m}(y_i j - \overline{y}_i)^2}{a(m-1)}. \tag{2.58}$$

Estimador da variância da média total por subunidade

$$\widehat{V}(\overline{\overline{y}}) = \frac{1-f_1}{a}s_1^2 + \frac{f_1(1-f_2)}{ma}s_2^2. \tag{2.59}$$

Para o caso do estudo envolver proporções, Cochran (1977) também apresenta algumas definições:

Variância amostral entre os elementos dos conglomerados para proporções

$$s_{1prop}^2 = \frac{\sum_{i=1}^{a}(p_i - \overline{p})^2}{a-1}, \tag{2.60}$$

na qual $p_i = \frac{a_i}{m}$ é a proporção de elementos na subunidade que possuem uma determinada característica C, a_i é a quantidade de elementos na subunidade que possuem esta característica e $\overline{p} = \frac{\sum_{i=1}^{a} p_i}{a}$.

Variância amostral entre os elementos das subunidades para proporções

$$s^2_{2prop} = \frac{m}{a(m-1)} \sum_{i=1}^{a} p_i q_i, \tag{2.61}$$

em que $q_i = 1 - p_i$.

Estimador da variância da média total por subunidade para proporções

$$\widehat{V}(\overline{p}) = \frac{1-f_1}{a(a-1)} \sum_{i=1}^{a} (p_i - \overline{p})^2 + \frac{f_1(1-f_2)}{a^2(m-1)} \sum_{i=1}^{a} p_i q_i. \tag{2.62}$$

Capítulo 3

AMOSTRAGEM POR ÁREA

3.1 INTRODUÇÃO

O processo de amostragem muitas vezes necessita de um bom registro (ou do inglês *frame*) da população que torne viável a aleatorização dos elementos que serão selecionados, mas nem sempre este registro existe ou está disponível. A amostragem por área propõe soluções para esse problema, possibilitando ao pesquisador a obtenção de dados da população pertencente a determinada área geográfica que não possui registros sobre os elementos (ou de possíveis grupos a que eles pertençam). Além disso, essa técnica tem o objetivo de atribuir probabilidades não nulas de seleção a todos os elementos, sendo possível então que qualquer um seja selecionado para a amostra.

A amostragem por área, portanto, é uma técnica que visa a construção de registros que ainda não existem e o mapeamento de áreas ainda não mapeadas. Por se tratar de uma técnica que envolve diversas camadas de área geográfica, Valliant et al. (2013) exemplifica uma hierarquia delas de acordo com suas respectivas extensões populacionais e áreasutilizada em um levantamento feito nos EUA (3.1).

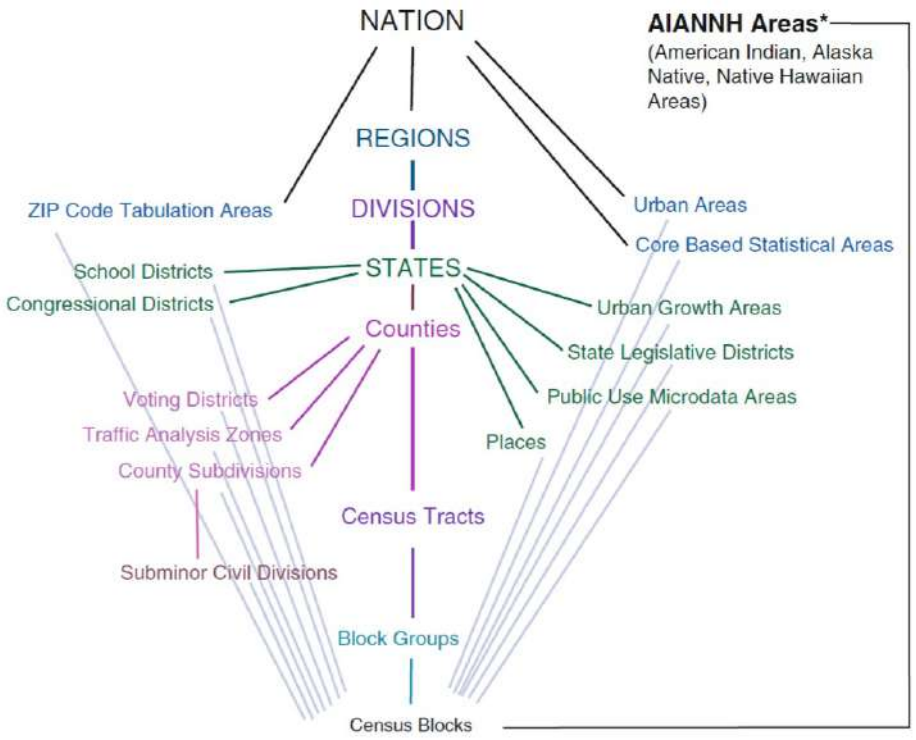


Figura 3.1: Hierarquia geográfica das unidades definida pelo Departamento do Censo dos Estados Unidos (Fonte: Valliant et al. (2013))

3.2 PREPARAÇÃO DOS MAPAS

Como os registros a serem feitos dependem da coleta dos dados, a primeira etapa consiste em mapear a área desconhecida. Portanto, Kish (1965) descreve diversos procedimentos para a preparação de um mapa que proporcione bons registros para domicílios, que serão esclarecidos nesta seção.

A área deve ser dividida em blocos, mas para que isso seja possível são necessárias fronteiras. Por isso, deve-se levar em conta o objetivo da pesquisa, para que não seja possível obter uma amostra com elementos que não pertençam à área que se deseja estudar. É conveniente também que sejam excluídas as áreas que, definitivamente, não contenham domicílios, como por exemplo estádios, lagos ou parques.

As fronteiras dos blocos devem ser muito bem definidas, permanentes e facilmente identificáveis pelo pesquisador de campo. Margens de rios, avenidas, ruas ou qualquer outro tipo de espaço físico que não contenha domicílio pode ser uma fronteira conveniente. Por fim, as fronteiras devem ser construídas de maneira que os domicílios pertençam a apenas um bloco e que possuam probabilidade maior do que zero de serem selecionados para a amostra.

Em seguida, deve-se considerar a estratificação ordinária dos domicílios, classificando-os em grupos semelhantes entre si, como vizinhanças, apartamentos e condomínios. Esta estratificação visa a diversidade da amostra, uma vez que ao menos um elemento seja retirado de cada estrato. Pode ser conveniente considerar estratos de domicílios humildes, hotéis e mansões. para diversificar ainda mais a amostra.

Para ser possível sortear os domicílios é necessário enumerá-los. Essa enumeração pode ser feita de forma crescente e com números inteiros (de 1 em diante, por exemplo) ou seguir alguma enumeração já estabelecida no local (como o número da casa ou do apartamento por exemplo). Uma maneira eficiente de enumerar domicílios é em forma de sepente, ou seja, enumerar os domicílios de acordo com o caminho que o próprio pesquisador de campo irá realizar. A Figura 3.2 exemplifica esse procedimento.

Entretanto, todas estas ideias foram propostas na década de sessenta. Na época, uma considerável parte dos recursos financeiros destinados à pesquisa eram utilizados na primeira etapa, a do reconhecimento aéreo do local. Para isso, sobrevoava-se a

Figura 3.2: Exemplo de enumeração de domicílios em um local hipotético

área desejada em um balão ou qualquer outro meio de transporte aéreo disponível para obter os primeiros desenhos da área em estudo. Hoje em dia, porém, existem disponíveis as fotos de satélite e o acesso a elas é gratuito. Além disso, as próprias imagens obtidas podem ser utilizadas diretamente, não havendo a necessidade do pesquisador desenhar a mão todo o local. O principal cuidado que deve ser tomado é o de verificar a área pessoalmente, dado que as fotos podem estar desatualizadas.

Um resumo de forma sistemática sobre o material primordial necessário para a realização de uma amostragem por área que Kish (1965) apresenta inclui: definir os limites da área de estudo, dividir a área de estudo em estratos, definir os limites

dos blocos e atribuir à eles uma medida de tamanho, que será apresentada na seção seguinte.

3.3 SELEÇÃO DE BLOCOS E DOMICíLIOS

Seja a fração amostral $f = \frac{n}{N}$, em que n é o tamanho da amostra e N o tamanho da população, da mesma forma que foi definida na Equação (2.7). Sejam também a o número de blocos que serão selecionados para a amostra e A o número total de blocos na população. Segundo Kish (1965), o procedimento mais simples de selecionar uma amostra de blocos é pela amostragem sistemática, usando o intervalo f_a. Como os blocos possuem a mesma probabilidade $f_a = \frac{1}{F_a} = \frac{a}{A}$ de serem selecionados, a probabilidade de seleção dos elementos dentro dos blocos é uniforme e igual a $f_b = \frac{1}{F_b} = \frac{f}{f_a}$. Portanto, a probabilidade uniforme f de seleção geral é o resultado do produto entre f_a e f_b, ou seja, $f = f_a \times f_b = \frac{1}{F_a} \times \frac{1}{F_b} = \frac{1}{F}$.

Entretanto, seria conveniente utilizar alguma medida de peso, dando assim maior probabilidade de seleção aos blocos que tivessem maior número de domicílios. Seguindo o raciocinio de Kish (1965), multiplica-se f_a por uma medida de tamanho do bloco ou do inglês *MoS* (*measure of size*), aumentando ou diminuindo, dessa forma, o valor de f_a. Porém, para manter o valor de f inalterado, multiplica-se f_b por $\frac{1}{MoS}$, alterando a probabilidade de seleção dos domicílios dentro do bloco. Além disto, dada a definição de *MoS*, tem-se que $f_b = \frac{b_i}{MoS}$, em que b_i é o número de elementos que serão selecionados dentro i-ésimo bloco. Logo, o valor de f em uma AAS em dois estágios seria:

$$f = \frac{1}{F} = \frac{a}{A} \times \frac{b_i}{MoS} = (MoS \times f_a) \times \left(\frac{f_b}{MoS}\right) = \frac{MoS}{F_a} \times \frac{1}{MoS \times F_b}. \quad (3.1)$$

Entretanto, se quisermos encontrar o valor de b_i, colocando-o em função dos demais, observa-se que, quanto maior for o valor do MoS, maior será b_i. Ou seja, quanto maior forem os blocos selecionados, maior será o tamanho da amostra. Para fazer com que de todos os blocos selecionados seja retirada uma amostra do mesmo tamanho, Kish (1965) apresenta o valor de f no caso em que se deseja uma amostra PPS (*Probability Proportional to Size* ou probabilidade de seleção proporcional ao tamanho) que é a seguinte:

$$f_e = \frac{n}{N} = \frac{a\ MoS}{N} \times \frac{b_i}{MoS} \Rightarrow b_i = \frac{n}{a} \quad (3.2)$$

Note que a probabilidade f_e enquadra-se em um tipo especial de amostragem, definido por Kish (1965) como *EPSEM* (*Equal Probability of Selection Method* ou método de seleção equiprovável), em que os elementos da população possuem a mesma probabilidade de serem selecionados para a amostra.

3.4 BLOCOS MUITO PEQUENOS OU MUITO GRANDES

Alguns dos problemas em amostragem por área são causados por causa de blocos muito grandes ou muito pequenos. Segundo Kish (1965), blocos são considerados muito pequenos se o valor da medida de tamanho MoS do bloco for menor do que a razão $\frac{n}{a}$. Uma das formas de evitar esta situação, sugerida por Kish (1965), é

agregar o bloco pequeno com outro adjacente, fazendo com que os dois formem um único bloco de tamanho semelhante aos demais.

Existem também os casos em que os blocos são muito grandes. Segundo Kish (1965), blocos são considerados muito grandes se o valor da medida de tamanho MoS do bloco for maior do que $\frac{N}{a}$. Ainda segundo Kish (1965), o procedimento para se lidar com esta situação é semelhante ao caso anterior, mas aqui divide-se o bloco em dois outros menores (se for possível), ou seleciona-se esse grande bloco com probabilidade 1 (nsse caso, fazendo esse bloco um extrato de tamanho unitário).

Após a identificação de blocos pequenos e grandes, se eles existirem, haverá uma nova quantidade de blocos. Deve-se então verificar novamente se ainda há blocos muito grandes ou muito pequenos. Caso eles existam, devem ser aplicadas as técnicas acima novamente. Caso contrário, este não será mais um problema. Este procedimento deve ser aplicado sucessivamente, até que não hajam mais blocos com problemas de tamanho.

Em resumo, tem-se as seguintes inequações para se identificar blocos muito grandes ou muito pequenos:

Condição para que o bloco i seja considerado muito grande

$$MoS_i > \frac{N}{a}. \tag{3.3}$$

Condição para que o bloco i seja considerado muito pequeno

$$MoS_i < \frac{n}{a}. \tag{3.4}$$

3.5 MODIFICANDO PARA TRÊS OU MAIS ESTÁGIOS

Mesmo com todas as técnicas listadas, algumas situações ainda podem exigir um intenso trabalho de listagem dos blocos e dos domicílios, além do alto custo. Uma alternativa proposta por Kish (1965) para diminuir estes problemas é a de criar um terceiro estágio para a amostra, onde a seleção dos blocos seria realizada após a seleção de uma área ainda maior que os contesse.

Kish (1965) sugere que esta área maior pode ser obtida através da observação físico do local, mantendo a ideia de que os limites geográficos facilmente observáveis delimitem cada uma destas localidades, ou da utilização de cadastros anteriores.

Entretanto, Kish (1965) ainda explícita outros casos em que a amostra pode adquirir novos estágios. Dentre elas, uma amostra de pessoas logo após a escolha dos domicílios ou a subdivisão de um bloco muito grande em outros menores. Além disso, Kish (1965) apresenta que, em algumas situações, uma etapa preliminar de seleção pode ser adicionada, ao invés de um novo estágio. Essa etapa preliminar seria a seleção de uma amostra de blocos maior do que a quantidade deles que realmente se deseja no plano amostral, podendo ser feita de forma *EPSEM*, dada a ignorância sobre o comportamento das variáveis locais. Após essa primeira seleção, a amostra que de fato será observada é realizada, levando em conta apenas os blocos que pertencem à seleção preliminar.

Capítulo 4

MATERIAL E MÉTODOS

4.1 MATERIAL

Para a realização do trabalho será utilizado um banco de dados gentilmente cedidos pelo professor Jim Lepkowski da Universidade de Michigan nos Estados Unidos como parte do *Sampling for Survey Statisticians*. Trata-se de um estudo amostral sobre a fertilidade e as práticas familiares no distrito de Madurai, localizado num estado da Índia chamado Tamil Nadu, ao sudeste da península do país.

Um exemplo dos tipos de questões a serem utilizadas no questionário estão disponíveis no modelo do *World Fertility Survey* familiar e individual. A amostra possui tamanho $n = 4.500$ e é composta por mulheres entre 15 e 44 anos, seguindo um método *EPSEM*.

Um segundo estudo de caso será feito na Unversidade de Brasília afim de estimar a proporção de alunos do sexo feminino (ou masculino), perfazendo uma amostra de tamanho $n = 402$.

4.2 MÉTODOS

Para a realização do estudo será utilizada a técnica de amostragem por área, na

qual as vilas serão selecionadas primeiro, depois as famílias e, finalmente, todas as pessoas aptas a responder o questionário dentro das famílias amostradas. No outro estudo de caso, na Universidade de Brasília, primeiro são selecionados os blocos de salas de aula e depois as próprias salas de aula.

As etapas para o processo de seleção foram as seguintes:

1. obter e apresentar os mapas: por tratar-se de uma técnica geográfica, a primeira etapa é a de obtenção dos mapas. Neste passo serão obtidas as imagens aéreas do local que se deseja estudar e a delimitação da área de interesse, seguindo os procedimentos apresentados na Seção 3.2;

2. determinar o tamanho da amostra: o pesquisador precisa esclarecer quantos elementos ele deseja - no mínimo - dentro da amostra para que, em seguida, seja escolhida a técnica amostral mais adequada;

3. calcular a fração amostral: segundo Lepkowski e Heeringa (2016), um método de seleção *EPSEM* necessita de uma fração amostral global, definida na Seção 3.3 fixada;

4. examinar a amostra do primeiro estágio: nesta etapa, a técnica de amostragem a ser utilizada precisa ser definida;

5. preparar as unidades amostrais primárias para seleção com probabilidade proporcional ao tamanho estimado dos blocos: identificar os conglomerados muito grandes e agregar os conglomerados muito pequenos, de acordo com as definições apresentadas na Seção 3.4. A preparação inclui também a enumeração

das unidades de cada uma das subdivisões das unidades primárias, para que elas possam ser selecionadas, além de satisfazer o prerequisito desejado da seleção seguir um método *EPSEM*. Finalmente, as *MoS*, definidas na Seção 3.3 devem ser atribuídas a cada bloco;

6. selecionar cada conglomerado utilizando probabilidade proporcional aos respectivos tamanhos estimados;

7. calcular as frações amostrais dentro das unidades primárias de amostragem;

8. criar os segmentos apenas das áreas amostradas: este procedimento diminui a intensidade do trabalho, pois apenas as áreas amostradas terão um desenho completo da quantidade de domicílios que a compõem. Os segmentos criados serão utilizados para a realização de uma amostra sistemática. Portanto, a criação de segmentos tem como objetivo facilitar a coleta;

9. listar as famílias e começar a realizar as entrevistas.

Por fim, os cálculos foram realizados utilizando o *software* SAS 9.4 (SAS, 2013).

Capítulo 5

ANÁLISE DOS RESULTADOS

5.1 INTRODUÇÃO

Este capítulo apresentará a análise dos resultados que foram obtidos a partir dos dados especificados na Seção 4.1. Para isto, foram seguidas as etapas apresentadas na Seção 4.2. Em seguida, uma outra seção apresentará a análise dos resultados obtidos num estudo de caso realizado na Universidade de Brasília (UnB).

5.2 ESTUDO DE CASO: DADOS DA ÍNDIA

As subseções seguintes estão enumeradas da mesma forma e na mesma ordem que os procedimentos apresentados na seção 4.2, para um melhor entendimento. Além disso, todas as etapas seguem os comandos, utilizam figuras e dados segundo Lepkowski e Heeringa (2016).

5.2.1 Apresentação dos Mapas

No estudo em questão, obteve-se uma imagem aérea do distrito de Madurai e uma outra para cada uma de suas divisões administrativas. Como a pesquisa é antiga, os mapas foram confecionados manualmente por meio de desenhos. A Figura 5.1

apresenta o desenho desenvolvido para o Distrito de Madurai, enquanto a Figura 5.2 refere-se ao desenho de uma de suas divisões administrativas, Melur Taluk - que foi realizada da mesma forma que as outras divisões.

É possível observar, na Figura 5.2, que há uma vasta região de florestas em Melur Taluk, principalmente ao norte e ao noroeste da região. Estas áreas não foram consideradas como áreas urbanas, que foi o foco do estudo.

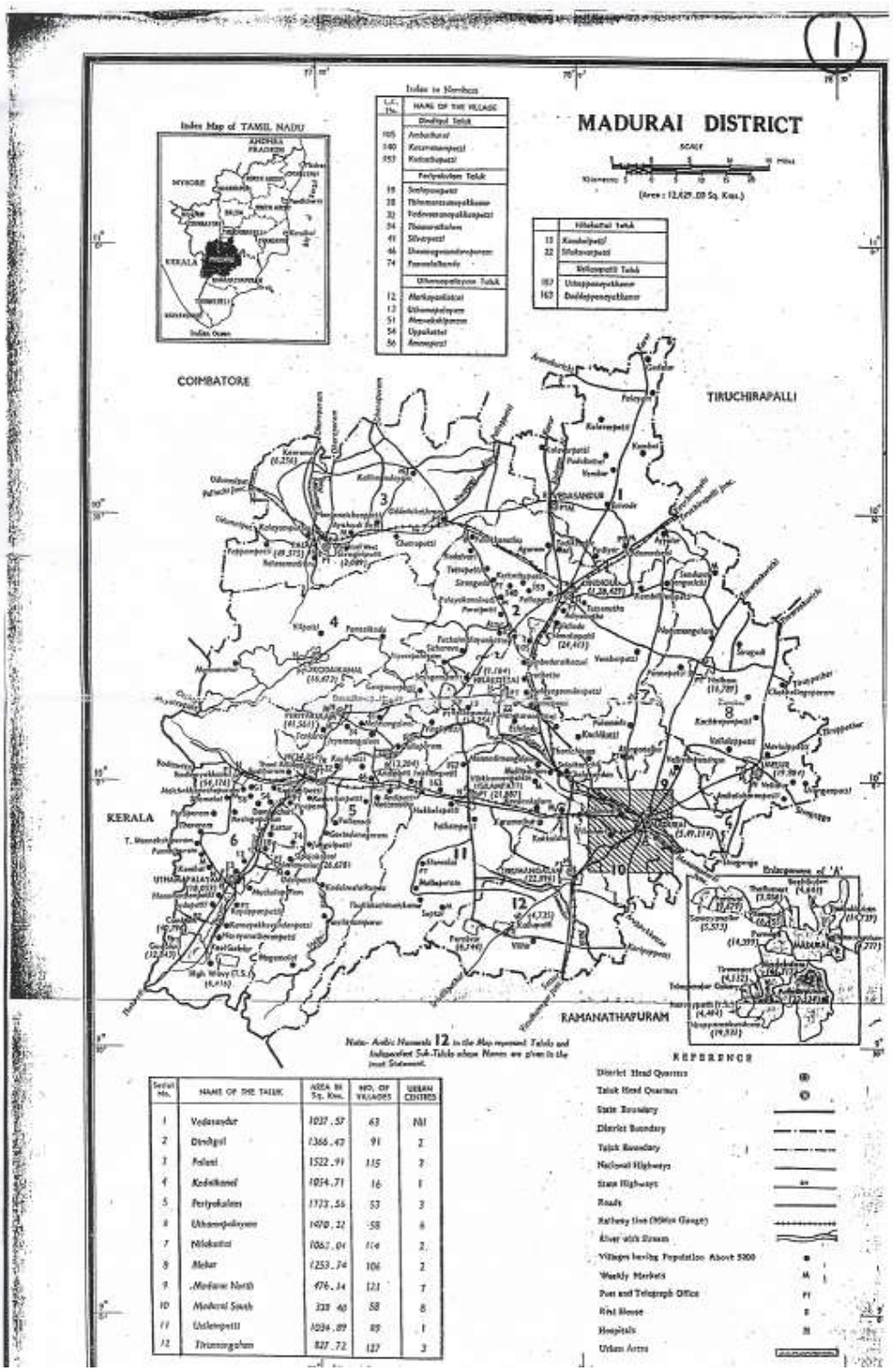


Figura 5.1: Mapa do distrito de Madurai

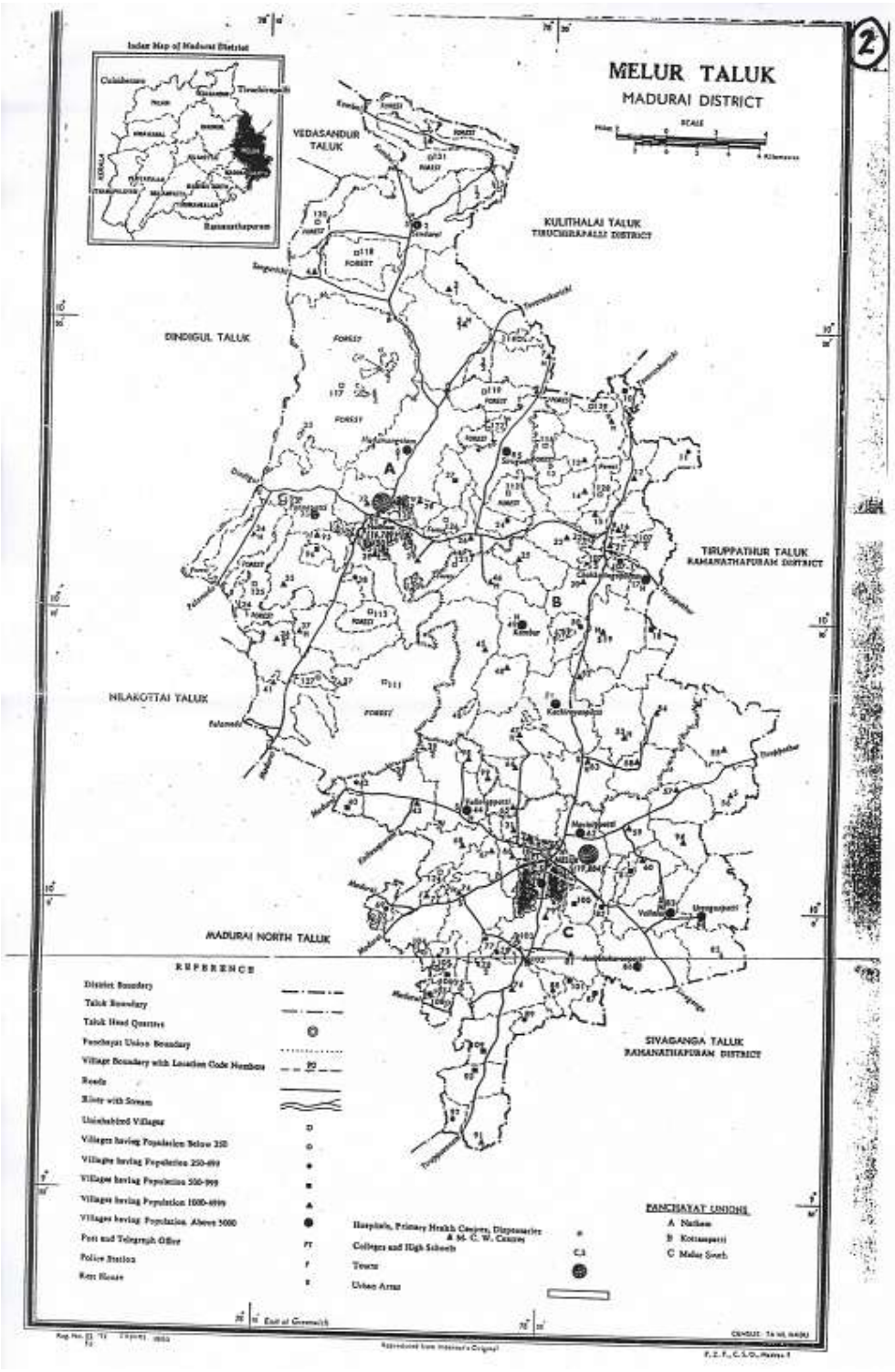


Figura 5.2: Mapa de Melur Taluk, divisão administrativa do distrito de Madurai

5.2.2 Tamanho da Amostra

Algumas considerações relacionadas à carga de trabalho sugerem que um total de $b = 25$ entrevistas completas devem ser realizadas, com mulheres elegíveis, em cada uma das vilas selecionadas e que essas considerações variam entre as localidades urbanas. Além disso, as estimativas serão obtidas para toda a amostra e para os subgrupos mais significativos dela. Erros padrão de até 5% para as estimativas das médias e das proporções dentro dos subgrupos, cujos tamanhos são próximos a $n = 400$, são almeijados.

5.2.3 Fração Amostral

Para a pesquisa em questão, são consideradas as seguintes informações obtidas num censo anterior:

- taxa esperada de resposta entre as pessoas elegíveis: $0,90$;
- número médio de pessoas por domicílio: $5,73$;
- proporção da população feminina entre 15 e 44 anos completos de idade (pessoas elegíveis): $0,22$;
- número de domicílios ocupados: 687.356;
- taxa de crescimento anual para os domicílios: $0,02$.

A taxa amostral a ser encontrada é: no numerador, o número de domicílios que devem ser visitados para se obter a quantidade de $n = 4.500$ entrevistas (quantidade definida pelo pesquisador); no denominador, o número estimado de domicílios ocupados cinco anos depois do censo anterior. Utilizando os valores considerados:

- em cada domicílio terão $(5,73 \times 0,22) = 1,2606$ pessoas elegíveis;

- dentre as pessoas elegíveis, a taxa de resposta é de $0,90$, fazendo com que em cada domicílio sejam realizadas $(1,2606 \times 0,90) = 1,13454$ entrevistas completas;

- como $n = 4.500$, então a quantidade necessária de domicílios ocupados que devem ser visitados é $\frac{4.500}{1,13454} = 3.966,36$, que, arredondada para cima, iguala-se a 3.967;

- atualizando o número total de domicílios ocupados durante cinco anos, considerando a taxa de crescimento anual de $0,02$, tem-se que a quantidade total prevista é de 758.897 domicílios ocupados;

- com os valores obtidos, a taxa amostral é $\frac{3.967}{758.897}$, que é aproximadamente igual a $0,00523$.

5.2.4 Análise da Amostra do Primeiro Estágio

A pesquisa foi realizada em todo o distrito de Madurai, que é composto por diversas divisões administrativas (dentre elas, Melur Taluk). Cada uma destas divisões é considerada heterogênea entre si. Uma técnica apropriada para esta situação, e que também foi a utilizada na pesquisa, é a estratificação das unidades amostrais primárias, definida na Seção 2.3. As vilas que compõem cada uma destas regiões foram consideradas homogêneas entre si e para selecioná-las foi utilizada a técnica de amostragem por conglomerados, definida na Seção 2.5. Finalmente, os domicílios que compõem as vilas foram selecionados de acordo com a técnica de amostragem

sistemática, definida na Seção 2.4.

5.2.5 Preparação das Unidades Amostrais Primárias

As vilas de cada uma das divisões administrativas foram enumeradas. No caso de Melur Taluk, o número total de vilas era 131. Entretanto, 21 destas vilas eram reservas florestais. Então, neste estrato entraram para o sorteio da amostra apenas as 110 vilas que possuíam domicílios.

5.2.6 Seleção dos Conglomerados

Para a seleção das vilas, foi utilizada a técnica de amostragem sistemática, apresentada na Seção 2.4. Um dos conglomerados selecionados na pesquisa foi a vila Pathinettangudy. Como as vilas eram compostas por aldeias, o pesquisador confeccionou um mapa para cada vila selecionada e, depois, um outro mapa para cada aldeia que ela continha. O mapa de Pathinettangudy é apresentado na Figura 5.3, onde é possível observar os nomes e a quantidade de domicílios de cada uma de suas três aldeias. Os mapas de cada uma das aldeias são apresentados nas Figuras 5.4, 5.5 e 5.6. Note que nesta vila a aldeia principal é Keelapathinetanguddi.

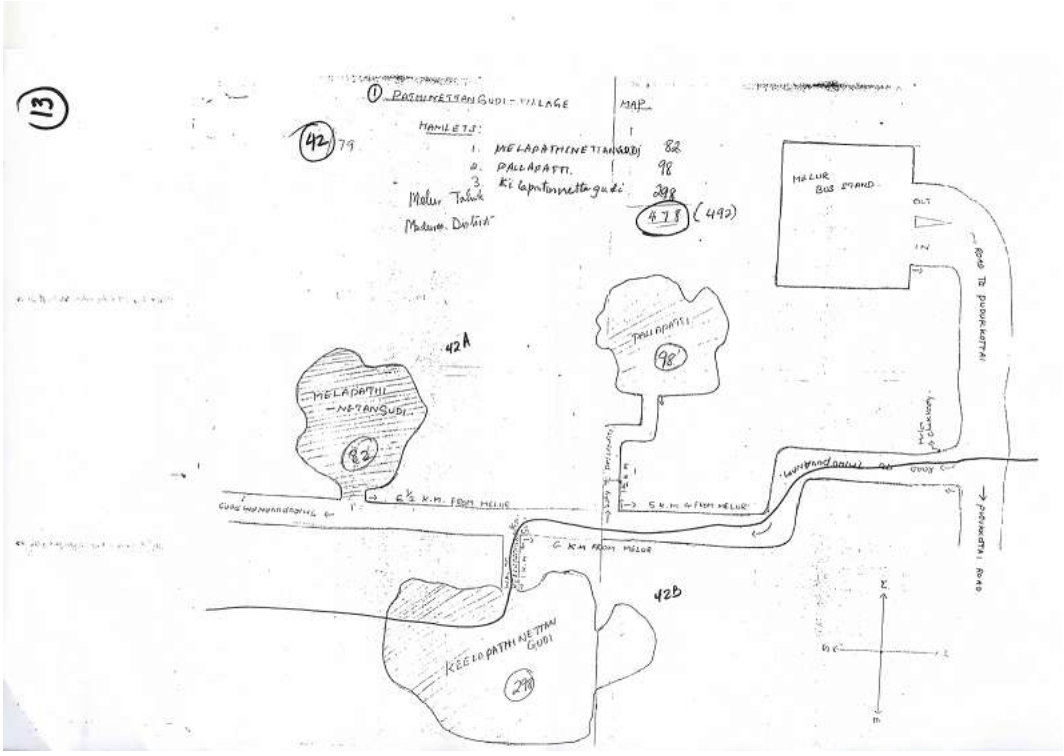


Figura 5.3: Mapa da vila Pathinettangudy

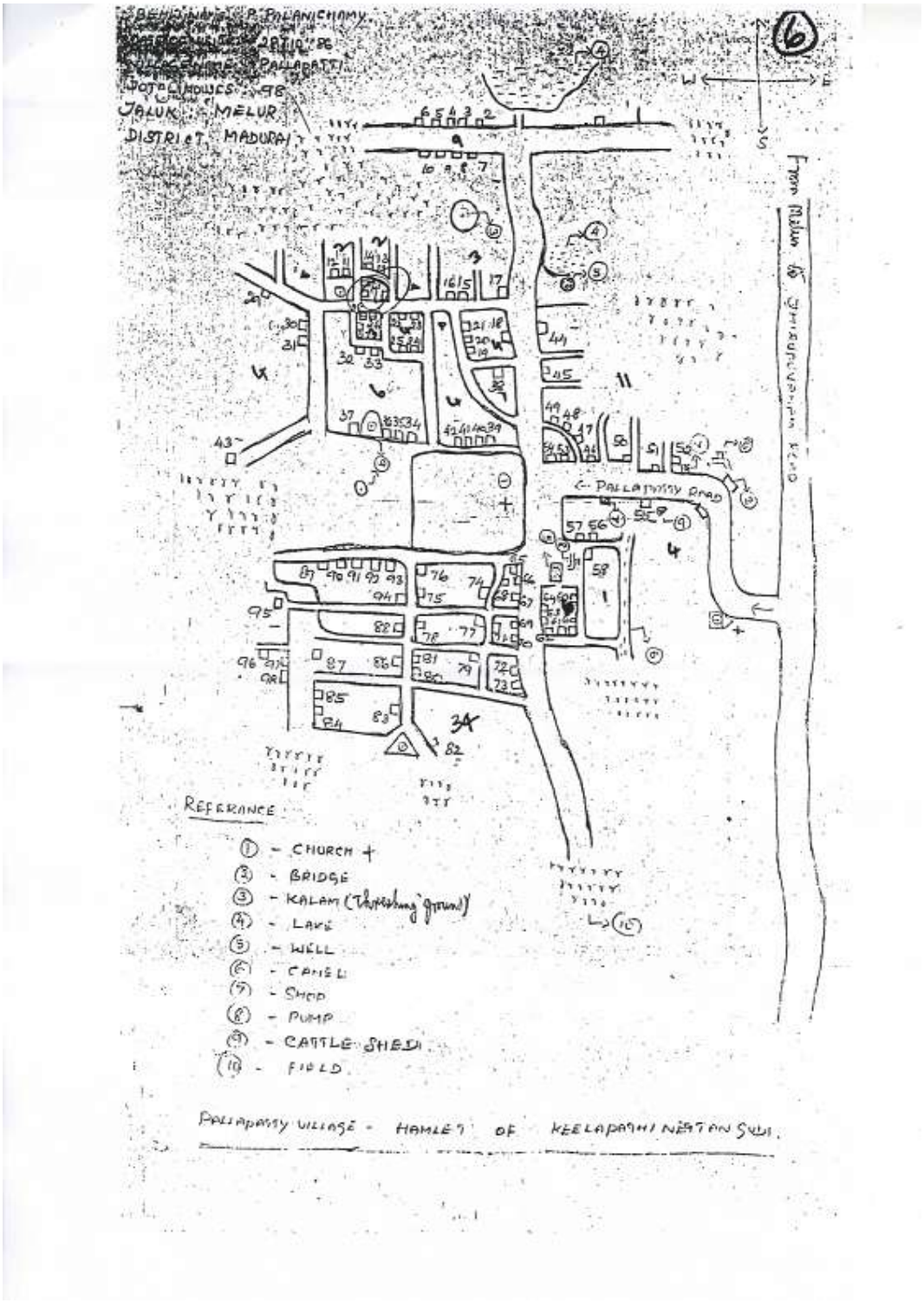


Figura 5.4: Mapa da aldeia Pallapatti

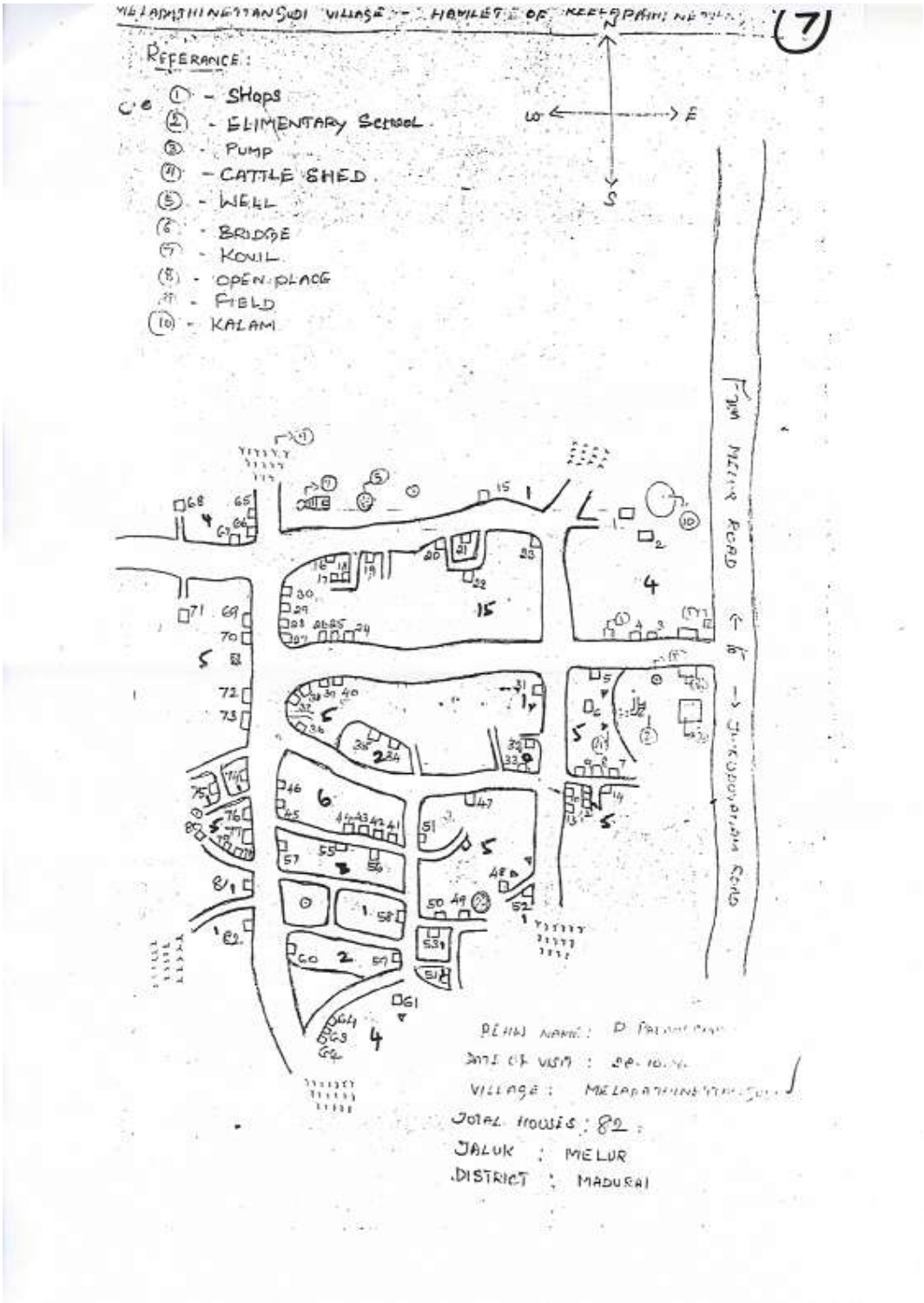


Figura 5.5: Mapa da aldeia Melapathinettangudui

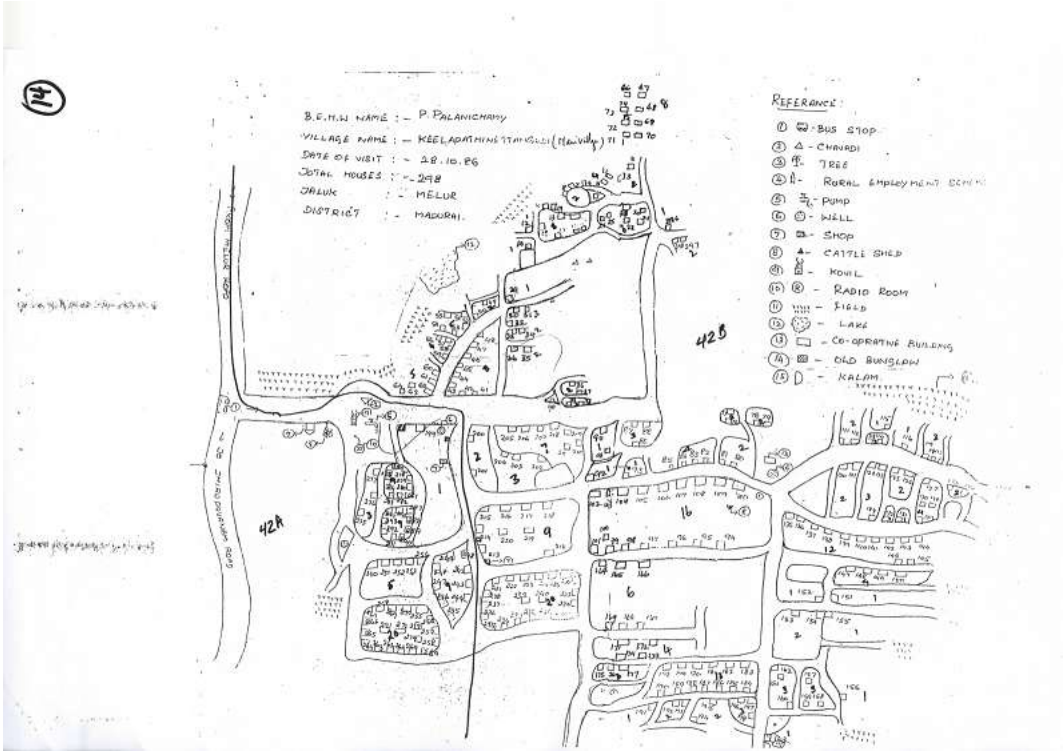
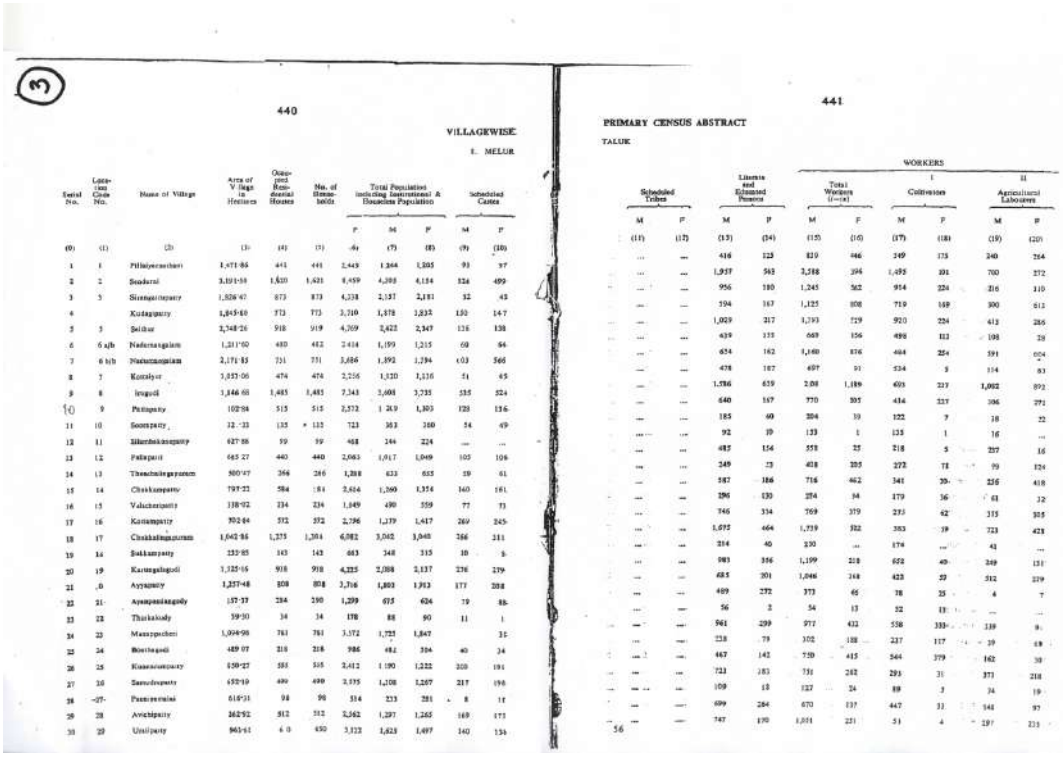


Figura 5.6: Mapa da aldeia Keelapathinetanguddi

5.2.7 Cálculo das Taxas Amostrais

A Figura 5.7 mostra como os dados de Melur Taluk foram organizados.



440

VILLAGEWISE

I. MELUR

Serial No.	Location Code No.	Name of Village	Area of Village in Hectares	Occupied Residential Houses	No. of Households	Total Population including Institutional & Houseless Population			Scheduled Castes	
						P	M	F	M	F
(0)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	1	Pillaiyarachari	1,471·86	441	441	2,449	1,244	1,205	91	97
2	2	Sendurai	3,191·58	1,620	1,621	8,459	4,298	4,154	326	499
3	3	Sirappuimpatty	1,826·47	873	873	4,238	2,157	2,181	32	43
4		Kudaipatty	1,645·80	773	773	3,710	1,878	1,832	130	147
5	5	Selkur	2,748·26	918	919	4,269	2,422	2,347	126	138
6	6 a/b	Nedunagalam	1,211·50	480	482	2,414	1,199	1,215	69	94
7	6 b/b	Nedunagalam	2,171·85	731	731	3,686	1,892	1,794	403	566
8	7	Kottaiyur	3,057·06	474	474	2,256	1,120	1,136	51	65
9	8	Irugudi	3,846·68	1,485	1,485	7,343	3,608	3,735	535	524
10	9	Pettipatty	102·94	515	515	2,572	1,269	1,303	128	136
11	10	Soorapatty	32·33	135	135	723	363	360	54	49
12	11	Illambokanapatty	627·86	99	99	468	244	224	...	...
13	12	Pallapatti	665·27	440	440	2,063	1,017	1,049	105	106
14	13	Thenkanipagapatti	500·47	266	266	1,288	633	655	59	61
15	14	Chinnampatty	797·22	584	584	2,614	1,260	1,354	140	161
16	15	Valichampatty	138·92	234	234	1,049	490	559	77	73
17	16	Kottampatty	302·84	572	572	2,796	1,379	1,417	269	245
18	17	Chinnalangapatti	1,642·86	1,275	1,284	6,082	3,042	3,040	266	211
19	18	Sokkampatty	232·85	143	143	663	348	315	10	5
20	19	Kattangalagudi	1,125·16	918	918	4,225	2,088	2,137	276	279
21	20	Ayyampatty	1,257·48	808	808	3,706	1,893	1,813	177	208
22	21	Ayampanianagudy	157·37	284	290	1,299	675	624	79	88
23	22	Thurkaiady	39·30	34	34	178	88	90	11	1
24	23	Manappochari	1,094·98	781	781	3,572	1,725	1,847		35
25	24	Bothagudi	489·07	218	218	986	482	504	40	34
26	25	Kumarampatty	650·27	585	585	2,412	1,190	1,222	209	191
27	26	Samudrampatty	652·19	490	490	2,375	1,108	1,267	217	198
28	27	Pannaiyanai	618·31	98	98	514	233	281	8	11
29	28	Avichipatty	362·92	512	512	2,362	1,097	1,265	169	173
30	29	Unaipatty	963·61	450	450	2,122	1,625	1,497	140	154

441

PRIMARY CENSUS ABSTRACT

TALUK

Scheduled Tribes		Literate and Educated Persons		WORKERS: Total Workers (i+ii)		I Cultivators		II Agricultural Labourers	
M	F	M	F	M	F	M	F	M	F
(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
...	...	416	125	819	446	249	175	240	264
...	...	1,957	543	3,588	394	1,495	191	700	272
...	...	956	180	1,245	562	914	224	216	110
...	...	594	167	1,125	808	719	169	300	612
...	...	1,029	217	1,793	729	920	224	412	286
...	...	439	155	669	156	488	113	198	28
...	...	654	162	1,160	876	484	254	591	604
...	...	478	187	497	91	334	5	114	83
...	...	1,586	659	2,08	1,189	693	217	1,002	892
...	...	640	167	770	395	434	227	306	271
...	...	185	40	204	10	122	7	18	22
...	...	92	19	133	1	135	1	16	...
...	...	485	154	553	25	218	5	257	16
...	...	249	23	408	205	272	72	99	124
...	...	587	186	716	462	341	20	236	418
...	...	296	130	274	34	179	36	61	32
...	...	746	334	769	379	273	42	315	305
...	...	1,672	464	1,739	522	383	19	723	423
...	...	214	40	230	...	174	...	43	...
...	...	983	356	1,199	218	652	49	249	131
...	...	685	201	1,046	248	423	39	512	229
...	...	489	272	373	65	78	25	4	7
...	...	56	2	54	13	52	13	...	...
...	...	941	299	977	422	558	333	139	9
...	...	238	79	302	188	227	117	39	68
...	...	467	142	750	415	564	379	162	30
...	...	721	183	731	262	293	31	371	218
...	...	109	18	127	24	89	3	34	19
...	...	699	284	670	137	447	31	148	97
...	...	747	190	1,001	221	51	4	197	213

Figura 5.7: Dados relativos às trinta primeiras vilas listadas de Melur Taluk

O número total de domicílios ocupados em Melur Taluk é a soma do número de domicílios ocupados em cada uma de suas vilas, que é 53.812. Da Seção 5.2.3, tem-se que o número total estimado de domicílios ocupados no Distrito de Mdurai seja de 758.897. Portanto, o número total de vilas que devem ser visitadas em Melur Taluk é $\frac{758.897}{53.812} \cong 14$. Para encontrar o valor de k, definido na Seção 2.4, faz-se o quociente entre o número de domicílios em Melur e o número de entrevistas que devem ser realizadas no local, resultando em $k = \frac{53.812}{14} = 3.843,714$. Sorteou-se um número aleatório RN entre 1 e k e obteve-se $RN = 534,42$. Na Figura 5.7, uma coluna com a distribuição acumulada da quantidade de domicílios em cada vila deve ser feita para que se escolham as vilas que contenham o RN-*ésimo* domicílio, o $(RN + k)$-*ésimo* domicílio, e assim sucessivamente. Entretanto, é importante lembrar que os blocos

(que, neste caso, são as vilas) podem sofrer algumas alterações em seus tamanhos antes da seleção por causa de situações como as citadas na Seção 3.4.

Finalmente, utilizando os valores obtidos, as 14 vilas selecionadas para a amostra foram as de números $2, 5, 9, 18, 26, 34, 41, 47, 52, 57, 63, 74, 84, 90$.

5.2.8 Criação dos Segmentos

Supondo que a vila Pathinettangudi, representada na Figura 5.3, fosse selecionada, sabe-se que o local possui 396 domicílios ocupados. O número total de domicílios ocupados em Melur Taluk é 53.812. Portanto, o número total de domicílios ocupados que devem ser visitados em Pathinettangudi é $\frac{53.812}{396} \cong 136$. Neste caso, $k = \frac{396}{136} = 2,911$. Sorteou-se um número aleatório entre 1 e k e obteve-se $RN = 2,911$. Seguindo o procedimento de uma amostragem sistemática, estes domicílios que serão selecionados devem estar enumerados em "forma de serpente", como é possível observar nas Figuras 5.4, 5.5 e 5.6. Apesar dos domicílios das aldeias da vila estarem enumeradas separadamente, é possível que todos eles sejam enumeradas conjuntamente de 1 a 396 para a seleção.

5.2.9 Listagem e Início das Entrevistas

Finalmente, nesta última etapa os domicílios que devem ser visitados foram listados e, em cada um deles, foi aplicado o questionário da pesquisa à todas as pessoas elegíveis.

5.3 ESTUDO DE CASO: UNIVERSIDADE DE BRASÍLIA

Neste estudo, o objetivo foi observar a proporção de indivíduos do sexo feminino (e, portanto, também do sexo masculino) que compõe as salas de aula da Universidade de Brasília (UnB) no período matutino. Tratou-se de um estudo observacional e foram estimadas as proporções para os alunos. O estudo foi realizado para ilustrar como a amostragem por área pode ser aplicada desde o início.

5.3.1 Apresentação dos Mapas

Para o estudo, primeiro obteve-se uma imagem de satélite de toda a UnB, para que pudessem ser selecionados os conglomerados. Foram considerados os blocos majoritariamente dedicados à salas de aula, listados a seguir: Instituto Central de Ciências (ICC), Pavilhão Anísio Teixeira (PAT), Pavilhão João Calmon (PJC), Bloco Eudoro de Souza (BAES), Bloco de Salas de Aula Sul (BSAS) e Bloco de Salas de Aula Norte (BSAN). Tem-se então que o número total de blocos é $A = 6$. As Figuras 5.8 e 5.9 mostram como foi a delimitação da área de estudo e a Figura 5.10 mostra os blocos considerados no estudo.

Figura 5.8: Foto de satélite da Asa Norte

Figura 5.9: Foto de satélite da Universidade de Brasília

Figura 5.10: Blocos considerados no estudo em ênfase

Pela Figura 5.9, é possível observar alguns dos conceitos de amostragem por área apresentados na Seção 3.2. Observa-se que os blocos são bastante distintos entre si e a imagem de satélite mostra que as ruas são as características geográficas observáveis pelo pesquisador em campo que delimitam a área de cada um dos conglomerados, além dos próprios limites físicos dos edifícios.

5.3.2 Tamanho da Amostra

Solicitou-se alguns dados à secretaria do Departamento de Estatística, sobre o número de matrículas do segundo semestre de 2017, para que pudessem ser estimados as quantidades de salas de aula e alunos em cada um dos blocos. A Tabela 5.1 mostra estes dados. A probabilidade de seleção do local é dada por $Probabilidade = 3 \times \frac{Quantidade\ de\ salas\ do\ local}{Quantidade\ total\ de\ salas}$ e $Peso = \frac{1}{Probabilidade}$.

Para obter o tamanho da amostra foi utilizada a Equação (2.23), apresentada na

Tabela 5.1: Distribuição de salas de aula por localidade, com respectivas probabilidades de seleção e pesos amostrais (valores aproximados)

Local	Quantidade de Salas	Probabilidade de Seleção	Peso
BAES	1	0,016	63,333
BSAN	18	0,284	3,519
BSAS	45	0,711	1,407
ICC	65	1,026	0,974
PAT	30	0,474	2,111
PJC	31	0,489	2,043
Total	190		

Seção 2.2. Considerou-se o quantil da normal correspondente ao nível de significância de 5%, ou seja, $z_{\alpha/2} = 1,96$ para os cálculos. Além disto, o erro máximo tolerado foi $\epsilon = 0,22$ e o valor de $N = 190$ foi obtido através da Tabela 5.1. Tem-se então que:

$$n_{0propmax} = \frac{1,96^2 \times 0,25}{0,22^2} = 19,843.$$

Em seguida, foi utilizada a Equação (2.24) e obteve-se o seguinte tamanho de amostra:

$$n = \frac{19,843}{1 + \frac{19,843}{190}} = 17,966,$$

que foi aproximado para $n = 18$.

Note que $n = 18$ refere-se ao número total de salas que foram observadas na pesquisa, e não ao número de pessoas.

Além disso, por uma questão de tempo, foi definido que três blocos deveriam ser visitados, ou seja, $a = 3$.

5.3.3 Fração Amostral

Utilizadno os valores obtidos na seção anterior, tem-se que a fração amostral

global é $f = \frac{n}{N} = \frac{18}{190} = 0,094737$.

5.3.4 Análise da Amostra do Primeiro Estágio

Para a realização da amostra do primeiro estágio, as localidades (blocos) foram consideradas homogêneas entre si e, por isto, foi utilizada a técnica de amostragem por conglomerados, definida na Seção 2.5. Depois, para a seleção das salas de aula, foi utilizada a técnica de amostragem sistemática, definida na Seção 2.4. Como o estudo foi observacional, foram listadas todas as pessoas dentro das salas, ou seja, tratou-se de uma amostragem em dois estágios: seleção do bloco e depois seleção das salas e aula, não necessitando de um processo aleatório para a seleção das pessoas.

5.3.5 Preparação das Unidades Amostrais Primárias

Na Tabela 5.1, é possível observar que a probabilidade de seleção do ICC é maior do que 1, o que significa um absurdo de acordo com a definição de probabilidade. A interpretação deste valor é a de que o bloco ICC é muito grande, situação apresentada na Seção 3.4. Entretanto, serão conferidos se todos os blocos não possuem problemas de tamanho. De acordo com a Seção 3.4 e as Equações (3.3) e (3.4), um bloco será muito grande se $MoS_i > \frac{190}{3} = 63,333$ e será muito pequeno se $MoS_i < \frac{18}{3} = 6$, onde $MoS_i = \textit{Quantidade de salas no bloco } i$. Observa-se então que há dois blocos com problemas de tamanho: o BAES (muito pequeno) e o ICC (muito grande).

Para solucionar o problema, uniram-se o BAES e o BSAN num único bloco e o ICC foi dividido em dois menores. O critério de união do BAES com o BSAN foi geográfico, dado que estão próximos. O ICC foi dividido em lado A e lado B, seguindo a própria nomenclatura dada às suas salas. Os anfiteatros e as salas com

código CSS foram consideradas como parte do lado B do ICC.

Com esta nova divisão, obteve-se um novo conjunto de dados, apresentados na Tabela 5.2.

Tabela 5.2: Distribuição de salas de aula por localidade, com respectivas probabilidades de seleção e pesos amostrais (valores aproximados) para a nova divisão dos blocos

Local	Quantidade de Salas	Probabilidade de Seleção	Peso
BSAN (BAES)	19	0,300	3,333
BSAS	45	0,711	1,407
ICC lado A	26	0,411	2,436
ICC lado B	39	0,616	1,624
PAT	30	0,474	2,111
PJC	31	0,489	2,043
Total	190		

Precisa-se verificar agora se ainda há blocos muito grandes ou muito pequenos. Utilizando os mesmos cálculos anteriores, tem-se que o bloco será muito grande se $MoS_i > 63,333$ e muito pequeno se $MoS_i < 6$. Desta vez, observa-se que não há mais problemas com os tamanhos dos blocos.

Coincidentemente, o número total de blocos $A = 6$ foi mantido.

5.3.6 Seleção dos Blocos

Para a seleção dos blocos, eles foram enumerados de 1 a 6. A ordem foi escolhida para que, propositalmente, os blocos de interesse fossem selecionados. Embora isto não possa ser feito na prática - por questões éticas e também por conta do viés que pode ser gerado - neste caso este recurso foi utilizado, mas apenas para facilitar os resultados, a compreensão e também para fins ilustrativos. Porém, será suposto que a enumeração dos blocos foi feita através de um sorteio e a ordem encontrada é dada pela Tabela 5.3.

Tabela 5.3: Distribuição de salas de aula por localidade, organizadas pela ordem do sorteio

Número	Local	Quantidade de Salas	Quantidade Acumulada
1	BSAS	45	45
2	BSAN (BAES)	19	64
4	ICC lado A	26	90
5	PJC	31	121
6	ICC lado B	39	160
7	PAT	30	190

O número total de salas de aula é a soma do número de salas de aula em cada um dos blocos, que é 190. Tem-se que o número de blocos que devem ser visitados é $a = 3$. Para encontrar o valor de k, definido na Seção 2.4, faz-se o quociente entre o número total de salas de aula e o número de blocos que devem ser visitados, resultando em $k = \frac{190}{3} = 63,333$. Obteve-se um número RN entre 1 e k, dado por $RN = 50$. Normalmente o valor de RN é obtido aleatoriamente, mas, assim como na ordenação dos blocos, o valor de RN foi escolhido propositalmente, pelos mesmos motivos já citados anteriormente. Utilizando a coluna dos valores acumulados da Tabela 5.3, serão escolhidos os blocos que contenham a RN-*ésima* sala de aula, a $(RN + k)$-*ésima* sala de aula, e assim sucessivamente.

Finalmente, utilizando os valores obtidos, os 3 blocos selecionadas para a amostra foram o BSAN (BAES incorporado), o PJC e o PAT. Com os valores obtidos até esta etapa é possível calcular a probabilidade de seleção de cada uma das salas dos blocos selecionados. Sejam $PS1_i$ a probabilidade de seleção do bloco i, $PS2_i$ a probabilidade de seleção de cada uma das salas dentro do bloco i, $PSF = PS1_i \times PS2_i$ a probabilidade de selecionar uma das salas do bloco i e $Peso = \frac{1}{PSF}$. Com os valores obtidos, construiu-se a Tabela 5.4.

Tabela 5.4: Probabilidades de seleção dos blocos da amostra e suas respectivas salas de aula

Local	Quantidade de Salas	PS1	PS2	PSF	Peso
BSAN (BAES)	19	0,300	0,316	0,094737	10,559
PAT	30	0,474	0,200	0,094737	10,559
PJC	31	0,489	0,194	0,094737	10,559

Note que PS1 é o mesmo valor que foi encontrado anteriormente na Tabela 5.2. Para o local BSAN (BAES incorporado), por exemplo, obteve-se $PS1 = 3 \times \frac{19}{190} = 0,300$ e $PS2 = \frac{6}{19} = 0,316$.

Note também que as probabilidades de seleção de cada sala de aula é $PSF = f = \frac{n}{N} = 0,094737$, que é a fração amostral global. Este resultado deixa explícito que a amostra realizada encaixa-se no tipo de amostragem *EPSEM*, definida na Seção 3.3. Além disso, o produto entre os pesos encontrados e o número de salas que foram parte da amostra resulta no número de salas totais, ou seja, $10,559 \times 18 = 190$. Os valores dos PSF's foram encontrados para fins ilustrativos, pois o tipo de amostragem *EPSEM* caracteriza-se por selecionar o mesmo número de unidades no último estágio - que neste caso eram as 6 salas de aulas selecionadas em cada bloco.

Antes de prosseguir é necessário enfatizar que, mesmo que alguns valores tenham sido escolhidos por conveniência, é possível que eles sejam gerados aleatoriamente, embora seja pouco provável.

5.3.7 Seleção das Salas de Aula

Para a seleção das salas de aula foi utilizada a técnica de amostragem sistemática, apresentada na Seção 2.4. Primeiro, obteve-se uma imagem dos blocos selecionados, no caso, o BSAN, o PJC e o PAT. As Figuras 5.11 e 5.12 apresentam os desenhos dos

blocos sorteados. Uma imagem do BAES não foi necessária porque ele possui apenas uma sala e ela pode ser considerada como a primeira (ou a última) sala do BSAN. Talvez fosse possível obter imagens semelhantes à estas solicitando os projetos ou as plantas dos respectivos locais na universidade, mas os desenhos foram feitos à mão para evidenciar o fato de que não são necessárias quaisquer informações anteriores do local e que a técnica de amostragem por área pode ser utilizada em populações e áreas completamente desconhecidas.

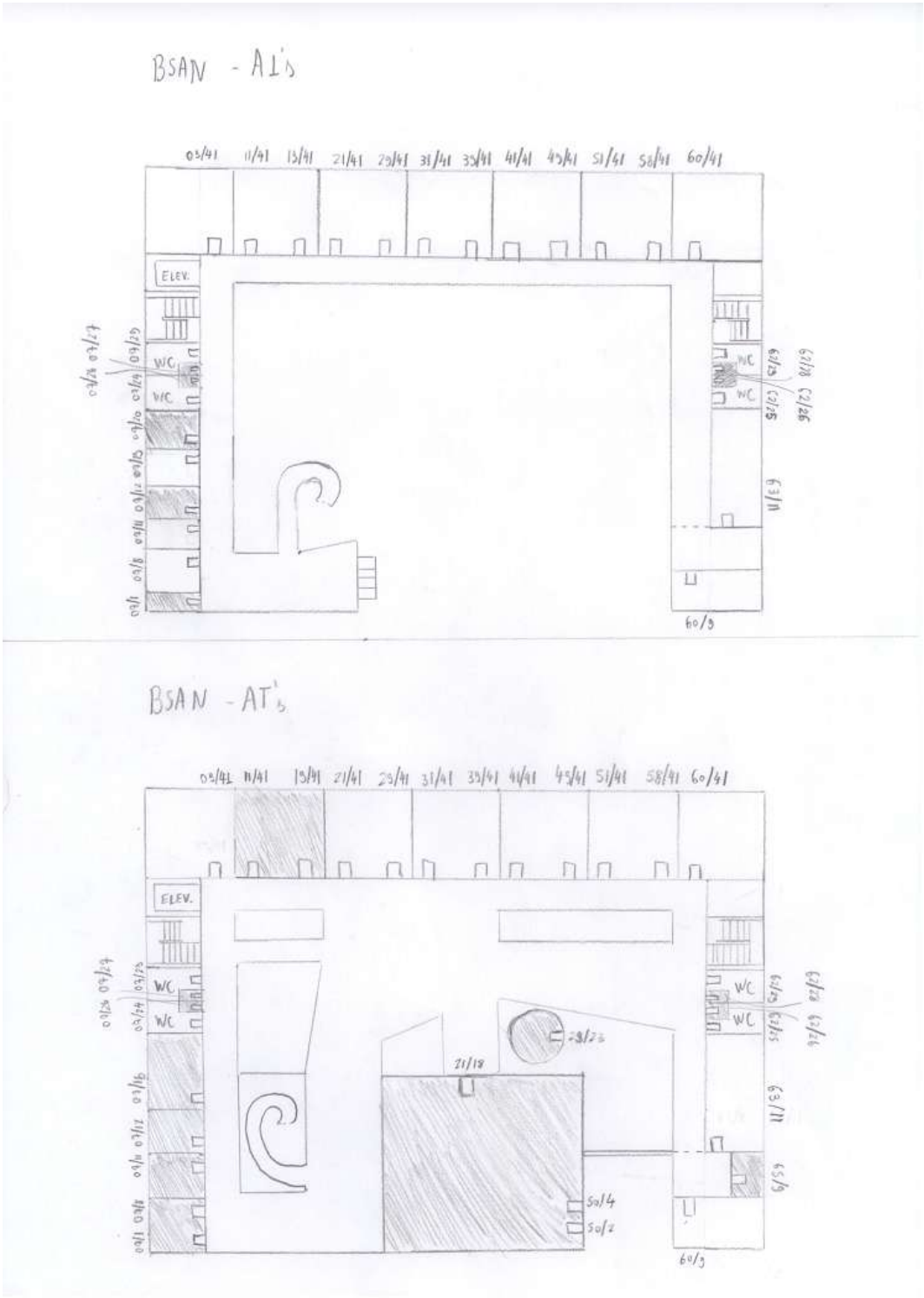


Figura 5.11: Desenho à Mão do Interior do BSAN

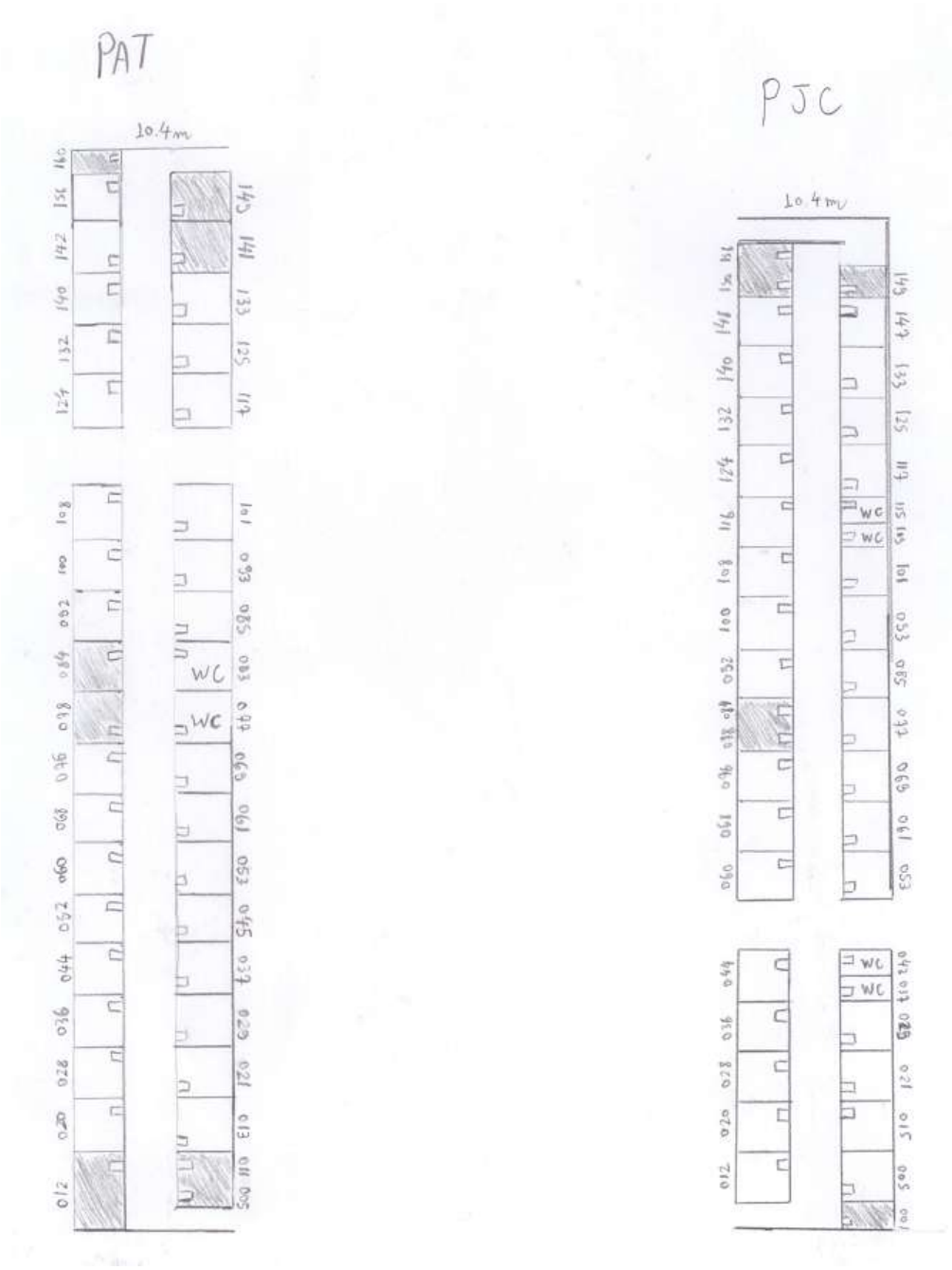


Figura 5.12: Desenho à Mão do Interior do PJC e do PAT

Utilizando os dados da Tabela 5.3 e a Equação (3.2), é possível descobrir os novos valores dos $b_i's$ de interesse. Obteve-se então que $b_i = 6$, ou seja, foram observadas 6 salas de aula em cada um dos blocos selecionados.

5.3.8 Criação dos Segmentos

Com as imagens e os valores obtidos na Seção 5.3.7, o próximo passo foi definir os segmentos. Primeiramente, foram enumeradas as salas dos blocos selecionados de acordo com a técnica de amostragem sistemática. Esta enumeração é apresentada nas Figuras 5.13 e 5.14. A única sala do BAES foi considerada como a de número 19 do bloco BSAN (BAES).

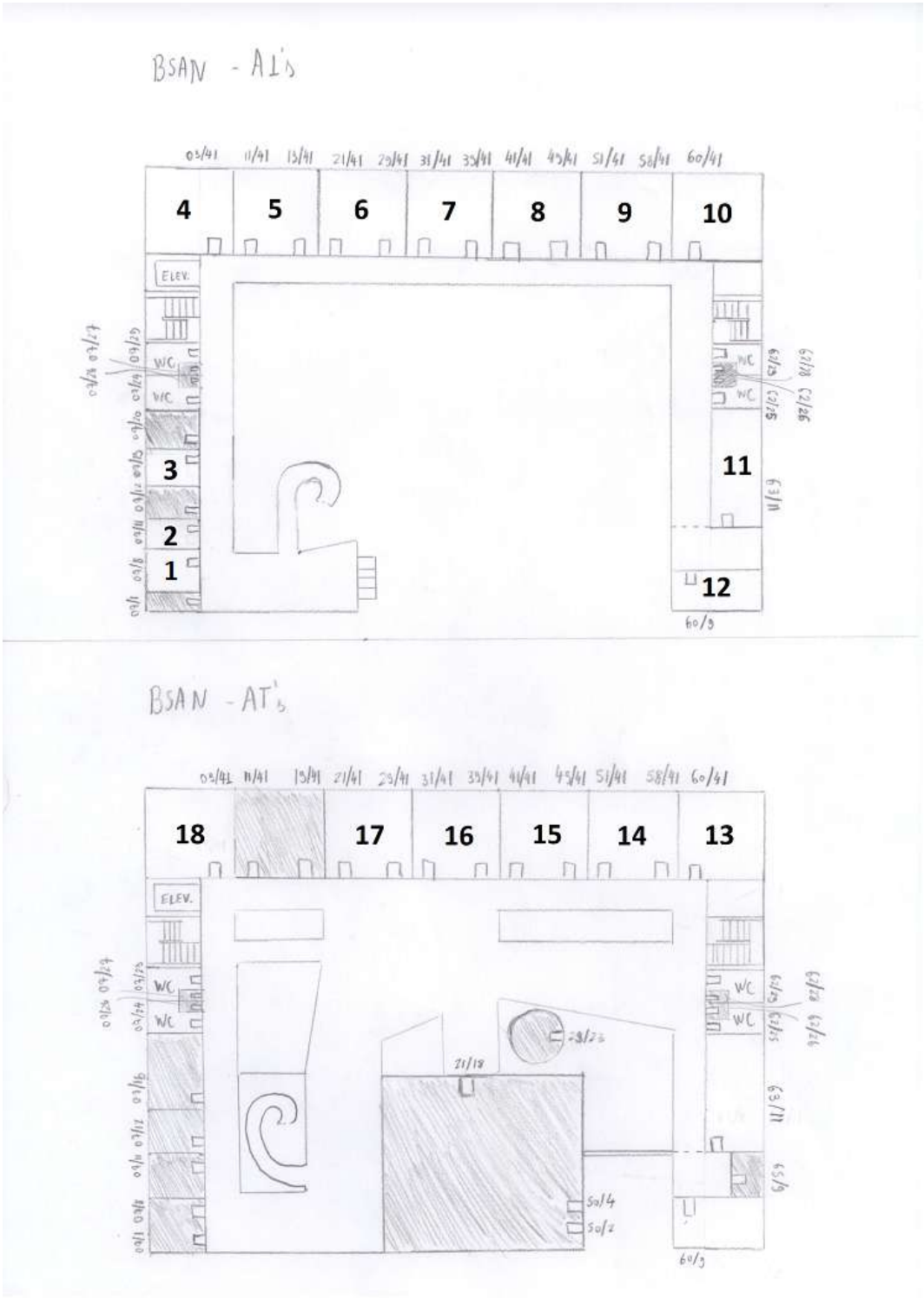


Figura 5.13: Enumeração das salas do BSAN

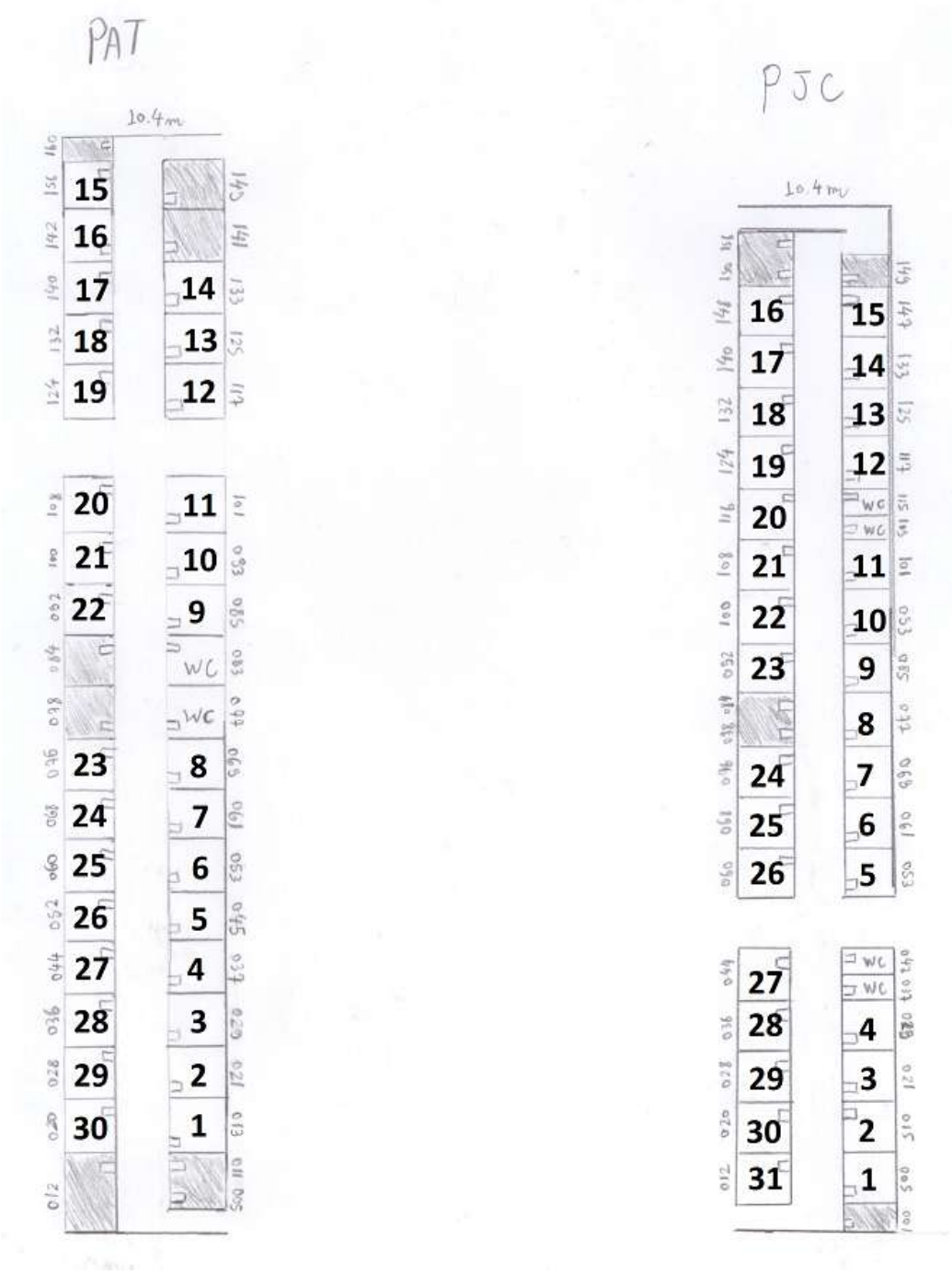


Figura 5.14: Enumeração das salas do PJC e do PAT

Em seguida, foram obtidos aleatoriamente os números das salas que seriam visitadas. Para isto, ainda seguindo as definições da Seção 2.4, primeiro foram obtidos os $k's$ de cada um dos blocos. Foram obtidos os valores:

- $k_{BSAN} = \frac{19}{6} = 3,17$;
- $k_{PAT} = \frac{30}{6} = 5$;
- $k_{PJC} = \frac{31}{6} = 5,17$.

Em seguida, foram obtidos os $RN's$ para cada bloco. Para isto, foi utilizada a tabela de números aleatórios encontrada em RAND (2001). Para os blocos em que o valor de k apresentou um resultado com casas decimais (k_{BSAN} e k_{PJC}), os números da tabela foram vistos de três em três, sendo que o primeiro número da sequência se refere à unidade e os dois seguintes às duas primeiras casas decimais. Note que os valores de k_{BSAN} e k_{PJC} foram arredondados para cima, para que fosse possível selecionar a última sala de cada bloco. Para o outro bloco, os números da tabela foram vistos de um em um. Para o BSAN (BAES incorporado), o PAT e o PJC foram selecionadas as linhas 19950, 19951 e 19952 da tabela, respectivamente. A Figura 5.15 mostra os números selecionados.

400 TABLE OF RANDOM DIGITS

19950	92630	78240	19267	95457	53497	23894	37708	79862	76471	66418
19951	79445	78735	71549	44843	26104	67318	00701	34986	66751	99723
19952	59654	71966	27386	50004	05358	94031	29281	18544	52429	06080
19953	31524	49587	76612	39789	13537	48086	59483	60680	84675	53014
19954	06348	76938	90379	51392	55887	71015	09209	79157	24440	30244
19955	28703	51709	94456	48396	73780	06436	86641	69239	57662	80181
19956	68108	89266	94730	95761	75023	48464	65544	96583	18911	16391
19957	99938	90704	93621	66330	33393	95261	95349	51769	91616	33238
19958	91543	73196	34449	63513	83834	99411	58826	40456	69268	48562
19959	42103	02781	73920	56297	72678	12249	25270	36678	21313	75767
19960	17138	27584	25296	28387	51350	61664	37893	05363	44143	42677
19961	28297	14280	54524	21618	95320	38174	60579	08089	94999	78460
19962	09331	56712	51333	06289	75345	08811	82711	57392	25252	30333
19963	31295	04204	93712	51287	05754	79396	87399	51773	33075	97061
19964	36146	15560	27592	42089	99281	59640	15221	96079	09961	05371
19965	29553	18432	13630	05529	02791	81017	49027	79031	50912	09399
19966	23501	22642	63081	08191	89420	67800	55137	54707	32945	64522
19967	57888	85846	67967	07835	11314	01545	48535	17142	08552	67457
19968	55336	71264	88472	04334	63919	36394	11196	92470	70543	29776
19969	10087	10072	55980	64688	68239	20461	89381	93809	00796	95945
19970	34101	81277	66090	88872	37818	72142	67140	50785	21380	16703
19971	53362	44940	60430	22834	14130	96593	23298	56203	92671	15925
19972	82975	66158	84731	19436	55790	69229	28661	13675	99318	76873
19973	54827	84673	22898	08094	14326	87038	42892	21127	30712	48489
19974	25464	59098	27436	89421	80754	89924	19097	67737	80368	08795
19975	67609	60214	41475	84950	40133	02546	09570	45682	50165	15609
19976	44921	70924	61295	51137	47596	86735	35561	76649	18217	63446
19977	33170	30972	98130	95828	49786	13301	36081	80761	33985	68621
19978	84687	85445	06208	17654	51333	02878	35010	67578	61574	20749
19979	71886	56450	36567	09395	96951	35507	17555	35212	69106	01679
19980	00475	02224	74722	14721	40215	21351	08596	45625	83981	63748
19981	25993	38881	68361	59560	41274	69742	40703	37993	03435	18873
19982	92882	53178	99195	93803	56985	53089	15305	50522	55900	43026
19983	25138	26810	07093	15677	60688	04410	24505	37890	67186	62829
19984	84631	71882	12991	83028	82484	90339	91950	74579	03539	90122
19985	34003	92326	12793	61453	48121	74271	28363	66561	75220	35908
19986	53775	45749	05734	86169	42762	70175	97310	73894	88606	19994
19987	59316	97885	72807	54966	60859	11932	35265	71601	55577	67715
19988	20479	66557	50705	26999	09854	52591	14063	30214	19890	19292
19989	86180	84931	25455	26044	02227	52015	21820	50599	51671	65411
19990	21451	68001	72710	40261	61281	13172	63819	48970	51732	54113
19991	98062	68375	80089	24135	72355	95428	11808	29740	81644	86610
19992	01788	64429	14430	94575	75153	94576	61393	96192	03227	32258
19993	62465	04841	43272	68702	01274	05437	22953	18946	99053	41690
19994	94324	31089	84159	92933	99989	89500	91586	02802	69471	68274
19995	05797	43984	21575	09908	70221	19791	51578	36432	33494	79888
19996	10395	14289	52185	09721	25789	38562	54794	04897	59012	89251
19997	35177	56986	25549	59730	64718	52630	31100	62384	49483	11409
19998	25633	89619	75882	98256	02126	72099	57183	55887	09320	73463
19999	16464	48280	94254	45777	45150	68865	11382	11782	22695	41988

Figura 5.15: Tabela de números aleatórios com os RN's obtidos para cada um dos blocos selecionados

Tem-se então que $RN_{BSAN} = 3,07$, $RN_{PAT} = 4$ e $RN_{PJC} = 1,96$. Portanto, as

salas selecionadas, de acordo com as enumerações das Figuras 5.13 e 5.14, foram:

- BSAN (BAES incorporado): 3, 6, 9, 12, 15 e 18;
- PAT: 4, 9, 14, 19, 24 e 29;
- PJC: 1, 7, 12, 17, 22 e 27.

Note que as salas de aula estão aleatoriamente distribuídas dentro dos blocos, ou seja, não há um número fixo anteriormente definido de alunos do sexo feminino ou masculino em cada uma ou correlação entre as salas visinhas. Segundo Cochran (1977), esta situação é suficiente para satisfazer o teorema que valida a igualdade $E(V_{Sis}) = E(V_{AAS})$, onde V_{Sis} é a variância do estimador da média em amostragem sistemática e V_{AAS} é a variância do estimador da média em AAS.

Note também que, se fosse um estudo completo, os horários das aulas seriam o terceiro estágio da amostra, ou seja, além de sortear o bloco e as salas que seriam visitadas, o horário deveria ser sorteado também, nos casos em que a sala possuísse aula às 8 horas e às 10 no mesmo dia. Entretanto, as salas foram visitadas em horários convenientes para o pesquisador, por se tratar de um estudo com fins ilustrativos.

5.3.9 Listagem e Início das Visitas

Finalmente, agora que as salas de aula que deveriam ser observadas foram listadas, foram feitas as observações. Os resultados são apresentados na Tabela 5.5. Para não ter que lidar com o fato de interromper as aulas ou depender da autorização dos

professores para fazer a contagem, as observações foram feitas do lado de fora das salas, pela janela da porta ou diretamente, nos casos em que a porta estava aberta.

Tabela 5.5: Quantidades observadas de alunos do sexo feminino e masculino para as salas da amostra

Bloco	Código da sala	Sexo feminino	Sexo masculino
BSAN (BAES)	A1 07/19	2	0
BSAN (BAES)	A1 21/41 - A1 29/41	10	27
BSAN (BAES)	A1 51/41 - A1 58/41	32	28
BSAN (BAES)	A1 60/9	4	10
BSAN (BAES)	AT 41/41 - AT 45/41	16	23
BSAN (BAES)	AT 09/41	19	18
PAT	AT 037	8	11
PAT	AT 085	8	6
PAT	AT 133	15	18
PAT	AT 124	18	13
PAT	AT 068	3	4
PAT	AT 028	2	18
PJC	BT 005	16	2
PJC	BT 069	9	12
PJC	BT 117	2	21
PJC	BT 140	2	4
PJC	BT 100	14	5
PJC	BT 044	0	2
Total		180	222

As salas A1 07/19, A1 21/41 - A1 29/41, A1 51/41 - A1 58/41, A1 60/9, AT 09/41, AT 037, AT 024, AT 068, AT 028, BT 117, BT 100 e BT 44 foram visitadas no dia 18/10/2017, durante a aula das 10 horas.

As salas AT 068 e BT 140 foram visitadas no dia 20/10/2017, durante a aula das 8 horas.

As salas AT 085 e BT 069 foram visitadas no dia 20/10/2017, durante a aula das 10 horas.

As salas AT 41/41 - AT 45/41 e BT 005 foram visitadas no dia 30/10/2017, durante a aula das 10 horas.

A sala A1 60/9 estava fechada no dia da visita e não havia janela ou qualquer outra abertura em sua porta. Para a contagem, esperou-se o término da aula e foram feitas as observações à medida que as pessoas deixavam o local.

Foram observadas as seguintes quantidades na amostra:

- Número de alunos do sexo feminino: 180;
- Número de alunos do sexo masculino: 222;

5.3.10 Resultados Para o Estudo na UnB

Como tratou-se de um estudo com fins ilustrativos, houve a vantagem de conhecer os valores populacionais, permitindo a comparação das estimativas obtidas na amostra com seus respectivos parâmetros. A quantidade de alunos na população era de 28.452, a qual 14.566 eram do sexo feminino e 13.886 do sexo masculino. Portanto, as proporções populacionais obtidas foram $FEMININO = \frac{14.556}{28.452} = 0,512$ e $MASCULINO = \frac{13.886}{28.452} = 0,488$.

Observou-se um total de 402 alunos na amostra. A proporção amostral pertencente ao sexo feminino foi de $feminino = \frac{180}{402} = 0,448$ e, portanto, para o sexo masculino foi de $masculino = 0,552$. Supondo que foi utilizada uma AAS, as variâncias estimadas das proporções amostrais seriam encontradas utilizando a Equação (2.19) e o valores obtidos seriam $\widehat{V_{AAS}}(aluno) = \widehat{V_{AAS}}(feminino) = \widehat{V_{AAS}}(masculino) = 0,000617$. Entretanto, deve-se utilizar a Equação (2.54), própria para os casos de amostragem por conglomerados. Desconsiderando a correção para populações finitas, tem-se então as variâncias $\widehat{V_{Cong}}(aluno) = \widehat{V_{Cong}}(feminino) = \widehat{V_{Cong}}(masculino) = 0,002071$.

Portanto, utilizando a Equação (2.55), tem-se que os intervalos de confiança são:

$$IC(FEMININO, 95\%) = [0,352; 0,544];$$

$$IC(MASCULINO, 95\%) = [0,456; 0,648].$$

Para os blocos destinados majoritariamente à salas de aula, conclui-se a partir da amostra que a proporção de alunos do sexo feminino e do sexo masculino não são diferentes, com nível de confiança de 95%, pois os respectivos intervalos de confiança se cruzam.

Observa-se também que a proporção de alunos do sexo feminino observada na amostra foi menor do que a respectiva proporção populacional, provavelmente porque os blocos visitados possuem mais aulas de disciplinas voltadas aos cursos da área de exatas, que normalmente possuem mais alunos do sexo masculino. Além disso, um dos três blocos visitados necessariamente deveria ser um dos lados do ICC, pois o ICC era um conglomerado muito grande. Deixar de visitá-lo também pode ter causado este pequeno viés na amostra.

Para fins de comparação, pode-se obter a medida *deff*, apresentada na Equação (2.35), realizando o quociente $\frac{\widehat{V_{Cong}}(aluno)}{\widehat{V_{AAS}}(aluno)}$. Tem-se então que $deff = \frac{0,002071}{0,000617} = 3,35$, ou seja, a variância obtida através da amostragem por conglomerados, neste caso, foi cerca de 3,35 vezes maior do que a variância que seria obtida a partir de uma AAS.

Capítulo 6

CONCLUSÕES

O foco do presente trabalho foi sistematizar a técnica de amostragem por área, apresentando todos os seus principais prociedimentos e características. Observando os estudos de caso, aplicados à dados reais, observou-se a eficiência do método até mesmo em situações onde não há registros anteriores. Observou-se também uma atualização do método, pois as imagens do exterior dos blocos puderam ser obtidas através de imagens de satélite, um recuso mais moderno e recente.

Kish (1965) colocou que a amostragem por área é uma técnica eficiente para a obtenção de registros populacionais que podem ser utilizados no futuro. No estudo de caso da Índia foi possível observar que, de fato, para cada uma das divisões administrativas do distrito de Madurai, foram listadas todas as vilas, mesmo as que não foram visitadas, obtendo assim um registro que pode ser usado futuramente.

Lepkowski e Heeringa (2016) mostraram também que a técnica de amostragem por área é eficiente em relação aos custos da pesquisa, dada a inviabilidade de realizar um procedimento completamente aleatório em locais geograficamente grandes. Em ambos estudos de caso do Capítulo 5 observou-se que a realização da pesquisa em toda a área de estudo demandaria muito tempo e muitas despesas, sendo necessários

os desenhos de todos os blocos, deslocar-se ao longo de toda a região, entre outros.

No estudo de caso da UnB, foi possível observar que os resultados amostrais aproximaram-se bastante dos valores populacionais previamente obtidos, evidenciando ainda mais a eficiencia da técnica de amostragem por área. Tanto na população quanto na amostra, observou-se que a proporção de alunos do sexo feminino é bastante semelhante à do sexo masculino. Além disso, foi possível observar a eficiência do tipo de amostragem *EPSEM* e suas características, uma vez que a mesma quantidade de salas foram selecionadas em cada bloco para a amostra.

Finalmente, a técnica de amostragem por área pôde ser aplicada desde o início e com todas as etapas, como apresentou o estudo de caso da UnB, atribuindo também probabilidades não nulas de seleção à todos os elementos da população, como foi explicado na Seção 3.1.

Referências Bibliográficas

Bussab, W. O. & Morettin, P. A. (2002). *Estatística Básica*, (5ª ed.). São Paulo: Editora Saraiva.

Cochran, W. G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley & Sons.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Lepkowski, J. & Heeringa, S. (2016). Workshop in sampling techniques, University of Michigan, EUA. Notas de Aula.

RAND (2001). *A Million Random Digits With 100,000 Normal Deviantes*. Glencoe: RAND.

SAS (2013). On Line DOC, Version 9.4. Cary. North Carolina: SAS Institute. Inc., Version 9.4.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Pratical Tools for Designing and Weighting Survey Samples*. New York: Springer.