Avaliação metodológica das pesquisas eleitorais brasileiras

Neale Ahmed El-Dash

Tese apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Doutor em Ciências

Programa: Doutorado em Estatística

Orientador: Prof. Dr. Sérgio Wechsler

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES/CNPq

São Paulo, 14 de dezembro de 2010

Avaliação metodológica das pesquisas eleitorais brasileiras

Este exemplar corresponde à redação
final da tese devidamente corrigida
e defendida por Neale Ahmed El-Dash
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Prof. Dr. Sérgio Wechsler (Presidente) - IME-USP.
- Prof. Dr. Carlos Alberto de Bragança Pereira - IME-USP.
- Prof. Dr. Josemar Rodrigues - UFScar.
- Prof. Dr. Francisco Louzada Neto - UFScar.
- Prof. Dr. Hélio dos Santos Migon - UFRJ.

Agradecimentos

Aos professores Sérgio e Carlinhos pelo apoio e, principalmente, por permitirem que eu seguisse meu próprio caminho.

Ao Clifford Young, por me ensinar sobre amostragem de populações humanas na prática, e por me incentivar a estudar o tema com mais profundidade.

À Lara, que mesmo sem entender do assunto, sempre prestou muita atenção nas minhas discussões sobre o tema da tese. Mas principalmente, por sempre apoiar e acreditar em mim, incondicionalmente. E também, é claro, pelo presente de Natal inesquecível (conceitual)!

À minha mãe e as minhas irmãs, por tudo, mas em especial por terem vindo a São Paulo assistir a minha defesa de tese. Esse realmente foi o melhor presente de Natal possível...

Aos meus amigos "estatísticos", em especial a Roberta, a Valéria, ao Brutus e ao Mariu, com os quais tive diversas discussões estatísticas (e futebolísticas!) as quais me ajudaram a amadurecer diversos argumentos utilizados nessa tese.

Aos meus amigos "não-estatísticos", em especial ao Maurício e ao Piteco, pelas animadas noites de terça-feira jogando Texas Hold'em. E, por mais incrível que pareça para eles, por me ajudarem a pensar sobre probabilidades condicionais...

À banca examinadora, pelas sugestões e correções.

À CAPES e ao CNPq pelo suporte financeiro.

Resumo

Nessa tese examinaremos a metodologia de amostragem e de inferência das pesquisas eleitorais feitas no Brasil. Um dos desenhos amostrais mais utilizados pelos institutos de pesquisa é a Amostragem Probabilística com Cotas. Esse desenho é de 2 estágios, onde no primeiro estágio selecionam-se conglomerados, usualmente setores censitários[1], e no segundo estágio, selecionam-se os entrevistados de maneira não-probabilística, através de cotas.

Esse desenho amostral é muito criticado no meio acadêmico pois nele não é possível calcular as probabilidades de inclusão π_i para todas as pessoas entrevistadas, e consequentemente não é possível obter as estimativas de quantidades de interesse usualmente recomendadas pela teoria de populações finitas.

O objetivo desse trabalho é apresentar uma justificativa teórica para amostragem probabilística com cotas e compará-la, do ponto de vista de inferência baseada no desenho **(ID)**, com um desenho totalmente probabilístico equivalente. A utilização do modelo Grupos de Resposta Homogênea (**GRH**) para modelar explicitamente as probabilidades de resposta individuais permite o uso dos estimadores usuais.

O mesmo modelo para as probabilidades de resposta possibilita também calcular as probabilidades de inclusão para o caso da amostragem probabilística, permitindo assim que ambos os desenhos amostrais sejam comparados sob as mesmas suposições. Para representar com mais precisão a amostragem probabilística na prática, foram incluídos nesse modelo dois parâmetros: κ_1 e κ_2, que determinam quantas tentativas serão feitas pelo entrevistador para fazer contato com o domicílio e com o morador selecionado, respectivamente.

Essa comparação será feita utilizando o erro quadrático médio (**EQM**) e o tempo até o término da coleta de dados (número de contatos). Serão comparados diferentes estimadores da probabilidade de resposta para cada um dos desenhos amostrais estudados. Também será feita uma avaliação empírica da qualidade da previsão de 898 pesquisa eleitorais realizadas no Brasil, entre os anos de 1989 e 2004.

Palavras-chave: Pesquisas Eleitorais, Amostragem por Cotas, Amostragem Probabilística, Erro de Não-Resposta, Inferência baseada no Desenho.

[1]Setor Censitário é a menor unidade geográfica para a qual existem informações oficiais do IBGE disponíveis.

Abstract

In this thesis we examine the sampling and inference methodology of polls taken in Brazil. One of the sampling designs most used by research institutes is the Probability Sampling with Quotas. This sample has two stages, where in the first stage clusters are selected, usually census tracts[2], and in the second stage, the selection of the actual respondents is done in a non-probabilistic form, using quotas.

This sampling design is very criticized in the academic world because it doesn't allow the inclusion probabilities π_i for all respondents to be calculated, and therefore it is not possible to obtain the estimates recommended by theory of finite populations of the usual quantities of interest.

The aim of this paper is to present a theoretical justification for probability sampling with quotas and compare it, from the point of view of design-based inference (**DI**), with an equivalent fully probabilistic design. The use of the response homogeneity group model (**RHG**) to explicitly model the probabilities of individual response allows the use of the estimators described above.

The same model for the probabilities of response allows calculation of the inclusion probabilities for the case of probabilistic sampling, thus allowing both sample designs to be compared under the same assumptions. To represent more accurately the probabilistic sampling in practice, two parameters were included in this model: κ_1 and κ_2, which determine how many attempts will be made by the interviewer to make contact with the selected household and resident, respectively.

This comparison will be done using the mean square error (**MSE**) and the time it takes to finish the collection of data (number of contacts). Different estimators of the probability of response for each of the studied sampling designs are compared. Also, an empirical assessment of the quality of the prediction of 898 electoral surveys conducted in Brazil between the years 1989 and 2004 is presented.

Keywords: Political Polls, Quota Sampling, probabilistic sampling, Non-Response error, Design-based Inference.

[2]Census Tract is the smallest geographical area where official information is available.

Sumário

Lista de Abreviaturas

AAS	Amostragem Aleatória Simples (*Simple Random Sample*).
AASc	Amostragem Aleatória Simples com Reposição (*Simple Random Sample with replacement*).
AASs	Amostragem Aleatória Simples sem Reposição(*Simple Random Sample without replacement*).
AC	Amostragem por Cotas (*Quota Sampling*).
AP	Amostragem Probabilística (*Probabilistic Sampling*).
APC	Amostragem Probabilística com Cotas (*Probabilistic Sampling with Quotas*).
APV	Amostragem Probabilística com Voltas (*Probabilistic Sampling with Callback's*).
APVS	Amostragem Probabilística com Voltas Simples (*Simple Probabilistic Sampling with Callback's*).
BN	Estimador baseado na Distribuição Binomial Negativa (*Negative-binomial distribution based Estimator*).
C	Estimador baseado no número de contatos e de pessoas abordadas (*Number of Contacts based Estimator*).
EM	Estimador do tipo EM para APV(*EM-type Estimator for APV*).
EMV	Estimador de Máxima-Verossimilhança (*Maximum-Likelihood Estimator*).
EPA	Efeito do planejamento amostral(*Sample Design Effect*).
EQM	Erro Quadrático Médio (*Mean Square Error*).
GRH	Grupos de Resposta Homogênea (*Homogeneous Response Groups*).
GT	Estimador baseado na Distribuição Geométrica Truncada (*Truncated Geometric distribution based Estimator*).
GTS	Estimador simplificado baseado na Distribuição Geométrica Truncada (*Truncated Negative-binomial distribution based Estimator*).
HH	Estimador de Hansen-Hurwitz (*Hansen-Hurwitz estimator*).
HT	Estimador de Horvitz-Thompson (*Horvitz-Thompson estimator*).
IBM	Inferência Bayesiana baseada no Desenho (*Bayesian Model-based inference*).
ID	Inferência baseada no Desenho (*Design-based inference*).
IM	Inferência baseada no Modelo (*Model-based inference*).
TSE	Tribunal Superior Eleitoral (*Superior Electoral Court*).

Lista de Símbolos

N	Número de unidades na população.
H	Número de estratos na amostra.
A	Número de unidades primárias na população.
k	k-ésima unidade primária.
n	Tamanho total da amostra.
a	Tamanho da amostra do primeiro estágio.
b	Tamanho da amostra do segundo estágio (conglomerado k).
b_h	Tamanho da amostra no estrato h do conglomerado k.
Y	Variável de interesse
Y_i	Valor da variável de interesse Y para a unidade i.
Y_{hi}	Valor da variável de interesse Y para a unidade i do estrato h.
Y_{khi}	Valor da variável de interesse Y para a unidade i do conglomerado k do estrato h.
τ_y	Total populacional da variável Y.
τ_k	Total populacional da variável Y no conglomerado k.
τ_k^h	Total populacional da variável Y no conglomerado k no estrato h.
$\tau_{\hat{HH}}$	Estimador de Hansen-Hurwitz do total populacional.
$\hat{\tau}_k$	Estimador do total populacional no conglomerado k.
$\hat{\tau}_k^h$	Estimador do total populacional no estrato h no conglomerado k.
$Var\left(\tau_{\hat{HH}}\right)$	Variância do estimador do total populacional.
$\hat{Var}\left(\tau_{\hat{HH}}\right)$	Estimador da Variância do estimador do total populacional.
V_{hk}^I	Variância do estimador do total populacional τ_k^h.
$V_{kk'}^E$	Parcela da variância Entre-Conglomerados do estimador de $\tau_{\hat{HH}}$.
$\hat{V}_{hk}^I$	Estimador da variância do estimador do total populacional τ_k^h.
p_i	Prob. da i-ésima unidade populacional ser incluída na pesquisa (sem reposição).
pi_i	Prob. da i-ésima unidade populacional ser incluída na pesquisa (com reposição).
p_i^{selec}	Prob. da i-ésima unidade populacional ser selecionada em um único sorteio.
p_i^{Resp}	Prob. da i-ésima unidade populacional responder.
D_k	Quantidade de domicílios no conglomerado k.
N_k	Quantidade de moradores no conglomerado k.
$D_{j,k}$	j-ésimo domicílio do conglomerado k.
N_k^h	Quantidade de moradores do estrato h do conglomerado k.
$N_{j,k}^h$	Quantidade de moradores do domicílio j do estrato h do conglomerado k.
$N_{j,k}$	Quantidade de moradores do domicílio j do conglomerado k.

π_k	Prob. de inclusão do conglomerado k.
$p_{hi/k}$	Prob. de selecionar a unidade i do estrato h, do conglomerado k, em um único sorteio.
$\pi_{kk'}$	Prob. de inclusão conjunta dos conglomerados k e k'.
s_I	Conjunto das unidades primárias pertencentes a amostra.
s_k^h	Conjunto das unidades secundárias do estrato h pertencentes a sub-amostra do cong. k.
p_k^h	Prob. de resposta do estrato h pertencente ao conglomerado k.

Lista de Figuras

Lista de Tabelas

Introdução

As pesquisas de opinião pública têm como objetivo avaliar a opinião das pessoas com respeito aos mais diversos temas, como por exemplo, no mundo corporativo é importante conhecer essas opiniões para desenvolver um produto que o consumidor goste e compre, descobrir se ele está satisfeito com uma empresa e se ele compra produtos da concorrência; no meio acadêmico existe o interesse em entender como as pessoas pensam e como elas reagem a determinadas situações; durante as eleições é importante para os candidatos entenderem quais são as necessidades e desejos do eleitorado para que possam definir suas plataformas de governo.

Essas pesquisas são usualmente realizadas por empresas especializadas, denominadas empresas de pesquisa. As pesquisas de opinião são baseadas em uma amostra, ou seja, somente numa parcela da população de interesse. As pessoas pertencentes a essa amostra usualmente respondem a questionários, formulários, ou são entrevistadas sobre o tema de interesse. Após coletadas, essas informações são utilizadas para explicar a opinião de todas as pessoas da população de interesse, e não somente daquelas pessoas pertencentes a amostra, num processo denominado inferência.

Quando essas pesquisas têm como objetivo saber como a população votará em uma eleição, elas são denomindas de pesquisas eleitorais. Ao serem divulgadas na mídia, o público conhece qual candidato têm a preferência do eleitorado naquele determindado instante de tempo, permitindo que cada pessoa possa usar essa informação na escolha racional de seu candidato. Além disso, elas também afetam aos candidatos e as suas campanhas, seja pela confiança e exposição na mídia adquirida pelos que estão na liderança, ou pela desmotivação daqueles que não têm chances reais de ganhar a eleição.

Pesquisas eleitorais, mesmo quando não divulgadas na imprensa, também servem para auxiliar os candidatos a definirem suas estratégias de campanha e suas plataformas de governo, ao conhecer os principais problemas da população e entender quais áreas sócio-econômicas são prioridades do eleitorado. Por causa do impacto que as pesquisas eleitorais podem ter nas eleições, a divulgação das mesmas na mídia é regulamentada pelo Tribunal Superior Eleitoral (TSE).

Existe muita polêmica com relação às pesquisas eleitorais realizadas no Brasil. Essa polêmica é fruto, em parte, da grande exposição na mídia que essas pesquisas recebem durante as eleições (de 2 em 2 anos), pelo fato desse ser um dos únicos cenários reais onde é possível verificar se pesquisas de opinião conseguiram "prever corretamente"[3] o resultado e pelos diversos interesses políticos

[3] A expressão "prever corretamente"está entre aspas pois pode ser interpretada de diferentes maneiras, como acertar a ordem de preferência aos candidatos na eleição, acertar quem ganhará as eleições ou prever dentro das margens de erro os percentuais obtidos por cada candidato, entre outras.

envolvidos.

Do ponto de vista dos estatísticos, também há muita polêmica envolvendo as pesquisas eleitorais, principalmente porque **elas representam um encontro da teoria com a prática**. A polêmica surge porque as metodologias utilizadas pelos institutos de pesquisa tendem a se preocupar principalmente com o lado prático da pesquisa, já os acadêmicos tendem a se preocupar apenas com a teoria. Não é possível fazer pesquisas eleitorais ignorando a teoria, porém também não é possível fazê-las sem considerar o lado prático. Citando uma frase do filósofo John Dewey: "A prática sem a teoria é cega. E a teoria sem a prática é vazia".

Os aspectos metodológicos fundamentais das pesquisas de opinião são naturalmente agrupados em duas categorias:

Obtenção dos dados: Inclui todas as etapas do processo de obtenção dos dados. Essas etapas são: planejamento da amostra, escolha da amostra e coleta dos dados. É importante ressaltar que, na prática, as amostras são planejadas levando-se em conta as dificuldades logísticas da coleta dos dados, com o objetivo de diminuir custo e tempo na obtenção dos dados e reduzindo o mínimo possível a qualidade dos mesmos.

Inferência: Após observados os dados, usualmente deseja-se fazer inferências sobre a população sendo estudada. Essa categoria inclue todas as etapas necessárias para se transformar a informação da amostra em informação sobre a população.

Na maioria dos institutos de pesquisa do Brasil, essas duas etapas são efetivamente independentes, ou seja, usualmente não é levado em consideração na etapa de inferência o desenho amostral utilizado. Essa é uma das críticas mais sérias feita pelos acadêmicos, pois é difícil dizer o impacto que essa omissão pode ter nas inferências feitas para toda a população.

Outro aspecto bastante criticado é a maneira com que as pessoas são selecionadas para fazerem parte da pesquisa. Na teoria estatística mais popular, conhecida como Inferência baseada no Desenho (**ID**), deve ser utilizada amostragem probabilística. Isso quer dizer que para ser possível fazer inferência nesse contexto, todas as pessoas da população de interesse devem ter uma probabilidade positiva e conhecida de serem selecionadas para pertencer a amostra. Apesar disso, os institutos de pesquisa utilizam, por uma necessidade prática que será discutida com mais detalhes ao longo da tese, um método de seleção denominado Amostragem por Cotas. Esse método é usualmente considerado não-probabilístico, pois as pessoas são escolhidas para pertencer a amostra sem a realização de um sorteio explícito e a probabilidade de cada pessoa ser selecionada é desconhecida.

Essa tese tem dois objetivos principais: **1)** fazer uma crítica metodológica às pesquisas eleitorais brasileiras, não só do ponto de vista de Inferência baseada no Desenho como também de outros tipos de inferência usualmente considerados, **2)** apresentar uma justificativa teórica para a Amostragem Probabilística com Cotas, que é um tipo de desenho amostral híbrido, que combina características da amostragem probabilística e da amostragem por cotas, o qual também é bastante utilizado nas pesquisas eleitorais.

Essa tese está organizada em 6 capítulos. No Capítulo 1, apresentamos a teoria usual de amostragem para populações finitas no contexto usual de inferência baseada no desenho (**ID**), porém também destacamos a importância dos erros não-amostrais e a existência de outros tipos de inferência. Atenção especial é dada a explicação de conceitos básicos de amostragem, com o objetivo de discutir questões importantes de forma não-técnica, facilitando o entendimento de leitores menos experientes. Essas explicações podem ser puladas por leitores mais experientes. Na Seção 1.2.2 é apresentado um resultado novo, que pode ser interpretado como uma forma de desempate técnico do ponto de vista de **ID**. No Capítulo 2 discutimos os diferentes tipos de amostragem por cota, especificamente a Amostragem Probabilística com Cotas (**APC**), os diferentes critérios utilizados nesse tipo de amostragem, as justificativas teóricas existentes e algumas comparações empíricas entre a amostragem por cotas e a amostragem probabilística. No Capítulo 3 discutimos a importância das pesquisas eleitorais no resultado das eleições, os motivos por trás das controvérsias envolvendo-as e explicitamos as diferentes críticas feitas às pesquisas eleitorais. No Capítulo 4 apresentamos uma justificativa teórica para a **APC** sob um modelo de resposta denominado Grupos de Resposta Homogênea (**GRH**) e estimadores para as quantidades de interesse sob esse tipo de amostragem. Esse novo resultado permite que inferência baseada no desenho possa ser feita a partir de **APC** com apenas uma suposição. Também são obtidas as probabilidades de seleção e respectivos estimadores de interesse para um desenho amostral denominado Amostragem Probabilística com Voltas (**APV**), o qual combina a amostragem probabilística usual com o modelo de resposta GRH. Esse novo resultado permite que a **APC** e a **APV** sejam comparadas no mesmo contexto e sob as mesmas suposições. No Capítulo 5, é realizado um estudo de simulação, para comparar a performance dos desenhos amostrais **APC** e **APV**, utilizando diferentes estimadores para o total populacional. Na Seção 5.1.3 é apresentado um novo resultado, a distribuição assintótica da média simples sob amostragem com probabilidades desiguais, o qual pode ser visto como uma versão, do ponto de vista de inferência baseada no desenho, do conceito de amostragem não-informativa. Também é realizado um estudo descritivo de 898 pesquisas eleitorais realizadas pelos institutos de pesquisa DataFolha e Ibope, entre os anos de 1989 e 2004, com o objetivo de avaliar empiricamente a performance das pesquisas eleitorais brasileiras. Esse estudo empírico é realizado utilizando diversos critérios de interesse. No Capítulo 6 são apresentadas as conclusões dessa tese, discutindo cada uma das críticas feitas aos institutos de pesquisa em detalhes, apresentando tanto resultados teóricos quanto evidências empíricas na argumentação.

Capítulo 1

Teoria de Amostragem para Populações Finitas

1.1 Introdução

Nesse capítulo, discutiremos os diferentes tipos de amostragem para populações finitas. Populações finitas são, como o próprio nome diz, populações que têm um número finito N de unidades populacionais. Essas unidades populacionais são identificadas pelos rótulos $i \in \{1, 2, ..., N\}$. Vamos supor que estamos interessados em estudar apenas uma característica Y nessa população. O valor de Y em cada unidade populacional é denotado por Y_i.

O foco é o valor de uma função $g(Y_1, Y_2, ...Y_N)$, usualmente denominada parâmetro populacional de interesse. Diferentes parâmetros populacionais podem ser de interesse, por exemplo, o pesquisador pode estar interessado no total populacional $\tau_Y = \sum_{i=1}^{N} Y_i$, na média populacional $\mu_Y = \frac{\sum_{i=1}^{N} Y_i}{N}$ e na variância populacional $\sigma_Y^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu_Y)^2}{N}$.

Dado que o valor de Y em cada unidade populacional é a princípio desconhecido, não é possível calcular o valor dos parâmetros populacionais de interesse sem antes descobrir os valores de um ou mais Y_i's. Existem duas opções para isso:

Censo Consiste em medir/observar os valores Y_i para todas as unidades populacionais. Essa opção é mais cara e demorada, e dependendo do tamanho N da população de interesse, muitas vezes é impossível de ser realizada.

Amostra Consiste em medir/observar os valores Y_i para apenas algumas unidades populacionais. Essa opção é usualmente bem mais barata e rápida e, em alguns casos, a qualidade da medição/observação pode ser melhor.

Se o pesquisador optar por realizar um censo da população, após o seu término as quantidades Y_i serão conhecidas para todas as unidades populacionais, bastando então calcular diretamente os parâmetros populacionais de interesse. Por outro lado, se o pesquisador optar por uma amostra de tamanho $n < N$, as quantidades Y_i serão conhecidas apenas para o sub-grupo da população pertencente à amostra. Esse grupo será denotado por $s = \{i_1, ..., i_n\}$, onde os índices i_j representam os rótulos das unidades populacionais examinadas. Como Y_i será conhecido apenas para as unidades $i_j \in s$, não será possível calcular os parâmetros populacionais de interesse. Nesse cenário, além da obtenção da amostra, existe mais uma etapa que deve ser realizada para avaliar esses parâmetros

Critério do Estatístico	Método de Seleção	
	Probabilístico	Não-Probabilístico
Objetivo	Amostras Probabilísticas	Amostras Criteriosas
Subjetivo	Amostras Quase-aleatórias	Amostras Intencionais

Tabela 1.2: Tipos de Desenhos Amostrais

populacionais, conhecida como inferência. Esta consiste em utilizar as informações amostrais, obtidas apenas para $i_j \in s$, para avaliar toda a população estudada, obtendo, por exemplo, estimadores ou preditores das quantidades de interesse.

Existem muitas formas diferentes de se obter uma amostra e de se fazer inferência a partir dela. Discutiremos inicialmente os diferentes procedimentos que podem ser utilizados para se obter uma amostra da população de interesse. O procedimento utilizado para selecionar uma amostra é denominado desenho amostral.

Usualmente os desenhos amostrais são classificados em dois grandes grupos segundo o método de seleção das unidades populacionais utilizado, probabilístico e não-probabilístico. Os desenhos amostrais probabilísticos são aqueles onde a probabilidade $p(s) \geq 0$ de cada particular amostra s ser selecionada é conhecida, e a soma das probabilidades de todas as amostras é 1, ou seja, $\sum_s p(s) = 1$. Qualquer desenho amostral que não possua essas características é classificado como não-probabilístico.

Em Jessen [1978], o autor também classifica os desenhos amostrais segundo o critério de escolha utilizado pelo estatístico: objetivo e subjetivo. O critério objetivo é um que é claro e não permite margem para dúvidas, e se for seguido rigorosamente por qualquer estatístico, produzirá a mesma amostra (ou uma com as mesmas propriedades), já o critério subjetivo permite que o estatístico utilize sua opinião, seu conhecimento e seu "feeling"para definir o que é uma boa amostra. Essa classificação é apresentada na tabela 1.2.

Para fixar melhor as diferenças entre critérios objetivos e subjetivos, discutiremos aqui as diferenças entre esses critérios dentro dos procedimentos de seleção não-probabilísticos e probabilísticos.

Diferença entre amostras probabilísticas e quase-aleatórias: A diferença é que na primeira, a amostra é selecionada utilizando explicitamente um mecanismo roleta com as probabilidades $p(s)$, o que podemos chamar de aleatorização explícita. Já na segunda, o estatístico responsável supõe a amostra foi selecionada utilizando as mesmas probabilidades $p(s)$ mesmo sem selecioná-la explicitamente através de um desses mecanismos, o que pode ser chamado de aleatorização implícita. Por exemplo, imagine uma situação onde o objetivo é descobrir qual é a proporção de chocolates M&M's® de cor marrom que são fabricados. Para isso, é necessário selecionar uma amostra. Se um estatístico for até uma banca de jornal e comprar um pacote qualquer de chocolates M&M's®,

ele pode considerar essa amostra como sendo probabilística, supondo que todos os pacotes têm a mesma chance de serem sorteados. Porém um segundo estatístico pode discordar. Para ele, a amostra só será probabilística se for realizada uma loteria com todos os pacotes de chocolates M&M's® produzidos.

Diferença entre Amostras Criteriosas e Intencionais: A diferença é que na primeira, seleciona-se a amostra usando um procedimento que pode ser repetido por qualquer outro estatístico e que resultará na mesma amostra (ou com as mesmas propriedades), mesmo que as probabilidade $p(s)$ sejam desconhecidas. Por exemplo, serão selecionadas para pertencer a amostra todas as pessoas nascidas na cidade de São Paulo no dia 10 de agosto, entre as 10 e 11 horas da manhã. Já na segunda, o procedimento utilizado para seleção da amostra é realizada por um critério que depende da opinião do estatístico responsável. Por exemplo, imagine uma pesquisa que tem o objetivo de entender a situação trabalhista da população. O estatístico responsável pode escolher a amostra baseado na maneira como as pessoas se vestem, por acreditar que existe uma relação entre a maneira como as pessoas se vestem e a situação trabalhista. Um segundo estatístico pode escolher a amostra de outra forma, pois associa a vestimenta das pessoas com a situação trabalhista de uma maneira diferente do que aquela na qual o primerio estatístico acredita.

Não existe um tipo de desenho amostral universalmente aceito como ideal ou ótimo, porém o método de inferência mais utilizado, chamado de inferência baseada no desenho (design-based inference) pressupõe que a amostragem tenha sido probabilística. Além disso, é comum que os outros tipos de amostragem sejam rotulados como teoricamente incorretos, justamente porque não é possível fazer inferência baseada no desenho a partir dessas amostras.

1.2 Amostragem Probabilística e Inferência baseada no Desenho

Nesta seção discutiremos a amostragem probabilística e a inferência baseada no desenho(**ID**) com mais detalhes, devido à importância que ambas têm no contexto dessa tese. Na **ID**, se a $p(s)$ de cada possível amostra ser selecionada não é conhecida, não é possível fazer inferência. Ou seja, o conhecimento dessas probabilidades, e consequentemente o uso de uma amostra probabilística, é fundamental para se fazer **ID**. A literatura para esse tipo de inferência é bastante extensa, sendo Cochran [1977], Kish [1965], Bolfarine and Bussab [2005] e Särndal et al. [1992] ótimas referências.

1.2.1 Amostragem Aleatória Simples (AAS) com e sem Reposição

A amostragem aleatória simples (**AAS**) é um dos desenhos amostrais probabilísticos mais simples, por isso será o primeiro desenho amostral apresentado nessa tese. Existem duas variações básicas desse desenho, uma denominada sem Reposição (**AASs**), onde uma unidade populacional pode ser incluida no máximo uma vez na amostra, e a outra denominada com Reposição (**AASc**), na qual uma unidade populacional pode ser incluída mais de uma vez na amostra.

O que caracteriza um desenho amostral como sendo **AAS** é que todas as unidades populacionais possuem a mesma probabilidade de pertencer à amostra e que o tamanho da amostra, n, seja fixado de antemão. Usualmente a inclusão de unidades populacionais na amostra é realizada unidade a unidade, até que a amostra esteja completa, ou seja, tenha atingido o tamanho desejado.

Nessa seção iremos comparar os desenhos amostrais **AASc** e **AASs** pois eles são fáceis de serem trabalhados matematicamente, e para mostrar quais princípios são utilizados do ponto de vista de **ID** para escolher desenhos mais eficientes. Para as **AASs** e **AASc** é fácil calcular as probabilidades $p(s)$, pois ela é constante para toda possível amostra s. No caso da **AASc**, usualmente se considera todas as possíveis amostras ordenadas por motivos que ficarão claros na Seção 1.2.6. Ou seja, duas amostras com as mesmas unidades populacionais, porém obtidas em ordem diferente, são consideradas distintas, por exemplo, a amostra $s_1 = (u_1, u_2)$ é diferente da amostra $s_1 = (u_2, u_1)$, onde u_i represnta a unidade populacional i . Nesse contexto, a probabilidade de cada possível amostra s é dada por:

$$p(s) = \frac{1}{N^n}. \tag{1.1}$$

Já no caso da **AASs**, usualmente não se considera as amostras ordenadas, assim só é relevante quais unidades populacionais pertencem a amostra e não a ordem em que essas unidades foram obtidas, ou seja, nesse caso as amostras s_1 e s_2 são consideradas a mesma amostra. Nesse contexto, a probabilidade de cada possível amostra s é dada por:

$$p(s) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}. \tag{1.2}$$

Os desenhos amostrais **AASs** e **AASc** possuem outras diferenças relevantes, além da diferença entre as probabilidades $p(s)$. No contexto de **ID** também é relevante conhecer duas outras características importantes de qualquer desenho amostral, as funções f_j e I_j. A função f_j indica quantas vezes a unidade populacional j foi incluída na amostra, assumindo valores em $\{0, 1, 2, ..., n\}$. Já I_j é uma função indicadora, que indica se a unidade populacional j pertence à amostra. No caso da **AASs** temos $f_j = I_j$, pois uma unidade populacional só pode ser incluída uma única vez na amostra.

No modo usual de se selecionar a amostra, são realizados n sorteios consecutivos. Duas probabilidades distintas são muito importantes nesse contexto de inferência baseada no desenho para populações finitas. Essas probabilidades são:

Probabilidade de Seleção é a probabilidade p_j da unidade populacional j ser selecionada para pertencer à amostra em um único sorteio. Essa probabilidade só faz sentido no contexto de um desenho amostral com reposição, pois em todo sorteio, essa probabilidade se mantém igual.

Probabilidade de Inclusão é a probabilidade π_j da unidade populacional j ser selecionada para pertencer à amostra (considerando o total de sorteios que serão realizados). Essa probabilidade é $\pi_j = \sum_{s \ni j} p(s)$, ou seja, a soma das probabilidades de todas as amostras que contêm a unidade populacional j.

No caso da **AASc**, a probabilidade de seleção p_j da unidade j ser selecionada em qualquer um dos n sorteios é igual a $\frac{1}{N}$. Uma consequência desse resultado é que, no caso da **AASc**:

$$f_j \sim Bin\left(n, \frac{1}{N}\right), \tag{1.3}$$

onde $Bin(n,p)$ indica a distribuição binomial com parâmetros n e p. Assim temos $E(f_j) = \frac{n}{N}$, $Var(f_i) = n\frac{1}{N}\left(1-\frac{1}{N}\right)$ e $Cov(f_i, f_j) = -\frac{n}{N^2}$. A probabilidade de inclusão π_j para a **AASc** é $1-\left(1-\frac{1}{N}\right)^n$. Também é importante calcular a probabilidade de inclusão simultânea π_{ij} das unidades populacionais i e j, que é dada por $\pi_{ij} = 1 - 2\left(1-\frac{1}{N}\right)^n + \left(1-\frac{2}{N}\right)^n$.

Já no caso da **AASs**, em cada sorteio as probabilidades de seleção se alteram. Essas probabilidades (condicionais) são dadas por $\frac{1}{N}$ no primeiro sorteio, $\frac{1}{N-1}$ no segundo sorteio, $\frac{1}{N-2}$ no terceiro sorteio e assim sucessivamente até $\frac{1}{N-(n-1)}$ no n-ésimo sorteio. Essas probabilidades são condicionais pois em cada um desses sorteios, a unidade populacional só participa se ela não tiver sido sorteada nos sorteios anteriores, assim a probabilidade de participação é dada por 1 no primeiro sorteio, $\frac{N-1}{N}$ no segundo sorteio, $\frac{N-2}{N}$ no terceiro sorteio e assim sucessivamente. Assim, obtemos que a probabilidade de inclusão na **AASs** é dada por:

$$\pi_j = E(f_j) = \frac{n}{N}. \tag{1.4}$$

Uma consequência desse resultado é que no caso da **AASs** temos

$$f_j \sim Bern(\pi_j), \tag{1.5}$$

onde $Bern(p)$ indica a distribuição de Bernoulli com probabilidade de sucesso p. Assim obtemos $E(f_j) = \frac{n}{N}$, $Var(f_j) = \frac{n}{N}\left(1-\frac{n}{N}\right)$ e $Cov(f_i, f_j) = -\frac{n}{N^2}\frac{N-n}{N-1}$. Como no caso da **AASc**, também é importante obter a probabilidade de inclusão sumltânea π_{ij} das unidades populacionais i e j, que é dada por $\pi_{ij} = \frac{n}{N}\frac{n-1}{N-1}$.

Até este ponto calculamos probabilidades e funções de interesse relacionadas aos desenhos amostrais **AASc** e **AASs**. Torna-se necessário agora considerar estimadores para as quantidades populacionais de interesse. Um estimador é qualquer função que dependa somente da amostra, ou seja, não pode depender de quantidades desconhecidas, como por exemplo de Y_i quando i não pertence a amostra. Exemplos de estimadores para AAS são o estimador do total populacional $\hat{\tau} = \frac{N}{n}\sum_{i\in s} Y_i$ e o da média populacional $\hat{\mu} = \frac{1}{n}\sum_{i\in s} Y_i$.

A distribuição amostral utilizada para se fazer inferência baseada no Desenho é concretizada, conceitualmente, imaginando-se infinitas replicações do desenho amostral, e avaliando como o estimador da quantidade populacional de interesse se comporta em cada uma dessas replicações. No contexto de populações finitas, a qualidade de um estimador é avaliada por meio de dois critérios

básicos resultantes de sua distribuição amostral:

Vício do Estimador É uma medida de quanto, ao longo das infinitas replicações, é a diferença média entre o estimador e o parâmetro populacional sendo estudado.

Variância do Estimador É uma medida de quanto, ao longo das infinitas replicações, os valores do estimador nas diferentes amostras variam, ou o oscilam, em torno de sua média.

O ideal é que o estimador tenha um vício pequeno, ou seja, que em média ele acerte o valor do parâmetro populacional, e que tenha uma variância pequena, ou seja, que os diferentes valores que o estimador assumiria em cada possível amostra não sejam muito diferentes uns dos outros. A maneira mais intuitiva de entender esses conceitos é fazendo uma analogia com um alvo, onde cada uma das quatro figuras em 1.1 representam os arremessos de dardos de quatro jogadores diferentes, denominados (a), (b), (c) e (d). O objetivo é acertar o ponto vermelho no alvo. O jogador (a), em média, erra o ponto vermelho, ou seja, ele têm um arremesso viciado no sentido de consistentemente arremessar o dardo mais para para o lado direito do alvo, e além disso, o resultado de cada arremesso varia bastante, ou seja, sua mira não é muito precisa. O jogador (a) é o pior dos 4 jogadores. Já o jogador (b), apesar de a sua mira não ser muito precisa, pois o resultado de seus arremessos varia bastante, em média acerta o alvo. O jogador (c) tem a mira bastante precisa, pois os resultados de seus arremessos sempre ficam muito próximos uns dos outros, porém em média ele erra seus arremessos. É difícil dizer entre os jogadores (b) e (c) qual é o melhor, porém eles são claramente melhores do que o jogador (a). O melhor jogador de todos é (d), pois ele é preciso e em média acerta o alvo.

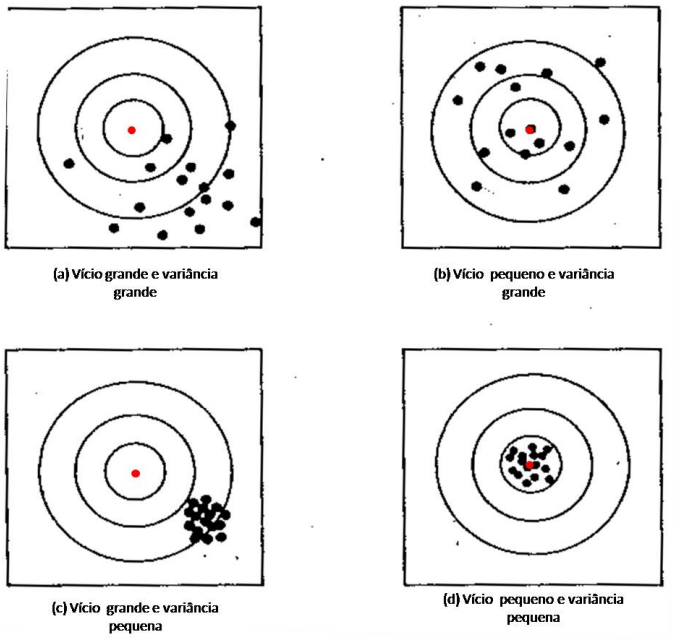

Figura 1.1: Vício e Variância sob diferentes desenhos amostrais

O alvo em questão no nosso caso é o parâmetro populacional de interesse. Como foi visto no exemplo, tanto o vício quanto a variância são importantes para avaliar um desenho amostral, ou seja, aqueles que em média têm uma performance melhor. Uma medida que leva o vício e a variância de um estimador em consideração simultaneamente é o erro quadrático médio (EQM), apresentado na definição 1.1.

Definição 1.1 (Erro Quadrático Médio (EQM)) *Seja $\hat{\theta}$ um estimador da quantidade populacional θ. Então o erro quadrático médio do estimador $\hat{\theta}$ é dado por:*

$$EQM(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var(\hat{\theta}) + \left[Vicio(\hat{\theta})\right]^2, \tag{1.6}$$

onde $Vicio(\hat{\theta}) = E\left(\hat{\theta}\right) - \theta$.

Nesse ponto, é importante mencionar que o $EQM(\hat{\theta})$ é uma função de θ e que para calcular o vício e a variância de um estimador, é necessário conhecer o desenho amostral em questão, pois essas quantidades são calculadas em termos da distribuição amostral mencionada anteriormente, por exemplo para o caso da esperança, temos $E(\hat{\theta}) = \sum_s \hat{\theta}(s)p(s)$. No contexto de **ID**, a única quantidade aleatória é o vetor de rótulos das unidades populacionais que compõem a amostra $(i_{j_1}, i_{j_2}, ..., i_{j_n})$. No caso da **AASc**, temos por exemplo que:

$$\begin{aligned} E\left(\hat{\tau}_{AASc}\right) &= E\left(\frac{N}{n}\sum_{i\in s} Y_i\right) = E\left(\frac{N}{n}\sum_{i=1}^{N} Y_i f_i\right) \\ &= \frac{N}{n}\sum_{i=1}^{N} Y_i E\left(f_i\right) = \frac{N}{n}\sum_{i=1}^{N} Y_i \frac{n}{N} = \sum_{i=1}^{N} Y_i = \tau, \end{aligned} \tag{1.7}$$

pois $E\left(f_i\right) = \frac{n}{N}$. Ou seja, o estimador $\hat{\tau}_{AASc}$ é não viciado para $\tau = \sum_{i=1}^{N} Y_i$, pois sua esperança é igual ao valor do parâmetro populacional o qual o estimador se propõe a estimar. Já para a variância do estimador, obtemos:

$$\begin{aligned} Var\left(\hat{\tau}_{AASc}\right) &= Var\left(\frac{N}{n}\sum_{i\in s} Y_i\right) = Var\left(\frac{N}{n}\sum_{i=1}^{N} Y_i f_i\right) \\ &= \frac{N^2}{n^2}\left(\sum_{i=1}^{N} Y_i^2 Var\left(f_i\right) + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} Y_i Y_j Cov(f_i, f_j)\right) \\ &= \frac{N^2}{n}\frac{\sum_{i=1}^{N}\left(Y_i - \mu_Y\right)^2}{N} = N^2\frac{\sigma_Y^2}{n}, \end{aligned} \tag{1.8}$$

onde $\sigma_Y^2 = \frac{\sum_{i=1}^{N}(Y_i-\mu_Y)^2}{N}$. Já para o caso da **AASs**, a esperança do estimador do total populacional é dada por:

$$\begin{aligned} E\left(\hat{\tau}_{AASs}\right) &= E\left(\frac{N}{n}\sum_{i\in s}Y_i\right) = E\left(\frac{N}{n}\sum_{i=1}^{N}Y_i f_i\right) \\ &= \frac{N}{n}\sum_{i=1}^{N}Y_i E\left(f_i\right) = \frac{N}{n}\sum_{i=1}^{N}Y_i\frac{n}{N} = \sum_{i=1}^{N}Y_i = \tau, \end{aligned} \tag{1.9}$$

e a variância é dada por:

$$\begin{aligned} Var\left(\hat{\tau}_{AASs}\right) &= Var\left(\frac{N}{n}\sum_{i\in s}Y_i\right) = Var\left(\frac{N}{n}\sum_{i=1}^{N}Y_i f_i\right) \\ &= \frac{N^2}{n^2}\left(\sum_{i=1}^{N}Y_i^2 Var\left(f_i\right) + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}Y_i Y_j Cov(f_i, f_j)\right) \\ &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{\sum_{i=1}^{N}\left(Y_i-\mu_Y\right)^2}{N-1} = N^2\left(1-\frac{n}{N}\right)\frac{S_Y^2}{n}, \end{aligned} \tag{1.10}$$

onde $S_Y^2 = \frac{N}{N-1}\sigma_Y$. Podemos ver que as propriedades do estimador do total populacional sob **AASs** e **AASc** são muito parecidas. As propriedades dos estimadores do total e da média populacional, na classe dos estimadores não-viciados, estão resumidas na tabela 1.3.

Tabela 1.3: Estimadores sob os desenhos amostrais **AASs** e **AASc**

Desenho Amostral	Parâmetro populacional	Notação	Estimador	Vício do Estimador	Variância do Estimador
AASc	Média	$\hat{\mu}_{AASc}$	$\frac{1}{n}\sum_{i\in s}Y_i$	0	$\frac{\sigma_Y^2}{n}$
	Total	$\hat{\tau}_{AASc}$	$\frac{N}{n}\sum_{i\in s}Y_i$	0	$N^2\frac{\sigma_Y^2}{n}$
AASs	Média	$\hat{\mu}_{AASs}$	$\frac{1}{n}\sum_{i\in s}Y_i$	0	$\left(1-\frac{n}{N}\right)\frac{S_Y^2}{n}$
	Total	$\hat{\tau}_{AASs}$	$\frac{N}{n}\sum_{i\in s}Y_i$	0	$N^2\left(1-\frac{n}{N}\right)\frac{S_Y^2}{n}$

A quantidade $\frac{n}{N}$ que aparece na variância dos estimadores da AASs é conhecida como fração amostral. Uma forma de se comparar os estimadores da **AASs** e da **AASc** é utilizando o efeito do planejamento (EPA), que é definido como a razão dos EQM dos estimadores em questão, obtendo:

$$EPA(\hat{\mu_{AASc}}, \hat{\mu_{AASs}}) = \frac{EQM(\hat{\mu}_{AASc})}{EQM(\hat{\mu}_{AASs})} = \frac{N-1}{N-n}, \tag{1.11}$$

o qual é maior do que 1 para todo $n > 1$, assim podemos afirmar que o estimador $\hat{\mu}_{AASc}$ é sempre

menos eficiente do que o estimador $\hat{\mu}_{AASs}$, o que é bastante intuitivo, visto que na amostragem com reposição podem existir informações redundantes, pois a mesma unidade populacional pode estar múltiplas vezes na amostra. Em Basu [1958] o autor provou que a **AASc**, desconsiderando as unidades repetidas, é mais eficiente do que a **AASc** mantendo todas as unidades repetidas se o tamanho da amostra for maior do que um. Sendo ν o número de unidades distintas aparecendo na amostra, o autor mostrou que **EQM** do estimador da média considerando somente as unidades amostrais distintas é dada, implicitamente, por $E\left(\frac{N-\nu}{N-1}\frac{\sigma^2}{\nu}\right)$, e é sempre menor do que $\frac{\sigma^2}{n}$ se $n > 1$. Ou seja, é melhor ter a amostra menor, sem repetição, do que a completa, maior, com unidades repetidas.

É importante notar que a qualidade de um estimador no contexto de **ID** não está relacionada com a estimativa obtida para uma amostra específica, mas somente com a média dos valores que poderiam ser obtidos em todas as possíveis amostras. É uma lógica pré-experimental que perdura após a experimentação, não importando qual particular estimativa foi efetivamente observada.

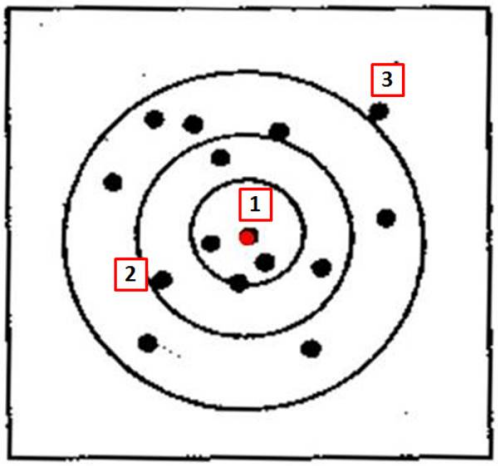

Figura 1.2: Alvo do jogador (b)

Para explicar melhor esse ponto, na figura 1.2 desenhamos somente o alvo do jogador (b), e destacamos 3 resultados de arremessos feitos por ele. O arremesso 1 foi o melhor deles, atingiu ao alvo vermelho em cheio, o arremesso 2 foi mediano e o arremesso 3 foi o pior de todos os arremessos feito pelo jogador. Apesar de o jogador (b) na média de todos os arremessos acertar ao alvo, isso não garante que um arremesso específico irá acertar o mesmo, mais que isso, seria possível que nenhum dos arremessos acertasse o alvo. Além disso, apesar do mesmo jogador ter feito todos os arremessos, claramente alguns arremessos são melhores do que os outros, ou seja, têm um erro observado menor.

Analogamente, no caso do desenho amostral, por mais que o estimador utilizado seja não-viciado, isso não quer dizer que uma amostra específica tenha "acertado o alvo". Além disso, algumas amostras serão melhores que outras, ou seja, terão um erro amostral menor, mesmo tendo sido (hipoteticamente) geradas pelo mesmo desenho amostral. O erro amostral para uma particular amostra pode ser definido como a distância entre a estimativa nessa particular amostra e o

parâmetro populacional o qual o estimador se propõe a estimar. O erro amostral será discutido com mais detalhes em 1.2.2.

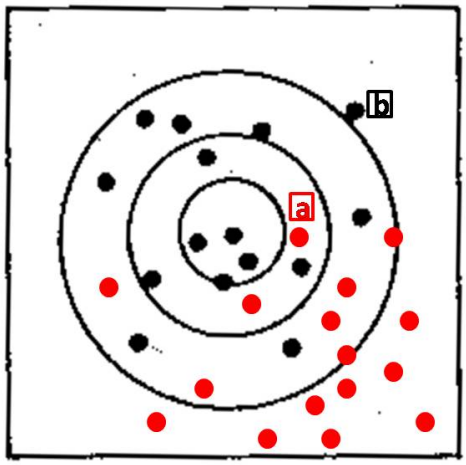

Figura 1.3: Alvos dos jogadores (a) e (b) sobre-postos

Seguindo a mesma linha de raciocínio, na figura 1.3 mostramos a sobreposição dos arremessos do jogador (b) em preto, e do jogador (a) em vermelho. Apesar do jogador (a) ser considerado o pior jogador pelo critério do vício, pois em média seus arremessos erram mais o alvo do que os outros jogadores, existem arremessos do jogador (a) que são melhores do que arremessos do jogador (b), como indicado na figura 1.3 por $[a]$ e $[b]$, respectivamente.

O mesmo pode ocorrer com os diferentes pares de desenhos amostrais e estimadores, um estimador mais eficiente pode ter uma estimativa em uma amostra específica pior do que uma estimativa obtida de um estimador menos eficiente. Por exemplo, o estimador do total populacional de uma particular amostra da **AASc** pode ser melhor, ou seja, ter um erro amostral menor, do que o estimador do total populacional de uma particular amostra da **AASs**. Independentemente dessas possibilidades, até porque é impossível saber se algo do tipo ocorreu pois o parâmentro populacional é desconhecido, estimadores são comparados segundo suas propriedades pré-experimentais, não importando o valor que o estimador tenha assumido em uma particular amostra, denominado estimativa.

1.2.2 Erro Amostral para AAS

O erro amostral é a diferença entre a estimativa e o parâmetro populacional o qual o estimador se propõe a estimar. Antes de selecionar a amostra, essa diferença pode ser representada por $\hat{\mu} - \mu_Y$ para o estimador $\hat{\mu}$ da média populacional μ. Como para cada amostra efetivamente selecionada a estimativa obtida é diferente e o parâmetro populacional é desconhecido, não é possível saber qual o erro amostral cometido por uma particular amostra.

Note que erro amostral se refere somente ao erro cometido por estimar uma quantidade populacional utilizando uma amostra. Existem outros tipos de erros não-amostrais que podem fazer com que mesmo um estimador não-viciado, não coincida em média, com o parâmetro populacional, os

quais serão discutidos na Seção 1.3. No exemplo dos alvos, é o mesmo que um jogador que erra o alvo por causa de uma rajada de vento, por exemplo, e não por causa de sua mira.

Na Seção 1.2.1 calculamos a esperança e a variância de $\hat{\mu}$, e consequentemente de $\hat{\mu} - \mu_Y$, para o caso das **AASs** e **AASc**. Nessa seção o interesse não reside somente nessas duas quantidades características da distribuição amostral, mas na distribuição amostral completa. Para populações infinitas, as quais não possuem uma quantidade finita unidades populacionais, e no contexto de AAS, não é difícil mostrar que a média amostral tem assintoticamente uma distribuição normal, ou seja, quanto maior for o tamanho da amostra n, melhor a distribuição amostral de $\hat{\mu}$ (ou $\hat{\tau}$) é aproximada pela distribuição normal. Esse resultado é consequência do Teorema Central do Limite (**TCL**). Porém, no contexto de populações finitas, esse resultado não vale, como é exemplificado em Plane and Gordon [1982]. Para verificar esse fato, primeiro é importante notar que para populações finitas, se a média amostral de uma particular amostra é denotada por $\overline{y}_n$, então a média das $N-n$ unidades populacionais restantes é dada por $\overline{y}_{N-n} = \frac{\tau - n\overline{y}_n}{N-n} = a_n - b_n\overline{y}_n$, onde $a_n = \frac{\tau}{N-n}$ e $b_n = \frac{n}{N-n}$. Se existem Q amostras distintas de tamanho n que resultam na média amostral $\overline{y}_n$, então a probabilidade de $\overline{y}_n$ é dada por:

$$P_n\left(\overline{y}_n\right) = \frac{Q}{\binom{N}{n}}, \tag{1.12}$$

porém isso também implica que devem haver Q amostras distintas de tamanho $N-n$ que resultam na média amostral $\overline{y}_{N-n}$, então a probabilidade de $\overline{y}_{N-n}$ é dada por:

$$P_{N-n}\left(\overline{y}_{N-n}\right) = \frac{Q}{\binom{N}{N-n}} = \frac{Q}{\binom{N}{n}}, \tag{1.13}$$

o que implica que $P_n\left(\overline{y}_n\right) = P_{N-n}\left(\overline{y}_{N-n}\right) = P_{N-n}(a_n - b_n\overline{y}_n)$. Ou seja, a distribuição amostral de $\overline{Y}_{N-n}$ tem a mesma forma que a distribuição amostral de $\overline{Y}_n$, porém os valores que $\overline{y}_{N-n}$ assume são uma transformação linear dos valores de $\overline{y}_n$, fazendo com que essa distribuição seja invertida e mais concentrada. Assim, se a distribuição de $\overline{Y}_n$ não tem a forma de uma distribuição normal, a de $\overline{Y}_{N-n}$ também não terá. Ou seja, para N grande, finito, mesmo que $n = N-1$, a distribuição amostral só será normal se ela também for para o caso $n = 1$.

São necessárias mais condições, além da usual de que $n \to \infty$, para encontrar um resultado similar ao TCL para **AASs** de populações finitas, especificamente, $N - n \to \infty$ e a população não pode ter observações muito discrepantes, que sejam responsáveis por uma contribuição exagerada à variância populacional. Detalhes do resultado e de suas condições necessárias podem ser encontradas em Hájek [1960] e em Bolfarine and Bussab [2005]. Para enunciar esse teorema, é necessário considerar uma sequência de populações $\{\mathcal{U}_\nu\}_{\nu \geq 1}$, tal que N_ν é o tamanho populacional da população ν, a qual tem média e variância populacional dadas por μ_ν e S^2_ν, respectivamente, onde $N_{\nu+1} > N_\nu$, $\nu \geq 1$. Da população $\mathcal{U}_\nu$ é selecionada uma amostra $\int_\nu$ de tamanho $n_\nu (n_{\nu+1} > n_\nu)$

segundo **AASs**. Para estimar a média populacional μ_ν é utilizada a média amostral $\overline{y}_\nu$ correspondente à amostra observada, a qual tem esperança μ_ν e variância $\left(1-\frac{n_\nu}{N_\nu}\right)\frac{S_\nu^2}{n_\nu}$.

Teorema 1.1 (TCL para AASs de populações finitas) *Suponha que $n_\nu \to \infty$ e $N_\nu - n_\nu \to \infty$ quando $\nu \to \infty$. Considere também que a sequência $\{Y_{i\nu}\}_{i\nu}$ satisfaz a condição de Lindenberg-Hajek,*

$$\lim_{\nu\to\infty} \frac{\sum_{T_\nu(\delta)} Y_{i\nu} - \mu_\nu}{\sum_{i=1}^{N_\nu} (Y_{i\nu}-\mu_\nu)^2} = 0, \tag{1.14}$$

para todo $\delta > 0$, onde $T_\nu(\delta)$ é o conjunto das unidades populacionais em $\mathcal{U}_\nu$ para as quais

$$\frac{|Y_{i\nu}-\mu_\nu|}{\sqrt{\sum_{i=1}^{N_\nu}(Y_{i\nu}-\mu_\nu)^2}} > n\delta. \tag{1.15}$$

*Então, com relação a **AASs**,*

$$\frac{\overline{y}_\nu - E(\overline{y}_\nu)}{\sqrt{Var(\overline{y}_\nu)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{1.16}$$

quando $\nu \to \infty$, onde $\mathcal{N}(0,1)$ representa a distribuição normal-padrão e $\xrightarrow{\mathcal{D}}$ significa convergência em distribuição. Se a sequência $\{Y_{i\nu}\}_{i\nu}$ também satisfizer a condição

$$\left(1-\frac{n_\nu}{N_\nu}\right)\frac{S_\nu^2}{n_\nu} \to 0, \tag{1.17}$$

quando $\nu \to \infty$, também temos que

$$\frac{\overline{y}_\nu - E(\overline{y}_\nu)}{\sqrt{\widehat{Var}(\overline{y}_\nu)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{1.18}$$

onde $\widehat{Var}(\overline{y}_\nu)$ e s_ν^2 são estimadores não-viciados de $Var(\overline{y}_\nu)$ e S_ν^2, respectivamente.

Com alguns pequenos ajustes, o teorema 1.1 também é válido para **AASc**. A grande vantagem da aproximação apresentada nesse teorema é que não é necessário calcular as probabilidades p(s) de todas as possíveis amostras s e suas respectivas estimativas para se conhecer a distribuição amostral do estimador da média populacional. Outra característica do resultado no teorema 1.1 é que quando $Var(\overline{Y}_n)$ é desconhecida, o qual geralmente é o caso, é necessário encontrar um estimador não-viciado para $Var(\overline{Y}_n)$ para poder utilizar o resultado. Os estimadores não-viciados dessa quantidade para a **AASs** e **AASc** são:

$$\begin{aligned}\widehat{Var}(\hat{\mu}_{AASc}) &= \frac{\sum_{i\in s}(Y_i-\overline{Y}_n)^2}{n(n-1)} = \frac{s^2}{n},\\ \widehat{Var}(\hat{\mu}_{AASs}) &= \left(1-\frac{n}{N}\right)\frac{s^2}{n} = \left(1-\frac{n}{N}\right)\frac{\sum_{i\in s}(Y_i-\overline{Y}_n)^2}{n(n-1)} = \left(1-\frac{n}{N}\right)\frac{s^2}{n},\end{aligned}$$

onde $s^2 = \frac{\sum_{i\in s}(Y_i-\mu_{A\hat{A}Sc})^2}{(n-1)}$ na **AASc** e $s^2 = \frac{\sum_{i\in s}(Y_i-\mu_{A\hat{A}Ss})^2}{(n-1)}$ na **AASs**. Note que na **AASc**, temos $E(s^2) = \sigma^2_Y$ e na **AASs**, temos $E(s^2) = S^2_Y = \frac{N}{N-1}\sigma^2_Y$.

O resultado em 1.16 nos diz qual é o comportamento assintótico de $\overline{Y}_n$ em todas as possíveis amostras de uma população que satisfaz todas as condições do teorema, e consequentemente como o erro amostral se comportaria ao longo de todas as possíveis amostras. Das probabilidades de cada intervalo da distribuição normal desenhados na figura 1.4, fica evidente que a maioria dos erros amostrais estarão concentrados em torno de zero, e quanto maior for o erro amostral de uma determinada amostra, menor é a probabilidade dessa amostra ser selecionada.

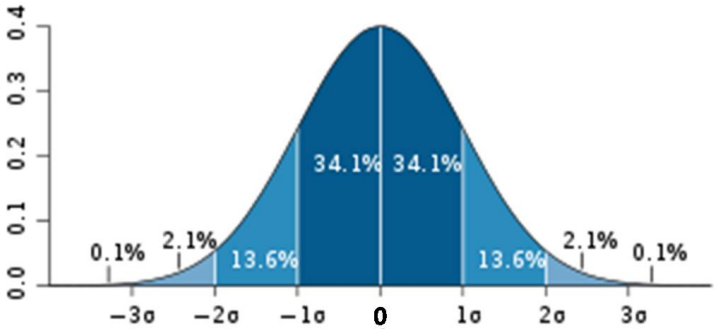

Figura 1.4: Distribuição dos erros amostrais na AAS

Esse resultado também é bastante importante para determinar o tamanho da amostra e para determinar os intervalos de confiança dos estimadores utilizados, os quais são essenciais para se fazer **ID**.

Intervalos de confiança e a margem de erro para AAS

Apesar dos resultados apresentados nessa seção serem válidos para estimadores de qualquer média populacional, estamos mais interessados no contexto de proporções populacionais, pois esse é o cenário usualmente considerado em pesquisas eleitorais. Nessa seção iremos denotar o estimador da proporção populacional por $\hat{P}_i$ e o parâmetro populacional por P_i. Supondo que existam C categorias (candidatos disputando uma eleição e também respostas como não sabe, não respondeu etc...), estamos interessados em estimar as quantidades populacionais, para $i = 1, 2, ..., C$:

$$P_i = \frac{\sum_{j=1}^{N} Y_j^{(i)}}{N}, \tag{1.19}$$

onde $Y_j^{(i)} = 1$ se a j-ésima unidade populacional votar no candidato i, e $Y_j^{(i)} = 0$ em caso contrário. É importante salientar que $\sum_{i=1}^{C} P_i = 1$. O estimador de P_i é dado por:

$$\hat{P}_i = \frac{\sum_{j \epsilon s} Y_j^{(i)}}{n}, \tag{1.20}$$

onde n é o tamanho da amostra. O erro amostral cometido em uma particular amostra, para uma categoria específica i, nesse contexto, é o valor da diferença $(\hat{P}_i - P_i)$[1]. Usualmente, é impossível avaliar essa diferença pois não conhecemos P_i. Ou seja, quando observamos uma amostra e obtemos uma estimativa, não sabemos qual o erro amostral sendo cometido. O resultado no teorema 1.1 quantifica, a priori, os diferentes tamanhos de erro amostral, porém não diz nada sobre aquele da amostra efetivamente observada.

Gostariamos de afirmar que "O valor absoluto do erro amostral dessa pesquisa é menor ou igual a $d\%$", onde d é uma constante arbitrariamente pequena. Em pesquisas eleitorais, usualmente chama-se d de margem de erro. Apesar de não ser possível avaliar a estimativa $(\hat{p}_i - P_i)$ para a amostra que foi efetivamente selecionada, o teorema de Hájek em 1.1 nos diz qual é o seu comportamento ao longo de todas as possíveis amostras. Utilizando esse resultado podemos quantificar $(\hat{P}_i - P_i)$ de uma forma um pouco diferente. Podemos afirmar, por exemplo que "Antes da amostra ser selecionada, o valor absoluto do erro amostral da pesquisa tinha uma probabilidade de $(1 - \alpha)$ % de ser menor do que d%". Apesar da segunda afirmação ter uma probabilidade associada, ela também tem muita força. Para formalizá-la matematicamente, três quantidades que necessitam ser especificadas são:

Confiança $(1 - \alpha)$ - a confiança que gostaríamos de ter, ou seja, especificar qual probabilidade queremos associar a d.

Erro amostral (d) - o erro amostral desejado, ou seja, qual d é considerado suficiente.

Tamanho amostral (n) - o tamanho da amostra desejada, ou seja, quantas entrevistas serão realizadas.

Matematicamente, a frase acima pode ser enunciada como:

$$P(|\hat{P}_i - P_i| \leq d) = 1 - \alpha, \tag{1.21}$$

de onde obtemos, como pode ser verificado em Bolfarine and Bussab [2005], que as quantidades α, n e d são relacionadas da seguinte forma:

[1]Também é possível definir o erro amostral total, como sendo a distância entre $\hat{P}$ e P, onde $\hat{P}$ e P são vetores definidos, respectivamente, por $(\hat{P}_1, \hat{P}_2, ..., \hat{P}_C)$ e $(P_1, P_2, ..., P_C)$.

$$d = z_{\frac{\alpha}{2}} \sqrt{Var(\hat{P}_i)}, \tag{1.22}$$

onde $z_{\frac{\alpha}{2}}$ é o quantil da distribuição normal-padrão, tal que $P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$, onde Z têm distribuição $\mathcal{N}(0, 1)$. Note que em 1.22, cada categoria i possui um erro amostral diferente associado a ela, pois d depende de $Var(\hat{P}_i)$. Lembrando que para a **AASc**, temos $Var(\hat{P}_i) = \frac{P_i(1-P_i)}{n}$ e para a **AASs** temos $Var(\hat{P}_i) = \left(\frac{N-n}{N-1}\right) \frac{P_i(1-P_i)}{n}$. Para evitar a inconveniência de ter que divulgar um d_i diferente para a intenção de voto de cada candidato e pelo fato de não conhecermos os valores de P_i, os institutos de pesquisa usualmente substituiem $Var(\hat{P}_i)$ em 1.22 pelo pior caso possível dado por $P_i = \frac{1}{2}$, ou seja, o valor que torna d maior possível para um mesmo nível de confiança $1 - \alpha$. Nesse caso, a probabilidade em 1.21 é modificada da seguinte forma:

$$P(|\hat{P}_i - P_i| \leq d) \geq 1 - \alpha. \tag{1.23}$$

É importante enfatizar aqui que a necessidade de trocar P_i por $\frac{1}{2}$ ocorre porque estamos olhando cada categoria i separadamente, ao invês de considerarmos todas as categorias conjuntamente. De 1.23, a forma como as quantidades α, n e d são relacionadas no caso da **AASc** é dada por:

$$d = z_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}, \tag{1.24}$$

e para o caso da **AASs** basta multiplicar o lado direito da equação por $\sqrt{\frac{N-n}{N-1}}$. No contexto de **ID**, são muito utilizadas estimativas intervalares dos parâmetros de interesse derivadas da pa probabilidade em 1.23, pois essa é uma forma de mostrar os resultados de uma pesquisa levando em conta o erro amostral. Essas estimativas intervalares são conhecidas como intervalos de confiança. Para derivar matematicamente esses intervalos de confiança, é necessário obter uma quantdade pivotal, que pode ser interpretada como uma variável aleatória que depende tanto de quantidades amostrais quanto de parâmetros populacionais, porém a sua distribuição amostral não depende de nenhum parâmetro. No contexto de proporções populacionais, a quantidade pivotal utilizada é $\frac{\hat{P}_i - P_i}{\sqrt{Var(\hat{P}_i)}}$. O intervalo de confiança individual para P_i com confiança de $(1 - \alpha)\%$ é dado por:

$$\left[\hat{P}_i - d; \hat{P}_i + d\right], \tag{1.25}$$

onde $d = z_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}$ para **AASc** e $d = z_{\frac{\alpha}{2}} \sqrt{\frac{N-n}{N-1}} \frac{1}{2\sqrt{n}}$ para **AASs**.

A interpretação do intervalo de confiança usualmente causa confusão. Teoricamente, o intervalo em 1.25 quer dizer que, se em toda possível amostra selecionada esse intervalo for calculado com a estimativa obtida $\hat{P}_i$, esses intervalos irão conter o real parâmetro populacional P_i em $(1-\alpha)\%$ das amostras. Novamente, essa probabilidade associada ao intervalo é pré-experimental, ou seja, antes

da seleção da amostra, porém depois de observada a amostra, o intervalo de confiança calculado para essa amostra específica ou contém ou não contém o parâmetro populacional.

Por exemplo, imagine que uma fábrica produz canetas, das quais $1,3\%$ são defeituosas, não escrevem. Imagine que o João foi numa loja comprar uma única caneta dessa marca. No ato da compra, ele tem uma probabilidade de $0,987$ de comprar uma caneta que funciona. Depois que ele comprou a caneta, levou-a pra casa e tentou escrever com ela, não existe mais probabilidade associada a caneta. Ou ela funciona, ou ela não funciona. Ou seja, depois de realizado o experimento, no caso a compra de uma particular caneta, não importa mais qual era a probabilidade da caneta ser defeituosa, pois aquela caneta é ou não defeituosa. A dificuldade com amostragem é que como o parâmetro populacional P_i é desconhecido na maioria dos casos, nunca sabemos se o intervalo de confiança daquela particular amostra contém ou não o parâmetro populacional.

Voltando aos exemplos dos alvos, imagine que os jogadores não estão mais arremessando dardos, que seriam os equivalentes as estimativas pontuais $\hat{P}_i$, agora eles estão arremessando argolas, que seriam o equivalente as estimativas intervalares $\hat{P}_i \pm z_{\frac{\alpha}{2}} \sqrt{Var(\hat{P}_i)}$, e o alvo pode ser interpretado como um pino. Se uma particular argola é arremessada e contém o pino no seu interior, seria o equivalente do intervalo de confiança conter o parâmetro populacional. Ou seja, apesar de existir uma determindada probabilidade à priori das argolas acertarem o alvo para cada jogador, cada arremesso contém ou não contém o pino.

Além disso, podemos imaginar o raio da argola como sendo determinado pela confiança $(1-\alpha)$. Se uma argola tiver um raio maior, a sua confiança é maior do que era com a argola menor, pois quando o jogador estiver arremessando a argola maior, ele terá uma probabilidade maior de acertar o alvo, como podemos ver na figura 1.5, onde as argolas são representadas por círculos vermelhos em torno do arremesso. Claramente é mais difícil de acertar o alvo usando as argolas em (a) do que utilizando as argolas em (b). O mesmo ocorre com o intervalo de confiança, aumentar a confiança do intervalo aumenta a largura do intervalo, e será mais fácil de um intervalo conter o parâmetro populacional.

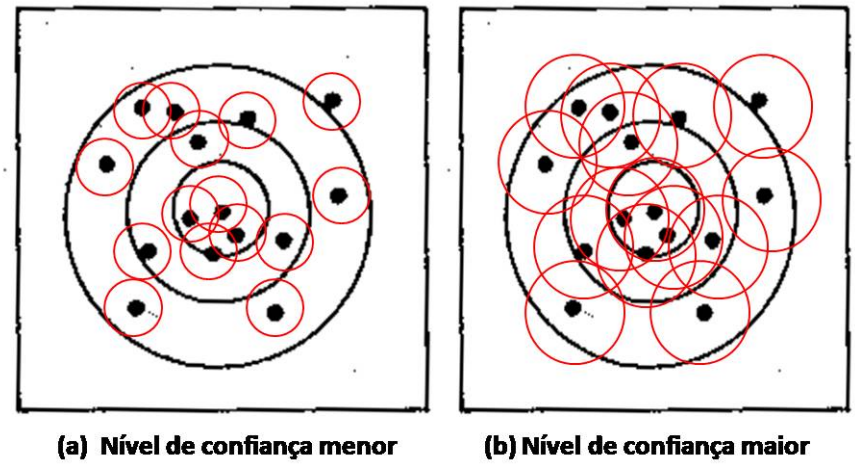

Figura 1.5: Precisão dos Intervalos de Confiança

Alterar a confiança do intervalo de confiança não tem nenhum ganho real, pois apesar de aumentar a probabilidade de acerto, na verdade a precisão do intervalo diminue, pois um erro amostral maior deixa de ser considerado um erro. Ou seja, uma pessoa que passa a arremesar uma argola maior, vai acertar mais o alvo, mas isso não quer dizer que a sua mira melhorou, o que aumentou foi o raio da argola, e uma distância que na argola menor era considerada como um erro, agora passará a ser considerada como um acerto.

A única forma de realmente melhorar a eficiência do intervalo de confiança é manter o mesmo nível de confiança porém diminuir a largura do intervalo. Isso só é possível se a variância do estimador diminuir, o que por sua vez só é possível aumentando o tamanho da amostra, uma vez que a variância do estimador $\hat{P}_i$ depende de $\frac{1}{n}$. Pensando no exemplo das argolas, manter a probabilidade de uma jogador acertar porém diminuir a largura da argola não parece ser fisicamente possível. Para faciliar a compreensão, imagine que aumentar o tamanho da amostra seria o equivalente do jogador dar um passo em direção ao alvo, aumentado assim a sua probabilidade de acertar o alvo, mesmo com um disco de mesmo raio. Nesse exemplo dos alvos, o comprimento do raio pode ser interpretado como o erro amostral d, definido em 1.22.

Os resultados obtidos nessa seção consideram cada categoria P_i separadamente, sem levar em consideração que $\sum_{i=1}^{C} P_i = 1$ e que os estimadores $\hat{P}_i$ e $\hat{P}_j$, para $i \neq j$, são correlacionados. Ou seja, é correto considerar simultaneamente os erros amostrais sendo cometidos na pesquisa, considerando todas as C categorias ao mesmo tempo. Nesse caso, os intervalos de confiança são denominados simultâneos, ou regiões de confiança. Maiores detalhes sobre inferência simultânea em um contexto mais geral podem ser obtidos em Miller [1980].

O problema de utilizar intervalos de confiança individuais de $(1-\alpha)\%$ para cada estimativa de uma mesma pesquisa com C categorias, é que a inferência simultânea não tem a confiança desejada de $(1-\alpha)\%$. A probabilidade pré-experimental de todos os intervalos de confiança individuais conterem os parâmetros populacionais, **supondo independência entre eles**, é $(1-\alpha)^C$, que é a probabilidade pré-experimental de todos intervalos conterem os parâmetros populacionais. Por exemplo, para o caso com $C = 2$ e $\alpha = 0,05$, a confiança simultânea dessas intervalos é $90,25\%$, quando as confianças individuais são 95%. Quanto maior for o número de categorias C de uma pesquisa, pior é a cobertura simultânea desses intervalos, como pode ser visto no gráfico 1.36.

Para evitar esse problema, e calcular intervalos de confiança simultâneos com a cobertura correta, é necessário re-escrever a probabilidade 1.23, utilizando uma margem de erro d_i diferente para cada categoria da pesquisa:

$$P\left(\bigcap_{i=1}^{C}\left\{|\hat{P}_i - P_i| \leq d_i\right\}\right) \geq 1-\alpha, \tag{1.26}$$

onde d_i é a margem de erro para cada categoria individual. Vamos trabalhar simultaneamente com o α global e também com α_i de cada categoria, onde $P(|\hat{P}_i - P_i| \leq d_i) \geq 1-\alpha_i$. A maneira mais simples de trabalhar com a probabilidade em 1.26 é encontrar um limite inferior. Utilizando

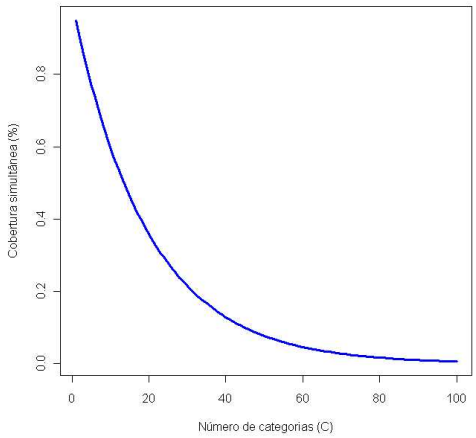

Figura 1.6: Cobertura simultânea utilizando IC individuais independentes com $1-\alpha = 0,95$

a desigualdade de Bonferroni, em Goodman [1965], o autor mostrou que:

$$\begin{aligned} P\left(\bigcap_{i=1}^{C}\left\{|\widehat{P}_i - P_i| \leq d_i\right\}\right) &\geq 1 - \sum_{i=1}^{C} P(|\widehat{P}_i - P_i| \geq d_i) \\ &\geq 1 - \sum_{i=1}^{C} \alpha_i. \end{aligned} \tag{1.27}$$

Ou seja, é possível obter um limite inferior para a probabilidade dos erros amostrais simultâneos em 1.26 somando as probabilidades dos erros amostrais individuais. Se quisermos a mesma confiança para o erro amostral de cada categoria, temos $\alpha_i = \frac{\alpha}{C}$, e assim obtemos que o nível de confiança simultâneo é dado por:

$$P\left(\bigcap_{i=1}^{C}\left\{|\widehat{P}_i - P_i| \leq d_i\right\}\right) \geq 1 - \sum_{i=1}^{C} \alpha_i = 1 - \frac{C\alpha}{C} = 1 - \alpha. \tag{1.28}$$

Por exemplo, se $\alpha_i = 5\%$ e $C = 5$, sabemos então que $1-\alpha \geq 0,75$, que é um nível de confiança muito baixo. Nesse mesmo exemplo, se quisermos que $1-\alpha$ seja próximo do nível usual de $0,95$, fazemos $\alpha_i = \frac{\alpha}{C} = \frac{0,05}{5} = 1\%$. Existem várias outras formas de se obter intervalos de confiança simultâneos, cada uma com suas vantagens e defeitos, algumas das quais serão apresentadas a seguir. Denotando $\mathbf{P} = (P_1, ..., P_{C-1})$ e $\hat{\mathbf{P}} = (\hat{P}_1, ..., \hat{P}_{C-1})$, em Gold [1963] o autor mostrou que para n grande o suficiente:

$$n\left(\hat{\mathbf{P}} - \mathbf{P}\right)' \hat{\mathbf{\Sigma}}^{-1} \left(\hat{\mathbf{P}} - \mathbf{P}\right) \sim \mathcal{X}^2_{C-1}, \tag{1.29}$$

onde $\mathcal{X}^2_{C-1}$ representa a distribuição qui-quadrado com $C-1$ graus de liberdade, e $\frac{1}{n}\hat{\mathbf{\Sigma}}^{-\mathbf{1}}$ é a matrix de covariância de $\hat{\mathbf{P}}$, onde os elementos da diagonal são dados por $\sigma_{ii} = P_i(1-P_i)$ e os fora da diagonal, são dados por $\sigma_{ij} = -P_iP_j$, com $i \neq j$. Desse resultado, a elipsóide

$$E_p = \left\{ \mathbf{P} : n\left(\hat{\mathbf{P}} - \mathbf{P}\right)' \hat{\mathbf{\Sigma}}^{-\mathbf{1}} \left(\hat{\mathbf{P}} - \mathbf{P}\right) \leq \mathcal{X}^2_{C-1,\alpha} \right\}, \tag{1.30}$$

é assintóticamente uma região de confiança de $(1-\alpha)\%$ para $\mathbf{P}$, onde $\mathcal{X}^2_{C-1,\alpha}$ é o α-quantil da distribuição $\mathcal{X}^2_{C-1}$. Projetando essa região em cada um dos eixos (com uma pequena adaptação para o caso $i = C$), obtemos que os intervalos de confiança simultâneos são:

$$\left[\hat{P}_i - \sqrt{\mathcal{X}^2_{C-1,\alpha}\left(\frac{\hat{P}_i(1-\hat{P}_i)}{n}\right)}; \hat{P}_i + \sqrt{\mathcal{X}^2_{C-1,\alpha}\left(\frac{\hat{P}_i(1-\hat{P}_i)}{n}\right)} \right] \qquad \forall i = 1,..,C. \tag{1.31}$$

Outra forma de se obter esses intervalos simultâneos foi apresentada em Quesenberry and Hurst [1964]. Utilizando a estatística qui-quadrado usual:

$$\sum_{i=1}^{C} \frac{(n_i - nP_i)^2}{nP_i}, \tag{1.32}$$

a qual assintoticamente tem uma distribuição $\mathcal{X}^2_{C-1}$, onde n_i é o número de observações na amostra pertencentes a categoria i, o autor mostra que os intervalos de confiança simultâneos assintóticos de $(1-\alpha)\%$ são dados por:

$$\left[\frac{\mathcal{X}^2_{C-1,\alpha} + 2n_j}{2(n + \mathcal{X}^2_{C-1,\alpha})} - d_i; \frac{\mathcal{X}^2_{C-1,\alpha} + 2n_j}{2(n + \mathcal{X}^2_{C-1,\alpha})} + d_i \right] \qquad \forall i = 1,..,C, \tag{1.33}$$

onde $d_i = \frac{\sqrt{\mathcal{X}^2_{C-1,\alpha}\left[\mathcal{X}^2_{C-1,\alpha} + 4n_j(n-n_j)/n\right]}}{2(n+\mathcal{X}^2_{C-1,\alpha})}$.

Em Goodman [1965], o autor mostra que os intervalos em 1.33 podem ser melhorados, ou seja, mantém a mesma cobertura ou confiança geral de $(1-\alpha)\%$, porém a região de confiança tem um volume menor, substituindo a quantidade $\mathcal{X}^2_{C-1,\alpha}$ por $\mathcal{X}^2_{1,\frac{\alpha}{C}}$.

Usualmente em pesquisas eleitorais, o erro amostral é divulgado como sendo "mais ou menos d %", ou seja, a prática comum dos institutos de pesquisa é divulgar uma única margem de erro para todas as categorias das pesquisas. No contexto de intervalos de confiança simultâneos, seria equivalente a fazer todas as margens de erro individuais iguais, com $d_i = d_G$ para toda categoria, ao invês de utilizar um d_i para cada categoria. Alguns autores denominam esses intervalos como intervalos de confiança simultâneos rápidos. Nesse caso, o interesse está na probabilidade:

$$P\left(\bigcap_{i=1}^{C}\left\{|\hat{P}_i - P_i| \leq d_G\right\}\right) \geq 1 - \alpha_G, \tag{1.34}$$

onde d_G é a margem de erro utilizada para todas as categorias e $1-\alpha_G$ é o nível de confiança global. Em Fitzpatrick and Scott [1987], os autores provaram que, para qualquer número de categorias C, se na probabilidade em 1.34 for utilizado a margem de erro d_G igual a margem de erro individual d obtida em 1.22, então $1-\alpha_G$ é aproximadamente $1-2\alpha$, sendo α aquele utilizado para obter d em 1.22, para n suficientemente grande. Esse resultado é enunciado no teorema 1.2.

Teorema 1.2 (Limite Inferior Assintótico para IC simultâneos) *Seja $P_1, P_2, ..., P_C$ as proporções populacionais definidas em 1.19, $\hat{P}_1, \hat{P}_2, ..., \hat{P}_C$ seus respectivos estimadores sob **AASc** definidos em 1.20, a margem de erro $d = z_{\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}$ e α o coeficiente de confiança utilizado para obter d. Para qualquer $C = 2, 3, ...$ temos que:*

$$\lim_{n\to\infty} P\left(\bigcap_{i=1}^{C}\left\{|\hat{P}_i - P_i| \leq d\right\}\right) \geq L(\alpha), \tag{1.35}$$

onde

$$L(\alpha) = \begin{cases} 1-2\alpha & \text{se } \alpha \leq 0,016 \\ 6\Phi\left(\frac{3z(1-\frac{\alpha}{2})}{\sqrt{8}}\right) - 5 & \text{se } 0,016 \leq \alpha \leq 0,15 \end{cases}$$

e $\Phi(.)$ representa a função de distribuição acumulada da normal-padrão.

Além disso, em Fitzpatrick and Scott [1987] os autores afirmam acreditar que no teorema 1.2 o limite inferior seja $1-2\alpha$ para qualquer valor de α, porém não conseguiram provar esse resultado ainda. Na prática, a diferença entre os dois limites é muito pequena. Supondo válidas as condições do teorema, obtemos que região de confiança simultânea para todos P_i's com confiança global de $(1-2\alpha)\%$ é dada por:

$$\left[\hat{P}_i - d; \hat{P}_i + d\right] \qquad \forall i = 1, .., C, \tag{1.36}$$

onde $d = z_{\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}$ para **AASc**. Se supormos que o teorema 1.2 também é válido para a **AASs**, a região de confiança é obtida usando $d = z_{\frac{\alpha}{2}}\sqrt{\frac{N-n}{N-1}}\frac{1}{2\sqrt{n}}$. Em Kwong [1998], o autor mostra que esses intervalos de confiança também podem ser melhorados ao se levar em conta explicitamente a covariância entre os estimadores considerados, porém não nos aprofundaremos mais nesse tema nessa tese.

Definir o tamanho de amostra considerando cada estimador separadamente já foi largamente discutido na literatura, como em Bolfarine and Bussab [2005], o mesmo não ocorrendo no caso simultâneo. Existem algumas alternativas para encontrar o tamanho da amostra levando em con-

sideração a probabilidade definida em 1.34, sendo os artigos Kwong [1998], Sison and Glaz [1995] e Bromaghin [1993] ótimas referências para um estudo mas aprofundando do tema. Nessa tese, consideramos apenas uma simples adaptação dos resultados usuais $n_{AASc} = \frac{1}{4\frac{d^2}{z^2_{\alpha/2}}}$ e $n_{AASs} = \frac{N}{4(N-1)\frac{d^2}{z^2_{\alpha/2}}+1}$ para **AAS** com e sem reposição, respectivamente, substituindo o percentil $z^2_{\alpha/2}$ por z^2_{α}, conforme discutido anteriormente. Assim, o tamanho da amostra considerando os erros amostrais simultaneamente, para o caso da **AASc** é:

$$n_{AASc} = \frac{1}{4\frac{d^2}{z^2_{\alpha}}}, \tag{1.37}$$

e para o caso da **AASs** é:

$$n_{AASs} = \frac{N}{4(N-1)\frac{d^2}{z^2_{\alpha}}+1}. \tag{1.38}$$

Intervalos de Confiança para AAS e o Empate Técnico

Os intervalos de confiança definidos em 1.25 são muito utilizados para se fazer **ID**, pois são uma forma simples de se comparar estimativas levando em consideração o desenho amostral. Cada estimativa de uma pesquisa têm um intervalo de confiança associado a ela. A forma como inferências para a população são realizadas nesse contexto de **ID** e dos intervalos de confiança pelos institutos de pesquisa é bem simples: se intervalos de confiança individuais com o mesmo nível de confiança de diferentes parâmetros populacionais tiverem alguma sobreposição, ou seja, contém alguns valores em comum, então conclui-se que não é possível afirmar que esses parâmetros populacionais são diferentes. Caso o contrário seja verdade, ou seja, não existir sobreposição dos difentes intervalos, então conclui-se que existe **uma diferença estatística significante**, ou seja, os parâmetros populacionais são diferentes.

A fundamentação teórica para os empates técnicos provém dos testes de hipóteses estatísticos da **ID**, que serão definidos a seguir.

Definição 1.2 *Um testes de hipóteses $\varphi : \chi \to 0, 1$ é uma regra de decisão para quais pontos $x \in \chi$ a hipótese $H : \theta \in \Theta_0$ deve ser rejeitada ($\varphi(x) = 1$) ou aceita ($\varphi(x) = 0$), onde φ é o conjunto em que o estimador da quantidade populacional de interesse assume valores, denominado espaço amostral, e Θ é o conjunto de valores onde o parâmetro populacional assume valores, e Θ_0 é o espaço paramétrico sob a hipótese H, com $\Theta = \Theta_0 \cup \Theta_1$. O subconjunto do espaço amostral que contém os pontos que levam a rejeição de H ($\varphi^{-1}(1)$) é chamado de região crítica ou região de rejeição do teste φ.*

Ao realizar um teste de hipóteses, pode-se cometer dois tipos de erros. Se $\theta \in \Theta_0$ mas, erroneamente, decide-se rejeitar H, então diz-se que foi cometido o erro do tipo I, por outro lado, se $\theta \in \Theta_1$

e decide-se pela não rejeiçao de H, o erro do tipo II foi cometido, como pode ser visto na tabela 1.4.

Tabela 1.4: tipos de erro em testes de hipóteses.

Decisão	$\theta \in \Theta_0$	$\theta \in \Theta_1$
Aceitar H	Decisão Correta	Erro Tipo II
Rejeitar H	Erro Tipo I	Decisão Correta

De acordo com a abordagem de Neyman-Pearson, uma forma de encontrar um teste razoável é fixar a probabilidade do erro tipo II, usualmente denotado por α, com $0 \geq \alpha \geq 1$, para todo $\theta \in \Theta_0$. Diz-se que tais testes tem nível de significância α. Usualmente a região crítica é também expressa por meio de uma estatística de teste $T(\{Y_i\}_{i\in s})$, que é utilizada para ordenar o espaço paramétrico de modo que grandes valores de T indicam pontos mais desfavoráveis a $H : \theta \in \Theta_0$, onde . Uma vez fixado o nível de significância α, pode-se determinar a região crítica de um teste encontrato o valor crítico k tal que

$$\alpha = \sup_{\theta\in\Theta_0} P(\varphi^{-1}(1)|\theta) = \sup_{\theta\in\Theta_0} P(T(\{Y_i\}_{i\in s}) \geq k|\theta). \tag{1.39}$$

O teorema 1.3, mostra que existe uma relação bem definida entre intervalos de confiança e testes de hipótese, que originou a forma de se fazer inferência pelos institutos de pesquisa, considerando somente os intervalos de confiança e sem explicitar qual teste de hipóteses está sendo realizado.

Teorema 1.3 (Relação entre Teste de Hipóteses e Intervalos de Confiança) *Para cada* $\theta_0 \in \Theta_0$ *seja* $A(\theta_0)$ *a região de aceitação (complementar a região de rejeição) do teste de nível* α *para testar* $H(\theta_0) : \theta = \theta_0$*, e para cada ponto do espaço amostral* x *seja* $S(x)$ *o conjunto de valores do parâmetro tal que*

$$S(x) = \{\theta : x \in A(\theta), \theta \in \Theta_0\} \tag{1.40}$$

Então $S(x)$ *é uma família de intervalos de confiança para* θ *com nível de confiança* $1 - \alpha\%$.

Ou seja, o intervalo de confiança com confiança de $1 - \alpha$ para um parâmetro P corresponde a região de não-rejeição de um teste de hipóteses para $H : P = P_0$ de nível α, e dessa forma, se um intervalo de confiança gerado a partir de uma amostra específica conter P_0, a decisão tomada é a mesma ao utilizar o teste, de não-rejeitar a hipótese H. E, analogamente, se o intervalo específico não conter o parâmetro, rejeita-se a hipótese H.

Nessa sessão estamos interessados no teste da hipótese de igualdade entre dois parâmetros populacionais P_i e P_j, com $i \neq i$, por exemplo. Essa hipótese pode ser definida como $H : P_i = P_j = P$ ou $H : P_i - P_j = 0$. A estatística de teste usualmente recomendada é $T(\{Y_i\}_{i\in s}) = \hat{P}_i - \hat{P}_j$.

A região de rejeição desse teste é dada por $|\hat{P}_i - \hat{P}_j| > k$. Nesse caso, rejeita-se a hipótese se a diferença em módulo entre $\hat{P}_i$ e $\hat{P}_j$ for maior que k, onde k depende do nível do teste α. Supondo válida-se as condições do teorema 1.1, a distribuição da estátistica do teste, supondo independência entre $\hat{P}_i$ e $\hat{P}_j$, pode ser aproximada por:

$$\hat{P}_i - \hat{P}_j \sim \mathcal{N}\left(P_i - P_j, V(\hat{P}_i) + V(\hat{P}_j) + Cov(\hat{P}_i, \hat{P}_j)\right). \tag{1.41}$$

Supondo independência entre $\hat{\mathbf{P}}_\mathbf{i}$ e $\hat{\mathbf{P}}_\mathbf{j}$, o mesmo teste pode ser formulado considerando o intervalo de confiança de cada estimador separadamente, pois quando $\hat{P}_i - \hat{P}_j > 0$, implica que $\hat{P}_i > \hat{P}_j$ e quando $\hat{P}_i - \hat{P}_j < 0$, implica que $\hat{P}_i < \hat{P}_j$. Ou seja, a região de rejeição de ambas as formulações para o teste são da mesma forma. Como a variância da diferença é idêntica a soma das duas variâncias individuais sob a suposição de independência, a decisão tomada em ambas as formulações é igual. No caso considerando os estimadores separadamente, a hipótese de igualdade não é rejeitada para uma amostra específica, quando existe uma intersecção não-vazia dos intervalos de confiança com confiança de nível $1-\alpha$ individuais de $\hat{\mathbf{P}}_\mathbf{i}$ e $\hat{\mathbf{P}}_\mathbf{j}$. **No linguajar dos institutos de pesquisa, os quais fazem inferência baseada no Desenho e utilizam os intervalos de confiança individuais supondo independência, esse tipo de situação é denominada de empate técnico.**

Diversos problemas com essa metodologia podem ser apontados, discutiremos apenas alguns deles aqui. Primeiramente, **se a amostra for grande o suficiente**, nenhum intervalo de confiança terá intersecção com outro intervalo, o que levaria a conclusão de que todos os parâmetros populacionais na população sendo estudada são diferentes. Mas o que, precisamente, quer dizer que os parâmetros populacionais são diferentes? Uma diferença detectada apenas na centésima casa decimal significa que os parâmetros populacionais são diferentes? Qual a relevância dessa diferença no contexto que ela está sendo estudada? Em Huff [1954], o autor cunhou a seguinte frase, bastante relevante nesse contexto: *"Uma diferença, só é uma diferença, se ela faz diferença."*. Em ensaios clínicos e de bioequivalência, por exemplo, define-se um parâmetro δ, denominado de margem clinicamente importante, e apenas consideram-se como relevantes, diferenças maiores do que δ, como pode ser visto em Wada and Andrade [2010]. No contexto expecífico de pesquisas eleitorais, essa crítica não tem muita relevância, pois numa eleição, se um candidato tiver apenas um único voto válido a mais que o candidato concorrente, ele será eleito. Ou seja, nesse caso, qualquer diferença faz diferença.

O segundo problema é a existência de empates técnicos na pesquisas eleitorais, quando é praticamente nula a possibilidade de dois candidatos estarem realmente empatados numa disputa eleitoral, ou seja, terem exatamente o mesmo número de votos válidos no momento em que a pesquisa é realizada. Essa crítica está relacionada com o tipo de inferência utilizada pelos institutos de pesquisa. Um exemplo prático desse problema ocorreu nas eleições presidenciais de 2010, quando a notícia

Figura 1.7: Notícia sobre empate técnico nas eleições presidenciais de 2010

na figura 1.7 foi veículada na mídia[2]. Utilizando a metodologia dos institutos de pesquisa, não é possível prever qual candidato está liderando a disputa pois segundo a pesquisa do CNT/Sensus, o intervalo de confiança de 95% do percentual de votos para Serra (P_{Serra}) dado por $[30,2; 36,2]$ e para Dilma (P_{Dilma}) dado por $[24,8; 30,8]$, têm uma intersecção não-vazia. Um ponto importante é que, mesmo quando uma pesquisa afirma que há um empate técnico entre dois candidatos, isso não quer dizer que ambos os candidatos têm a mesma chance de ganhar as eleições. Em Zabala [2009], o autor discute com profundidade o que é o empate técnico, e prova que, do ponto de vista de inferência Bayesiana, quando ocorre empate técnico a probabilidade do candidato que com proporção estimada maior ganhar converge para $1-\alpha$ quando $n \to \infty$. Esse outro tipo de inferência será abordado na Seção 1.4. A importância desse resultado é que permite aos institutos de pesquisa sempre inferirem quem está ganhando a disputa eleitoral, nunca havendo a necessidade de recorrer ao empate técnico.

O terceiro problema, novamente, é que os intervalos de confiança utilizados pelos institutos de pesquisa consideram cada categoria P_i separadamente, sem levar em consideração que $\sum_{i=1}^{C} P_i = 1$ e que os estimadores $\hat{P}_i$ e $\hat{P}_j$, para $i \neq j$ são correlacionados. Existem diferentes formas de levar em conta essa covariância ao se fazer **ID**. Uma maneira bem simples de se comparar duas estimativas, apresentada em Scott and Seber [1983] e discutida no início dessa sessão, é considerar o intervalo de confiança para a diferença $\hat{P}_i - \hat{P}_j$. Nesse caso, temos que a variância da diferença é dada, no caso da **AASc**, por:

$$\begin{aligned} Var(\hat{P}_i - \hat{P}_j) &= Var(\hat{P}_i) + Var(\hat{P}_j) - 2Cov(\hat{P}_i, \hat{P}_j) \\ &= \frac{P_i + P_j - (P_i - P_j)^2}{n}, \end{aligned}$$

a qual incorpora explicitamente a covariância dos dois estimadores. No caso da **AASs**, essa variância deve ser multiplicada por $\frac{N-n}{N-1}$.

No contexto de intervalos de confiança da diferença, o resultado é considerado como empate técnico se o zero estiver contido no intervalo. Utilizando os mesmos resultados obtidos na pesquisa da CNT/Sensus na figura 1.7, porém considerando a covariância entre os estimadores, obtemos um intervalo de confiança de 95% para a diferença $P_{Serra} - P_{Dilma}$ dado por $[1,99; 8,81]$, o qual não

[2]Site: *www.jusbrasil.com.br/politica/4775723/cnt − sensus − da − empate − tecnico − entre − serra − e − dilma*

contém o 0, ou seja, utilizando explicitamente a covariância, é possível afirmar que o candidato Serra está na liderança da disputa.

Além dos institutos de pesquisa estarem fazendo uma suposição claramente errada ao utilizar intervalos de confiança individuais e independentes, é possível mostrar que dessa forma eles inflacionam a incidência de empates técnicos. Para avaliar quando o empate técnico ocorre em cada um dos casos (com e sem levar em conta a covariância dos estimadores), é preciso definir formalmente o que é o empate técnico. O empate técnico entre os candidatos i e j ocorre, no primeiro caso (sem considerar a covariância entre os candidatos) quando os extremos dos intervalos de confiança para as quantidades P_i e P_j, com $P_i + P_j \leq 1$, dados respectivamente por $(\hat{P}_i - \xi_i; \hat{P}_i + \xi_i)$ e $(\hat{P}_j - \xi_j; \hat{P}_j + \xi_j)$ com $\xi_i = z_{\frac{\alpha}{2}}\sqrt{\frac{P_i(1-P_i)}{n}}$ e $\xi_j = z_{\frac{\alpha}{2}}\sqrt{\frac{P_j(1-P_j)}{n}}$ no caso da **AASc**, se intersectam, ou seja, quando $\hat{P}_i + \xi_i > \hat{P}_j - \xi_j$ se $\hat{P}_i < \hat{P}_j$ ou $\hat{P}_j + \xi_j > \hat{P}_i - \xi_i$ se $\hat{P}_j < \hat{P}_i$. De ambas essas condições, obtemos que se $\frac{z_{\frac{\alpha}{2}}\left(\sqrt{\frac{P_i(1-P_i)}{n}}+\sqrt{\frac{P_j(1-P_j)}{n}}\right)}{|\hat{P}_i - \hat{P}_j|} > 1$, ocorrerá o empate técnico.

Já no segundo caso (considerando a covariância entre os candidatos), o empate técnico ocorre quando o intervalo de confiança para a quantidade $P_i - P_j$, com $P_i + P_j \leq 1$, dado por $(\hat{P}_i - \hat{P}_j - \xi_{ij}; \hat{P}_i - \hat{P}_j + \xi_{ij})$, onde $\xi_{ij} = z_{\frac{\alpha}{2}}\sqrt{\frac{P_i(1-P_i)}{n} + \frac{P_j(1-P_j)}{n} + \frac{2P_iP_j}{n}}$ no caso da **AASc**, contém o 0, ou seja, se $\hat{P}_i - \hat{P}_j - \xi_{ij} < 0$ quando $\hat{P}_i > \hat{P}_j$ ou se $\hat{P}_i - \hat{P}_j + \xi_{ij} > 0$ quando $\hat{P}_i < \hat{P}_j$. De ambas essas condições, obtemos que se $\frac{z_{\frac{\alpha}{2}}\sqrt{\frac{P_i(1-P_i)}{n}+\frac{P_j(1-P_j)}{n}+\frac{2P_iP_j}{n}}}{|\hat{P}_i - \hat{P}_j|} > 1$, ocorrerá o empate técnico.

Assim, resumindo, quanto maior for a quantidade $d_{indep} = \sqrt{P_i(1-P_i)} + \sqrt{P_j(1-P_j)}$, maior a chance de ocorrer empate técnico no primeiro caso, sem a covariância entre os candidatos. De forma equivalente, quanto maior for a quantidade $d_{cov} = \sqrt{(P_i(1-P_i)) + (P_j(1-P_j)) + (2P_iP_j)}$, maior a chance de ocorrer empate técnico no segundo caso, considerando a covariância entre os candidatos. Numericamente é fácil ver que $d_{indep} > d_{cov}$ para quaisquer P_i e P_j tais que $P_i + P_j \leq 1$, ou seja, sempre é mais fácil a ocorrência do empate técnico se a covariância entre as categorias não for considerada. Na figura 1.8, onde desenhamos $\frac{d_{indep}}{d_{cov}}$ para todas as possíveis combinações de P_i e P_j, é fácil de ver que essa razão sempre é maior do que 1. Dessa figura, é evidente que quanto menor forem as quantidades P_i e P_j, mais importante é a inclusão da covariância entre os estimadores. Já no caso onde essas quantidades são próximas de $0,5$, não há um ganho efetivo ao se considerar essa covariância.

Existem outras formas de calcular os intervalos de confiança levando em consideração a correlação entre as categorias. Por exemplo, em Cochran [1977] na página 60, o autor considera a distribuição condicional de $\hat{P}_i/(\hat{P}_i + \hat{P}_j)$. O empate técnico ainda pode ocorrer mesmo levando em consideração a covariância entre os estimadores, e o problema pode ser mais grave do que declarar empate técnico entre dois candidatos, é possível que todos os candidatos sejam declarados empatados, mesmo existindo uma grande diferença entre o primeiro e o último colocado. Essa situação ocorre quando intervalos de confiança individuais são utilizados e pares de candidatos consecutivos estão empatados, ou seja, quando o candidato de ranking j está empatado com o candidato de ranking $j+1$, para todo $j \in \{1, ..., C-1\}$. A única forma de efetivamente acabar com o problema do empate técnico de forma completa é utilizando inferência Bayesiana baseada em modelos

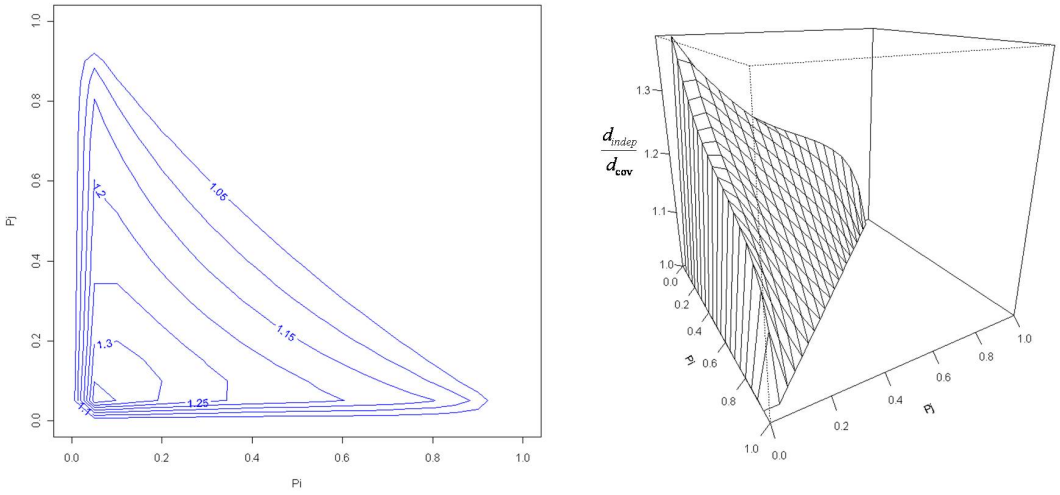

Figura 1.8: $\frac{d_{indep}}{d_{cov}}$ para todas as possíveis combinações de P_i e P_j

(IBM), descrita na Seção 1.4.2, pois mesmo considerando a covariância entre os estimadores ou os intervalos de confiança simultâneos da **ID**, esse problema continua existindo.

1.2.3 Estratificação e Pós-Estratificação

Nas seções anteriores, as únicas quantidades conhecidas eram Y_i para as unidades populacionais pertencentes a amostra. Nessa seção, iremos supor que existe uma covariável X, cujos valores são conhecidos para todas as unidades populacionais, independentemente destas pertencerem à amostra ou não.

Amostragem estratificada, ou estratificação, é uma forma de selecionar a amostra levando em consideração a existência de grupos na população. Esse tipo de amostragem é denominada de estratificada pois usualmente se divide a população em H grupos ou estratos, e esses grupos são controlados durante a seleção da amostra. Ou seja, usualmente a covariável X sendo controlada divide a população em grupos. Se X for uma variável contínua, ela pode ser discretizada em um número conveniente de categorias, permitindo que amostragem estratificada seja utilizada.

Existem muitas formas diferentes de se estratificar uma amostra. Essa estratificação pode ser explícita, onde os H grupos da população são exatamente replicados na amostra ao se fazer inferência, ou implícita, onde os valores de X são aproximadamente controlados na amostra, esse tipo de amostragem também é conhecida amostragem sistemática, e será discutida na Seção 1.2.4. Os estratos podem ter seus tamanhos n_h fixados antes da seleção da amostra, ou seja, o tamanho de cada estrato não é aleatório, ou esses tamanhos podem ser aleatórios, e após a seleção da amostra utiliza-se uma técnica conhecida como pós-estratificação, ou ponderação, para fazer com que esses estratos representem corretamente a covariável X.

Estratificação

Vamos supor que a covariável X determina os H grupos, sendo que cada grupo h tem N_h unidades populacionais pertencentes a ele, e total de cada grupo é dado por τ_h. Dessa forma

podemos escrever o total populacional como sendo $\tau_Y = \sum_{h=1}^{H}\sum_{i=1}^{N_h} Y_{h,i} = \sum_{h=1}^{H}\tau_h$, onde $Y_{h,i}$ indica o valor da variável Y para a i-ésima unidade populacional do h-ésimo estrato. Supondo que as quantidades n_h, que indicam o tamanho da amostra para cada estrato, sejam fixadas de antemão de forma que $n = \sum_{h=1}^{H} n_h = n$, podemos escrever o estimador do total populacional como:

$$\hat{\tau}_{est} = \sum_{h=1}^{H} N_h \sum_{i\in s_h} \frac{Y_{h,i}}{n_h} = \sum_{h=1}^{H} \hat{\tau}_h, \tag{1.42}$$

onde $i \in s_h$ indica as unidades populacionais pertencentes a amostra do estrato h e $\hat{\tau}_h$ é o estimador do parâmetro populacional τ_h. É fácil mostrar que o estimador $\hat{\tau}_{est}$ é não-viciado e que a sua variância é dada por:

$$Var(\hat{\tau}_{est}) = \sum_{h=1}^{H} Var(\hat{\tau}_h), \tag{1.43}$$

onde $Var(\hat{\tau}_h) = N_h^2 \frac{\sigma_h^2}{n_h}$ se for utilizado o desenho **AASc** e $Var(\hat{\tau}_h) = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$ se for utilizado o desenho **AASs**, sendo a variância populacional do estrato h dada por $\sigma_h^2 = \frac{\sum_{i=1}^{N_h}(Y_{hi}-\mu_h)^2}{N_h}$, a média populacional do estrato h dada por $\mu_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h}$ e $S_h^2 = \frac{N_h}{N_h-1}\sigma_h^2$.

Como vimos na Seção 1.2.2, também é importante obter um estimador de $Var(\hat{\tau}_{est})$. Esse estimador é dado por:

$$\hat{Var}(\hat{\tau}_{est}) = \sum_{h=1}^{H} \hat{Var}(\hat{\tau}_h), \tag{1.44}$$

onde $\hat{Var}(\hat{\tau}_h) = N_h^2 \frac{s_h^2}{n_h}$ se for utilizado o desenho **AASc** e $\hat{Var}(\hat{\tau}_h) = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}$ se for utilizado o desenho **AASs**, com $s_h^2 = \frac{\sum_{i\in s_h}(Y_{hi}-\hat{\mu}_h)^2}{n_h-1}$ e $\hat{\mu}_h = \frac{\sum_{i\in s_h} Y_{hi}}{n_h}$.

Os tamanhos amostrais fixos n_h podem ser definidos, ou alocados, de várias formas diferentes. As formas mais comuns são a alocação proporcional ao tamanho, onde $n_h = n\frac{N_h}{N}$, alocação uniforme, onde $n_h = \frac{n}{H}$ e a alocação ótima de Neyman, que minimiza a variância do estimador estratificado do total populacional, ou seja, essa forma de alocação nos indica quais devem ser as quantidades n_h de forma a obter o estimador da média populacional com a menor variância possível considerando um custo fixo. Supondo que a pesquisa tem um custo fixo de forma linear, dado por $C' = \sum_{h=1}^{H} c_h n_h$, onde c_h é o custo por unidade observada no estrato h, a alocação ótima é dada por $n_h = n\frac{N_h \frac{\sigma_h}{\sqrt{c_h}}}{\sum_{h=1}^{H} N_h \frac{\sigma_h}{\sqrt{c_h}}}$, para o caso da **AASc**. Esses resultados podem ser facilmente adaptados para o caso onde o interesse está em estimar o total populacional. Mais detalhes sobre as diferentes formas de alocação podem ser obtidos em Bolfarine and Bussab [2005]. As principais vantagens da amostragem estratificada são:

Controle de Covariável Controlar covariáveis importantes que podem estar correlacionadas com

a variável de interesse Y.

Variância do Estimador Na maioria dos casos, para um mesmo tamanho de amostra n, a variância do estimador estratificado usando alocação ótima ou proporcional é menor ou igual a variância do estimador sem estratificação.

Estratos de Leitura Garantir um tamanho de amostra mínimo para um sub-grupo da população para o qual se deseja fazer inferência.

A importância de se controlar covariáveis correlacionadas com a variável Y fica bastante evidente num exemplo do Paradoxo de Simpson, apresentado em Lindley and Novick [1981]. Consideremos que está sendo testado um tratamento T, para reduzir a taxa de morte causada por uma certa doença, e foi realizado um estudo comparando os efeitos do tratamento T com um placebo. Os resultados foram resumidos na tabela 1.5. Por esses resultados, o tratamento T parece reduzir em 10% a taxa de morte comparado com o placebo, ou seja, o tratamento T parece ser eficiente.

Tabela 1.5: Efeito do tratamento T nas taxas de morte

Tratamento Utilizado	Número de Mortes	Número de Sobreviventes	Total Pacientes	Taxa de Morte
T	20	20	40	50%
Placebo	24	16	40	60%

Existe, no entanto, uma covariável X que não foi considerada no desenho desse estudo. Na tabela 1.6 foram resumidos os resultados do experimento considerando-se a covariável X. O tratamento utilizado está correlacionado com a covariável X, pois na ausência da covariável $(X = 0)$ foram tratados com o tratamento T apenas 10 pacientes, e na presença da covariável $(X = 1)$, 30 pacientes foram tratados com o tratamento T. Essa covariável também está associada com a variável Y, pois quando $(X = 0)$ as taxas de morte são bem maiores do que quando $(X = 1)$.

Analisando as taxas de morte dentro das duas classes, presença (X=1) e ausência (X=0) da covariável, percebemos que na verdade o tratamento T aumenta a taxa de morte. Para o caso $X = 1$, temos que ela aumenta de 30% com o placebo para 40% com o tratamento T, e para o caso $X = 0$, temos que ela aumenta de 70% com o placebo para 80% com o tratamento T. Ou seja, na realidade, o tratamento T aumenta a taxa de morte em exatamente a mesma quantidade que ele supostamente diminuía a taxa de morte ao analisarmos somente os dados da tabela 1.5, desconsiderando a covariável X.

Desse exemplo fica evidente que a alocação de tratamentos nesse experimento deveria ser realizada garantindo o controle da covariável X, assim evitando esse tipo de efeito de confundimento, que compromete as inferências realizadas a partir dessa amostra. O mesmo tipo de situação pode ocorrer em amostragem, assim existindo o interesse de se utilizar amostragem estratificada controlando a covariável X. Apesar desse exemplo ser artificial, existem diversos exemplos reais onde o Paradoxo de Simpson foi observado, como em Appleton et al. [1996] e em Freedman et al. [1978].

Um exemplo bastante simples e intuitivo desse paradoxo é na atuação dos bombeiros, onde correlação entre o número de bombeiros enviados para um incêndio e os prejuízos causados pelos incêndios é positiva, o que parece paradoxal. Porém, são enviados mais bombeiros para os grandes incêndios. Se você controle para o tamanho do incêndio, o sinal da correlação se torna negativo, ou seja, considerando somente incêndios de mesmo tamanho, quanto mais bombeiros, menor é o prejuízo.

Tabela 1.6: Efeito do tratamento T nas taxas de morte, controlando a covariável X

Covariável X	Tratamento Utilizado	Número de Mortes	Número de Sobreviventes	Total Pacientes	Taxa de Morte
1	T	12	18	30	40%
	Placebo	3	7	10	30%
0	T	8	2	10	80%
	Placebo	21	9	30	70%

Quanto a redução da variância do estimador, é possível mostrar que utilizando alocação proporcional, com $n_h = n\frac{N_h}{N}$, no caso da **AASc** temos:

$$Var(\tau_{A\hat{A}Sc}) = Var(\hat{\tau}_{est}) + \frac{N}{n}\sum_{h=1}^{H} N_h\left(\mu_h - \mu\right)^2, \tag{1.45}$$

ou seja, a variância do estimador $\hat{\tau}_{est}$ é sempre menor ou igual a variância do estimador $\tau_{A\hat{A}Sc}$, pois $\frac{N}{n}\sum_{h=1}^{H} N_h\left(\mu_h - \mu\right)^2$ é uma quantidade positiva. Mais que isso, quanto mais diferentes forem as médias μ_h entre si, mais eficiente será o estimador estratificado. Para o caso da **AASs** utilizando alocação proporcional, é possível se chegar a mesma conclusão supondo que N e N_h sejam grandes o suficiente para que $\frac{N}{N-1} \approx 1$ e $\frac{N_h}{N_h-1} \approx 1$, assim obtendo:

$$Var(\tau_{A\hat{A}Ss}) = Var(\hat{\tau}_{est}) + \frac{N}{n}\left(1 - \frac{n}{N}\right)\sum_{h=1}^{H} N_h\left(\mu_h - \mu\right)^2. \tag{1.46}$$

No caso do estimador da média populacional para **AASc**, é fácil mostrar que

$$Var(\hat{\tau_{ot}}) = Var(\hat{\tau}_{prop}) + \frac{1}{n}\sum_{h=1}^{H} \frac{N_h}{N}\left(\sigma_h - \overline{\sigma}\right)^2, \tag{1.47}$$

onde $\overline{\sigma} = \sum_{h=1}^{H} \frac{N_h}{N}\sigma_h$. Porém, encontrar resultados similares para o caso do estimador do total e de **AASs** é bem mais complicado, apesar de empiricamente sabermos que resultados similares são válidos.

Pós-Estratificação ou Ponderação

A pós-estratificação não deve ser considerada como um plano amostral em si, apenas uma forma de fazer inferência baseada no desenho. Apesar disso, ficará evidente dos resultados que serão apresentados nessa seção a vantagem de se selecionar uma amostra estratificada a priori comparado a tentar corrigir possíveis viéses a posteriori, ou seja, discutir a pós-estratificação é uma forma de mostrar a vantagem de se utilizar amostragem estratificada.

A diferença entre pós-estratificação e estratificação é que no caso da pós-estratificação, as quantidades n_h não são pré-fixadas. Ou seja, a estratificação da amostra não é realizada antes da mesma ser selecionada. Nesta seção, trabalharemos somente com o caso da **AASs** numa população estratificada, porém sem utilizar amostragem estratificada, ou seja, a covariável de estratificação X é conhecida para toda a população e as quantidades N_h também são conhecidas, porém opta-se por não utilizar um desenho amostral estratificado. De 1.42, temos que a variância do estimador $\hat{\tau}_{est}$ é dada por

$$Var(\hat{\tau}_{est}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2. \tag{1.48}$$

Para o caso da alocação proporcional onde $n_h = n\frac{N_h}{N}$, obtemos

$$Var(\hat{\tau}_{est}) = N \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} N_h S_h^2. \tag{1.49}$$

O estimador de pós-estratificação, denotado aqui por $\hat{\tau}_{ps}$, tem exatamente a mesma forma de $\hat{\tau}_{est}$ condicionado a $\mathbf{n} = (n_1, ..., n_H)$, é não viciado, porém a sua variância é diferente de 1.48. Isso ocorre porque, no caso do $\hat{\tau}_{ps}$, o vetor $\mathbf{n}$ também é uma variável aleatória. Para ver isso, primeiro note que a variância de $\hat{\tau}_{ps}$ condicionada a $\mathbf{n}$ é dada por:

$$Var(\hat{\tau}_{ps} \mid \mathbf{n}) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2, \tag{1.50}$$

onde $\mathbf{n}$ denota o vetor $(n_1, ..., n_H)$ e $Var(X \mid \mathbf{n})$ é a variância de X condicionada a $\mathbf{n}$. Ou seja, a variância condicionada a $\mathbf{n}$ de $\hat{\tau}_{ps}$ é igual a variância de $\hat{\tau}_{est}$. Porém, pensando em **ID** e na variância de $\hat{\tau}_{ps}$ na replicação de todas as possíveis amostras, é necessário calcular variância incondicional. Em Smith [1991], o autor mostra que:

$$
\begin{aligned}
Var(\hat{\tau}_{ps}) &= \sum_{h=1}^{H} N_h^2 \left(E\left(\frac{1}{n_h}\right) - \frac{1}{N_h} \right) S_h^2 \\
&\approx N\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} N_h S_h^2 + \frac{1}{n}\left(\frac{1}{n} - \frac{1}{N}\right) \sum_{h=1}^{H} (N - N_h) S_h^2, \qquad (1.51)
\end{aligned}
$$

ou seja, a variância do estimador $\hat{\tau}_{ps}$ é sempre maior ou igual a variância de $\hat{\tau}_{est}$ para o caso da alocação proporcional, ou seja, é mais eficiente controlar a amostra á priori, antes de selecioná-la, do que corrigí-la posteriormente a sua seleção. Apesar disso, a pós-estratificação é uma forma bastante utilizada de corrigir distorções na amostra efetivamente observada e evitar problemas como o Paradoxo de Simpson. Existem diversos artigos questionando se a variância em 1.51 é correta para o estimador. Algums acreditam que deveria ser utilizada a variância condicionada a $\mathbf{n}$, e nesse caso a variâncias de $\hat{\tau}_{ps}$ e de $\hat{\tau}_{est}$ seriam iguais. Maiores detalhes podem ser obtidos em Holt and Smith [1979].

É comum, na prática, usar a pós-estratificação para corrigir as distorções amostrais de todas as covariáveis conhecidas na população. Uma aplicação dessa técnica, também conhecida como ponderação iterativa[3], pode ser encontrada em Izrael et al. [2000].

1.2.4 Amostragem por Conglomerados e Amostragem Sistemática

Os planos amostrais vistos até agora sorteiam diretamente as unidades populacionais. Porém na prática, em muitas situações é impossível obter uma listagem que permita a seleção de cada unidade separadamente, principalmente em se tratando de populações humanas.

Nas pesquisas de intenção de voto, onde a unidade populacional é o ser humano, é muito difícil obter uma listagem de todos os eleitores de um determinado município, para não se falar do caso das eleições presidenciais, onde seria necessária uma listagem com todos os eleitores do Brasil.

Nesses casos, é comum selecionar conglomerados ou grupos de pessoas, denominados unidades primárias, ao invés de selecionar cada pessoa uma a uma para a amostra. Usualmente os conglomerados são geográficos, como por exemplo um quarteirão ou um setor específico de um mapa cartográfico de uma cidade. A amostragem por conglomerados consiste em selecionar a conglomerados e entrevistar todas as pessoas residentes nesses conglomerados. Nesse contexto, as pessoas são denominadas unidades secundárias. Existe ainda uma variação desse desenho amostral, denominada Amostragem por Conglomerados em dois estágios, onde somente parte de cada conglomerado é entrevistada.

A amostragem por conglomerados usualmente é menos eficiente do que a AAS, mesmo assim ela é bastante utilizada. Uma grande vantagem desse desenho amostral é a logística de coleta de dados, que é amplamente facilitada, pois o entrevistador só precisa se locomover dentro de um mesmo conglomerado para realizar as entrevistas.

Em pesquisas eleitorais realizadas no Brasil, usualmente é utilizado o setor censitário, definido

[3] Em inglês, essa técnica é conhecida como Raking.

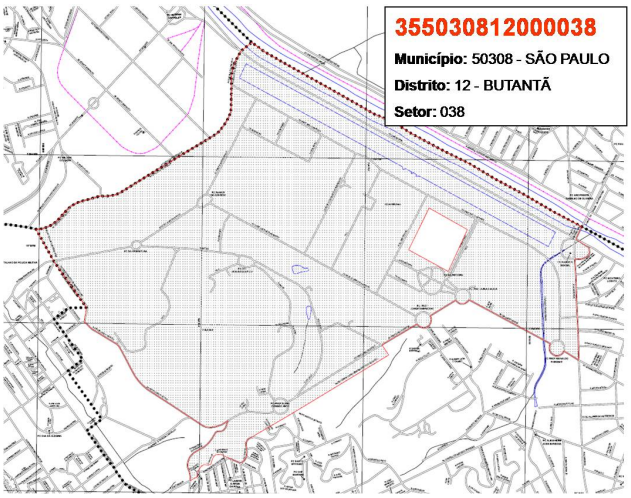

Figura 1.9: Mapa do Setor Censitário que contém a USP (área hachurada)

pelo Instituto Brasileiro de Geografia e Estatística (IBGE), como o conglomerado do desenho amostral, pois essa é a menor unidade geográfica do Brasil para as quais existem informações oficiais, permitindo que desenhos amostrais possam utilizar informações populacionais para sortear uma amostra mais eficiente. O Setor Censitário é definido pelo **IBGE** como:

> O território brasileiro está subdividido em recortes administrativos (estados, municípios, distritos e subdistritos). Por conta de suas dimensões, o que implicaria grandes deslocamentos por parte dos agentes censitários e pesquisadores, estes recortes são ainda mais subdivididos, em aproximadamente 150 estabelecimentos agropecuários, 200 domicílios ou área de 500 km^2 para os setores rurais e 300 domicílios para os setores urbanos. **Estas porções menores do território brasileiro são denominadas Setores Censitários**. Seus limites são estabelecidos pelo IBGE e também obedecem aos alinhamentos dos recortes anteriores, isto é, um limite de setor não cruza um limite de distrito, por exemplo.

Para se ter uma idéia da quantidade de setores censitários existentes no Brasil, a cidade de São Paulo é composta por mais de 13.000 setores. Na figura 1.9, destacamos no mapa o setor censitário (SC) que contém a Universidade de São Paulo (USP). Cada SC é identificado por um número de 15 algarismos, denominado código do setor censitário. Na figura 1.9, o código do setor desenhado aparece destacado.

As propriedades teóricas dos estimadores sob amostragem por conglomerados são facilmente derivadas a partir da AAS. Nessa seção, iremos apresentar a teoria somente no contexto de **AASs**, porém as propriedades teóricas para o caso da **AASc** podem ser facilmente obtidas. Denotando por A o número de conglomerados e por B_α o total de unidades secundárias no conglomerado α, temos $\tau_Y = \sum_{\alpha=1}^{A} \sum_{i=1}^{B_\alpha} Y_{\alpha i} = \sum_{\alpha=1}^{A} \tau_\alpha$, onde $Y_{\alpha i}$ representa a i-ésima unidade populacional do conglomerado α e $\tau_\alpha = \sum_{i=1}^{B_\alpha} Y_{\alpha i}$ é o total do conglomerado α.

Nesse contexto, estimador do total populacional é análogo ao estimador do total no caso da **AASs**, porém imaginando que as unidades populacionais são os conglomerados. Dessa forma, para uma amostra de tamanho a, obtemos:

$$\hat{\tau}_{conglo} = A\frac{\sum_{\alpha\in s_a}\tau_\alpha}{a} = A\hat{\overline{\tau}}, \tag{1.52}$$

onde s_a é o conjunto do índices dos conglomerados que pertencem a amostra e $\hat{\overline{\tau}}$ é o total populacional médio dos conglomerados pertencentes a amostra. Temos então que $\hat{\tau}_{conglo}$ é um estimador não-viciado e que a sua variância é dada por:

$$Var(\hat{\tau}_{conglo}) = A^2\left(1-\frac{a}{A}\right)\frac{\sum_{\alpha=1}^{A}\left(\tau_\alpha-\overline{\tau}\right)^2}{(A-1)a} = A^2\left(1-\frac{a}{A}\right)\frac{S_A^2}{a}, \tag{1.53}$$

onde $\overline{\tau} = \frac{\sum_{\alpha=1}^{A}\tau_\alpha}{A}$ é o total populacional médio dos conglomerados e $S_A^2 = \frac{\sum_{\alpha=1}^{A}(\tau_\alpha-\overline{\tau})^2}{(A-1)}$ é a variância dos totais populacionais. Um estimador não-viciado para $Var(\hat{\tau}_{conglo})$ é dado por:

$$\hat{Var}(\hat{\tau}_{conglo}) = A^2\left(1-\frac{a}{A}\right)\frac{\sum_{\alpha\in s_a}\left(\tau_\alpha-\hat{\overline{\tau}}\right)^2}{(a-1)a} = A^2\left(1-\frac{a}{A}\right)\frac{s_A^2}{a}. \tag{1.54}$$

Note que nesse contexto, o tamanho da amostra é aleatório, pois dependerá de quais conglomerados foram incluídos na amostra. O tamanho da amostra é dado por $n = \sum_{\alpha\in s_a} B_\alpha$.

No contexto de amostragem por conglomerados, é importante avaliar a sua eficiência comparada a **AAS**. Para comparar os dois desenhos amostrais, primeiro é necessário decompor S^2, quantidade que determina a variância do estimador de total da **AASs**, da seguinte forma:

$$\begin{aligned}
S^2 &= \frac{1}{N-1}\sum_{\alpha=1}^{A}\sum_{i=1}^{B_\alpha}(Y_{\alpha i}-\mu)^2 \\
&= \frac{1}{N-1}\sum_{\alpha=1}^{A}\sum_{i=1}^{B_\alpha}(Y_{\alpha i}-\mu_\alpha)^2 + \frac{1}{N-1}\sum_{\alpha=1}^{A}\sum_{i=1}^{B_\alpha}(\mu_\alpha-\mu)^2 \\
&= \frac{1}{N-1}\sum_{\alpha=1}^{A}B_\alpha\sigma_\alpha^2 + \frac{1}{N-1}\sum_{\alpha=1}^{A}B_\alpha(\mu_\alpha-\mu)^2 \\
&= \sigma_{dc}^2+\sigma_{ec}^2,
\end{aligned} \tag{1.55}$$

onde $\mu_\alpha = \frac{1}{B_\alpha}\sum_{i=1}^{B_\alpha}Y_{\alpha i}$ e $\sigma_\alpha^2 = \frac{1}{B_\alpha}\sum_{i=1}^{B_\alpha}(Y_{\alpha i}-\mu_\alpha)^2$. Assim, podemos perceber que S^2 é composta de duas componentes, σ_{dc}^2 representando a variância dentro dos conglomerados e por σ_{ec}^2 representando a variância entre os conglomerados. Pode-se mostrar também que

$$S_A^2 = \frac{\sum_{\alpha=1}^{A} \overline{B}^2 \left(\frac{B_\alpha}{\overline{B}}\mu_\alpha - \mu\right)^2}{(A-1)}, \tag{1.56}$$

onde $\overline{B}$ é o tamanho médio dos conglomerados. Ou seja, a quantidade S_A^2 é muito parecida com o termo σ_{ec}^2 de S^2, assim quanto maior for a contribuição de σ_{ec}^2 para S^2, menor será a eficiência da amostragem de conglomerados com relação a **AASs**. Essa característica desse tipo de amostragem é conhecida como efeito de conglomeração.

Para avaliar quando isso ocorre, é comum utilizar o coeficiente de correlação intra-classe ρ_{int}, definido em Bolfarine and Bussab [2005]. Esse coeficiente mede o grau de similaridade das unidades populacionais dentro dos conglomerados. Quanto maior for esse coeficiente, mais similares são as unidades dentro de um mesmo conglomerado, e mais diferentes são as unidades de diferentes conglomerados.

Para o caso onde todos os conglomerados têm o mesmo tamanho, ou seja, quando $B_\alpha = B$, esse coeficiente assume valores entre 1 e $-\frac{1}{B-1}$, sendo que o valor 1 indica homogeneidade máxima entre as unidades de um mesmo conglomerado, ou seja, todos são iguais, e o valor $-\frac{1}{B-1}$ indica heterogeinidade máxima, onde cada conglomerado é uma microrepresentação do universo. É possível mostrar para o caso com reposição, que $Var(\hat{\tau}_{conglo}) = [1 + (B-1)\rho_{int}]\frac{\sigma^2}{aB}N^2$. Desses resultados, obtemos que:

$$EPA(\hat{\tau}_{AASc}, \hat{\tau}_{conglo}) = 1 + \rho_{int}(B-1), \tag{1.57}$$

assim quanto mais homogêneos forem os conglomerados, menos eficiente é a amostragem por conglomerados. Geralmente esse é o caso, ou seja, as unidades populacionais dentro de um mesmo conglomerado são mais aprecidas entre si, como pode ser visto em Pessoa and Silva [1998]. Esse é o motivo que, do ponto de vista da **ID**, ao analisar uma amostra proveniente da amostragem de conglomerados, deve se levar em consideração o efeito de conglomeração. Se o desenho amostral for ignorado, que equivale nesse caso a estimar $Var(\hat{\tau}_{conglo})$ como se fosse uma **AAS**, a variância do estimador provavelmente será sub-estimada pois geralmente $\rho_{int} > 0$, o que implica que as conclusões obtidas do ponto de vista da **ID** podem ser equivocadas[4]. Resultados similares aos apresentados nessa seção podem ser obtidos para o caso geral, onde o tamanho do conglomerado α é dado por B_α.

Amostragem Sistemática (Estratificação Implícita)

Suponha que o tamanho da populaçao N seja igual a kn, onde k e n são números inteiros, sendo n o tamanho da amostra e também suponha que essas unidades populacionais estão ordenadas segundo algum critério. A amostragem sistemática consiste em selecionar, utilizando AAS,

[4]Como consequência disso, os intervalos de confiança terão amplitude menor do que deveriam ter, o que fará com que algumas diferenças não-significativas pareçam significativas.

uma unidade populacional entre as k primeiras unidades, e depois selecionar sistematicamente as unidades populacionais em intervalos de k unidades. Supondo que a primeira unidade selecionada foi k', onde $1 \leq k' \leq k$, serão selecionadas as unidades $k', 2k', ..., nk'$.

Em Madow and Madow [1944], os autores mostram que a amostragem sistemática pode ser vista como um caso especial da amostragem por conglomerados. Isso ocorre pois existem k possíveis amostras, cada uma delas podendo ser interpretada como um conglomerado, e a amostra consiste de apenas um desses conglomerados. A vantagem desse desenho amostral é que os conglomerados são criados pelo estatístico, através da ordenação das unidades populacionais segundo alguma variável de interesse, permitindo que características conhecidas da população sendo estudada possam ser controladas durante a seleção da amostra. Cada conglomerado terá unidades populacionais de todas as faixas de valores da covariável utilizada na ordenação. Como os conglomerados são criados pelo estatístico, de forma indireta, o estatístico também controla ρ_{int}. Ou seja, dependendo da relação entre a covariável e a variável de interesse Y, a amostragem sistemática pode ser mais eficiente que a AAS. Por causa dessa possibilidade de controle de covariáveis de interesse, a amostragem sistemática também é conhecida como Estratificação Implícita.

Dessa forma, para obter as propriedades teóricas de uma amostra sistemática de tamanho n, utilizamos os resultados de amostragem com conglomerados com $a = 1$, $A = k$ e $\overline{B} = n$ obtemos:

$$\hat{\tau}_{sis} = k \sum_{i=1}^{n} Y_{k'i}, \tag{1.58}$$

onde k' indica o conglomerado que foi selecionado. Temos então que $\hat{\tau}_{sis}$ é um estimador não-viciado para o total populacional e que a sua variância é dada por:

$$Var(\hat{\tau}_{sis}) = k^2 \left(1 - \frac{1}{k}\right) \frac{\sum_{\alpha=1}^{k} (\tau_\alpha - \overline{\tau})^2}{k-1} = k(k-1)S_k^2, \tag{1.59}$$

onde $\overline{\tau} = \frac{\sum_{\alpha=1}^{k} \tau_\alpha}{k}$ é o total populacional médio dos conglomerados. Como $a - 1 = 0$, nesse contexto, não existe estimador não-viciado para $Var(\hat{\tau}_{sis})$. Apesar de ser um estimador viciado, usualmente utiliza-se:

$$\hat{Var}(\hat{\tau}_{sis}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_Y^2}{n}. \tag{1.60}$$

Existem formulações alternativas da amostragem sistemática que permitem obter estimadores não-viciados para a variância, e também permitem que as propriedades teóricas dos estimadores da amostragem sistemática possam ser obtidas mesmo quando a suposição de que N seja igual a kn, onde k e n são números inteiros, não é satisfeita. Algumas dessas alternativas são encontradas em Bolfarine and Bussab [2005] e Wolter [1985].

Amostragem por Conglomerados em 2 estágios

A amostragem por conglomerados em 2 estágios consiste em selecionar a conglomerados no primeiro estágio, e dentro de cada conglomerado selecionado, selecionar b pessoas residentes nesses conglomerados. Dessa forma, o tamanho da amostra é dado por $n = a * b$ unidades populacionais. Esse tipo de amostragem também pode ser definida selecionando-se b_α unidades populacionais no conglomerado α pertencente a amostra, mas aqui apresentaremos somente o caso mais simples com $b_\alpha = b$.

A principal vantagem desse tipo de amostragem é que, para um mesmo tamanho amostral, muito mais conglomerados são selecionados do que no caso da amostragem de conglomerados usual, o que garante uma maior dispersão das unidades da amostra, e mais especificamente no caso de pesquisas eleitorais, quando o setor censitário é utilizado como conglomerado, uma maior dispersão geográfica. Além disso, quando as unidades populacionais dentro de um mesmo conglomerado são muito parecidas, selecionar todas as unidades de um mesmo conglomerado trás informação redundante. Nesse caso também é melhor selecionar mais conglomerados porém com menos unidades populacionais dentro de cada um.

Aqui discutiremos somente a amostragem por conglomerados em 2 estágios para o caso da **AASc**, porém resultados similares podem ser obtidos tanto para mais estágios, quanto para a **AASs** em Nascimento [1981]. Além disso, o caso mais geral de amostragem por conglomerados em 2 estágios com probabilidades desiguais será discutido com bastante detalhe na Seção 4.2.1.

Segundo as especificações amostrais discutidas aqui, para estimar o total populacional da variável Y de uma amostra de a conglomerados e b unidades populacionais em cada conglomerado, usuamente utiliza-se o seguinte estimador não-viciado:

$$\hat{\tau}_{conglo2} = A\frac{\sum_{\alpha\in s_a^1}\hat{\tau}_\alpha}{a} = \frac{A}{a}\sum_{\alpha\in s_a^1}\frac{B_\alpha}{b}\sum_{i\in s_{b,\alpha}^2} Y_{\alpha i}, \tag{1.61}$$

onde s_a^1 é o conjunto de a índices dos conglomerados que pertencem a amostra do estágio 1, $s_{b,\alpha}^2$ é o conjunto dos b índices das unidades populacionais do conglomerado α que pertencem a amostra do estágio 2 e $\hat{\tau}_\alpha = \frac{B_\alpha}{b}\sum_{i\in s_{b,\alpha}^2} Y_{\alpha i}$ é o estimador do total populacional do conglomerado α. A variância de $\hat{\tau}_{conglo2}$ é dada por:

$$Var(\hat{\tau}_{conglo2}) = \frac{A}{a}\sum_{\alpha=1}^{A}(\tau_\alpha - \overline{\tau})^2 + \frac{A}{a}\sum_{\alpha=1}^{A} B_\alpha^2\frac{\sigma_\alpha^2}{b}, \tag{1.62}$$

onde $\sigma_\alpha^2 = \frac{\sum_{i=1}^{B_\alpha}(Y_{\alpha i}-\mu_\alpha)^2}{B_\alpha}$ é a variância dentro do conglomerado α e $\mu_\alpha = \frac{1}{B_\alpha}\sum_{i=1}^{B_\alpha} Y_{\alpha i}$ é a média do conglomerado α. Note que a variância do estimador $\hat{\tau}_{conglo2}$ é composta por duas partes, uma relativa a variância entre os conglomerados, e outra relativa a média das variâncias dentro dos conglomerados. Um estimador não-viciado para essa variância é dado por:

$$\hat{Var}(\hat{\tau}_{conglo2}) = \frac{A^2}{a}\frac{\sum_{\alpha\in s_a^1}\left(\hat{\tau}_\alpha - \frac{\hat{\tau}_{conglo2}}{A}\right)^2}{(a-1)}. \quad (1.63)$$

Para entender melhor o comportamento do estimador $\hat{\tau}_{conglo2}$, ele pode ser re-escrito como função do coeficiente de correlação intra-classe ρ_{int} discutida na Seção 1.2.4. Supondo que $B_\alpha = B$, ou seja, que todos os conglomerados têm o mesmo tamanho, e além disso que $\frac{B-1}{B} \simeq 1$ e $\frac{1}{B} \simeq 0$, obtemos que:

$$Var(\hat{\tau}_{conglo2}) \simeq (1+\rho_{int}(b-1))\frac{\sigma^2}{ab}N^2, \quad (1.64)$$

ou seja, quanto mais homogêneos forem os conglomerados, maior será a variância do estimador, e quanto mais heterogêneos forem, menor será a variância do estimador. É interessante, nesse contexto, comparar a amostragem por conglomerados em 1 e 2 estágios. Para que essa comparação seja justa, é preciso comparar amostras de mesmo tamanho, pois no primeiro caso selecionamos a conglomerados com amostra de tamanho B, e no segundo selecionamos a' conglomerados com amostras de tamanho b. Supondo que $aB = a'b$ obtemos:

$$EPA(\hat{\tau}_{conglo2}, \hat{\tau}_{conglo}) \simeq \frac{[1+\rho_{int}(b-1)](\frac{\sigma^2}{a'b}N^2)}{[1+\rho_{int}(B-1)](\frac{\sigma^2}{aB}N^2)} = \frac{[1+\rho_{int}(b-1)]}{[1+\rho_{int}(B-1)]}, \quad (1.65)$$

onde é fácil ver que, se $\rho_{int} > 0$, a amostragem por conglomerados em 2 estágios é mais eficiente do que em 1 estágio. Esse resultado é bastante intuitivo, pois se as unidades populacionais dentro de um conglomerado são mais parecidas entre si, é melhor incluir mais conglomerados na amostra com um amostra menor dentro de cada um deles.

1.2.5 Amostragem Inversa

A amostragem inversa é justificada no contexto de estimar uma proporção P_i na população de interesse, como descrito na Seção 1.2.2. Assim, supondo que existam C categorias, estamos interessados em estimar as quantidades populacionais:

$$P_i = \frac{\sum_{j=1}^N Y_j^{(i)}}{N},$$

onde $Y_j^{(i)} = 1$ se a j-ésima unidade populacional pertencer a categoria i, e $Y_j^{(i)} = 0$ caso contrário. O estimador de P_i é dado por:

$$\hat{P}_i = \frac{\sum_{j\epsilon s} Y_j^{(i)}}{n},$$

onde n é o tamanho da amostra,que é fixo. Porém, se a proporção de interesse P_i for muito pequena, por exemplo quando estamos interessados em estimar a proporção de alguma característica populacional muito rara, Haldane [1945] propôs um método inverso, conhecido como amostragem inversa, onde ao invês de fixar n e observar quantos $Y_j^{(i)} = 1$, fixa-se o número de unidades populacionais com essa característica que deseja-se observar, digamos m com $(m > 1)$, e observa-se o tamanho da amostra necessária para obter essas m unidades populacionais. Procedendo dessa forma, um estimador não-viciado é dado por $\hat{P}_i^{Hald} = \frac{m-1}{n-1}$. Se N for grande, e P_i for pequeno, temos que $Var(\hat{P}_i^{Hald})$ é aproximadamente $\frac{mP_i^2(1-P_i)}{(m-1)^2}$. Um estimador para essa variância é dado por $\hat{Var}(\hat{P}_i^{Hald}) = \frac{m(\hat{P}_i^{Hald})^2(1-\hat{P}_i^{Hald})}{(m-1)^2}$.

As vantagens desse tipo de amostragem são garantir a existência de um número mínimo de unidade amostrais contento a característica de interesse e diminuir o erro relativo do estimador de P_i, conforme mostrado em Cochran [1977], na página 76.

1.2.6 Amostragem com Probabilidades Desiguais

Nesta seção discutiremos desenhos amostrais onde as probabilidades de seleção das unidades populacionais são desiguais. Existe o interesse nesse tipo de desenho amostral pois ele é bastante utilizado na prática. Por exemplo, em pesquisas eleitorais, é comum que a seleção de setores censitários seja feita de forma que as probabilidades de seleção de cada setor sejam proporcionais ao tamanho.

Vamos supor que as probabilidades de seleção em um único sorteio p_i e as probabilidades de inclusão π_i, definidas na Seção 1.2.1, são conhecidas para todas as unidades populacionais. Note que essas probabilidades podem ser geradas por qualquer desenho amostral, não precisam ser aquelas calculadas para a **AASc** e **AASs**.

Essas probabilidades são obtidas do desenho amostral utilizado. Por exemplo, suponha que o interesse está em um desenho amostral com reposição, onde as probabilidade de seleção sejam proporcionais a uma covariável Z conhecida para toda a população. Nesse caso, temos então que $p_i = \frac{Z_i}{\sum_{j=1}^{N} Z_j}$, ou seja, as probabilidades de seleção em um único sorteio são geralmente fáceis de serem obtidas. As probabilidades de inclusão são usualmente mais difíceis de serem calculadas, porém são necessárias se existe o interesse em se obter estimadores não-viciados para desenhos amostrais sem reposição. Em Fellegi [1963], o autor mostra que uma amostra pode ser selecionada sequencialmente sem reposição utilizando as mesmas probabilidades p_i da amostragem com reposição. Primeiramente, a probabilidade da unidade populacional i_2 ser a segunda unidade selecionada condicionado a unidade i_1 ter sido a primeira unidade selecionada é dada por:

$$\frac{p_{i_2}}{1 - p_{i_1}}, \tag{1.66}$$

e analogamente, a probabilidade da unidade i_n ser a n-ésima unidade populcional selecionada condicionado as unidades i_1, i_2,...,i_{n-1} terem sido previamente selecionadas é dada por:

$$\frac{p_{i_n}}{1 - p_{i_1} - ... - p_{i_{n-1}}} \tag{1.67}$$

Assim, a probabilidade incondicional $\delta_i(k)$ de que a i-ésima unidade populcaional seja a k-ésima unidade selecionada é dada por:

$$\delta_i(k) = \sum_{k-1,i} p_{i_1} \frac{p_{i_2}}{1 - p_{i_1}} ... \frac{p_{i_{k-1}}}{1 - p_{i_1} - ... - p_{i_{k-2}}} \frac{p_i}{1 - p_{i_1} - ... - p_{i_{k-1}}} \tag{1.68}$$

onde $\sum_{k-1,i}$ é o somatório de todas as $(k-1)$-tuplas $(i_1, ..., i_{k-1})$ tais que $i_1, ..., i_{k-1}$ são diferentes inteiros entre 1 e N, nenhum deles sendo iguais a i. Dessa forma, a probabilidade de inclusão π_i é dada por:

$$\pi_i = \sum_{k=1}^{n} \delta_i(k). \tag{1.69}$$

No contexto de amostragem com probabilidades desiguais, os estimadores não-viciados do total populacional utilizados para amostragem com e sem reposição são diferentes. O estimador $\hat{\tau}_{HH}$ para o caso com reposição, conhecido como estimador de Hansen-Hurwitz (HH), foi apresentado em Hansen and Hurwitz [1943], já o estimador $\hat{\tau}_{HT}$ para o caso sem reposição, conhecido como estimador de Horvitz-Thompson (HT), foi apresentado em Horvitz and Thompson [1952]. Os estimadores, suas variâncias e estimadores para as suas variâncias são apresentados na tabela 1.7.

Tabela 1.7: Estimadores Não-Viciados para Amostragem com Probabilidades Desiguais

Desenho Amostral	Com Reposição (HH)	Sem Reposição (HT)
Estimador	$\hat{\tau}_{HH} = \sum_{i \in s} \frac{Y_i}{np_i}$	$\hat{\tau}_{HT} = \sum_{i \in s} \frac{Y_i}{\pi_i}$
$Var(\hat{\tau})$	$\frac{1}{n} \sum_{i=1}^{N} p_i \left(\frac{Y_i}{p_i} - \tau \right)^2$	$\sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} Y_i^2 + \sum_{i=1}^{N} \sum_{j=1 \neq i}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j$
$\hat{Var}(\hat{\tau})$	$\frac{1}{n(n-1)} \sum_{i \in s} \left(\frac{Y_i}{p_i} - \hat{\tau}_{HH} \right)^2$	$\sum_{i \in s} \frac{1-\pi_i}{\pi_i^2} Y_i^2 + \sum_{i \in s} \sum_{j \in s \neq i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} Y_i Y_j$

Para o estimador $\hat{\tau}_{HH}$ é fácil ver que se as probabilidades de seleção p_i forem proporcionais a Y_i, ou seja, $p_i = \frac{Y_i}{\tau}$, então a variância do estimador é 0. Isso que dizer que se a variável de interesse Y fosse conhecida e utilizada na seleção da amostra, esse estimador seria muito eficiente, pois para toda possível amostra a estimativa seria igual ao total populacional. Claramente, essa situação nunca ocorrerá, pois se Y fosse conhecido, não haveria a necessidade de se fazer amostragem. Apesar disso, é evidente que utilizar probabilidades de seleção proporcionais a uma covariável conhecida é uma forma de controlar essa covariável na amostra. O mesmo resultado pode ser obtido para o estimador $\hat{\tau}_{HT}$ quando o tamanho da amostra é fixo.

Os estimadores $\hat{\tau}_{HH}$ e $\hat{\tau}_{HT}$ são usualmente vistos como uma forma de unificar a amostragem para populações finitas no contexto de **ID**. Para ver isso, note que substituindo as quantidades p_i, π_i e π_{ij} da **AASc** e **AASs** nos respectivos estimadores com e sem reposição, obtemos os mesmos resultados derivados na Seção 1.2.1.

Ao comparar as propriedades dos dois estimadores, percebe-se que a variância de $\hat{\tau}_{HT}$ é muito mais complicada de ser estimada, pois depende também das probabilidades de inclusão conjuntas π_{ij}. Por causa da dificuldade em se estimar essa variância, é bastante comum utilizar $\hat{Var}(\tau_{\hat{HH}})$ como uma aproximação para $\hat{Var}(\tau_{\hat{HT}})$. Em Wolter [1985], são apresentadas diversas correções que podem ser utilizadas para melhorar essa aproximação.

1.2.7 Amostragem Complexa

A amostragem complexa é a combinação de diferentes tipos de desenhos amostrais em uma única amostra. Por exemplo, é bastante comum que desenhos amostrais sejam de conglomerados em dois estágios, com a seleção das unidades primárias sendo estratificada, com probabilidades proporcionais ao tamanho e seleção sistemática.

A motivação para utilizar amostragem complexa provém das informações disponíveis, as quais geralmente tornam impossível a utilização de um desenho **AAS**, e da necessidade de controlar covariáveis de interesse buscando tornar amostra o mais parecida possível com a população de interesse.

Por exemplo, no contexto de pesquisas eleitorais, usualmente é possível obter informações de orgãos oficiais, como o IBGE, permitindo que a seleção dos setores censitários no primeiro estágio seja feita levando em consideração diversas características de interesse, como número de moradores, renda média do setor, etc... Já na seleção das unidades secundárias, usualmente não existe informação, assim o estatístico é obrigado a selecionar amostras sem controlar covariáveis de interesse. Ou seja, em uma mesma amostra complexa, diferentes desenhos amostrais podem ser utilizados em cada estágio.

Todas as características de um desenho amostral complexo podem ser incluídas nas probabilidades de seleção p_i e de inclusão π_i, permitindo que os estimadores $\hat{\tau}_{HH}$ e $\hat{\tau}_{HT}$ apresentados na tabela 1.7 sejam utilizados no contexto de amostragem complexa. Em Särndal et al. [1992], mostra-se como obter essas probabilidades para diversos desenhos amostrais de interesse.

Uma forma de evitar utilizar o estimadores de HH e de HT ao se desenhar uma amostra complexa são as amostras auto-ponderdas, que são amostras onde todas as unidades populacionais têm a mesma probabilidade de resposta. Esse tipo de amostra é bastante discutida em Kish [1965]. Pensando em amostragem por conglomerados em 2 estágios de populações humanas, a forma mais comum de selecionar uma amostra auto-ponderada é selecionar as unidades populacionais em todos os estágios, menos no último, com probabilidades proporcionais ao tamanho. No último estágio, utilizam-se probabilidades uniformes para todos os moradores. Por exemplo, no contexto de amostragem de conglomerados, denotando o tamanho dos conglomerados por B_j e o tamanho dos domicílios é dados por D_{ji}, onde $\sum_i D_{ji} = B_j$, obtemos que a probabilidade de selecionar uma determinada pessoa do domicílio i do conglomerado j é dada por:

$$n\frac{B_j}{\sum_j B_j}\frac{D_{ji}}{B_j}\frac{1}{D_{ji}} = \frac{n}{\sum_j B_j} = \frac{n}{N} \qquad \forall i, j,$$

pois $\sum_j B_j = N$, ou seja, a probabilidade de seleção é constante para toda unidade populacional, e assim os estimadores pontuais da **AAS** podem ser utilizados, porém para o cálculo das variâncias ainda é necessário utilizar as fórmulas específicas para o desenho utilizado. A mesma idéia pode ser aplicada para desenhos amostrais com 3 estágios ou mais.

1.3 Amostragem Probabilística na Prática e o Erro Não Amostral

Na prática, a amostra selecionada utilizando um desenho amostral probabilístico quase nunca é igual a amostra efetivamente observada, principalmente no contexto de populações humanas. Essa diferença ocorre por causa dos erro não-amostrais. Ou seja, erros não-amostrais são todos os erros que ocorrem em uma pesquisa exceto aqueles justificados pela seleção de uma amostra da população de interesse. Na Seção 1.2 discutimos como calcular e estimar o erro amostral, porém devida a natureza do erro não-amostral, não é possível fazer o mesmo com o erro não-amostral.

Nessa seção discutiremos as diferentes fontes de erros não-amostrais, as diferentes formas de medir e classificar o mesmo, além dos procedimentos usualmente utilizados para evitá-los e corrigí-los. A intenção é apresentar uma breve introdução sobre o tema, para o leitor mais interessado recomendamos a leitura de Groves [1989], Lessler and Kalsbeek [1992] e dos Capítulos 14 a 17 de Särndal et al. [1992].

Diferentes erros não-amostrais ocorrem em diferentes etapas de uma pesquisa. Para facilitar a distinção entre os diferentes tipos de erros, cada uma das etapas de uma pesquisa na pratica serão brevemente descritas a seguir:

Planejamento da Pesquisa Nessa etapa define-se qual tipo de pesquisa será realizada, quais são os objetivos, os recursos disponíveis e prazos para execução de cada etapa. Os tipos de pesquisa usualmente consideradas são por correio, pela internet, pelo telefone e pessoais.

Seleção da Amostra Nessa etapa define-se quais covariáveis serão utilizadas para selecionar a amostra, qual tipo de amostragem será utilizada e quais domínios de leitura devem ser considerados ao determinar os estratos da amostra, além de efetivamente selecionar a amostra.

Preparação do Questionário Essa etapa compreende o desenvolvimento dos questionários e formulários que serão utilizados para mensurar as variáveis de interesse. Fatores relevantes como tempo levado para completar o questionário, se será auto-preenchimento ou não devem ser levados em consideração.

Coleta dos Dados Nessa etapa inclui-se todos os esforços para a coleta de dados, como as entrevistas em si, verificações feitas por supervisores para evitar fraudes na coleta e no procedimento de seleção dos respondentes, as intruções recebidas pelos entrevistadores sobre

como selecionar e contactar os respondentes, sobre como aplicar o questionário e sobre como conduzir as entrevistas, etc...

Entrada dos Dados e Codificação Nessa etapa inclui-se a codificação de perguntas abertas, digitar os dados para que possam ser utilizados em um computador, verificação de consistência interna das perguntas dos questionários, limpeza dos dados, etc...

Análise dos Dados Nessa etapa inclui-se tentar responder as perguntas objetivo da pesquisa, determinar quais estimadores e pesos amostrais devem ser utilizados, quais tipos de análises estatísticas e modelos probabilísticos serão utilizados., etc...

Documentação e Publicação Quando for o caso, publicar os resultados, disponibilizar os dados e os pesos necessários para se fazer inferência, além de documentar detalhadamente todas as etapas envolvidas na pesquisa.

Como pode ser visto, existem muito mais etapas em uma pesquisa do que meramente o desenho amostral e a inferência estatística, que foram descritos com detalhes na Seção 1.2. E erros não-amostrais ocorrem em todas as etapas descritas, e apesar do nome, não significam que ocorrem somente porque a pesquisa foi mal executada, existem várias situações onde eles são inevitáveis.

1.3.1 Tipos de Erro

De uma forma geral, os erros podem ser classificados em duas grandes categorias, os erros de observação e não-observação. Os de observação ocorrem quando a unidade populacional é de fato observada, porém a quantidade de interesse Y não é corretamente mensurada. Já os erros de não-observação ocorrem porque unidades populacionais não foram observadas, ou seja, a quantidade de interesse Y não chegou a ser mensurada. Na figura 1.10 são resumidas as possíveis fontes de erro usualmente consideradas em uma pesquisa.

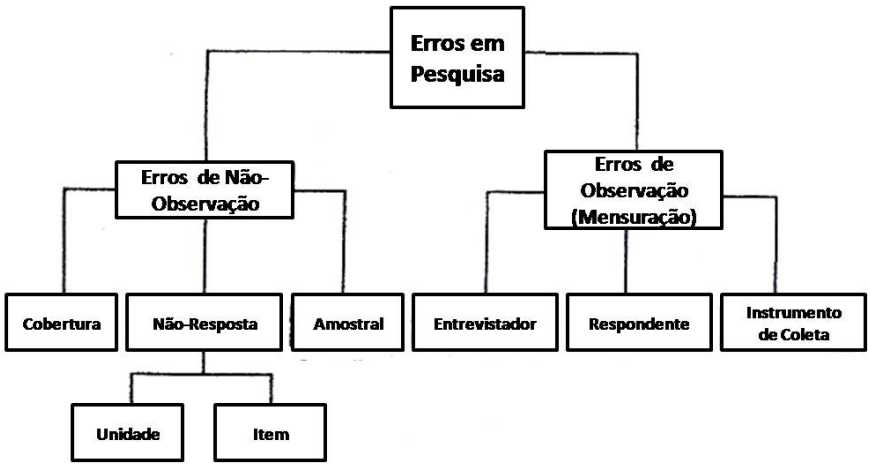

Figura 1.10: Tipos de Erros em Pesquisas

A seguir, faremos uma breve descrição de cada fonte de erro, com alguns exemplos de erros que podem ocorrer:

Cobertura Erros de Cobertura ocorrem quando unidades populacionais são excluídas da população a qual pertencem, consequentemente não tendo chances de pertencer a amostra. Isso geralmente ocorre quando a listagens das unidades populacionais utilizadas para selecionar as amostras estão incompletas ou desatualizadas. Por exemplo, ao selecionar setores censitários na amostragem por conglomerados, novos bairros acabam sendo excluídos do universo e moradores desses locais não têm chance de pertencer a amostra, mesmo fazendo parte do universo de interesse.

Não-Resposta O erro de não-resposta ocorre quando não se consegue mensurar as quantidades de interesse Y de unidades populacionais selecionadas para pertencer a amostra. Esse tipo de erro pode ser separado em duas categorias: não resposta da unidade, quando nenhuma das variáveis de interesse são mensuradas, e do item, quando algumas das variáveis de interesse não são mensuradas. A não-resposta do item ocorre quando uma unidade populacional selecionada não é encontrada ou se recusa a responder. A não-resposta do item ocorre quando a unidade selecionada se recusa ou não sabe a resposta para um determinado item do instrumento de coleta.

Amostral O erro amostral ocorre porque unidades populacionais não são incluídas na amostra.

Erro de Mensuração Esse tipo de erro ocorre quando a variável Y não é mensurada corretamente. Esse erro pode ter diversas fontes, como o entrevistador, o respondente e o instrumento de coleta (questionário). O entrevistador pode induzir a resposta do respondente ou simplesmente não aplicar corretamente o questionário, o respondente pode responder equivocadamente (com ou sem intenção) e o questionário pode ser mal formulado, não permitindo que o respondente compreenda a pergunta ou causando confusão no mesmo. Esse erro também pode ser causado por problemas decorrentes da entrada de dados e codificações.

Apesar da importância de todos os erros apresentados nessa sessão, além do erro amostral que foi discutido com detalhes em 1.2, apenas discutiremos com alguma profundidade o erro de não-resposta da unidade. Um dos grandes problemas referentes aos erros não-amostrais é a dificuldade de avaliar o grau de importância de cada um, e quanto os resultados da pesquisa são afetados por eles. É comum associar a uma determinada pesquisa somente o erro amostral cometido, porém é bem possível que os erros não-amostrais tenham um impacto muito maior sobre a precisão das estimativas do que o próprio erro amostral.

1.3.2 Erro Não-Resposta da Unidade e a Probabilidade de Resposta

Nessa seção serão discutidas questões relacionadas ao erro de não-resposta da unidade, especificamente, quando ele ocorre, como pode ser mensurado e como lidar com o erro de não-resposta de forma a minimizar seu impacto nas estimativas populacionais.

Razões para a Não-Resposta

Dependendo do tipo de pesquisa realizada, diferentes motivos para a não-resposta podem existir. Por exemplo, em pesquisas domiciliares, uma entrevista com uma pessoa selecionada pode não ser

realizada pois o domicílio onde a pessoa selecionada reside não foi localizado, sendo que em uma pesquisa telefônica esse motivo não ocorre.

Na literatura muitas nomenclaturas diferentes são utilizadas para especificar os mesmos erros de não-reposta. Para faciliar a explicação dos diferentes tipos de erro, utilizaremos a nomenclatura apresentada a seguir:

Inelegível A unidade populacional é classificada como inelegível se ela não satisfaz aos filtros da pesquisa, ou seja, se ela não pertence a população de interesse. Por exemplo, em pesquisas eleitorais não há o interesse em entrevistar pessoas que terão menos de 16 à época da eleição.

Não-Contactado A unidade populacional é classificada como Não-Contactada quando ela não é localizada, de forma que o contato nunca é realizado. Em pesquisas domiciliares, isso ocorre quando o domicílio não é encontrado, quando a pessoa reside em um apartamento ou condomínio no qual não é permitido acesso ao entrevistador, ou quando a pessoa não está em casa.

Contactado porém Recusou A unidade populacional é classificada como Contactada porém Recusou quando o contato com a pessoa é feito, porém ela se recusa a responder ao questionário.

Contactado porém Incapaz A unidade populacional é classificada como Contactada porém Incapaz quando o contato com a pessoa é realizado porém a pessoa é incapaz de responder ao questionário, seja por não ter disponibilidade de tempo no momento ou por ter algum problema mental, emocional, físico ou de comunicação.

Outras Razões A unidade populacional é classificada como outras Razões quando não se encaixa nas categorias anteriores.

As categorias **Inelegível** e **Contactado porém Recusou** usualmente são consideradas definitivas, ou seja, quando uma pessoa é classificada dessa forma, não se tenta mais fazer contato com a mesma. Já no caso das categorias **Não-Contactado**, **Contactado porém Incapaz** e **Outras Razões** é possível tentar fazer contato em algum outro momento, ou seja, é comum que o entrevistador volte a tentar fazer contato/entrevistar essa pessoa, como será discutido na Seção 1.3.3.

Em Groves [1989], o autor analisa diversas pesquisas com o objetivo de avaliar quando é mais provavél que uma pessoa esteja em casa, ou seja, como é possível evitar o erro de não-resposta da categoria **Não-Contactado**. Para exemplificar um dos resultados empíricos obtidos pelo autor, na figura 1.11 apresentamos a proporção de domicílios pesquisados, por hora do dia, onde algum morador declarou estar em casa e acordado, ou seja, horários do dia nos quais seria possível contactar o morador. Dessa figura fica evidente que os melhores horários para tentar fazer contato com as pessoas selecionadas em pesquisas domiciliares são de manhã cedo e no fim da tarde, pelo menos com relação aos erros da do tipo **Não-Contactado**.

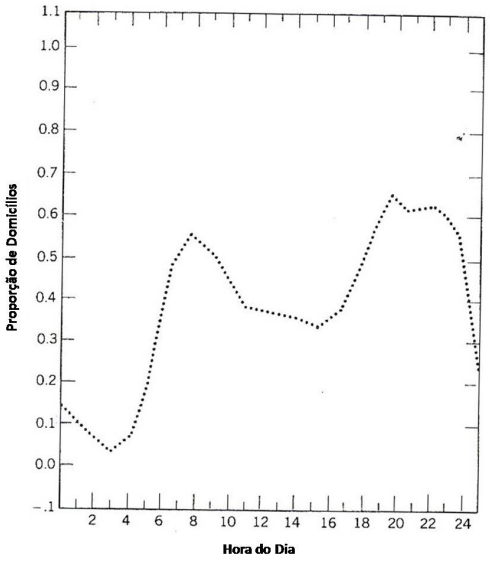

Figura 1.11: Proporção estimada de Respondentes que afirmaram estar em casa e acordado.

Também em Groves [1989], existem evidências empíricas de que, em alguns horários, é mais fácil localizar moradores em casa no sábado do que durante a semana. Na figura 1.12 apresentamos uma tabela resumo, com as proporções de domicílios pesquisados onde pelo menos um morador com mais de 14 anos estava em casa, por hora do dia.

Conhecimentos empíricos como os descritos aqui, ajudam a diminuir os erros de não-resposta, pois essas considerações podem ser utilizadas na logística de coleta de dados. É difícil dizer quanto é possível reduzir esse impacto, porém uma forma de avaliar se uma pesquisa encontrou muitos erros de não-resposta é através da taxa de resposta, que será discutida na Seção 1.3.2.

Taxas de Resposta

Uma das formas mais utilizadas para avaliar a qualidade de uma pesquisa com relação ao erro de não-resposta é a taxa de não-resposta, que de maneira geral, avalia o percentual das pessoas contactadas para participar da pesquisa que efetivamente responderam-na. Usualmente, quanto maior essa taxa menor o impacto da não-resposta nos resultados da pesquisa. Além disso, o fato de uma pesquisa permitir que a taxa de resposta seja calculada já demonstra uma preocupação dos executores da pesquisa com a não-resposta, que já é um indicativo de qualidade da pesquisa.

Existem muitas formas diferentes de se calcular a taxa de resposta, e além disso, dependendo do tipo de pesquisa, algumas taxas fazem mais sentido do que outras. Diferentes indicadores classificando o resultado de cada tentativa de entrevista são utilizados no cálculo dessas taxas. Alguns dos indicadores mais comuns são:

Entrevistas Completadas (C) Número de pessoas que foram entrevistadas.

Entrevistas Parciais (P) Número de pessoas que encerraram a entrevista antes de terminar o questionário.

	Proporção por dia da Semana						
Hora do Dia	Dom.	Seg.	Ter.	Qua.	Qui.	Sex.	Sáb.
8:00–8:59a.m.	–	–	–	–	–	–	–
9:00–9:59a.m.	–	–	–	0.55	0.28	0.45	–
10:00–10:59a.m.	–	0.47	0.42	0.38	0.45	0.40	0.55
11:00–11:59a.m.	0.35	0.41	0.49	0.46	0.43	0.50	0.62
12:00–12:59p.m.	0.42	0.53	0.49	0.56	0.45	0.55	0.60
1:00–1:59p.m.	0.49	0.44	0.50	0.48	0.43	0.51	0.63
2:00–2:59p.m.	0.49	0.50	0.52	0.47	0.45	0.45	0.59
3:00–3:59p.m.	0.54	0.47	0.49	0.54	0.50	0.50	0.65
4:00–4:59p.m.	0.52	0.58	0.55	0.57	0.57	0.56	0.53
5:00–5:59p.m.	0.61	0.67	0.65	0.67	0.59	0.57	0.56
6:00–6:59p.m.	0.75	0.73	0.72	0.68	0.65	0.64	0.59
7:00–7:59p.m.	0.73	0.74	0.75	0.64	0.61	0.57	0.66
8:00–8:59p.m.	–	0.51	0.51	0.59	0.74	0.52	–
9:00–9:59p.m.	–	–	–	0.64	–	–	–

Figura 1.12: Proporção estimada de Domicílios com pelo menos um morador com mais de 14 anos em casa.

Unidades Não-Contactadas porém conhecidas e elegíveis (NC) Número de pessoas que não foram contactadas, porém eram conhecidas e elegíveis.

Unidades Elegíveis que recusaram (R) Número de pessoas elegíveis que foram contactadas porém se recusaram a responder.

Unidades não-elegíveis (NE) Número de pessoas contactadas porém não-elegíveis.

Outras unidades não-entrevistadas (O) Outras pessoas contactadas porém não-entrevistadas.

Podem ser considerados muitos outros indicadores, porém para o propósito dessa seção, os indicadores descritos aqui são suficientes. Uma das taxas de resposta mais utilizada na prática é definida como:

$$\frac{C}{C+P+NC+R+O}, \tag{1.70}$$

onde o denominador inclui todas as pessoas com as quais uma entrevista completa poderia ter sido realizada. Em casos onde estratos ou conglomerados têm probabilidades de seleção desiguais, pode ser de interesse calcular uma taxa de resposta ponderada pelos inversos das probabilidades de seleção. Denotando por h os estratos ou conglomerados considerados e por w_h o peso do estrato ou conglomerado h, podemos calcular a seguinte taxa de resposta ponderada:

$$\sum_h w_h \frac{C_h}{C_h+P_h+NC_h+R_h+O_h}, \tag{1.71}$$

onde indicadores com o índice h representam os indicadores dos respectivos conglomerados ou estratos h.

Lidando com a Não-Resposta da Unidade

A não-resposta pode ser ignorável ou não, conforme discutido com detalhes em Rubin [1976], no contexto de dados faltantes, e em Little [1982], no contexto de modelos de super-população. Como o prórpio nome já diz, se a não-resposta for ignorável, ela pode ser ignorada, ou seja, o único impacto proveniente da mesma é a redução do tamanho da amostra. Porém na prática, raramente sabe-se se a não-resposta é ignorável, assim é sempre recomendável lidar com ela, seja no planejamento e na coleta dos dados, seja na etapa de estimação.

As principais estratégias para lidar com a Não-Resposta da unidade podem ser separadas em 3 categorias:

1-Antes e Durante a Coleta de Dados Durante o planejamento e a coleta de dados são definidas estratégias para reduzir a incidência de Não-Resposta.

2-Técnicas especiais de Estimação Técnicas especias (e provavelmente caras) são utilizadas na coleta de dados e na estimação que permitem estimação não-viciada.

3-Modelos para a Não-Resposta O mecanismo de não-resposta é modelado explicitamente, permitindo que a não resposta seja "corrigida"sob as suposições do modelo.

Os procedimentos utilizados antes e durante a coleta de dados consistem em ter uma estratégia elaborada, ou seja, um plano de ação, o qual será colocado em prática toda vez que ocorrer uma não-resposta da unidade. Usualmente essas estratégias consistem em repetidas tentativas (voltas) do entrevistador fazer contato com a pessoa selecionada que não-respondeu. Existem algumas estratégias comunmente utilizadas na prática, as quais serão discutidas em 1.3.3.

Essas repetidas tentativas são usualmente caras e demoradas, assim outras estratégias podem ser mais interessantes por terem um custo/benefício maior. Uma alternativa, a qual pertence a categoria de técnicas especiais de estimação, é selecionar uma sub-amostra de todas as pessoas que não responderam, e fazer todos os esforços necessários para entrevistar as pessoas pertencentes a essa sub-amostra. Se for realizado corretamente, este procedimento permite obter estimadores não viciados para as quantidades populacionais de interesse, como pode ser visto na página 567 de Särndal et al. [1992]. Apesar de teoricamente interessante, esse tipo de estratégia é usualmente bastante cara (e as vezes impossível), pois é necessário entrevistar toda as pessoas selecionadas na sub-amostra de não-respondentes.

Uma outra alternativa, mais barata e rápida, é modelar a não-resposta. A principal desvantagem dessa técnica é que se o modelo considerado estiver errado, as estimativas populacionais podem ser afetadas. Existem duas grandes classes de modelos para a não-resposta. A primeira, denominada determinística, divide a população em dois estratos, um estrato de respondentes, e outro de não-respondentes. Nessa classe de modelos, unidades populacionais no estrato dos não-respondentes têm probabilidade 0 de responder a pesquisa, e no estratos de respondentes, essa probabilidade é 1.

Essa classe de modelos é bastante limitada, como pode ser visto na página 360 de Cochran [1977], onde o autor afirma:

> A divisão em dois estratos é, claramente, uma simplificação. O acaso tem um papel importante em determinar se uma unidade é encontrada e mensurada em um determinado número de tentativas. Numa especificação mais completa do problema, nós associaríamos a cada unidade populacional **uma probabilidade representando a chance que ela tem de ser mensurada por um determinado método de coleta se for selecionada para pertencer a amostra**.

A classe de modelos que permite que o acaso tenha influência para determinar quais pessoas respondem a uma pesquisa é denominada estocástica. Nessa classe, usualmente modela-se explicitamente a(s) **probabilidade(s) de resposta, que é a probabilidade de uma unidade populacional ser mensurada dado que foi selecionada para pertencer a amostra**.

Tipicamente, na classe de modelos estocásticos, 3 técnicas são utilizadas para lidar com a não-resposta: modelo explícito para a probabilidade de resposta, ponderação e imputação. Em todos os casos, usualmente algum tipo de hipótese simplificadora é feita, permitindo que pessoas sejam agrupadas e um mesmo procedimento seja utilizado para um grupo de pessoas, como no caso do modelo de grupos de respostas homogêneas (GRH), que será discutido em 4.1.1, ou no caso do estimador de Politz-Simmons, apresentado em Politz and Simmons [1949a], onde pergunta-se ao respondente se ele estava em casa nas últimas 5 noites, e a sua resposta é utilizada para determinar o seu peso: um respondente que esteve em casa as 5 noites recebe peso 1, enquanto que um respondente que não esteve em casa nas 5 noites anteriores recebe o peso de 6, pois somente $\frac{1}{6}$ desses respondentes seriam encontrados em casa numa noite aleatória qualquer. Mais detalhes sobre as técnicas da classe de modelos estocásticos utilizados para lidar com a não-resposta podem ser obtidos em Särndal et al. [1992].

É comum que diversas das estratégias discutidas aqui sejam utilizadas em conjunto, como forma de minimizar o impacto do erro de não-resposta de uma pesquisa.

1.3.3 Amostragem Probabilística com Voltas (APV)

A amostragem probabilística, como discutida na Seção 1.2, não pode ser implementada na prática exatamente como a teoria determina, por causa dos erros não-amostrais, mais especificamente por causa do erro de não-resposta.

Algumas alternativas práticas para a **AP** são usualmente utilizadas. Essas alternativas são adaptações da **AP**, onde é definido um critério explícito para se lidar com o erro de não-resposta da unidade durante a etapa de coleta de dados. Três principais alternativas são usualmente consideradas:

Substituições Nesse critério, quando uma pessoa selecionada não é efetivamente entrevistada, a pessoa selecionada iniciamente é substituída por outra similar, segundo algumas covariáveis, usualmente sócio-demográficas.

Voltas (Callbacks) Nesse critério, quando uma pessoa selecionada não é efetivamente entrevistada, são feitas novas tentativas (denominadas voltas) de completar a entrevista com a mesma pessoa. Usualmente um número máximo de κ voltas é definido por pessoa selecionada. Quando esse número é atingido, dois procedimentos são adotados, ou a pessoa é substituída, conforme o critério anterior, ou então é realizada uma nova seleção probablística.

Over-Sampling Nesse critério, se existe uma estimativa da taxa de resposta p, inflaciona-se o tamanho da amostra desejada de n para $\frac{n}{p}$, de forma que o número esperado de entrevistas realizadas seja n. Assim, quando uma pessoa selecionada não é efetivamente entrevistada, não são feitas novas tentativas de completar a entrevista e nem a pessoa selecionada iniciamente é substituída por outra similar.

Os critérios de **Substituição** e de **Over-Sampling** supõe que a probabilidade de resposta das unidades populacionais que foram inicialmente selecionadas porém não entrevistadas são as mesmas das pessoas que não foram selecionadas. Já o critério de **Voltas**, apesar de ser mais demorado pois requer que o entrevistador volte ao mesmo local mais de uma vez, requer uma suposição sobre as probabilidades de resposta somente para as pessoas que não forem contactadas em κ voltas.

Ao longo dessa tese, quando o critério com Voltas for utilizado, esse desenho amostral será denominado de Amostragem Probabilística com Voltas (**APV**). As propriedades teóricas desse desenho amostral, sob a suposição do modelo de não-resposta GRH, serão discutidas no Capítulo 4. Na Seção 2.5, serão apresentados mais detalhes sobre a implementação da **APV** na prática.

1.4 Outros Tipos de Inferência

Usualmente, quando se fala de amostragem de populações finitas, se pensa em inferência baseada no desenho **(ID)**, como foi discutido na Seção 1.2, porém existem 2 outros tipos de inferência que também são bastante utilizadas, cada uma com prós e contras que serão discutidos a seguir. Para facilitar a compreensão das diferenças entre esses tipos de inferência, alteraremos a notação utilizada na Seção 1.2. Nessa seção, denotaremos a valor observado da variável de interesse Y para a i-ésima unidade populacional como y_i o qual será interpretado como a realização da variável aleatória Y_i.

Na **ID**, assumimos que o valor da variável de interesse Y para a cada unidade populacional é fixo, ou seja, não existe incerteza associada esse valor. Outra forma de expressar essa situação é supondo que a distribuição de Y_i é degenerada, ou seja, $P(Y_i = y_i) = 1$. Por isso que na Seção 1.2 omitimos a notação y_i e utilizamos apenas Y_i.

Para se fazer inferência, é necessário recorrer a uma distribuição de referência. Essa distribuição é necessária para que as inferências feitas a partir da amostra possam ser mensuradas com uma régua probabilística. Essa distribuição pode ser interpretada de diferentes formas dependendo do tipo de inferência considerada.

No contexo de **ID**, a única distribuição de probabilidade relevante para realizar inferência decorre da seleção da amostra s do universo $\{1, 2, ..., N\}$ com probabilidade $p(s)$. **Nesse caso, a distribuição de referência é a distribuição amostral do estimador de interesse, ou seja, o comportamento do estimador ao longo de todas as possível amostras.** Por isso

é necessário, no contexto de **ID**, que a seleção da amostra seja probabilística, caso contrário, não existe uma distribuição de referência, e consequentemente, não há como mensurar a inferência com uma régua probabilística.

Além disso, na **ID**, o interesse está em estimar quantidades descritivas populacionais da forma $g(y_1, ..., y_N)$, também conhecidas como parâmetros populacionais. Usualmente, as quantidades populacionais de interesse são totais populacionais $g(y_1, ..., y_N) = \sum_{i=1}^{N} y_i$ e médias populacionais $g(y_1, ..., y_N) = \frac{1}{N}\sum_{i=1}^{N} y_i$. Quando o objetivo está em fazer inferência de quantidades populacionais descritivas, ela é denominada inferência descritiva.

Discutiremos nessa seção dois outros tipos de inferência: Inferência baseada no Modelo (**IM**) e Inferência Bayesiana baseada no Modelo (**IBM**). Esses dois tipos de inferência são baseados em um modelo. Nesse modelo, denominado de modelo superpopulação, os valores $\mathbf{y} = (y_1, ..., y_N)$ da variável de interesse Y são observações ou realizações do vetor de variáveis aleatórias $(Y_1, ..., Y_N)$, o qual tem distribuição probabilística com densidade $f(\mathbf{y}/\mathbf{x}, \theta)$, $\theta \in \Theta$, onde Θ é denominado espaço paramétrico e $\mathbf{x}$ é o vetor de covariáveis. Ou seja, o modelo superpopulação é indexado pelo parâmetro θ, o qual é desconhecido, e uma particular população observada é uma de muitas populações potenciais que poderiam ter sido geradas por este modelo.

Nesse contexto, usualmente o interesse está em utilizar uma amostra s, obtida com probabilidade $p(s)$, para fazer inferência sobre o parâmetro θ, e não sobre as quantidades descritivas populacionais. Um exemplo, descrito em Graubard and Korn [2002], é o caso onde as médias de dois domínios (estratos) são comparadas, e o interesse é saber se os parâmetros superpopulacionais são iguais, pois raramente existe o interesse em saber se as médias populacionais em si são iguais, pois elas raramente serão. Quando o objetivo está em fazer inferência sobre os parâmetros do modelo superpopulação, e não em quantidades descritivas populacionais, ela é denominada inferência analítica.

Note que em qualquer tipo de inferência baseada em modelos (**IM** e **IBM**), sempre existe a suposição de que o modelo está correto, além disso diferentes analistas podem postular diferentes modelos. Um problema evidente com esse tipo de inferência ocorre quando o modelo não está correto. Em Hansen et al. [1983], os autores discutem o impacto que a má-especificação do modelo de superpopulação pode ter nas estimativas. O fato da **ID** não depender da suposição de um modelo paramétrico é um dos principais motivos pelo qual esse tipo de inferência é mais popular, e por causa disso as vezes é denominada de inferência não-paramétrica.

Do ponto de vista de inferência baseada no desenho (**ID**), também é comum utilizar modelos, porém com o objetivo de melhorar os estimadores, como é discutido em Särndal et al. [1992], e/ou para considerar a não resposta durante o processo de inferência, como foi discutido na Seção 1.3, porém as inferências sempre são realizadas com relação a distribuição gerada pelo desenho amostral. Nesse contexto, quando modelos são utilizados, diz-se que a **inferência é assistida por modelos**.

Quando a densidade $f(\mathbf{y}/\mathbf{x}, \theta)$ é vista como uma função de θ, ela é usualmente denominada de função de verossimilhança, e é denotada por $L(\theta/\mathbf{y})$. Sob esse ponto de vista, a função de verossimilhança num ponto específico θ' pode ser interpretada como a probabilidade, sob o modelo superpopulação, das unidades populacionais pertencentes a amostra terem sido geradas pelo modelo

superpopulação se o parâmetro populacional for igual a θ'.

Se o desenho amostral for **AAS** e se o modelo de superpopulação for independente e identicamente distribuido (iid) com distribuiçao $f(y/\theta)$ para cada Y_i, então pode-se utilizar técnicas usuais de inferência, desconsiderando o desenho amostral, como por exemplo o estimador de máxima-verossimilhança.

1.4.1 Inferência baseada no Modelo (IM)

Esse tipo de inferência é bastante utilizada no contexto de pequenos domínios, quando o tamanho amostral é muito pequeno nas áreas onde se deseja fazer inferência, e consequentemente a suposição de tamanho amostral grande o suficiente da **ID** passar a ser questionável. Uma ótima referência para **IM** no contexto de pequenos domínios é Moura [2008].

No caso da **IM**, a distribuição de referência utilizada para fazer inferência é obtida, conceitualmente, supondo-se que infinitas populações diferentes serão geradas do modelo superpopulação com a amostra observada mantida fixa, ou seja, somente utiliza-se o modelo superpopulação para gerar os valores Y_i das unidades populacionais não pertencentes a amostra.

Suponha que o interesse está em estimar uma combinação linear dos valores populacionais $\mathbf{Y} = (Y_1, ..., Y_N)'$, definida por $h = c'\mathbf{Y}$, onde $c = (c_1, ..., c_N)$ é um vetor de coeficientes conhecidos. Se $c = (1, ..., 1)$, então a quantidade de interesse é o total populacional e se $c = (1/N, ..., 1/N)$, a quantidade de interesse é a média populacional. Note que o vetor populacional $(Y_1, ..., Y_N)$ é uma quantidade aleatória, e consequentemente, a função linear h também é. O estimador de uma quantidade de interesse que é aleatória é denominado preditor.

Nesse contexto, o preditor linear $\hat{h}$ de h é dado por $\hat{h} = c_s'\mathbf{Y_s}$, onde $\mathbf{Y_s} = (Y_1, ..., Y_n)$ é o vetor aleatório associado aos valores de Y observados na amostra e c_s é um vetor de coeficientes conhecidos.

Para exemplificar como o procedimento de inferência é feito do ponto de vista de **IM**, apresentaremos um exemplo de Moura [2008]. Suponha que não haja nenhuma estrutura relevante na população de interesse, e que o analista acredite que a variável de interesse Y têm o mesmo comportamento para todas as unidades da população, então essa população pode ser representada pelo seguinte modelo de superpopulação, denominado de ξ:

$$\begin{aligned} E_\xi(Y_i) &= \mu; \quad \forall i = 1, ..., N \\ Cov_\xi(Y_i) &= Var_\xi(Y_i) = \sigma^2; \quad \forall i = j = 1, ..., N \\ Cov_\xi(Y_i) &= 0; \quad \forall i \neq j = 1, ..., N \end{aligned} \tag{1.72}$$

Estamos interessados em estimar o total populacional $T = \sum_{i=1}^{N} y_i$, utilizando os dados $y_s = (y_1, ..., y_n)$ obtidos na amostra. Vamos utilizar o preditor $\hat{T} = c_s'\mathbf{Y_s}$, com $c_s = (\frac{N}{n}, ...\frac{N}{n})$, ou seja, temos que $\hat{T} = N\overline{y}_s$, onde $\overline{y}_s = \frac{1}{n}\sum_{i\in s} y_i$. Note que o preditor $\hat{T}$, sob o modelo ξ, pode ser obtido da seguinte forma:

$$\begin{aligned}\hat{T} &= \sum_{i\in s} y_i + \sum_{i\notin s} \hat{y}_i = \sum_{i\in s} y_i + \sum_{i\notin s} \overline{y}_s \\ &= n\overline{y}_s + (N-n)\overline{y}_s = N\overline{y}_s.\end{aligned}$$

Nesse contexto, para avaliar a eficiência do preditor, o interesse está em calcular a esperança e a variância do erro de predição $(\hat{T} - T)$, ou seja, apenas da parcela $(N-n)(\overline{y}_s)$ do preditor que foi utilizada para prever $\sum_{i\notin s} y_i$, referente as unidades populacionais que não foram observadas na amostra. Temos então, supondo que o modelo ξ está correto:

$$\begin{aligned}E_\xi(\hat{T} - T) &= (N-n)E_\xi(\overline{y}_s) - (N-n)E_\xi(Y) = (N-n)\mu - (N-n)\mu = 0, \\ V_\xi(\hat{T} - T) &= (N-n)^2 V_\xi(\overline{y}_s) + (N-n)Var_\xi(Y) = N^2\frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right).\end{aligned}$$

Note que a esperança e a variância do erro de predição, para o modelo ξ, são muito parecidas com a esperança e a variância do estimador do total populacional do desenho amostral **AASs**, com a diferença de que aparece σ^2 no lugar de S^2. Para obter esses resultados, não consideramos como os dados da amostra foram obtidos, pois a **distribuição de referência utilizada é baseada no modelo superpopulação utilizado**. Nesse exemplo, o modelo considerado é similar ao desenho amostral AAS, por isso a semelhança dos resultados.

Esse fato pode levar a conclusão errônea de que o desenho amostral utilizado é irrelevante, desde que algum modelo superpopulação seja assumido. As duas fontes de variação devem ser consideradas para realizar inferência: a primeira decorrente da seleção da amostra e a segunda referente ao modelo superpopulação considerado. Se o desenho amostral utilizado para selecionar a amostra, probabilístico ou não, for ignorável como definido em 1.3 (veja mais detalhes em Sugden and Smith [1984]), ou seja, se as probabilidades das unidades populacionais pertencerem a amostra não dependem da quantidade populacional de interesse Y, pode-se fazer inferência sem considerar o desenho amostral utilizado.

É ímportante explicitar que dependendo da variável de interesse Y considerada, um mesmo desenho amostral pode ser ignorável ou não. Por exemplo, imagine que selecionou-se para pertencer a uma amostra todas as pessoas que nasceram em um determinado dia do ano, na cidade de São Paulo, e o dia do ano foi selecionado com probabilidades uniformes. Se o objetivo da pesquisa for estimar a idade média das pessoas da cidade de São Paulo ou a proporção de torcedores corinthianos na cidade, essa amostra pode ser considerada ignorável, ou não-informativa, pois não existe relação entre as probabilidades das pessoas serem selecionadas para participar da amostra com a variável de estudo. Porém, se o objetivo da pesquisa for estimar a proporção de pessoas que nascem no mês de novembro em São Paulo, essa amostra não é ignorável (é informativa), pois nesse caso existe uma relação clara entre as probabilidades das pessoas pertencerem a amostra com a variável de interesse. Se um dia do mês de novembro for selecionado, essa proporção será de 100% e se for

selecionado um dia de qualquer outro mês, essa proporção será de 0%.

Definição 1.3 (Amostragem Ignorável) *Seja* $I = (I_1(s), ..., I_N(s))'$ *o vetor indicador de pertinência de cada unidade populacional na amostra, onde* $I_j(s) = 1$ *se a unidade populacional* j *pertence a amostra* s *e* $I_j(s) = 0$ *caso contrário, e* $f_I(I|\mathbf{y}, \mathbf{x}, \phi)$ *a função de probabilidade conjunta de* I *condicionada a* $\mathbf{x}$*,* $\mathbf{y}$ *e* ϕ*, onde* $\mathbf{x}$ *é um vetor de covariáveis,* $\mathbf{y}$ *é um vetor das variáveis de interesse e* ϕ *os parâmetros utilizados no desenho amostral. O desenho amostral definido por* I *é dito ignorável se:*

$$f_I(I|\mathbf{y}, \mathbf{x}, \phi) = f_I(I|\mathbf{x}, \phi). \tag{1.73}$$

Por exemplo, para o caso da AAS, temos que

$$f_I(I|\mathbf{y}, \mathbf{x}, \phi) = f_I(I) = \begin{cases} \binom{N}{n}^{-1} & \text{se } \sum_{i=1}^{N} I_j(s) = n \\ 0 & \text{caso contrário} \end{cases}, \tag{1.74}$$

ou seja, o vetor de indicador de pertinência amostral não dependende nem de $\mathbf{y}$ e nem de $\mathbf{x}$, assim no caso da AAS, o desenho amostral é ignorável. Usualmente, quando a amostragem é probabilística, o desenho amostral é ignorável, pois não depende de $\mathbf{y}$, apenas das covariáveis $\mathbf{x}$ e do parâmetro ϕ.

Mas no caso geral, quando o desenho amostral não é ignorável, o modelo utilizado para fazer inferência deve considerar o desenho amostral explicitamente, pois a distribuição dos valores da amostra $f_s(\mathbf{y}/\mathbf{x}, \theta)$ é diferente da distribuição populacional $f(\mathbf{y}/\mathbf{x}, \theta)$. Ou seja, nesse caso, os dados não observados (voluntaria ou involuntariamente) $\mathbf{y}_{\bar{\mathbf{s}}}$ também fornecem informações relevantes ao modelo proposto.

No caso de desenho informativos, então a verossimilhança completa dos dados deve ser especidicada como $f(\mathbf{y}, I/\mathbf{x}, \theta, \phi)$. Em Ravines [2003] e no Capítulo 7 de Gelman et al. [2003] descreve-se com muito mais detalhes como modelar os dados utilizando a verossimilhança completa. Nessa seção, faremos apenas um breve esboço. Utilizando a propriedade $P(A \cap B) = P(A/B)P(B)$ da probabilidade condicional, a verossimilhança completa pode ser particionada da seguinte forma:

$$f(\mathbf{y}, I/\mathbf{x}, \theta, \phi) = f(\mathbf{y}/\mathbf{x}, \theta) f_I(I|\mathbf{y}, \mathbf{x}, \phi), \tag{1.75}$$

ou seja, depende tanto do modelo superpopulação quanto do desenho amostral, representado pelo vetor I. Embora a esquação em 1.75 seja útil para desenvolver o modelo para se fazer inferência, essa não é a verossimilhança dos dados a não ser que $\mathbf{y}$ tenha sido completamente observada, ou seja, se um censo foi realizado. Usualmente, não se conhece o vetor $\mathbf{y}$ completo, pois ele depende tanto dos valores observados $\mathbf{y}_{\mathbf{s}}$ quanto dos valores não observados $\mathbf{y}_{\bar{\mathbf{s}}}$. Assim, a verossimilhança

dos dados observados é dada por:

$$f(\mathbf{y_s}, I/\mathbf{x}, \theta, \phi) = \int f(\mathbf{y_s}, \mathbf{y_{\bar{s}}}/\mathbf{x}, \theta) f_I(I|\mathbf{y_s}, \mathbf{y_{\bar{s}}}, \mathbf{x}, \phi) d\mathbf{y_{\bar{s}}}. \tag{1.76}$$

Com relação ao desenho ser ignorável, um detalhe bastante importante é discutido em Sugden [1985]. É comum que o estatístico que desenhou a amostra (amostrista) não seja o mesmo que irá analisar os resultados (analista). Nesse cenário, para o amostrista que conhece as covariáveis $\mathbf{x} = (\mathbf{x_s}, \mathbf{x_{\bar{s}}})$ para toda a população em questão, um desenho amostral onde $f_I(I|\mathbf{y}, \mathbf{x}, \phi) = f_I(I|\mathbf{x}, \phi)$ é ignorável, porém o mesmo desenho amostral só é ignorável para o analista se ele conhecer as covariáveis $\mathbf{x}$ para toda a população. Por exemplo, no caso de uma amostra estratificada com estratos não proporcionais, é necessário conhecer $\mathbf{x} = (\mathbf{x_s}, \mathbf{x_{\bar{s}}})$ para que o desenho amostral seja ignorável.

1.4.2 Inferência Bayesiana baseada no Modelo (IBM)

A principal diferença da **IBM** para a **IM**, matematicamente, é que ela utiliza uma distribuição $\pi(\theta)$, denominada distribuição a priori, para todos os parâmetros θ desconhecidos do modelo superpopulação conisderado. Conceitualmente, a distribuição $\pi(\theta)$ representa a opinião do estatístico que está analisando os dados sobre o parâmetro desconhecido. Ou seja, cada estatístico que for fazer inferência pode ter sua própria opinião sobre o parâmetro θ, a qual será representada matematicamente na distribuição $\pi(\theta)$. Essa é a principal crítica que esse tipo de inferência recebe, pois além do modelo de superpopulação, ela tem outra fonte de subjetividade: a opinião de quem está fazendo inferência. Por outro lado, essa é uma das principais vantagens, pois permite que os estatístico combine as informações obtidas da amostra com o seu conhecimento a priori para fazer inferência. Além disso, a interpretação de inferências do tipo **IBM** são mais intuitivas do que as do tipo **ID**. Uma ótima referência sobre probabilidades subjetivas é o livro de Finetti [1974].

No caso da **IBM**, **a distribuição de referência utilizada para fazer inferência é denominada distribuição à posteriori**, usualmente denotada por $\pi(\theta/\mathbf{y})$. Essa distribuição é obtida utilizando o Teorema de Bayes. Matematicamente, sob o modelo de superpopulação $f(\mathbf{y}/, \theta)$, essa distribuição é dada por:

$$\pi(\theta/\mathbf{y}) = \frac{f(\mathbf{y}/\theta)\pi(\theta)}{\int f(\mathbf{y}/\theta)\pi(\theta)d\theta} \propto f(\mathbf{y}/\theta)\pi(\theta). \tag{1.77}$$

A forma mais fácil de explicar esse tipo de inferência é através de um exemplo. Vamos supor que uma amostra de tamanho n foi observada, e y indica o número de pessoas que afirmaram que votarão no atual prefeito em uma determinada cidade. O estatístico decide utilizar o modelo superpopulação binomial $Bin(n, \theta)$, o qual depende do parâmetro θ, que pode ser interpretado como a proporção real de pessoas que votarão no atual prefeito. Para representar a sua opinião a priori sobre essa proporção, o estatístico utiliza a distribuição beta, dada por $\mathcal{B}(a, b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$ com

$0 < \theta < 1$, onde $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta$ é a função beta de 2 parâmetros. Os parâmetros a e b são denominados hiper-parâmetros, e são utilizados pelo estatístico para especificar a sua opinião a priori sobre o parâmetro de interesse.

Nesse exemplo, deixaremos a e b genéricos, porém eles podem ser especificados da seguinte forma. Sabe-se que se $X \sim \mathcal{B}(a,b)$, então $E(X) = \frac{a}{a+b}$ e $V(X) = \frac{ab}{(a+b+1)(a+b)^2}$. O estatístico deve utilizar valores de a e b que reproduzam aquela que ele acredita ser a proporção real da população θ, por exemplo, se ele acredita que ela está próxima de $\frac{1}{2}$, ele deve fazer $a = b$. Do ponto de vista de inferência bayesiana, a variância da distribuição a priori pode ser interpretada como o inverso da certeza que o estatístico tem com relação a esse parâmetro, ou seja, se existe muita certeza de que $\theta = \frac{1}{2}$, a variância da priori deve ser pequena, nesse caso representada por valores grandes de a e b, porém se o estatístico não tem muita certeza, a variância da priori deve ser grande, e nesse caso ele deve utilizar valores pequenos para a e b.

Assim, substituindo essas distribuições em 1.77, obtemos que:

$$\pi(\theta/\mathbf{y}) \propto \theta^{a-1+y}(1-\theta)^{b-1+(n-y)}, \tag{1.78}$$

ou seja, obtemos que $\theta/y \sim \mathcal{B}(y+a, n-y+b)$, assim , nesse exemplo, a distribuição $\mathcal{B}(y+a, n-y+b)$ resume todo o conhecimento do estatístico com relação ao parâmetro θ, sobre o qual ele deseja fazer inferência. Dessa forma, obtemos que $E(\theta/y) = \frac{y+a}{n+a+b}$ e sua variância é dada por $Var(\theta/y) = \frac{(y+a)(n-y+b)}{(n+a+b+a)(n+a+b)^2}$. O livro Gelman et al. [2003] é uma ótima referência para mais detalhes sobre como utilizar $\pi(\theta/\mathbf{y})$ para fazer inferência.

Note que no exemplo apresentado, a informação contida na amostra a respeito de θ só foi incluída no processo de estimação através da função de verossimilhança, ou seja, o processo de estimação é realizado condicionalmente a amostra obtida, não importando todas as outras possíveis amostras que poderiam ter sido observadas. Essa característica faz parte do que é conhecido como princípio da verossimilhança:

> **Princípio da Verossimilhança:** Toda a informação sobre θ existente em uma amostra (ou experimento) está contida na função de verossimilhança de θ dado $\mathbf{y}$. Duas funções de verossimilhança para θ (do mesmo ou de diferentes experimentos) contém a mesma informação para θ se são proporcionais.

Esse princípio é considerado fundamental por muitos estatísticos, não só do ponto de vista da **IBM**, porém é nesse contexto onde ele tem mais força. Mais detalhes sobre o princípio da verossimilhança e a sua importância para a inferência Bayesiana podem ser obtidos em Berger and Wopert [1988].

Muitas vezes, o princípio da verossimilhança é evocado para afimar que a forma como a amostra foi coletada (desenho amostral, não resposta, seleção intencional, etc...) não é relevante para se fazer inferência, apenas importam os dados observados. O problema com esse argumento, segundo os autores em Gelman et al. [2003], é que a definição de *dados observados* deve incluir informação

sobre a forma como esses dados foram selecionados, ou seja, para fazer inferência deve ser utilizada a verossimilhaça dos dados observados, apresentada em 1.76.

Um exemplo simples é apresentado pelos autores para desmistificar a noção de que o método de coleta de dados é irrelevante do ponto de vista Bayesiano. Imagine que um analista recebe uma amostra de um pesquisador, nessa amostra contém os resultados de 10 arremessos de dados e todos esses resultados são 6. Com certeza a atitude do analista com relação a natureza do dado depois de analisar esses resultados seria diferente se ele fosse avisado pelo pesquisador de que **1)** esses foram os únicos arremessos realizados, versus **2)** foram realizados 60 arremessos porém só foram incluídos nas amostra os arremessos iguais a 6, versus **3)** foi decidido a priori que seriam incluídos na amostra 10 arremessos com o resultado 6, porém que não seria registrado quantos arremessos seriam necessários para obter esses resultados, e foram necessários 500 arremessos para obter a amostra. Em exemplos simples como esse, é fácil de ver que a distribuição dos dados observados têm uma distribuição diferente daquela para os dados completos.

A utilização da **IBM** no contexto de amostragem de populações finitas, começou a ser pesquisada no início da década de 60. Um dos principais artigos nesse contexto é Ericson [1969], no qual o autor considera modelos paramétricos conjugados na família exponencial. Assumindo apenas a condiçao de linearidade à posteriori e permutabilidade, o autor obtém os estimadores de Bayes de quantidades de interesse na população, sob perda quadrática. No contexto de **IBM**, supor permutabilidade das unidades populacionais equivale a supor que os índices que identificam as unidades populacionais são não-informativos, ou seja, a informação obtida dos y_i independe das unidades efetivamente observadas. Esse conceito é explicado com mais detalhes em Bernardo [1996] e Cordani and Wechsler [2006].

A partir da verossimilhança dos dados observados em 1.76, a distribuição à posteriori conjunta dos parâmetros θ e ϕ é dada por:

$$\begin{aligned} \pi(\theta,\phi/\mathbf{x},\mathbf{y_s},I) &= \pi(\theta,\phi/\mathbf{x})\int f(\mathbf{y_s},\mathbf{y_{\bar{s}}}/\mathbf{x},\theta)f_I(I|\mathbf{y_s},\mathbf{y_{\bar{s}}},\mathbf{x},\phi)d\mathbf{y_{\bar{s}}} \\ &\propto \pi(\theta,\phi/\mathbf{x})f(\mathbf{y_s},I/\mathbf{x},\theta,\phi), \end{aligned} \tag{1.79}$$

onde $\pi(\theta,\phi/\mathbf{x})$ é a distribuição a priori dos parâmetros θ e ϕ condicionada a $\mathbf{x}$. Usualmente, não há interesse analítico no parâmetro ϕ, assim a distribuição à posteriori somente de θ é de maior interesse, a qual é dada por:

$$\pi(\theta/\mathbf{x},\mathbf{y_s},I) = \pi(\theta/\mathbf{x})\int\int \pi(\phi/\mathbf{x},\theta)f(\mathbf{y_s},\mathbf{y_{\bar{s}}}/\mathbf{x},\theta)f_I(I|\mathbf{y_s},\mathbf{y_{\bar{s}}},\mathbf{x},\phi)d\mathbf{y_{\bar{s}}}d\phi, \tag{1.80}$$

onde $\pi(\phi/\mathbf{x},\theta)$ é a distribuição a priori do parâmetro ϕ condicionada a $\mathbf{x}$ e a θ.

No contexto de **IBM**, um desenho amostral é dito ignorável se $f_I(I|\mathbf{y_s},\mathbf{y_{\bar{s}}},\mathbf{x},\phi) = f_I(I|\mathbf{x},\phi)$, e nesse caso a distribuição à posteriori de θ passa a ser:

$$\begin{aligned}\pi(\theta/\mathbf{x},\mathbf{y_s}) &= \pi(\theta/\mathbf{x})\int\int \pi(\phi/\mathbf{x})f(\mathbf{y_s},\mathbf{y_{\bar{s}}}/\mathbf{x},\theta)f_I(I|\mathbf{x},\phi)d\mathbf{y_{\bar{s}}}d\phi \\ &\propto \pi(\theta/\mathbf{x})\int f(\mathbf{y_s},\mathbf{y_{\bar{s}}}/\mathbf{x},\theta)d\mathbf{y_{\bar{s}}}. \end{aligned} \quad (1.81)$$

Duas condições são necessárias e suficientes para que o desenho amostral seja ignorável:

Faltantes ao Acaso (Missing at Random) Dado o parâmetro ϕ, a distribuição de I depende somente de $\mathbf{x}$ e de $\mathbf{y_s}$, de forma que $f_I(I|\mathbf{y_s},\mathbf{y_{\bar{s}}},\mathbf{x},\phi) = f_I(I|\mathbf{y_s},\mathbf{x},\phi)$.

Parâmetros Distintos O parâmetro da coleta de dados ϕ, dado $\mathbf{x}$, é independente de θ, ou seja, $\pi(\phi/\mathbf{x},\theta) = \pi(\phi/\mathbf{x})$.

1.4.3 Amostragem e Aleatorização

Do ponto de vista da **ID**, a necessidade de que a amostragem seja probabilística, ou seja, que a seleção das pessoas seja feita utilizando alguma forma de aleatorização é evidente, pois é dessa aleatorização que a distribuição de referência utilizada para fazer inferência é gerada. Porém como foi apontado nas seções 1.4.1 e 1.4.2, do ponto de vista da **IM** e da **IBM**, a aleatorização não têm um papel explícito no processo de inferência.

Nessa seção, o objetivo é discutir qual o papel da aleatorização no contexto dessas outras formas de se fazer inferência. Mesmo sem ter um papel explícito, a amostra deve ser obtida a partir de um método de aleatorização? E depois de selecionada a amostra, as probabilidades de seleção devem ser consideradas para se fazer inferência, como no caso dos estimadores de HH e de HT? Para discutir a sua utilização, é conveniente distinguir as seguites etapas do processo de inferência:

Pré-Experimental A aleatorização é utilizada para selecionar as unidades populacionais que serão incluídas na amostra ou para determinar qual tratamento será associado a cada unidade amostral.

Pós-Experimental Após selecionada a amostra, as probabilidades de seleçao das unidades populacionais incluídas na amostra são utilizadas no processo de inferência.

Não existem respostas concretas para essas perguntas. No geral, parece haver um concenso de que a amostra deve ser obtida através de um mecanismo probabilístico, de forma que todas as unidades populacionais tenham alguma chance de serem selecionadas. Do ponto de vista da **IM**, como foi vimos na Seção 1.4.1, essa forma de seleção garante que o desenho amostral seja ignorável, se as covariáveis $\mathbf{x}$ forem conhecidas pelo analista. Do ponto de vista da **IBM**, uma distinção importante deve ser feita, entre *experimentos para aprender* e *experimentos para provar*, como discutido em Kadane and Seidenfeld [1990]. No primeiro caso, quando o objetivo do pesquisador é apenas aprender sobre Y, apenas um único tomador de decisões está envolvido, e nesse caso os autores argumentam que não existe a necessidade da aleatorização. Já no segundo caso, quando o

objetivo é provar um resultado sobre Y, mais de um tomador de decisões utilizarão os dados da amostra para fazer inferência, possivelmente até alguém que utilizará **ID**, e nesse caso os autores acreditam que se justifique o uso da aleatorização.

Uma outra justificativa para a aleatorização pré-experimental é que dessa forma o pesquisador se isenta da responsabilidade de selecionar uma particular amostra, pois a seleção foi aleatória. Sob esse ponto de vista, no caso particular de experimentos médicos, pode ser considerado anti-ético um médico receitar o tratamento B porque foi o resultado da aleatorização, quando ele acredita que o paciente pode se beneficiar mais do tratamento A, conforme é discutido em Wajsbrot [1997].

Alguns argumentos a favor de amostras não-probabilísticas são descritos em Jessen [1978]. São eles:

Critério da Correlação Se uma amostra pode ser selecionada de forma que a média amostral seja equivalente a média populacional com relação a covariáveis X conhecidas para toda a população, então essa amostra deve ter a média da variável de interesse Y próxima com o valor populacional, se X e Y forem correlacionadas.

Critério da Persistência Se os resultados das eleições em uma particular cidade sempre coincidem com os resultados da eleição presidencial, por exemplo, então espera-se que essa cidade continue reproduzindo os resultados gerais (persistência dos resultados), e nesse contexto, pode ser mais interessante (custo/benefício) selecionar uma amostra nessa cidade do que no país todo.

O critério da persistência para selecionar amostras já foi utilizado tanto em eleições brasileiras, como em Mendonça and Migon [1987], quanto em eleições estrangeiras, como em Bernardo [1984], em ambos os casos com sucesso. Ou seja, são exemplos de amostras determinísticas que obtiveram comprovadamente bons resultados.

Com relação a etapa pós-experimental, existe o concenso de que informações relacionadas ao desenho amostral devem ser consideradas na análise, porém há divergências sobre como fazer isso. Quando os desenhos amostrais não são ignoráveis, alguns autores consideram que as probabilidades de seleção devem ser utilizadas explicitamente nas análises, enquanto outros consideram que basta incluir as covariáveis utilizadas para seleção da amostra no modelo utilizado para se fazer inferência. Uma visão conciliadora é expressada em Rubin [1985], onde o autor considera que as probabilidades de seleção podem ser consideradas como uma forma de resumir as informações das covariáveis X utilizadas no desenho amostral, as quais são mais facilmente incluídas num modelo do que as covariáveis originais, porém o autor enfatiza que isso não quer dizer que no modelo utilizado as covariáveis não podem/devem ser explicitamente modeladas também.

Capítulo 2

Amostragem por Cotas (AC)

A Amostragem por Cotas foi utilizada pela primeira vez, com sucesso, em 1932 na eleição presidencial americana. Esse desenho amostral foi desenvolvido, na época, por George Gallup, Archibald Crossley e Elmo Roper. Esse sucesso decorreu do fato de que as amostras com cotas acertaram o nome do vencedor dessa eleição. Em contraste com essa pesquisa, a pesquisa do Literary Digest que era referência na época, pois havia acertado o vencedor da eleição presidencial anterior[1], errou o nome do vencedor. Mais detalhes dessa história podem ser obtidos em Ferraz [1996], onde o autor resume de forma interessante a performance das primeiras pesquisas eleitorais americanas.

A amostragem por cotas é um termo genérico, no geral relacionado a forma como as unidades populacionais são selecionadas para pertencer a amostra. Esse tipo de amostragem quase que exclusivamente é aplicado a populações humanas. Superficialmente, podemos imaginar a amostragem por cotas como sendo uma amostra estratificada, ou seja, se divide a população em H grupos, ou cotas, e esses grupos são controlados durante a seleção da amostra. Para criar essas cotas ou grupos, utiliza-se uma ou mais covariáveis conhecidas para toda a população. As cotas são formadas combinando as categorias das covariáveis. Por exemplo, supondo que as covariáveis "sexo"e "idade"são utilizadas, uma possível categoria da cota é "Homens de até 30 anos de idade". Na Seção 2.2 serão discutidos com mais detalhes como as cotas são determinadas e os diferentes tipos de cota.

Usualmente as covariáveis, que também são conhecidas como variáveis de cota, são obtidas do censo populacional, ou de grandes amostras governamentais, como por exemplo a Pesquisa Nacional por Amostragem de Domicílios (PNAD), ou os Micro-dados do Censo, ambas realizadas pelo IBGE. É importante que as cotas definidas para um determinada pesquisa não sejam ambíguas, para garantir que o entrevistador classifique corretamente as pessoas entrevistadas.

Existem muitas formas diferentes de selecionar as unidades populacionais dentro de cada cota, algumas mais rigorosas, como a amostragem probabilística com cotas onde o entrevistador é obrigado a tentar fazer contato com todos os moradores, e outras mais flexiveis, como as pesquisas de ponto de fluxo, deixando a seleção do respondente a critério do entrevistador. Os diferentes tipos de amostragem por cotas serão discutidos na Seção 2.3. De maneira geral, o procedimento é simples, definidas as cotas, ou seja, o tamanho das amostras de cada grupo, o entrevistador deve

[1]Nessa pesquisa, a participação era voluntária e o tamanho de amostra que chegava a 2 milhões de pessoas.

entrevistar essa quantidade de pessoas em cada cota.

A principal diferença com relação a amostragem estratificada, é que dentro de cada cota não é realizada uma amostra probabilística, ao invês disso o entrevistador seleciona qualquer pessoa que ele encontrar que se encaixe nas cotas, e consequentemente as probabilidade de cada unidade populacional pertencer a amostra não são conhecidas. Ou seja, do ponto de vista de **ID**, não é possível obter os estimadores do parâmetro populacional e da variância desse estimador à partir desse tipo de amostragem.

A principal justificativa para utilizar a amostragem por cotas é que elas são consideradas mais rápidas e baratas do que a amostragem probabilística. Em Sudman [1967], o autor compara os custos médios por entrevista de 6 amostras probabilísticas, o qual foi de US$14 com o de 4 amostras por cotas, o qual foi de US$7.50. Ou seja, a custo por entrevista chega a reduzir praticamente 50%. Claro que existem diversos fatores que podem estar influenciando nessa redução, como tempo de aplicação do questionário e questões logísticas, além de características mais específicas de cada desenho amostral, como por exemplo, a quantidade de cotas de uma pesquisa, pois quanto mais cotas houverem, mais difícil será para o entrevistador concluir as entrevistas, ou seja, ele terá que trabalhar por mais tempo para conseguir completar as entrevistas, como é destacado pelo autor na página 563 de Kish [1965].

O principal motivo pelo qual a amostragem por cotas é mais rápida e barata do que a amostragem probabílistica é por causa do erro de não-resposta, o qual foi tratado com mais detalhes em 1.3.2. Esse erro ocorre quando a pessoa selecionada não responde ao questionário ou a um item do questionário. Quando o universo de interesse é uma população humana, isso pode ocorrer por diversos motivos: pode ser difícil de localizá-la, como com pessoas que moram em favelas ou em cidades pequenas ou que trabalham o dia todo, o acesso a pessoa pode ser difícil, como com moradores de prédios ou condomínios fechados, e a pessoa selecionada pode se recusar a responder o questionário.

Na amostragem probabilística, o entrevistador tenta por diversas vezes entrevistar a pessoa selecionada. Esse processo de procurar a pessoa selecionada pode ser bastante demorado e caro. Para evitar um gasto excessivo, é comum limitar o número de vezes que o entrevistador deve tentar fazer contato com a pessoa selecionada. Já na amostragem por cotas, como entrevistador seleciona qualquer pessoa que ele consiga encontrar e que se encaixe nas cotas, pois ninguém foi especificamente selecionado, o tempo necessário para completar uma entrevista é usualmente bem menor, e o custo também.

Um outro motivo para utilizar a amostragem por cotas, porém raramente citado, é que controlando as covariáveis através das cotas é possível diminuir a incidência de erros não-amostrais no momento da coleta de dados. Para que isso ocorra, as variáveis da cota devem ser correlacionadas com a fonte do erro. Essa questão será discutida com mais detalhes na Seção 2.1.

A principal crítica a amostragem por cotas é que ela não é uma amostra probabilística e assim não é possível fazer **ID**. Numa tentativa de contornar esse problema, usualmente utiliza-se estimadores da **AAS** para realizar as inferências necessárias, fato que também gera críticas.

Historicamente, há muita discussão sobre qual desenho amostral deve ser utilizado em pesquisas

de populações humanas, por cotas ou probabilístico. Alguns acreditam que a amostragem por cotas é tão imprecisa e tão suscetível a vícios que a consideram quase inútil. Outros acreditam que ela não é tão precisa quanto a amostragem probabilística porém que pode ser utilizada seguramente em algumas situações, e ainda há aqueles que acreditam que se as instruções forem bem precisas e se os controles impostos sobre a liberdade dos entrevistadores ao selecionarem os respondentes forem suficientes, que a amostragem por cotas pode ser bastante precisa, o suficiente para que a redução do custo seja suficientemente maior do que a redução da precisão a ponto de justificar sua utilização.

Não parece haver uma resposta única e definitiva para a questão, no sentido de que qualquer um dos dois desenhos pode ser considerado melhor em todas as situações, dependendo dos argumentos utilizados. Mais que isso, existem situações onde ninguém sugeriria o uso de uma amostra por cotas, e outras onde é impossível fazer a amostragem probabilística. Apesar disso, existe uma vasta gama de cenários onde qualquer um dos dois métodos pode ser utilizado. Na Seção 2.4 serão apresentadas com mais detalhes as principais críticas à amostragem por cotas, e na Seção 2.6 as justificativas teóricas existentes. Também são de bastante interesse as comparações empíricas entre amostragem probabilística e amostragem por cotas, as quais serão apresentadas na Seção 2.5.

2.1 Variáveis de cota

Determinar quais variáveis devem ser utilizadas para fazer as cotas de uma pesquisa está relacionado a forma como a amostragem por cotas é justificada. As justificativas teóricas e empíricas existentes serão discutidas com mais detalhes nas seções 2.6 e 2.5. Das justificativas que usualmente são consideradas para a amostragem por cotas, 3 critérios diferentes podem ser utilizados para formular as cotas. São eles:

1- Correlacionadas com a variável Y Determinar cotas que são correlacionadas com a variável de interesse Y, garantindo assim que a variável Y seja bem representada pela amostra.

2- Miniatura da População Determinar o máximo de cotas possíveis, de forma a garantir que a amostra seja "representativa", no sentido de que ela se pareça com o universo de interesse em todas as covariáveis que são conhecidas para a população, supostamente garantindo que a variável Y também seja bem representada na amostra.

3- Correlacionadas com a probabilidade de Resposta Determinar cotas que sejam correlacionadas com as probabilidades de resposta das pessoas, para garantir que perfis populacionais que são difíceis de serem encontrados/entrevistados pertençam a amostra. Dessa forma, dentro de cada cota as pessoas teriam a mesma probabilidade de pertencer a amostra, possibilitando fazer inferência do ponto de vista de **ID**.

Para definir as cotas, é necessario utilizar variáveis que sejam conhecidas para toda a população, ou seja, é necessário um compromisso entre quais cotas seriam teoricamente desejáveis e quais são possíveis, tanto do ponto de vista da disponibilidade das informações quanto da logística da coleta de dados, pois variáveis difíceis de serem mensuradas podem representar mais uma dificuldade

para o entrevistador. Geralmente utilizam-se covariáveis sócio-demográficas, como por exemplo sexo, idade, grau de instrução, situação trabalhista, faixas de renda e classe social. Usualmente as cotas são proporcionais as quantidades populacionais dessas covariáveis.

Do ponto de vista do critério 1, para cada pesquisa a variável de interesse Y pode ser diferente, assim é muito difícil afirmar se as cotas são realmente correlacionadas com a variável de interesse de uma particular pesquisa. Pode ser que alguma pesquisa consiga atingir esse objetivo, mas provavelmente a maioria das pesquisas não consegue. Na Seção 3.1.3, discutiremos se no caso das pesquisas eleitorais covariáveis sócio-demográficas podem ser consideradas relacionadas ao voto eleitoral.

Do ponto de vista do critério 2, a princípio não existe um interesse especial em uma determinada covariável, todas que forem conhecidas a nível populacional podem ser utilizadas. Um ponto é importante nesse cenário: não é possível controlar todas as variáveis, pois isso impossibilitaria a coleta de dados, como será discutido na Seção 2.2, ou seja, é preciso limitar a quantidade de covariáveis utilizadas. Dependendo de como o conceito subjetivo de "amostra representativa"seja definido, pode ser possível priorizar algumas variáveis, que podem tornar a amostra mais "representativa"sob esse ponto de vista. Note que o conceito de "amostra representativa"não foi definido nessa seção intencionalmente, pois não existe uma definição universalmente aceita, como pode ser visto em Cochran et al. [1954]. Essa definição é ainda mais questionável quando existe mais de uma variável de interesse Y.

Já do ponto de vista do critério 3, existem muitas evidências empíricas de que as variáveis sócio-demográficas estão correlacionadas com as probabilidades de resposta das pessoas. Como foi discutido na Seção 1.3.2, existem diversos tipos de erro de não-resposta. Dependendo do tipo de pesquisa sendo realizada, a probabilidade de resposta pode considerar diferentes tipos de não-resposta. Por exemplo, numa pesquisa por telefone, não existe a necessidade de encontrar o domicílio da pessoa selecionada ou de conseguir autorização do porteiro para contactar um morador, já em pesquisas pessoais, esses pontos podem ser determinantes para a probabilidade uma entrevista ser completada.

Covariável	Categoria	Probabilidade de Resposta Estimada
Sexo	Masculino	72%
	Feminino	79%
Idade	17-24	61%
	25-44	86%
	45-64	81%
	65-74	57%
Educação	0-11 anos	58%
	12 anos	72%
	13 anos ou +	96%
Atividade	Trabalhando	80%
	Dona de Casa	85%
	Outras	47%
Total		75%

Figura 2.1: Probabilidades de resposta estimadas por categoria de covariáveis sócio-demográficas

Em Groves [1989], o autor analisa diversas pesquisas realizadas com o objetivo de avaliar covariáveis correlacionadas com a probabilidade de resposta. Alguns dos resultados são apresentados na figura 2.1. Analisando a tabela, é evidente que dependendo das categorias das covariáveis sócio-demográficas, as probabilidades de resposta se alteram. Por exemplo, nessa pesquisa, a probabilidade de resposta das mulheres é maior do que dos homens e quanto mais anos de estudo, maior a taxa de resposta. Essas relacões não são necessariamente as mesmas em pesquisas diferentes, quanto mais em diferentes países, mas provavelmente algum tipo de relação entre covariáveis sócio-demográficas e a probabilidade de resposta existirá.

Pensando especificamente em erros de não resposta que ocorrem em pesquisas domiciliares, usualmente dos tipos "Não-contactada porém conhecida e elegível (NC)"e "Elegível que recusou (R)", discutiremos o racional para inclusão de cada variável sócio-demográfica na definição das cotas:

Sexo No geral, é mais fácil encontrar mulheres em casa, como donas de casa e idosas.

Idade No geral, é mais fácil encontrar pessoas mais jovens e mais idosas em casa, ou seja, pessoas fora da idade produtiva.

Ocupação No geral, pessoas que trabalham são mais difíceis de serem encontradas em casa do que pessoas que não trabalham.

Escolaridade No geral, pessoas com maior escolaridade se recusam mais a responder as pesquisas. Esse efeito pode ser confundido com a covariável Ocupação, pois estudantes não trabalham.

Classe Social ou Renda No geral, pessoas de maior classe social ou renda mais alta são mais inacessíveis, pois moram em condomínios fechados ou em apartamentos, ou se recusam a responder.

Tipo de Residência No geral, pessoas que residem em apartamentos são mais difíceis de serem contactadas do que moradores de casas, com exceção de condomínios fechados e de mansões.

Como dependendo da pesquisa sendo realizada diferentes tipos de erros de não-resposta podem influenciar, claramente outras variáveis sócio-demográficas também podem ser consideradas.

2.2 Tipos de cotas

Existem dois tipo de cotas, as cotas cruzadas, ou interrelacionadas e as cotas marginais ou independentes. As primeiras são o tipo de cota que mais se assemelha com a amostragem estratificada. A vantagem dessas cotas é que elas preservam as relações entre as categorias das diferentes variáveis de cota, ou seja, a estrutura de correlação dessas variáveis são reproduzidas na amostra. A dificuldade com essas cotas é que se muitas categorias das covariáveis forem controladas, será muito difícil o entrevistador conseguir completar todas as entrevistas.

As cotas marginais, ou independentes, são uma alternativa, pois permitem controlar diversas covariáveis, porém somente as marginais dessas variáveis são reproduzidas na amostra, ou seja, as

Entrevistas	SEXO	
IDADE	MASCULINO	FEMININO
De 16 a 24 anos	2	2
De 25 a 34 anos	1	1
De 35 a 44 anos	1	1
De 45 a 59 anos	1	1
De 60 anos ou mais	0	0

(a) Cotas Cruzadas

	Entrevistas
IDADE	
De 16 a 24 anos	4
De 25 a 34 anos	2
De 35 a 44 anos	2
De 45 a 59 anos	2
De 60 anos ou mais	0

SEXO	
MASCULINO	5
FEMININO	5

(b) Cotas Marginais

Figura 2.2: Tipos de Cotas

cotas não preservam a estrutura de correlação dessa variáveis. Na figura 2.2, exemplos dos dois tipos de cotas são apresentados.

Também é possível fazer cotas híbridas, ou seja, que tenham os dois tipos de estrutura. Por exemplo, numa pesquisa onde deseja-se controlar 3 variáveis: sexo, idade e escolaridade. É possível fazer as cotas de maneira a preservar apenas as estruturas de correlação entre sexo e idade e também entre sexo e escolaridade. Porém as estruturas de correlação conjunta das 3 variáveis não são preservadas, nem a estrutura de correlação entre a idade e a escolaridade. Essa é uma forma de tentar manter uma parte das correlações porém também permitir o controle de mais covariáveis. Um exemplo de uma cota híbrida pode ser visto na figura 2.3.

Entrevistas	SEXO	
IDADE	MASCULINO	FEMININO
De 16 a 24 anos	1	4
De 25 a 34 anos	1	0
De 35 a 44 anos	1	0
De 45 a 59 anos	0	0
De 60 anos ou mais	0	1
Total	3	5

Entrevistas	SEXO	
ESCOLARIDADE	MASCULINO	FEMININO
Primário Inc. / Comp.	1	1
Ginásio Inc. / Comp.	0	0
Colegial Inc. / Comp.	1	2
Superior Inc. ou mais	1	2
Total	3	5

Figura 2.3: Cotas Híbridas - Combinado cotas cruzadas com marginais

Também é importante mencionar, em pesquisas que têm alguma dispersão geográfica, ou seja, onde a amostra é espalhada em diferentes áreas geográficas, usualmente calcula-se uma cota por área. Dependendo da quantidade de áreas geográficas, as cotas de cada área podem ter uma quantidade pequena de entrevistas, o que também pode restringir bastante a quantidade de variáveis de cota, além dos problemas de arredondamento decorrentes de se tentar distribuir proporcionalmente um pequena quantidade de entrevistas entre as cotas. Um exemplo onde esse tipo de problema

pode ocorrer é na amostragem probabilística com cotas, que será descrita em 2.3. Uma descrição bastante detalhada de uma amostra por cotas discutindo problemas enfrentados na prática é apresentada no Capítulo 12 de Stephan and McCarthy [1958]. Uma comparação empírica das cotas cruzadas com as marginais é apresentada na Seção 2.5.1.

2.3 Tipos de Desenhos Amostrais com Cotas

Existem muitas formas diferentes de se planejar e executar uma amostra por cotas, considerando outras características amostrais além das questões especificamente relacionadas as cotas, as quais foram discutidas nas seções 2.1 e 2.2. Existe uma tendência a se classificar todas as amostras por cotas como equivalentes, porém esse claramente não é o caso. Essa questão da diversidade de desenhos amostrais por cotas é muito bem explicada em Stephan and McCarthy [1958], onde o autor afirma que *"Não é suficiente simplesmente dizer que amostragem por cotas foi utilizada e esperar que alguém tenha mais do que uma vaga idéia de como a amostra foi selecionada"*.

Diversas características da amostra por cotas podem ter muita relevância nos resultados obtidos. De maneira geral, os 4 fatores descritos a seguir podem ser considerados como mais importantes:

1- Controle do Entrevistador Esse fator diz respeito as instruções que o entrevistador recebeu para selecionar o respodente. Ele pode ser instruído a encontrar qualquer pessoa para completar as cotas ou utilizar um procedimento sistemático para selecioná-las, o qual seria executado da mesma forma por todos os entrevistadores de uma pesquisa.

2- Distribuição Geográfica Esse fator diz respeito a distribuição geográfica da amostra. Ela pode ser totalmente concentrada em somente um local, em alguns poucos locais, ou ter uma grande dispersão geográfica.

3- Estágios probabilísticos Esse fator está relacionado com a distribuição geográfica. As áreas geográficas pertencentes a amostra podem ter sido selecinadas de duas maneiras: probabilística ou determinística.

4- Entrevistas Domiciliares Esse fator indica se as entrevistas foram realizadas com as pessoas selecionadas enquanto elas estavam em seu domicílio ou não.

Esses fatores são fundamentais para avaliar a qualidade de uma amostra por cotas. O fator 1 é importante porque a liberdade do entrevistador em selecionar as pessoas que pertencerão a amostra talvez seja a maior fonte de vícios. Existem entrevistadores que têm menor ou maior capacidade de abordar pessoas de diferentes perfis sócio-demográficos, o que claramente terá impacto nas probabilidades de resposta. A vantagem de haver um procedimento sistemático para a seleção das pessoas é justamente evitar que as probabilidades de resposta, que já são afetadas pelos perfis das pessoas, também sejam afetadas pelos entrevistadores. Mesmo supondo que os vícios induzidos pelos diferentes entrevistadores se anulem, a existência de um procedimento sistemático diminue a variância dos estimadores obtidos.

O fator 2 é também muito importante, pois sabe-se que em áreas urbanas pessoas de diferentes localidades são muito diferentes entre si, tanto do ponto de vista de covariáveis, quanto provavelmente também com relação a variável Y de interesse. Distribuir a amostra geograficamente, nesse sentido, está relacionada as vantagens que obtemos em utilizar uma amostra estratificada. Quanto ao fator 3, que está relacionado a forma como essa distribuição é feita, é importante que covariáveis conhecidas para as diferentes áreas sejam controladas, e a vantagem da seleção probabilística é que ela permite, juntamente com algumas suposições sobre as probabilidades de resposta que serão apresentadas no Capítulo 4, que inferências baseadas no desenho amostral sejam realizadas. Já o fator 4 é importante para garantir que essas suposições mencionadas sejam mais plausíveis, além de provavelmente diminuir erros de mensuração, conforme será descrito na Seção 2.4.

O ideal é que uma amostra tenha todas as características descritas acima, ou seja, exista um procedimento sistemático a ser seguido pelo entrevistador, que a amostra seja distribuída geograficamente, controlando carcaterísticas importantes e através de amostragem probablística, e que as pessoas sejam entrevistadas em suas residências. Um exemplo desse tipo de amostragem por cotas é a amostragem probabilística por cotas, que será discutida com mais detalhes na Seção 2.3.1.

Muitas das críticas e dos erros da amostragem por cota provém da ausência de algumas dessas características. Um exemplo são as chamadas pesquisas em pontos de fluxo, as quais apesar de serem de rápida execução, usualmente têm uma distribuição geográfica bastante limitada, a liberdade dos entrevistadores para selecionar as pessoas é muito grande, não existe nenhum estágio probabilístico e as entrevistas são realizadas fora de casa, em locais de grande fluxo de pessoas. Esse tipo de amostragem é comum em pesquisas eleitorais (principalmente aquelas realizadas em cidades pequenas), e será discutida com um pouco mais de detalhe na Seção 3.2.2.

2.3.1 Amostragem Probabilistica por Cotas (APC)

Nessa seção faremos um resumo detalhado do Capítulo 2 do livro Sudman [1967]. O autor foi quem cunhou o termo Amostragem Probabilística com Cotas, e formalizou as principais vantagens e desvantagens desse desenho amostral com relação a amostragem probabilística **(AP)**.

É importante ressaltar que esse livro foi escrito em 1967, e nessa época utilizar estimadores que são médias ponderadas, como os estimadores HH e HT apresentados na Seção 1.2.6, não era uma tarefa trivial. A forma mais comum de evitar esse problema era fazer desenhos amostrais complexos porém nos quais as probabilidades de seleção/inclusão fossem iguais para todas as unidades populacionais (conhecidas como amostras auto-ponderadas, discutidas na Seção 1.2.7), permitindo que os estimadores da **AAS** fossem utilizados. Claro que procedendo dessa forma ainda existe a questão de estimar a variância dos estimadores, principalmente por causa do efeito de conglomeração, porém essas questões, quando pertinentes, são discutidas pelo autor no próprio texto. Assim, quando o autor afirma que um desenho é viciado, ele na verdade quer dizer que usando o estimador média simples da **AAS** o desenho é viciado, porém se as probabilidades de seleção/inclusão forem conhecidas e o estimador de HH/HT for utilizado, os desenhos amostrais discutidos não são viciados.

Segundo o autor, a principal diferença entre o amostragem probabilística por cotas e a amostragem por cotas é a existência de controles geográficos rígidos que devem ser respeitados pelo entrevis-

tador, o qual deve seguir um trajeto específico, visitando domicílios pré-designados. Também, o autor afirma que talvez a principal diferença entre **APC** e **AP** seja a velocidade de conclusão das entrevistas, pois **APC** sem muito urgência podem ser concluídas em 2 ou 3 semanas, já usualmente a amostragem probabilística (**AP**) leva 6 semanas ou mais.

Suposições da APC

Em amostragem probabilística com voltas **(APV)**, conforme discutido na Seção 1.3.3, a cada entrevistador é designado um domicílio ou pessoa específica para entrevistar. Se o indivíduo não está disponível na primeira tentativa de contato do entrevistador, repetidas voltas são realizadas até que a entrevista seja realizada ou o respondente recuse conceder a entrevista.

Na APC, a suposição básica é que é possível dividir os respondentes em estratos nos quais a probabilidade de uma pessoa estar disponível para ser entrevistada é conhecida e é a mesma para todas as pessoas de um mesmo estrato, porém diferente entre estratos. A probabilidade de qualquer pessoa ser entrevistada é dada pelo produto da sua probabilidade inicial de seleção vezes a sua probabilidade de ser entrevistado.

Existe uma suposição implícita de que o entrevistador em uma determinada área geográfica realiza entrevistas nos mesmos horários em pesquisas repetidas e que a probabilidade de estar disponível para ser entrevistado dos respondentes depende de características conhecidas. Dessa forma, **as cotas devem ser claramente relacionadas com as probabilidade das pessoas estarem disponíveis para serem entrevistadas.** Essencialmente, as cotas devem ser baseadas nos inversos das probabilidades de estar disponível. Se a probabilidade de pessoas do Estrato A é duas vezes maior do que a probabilidade de pessoas do Estrato B, então o tamanho amostral do estrato A deve ser metade do tamanho amostral do Estrato B, **para que a amostra seja auto-ponderada**.

Usualmente, as cotas são definidas para um determinado estrato baseado no tamanho das probabilidades de resposta e nas estimativas do tamanho populacional do estrato. Além disso, as cotas são usualmente determinadas para a menor unidade geográfica para a qual existe informação. Esse método introduz a possibilidade de erro por causa de estimativas inadequadas dos tamanho dos estratos.

A **APC** tem sido utilizada primariamente para amostras de respondentes individuais. Se o interesse fosse comportamento ou opiniões domiciliares, seria possível utilizar o mesmo procedimento, mas como tamanho do domicílio está altamente correlacionado com a disponibilidade para responder, seria necessário utilizá-lo como uma cota importante. Como qualquer adulto é aceitável como respondente em pesquisas domiciliares, **APV** dos domicílios é mais barata do que dos indivíduos, e nesse contexto talvez a **APC** não seja mais rápida e/ou barata do que a **APV**.

A **APC** depende de uma suposição muito importante, enquanto que a **APV** não. Felizmente, na próxima sessão serão apresentadas evidências de que essa suposição é quase-verdade na maioria dos tipos de pesquisas realizadas nos Estados Unidos. Mesmo na **APV** existem vícios, devido as pessoas que se recusam a participar ou não são encontradas, os quais provavelmente também existem na **APC**.

É interessante comparar o racional da **APC** e da **APV** com o método de ponderação de Politz-Simmons, descrito em Politz and Simmons [1949a] e Politz and Simmons [1949b], utilizado para ajustar o vício causado pelas pessoas que não ficam em casa (Not-At-Homes), ou seja, que não estão disponíveis para serem entrevistadas. Nesse procedimento não são feitas voltas e nem cotas são utilizadas. Tipicamente, pergunta-se ao respondente se ele estava em casa nas últimas 5 noites, e a sua resposta é utilizada para determinar o seu peso. Assim, um respondente que esteve em casa as 5 noites recebe peso 1, enquanto que um respondente que não esteve em casa nas 5 noites anteriores recebe o peso de 6, pois somente $\frac{1}{6}$ desses respondentes seriam encontrados em casa numa noite aleatória qualquer.

O uso do esquema de ponderação de Politz-Simmons tem duas desvantagens principais: depende da memória do respondente sobre sua disponibilidade nas noites anteriores, a qual usualmente é super-estimada, e o uso de pesos da pós-estratificação aumenta a variância amostral, como foi mostrado na Seção 1.2.3. Se a pergunta sobre disponibilidade em casa fosse considerada confiável, e não houvesse preocupações com o custo de se ponderar a amostra, seria possível desenvolver um método de amostragem combinado, utilizando cotas e a pergunta sobre disponibilidade em casa para remover pequenos vícios decorrentes da suposição utilizada pela **APC**.

Características do Respondente relacionadas com Disponibilidade para ser entrevistado

Como definir os estratos dentro dos quais as pessoas têm a mesma probabilidade de resposta? É necessário recorrer a experiências anteriores com **APV**. Muitos estudos anteriores mostram que mulheres geralmente estão mais disponíveis do que os homens, primariamente por que existem mais homens trabalhando do que mulheres. Se também for controlada a variável de situação trabalhista, existe uma grande diferença de disponibilidade entre mulheres que não-trabalham e que trabalham. Também, a idade dos homens parece ser relevante.

Estatística	Amostra	Todos os Respondentes	Homens			Mulheres		
			Total	Menos de 30	30 ou mais	Total	Empregadas	Desempregadas
Número de contatos até Completar a entrevista	Todos os lugares	2.7	3	3.2	2.9	2.5	3	2.2
	10 maiores RM's	3.2	3.2	3.4	3.2	3.3	3.9	2.8
	Outras RM's	2.9	3.3	3.5	3.2	2.5	2.9	2.2
	Fora das RM's	2.3	2.4	2.8	2.4	2.1	2.6	1.9
Probabilidade de Completar a Entrevista	Todos os lugares	28%	23%	24%	22%	31%	19%	40%
	10 maiores RM's	19%	18%	36%	14%	20%	10%	27%
	Outras RM's	26%	21%	30%	18%	30%	16%	41%
	Fora das RM's	35%	28%		31%	40%	30%	45%

Figura 2.4: Média de voltas (contatos) necessárias para conseguir realizar uma entrevista e a Probabilidade de Completar uma Entrevista.

O Centro Nacional de Pesquisas de Opinião (CNPO)[2] dos Estados Unidos, desenvolveu um sistema com 4 cotas, compostas por homens com menos de 30 anos, homens com mais de 30 anos, mulheres que trabalham e mulheres que não trabalham. Outro fator muito importante para determinar a disponibilidade é dado pelo tamanho da comunidade onde o respondente reside, porém esse fator usualmente é controlado ao se selecionar a amostra com probabilidades proporcionais ao

[2]Em inglês, National Opinion Reserach Center (NORC)

tamanho nos primeiros estágios da amostra. Na figura 2.4, são apresentadas a média de voltas (contatos) necessárias para conseguir realizar uma entrevista e a probabilidade de completar uma entrevista segundo sexo, idade, trabalho e localidade, para diversas **APV** realizadas pelo CNPO. Não se afirma que essas cotas são ótimas, ou ideais, apenas que elas têm funcionado bem na prática. Sempre é possível refinar essas cotas, porém a coleta de dados fica mais difícil, ou seja, mais cara e demorada.

Para melhor compreender a **APC**, é útil considerar a probabilidade de se completar uma entrevista em somente uma tentativa. Existe um grande aumento na probabilidade de se encontrar um respondente depois da primeira volta, assim utilizar a média desconsiderando o número de voltas (tentativas) superestima a probabilidade do respondente estar disponível. Na figura 2.5 são apresentadas as probabilidades de entrevistar o respondente segundo o número de tentativas.

Pesquisa do CNPO	Número de Tentativas				
	1	2	3	4	5
Pesquisa 1	36%	66%	56%	54%	50%
Pesquisa 2	42%	44%	52%	48%	48%

Figura 2.5: Probabilidade de completar a entrevista segundo o número de tentativas.

Outro fator claramente correlacionado com a probabilidade de realizar uma entevista em uma única tentativa é o número de moradores no domicílio. Pessoas que moram num domicílio maior têm uma chance maior de serem selecionadas. Essa característica pode ser observada na figura 2.6.

Amostra	Todos os Respondentes	Número de moradores no Domicílio					
		1	2	3	4	5	6 ou mais
Todos os lugares	56%	46%	52%	56%	58%	63%	67%
10 maiores RM's	44%	29%	36%	44%	50%	58%	59%
Outras RM's	56%	47%	51%	55%	58%	62%	67%
Fora das RM's	63%	55%	62%	64%	61%	66%	70%

Figura 2.6: Probabilidade de completar a entrevista segundo o número de moradores do domicílio.

Verificando Homogeneidade das Cotas

Para avaliar se as cotas da **APC** realmente são homogêneas, é razoável utilizar a distribuição geométrica como a distribuiçao teórica contra a qual a distribuição empírica das probabilidades de completar uma entrevista são comparadas.

Assumindo que o entrevistador da **APC** está procurando respondentes aleatoriamente e que a sua probabilidade de completar uma entrevista em qualquer domicílio em uma determinada área é a mesma que a probabilidade de completar a entrevista na primeira tentativa da **APV** na mesma área. Pensando um pouco sobre essas suposições, é evidente que o entrevistador da **APC** não está buscando domicílios aleatóriamente, porém por causa das cotas o entrevistador é obrigado a buscar respondentes de dia e de noite, e até nos fins de semana, de forma que os horários de busca do entrevistador se aproximam de um procedimento aleatório. Com relação a segunda suposição, ela é realista pois é o mesmo entrevistador que usualmente realiza entrevistas no mesmo período, independente de estar trabalhando com **APC** ou **APV**.

Amostra	Valor	Todos os Respondentes	Homens			Mulheres		
			Total	Menos de 30	30 ou mais	Total	Empregadas	Desempregadas
Todos os lugares	Observado	3.6	3.8	3.6	3.9	3.7	4.4	3.4
	Esperado	3.6	4.3	4.2	4.5	3.2	5.3	2.5
10 maiores RM's	Observado	5.8	5.6	5.6	5.5	6.0	7.0	5.3
	Esperado	5.3	5.6	3.8	7.1	5.0	10.0	3.7
Outras RM's	Observado	3.4	3.8	3.9	3.8	3.4	3.6	3.3
	Esperado	3.8	4.8	3.3	5.6	3.3	6.3	2.4
Fora das RM's	Observado	2.5	2.5	2.0	2.7	2.5	2.9	2.3
	Esperado	2.9	3.6		3.2	2.5	3.3	2.2

Figura 2.7: Comparação entre a média observada e a média esperada segundo o modelo Geométrico.

Utilizando essas suposições, o número de contatos necessários para completar uma entrevista na **APC** tem uma distribuição Geométrica, com valor esperado $\frac{1}{p}$ e variância $\frac{(1-p)}{p^2}$. Na figura 2.7, compara-se o valor observado obtido da **APC**, com o valor previsto utilizando o modelo Geométrico e as probabilidades de se entrevistar o respondente em uma única tentativa, obtida da **APV**. Para obter esses valores da **APC**, durante a coleta de dados foi mantida uma lista detalhada por todos os entrevistadores, onde detalhou-se todas os domicílios visitados, e os resultados obtidos. Nas estimativas, não foram incluídos domicílios vazios e estabelecimentos comerciais. Nessa tabela, percebe-se que os valores são próximos, dando credibilidade ao modelo. Com relação a variância observada e esperada, algumas diferenças são discrepantes, porém esse estimador é muito sensível a pequenos desvios na probabilidade estimada, como pode ser visto em Sudman [1967], onde um teste para verificar a homogeneidade das cotas também é apresentado.

Comparação entre os custos da APC e da APV

O argumento principal feito a favor da **AC** era o seu custo. No caso da **APC**, o custo é um pouco mais alto, porém continua sendo menor do que o da **APV**. Nessa seção, são comparados os custos de **APC** e da **APV** realizadas pelo CNPO Americano. Esses resultados são apresentados na figura 2.8. Uma breve análise dessa tabela revela que uma grande parte da diferença de custo entre os dois tipos de desenhos amostrais se deve a diferenças no planejamento, processamento e análise das amostras. Quase sempre, planejamento e análise da **APV** são mais caras, e consomem a maioria dos recursos desse tipo de pesquisa.

Custo	Amostra Probabilística com Voltas						Amostragem Probabilística com Cotas			
	Pesquisa 1	Pesquisa 2	Pesquisa 3	Pesquisa 4	Pesquisa 5	Pesquisa 6	Pesquisa 1	Pesquisa 2	Pesquisa 3	Pesquisa 4
Custo Direto de Coleta de Dados	$ 31,800	$ 21,000	$ 19,500	$ 5,000	$ 22,000	$ 16,900	$ 8,900	$ 9,900	$ 8,500	$ 9,000
Supervisão da Coleta de Dados	$ 8,100	$ 29,500	$ 4,900	$ 2,500	$ 9,500	$ 6,000	$ 1,900	$ 1,200	$ 1,200	$ 1,900
Outros Custos da Pesquisa	$ 173,100	$ 106,200	$ 93,400	$ 31,400	$ 38,500	$ 26,500	$ 16,000	$ 14,100	$ 14,100	$ 14,800
Custo Total	$ 213,000	$ 156,700	$ 117,800	$ 38,900	$ 70,000	$ 49,400	$ 26,800	$ 25,200	$ 23,800	$ 25,700
Tamanho da Amostra	2380	2810	2200	760	2500	1500	1200	1500	1300	1500
Custo Total por Entrevista	$ 89.5	$ 55.8	$ 53.5	$ 51.2	$ 28.0	$ 32.9	$ 22.3	$ 16.8	$ 18.3	$ 17.1
% do Custo por Entrevista proveniente da Coleta de Dados	19%	32%	21%	19%	45%	46%	40%	44%	41%	42%

Figura 2.8: Comparação entre o custo da **APC** e da **APV**, em dolares (US$).

Comparação Empírica entre APC e a APV

Essa última seção sobre **APC**, compara os resultados entre 3 pesquisas, a primeira sendo um **APV** realizada em junho de 1963, a segunda sendo **APC** realizada em dezembro de 1963 e uma pesquisa híbrida, sendo metade **APC** e metade **APV**, realizada em dezembro de 1963. Os resultados encontrados não provam que a **APC** é não-viciada, porém suportam a visão de que na maioria dos itens avaliados esses vícios são pequenos. As pequenas diferenças observadas podem ser justificadas pela baixa correlação entre a probabilidade de resposta e as variáveis analisadas. Foram comparadas 5 perguntas, nas quais as diferenças observadas podem ser justificadas pelo erro amostral e por detalhes específicos de cada amostra, e também foram comparadas 17 variáveis demográficas, onde só foram encontradas diferenças entre sexo e tamanho do domicílio. Na primeira, porque os homens têm probabilidade de estarem disponíveis para responder menor do que as mulheres, assim são subestimados na **APV**, problema que não ocorre na **APC** pois sexo é uma variável da cota. Na segunda, os resultados sugerem que a **APC** é deficiente na quantidade de domicílios com 1 e 2 pessoas, pois domicílios menores têm menos chance de ter pessoas disponíveis para serem entrevistadas nele, sugerindo que tamanho do domicílio talvez deva ser considerado como variável de cota. Detalhes dessa comparação empírica podem ser encontrados em Sudman [1967].

2.4 Críticas à Amostragem com Cotas

Aa críticas a **AC** serão apresentadas sem especificar a qual tipo de **AC** elas dizem respeito, porém ficará evidente que muitas das críticas não dizem respeito a **APC** como apresentada aqui. Em Moser [1952], o autor destaca as 6 principais críticas à **AC**, as quais serão reproduzidas aqui:

Seleção Probabilística Como a seleção da **AC** não é probabilística, não é possível calcular as precisões das estimativas do ponto de vista da inferência baseada no desenho.

Classe Social É comum utilizar a variável Classe Social nas cotas, porém 2 críticas podem ser feitas a essa prática: não existem informações disponíveis de orgãos oficiais e existem problemas com a definição das classes sociais, pois por exemplo, uma pessoa de renda alta pode pertencer a classe C se ela não possuir muitos bens, e um domicílio pobre porém com muitos moradores pode ser classificado na classe B.

Liberdade dos Entrevistadores Se os entrevistadores tiverem muita liberdade para selecionar os respondentes, pessoas que não-cooperam podem ser sub-representadas na amostra.

Disponibilidade Se as cotas utilizadas não forem homogêneas, podem existir sub-grupos com maior ou menor disponibilidade de serem entrevistados, viciando os resultados.

Local das entrevistas Vícios podem ser introduzidos pela peculiaridade do local de entrevista. Respostas de pessoas sendo entrevistadas nas ruas ou pontos de fluxo podem ser viciadas. Por exemplo, uma pessoa pode estar menos disposta a declarar sua renda sendo entrevistada na rua do que se for entrevistada na sua casa.

Verificação Dependendo de onde e como são realizadas as entrevistas, pode ser difícil de verificar o trabalho do entrevistador, tanto do ponto de vista de aplicação do questionário quanto do preenchimento das cotas.

Usualmente, as críticas a **AC** e as pesquisas eleitorais se confundem, como por exemplo no artigo Lynn and Jowell [1996]. Não entraremos em mais detalhes aqui, porém sempre que as previsões das pesquisas eleitorais falham, a crítica é direcionada ao desenho amostral, porém muitas vezes o problema não é a amostra, e sim como ela é utilizada para fazer inferência, como foi mostrado no final da Seção 1.2.2, onde a impossibilidade de se fazer uma previsão do resultado da eleição foi decorrente dos institutos de pesquisa não levarem em consideração a covariância entre os estimadores dos percentuais de voto dos candidatos. Na Seção 3.3 discutiremos as críticas feitas as pesquisas eleitorais, tomando o cuidado de separá-las em duas classes: críticas relacionadas a amostra e as inferências realizadas.

Nessa mesma linha, existem algumas publicações brasileiras, com críticas dirigidas especificamente as pesquisas eleitorais brasileiras. Os trabalhos mais críticos e sérios, nesse sentido, são Ferraz [1996] e de Souza [1990]. Em ambas, as críticas são direcionadas as amostras por cotas, porém não se faz distinção entre **AC** e **APC**, as quais são muito diferentes entre si, com foi discutido na Seção 2.3.1, e como também pode ser observado na Seção 2.5.

Muitas vezes as críticas as pesquisas eleitorais e a **AC** são bastante inflamadas e radicais. Usualmente, essas críticas não distinguem a **APC** da **AC**, e também são feitas do ponto de vista de inferência baseada no desenho **(ID)**. Um bom exemplo desse tipo de crítica pode ser visto no texto "A Falsidade das Margens de Erro em Pesquisas Eleitorais baseadas em Amotragem por Quotas"[3] em Carvalho and Ferraz [2006], onde os autores afirmam:

> "Se não se pode fazer um levantamento de modo adequado, melhor não fazer! E se não se garante a precisão das estimativas, melhor não se divulgar algo, possivelmente errado, que pode influenciar o processo eleitoral. É de todo recomendável que as autoridades que conduzem o processo eleitoral exijam não apenas a **declaração** da margem de erro, mas que demandem a **prova** de que a margem é mesmo aquela declarada."

2.5 Comparações empíricas entre APV e a AC

Algumas comparações empíricas já foram feitas entre a amostragem por cotas e a amostragem probabilística. Nessa seção serão apresentados os resultados das duas principais comparações realizadas. Na Seção 2.5.1, compara-se a amostragem probabilística com voltas (**APV**), definida na Seção 2.3.1, com diversas variações da **AC**, para avaliar quais pontos são relevantes ao se utilizar a **AC**. Já na Seção 2.5.2, a **APV** é comparada com a **APC**, que é versão mais atual e eficiente da **AC**, na qual diversas críticas feitas a **AC** deixam de ser relevantes.

Para o leitor mais interessado, outras comparações empíricas já foram discutidas na literatura, como por exemplo em Meier and Burke [1947], Hochstim and Smith [1948] e nos Capítulos 7 e 8 de

[3] Esse boletim pode ser encontrado no site *www.redeabe.org.br/Boletins/Boletim_64.pdf*

Stephan and McCarthy [1958], porém os artigos que serão discutidos nessa seção abordam todos os temas importantes.

2.5.1 Comparação Empírica 1: AC versus APV

Nessa seção apresentaremos os principais resultados de Moser and Stuart [1953], onde o autor faz uma comparação empírica preliminar entre algumas pesquisas **APV** e **AC** já realizadas. As informações obtidas dessa comparação foram utilizadas para planejar e executar um experimento que tinha o objetivo de comparar os desenhos amostrais citados, considerando vários aspectos relevantes da **AC** que foram observados nesse estudo preliminar. A apresentação desse artigo não será muito detalhada pois ele já foi devidamente discutido em Ferraz [1996].

No estudo preliminar, compara-se uma **AC** do *British Market Research Bureau* e um levantamento social utilizando **APV** do governo britânico. Entre todas as variáveis comparadas, foram detectadas apenas três diferenças relevantes: na distribuição geográfica das entrevistas, na distribuição de renda e no tipo de ocupação dos entrevistados. A distribuição geográfica da **AC** era mais aglomerada, na **APV** havia muita não-resposta na variável de renda, além de haver mais pessoas na categoria sem renda/com renda baixa e renda alta do que na **AC**, e na **AC** haviam mais pessoas sem ocupação, trabalhando meio período e aposentados. Não há mais detalhes sobre a **AC**, porém há uma explicação plausível para todas essas diferenças que não é discutida pelos autores: se as entrevistas foram realizadas em pontos de fluxo, ou seja, locais de grande convergência de pessoas, a distribuição geográfica provavelmente seria mais concentrada, usualmente pessoas sem ocupação e/ou aposentadas não precisam ir as esses locais, consequentemente pessoas sem renda também serão sub-representadas, e pessoas com renda alta não se sentiriam a vontade de responder a pergunta de renda nesses locais, também sendo sub-representadas.

Por causa dessas observações preliminares, e de diferentes críticas recebidas pela **AC**, o estudo foi planejado para avaliar as seguintes características da **AC**:

Variáveis de Cota Avaliar até que ponto as **AC** são afetadas por incluir variáveis de controle adicionais, além das usualmente utilizadas (sexo, idade e classe social).

Tipo de Cota Avaliar o impacto dos diferentes tipos de cotas (cruzadas versus marginais).

Entrevistadores A magnitude da variabilidade causada pela mudança dos entrevistadores.

AC versus APV Avaliar a magnitude das diferenças entre **AC** e **APV**.

Localização das Entrevistas Avaliar o impacto nos resultados das entrevistas serem realizadas nas ruas, nos domícílios ou no local de trabalho.

Não-Resposta Avaliar a quantidade de recusa e de entrevistas não-produtivas.

Distribuição Geográfica Comparar a distribuição geográfica da **AC** e da **APV**.

Ordem Impacto da ordem de preenchimento das cotas.

Classe Social Avaliar os erros cometidos ao se utilizar a variável de classe social nas cotas.

Com relação aos diferentes tipos de cota de sexo, idade e classe social, verificou-se que as cotas marginais tiveram problemas para representar diversos cruzamentos das variáveis de cota, com exceção do cruzamento de Sexo com Classe Social. Na figura 2.9, os resultados das cotas marginais são comparados com os resultados esperados, entre parêntesis, segundo o censo britânico.

		Idade			
Classe Social	Sexo	20–29	30–44	45–64	65 +
Alta	M	8 (17·6)	34 (23·5)	30 (28·4)	18 (15·7)
	F	6 (20·6)	45 (24·6)	39 (42·2)	17 (18·7)
Média	M	40 (37·3)	68 (62·8)	59 (64·9)	20 (24·6)
	F	51 (38·4)	92 (65·8)	67 (71·7)	31 (38·4)
Baixa	M	140 (149·3)	220 (228·9)	195 (230·8)	124 (85·5)
	F	150 (161·0)	261 (256·2)	305 (277·0)	101 (137·4)

Figura 2.9: Comparação das cotas marginais com os valores esperados (entre parêntesis).

Não serão apresentados outros detalhes específicos desse estudo, pois o interesse maior dessa tese é na **APC**. Apresentaremos com mais detalhes apenas as conclusões dos autores.

Conclusões

Com relação as variáveis de cotas utilizadas, não foram detectadas grandes diferenças ao se utilizar cotas geográficas e de ocupação, além das já tradicionais cotas de sexo, idade e classe social.

Com relação aos tipos de cotas, apesar das cotas marginais não representarem corretamente a maioria dos cruzamentos de sexo, idade e classe social, as cotas cruzadas não obtiveram resultados melhores com relação as variáveis de interesse.

Já ao comparar a **AC** com a **APV** com relação a diversas variáveis, foram observadas algumas diferenças, porém no geral não muito substânciais. As únicas diferenças que merecem destaque são as com relação as variáveis de ocupação e educação, que apresentaram grandes diferenças, indicando que talvez essas variáveis devam ser utilizadas como variáveis de cota.

A variância amostral da **AC** foi comparada com a da **APV**, e ela foi estimada como sendo de 1 a 3 vezes maior do que das **APV**. Supõem-se que essas diferenças ocorrem por causa da variabilidade entre os entrevistadores.

De forma geral, os autores concluem que relativamente poucas diferenças significativas foram encontradas, porém que isso não justifica a **AC**. Como ponto final da argumentação, os autores enfatizam que a sensibilidade do experimento foi baixa, por causa da grande quantidade de fatores considerados.

2.5.2 Comparação Empírica 2: APC versus APV

Nessa seção, será apresentado um resumo detalhado de Stephenson [1979], onde o autor compara e discute várias questões bastante relevantes sobre a amostragem probabilística com cotas (**APC**) e a amostragem probabilística com voltas **(APV)**.

Desde os primeiros estudos e comparações da amostragem com cotas e da amostragem probabilística realizados por Moser and Stuart [1953] e Stephan and McCarthy [1958], muita coisa mudou. Um novo tipo de desenho amostral com cotas foi desenvolvido, conhecido como **APC**, o qual foi apresentado na Seção 2.3.1. Na **APC**, muitas das questões que eram criticadas na **AC** foram modificadas, deixando de serem relevantes. Por isso o interesse nessa nova comparação, porém agora da **APC** com a **APV**.

Os dados utilizados nessa comparação são do General Social Surveys (GSS) de 1975 e 1976. O GSS é um estudo que contém uma grande gama de variáveis demográficas, atitudinais e comportamentais, o qual é repetido anualmente ou segundo um planejamento regular. Os primeiros três estudos da série foram realizados de 1972 a 1974, utilizando **APC**, e de 1977 em diante utilizou-se a **APV**. Nos anos de transição de 1975 e 1976, para permitir comparações metodológicas, a amostra foi particionada, sendo utilizados ambos os tipos de desenho amostral.

Desenhos amostrais do GSS

A amostra do GSS é uma amostra estratificada de multi-estágios de conglomerados do Estados Unidos. As unidades primárias são conglomerados de domicílios similares aos setores censitários. Para cada pesquisa, foram utilizadas duas sub-amostras independentes de conglomerados. Dentro de cada conglomerado selecionado, as unidades secundárias foram particionadas em 3 segmentos. Na primeira sub-amostra, 2 segmentos foram selecionados para realizar a **APV**, e o terceiro para a **APC**, e na segunda sub-amostra utilizaram-se 2 segmentos para **APC** e o terceiro para a **APC**. Dois entrevistadores trabalharam em cada conglomerado, um deles fazendo a **APC** e a **APV**, e o segundo fazendo o tipo de amostragem que faltava para completar os 3 segmentos. Uma média de 5 respondentes por segmento foram entrevistados.

Nos segmentos onde foi utilizada a **APV**, a seleção dos domicílios foi realizada com probabilidades iguais. Os entrevistadores foram aos domicílios selecionados para completar um questionário simples de 4 páginas com algum morador, obtendo informações sobre todos os residentes. Essa informação permitiu que o entrevistador seleciona-se com probabilidades iguais um morador para responder ao questionário principal. Se necessário, um horário era agendado para entrevistar a pessoa selecionada. As decisões em todos os estágios foram independentes da disponibilidade dos potenciais respondentes.

Nos segmentos onde foi utilizada a **APC**, os entrevistadores receberam cotas de sexo, idade e situação trabalhista. Eles iniciavam a busca por respondentes em um ponto específico, e percorriam os quarteirões procurando moradores que se encaixassem nas cotas em todos os domicílios do trajeto. Somente uma pessoa podia ser entrevistada por domicílio. Foram utilizadas as cotas descritas em Sudman [1967], ou seja, compostas por homens com menos de 30 anos, homens com mais de 30 anos, mulheres que trabalham e mulheres que não trabalham, as quais foram desenvolvidas para obter amostras de grupos populacionais mais dificeis de serem encontrados: homens jovens e mulheres empregadas.

Resultados da Comparação

Para analisar os resultados, 2 tipos de vícios devem ser distinguidos. Muitos desenhos amostrais têm vícios embutidos em seus procedimentos de seleção, ou seja, seriam viciados mesmo que nenhum problema ocorresse durante a execução do desenho. Além disso, as pessoas não ficam sentadas em casa esperando entrevistadores fazerem contato, e também não cooperam toda vez que são requisitadas para participar de uma pesquisa. Vamos denominar esses dois tipos de vícios, respectivamente, de vício do desenho e vício de participação.

O desenho da **APV** têm o mérito de permitir que o seu vício de desenho seja estimado, e também de parcialmente avaliar o vício de participação. De fato, se a unidade de análise for o domicílio, e a amostra for selecionada com probabilidades proporcionais ao tamanho, a probabilidade de seleção para todos os domicílios da população serão iguais, de forma que o vício do desenho seja zero, ou seja, não é necessário utilizar os estimadores de HH ou de HT nesse contexto. Porém, é mais comum que a unidade de análise seja o indivíduo. Nesse caso, geralmente a **APV** é viciada, pois é impraticável selecionar os domicílios com probabilidades proporcionais ao número de moradores, pois essas quantidades são desconhecidas. Assim, ao invês de entrevistar todos os domicílios para obter essa informação, usualmente seleciona-se tanto os domicílios quanto os moradores com probabilidades iguais. O vício ocorre porque uma entrevista tem a mesma probabilidade de ocorrer em um domicílio grande ou pequeno, porém uma pessoa que reside num domicílio grande tem uma probabilidade menor de ser selecionada do que uma pessoa que reside num domicílio pequeno. Nesses casos, a única forma dos resultados não serem viciados é utilizando os estimadores de HH ou de HT, ou seja, é necessário ponderar os dados.

Já no caso da **APC**, é difícil avaliar o vício de desenho, e além disso, os dois tipos de vícios se confundem. No geral, domicílios maiores terão uma probabilidade maior de serem selecionados, porém essas probabilidades não são necessariamente proporcionais ao número de moradores, pois dependem das probabilidades de participação cada morador. Assim é possível que um domicílio com 2 moradores bastante cooperativos e disponíveis tenha uma probabilidade maior de ter um morador selecionado do que um domicílio com 4 moradores que não cooperam e usualmente estão indisponíveis. Mais que isso, para retirar o vício ponderando os resultados da amostra com os estimadores de HH ou de HT seria necessário conhecer as probabilidades de participação de todos os moradores de cada domicílio pertencente a amostra[4].

Os vícios da **APC** e da **APV**, para os anos de 1975 e 1976 podem ser observados nos resultados apresentados na figura 2.10. As colunas do Current Population Survey (CPS) são útilizadas como referência para o verdadeiro valor populacional, mostrando a distribuição de domicílios de cada tamanho e do número de pessoas residentes em domicílios de cada tamanho. Já nas colunas da **APV**, apresentamos a distribuição empírica dos resultados segundo tamanho de domicílio para os respondentes, que são os resultados sem ponderação, e para os respondentes, nesse caso os resultados são ponderados pelo número de moradores. Ou seja, no caso da **APV**, é possível ver que sem ponderar os dados, a distribuição é parecida com a do número de domicílios do CPS,

[4]No Capítulo 4, é apresentada uma forma de utilizar esses estimadores e retirar o vício da **APC**.

já ponderando, a distribuição se assemelha a do número de residentes do CPS. Já para o caso da **APC**, só apresentamos uma coluna, por causa da dificuldade de ponderar os resultados. Percebe-se que a **APC** super-estima o número de domicílios grandes e sub-estima o número de moradores em domicílios grandes. Se os dados forem analisados sem ponderação, como geralmente é feito, e o objetivo for obter informações dos residentes, a **APC** é menos viciada do que **APV**.

Ano	Número de Adultos	CPS (Referência)		APV		APC
		Domicílios	Pessoas	Respondentes Não-Ponderados	Respondentes Ponderados	Respondentes Não-
1975	1	23%	12%	22%	11%	13%
	2	57%	56%	59%	58%	61%
	3	14%	21%	14%	21%	18%
	4 ou mais	6%	11%	5%	10%	8%
1976	1	25%	13%	25%	14%	16%
	2	58%	58%	63%	66%	59%
	3	12%	18%	10%	15%	16%
	4 ou mais	5%	11%	3%	6%	9%

Figura 2.10: Proporção de Domicílios de diferentes tamanhos

Usualmente, o vício de desenho causado pelos diferentes tamanhos de domicílios não é muito relevante, no sentido de que usualmente ele é bastante diluído em variáveis que estão relacionadas ao tamanho do domicílio. Mesmo assim, em todas as comparações realizadas entre **APV** e **APC**, os resultados da **APV** nos quais o interesse é no indivíduo foram ponderados, e naqueles nos quais o interesse era o domicílio não foram ponderados.

Outros vícios que ocorrem nas amostras são originados pelas pessoas não estarem disponíveis ou se negarem a participar da pesquisa. Na maioria dos casos, esse tipo de vício deve ocorrer menos na **APV** do que na **APC**, pois os entrevistadores e os verificadores se esforçam mais para contactar as pessoas em pesquisas desse tipo. Pessoas que são selecionadas porém não encontradas ou que se negam a participar são contadas e incluídas na taxa de resposta.

Já na **APC**, os entrevistadores simplesmente vão para o próximo domicílio quando nenhum pessoa está acessível no domicílio, assim no geral espera-se que pessoas difíceis de serem encontradas ou difíceis de serem entrevistadas estarão mais seriamente sub-representadas na **APC** do que na **APV**, e a **APC** não permite estimar a taxa de resposta[5]. Essa situação é um pouco diferente para variáveis de cota utilizadas pela **APC**. As categorias utilizadas para as cotas serão inteiramente completadas por causa da forma como a amostra é selecionada, ou seja, a amostra não será sub-representada nessas categorias. Mas isso não garante a eliminação do vício de participação, por exemplo, completar as cotas de homens com menos de 30 só cumpre essa função se esse grupo for homogêneo, caso contrário um sub-grupo de homens dentro dessa cota pode ser sub-representado, como homens que trabalham. Nesse mesmo exemplo, se a **APV** tiver menos homens por causa dessa taxa de resposta diferente entre os homens, essa amostra pode ser pós-estratificada (ponderada) como discutido na Seção 1.2.3, mas é evidente que um estrato amostral que apresenta a proporção correta por causa de cotas ou de pós-estratificação não irá necessariamente representar o

[5]Na Seção 4.3, é apresentada uma forma de estimar a probabilidade de resposta na **APC** sob a suposição do modelo GRH.

estrato populacional corretamente. Nenhuma amostra pode ser feita para representar corretamente características de grupos populacionais que não estão incluídos nela.

Voltando aos dados, o GSS incluiu uma pergunta avaliando o nível de cooperação dos respondentes que realmente participaram da pesquisa, para ser respondida pelos entrevistadores. Essa pergunta tinha 4 categorias, indo de "amigável e interessado"a "hostil". Historicamente 80% das entrevistas foram classificadas como "amigável e interessado". Em cada uma das pesquisas de 1975 e 1976, a sub-amostra da **APV** recebeu nas categorias menos cooperativas do que "amigável e interessado", 4% a mais de respostas do que a **APC**. Isso pode representar a irritação dos respondentes com a insistência do entrevistador na **APV**, porém confirma as expectativas de que a **APV** obtenha mais pessoas que não cooperam.

Uma forma menos óbvia na qual a **APC** controla a composição dos resultados é que nas sub-amostras da **APC**, necessariamente foram realizadas 5 entrevistas em cada conglomerado, já na **APV**, conglomerados em locais com baixa cooperação ou dificuldade de encontrar as pessoas terão um número menor de entrevistas. Ou seja, de maneira geral, se a taxa de resposta for mais baixa em certas áreas geográficas, essas áreas tendem a ser melhor representadas pela **APC** do que pela **APV**, por exemplo, em áreas centrais de grandes cidades, onde usualmente a taxa de resposta é menor. Para ilustrar esse fato, nas pesquisas do GSS de 1975 e 1976, na sub-amostra da **APV**, nas áreas centrais de cidades com mais de 250.000 habitantes, a média de contatos por entrevista por conglomerado foi de 4.1, porém nas 12 maiores regiões metropolitanas dos EUA, a média foi de 3.1 em 1975 e 4.3 em 1976.

Essa situação deve ser analisada com cautela, na **APV** o efeito da não-resposta é o mesmo em características da vizinhaça (conglomerado) e nas características individuais das pessoas: qualquer atruibuto concentrado em pessoas que são difíceis de serem encontradas ou entrevistadas será sub-representado. Já na **APC**, por obrigar que um número específico de entrevistas sejam realizadas em uma vizinhança, garante que características dessa área sejam bem representadas pela amostra, independente dos moradores cooperarem ou não. Assim características de uma vizinhança com muitas pessoas difíceis de serem encontradas ou que não-cooperam serão corretamente representadas, porém as pessoas em si não serão, pois as pessoas entrevistadas podem ser as mais amigáveis e cooperativas da vizinhança.

Existem centenas de variáveis das pesquisas GSS, e existia o interesse do autor em analisar cada uma delas com o objetivo de identificar em quais havia um diferença causada pelo desenho amostral utilizado. O procedimento utilizado para identificar as variáveis onde ocorreram essas diferenças é descrito em Stephenson [1979]. Além de considerar procedimentos estatísticos, como testes de hipóteses, foram feitas análises qualitativas, procurando explicações plausíveis para as diferenças, como as fontes de vício descritas nessa seção. Das 141 comparações efetivamente realizadas, apenas 9 foram significativas a um nível de 0.05 e apenas 2 a um nível de 0.01. Além disso, o autor afirma que não há nenhum motivo estatístico para acreditar que as 130 variáveis restantes são distribuidas diferentemente entre os desenhos amostrais **APC** e **APV**, apesar de não haver como provar isso.

Conclusão

Os dois principais resultados são de que a **1) APC** sub-representa domicílios grandes, e tanto a **APC** quanto a **APV** (principalmente sem ponderação) sub-representam pessoas de domicílios grandes e que **2)** Respondentes difíceis de serem encontrados ou que não cooperam serão provavelmente pior representados na **APC** do que na **APV**, com exceção de características dos bairros onde eles residem, as quais podem ser pior representadas na **APV**.

É importante enfatizar que esses resultados dizem respeito somente a **APC** como foi definida por Sudman [1967], e não outros tipos de **AC** que não tem o controle geográfico necessário, as quais não pode se assumir que se comportam tão bem quanto a **APC**.

O autor conclui a comparação afirmando que, sem dúvida, a principal conclusão dessa comparação empírica é de que os resultados de uma **APC** são bem comportados. Pesquisas por **APC** já foram realizadas, e continuarão a ser realizadas, com resultados bem menos questionáveis do que alguém esperaria ao ler os indiciamentos mais exaltados contra a **AC**.

2.6 Justificativas Teóricas para Amostragem por Cotas

Em resposta as críticas apresentadas na Seção 2.4, usualmente os argumentos utilizados a favor da **AC** podem ser resumidos nos 8 itens a seguir:

AP Viciada Argumenta-se que a **AP** na prática também é viciada, pois a amostra selecionada não é a mesma que a amostra observada, por causa das recusas e das pessoas não encontradas.

Custo A **AC** é mais barata do que a **AP**.

Facilidade A **AC** é mais fácil do ponto de vista administrativo da pesquisas, não existe a necessidade de se preocupar com pessoas que não cooperam, com voltas ou substituições.

Tempo Se a pesquisa tem que ser completada rapidamente, pode não haver outra opção.

Classe Social Em resposta parcial a crítica da classe social, é argumentado que são utilizadas apenas poucas classes sociais nas cotas, assim não haveria problema em associar um respondente a uma particular classe.

Controle Argumenta-se que instruções e restrições ao entrevistadores são suficientes para se prevenir contra grandes vícios na seleção da amostra.

Listagem Não é necessária uma listagem para selecionar a amostra. Quando tal listagem existe, esse argumento deixa de ser relevante.

Vícios Apesar da **AC** potencialmente ser viciada em algumas tipos de variáveis, ela claramente não é em outros tipos de variáveis, e pode ser bastante útil para quem está interessado nas variáveis não-viciadas.

Apesar desses argumentos práticos, o objetivo dessa seção é apresentar justificativas teóricas que já foram feitas para a **AC**. Do ponto de vista de inferência baseadas no modelo (**IM**) ou de

inferência bayesiana baseada no modelo (**IBM**), basta que o desenho amostral seja ignorável e/ou que a não-resposta seja ignorável, como foi discutido nas seções 1.4 e 1.3.1 para que se possa realizar inferência. Em Smith [1983], o autor discute as condições para que o desenho amostral de amostras não-probabilísticas sejam consideradas ignoráveis, com particular atenção a amostragem por cotas. Com relação a amostragem por cotas, o autor afirma que o maior problema se refere a não-resposta. Além disso, no contexto de **IBM**, é possível comparar amostragem probabilística e amostragem por cotas com o objetivo de encontrar um regra de decisão ótima para a escolha de um dos dois desenhos ou até combiná-los, como é discutido em King [1985].

O interesse principal nessa seção é discutir as justificativas do ponto de vista da inferência baseada no desenho **(ID)**. A primeira justificativa é empírica, e foi dada por Stephan and McCarthy [1958]. Segundo algumas suposições razoáveis, que serão descritas a seguir, o autor afirma que repetidas aplicações de um determinado procedimento de seleção da **AC** gera uma distribuição amostral empírica para o estimador considerado e que é possível estimar a variância amostral desse estimador através de uma ou mais amostras geradas por esse mesmo processo.

Para justificar a existência da distribuição amostral empírica, supõem-se que a população, o desenho amostral, o processo de mensuração e processo de estimação permanecem constantes na replicão das amostras. Dessa forma, a variação de amostra para amostra ocorreria porque os entrevistadores selecionam diferentes pessoas para pertencer a amostra, ou por causa de diferentes pontos amostrais, ou locais de coleta de dados. Nessas amostras replicadas, não se deve permitir que as instruções aos entrevistadores, as categorias das cotas, o tamanho das cotas e outros similares a se alterarem. Mudanças nesses fatores constituem mudança do desenho amostral. Dependendo das instruções da pesquisa, mudança de entrevistadores também pode influenciar na distribuição amostral. Também é importante mencionar que se amostras forem replicadas em instantes de tempo muito distantes, elas podem não ser mais replicações da mesma distribuição amostral, pois o universo pode ter se alterado.

Com base nessas suposições, dois procedimentos diferentes podem ser utilizados para estimar a variância: 1) repetições do mesmo procedimento de seleção e 2)procedimento de seleção dividido em sub-amostras. Também existe um terceira alternativa, que não será discutida aqui, porém ela pode ser encontrada na página 220 de Stephan and McCarthy [1958].

Para o caso 1), vamos supor que o procedimento amostral foi repetido k vezes, com todos os elementos relevantes como as cotas, o tamanho amostral, etc mantidos constantes. Para cada amostra, foi obtida a estimativa p_i com $i = 1, .., k$, do estimador p_1. Assim, podemos assumir que esses valores são uma **AASc** do universo de todos os possíveis valores das estimativas, sendo que esse universo é descrito pela distribuição amostra. Dessa forma, podemos estimar a variância desse universo utilizando o estimador $\hat{V}(p_1)$, onde:

$$\hat{V}(p_1) = \frac{\sum_{i=1}^{k}(p_i - \overline{p})^2}{k-1}, \tag{2.1}$$

onde $\overline{p} = \frac{1}{k}\sum_{i=1}^{k} p_i$. É importante destacar que se o estimador p_1 utilizado for viciado, não é possível estimar o seu vício, porém o esimador da variância definido em 2.1 permanece inalterado.

O problema com o procedimento 1) é que são necessárias replicações do desenho amostral, o que pode ser difícil de obter na prática. Além disso, obter essas k amostras pode levar bastante tempo, e se durante esse período o universo se alterar, ou seja, a quantidade populacional de interesse não for mais a mesma, não podemos considerar as replicações do desenho amostral como pertencentes a mesma distribuição amostral. O procedimento 2) evita esse problema. Nesse caso, vamos supor que o desenho amostral considerado é particionado em H amostras independentes, e que todos as sub-amostras possuem a mesma distribuição amostral. Denotando o estimador do procedimento 2) por p_2, com as respectivas estimativas de cada sub-amostra h dadas por p_h, podemos estimar a variância desse universo utilizando o estimador $\hat{V}(p_2)$, onde:

$$\hat{V}(p_2) \quad = \quad \frac{\sum_{h=1}^{H}(p_h - \overline{p})^2}{H-1}, \tag{2.2}$$

onde $\overline{p} = \frac{1}{H}\sum_{h=1}^{H} p_h$. Ou seja, a forma dos estimadores da variância os dois procedimentos são iguais, o que muda é a origem das estimativas utilizadas nas mesmas.

Além da justificativa empírica apresentada aqui, existem 2 justificativas teóricas, ambas questionáveis no sentido de que supõem que dentro de cada cota a seleção das unidades populacionais é feita com probabilidades iguais. Baseado nas discusões apresentadas nas seções 2.3.1 e 2.5.2, essa suposição só se aproximaria da realidade se o desenho amostral utilizado fosse a **APC**, se as cotas utilizadas fossem realmente homogêneas e o número de moradores pertencentes a cada cota fosse igual em todos os domicílios, como ficará claro no Capítulo 4, onde será apresentada uma teoria geral para amostragem probabilística por cotas considerando esses fatores destacados.

No artigo Deville [1991], o autor obtem a distribuição amostral do estimador supondo que a **AC** é equivalente a AAS, porém com a restrição de que a amostra obtida tem que respeitar os tamanhos das cotas definidas. É como se amostras da **AAS** fossem selecionadas, e todas aquelas em que a quantidade de entrevistas nas diferentes categorias das cotas não coincidirem com o valor pré-especificado para cada cota são descartadas. Outro problema é que não se considera a probabilidade de resposta das unidades populacionais, apesar do autor sugerir uma forma de corrigir o erro de não-resposta, utilizando idéias da pós-estratificação.

Já no artigo Salehi and Chang [2005], sob a suposição de que dentro das categorias das cotas a seleção dos respondentes é equivalente a AAS, os autores obtém a distribuição amostral do estimador utilizando a amostragem inversa para o caso multi-variado, que é uma generalização do que foi apresentado na Seção 1.2.5. Além da suposição da AAS, outro problema com essa formulação é que a quantidade de entrevistas pré-definidas em cada cota não é respeitada, a única garantia é que em todas as categorias haverão pelo menos a quantidade pré-definida, ou seja, por esse método, o tamanho da amostra será maior do que o desejado.

Também existe o interesse em avaliar o tempo que leva para completar as entrevistas ao se utilzar a **AC**. Nesse sentido, os artigos Sobel and Ebneshahrashoob [1992] e Young [1961] apresentam soluções aproximadas, para o caso **AAS** dentro de cada cota, onde o preenchimento das cotas também é visto como uma amostragem inversa multi-variada.

No Capítulo 4 será apresentada uma teoria geral para **APC**, a qual inclue a **AC** como um caso especial. Para obter a distribuição amostral do estimador de interesse, será necessário fazer uma suposição sobre as probabilidades de resposta dentro de cada cota. Além disso, no mesmo capítulo calcula-se o tempo esperado para se completar ambos os procedimentos amostrais. Já no Capítulo 5, a **AP** e a **APC** serão comparadas sob as mesmas suposições, através de um estudo de simulação.

Capítulo 3

Pesquisas Eleitorais e Amostragem na Prática

Nesse capítulo discutiremos as pesquisas eleitorais, tanto do ponto de vista estatístico, no que diz respeito aos diferentes tipos de desenhos amostrais utilizados e suas respectivas críticas, e principalmente se esses desenhos amostrais respeitam a legislação eleitoral existente, quanto do ponto de vista sociológico, discutindo se as pesquisas eleitorais realmente influênciam nos resultados das eleições e porque elas causam tanta polêmica na mídia.

3.1 Controvérsias envolvendo as Pesquisas Eleitorais

Nessa seção será feito um resumo de alguns capítulos de dois livros que discutem, do ponto de vista sociológico, as pesquisas eleitorais. São eles Almeida [2008] e Almeida [2002].

Nesses livros, o autor afirma que existem duas principais razões para que surjam controvérsias acerca dos resultados das pesquisas eleitorais e de sua divulgação na mídia. A primeira é que políticos, jornalistas e empresas de pesquisa tendem a atuar de acordo com lógicas diferentes. A segunda é que há uma suposição básica, ainda não verificada, de que as pesquisas eleitorais podem influenciar o voto do eleitorado, e consequentemente, o resultado das eleições. Essas duas razões serão discutidas, respectivamente, nas seções 3.1.1 e 3.1.3. Outra razão evidente são os erros que já foram cometidos pelas pesquisas eleitorais, alguns dos quais são apresentados em Ferraz [1996] e também na Seção 3.1.2.

3.1.1 Políticos, Jornalistas e Empresas de Pesquisa: Diferentes pontos de vista

Tudo que um político deseja é ganhar eleições, e o que nunca quer que ocorra é perdê-las. Numa campanha eleitoral, o melhor dos mundos para um político é aquele no qual ele possa contar com pesquisas precisas, que lhe dêem a melhor informação tanto do ponto de vista substantivo (provendo informações para a estratégia de campanha) quanto do ponto de vista metodológico (com resultados corretos). Além disso, o melhor é que as pesquisas só forneçam boas notícias. Já o pior dos mundos é aquele no qual as pesquisas eleitorais também sejam de excelente qualidade, mas que só fornecem más notícias, indicando que suas chances de vitória são mínimas.

Ainda que o político deseje pesquisas de excelente qualidade, é duro ouvir "más notícias". Em outras palavras, por mais racional que seja um candidato, o lado emocional da política se manifesta no horror de ouvir "más notícias". Assim, se um candidato estiver mal nas pesquisas, sempre haverá de sua parte uma *disposição favorável* a aceitar pesquisas nas quais apresente melhora; por

outro lado, se o candidato estiver bem posicionado, sempre haverá uma *disposição desfavorável* a aceitar resultados que apresentem um piora em seu desempenho.

Pensando em termos estatísticos, onde erro do tipo 1 é aceitar erradamente a hipótese falsa, e erro do tipo 2 é rejeitar erradamente a hipótese verdadeira, no contexto de pesquisas eleitorais, ocorre um **erro do tipo** 1 **quando alguém afirma que houve uma real alteração na situação eleitoral de uma candidato e isso de fato não ocorreu**, e ocorre um **erro do tipo** 2 **quando alguém afirma que nada se alterou, quando na realidade aconteceu uma mudança significativa**.

No caso de um político e de sua leitura de resultados de pesquisas, ele estará sempre pronto a cometer um erro do tipo 1 quando os índices de intenção de voto se alterarem favoravelmente a ele, e tenderá a cometer um erro do tipo 2 quando as notícias forem ruins, ou seja, quando as pesquisas indicarem que a sua situação eleitoral piorou.

Já o jornalista não é regido pela lógica eleitoral do político. A finalidade básica do trabalho jornalístico é informar a população e o público, de modo que atraia leitores, ouvintes e telespectadores. Há notícias mais e menos atraentes, que cativam mais leitores ou que podem sequer serem lidas ou ouvidas. Quando se aplica essa lógica à divulgação de pesquisas, isso resulta na tendência de noticiar mais mudanças nos índices de intenção de voto nos candidatos e menos a estabilidade do cenário eleitoral.

Ou seja, para um jornalista, em geral a mudança é notícia e a continuidade não é. Aplicada a uma capanha eleitoral e a resultados de pesquisas, isso indica que para os jornalistas, nada é mais frustante do que uma campanha eleitoral em que os índices de intenção de voto dos candidatos não se alteram. Ao comparar pesquisas eleitorais consecutivas, a tendência é que o jornalista destaque o que mudou de uma pesquisa para a outra. Assim, sempre haverá de sua parte uma *disposição favorável* a registrar qualquer mudança, para cima ou para baixo, nos índices de intenção de voto dos candidatos, e sempre haverá uma *disposição desfavorável* a aceitar a inexistência de mudanças. Ou seja, um jornalista estará propenso a cometer um erro do tipo 1, em quaisquer circuntâncias, independentemente do crescimento ou piora dos índices deste ou daquele candidato, e muito dificilmente ele cometerá um erro do tipo 2.

O terceiro personagem do mundo das pesquisas eleitorais são as empresas de pesquisa. Elas não têm o interesse de ganhar a eleição nem estão preocupadas se o resultado da pesquisa será ou não matéria-prima para as notícias. É claro que para fins de divulgação da empresa de pesquisa, o melhor é que toda pesquisa vire notícia, mas isso, do seu ponto de vista, depende exclusivamente da dinâmica da campanha eleitoral. O mais importante é que os resultados das pesquisas realizadas por elas sejam o mais correto e preciso possível, pois a sua reputação está em jogo. Assim, sempre haverá uma *disposição favorável* dos institutos de pesquisa a aceitar a inexistência de mudanças, e uma *disposição desfavorável* a registrar qualquer mudança, para cima ou para baixo, dos índices de intenção de voto dos candidatos, assim as empresas de pesquisa têm um predisposição muito maior de cometer um erro do tipo 2, e dificilmente cometeriam um erro do tipo 1. Ou seja, para elas é melhor não fazer uma descoberta do que aceitar uma hipótese falsa.

Ou seja, de forma resumida, temos que:

Políticos *Boas Notícias:* tendência a aceitar uma mudança quando na realidade ela não ocorreu.
Más Notícias: tendência a rejeitar uma mudança quando na realidade ela ocorreu.

Jornalistas Tendência a aceitar uma mudança quando na realidade ela não ocorreu.

Empresas de Pesquisa Tendência a rejeitar uma mudança quando na realidade ela ocorreu.

Quando se comparam políticos, jornalistas e empresas de pesquisa, as lógicas utilizadas para interpretar os resultados das pesquisas são bastante diferentes, para não dizer opostas. O resultado disso é que durante uma campanha eleitoral, e depois do seu término, os protagonistas das polêmicas acerca das pesquisas e de sua divulgação têm preocupações diferentes, o que acaba por gerar muita polêmica em torno das pesquisas eleitorais, principalmente sobre a credibilidade das mesmas. Muitos exemplos dessa controvérsia, principalmente do ponto de vista dos políticos, são apresentados em Ferraz [1996]. É interessante notar que **mesmo que as pesquisas eleitorais nunca errassem, elas seriam criticadas da mesma forma por causa das diferentes lógicas dos protagonistas em questão**.

Para piorar a situação, sabe-se que, do ponto de vista de inferência baseada no desenho, as pesquisas eleitorais vão errar um determinado percentual das previsões. Na Seção 3.1.2 serão apresentados alguns exemplos de pesquisas que erraram a previsão dos resultados da eleicao, e que por causa disso geraram polêmica na mídia.

Além disso, diferentes critérios de erro podem ser considerados, como por exemplo, errar o vencedor da eleição, errar a ordem dos candidatos ou se a pesquisa estava fora da margem de erro, permitindo que algumas pessoas considerem que uma mesma pesquisa errou a previsão dos resultados de uma eleicao quando outras consideram que ela acertou. Uma avaliação detalhada de aproximadamente 900 pesquisas eleitorais realizadas entre os anos de 1989 e 2004, considerando diversos critérios de erro, será apresentada na Seção 5.2.

3.1.2 Erros das Pesquisas Eleitorais

Fazendo um busca rápida na internet, é possível enontrar muitos exemplos de pesquisas eleitorais que não conseguiram prever corretamente o resultado das eleições. Alguns exemplos interessantes são:

> **Notícia 1: Eleições em Contagem - como erram as pesquisas!**
>
> Na última sexta-feira, 03/10, foi publicado no jornal SUPERNOTÍCIA pesquisa do instituto Sensus Pesquisa e Consultoria, cujas projeções indicavam uma vantagem de 6 pontos a favor do candidato Ademir Lucas-PSDB que teria 34,4% das intenções de voto sobre a Petista Marilia Campos que estaria com 28,4%. Dentre os argumentos do texto está à vaticinação de que Ademir Lucas estaria eleito com aquele resultado. O instituto Sensus só esqueceu do detalhe de que o município de Contagem conta com mais de 200.000 eleitores e, para que isto ocorresse, o candidato Ademir Lucas precisaria

de mais de 50% dos votos válidos. O resultado das urnas em primeiro turno mostrou outra coisa, Marília Campos ficou em 1º. Lugar com 43,87% e Ademir Lucas em 2º. com 37,39% dos votos válidos e a eleição será definida só no segundo turno de votações.

Reafirmo aqui que pesquisas eleitorais e textos cuidadosamente dissimulados tem sido utilizados para ludibriar o eleitor e como forma de alavancar candidaturas. Isto é um risco para a lisura dos pleitos democráticos.

site: linguadetrapo.blogspot.com/2008/10/eleies-em-contagem-como-erram-as.html

Notícia 2: As pesquisas baianas

Não há explicação plausível para o erro do IBOPE nas pesquisas da Bahia. O Instituto acertou na mosca uma pesquisa preparada para todo o país. E na Bahia errou por vinte pontos percentuais, um erro humanamente impossível. Jacques Wagner saltar de 34% na última pesquisa para 54% na boca de urna não é tecnicamente compreensível. Ainda mais sabendo-se que nas eleições anteriores, entre a última pesquisa e a boca-de-urna as intenções de voto para ele saltaram de 18% para 38,5%. Uma explicação do Roberto, nos comentários, seria o "voto atemorizado"do eleitor baiano nas pesquisas.

site: colunistas.ig.com.br/luisnassif/2006/10/03/as-pesquisas-baianas/

Notícia 3: Erro pode levar a CPI - Marconi Perillo quer que o Congresso investigue os erros nas pesquisas de intenção de voto

O senador eleito Marconi Perillo (PDSDB) pretende propor, assim que tomar posse, uma CPI para investigar as pesquisas. Isso porque, mais uma vez, os institutos erraram em Goiás. Ibope, Brasmark e Fortiori divulgaram no sábado, 30 de setembro, um dia antes do primeiro turno da eleição, pesquisas totalmente divergentes dos resultados das urnas. O cálculo que os governistas fazem é que, se as pesquisas tivessem apontado para a possibilidade de vitória de Alcides Rodrigues (PP) no primeiro turno, ele poderia realmente ter vencido a eleição, uma tese que tem fundamento. Segundo Mário Rodrigues, do instituto de pesquisa Grupom, 2 por cento dos eleitores tendem a votar no candidato que apresenta maiores chances de vitória. A diferença de Alcides Rodrigues e Maguito Vilela (PMDB) foi de menos de 2 por cento dos votos.

Todavia, errar pesquisa não é crime. Há crime quando o instituto frauda a pesquisa, o que é muito difícil comprovar. Mas, para atender os interesses de seus clientes, os institutos não precisam fraudar o levantamento e, conseqüentemente, incorrer em crime. Basta trabalhar dentro da margem de erro. Foi o que o Fortiori fez. O instituto efetivamente não errou, apesar de ter apontado que Alcides Rodrigues teria 36,2 por cento dos votos e Maguito Vilela, 41,7 por cento. O resultado da eleição foi previsto pelo instituto dentro da margem de erro e do número de eleitores indecisos.

Pesquisa é estatística e, de acordo com uma lei dessa ciência, de cada 100 pesquisas realizadas com metodologia e amostragem semelhantes, 95 por cento tenderão a ter o

mesmo resultado e 5 por cento podem ter resultados totalmente divergentes. Ou seja, observa Mário Rodrigues, se a pesquisa obedece aos critérios, a tendência é ela não divergir. "Pode haver pequenos senões, como uma amostragem errada, e, em pesquisa, qualquer erro influencia o resultado."Se o questionário, por exemplo, não segue o modelo preconizado, isto pode criar um viés de resposta e apontar uma tendência diversa. Se o entrevistador não segue rigorosamente o mapa amostral, ou seja, não entrevista as pessoas definidas, isto também pode influenciar no resultado do levantamento.

O que os institutos às vezes argumentam é que a mudança na intenção de voto se dá de uma forma tão rápida que a pesquisa não consegue detectar. Seria o caso do eleitor de Maguito Vilela ter migrado para Alcides Rodrigues nos últimos dois dias que antecederam a votação. Entretanto, no dia 25 de setembro o Grupom divulgou uma pesquisa que já apontava a virada de Alcides Rodrigues. Naquela data, Maguito Vilela tinha 39,7 por cento dos votos e Alcides Rodrigues já estava com 38,9 por cento. "Nos últimos três dias que antecedem a eleição é mais difícil detectar qualquer alteração no quadro, mas por isso é importante observar o mapa das tendências."O Grupom não errou o resultado da eleição.

Mário Rodrigues, no entanto, isenta os institutos de responsabilidade sobre o resultado da eleição. "As pesquisas não influenciaram, senão Maguito Vilela não teria perdido a eleição."Na sua opinião, pesquisa errada prejudica o instituto, que perde a credibilidade. No caso do Ibope, por exemplo, as pesquisas eleitorais respondem por apenas 8 por cento de seu faturamento. "O próprio mercado se incube de expurgar os ruins", afirma Mário Rodrigues.

Importância excessiva - Na opinião do deputado federal Vilmar Rocha, o grande problema das pesquisas de intenção de voto está no fato de se dar a elas mais importância do que elas merecem. "Já é ponto pacífico que as pesquisas refletem um momento específico e que aquele quadro pode mudar no momento seguinte", afirma. Segundo o deputado, levantamentos dessa natureza são elementos de avaliação, instrumentos importantes de trabalho, mas não podem ser vistos como definitivos.

Além disso, observa Vilmar Rocha, as pesquisas permitem algum grau de manipulação, o que deve influir no nível de credibilidade que o eleitor dá aos levantamentos. Porém, observa, não é caso para a abertura de uma CPI no Senado. "Trata-se de um instrumento privado e creio que, com o amadurecimento da democracia e do processo político, a tendência é a importância das pesquisas cair muito."Além disso, o deputado não acredita que as pesquisas influenciem o voto de uma forma essencial. "Apenas de maneira marginal, como um termômetro que mede a febre mas não a faz subir nem descer."

O advogado e professor de direito da UFG e da UCG Pedro Sérgio dos Santos chama atenção para a dificuldade de se provar que os institutos fraudaram determinado levan-

tamento: "No direito, alegar não é provar". Ou seja, pode-se alegar que os institutos manipularam os dados, mas é muito difícil provar que eles efetivamente manipularam as informações. Sendo assim, a CPI não teria efeito jurídico. "Pode funcionar como pressão política", observa o advogado. Levar o instituto ao descrédito. "Considerando que estes institutos não fazem apenas pesquisas eleitorais, mas também comerciais, a desmoralização, nesse caso, teria o efeito esperado."

site: jornalopcao.com.br/index.asp?secao=Reportagens&idjornal=210& idrep=2079

Notícia 4: Pesquisas: institutos erram e trocam acusações

Os dois principais institutos de pesquisas do país, Ibope e Datafolha, não acertaram os resultados das urnas em algumas cidades. Os casos mais notórios e distantes da média de 2 a 3 pontos percentuais de margem de erro ocorreram no Rio, em São Paulo e Belo Horizonte, segundo analistas políticos. Para o cientista Marco Antonio Villa, as três capitais se tornaram um verdadeiro "Triângulo das Bermudas"na política nacional por causa das pesquisas. Ele avalia que os eleitores podem ter alterado suas opções porque receberam informações erradas, principalmente no Rio. - As pesquisas são fontes importantes para a democracia, são parte do processo eleitoral. Podemos chegar a 2010 em uma situação de descrédito sobre as pesquisas de opinião, o que é muito negativo - disse Villa. Polarização leva eleitor a voto útil O caso mais grave foi registrado no Rio. Quando o Datafolha já mostrava o crescimento de Fernando Gabeira (PV), o Ibope continuou informando o eleitor que a polarização se dava entre Eduardo Paes (PMDB) e Crivella (PRB). Gabeira chegou a acusar o Ibope de estar a serviço do PMDB. - O que aconteceu agora, nas urnas, mostra que o candidato tinha alguma razão. Há algo de errado na metodologia do Ibope. Quando se aponta uma polarização, o eleitor tende a exercer o voto útil - disse Villa. "Parece que em Minas os caciques se acertaram, mas se esqueceram de combinar com os eleitores "

Já em São Paulo, os dois institutos colocaram Marta Suplicy (PT) na liderança, e as urnas deram a primeira colocação para o prefeito Gilberto Kassab (DEM). O Ibope chegou a confirmar a liderança de Marta em pesquisa de boca-de-urna (feita no dia da eleição). Em Belo Horizonte, Márcio Lacerda (PSB) aparecia muito à frente de Leonardo Quintão (PMDB), nos dois institutos, mas, nas urnas, a diferença ficou em apenas dois pontos. - Parece que em Minas os caciques se acertaram, mas se esqueceram de combinar com os eleitores. E os institutos não captaram os eleitores. Creio que os institutos terão que rever suas metodologias - disse Villa, lembrando que ocorreram diferenças acentuadas também em Porto Alegre e Salvador.

Em entrevistas ao Globo, os diretores dos dois institutos negaram erros e trocaram acusações, principalmente por causa do Rio. O Datafolha avalia que foi o único a detectar o crescimento de Gabeira na reta final. - Se não houvesse o Datafolha, a surpresa no Rio seria muito maior. E também seria outro resultado, porque eleitores

> da Jandira migraram para Gabeira. É claro que a pesquisa influenciou e isso é bom, é informação. No Rio, além de definir o candidato, o eleitor escolheu em qual instituto de pesquisa deve confiar - disse o diretor do Datafolha, Mauro Paulino.
>
> **Site:** oglobo.globo.com/pais/eleicoes2008/mat/2008/10/06/pesquisas_institutos _erram_ trocam_acusacoes-548586940.asp

É interessante que os institutos de pesquisa sempre têm uma justificativa não-amostral para explicar porque uma pesquisa eleitoral errou a previsão. Para o leitor interessado, o livro Almeida [2009] apresenta exemplos e discute os erros eleitorais em pesquisas brasileiras, além de considerar uma justificativa polêmica para a maioria dos erros detectados pelo autor: quanto menor a escolaridade no local sendo pesquisado, maior o erro cometido pelas pesquisas. Segundo o autor, isso ocorre por dois motivos principais: pessoas de escolaridade mais baixa declaram falsamente que votarão no candidato que está em primeiro lugar nas pesquisas e/ou ao votar, essas pessoas erram o voto com mais frequência, acidentalmente votando branco e/ou nulo. Na análise realizada na Seção 5.2, não encontramos evidência empírica de tal fato.

3.1.3 Influência das Pesquisas Eleitorais no resultado da eleição

Boa parte da preocupação com os resultados das pesquisas eleitorais e com a forma de divulgá-las deriva da crença de que a divulgação dos resultados das pesquisas influência o eleitor. Se a crença fosse oposta, os políticos não se queixariam dos jornalistas nem processariam os institutos de pesquisa, os jornalistas se preocupariam menos em estampar manchetes com resultados de pesquisas e as empresas de pesquisa teriam menos espaço na imprensa.

Dois tipos de influência das pesquisas eleitorais sobre o resultado das eleições usualmente são consideradas. A *influência direta* sobre o eleitor, que ao ter conhecimento dos resultados da pesquisa, decide votar nos candidatos favoritos, e a *influência indireta*, onde os resultados das pesquisas eleitorais exercem forte impacto sobre o ânimo e o moral das campanhas eleitorais (na capacidade de arrecadar recursos financeiros para a campanha) e sobre a cobertura da mídia, influênciando na eficiência das campanhas e consequentemente na disposição do eleitor votar neste ou naquele candidato.

Nessa seção, serão discutidos quais fatores as pessoas consideram para decidir em qual candidato votar, possibilitando uma discussão mais clara sobre o impacto das pesquisas eleitorais nos resultados das eleições. Também serão discutidas algumas pesquisas realizadas com o objetivo de avaliar empiricamente essa hipótese.

A literatura especializada explica de inúmeras maneiras a decisão de votar em diferentes candidatos ou partidos. Os fatores explicativos usualmente considerados podem ser separados em dois grupos: fatores de caráter estrutural e conjuntural. As primeiras se alteram de forma lenta, ao longo de anos ou até décadas, e funcionam como fatores que estruturam as escolhas. Já as variáveis conjunturais dizem respeito a cada campanha eleitoral em particular. As variáveis estruturais usualmente consideradas são:

Predisposições individuais Devido a fatores sociais e psicológicos, cada indivíduo tem predisposições e valores diferentes. Pessoas que julgam que o mérito e o empenho individuais são o fator-chave para a melhoria de vida tendem a escolher candidatos liberais do ponto de vista econômico, já as que acham que a melhoria do bem-estar é um problema mais coletivo do que individual tendem a escolher candidatos que defendem a intervenção do Estado na economia.

Preferência partidária A preferência partidária varia muito dependendo da pessoa, porém ela não é determinada apena pelas predisposições individuais. Há aqueles que votam somente em candidatos de um determinado partido, e há aqueles que não votam em candidatos de um determinado partido.

Nível de informação política Quanto mais informado é o eleitor sobre o que acontece na política e sobre por que as coisas acontecem do jeito que acontecem, mais ele tende a exigir de seus representantes e mais crítico tende a ser em relação ao governo. O oposto também é verdade.

Representatividade Social É comum que escolhas eleitorais sejam determinadas pelo grupo social: candidatos originários de grupos sociais específicos acabão obtendo votos desses grupos.

Status Socioeconômico Diferentes grupos sociais apresentam comportamentos eleitorais diferentes. O exemplo mais corriqueiro é o das campanhas eleitorais caracterizadas pela disputa entre o "candidato dos pobres"e o "candidato dos ricos".

Já as variáveis conjunturais usualmente consideradas são:

Avaliação do desempenho do governo O desempenho do governo é um fator fundamental. Eleitores que têm uma boa avaliação do governo tendem a votar no candidato que representa a continuidade, já a má avaliação resulta em voto na oposição.

Propostas dos candidatos O conjunto de propostas de um candidato em uma particular eleição é relevante para definir os votos, em que pese as diferenças ideológicas entre liberais e conservadores e entre esquerda e direita.

Redes Sociais Existem evidências empíricas de que algunas eleitores influenciam e outros são influenciados por pessoas próximas.

Imagem dos candidatos Numa campanha política, grande parte da propaganda eleitoral divide-se entre apresentação de propostas e a tentativa de transmitir ao eleitor mensagens que virão a compor a imagem do candidato.

Para avaliar o impacto que as pesquisa eleitorais podem ter no resultado das eleições, é interessante imaginar o processo pelo qual um eleitor médio decide em quem votar. Não é necessário abordar o papel dos fatores que regem a decisão de voto do eleitor, mas apenas avaliar o peso relativo das variáveis estruturais com relação as variáveis conjunturais. Quanto maior for a importância dos fatores estruturais, menos importância terão os resultados da pesquisa eleitoral.

É razoável supor que um eleitor será submetido a um grande volume de informações sobre o Poder Executivo, sobre críticas de opositores, denúncias e escândalos, e formará uma imagem de cada candidato que disputa a eleição. Algumas informações serão novas, outras antigas, e elas serão filtradas e lidas de acordo com as preferências, visões do mundo e simpatias do eleitor. Uma das informações que poderá chegar (ou não) a esse eleitor é a dos resultados das pesquisas. Note que essa será apenas uma das inúmeras informações a serem recebidas e filtradas.

Analisando todos os pressupostos necessários para que um eleitor obtenha as informações das pesquisas eleitorais e utilize-a para alterar o seu voto, não é óbvio que as pesquisas influenciem o voto, é algo que deve ser provado.

Algumas pesquisas, experimentais e observacionais, foram realizadas, principalmente nos Estados Unidos e na Grã-Bretanha, com o objetivo de mensurar se há ou não influência das pesquisas sobre o comportamento eleitoral. Existem muitas dificuldades metodológicas e de recursos para realizar um pesquisa desse tipo. Por exemplo, quando estudos experimentais são realizados eles são apenas uma simulação de uma eleição real, e dessa forma é difícil dizer se os seus resultados podem ser generalizados para uma eleição real. Além disso, ao realizar um estudo desse tipo existe o risco bastante real de se super-dimensionar a relevância das informações obtidas das pesquisas eleitorais pelos eleitores.

Os estudos realizados, descritos em Almeida [2008], não são conclusivos com relação a influência direta na escolha do eleitor. Há indícios de que existe influência, mas esses indícios são obtidos a partir de pesquisas inadequadas. Além disso, não se sabe se o efeito líquido entre a mudança de voto dos eleitores que querem votar no candidato que está na frente, porque é o que vai ganhar, e os eleitores que decidem votar no candidato que está atrás, para aumentar as suas chances de ganhar. Ou seja, as pesquisas podem influenciar nessas duas direções, e em todas as pesquisas em que houveram indícios de que isso aconteceu, foi impossível estimar o efeito.

Com relação a influência indireta das pesquisas no resultado de eleições não parece haver dúvidas de que ela existe, pois os principais candidatos apontados pelas pesquisas têm mais espaço na mídia, conseguem mais recursos, e animam mais facilmente seus partidários. Em geral, eles também são candidatos dos principais partidos políticos, assim provavelmente já teriam as vantagens de qualquer forma, o que também torna muito difícil avaliar o efeito da influência indireta.

3.2 Amostragem e Pesquisas Eleitorais

Como vimos nos Capítulos 1 e 2, existem diversos desenhos amostrais diferentes, e também existem diferentes formas de fazer inferência a partir dos dados de uma particular amostra. Nessa seção discutiremos a legislação das pesquisas eleitorais, ou seja, legalmente o que é necessário, discutiremos os diferentes desenhos amostrais e as diferentes metodologias de coleta de dados usualmente utilizadas pelo institutos de pesquisa.

3.2.1 Legislação das Pesquisas Eleitorais

A resolução número 23.190[1] do Tribunal Superior Eleitoral (TSE), dispõe sobre a legislação eleitoral para o ano de 2010, para pesquisas que serão divulgadas. Dois artigos determinam quais informações referentes as pesquisas eleitorais devem ser divulgadas: no Artigo 1°, são descritas as informações relativas ao registro das pesquisas no TSE, e no artigo 10°, são descritas as informações relativas a publicação das pesquisas. Reproduziremos apenas parte da resolução aqui, porém a íntegra dotexto está no apêndice A:

> **Artigo 1º - (DO REGISTRO)** A partir de 1° de janeiro de 2010, as entidades e empresas que realizarem pesquisas de opinião pública relativas às eleições ou aos candidatos, para conhecimento público, são obrigadas, para cada pesquisa, a registrar no tribunal eleitoral ao qual compete fazer o registro dos candidatos, com no mínimo 5 dias de antecedência da divulgação, as seguintes informações:
>
> **I** quem contratou a pesquisa.
>
> **II** valor e origem dos recursos despendidos no trabalho.
>
> **III metodologia e período de realização da pesquisa.**
>
> **IV plano amostral e ponderação quanto a sexo, idade, grau de instrução e nível econômico do entrevistado; área física de realização do trabalho, intervalo de confiança e margem de erro.**
>
> **V** sistema interno de controle e verificação, conferência e fiscalização da coleta de dados e do trabalho de campo.
>
> **VI** questionário completo aplicado ou a ser aplicado.
>
> **VII** nome de quem pagou pela realização do trabalho.
>
> **VIII** contrato social, estatuto social ou inscrição como empresário, que comprove o regular registro da empresa, com a qualificação completa dos responsáveis legais, razão social ou denominação, número de inscrição no Cadastro Nacional da Pessoa Jurídica (CNPJ), endereço, número de fac-símile em que receberão notificações e comunicados da Justiça Eleitoral.
>
> **IX** nome do estatístico responsável pela pesquisa - e o número de seu registro no competente Conselho Regional de Estatística -, que assinará o plano amostral de que trata o inciso IV retro e rubricará todas as folhas.
>
> **X** número do registro da empresa responsável pela pesquisa no Conselho Regional de Estatística (CONRE), caso o tenham.
>
> **Artigo 10º - (DA DIVULGAÇÃO)** Na divulgação dos resultados de pesquisas, atuais ou não, serão obrigatoriamente informados:
>
> **I** o período de realização da coleta de dados
>
> **II a margem de erro**

[1] www.tse.gov.br/internet/eleicoes/2010/Pesquisas_eleitorais.html

III o número de entrevistas

IV o nome da entidade ou empresa que a realizou, e, se for o caso, de quem a contratou

V o número do processo de registro da pesquisa.

Note que a legislação vigente não impõe nenhum tipo metodologia, tamanho de amostra mínimo, confiança ou outro critério de qualidade. Ou seja, não existe nenhum tipo de recomendação ou imposição explícita sobre qual tipo de inferência ou de desenho amostral deve ser utilizado, apenas define-se quais informações a respeito das pesquisas devem ser de domínio público. Dessas informações, foram destacadas acima, em negrito, aquelas que se referem ao desenho amostral.

Implicitamente, ao exigir a divulgação do intervalo de confiança e da margem de erro, parece haver a suposição de que será utilizada inferência baseada no desenho (**ID**). Porém é estranho pedir a divulgação de ambos, pois são informações redundantes. Levando o texto da legislação ao pé da letra, como não há a necessidade de informar o α utilizado para obter as margens de erro, é evidente que sempre haverá um α segundo o qual as margens de erro divulgadas serão respeitadas se a amostragem fosse repetida indefinidamente, mesmo no caso da amostragem por cotas, pois apesar das probabilidades de seleção serem desconhecidas, elas existem. Ou seja, de acordo com a legislação vigente, nunca se poderá afirmar que um desenho amostral específico não respeita a legislação vigente.

Talvez o relator quissese dizer apenas confiança, não intervalo de confiança, mas se esse for o caso, o que causa mais espanto é que não é necessário divulgar na mídia, juntamente com os resultados da pesquisa, a confiança das margens de erro que estão sendo divulgadas, ou seja, é impossível alguém saber a precisão das estimativas que estão sendo divulgadas, e consequentemente utilizar aquela informação de maneira racional para decidir em qual candidato votará, conforme discutido em 3.1.3.

Como não se faz menção explícita a **ID** e muito menos a distribuição de referência que deve ser utilizada para se fazer inferência, nada impede que outros tipos de inferência sejam utilizadas para calcular as margens de erro, como por exemplo utilizar intervalos de credibilidade no caso de inferência Bayesiana baseada no Modelo (**IBM**).

Um detalhe importante sobre a legislação vigente é que em nenhum momento se faz alguma menção a erros não-amostrais e a não-resposta, nem exigi-se que a taxa de resposta ou medida similar seja divulgada.

3.2.2 Qualidade das Pesquisas Eleitorais

As pesquisas eleitorais usualmente têm que ser realizadas com muita rapidez, afinal pesquisas com a intenção de voto do eleitorado há algumas semanas atrás dificilmente seriam divulgadas. Esse é um dos principais motivos pelo qual usualmente o desenho amostral utilizado é alguma variação da amostragem por cotas ou amostragem probabilística com cotas, os quais permitem que a coleta dos dados seja realizada com maior rapidez.

Por causa da rapidez com que as pesquisas eleitorais são realizadas, algumas características do desenho amostral e da coleta de dados são essênciais para determinar a qualidade das mesmas.

Especificamente, além do desenho amostral o qual já foi discutido nos Capítulos 1 e 2, algumas características da coleta de dados são fundamentais para avaliar a qualidade das pesquisas eleitorais: Tipo da entrevista, Local da Entrevista e Controle de Campo.

As entrevistas podem ser realizadas pessoalmente, por telefone, pela internet ou por correio. É difícil avaliar qual o impacto que cada tipo de entrevista terá nos resultados, mas pesquisas por telefone e pela internet notoriamente têm problemas de cobertura, pois pessoas que não tem acesso a esses meios de comunicação são excluídas da amostra. Além disso, outro problema tipicamente associado com pesquisas pela internet e pelo correio, porém que ocorre com todos os tipos de pesquisa, é a auto-seleção das pessoas, ou seja, só participa da pesquisa quem quiser. Esse problema aparentemente é pior nesses dois tipos de pesquisa pois não há acesso direto a pessoa selecionada, tornando mais fácil uma recusa. Novamente no caso de pesquisas pela internet e pelo correio, outro problema em potencial é que o questionário é preenchido pelo próprio respondente, sem o intermédio do entrevistador, o que pode diminuir a qualidade das respostas se o respondente não compreender alguma pergunta ou se não seguir as instruções de preenchimento corretamente.

Usualmente, as pesquisas eleitorais são realizadas em pontos de fluxo ou no domicílio dos respondentes. As pesquisas em ponto de fluxo são realizadas em locais de grande convergência de pessoas em áreas urbanas, como por exemplo praças centrais, terminais de transporte urbano e áreas comerciais, elas são geralmente mais rápidas que entrevistas domiciliares, porém o controle sobre o respondente e sobre a amostra de uma forma geral é muito menor, e o entrevistador tem muito mais liberdade para selecionar o respondente, provavelmente segundo algum critério de conveniência. Para que pesquisas desse tipo sejam justificadas com a teoria apresentada no Capítulo 4, as suposições necessárias são muito mais fortes do que as necessárias para pesquisas domiciliares, como será discutido na Seção 4.2.2.

Com relação ao controle de campo, muitas questões são importantes para determinar a qualidade da pesquisa: treinamento adequado dos entrevistadores; instruções de campo para cada pesquisa; técnicas de abordagem para que o entrevistador não influencie/espante o respondente; fiscalização de entrevistas realizadas, para combater fraudes e verificar as informações mensuradas e críticas aos questionários, para avaliar as incoerências dos mesmos e corrigí-las quando possível.

Ou seja, pesquisas utilizando o mesmo desenho amostral, porém utilizando diferentes critérios para a coleta de dados podem ter qualidades muito diferentes. Sempre deve se levar em conta como foi o processo de coleta de dados para avaliar a qualidade de uma pesquisa, e não somente o desenho amostral.

3.3 Críticas Metodológicas as Pesquisas Eleitorais

Nesta sessão, o interesse está em especificar quais são as principais críticas que as pesquisas eleitorais e os institutos de pesquisa recebem, para que possamos discutir cada uma delas com um pouco mais de detalhe. Elas são:

1 - Seleção Não-Probabilística da Amostra A metodologia de seleção da amostra, usualmente utilizando-se cotas quando a **ID** supõe seleção probabilística.

2 - Inferência baseada na AAS Ao analisar os resultados da amostra, não se leva em conta o desenho amostral, supondo-se que foi realizada uma **AAS**.

3 - Desconsiderar a Correlação entre Candidatos Ao analisar os resultados da amostra, não leva-se em conta a correlação entre categorias da multinomial, ou seja, desconsidera-se que em média, quanto maior o percentual de votos de um candidato, menor o percentual de votos nos outros candidatos.

4 - Empate Técnico entre Candidatos A ocorrência de empates técnicos entre candidatos, ou seja, depois de observada a amostra, não ser possível inferir qual candidato está na frente.

Dessas críticas, apenas a primeira se refere a amostra em si, as outras estão relacionadas a forma como se faz inferência. Quanto a seleção da amostra usualmente não ser proabilística, já vimos na Seção 1.4 que existem outros tipos de inferência que podem ser utilizados para se fazer inferência, os quais não necessitam de uma amostra probabilística. Nesse caso, usualmente uma crítica permanece: nesses outros tipos de inferências é necessário fazer suposições. Porém, no caso da amostragem probabilística combinada com a inferência baseada no desenho (**ID**) também é necessário fazer as seguintes susposições: 1- que o teorema central do limite é válido, conforme discutido em 1.2.2; 2- que o erro de não-resposta seja ignorável, conforme discutido em 1.3; 3- no caso de amostragem complexa, é necessário utilizar aproximações para estimar a variância do estimador, conforme discutido em Pessoa and Silva [1998] e em Wolter [1985]. Pensando somente na amostragem e não no tipo de inferência que será realizada, existem argumentos para que a amostra seja probabilística, conforme discutido na Seção 1.4.3.

Usualmente a crítica de que a inferência é baseada na **AAS** é feita quando utiliza-se a amostragem por cotas (**AC**), porém essa simplificação com certeza também é feita pelos institutos de pesquisa quando é utilizada a amostragem probabilística (**APV**) no caso onde a amostra não é auto-ponderada, e mesmo nesses casos as variâncias são usualmente estimadas supondo-se que a amostra é um **AAS**. No caso da **APC**, essa suposição é feita, pelo menos no último estágio de seleção, por falta de opção, pois os insitutos de pesquisa utilizam geralmente a **ID**. No Capítulo 4 é apresentada uma solução, no contexto de **ID**, para estimar a variância do estimador sob uma suposição muito mais realista e flexível do que supor que a amostra é uma **AAS**.

Note que essa é uma crítica importante, talvez a mais importante de todas as críticas feitas às pesquisas eleitorais, quando as probabilidades de inclusão/seleção são desconsideradas (mesmo que conhecidas) para se fazer inferência. Quando utiliza-se o estimador simples da **AAS** ao invês dos estimadores de HH ou de HT, isso não quer dizer que as inferências feitas estão erradas. Apenas quer dizer que os estimadores utilizados são viciados, porém o EQM do mesmo pode ser menor do que os estimadores de HH ou de HT, conforme será discutido na Seção 5.1.

Já a crítica feita aos institutos de pesquisa por eles desconsiderarem a correlação entre candidatos é relevante. Inclusive, vimos um exemplo real na Seção 1.2.2 onde considerar essa correlação permitiu que inferências corretas fossem feitas numa situação real. Duas possíveis explicações existem para que os mesmos não considerem a correlação ao fazer inferências: 1) deficiência técnica e/ou

desconhecimento da teoria ou 2) no contexto discutido na Seção 3.1.1, é uma forma das empresas de pesquisa serem conservadoras, evitando divulgar uma diferença.

As críticas sobre as pesquisas eleitorais discutidas até agora são usualmente feitas por defensores da **ID**. Já a crítica a respeito do empate técnico é diferente das outras, pois ela é feita por defensores da inferência Bayesiana baseada em Modelos (**IBM**) quando se faz inferências do tipo **ID** ou **IM**, ou seja, essa crítica na verdade não é especificamente para os institutos de pesquisa, mas sim ao tipo de inferência utilizada por eles. Do ponto de vista dos institutos, é mais uma forma de evitar a divulgação de uma diferença, conforme discutido na Seção 3.1.1.

Quando ocorre o empate técnico, do ponto de vista da **ID** e da **IM**, não é possível afirmar nada sobre qual candidato está na frente, já do ponto de vista da **IBM**, a probabilidade de cada candidato ganhar sempre é conhecida. Mais que isso, pelo menos para o caso com 2 candidatos, e talvez para o caso geral, em Zabala [2009] o autor provou que quando ocorre a iminência do empate técnico[2], que a probabilidade à posteriori de vitória do candidato com mais votos converge, conforme $n \to \infty$, para $(1-\alpha)$, onde α é a confiança do intervalo de confiança, independente da priori utilizada. Ou seja, utilizando **IBM** sempre é possível saber qual a chance de cada candidato ganhar, essa possibilidade sendo claramente mais interessante do que a existência de empates técnicos.

[2] A iminência do empate técnico entre os candidatos A e B ocorre quando os extremos dos intervalos de confiança para as quantidades P_A e $P_B = 1 - P_A$, dados respectivamente por $(\hat{P_A} - \xi; \hat{P_A} + \xi)$ e $(\hat{P_B} - \xi; \hat{P_B} + \xi)$ com $\xi = z_{\frac{\alpha}{2}} \sqrt{\frac{P_A P_B}{n}}$, se intersectam, ou seja, quando $P_A - \xi = P_B + \xi$ ou $P_A + \xi = P_B - \xi$.

Capítulo 4

Amostragem considerando a Probabilidade de Resposta

Quando a unidade populacional é o ser humano, o fato de uma pessoa ter sido selecionada para participar de uma pesquisa não garante sua participação na mesma, pois ela pode se recusar a responder ou não ser encontrada pelos entrevistadores. Nesse capítulo será discutida a amostragem quando a probabilidade de resposta, que é a probabilidade de uma unidade populacional responder dado que ela foi selecionada, é levada explicitamente em consideração.

Nesse contexto, serão calculadas as probabilidades de seleção tanto para o caso da amostragem probabilística com cotas quanto para o caso da amostragem probabilística com voltas equivalente.

Esse capítulo será desenvolvido no contexo de amostragem probabilística com reposição, e o parâmetro de interesse é o total populacional da variável Y, denotado como $\tau_y = \sum_{i=1}^{N} Y_i$, onde N é o número de unidades na população e Y_i é o valor da variável de interesse Y para a unidade populacional i.

4.1 Probabilidade de Resposta

Na teoria usual de amostragem, onde utiliza-se inferência baseada no desenho (**ID**), a seleção da amostra deve ser probabilística pois a distribuição de referência utilizada para fazer inferência é gerada pela própria seleção da amostra, ou seja, depende das probabilidades de seleção das unidades populacionais. No caso da amostragem com reposição, isso quer dizer que todas as pessoas da população de interesse devem ter uma probabilidade de seleção p_i^{Selec} positiva e conhecida (ou pelo menos calculável) para que as informações observadas na amostra possam ser inferidas para toda população, onde p_i^{Selec} deve ser interpretada como a probabilidade da unidade populacional i ser selecionada para pertencer a amostra em um único sorteio. Essa probabilidade está sob controle do estatístico responsável pela seleção da amostra.

Essas probabilidades são necessárias para estimar o parâmetro populacional de interesse utilizando o estimador $\tau_{\hat{HH}}$, definido em 1.2.6, o qual é usualmente recomendado para amostragem com reposição. Nessa seção, utilizaremos uma notação um pouco diferente, re-escrevendo o estimador $\tau_{\hat{HH}}$, definido em 1.2.6, como:

$$\tau_{\hat{HH}} = \sum_{i \epsilon s} \frac{Y_i}{n p_i^{Selec}}, \tag{4.1}$$

onde s representa o conjunto dos índices pertencentes a amostra. Nessas condições, vimos na Seção 1.2.6, que:

$$E\left(\tau_{\hat{HH}}\right) = \sum_{i=1}^{N} Y_i \qquad e \qquad Var\left(\tau_{\hat{HH}}\right) = \sum_{i=1}^{N} \frac{p_i^{Selec}}{n} \left(\frac{Y_i}{p_i^{Selec}} - \tau_y\right)^2.$$

Na prática, quando a unidade populacional é o ser humano, é impossível conhecer a probabilidade de todas as pessoas serem selecionadas e efetivamente selecionadas, pois elas podem se recusar a responder ou podem não ser encontradas. Ou seja, na realidade, a probabilidade p_i da pessoa i ser incluída na pesquisa é igual ao produto da sua probabilidade de seleção pela sua probabilidade de resposta dado que foi selecionada:

$$p_i = p_i^{Selec} * p_i^{Resp}, \tag{4.2}$$

onde p_i^{Resp} é a probabilidade condicional da i-ésima unidade populacional responder dado que foi selecionada (geralmente desconhecia). Podemos interpretar a p_i^{Resp} em pesquisas domiciliares como a proporção de tempo que o entrevistador está fazendo entrevistas no qual a pessoa i está no seu domicílio com propensão a responder uma pesquisa. Essa interpretação pode parecer um pouco "esotérica", porém utilizar essas probabilidades como uma forma de levar em consideração o erro de não-resposta no estimador pode melhorar significativamente as inferências realizadas.

Note que, ao utilizar o estimador 4.1 desconsiderando as p_i^{Resp} quando elas realmente fazem parte das probabilidades p_i, faz com que o mesmo seja viciado e também implica em uma alteração na forma variância, como pode ser visto abaixo:

$$\begin{aligned} E\left(\sum_{i \epsilon s} \frac{Y_i}{n p_i^{Selec}}\right) &= \sum_{i=1}^{N} Y_i p_i^{Resp} \\ Var\left(\sum_{i \epsilon s} \frac{Y_i}{n p_i^{Selec}}\right) &= \frac{1}{n}\left(\sum_{i=1}^{N} \frac{Y_i p_i^{Resp}}{p_i^{Selec}} - \left(\sum_{i=1}^{N} Y_i p_i^{Resp}\right)^2\right). \end{aligned} \tag{4.3}$$

4.1.1 Modelando a probabilidade de resposta

Como as probabilidades p_i^{Resp} são desconhecidas, é necessário fazer alguma suposição sobre elas para que se possa fazer inferência. Será apresentado nessa seção um modelo de não-resposta bastante útil, o qual permite que o estimador 4.1 seja utilizado, sem as conseqüências descritas em

4.3 causadas pela omissão da p_i^{Resp} nas probabilidades p_i.

Um modelo bastante simples e ingênuo, seria supor que todas as pessoas da população de interesse têm a mesma probabilidade de resposta

$$p_i^{Resp} = p. \tag{4.4}$$

Esse modelo tem pouca utilidade prática. Poucas são as situações onde parece razoável supor que todas as pessoas têm a mesma probabilidade de resposta. Uma varição mais útil desse modelo é supor que existem H grupos, e todas as pessoas dentro de um mesmo grupo têm a mesma probabilidade de resposta:

$$p_i^{Resp} = p_h. \tag{4.5}$$

O modelo em 4.5 é conhecido na literatura como Grupos de Resposta Homogênea (GRH). Em Särndal et al. [1992] há uma discussão mais aprofundada sobre as virtudes e defeitos desse modelo. O interesse nesse modelo é que ele pode ser bastante flexível, bastando para isso alterar o número de grupos H. No caso extremo onde existe somente uma unidade populacional em cada grupo ($H = N$), o modelo **GRH** equivale a não fazer suposição nenhuma sobre as probabilidades de resposta e permitir que cada respondente tenha uma probabilidade de resposta diferente, já no caso onde temos somente um grupo ($H = 1$), obtemos ao modelo ingênuo descrito em 4.4.

Utilizar o modelo **GRH** no contexto de amostragem é equivalente a estratificar a amostra segundo os H grupos, pois uma unidade amostral só pode estar em um dos grupos. Usualmente, para facilitar a notação, identificam-se as unidades populacionais também com um índice para o grupo h o qual pertence, conforme foi visto na Seção 1.2.3. Assim, podemos escrever o estimador em 4.1 como:

$$\hat{\tau}_{HH} = \sum_{h=1}^{H} \sum_{i \epsilon s_h} \frac{Y_{hi}}{n_h p_{hi}^{Selec} p_h}, \tag{4.6}$$

onde s_h é o conjunto dos índices das unidades populacionais pertencentes a amostra do estrato (ou grupo) h, Y_{hi} é o valor da variável de interesse Y para a i-ésima unidade populacional do estrato h, p_{hi}^{Selec} é a probabilidade de seleção em um único sorteio da unidade populacional i pertencente ao estrato h e n_h é o número de entrevistas realizadas no estrato h.

4.2 Inferência condicionada ao conhecimento de p_k^h

Usualmente a Amostragem por Cotas (**AC**) é descrita como uma forma de selecionar um sub-conjunto da população de maneira não-probabilística, selecionando-se um elemento da pop-

ulação se ele satisfizer alguma cota pré-especificada. Por exemplo, se a cota for de 10 pessoas do sexo masculino, as primeiras 10 pessoas do sexo masculino encontradas pelos entrevistadores pertencerão a amostra. Por causa do desconhecimento das probabilidades de seleção para esse tipo de amostragem, não se pode fazer inferência baseada no desenho (**ID**) utilizando o estimador em 4.1.

Nessa seção serão calculadas as probabilidades p_i da unidade populacional i pertencer a amostra (ser selecionada e observada) para a Amostragem Probabilística com Cotas supondo o modelo **GRH** em 4.5 e também **supondo que as H probabilidades definidas em 4.5 sejam conhecidas**, permitindo assim que se faça inferência baseada no desenho (**ID**) a partir de uma amostra probabilística com cotas utilizando o estimador em 4.6.

Também serão calculadas essas probabilidades para dois desenhos amostrais totalmente probabilísticos do tipo **APV**, levando em conta as probabilidades de resposta, um utilizando o modelo em 4.5 e outro supondo o modelo mais simples em 4.4. O objetivo é poder comparar esses 3 modelos sob as mesmas suposições e em condições equivalentes.

4.2.1 Amostragem por Conglomerados em dois estágios

Os dois tipos de amostragem que serão discutidos nessa tese (**APC** e **APV**) são desenhos amostrais em dois estágios, onde no primeiro estágio selecionam-se a conglomerados, os quais são um conjunto de domicílios como o setor censitário, definido na seção 1.2.4, de maneira totalmente probabilística, e no segundo estágio, selecionam-se b moradores residentes em cada um dos conglomerados selecionados no primeiro estágio. Esse tipo de desenho amostral é denominado Amostragem por Conglomerados em dois estágios, conforme discutido na Seção 1.2.4.

Note que tanto a **APC** quanto a **APV** têm um desenho amostral em dois estágios principalmente por causa das informações disponíveis para selecionar a amostra (existente somente para os conglomerados), mas também por causa da logística de coleta de dados, que é muito facilitada por esse tipo de desenho amostral. A diferença entre os dois tipos de desenho amostral só ocorre no último estágio, ou seja, na seleção das pessoas, e não na seleção das unidades primárias (conglomerados).

Nessa tese, a seleção no primeiro estágio (dos conglomerados) será considerada sem reposição (caso mais geral), e no segundo estágio (dos moradores) será considerada com reposição devido a dificuldade do cálculo das probabilidades de seleção p_i^{selec}. Também, no segundo estágio, a amostragem será estratificada, de forma a permitir o uso do modelo **GRH** apresentado em 4.5.

Para formalizar os estimadores utilizados nesse tipo de desenho amostral, precisamos definir alumas quantidades populacionais e deixar claro qual tipo de informação pode ser utilizada para selecionar a amostra. No contexto de amostragem em dois estágios, estaremos supondo que somente existe informação disponível para as unidades primárias (usualmente setores censitários). **Ou seja, sabemos quantos domicílios, ou pessoas, moram naquele conglomerado, e outras informações agregadas desse tipo, mas não existem informações sobre cada domicílio ou morador**. Se existe o interesse em utilizar informações das unidades secundárias (domicílios ou pessoas) para selecionar a amostra, essas informações têm que ser obtidas durante a coleta de

dados.

Vamos supor que existem A unidades primárias, e cada uma é denotada por k, $k = 1, .., A$. Serão selecionadas a unidades primárias, e b pessoas em cada unidade primária selecionada. Assim, o tamanho da amostra final será de $n = a * b$. Dentro de cada unidade primária k, os totais de domicílios e pessoas residentes, denotados por D_k e N_k, são conhecidos. Cada domicílio j do conglomerado k, denotado por $D_{j,k}$, $j = 1, ..., D_k$ possue $N_{j,k}$ moradores, podemos escrever então $N_k = \sum_{j=1}^{D_k} N_{j,k}$. Os domicílios são determinados de maneira única através dos índices $j = 1, ...D_k$, porém o valor $N_{j,k}$ é considerado desconhecido para todo j.

O número de pessoas residentes em cada estrato em cada conglomerado k será denotado por N_k^h, e essas quantidades são conhecidas. Dentro de cada domicílio $D_{j,k}$ podem existir moradores de cada um dos H estratos (ou cotas). O número de moradores do domicílio $D_{j,k}$ pertencentes a cada estrato será denotado por $N_{j,k}^h$, com $h = 1, .., H$. Assim, o total de moradores do conglomerado k do estrato h é dado por $N_k^h = \sum_{j=1}^{D_k} N_{j,k}^h$. O tamanho da amostra dentro de cada estrato do conglomerado k será denotada por b_h, onde $b = \sum_{h=1}^{H} b_h$. Usualmente teremos $b_h = b * \frac{N_k^h}{N_k}$.

Como em cada estágio a forma da seleção é diferente (com e sem reposição), trabalharemos com duas probabilidades diferentes:

- **1º Estágio - Probabilidade de Inclusão (Sem Reposição)-** π_k: é a probabilidade do conglomerado k pertencer a amostra. É definida como sendo $\pi_k = \sum_{s_I \ni k} p(s_I)$, onde s_I é o conjunto das unidades primárias pertencentes a amostra, $p(s_I)$ é a probabilidade da amostra s_I ser selecionada e $s_I \ni k$ indica todas as amostras do primeiro estágio que contém o conglomerado k. Não inclue a probabilidade de resposta, pois no primeiro estágio seleciona-se conglomerados, e não pessoas.

- **2º Estágio - Probabilidade de Seleção (Com Reposição)-** $p_{hi/k}$: é a probabilidade de se selecionar a i-ésima unidade populacional do estrato h, do conglomerado k, em um único sorteio dado que o conglomerado k foi selecionado no primeiro estágio. Como a seleção nesse estágio é com reposição, faz sentido pensar em replicação do sorteio, pois em cada replicação as probabilidades de seleção se mantém. Utiliza-se a notação $/k$ para indicar que essa probabilidade é condicional a seleção do conglomerado k no primeiro estágio. A probabilidade de resposta é considerada nesse estágio.

Se a seleção das unidades populacionais no segundo estágio têm as seguintes propriedades:

1. **Independente** - As sub-seleções das unidades populacionais em cada conglomerado sorteado no primeiro estágio são independentes entre si.

2. **Invariante** - A sub-seleção utilizada em um conglomerado sorteado no primero estágio independe de quais conglomerados foram sorteados no primeiro estágio.

então a probabilidade de seleção da í-ésima unidade populacional, do estrato h e do conglomerado k é dada por:

$$p_{khi} = \pi_k p_{hi/k}. \tag{4.7}$$

Por causa das diferentes probabilidades, em cada estágio também são utilizados diferentes estimadores. Essas diferenças podem ser vistas com mais detalhes em Särndal et al. [1992]. Quando utiliza-se amostragem em 2 estágios, o estimador da quantidade populacional de interesse combina os dois tipos de estimadores, como vemos a seguir:

$$\begin{aligned} \tau_{\hat{HH}} &= \sum_{k \epsilon s_I} \sum_{h=1}^{H} \frac{\hat{\tau}_k^h}{\pi_k} \\ &= \sum_{k \epsilon s_I} \sum_{h=1}^{H} \sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h \pi_k p_{hi/k}}, \end{aligned} \tag{4.8}$$

onde $\hat{\tau}_k^h$ é o estimador do total populacional da variável Y no estrato h no conglomerado k (τ_k^h), s_k^h é o conjunto das unidades amostrais secundárias pertencentes a sub-amostra do estrato h e do conglomerado k e Y_{khi} indica o valor da variável Y de interesse para a i-ésima unidade populacional do estrato h e do conglomerado k. O estimador em 4.8 é não-viciado e a sua variância é dada por:

$$\begin{aligned} Var\left(\tau_{\hat{HH}}\right) &= \sum_{k=1}^{A} \sum_{k'=1}^{A} \frac{\pi_{kk'} - \pi_k \pi_{k'}}{\pi_k \pi_{k'}} \tau_k \tau_{k'} \\ &+ \sum_{k=1}^{A} \frac{\left(\sum_{h=1}^{H} \sum_{i=1}^{N_k^h} \frac{p_{hi/k}}{b_h} \left(\frac{Y_{khi}}{p_{hi/k}} - \tau_k^h \right)^2 \right)}{\pi_k} \\ &= \sum_{k=1}^{A} \sum_{k'=1}^{A} V_{kk'}^E + \sum_{k=1}^{A} \sum_{h=1}^{H} \frac{V_{hk}^I}{\pi_k}, \end{aligned} \tag{4.9}$$

onde $\pi_{kk'}$ é a probabilidade de inclusão conjunta dos conglomerados k e k', definida como sendo $\pi_{kk'} = \sum_{s_I \ni k \& k'} p(s_I)$, τ_k é total populacional do conglomerado k e a quantidade V_{hk}^I é a variância do estimador do total populacional τ_k^h da variável Y no estrato h do conglomerado k.

Da variância do estimador em 4.9 fica evidente que a variância de $\tau_{\hat{HH}}$ pode ser decomposta em duas parcelas, denominadas variâncias entre (V_{entre}) e intra (V_{intra}) conglomerados. A primeira, representada pelo termo $\sum_{k=1}^{A} \sum_{k'=1}^{A} V_{kk'}^E$ e a segunda pelo termo $\sum_{k=1}^{A} \sum_{h=1}^{H} \frac{V_{hk}^I}{\pi_k}$. Note que a variância intra-conglomerados é na verdade uma média ponderada, pelo inverso das probabilidades de inclusão dos conglomerados, das variâncias das estimativas dos totais populacionais de cada conglomerado. Utilizando que:

$$\hat{\tau}_k^h = \sum_{i\epsilon s_k^h} \frac{Y_{khi}}{b_h p_{hi/k}} \tag{4.10}$$

$$\hat{\tau}_k = \sum_{h=1}^{H} \hat{\tau}_k^h = \sum_{h=1}^{H} \sum_{i\epsilon s_k^h} \frac{Y_{khi}}{b_h p_{hi/k}} \tag{4.11}$$

$$\hat{V}(\hat{\tau}_k^h) = \hat{V_{hk}^I} = \frac{1}{b_h(b_h-1)} \sum_{i\epsilon s_k^h} \left(\frac{Y_{khi}}{p_{hi/k}} - \hat{\tau}_k^h \right)^2 \tag{4.12}$$

temos que os estimadores não-viciados dessas variâncias são dados por:

$$\hat{V}_{entre} = \sum_{k\epsilon s_I} \sum_{k'\epsilon s_I} \left(\frac{\pi_{kk'} - \pi_k \pi_{k'}}{\pi_{kk'}} \right) \frac{\hat{\tau}_k}{\pi_k} \frac{\hat{\tau}_{k'}}{\pi_{k'}} - \sum_{k\epsilon s_I} \frac{1}{\pi_k} \left(\frac{1}{\pi_k} - 1 \right) \left(\sum_{h=1}^{H} \hat{V_{hk}^I} \right) \tag{4.13}$$

$$\hat{V}_{intra} = \sum_{k\epsilon s_I} \sum_{h=1}^{H} \frac{\hat{V_{hk}^I}}{\pi_k^2}. \tag{4.14}$$

Somando os estimadores em 4.13 e 4.14 pode-se mostrar que:

$$\hat{Var}\left(\tau_{\hat{HH}}\right) = \sum_{k\epsilon s_I} \sum_{k'\epsilon s_I} \left(\frac{\pi_{kk'} - \pi_k \pi_{k'}}{\pi_{kk'}} \right) \frac{\hat{\tau}_k}{\pi_k} \frac{\hat{\tau}_{k'}}{\pi_{k'}} + \sum_{k\epsilon s_I} \sum_{h=1}^{H} \frac{\hat{V_{hk}^I}}{\pi_k}. \tag{4.15}$$

4.2.2 Amostragem Probabilística com Cotas (APC)

O desenho utilizado aqui é inspirado no desenho proposto por Sudman [1967], denominado Amostragem Probabilística com Cotas (APC), o qual foi apresentado na Seção 2.3.1. Fica claro de 4.8 que só é necessário calcular as probabilidades $p_{hi/k}$ da **APC**, pois as probabilidades π_k são conhecidas. Para conseguir calculá-las, dois fatores são muito importantes:

Probabilidade de Responder Em Sudman [1967], o autor afirma que "Na **APC**, a suposição básica é de que é possível dividir os respondentes em estratos nos quais a probabilidade de estar disponível para ser entrevistado é conhecida e é a mesma para todas as pessoas dentro do estrato, porém variando entre estratos".

Tamanho dos domicílios Em Stephenson [1979], amostras selecionadas através de **APC** e de amostragem probabilística são comparadas, e o autor conclui que "Os procedimentos de seleção selecionam amostras marcadamente diferentes no que diz respeito ao tamanho dos domicílios. A amostra da **APC** superestima a quantidade de domicílios grandes".

Considerar esses dois fatores no cálculo da probabilidade $p_{hi/k}$ é fundamental. A probabilidade

de responder é essêncial pois como no caso da **APC** não existe explicitamente a p_i^{Selec}, assim as probabilidades $p_{hi/k}$ dependem somente da probabilidade p_i^{Resp}. O tamanho dos domicílios, e também a quantidade de moradores de cada domicílio que pertecem a uma cota específica, são quantidades importantes, pois claramente um domicílio maior tem um probabilidade maior de ser selecionado, e esse efeito deve ser considerado nas probabilidades $p_{hi/k}$.

Sob as suposições do modelo **GRH** em 4.5 as probabilidades de seleção da amostragem probabilística com cotas podem ser calculadas, ou seja, quando existem H estratos dentro do conglomerado nos quais a probabilidade de resposta das unidades populacionais são constantes. A probabilidade de resposta segundo o modelo **GRH** da unidade populacional i pertencente ao estrato h do conglomerado k será denotada por p_k^h. Note que estamos utilizando um modelo **GRH** diferente para cada conglomerado k pertencente a amostra, ou seja, as suposições em 4.5 sobre as probabilidades de resposta só são feitas para pessoas residindo dentro de um mesmo conglomerado e pertencentes a uma mesma cota. Essa é uma suposição bem mais fraca, pois somente pessoas que residem num mesmo conglomerado e pertencem a uma mesma cota estão sujeitas a ela. Em Silva and Moura [1990], os autores mostram que em todas as 39 variáveis estudadas, pessoas residentes em um mesmo conglomerado são mais parecidas entre si do que pessoas que residem em diferentes conglomerados. O mesmo deve ocorrer com a probabilidade de resposta, ou seja, nesse cenário, a suposição do modelo **GRH** parece ser bastante aceitável.

Vamos supor também que as cotas utilizadas na pesquisa definem os grupos do modelo 4.5, assim o mesmo índice será utilizado para indicar simultanêamente a cota e o grupo de interesse, ou seja, estamos considerando que as cotas são cruzadas, conforme discutido na Seção 2.2. Além disso, para facilitar a comparação com o desenho amostral probabilístico que será apresentado na Seção 4.2.4 e padronizar a terminologia utilizada nesse capítulo, denominaremos as cotas de estratos. Para que possamos formalizar as probabilidades de seleção da **APC**, vamos definir como é realizado o procedimento para seleção pelo entrevistador na etapa de coleta de dados, sendo que o mesmo procedimento é realizado em todos os conglomerados k selecionadas no primeiro estágio:

1. Ponto Inicial: Seleciona-se um domicílio para iniciar o trajeto em busca de respondentes.

2. Trajeto: O entrevistador percorre todo o trajeto de maneira sequêncial, procurando fazer contato com todos os domicílios do trajeto. Sempre que o trajeto for percorrido pelo entrevistador, os domicílios serão abordados na mesma ordem.

3. Entrevista: Quando o entrevistador encontrar uma pessoa do estrato h acessível e disposta a responder, ela será entrevistada. Se houver mais de uma pessoa em um mesmo domicílio, o entrevistador seleciona o respondente com probabilidades uniformes.

4. Próxima Entrevista: volta ao passo 2, porém existem duas alternativas. **Caso 1:** o entrevistador seleciona novamente com probabilidades uniformes em qual domicílio ele re-iniciará o trajeto ou **Caso 2:** o entrevistador continua o trajeto no domicílio seguinte aquele onde conseguiu completar a entrevista. O entrevistador continua fazendo isso até completar todas

as entrevistas do estrato. Se o entrevistador completar todo o trajeto, ele re-inicia o trajeto, procedendo da mesma maneira.

Primeiramente, vamos definir a sequência de variáveis aleatórias $\{X_{j,k}^h\}$, onde $h = 1, ..., H$ indica o estrato, $j = 1, .., D_k$ indica o domicílio e $k = 1, ..., A$ indica o conglomerado. As v.a's $\{X_{j,k}^h\}$ assumem o valor 0 se o entrevistador não conseguir entrevistar, em uma tentativa, uma pessoa da h-ésimo estrato do j-ésimo domicílio do conglomerado k e 1 caso contrário. Note que estamos implicitamente assumindo que essas probabilidades se mantém constantes, não importando quantas tentativas de contato forem realizadas, ou a hora do dia que foi feita uma tentativa. Temos então:

$$\begin{aligned} P(X_{j,k}^h = 0) &= \left(1 - p_k^h\right)^{N_{j,k}^h} \\ P(X_{j,k}^h = 1) &= 1 - \left(\left(1 - p_k^h\right)^{N_{j,k}^h}\right). \end{aligned} \tag{4.16}$$

A probabilidade de seleção de uma pessoa de um domicílio depende de onde o entrevistador inicia o trajeto (por causa que os domicílios do trajeto têm um número diferente de moradores) e de qual estrato h está sendo procurada uma pessoa. Uma determinada pessoa pode ser selecionada para pertencer a amostra em qualquer um dos b_h sorteios, onde $P_{hi/k}(a^oSelec)$ é a probabilidade da pessoa i do estrato h, que reside no domicílio $D_{j[i],k}$, ser a a-ésima pessoa que o entrevistador consiga entrevistar do estrato h ao percorrer seu trajeto no conglomerado k, onde **o índice** $j[i]$ **indica o domicílio** j **no qual o morador** i **reside** e $a = \{1, .., b_h\}$.

No **caso 1**, onde o entrevistador seleciona com probabilidades uniformes o domicílio onde ele reiniciará o trajeto, a probabilidade de uma pessoa ser selecionada é a mesma em todos os sorteios, o que implica que $P_{hi/k}(a^oSelec) = P_{hi/k}(1^oSelec)$ e consequentemente temos que $p_{hi/k} = P_{hi/k}(1^oSelec)$, ou seja, nesse caso só é necessário calcular a probabilidade $P_{hi/k}(1^oSelec)$. Assim é fácil ver que as variáveis $f_{hi/k}$ que representam a frequência com que a unidade populacional i do estrato h do conglomerado k será incluída na amostra, definidas na Seção 1.2.1, são dadas por:

$$f_{hi/k} \sim Bin\left(b_h, P_{hi/k}(1^oSelec)\right), \tag{4.17}$$

e assim obtemos que

$$\begin{aligned} E(f_{hi/k}) &= b_h \frac{\sum_{a=1}^{b_h} P_{hi/k}(a^oSelec)}{b_h} = b_h P_{hi/k}(1^oSelec) \\ V(f_{hi/k}) &= b_h P_{hi/k}(1^oSelec)\left(1 - P_{hi/k}(1^oSelec)\right) \\ Cov(f_{hi/k}, f_{hj/k}) &= -b_h P_{hi/k}(1^oSelec) P_{hj/k}(1^oSelec). \end{aligned}$$

Já no **caso 2**, onde o entrevistador continua o trajeto no domicílio seguinte aquele onde conseguiu completar a entrevista, em cada seleção as probabilidades são diferentes, pois dependem de qual pessoa foi selecionada no sorteio anterior. Nesse caso, para calcular as variáveis $f_{hi/k}$ será necessário utilizar as variáveis indicadoras auxiliares $Z^a_{hi/k}$ que assumem valores 1 se a unidade populacional i do estrato h do conglomerado k foi selecionada para pertencer a amostra na a-ésima seleção e 0 caso contrário. Temos então que:

$$Z^a_{hi/k} \sim Bin\left(1, P_{hi/k}(a^o Selec)\right).$$

Dessa forma, podemos ver que no **caso 2** a variável $f_{hi/k} = \sum_{a=1}^{b_h} Z^a_{hi/k}$ e assim obtemos que

$$\begin{aligned}
E(f_{hi/k}) &= \sum_{a=1}^{b_h} E\left(Z^a_{hi/k}\right) = \sum_{a=1}^{b_h} P_{hi/k}(a^o Selec) = b_h \frac{\sum_{a=1}^{b_h} P_{hi/k}(a^o Selec)}{b_h} \\
V(f_{hi/k}) &= \sum_{a=1}^{b_h} P_{hi/k}(a^o Selec)\left(1 - P_{hi/k}(a^o Selec)\right) + \sum_{a=1}^{b_h}\sum_{b\neq a}^{b_h} Cov\left(Z^a_{hi/k}, Z^b_{hi/k}\right) \\
Cov(f_{hi/k}, f_{hj/k}) &= -\sum_{a=1}^{b_h} P_{hi/k}(a^o Selec) P_{hj/k}(a^o Selec) + \sum_{a=1}^{b_h}\sum_{b\neq a}^{b_h} Cov\left(Z^a_{hi/k}, Z^b_{hj/k}\right).
\end{aligned}$$

O estimador **HH** apresentado na seção 1.2.6, o qual é utilizado quando a amostragem é com reposição e com probabilidades desiguais, para o total do conglomerado k e do estrato h pode ser escrito como:

$$\hat{\tau}^h_k = \sum_{i\epsilon s^h_k} \frac{Y_{khi}}{E\left(f_{hi/k}\right)}, \tag{4.18}$$

assim, notando que tanto para o **caso 1** quanto para o **caso 2** temos que a esperança de $f_{hi/k}$ é $\frac{b_h}{b_h}\sum_{a=1}^{b_h} P_{hi/k}(a^o Selec)$, podemos utilizar o mesmo estimador para ambos os casos, o qual é dado por:

$$\hat{\tau}^h_k = \sum_{i\epsilon s^h_k} \frac{Y_{khi}}{b_h p_{hi/k}}, \tag{4.19}$$

onde a probabilidade $p_{hi/k}$ é dada por:

$$p_{hi/k} = \frac{\sum_{a=1}^{b_h} P_{hi/k}(a^o Selec)}{b_h}. \tag{4.20}$$

Apesar do estimador do total ser o mesmo nos dois casos, as propriedades dele são diferentes

pois a $V(f_{hi/k})$ e a $Cov(f_{hi/k}, f_{hj/k})$ são diferentes para cada caso. É importante ressaltar que em ambos os casos os estimadores são não-viciados, a diferença se dá na variância desses estimadores. Esse estimador só terá o mesmo comportamento no **caso 1** e no **caso 2** se duas condições forem satisfeitas:

$$1 \quad - \quad Cov\left(Z^a_{hi/k}, Z^b_{hj/k}\right) = 0 \qquad \forall i, j, a, b$$

$$2 \quad - \quad \sum_{a=1}^{b_h} P_{hi/k}(a^o Selec) P_{hj/k}(a^o Selec) = b_h P_{hi/k}(1^o Selec) P_{hj/k}(1^o Selec) \qquad \forall i, j.$$

Nessa tese, apresentaremos toda a teoria para o **caso 1**, porém todos os resultados são válidos para o **caso 2** se ambas as condições forem satisfeitas. Para calcular a probabilidade de interesse $p_{hi/k}$, estamos interessados inicialmente em calcular a probabilidade $P_{hi/k}(1^o Selec)$, pois no **caso 1** elas são iguais, e no **caso 2**, a probabilidade $p_{hi/k}$ pode ser obtida recursivamente à partir da probabilidade $P_{hi/k}(1^o Selec)$, como pode ser visto em 4.30. Para calcular essa probabilidade, é necessário o cálculo de outras probabilidades (nas quais foram omitidos os índices k para não sobrecarregar a notação desse capítulo):

- P_i^h - Probabilidade da pessoa i do estrato h ser a pessoa selecionada dentre todos os $N^h_{j[i],k}$ moradores do domicílio $D_{j[i],k}$ pertencentes ao estrato h dado que pelos menos um morador do domicílio $D_{j[i],k}$ do estrato h está disponível para ser entrevistado.

- $P^h_{D_j}(1^o selec)$ - Probabilidade do domicílio $D_{j,k}$ ser o primeiro domicílio no conglomerado k no qual uma pessoa do estrato h foi entrevistada.

- $P^h_{D_j}(1^o selec/inic.m)$ - Probabilidade do domicílio $D_{j,k}$ ser o primeiro domicílio no conglomerado k no qual uma pessoa do estrato h foi entrevistada, dado que o entrevistador iniciou seu percurso no domicílio $D_{m,k}$.

- $P(inicio.m)$ - Probabilidade do entrevistador iniciar o trajeto no domicílio $D_{m,k}$.

- $P^h_{D_j}(1^o selec.z^a volta/inic.m)$ - Probabilidade do domicílio $D_{j,k}$ ser o primeiro domicílio no conglomerado k no qual uma pessoa foi entrevistada no trajeto, e isso ocorrer na z-ésima volta do entrevistador, dado que o entrevistador iniciou seu percurso no domicílio $D_{m,k}$.

A i-ésima pessoa do estrato h do conglomerado k é selecionada se o domicílio onde ela reside $j[i]$ for selecionado, e se dentre os moradores desse domicílio, ela for a pessoa selecionada. Assim a probabilidade $P_{hi/k}(1^o Selec)$ é dada pelo produto das probabilidades P_i^h e $P^h_{D_j}(1^o selec)$. Inicialmente estamos interessados em calcular a P_i^h. Para isso é importante notar que, apesar da notação não explicitar isso, essa é uma probabilidade condicional ao domicílio $j[i]$ ter sido selecionado, o que implica que sabemos que pelo menos um morador do domicílio estava acessível e disposto a responder a pesquisa. Assim estamos interessados na esperança de uma variável aleatória $\frac{1}{x}$ onde

x é o número de moradores acessíveis e dispostos a responder a pesquisa, condicionado a existência de pelo menos um morador com essas características ($x > 0$). Podemos escrever então:

$$\begin{aligned} P_i^h &= P_i^h(resp/D_{j[i]}^h resp) = \frac{P_i^h(resp \cap D_{j[i]}^h resp)}{P(D_{j[i]}^h resp)} \\ &= \frac{\sum_{x=1}^{N'} \frac{1}{x}\binom{N'-1}{x-1} p^x (1-p)^{N'-x}}{1-(1-p)^{N'}} \\ &= \frac{1}{N'}, \end{aligned} \tag{4.21}$$

onde $N' = N_{j[i],k}^h$ e $p = p_k^h$. Note que utilizamos $\binom{N'-1}{x-1}$ pois estamos supondo que a unidade populacional i é uma das unidades acessíveis e dispostas a responder. Ou seja, dado que uma pessoa ou mais do domicílio $j[i]$ esteja(m) disposta(s) a responder, se escolhermos uma delas com probabilidade uniforme, o respondente selecionado tem a probabilidade P_i^h dada por $\frac{1}{N_{j[i],k}^h}$.

Assim, utilizando o teorema das probabilidades totais e o resultado em 4.21, a probabilidade $P_{hi/k}(1^o Selec)$ pode ser decomposta em:

$$\begin{aligned} P_{hi/k}(1^o Selec) &= P_i^h P_{D_{j[i]}}^h(1^o selec) \\ &= P_i^h \sum_{m=1}^{D_k} P_{D_{j[i]}}^h(1^o selec/inic.m) P(inicio.m) \\ &= \frac{1}{N_{j[i],k}^h} \sum_{m=1}^{D_k} P_{D_{j[i]}}^h(1^o selec/inic.m) \frac{1}{D_k} \\ &= \frac{1}{D_k} \frac{1}{N_{j[i],k}^h} \sum_{m=1}^{D_k} P_{D_{j[i]}}^h(1^o selec/inic.m), \end{aligned} \tag{4.22}$$

supondo que o domicílio onde o trajeto se iniciará seja escolhido com probabilidade uniforme $\frac{1}{D_k}$, pois não existem informações disponíveis para sortear o domicílio com probabilidade proporcional ao tamanho.

Considerando que o domicílio j pode ser selecionado depois que o entrevistador der z voltas completas no trajeto , com $z \epsilon \{0, 1, 2, ...\}$, temos que:

$$P_{D_j}^h(1^o selec.z^a volta/inic.m) = P_{D_j}^h(1^o selec.1^a volta/inic.m) * \left[\left(1 - p_k^h\right)^{N_k^h}\right]^z \tag{4.23}$$

onde o termo $\left[\left(1 - p_k^h\right)^{N_k^h}\right]$ é a probabilidade do entrevistador dar uma volta completa no trajeto sem conseguir entrevistar nenhuma pessoa do estrato h. Observando que o número de voltas dadas pelo entrevsitador forma uma partição, temos que a probabilidade $P_{D_j}^h(1^o selec/inic.m)$ pode ser

decomposta em:

$$\begin{aligned} P_{D_j}^h(1^{\text{o}}selec/inic.m) &= \sum_{z=0}^{\infty} P_{D_j}^h(1^{\text{o}}selec.z^{\text{a}}volta/inic.m) \\ &= \sum_{z=0}^{\infty} P_{D_j}^h(1^{\text{o}}selec.1^{\text{a}}volta/inic.m) * \left[\left(1-p_k^h\right)^{N_k^h}\right]^z \\ &= \frac{P_{D_j}^h(1^{\text{o}}selec.1^{\text{a}}volta/inic.m)}{\left[1-\left(1-p_k^h\right)^{N_k^h}\right]} \end{aligned} \tag{4.24}$$

Assim, para calcular a probabilidade $P_{D_j}^h(1^{\text{o}}selec/inic.m)$ em 4.22 é preciso inicialmente calcular a probabilidade $P_{D_j}^h(1^{\text{o}}selec.1^{\text{a}}volta/inic.m)$. Para isso, é útil separar essas probabilidades em três grupos:

- $(j > m)$ - O domicílio $D_{j,k}$ para o qual queremos calcular a probabilidade está localizado depois do domicílio $D_{m,k}$ onde entrevistador iniciará o trajeto.
- $(j = m)$ - O domicílio $D_{j,k}$ para o qual queremos calcular a probabilidade é o mesmo domicílio onde entrevistador iniciará o trajeto.
- $(j < m)$ - O domicílio $D_{j,k}$ para o qual queremos calcular a probabilidade está localizado antes do domicílio $D_{m,k}$ onde entrevistador iniciará o trajeto,

onde $m = 1, .., D_k$ e $j = 1, .., D_k$.

Se $j > m$, quer dizer que para uma pessoa do estrato h do domicílio $D_{j,k}$ ser a primeira pessoa entrevistada na primeira volta, então nenhuma pessoa dos domicílios $D_{a,k}$, $a = m, .., j-1$ pode ser entrevistada na primeira volta do entrevistador. Ou seja, essa probabilidade é dada por:

$$\begin{aligned} P_{D_j}^h(1^{\text{o}}selec.1^{\text{a}}volta/inic.m) &= P\left(X_{j,k}^h = 1\right) * P\left(\sum_{a=m}^{j-1} X_{a,k}^h = 0\right) \\ &= \left[1-\left(\left(1-p_k^h\right)^{N_{j,k}^h}\right)\right] * \left(1-p_k^h\right)^{\sum_{a=m}^{j-1} N_{a,k}^h} \end{aligned} \tag{4.25}$$

Se $j = m$, quer dizer que para uma pessoa do estrato h do domicílio $D_{j,k}$ ser a primeira pessoa entrevistada na primeira volta, então uma pessoa desse domicílio tem que ser entrevistada nessa primeira tentativa. Ou seja, essa probabilidade é dada por:

$$\begin{aligned} P_{D_j}^h(1^{\text{o}}selec.1^{\text{a}}volta/inic.m) &= P\left(X_{j,k}^h = 1\right) \\ &= \left[1-\left(\left(1-p_k^h\right)^{N_{j,k}^h}\right)\right] \end{aligned} \tag{4.26}$$

Se $j < m$, quer dizer que para uma pessoa do estrato h do domicílio $D_{j,k}$ ser a primeira pessoa entrevistada na primeira volta, então nenhuma pessoa dos domicílios $D_{a,k}$, $a = j, .., D_k, 1, .., m-1$ pode ser entrevistada na primeira volta do entrevistador. Ou seja, essa probabilidade é dada por:

$$\begin{aligned} P^h_{D_j}(1^o selec.1^a volta \quad / \quad inic.m) &= P\left(X^h_{j,k} = 1\right) * P\left(\sum_{a=m}^{D_k} X^h_{a,k} + \sum_{a=1}^{j-1} X^h_{a,k} = 0\right) \\ &= \left[1 - \left(\left(1-p^h_k\right)^{N^h_{j,k}}\right)\right] * \left(1-p^h_k\right)^{\left(\sum_{a=m}^{D_k} N^h_{a,k} + \sum_{a=1}^{j-1} N^h_{a,k}\right)} \end{aligned} \tag{4.27}$$

Com as probabilidades em 4.25, 4.26 e 4.27, é possível calcular $P^h_{D_j}(1^o selec/inic.m)$. Substituindo em 4.22 essas probabilidades alteradas conforme 4.24 , obtemos:

$$\begin{aligned} P_{hi/k}(1^o Selec) &= \frac{1}{D_k}\frac{1}{N^h_{j[i],k}} \sum_{m=1}^{D_k} P^h_{D_{j[i]}}(1^o selec/inic.m) \\ &= \frac{1}{D_k}\frac{1}{N^h_{j[i],k}} \frac{\left[1 - \left(\left(1-p^h_k\right)^{N^h_{j[i],k}}\right)\right]}{\left[1 - \left(1-p^h_k\right)^{N^h_k}\right]} * C_{j[i]}(p^h_k) \end{aligned} \tag{4.28}$$

onde $C_{j[i]}(p^h_k)$ é igual a:

$$\sum_{inic=1}^{j[i]-1} \left(1-p^h_k\right)^{\sum_{a=inic}^{j[i]-1} N^h_{a,k}} \quad + \quad 1 + \left(1-p^h_k\right)^{\sum_{a=1}^{j[i]-1} N^h_{a,k}} * \left(\sum_{inic=j[i]+1}^{D_k} \left(1-p^h_k\right)^{\sum_{a=inic}^{D_k} N^h_{a,k}}\right).$$

Para calcular as probabilidades de seleção $p_{hi/k}$ para o **caso 2** precisamos calcular $P_{hi/k}(a^o Selec)$ para $a > 1$. Para isso será necessário definir mais duas probabilidades de interesse:

- $P^h_{D_j}(a^o selec)$ - Probabilidade de que um morador do domicílio $D_{j,k}$ do estrato h do conglomerado k seja a a-ésima pessoa entrevistada do estrato h do conglomerado k.
- $P^h_{D_j}(a^o selec/D_i.z^o selec)$ - Probabilidade de que um morador do domicílio $D_{j,k}$ do estrato h do conglomerado k a seja a a-ésima pessoa entrevistada do estrato h do conglomerado k dado que um morador do domicílio $D_{i,k}$ do estrato h do conglomerado k foi a z-ésima pessoa entrevistada do estrato h do conglomerado k, com $a > z$.

Utilizando essas probabilidades, de forma análoga a 4.20, podemos escrever $p_{hi/k}$ da seguinte forma:

$$p_{hi/k} = \frac{1}{b_h N^h_{j[i],k}} \sum_{a=1}^{b_h} P^h_{D_{j[i]}}(a^o selec). \tag{4.29}$$

Note que para $a > 1$, a probabilidade $P^h_{D_j}(a^o selec)$ em 4.29 pode ser escrita como:

$$\begin{aligned} P^h_{D_j}(a^o selec) &= \sum_{m=1}^{D_{Ak}} P^h_{D_j}(a^o selec/D_m.(a-1)^o selec) * P^h_{D_m}((a-1)^o selec) \\ &= \sum_{m=1}^{D_{Ak}} P^h_{D_j}(1^o selec/inic.m) * P^h_{D_m}((a-1)^o selec) \end{aligned} \tag{4.30}$$

Para o caso com $a = 2$, de 4.22 sabemos que $P^h_{D_{j[i]}}(1^o selec)$ se reduz a $N^h_{j[i],k} P_{hi/k}(1^o Selec)$, quando o morador i do estrato h reside no domicílio $j[i]$. Multiplicamos $P_{hi/k}(1^o Selec)$ nesse caso por $N^h_{j[i],k}$ pois estamos interessados aqui somente na probabilidade do domicílio ser selecionado, e não da pessoa. Ou seja, com $a = 2$, obtemos de 4.30 que:

$$\begin{aligned} P^h_{D_j}(2^o selec) &= \sum_{m=1}^{D_{Ak}} P^h_{D_j}(1^o selec/inic.m) * P^h_{D_m}(1^o selec) \\ &= \sum_{m=1}^{D_{Ak}} P^h_{D_j}(1^o selec/inic.m) * N^h_{m,k} P_{hi(m)/k}(1^o Selec), \end{aligned} \tag{4.31}$$

onde o índice $i(m)$ indica uma pessoa i que reside no domicílio m. É possível ver que, recursivamente, partindo do caso $a = 2$ em 4.31 e utilizando 4.30 podemos obter todas as b_h parcelas necessárias para se obter a probabilidade em 4.29. Porém, analiticamente, é bastante difícil calcular essas probabilidades.

A fim de permitir uma comparação analítica, vamos trabalhar com o **caso 1**, onde o domicílio onde o entrevistador continuará o trajeto após entrevistar uma pessoa é selecionado com probabilidades uniformes. Nesse caso, conforme discutido anteriormente, temos que 4.29 se reduz a:

$$p_{hi/k} = \frac{1}{b_h N^h_{j[i],k}} \sum_{a=1}^{b_h} P^h_{D_{j[i]}}(a^o selec) = P_{hi/k}(1^o Selec). \tag{4.32}$$

Uma consequência interessante do resultado em 4.32 e de 4.28, é que se $p^h_k = 1$, então as probabilidades de seleção $p_{hi/k}$ são iguais a

$$p_{hi/k} = \frac{1}{D_k} \frac{1}{N^h_{j[i],k}}. \tag{4.33}$$

Ou seja, no caso da **APC**, a probabilidade de seleção desconsiderando a probabilidade de resposta seria definida pela probabilidade do entrevistador iniciar o trajeto em um determinado domicílio e pelo número de moradores do domicílio pertencentes ao estrato h. Quando não existe

a não-resposta, que é o caso quando $p_k^h = 1$, a seleção pela **APC** é equivalente a entrevistar um morador do primeiro domicílio selecionado.

Um problema com as probabilidades em 4.28 e 4.32 é que essas probabilidades de seleção dependem de todas as quantidades $N_{j,k}^h$, que são desconhecidas. Essa dependência ocorre porque a probabilidade de seleção de um domicílio depende do tamanho dos domicílios que o precedem no trajeto definido. Essa questão será discutida na Seção 4.4. Na Seção 4.2.4 veremos que o mesmo problema ocorre na amostragem probabilística com voltas (**APV**) quando o tamanho da amostra é fixo.

Amostragem por Cotas (AC)

Também é muito comum falar em Amostragem por Cotas (AC). Podemos pensar a AC como um caso específico da **APC**, porém sem o primeiro estágio de seleção. No contexto de pontos de fluxo, também é usual considerar que todos os domicílios têm apenas um morador, pois nesse contexto não faz sentido considerar o efeito dos domicílios. Os resultados da **APC** podem ser facilmente adaptados para a AC, bastando para isso supor que a população é composta somente por um conglomerado ($A = 1$) com probabilidade de inclusão $\pi_{A_1} = 1$.

Utilizando as mesmas probabilidades de seleção $p_{hi/k}$ em 4.29, porém alterando a notação para p_{hi} pois nesse contexto não há a necessidade de indicar dependência no conglomerado selecionado no primeiro estágio, é fácil ver que:

$$\hat{\tau_{HH}} = \sum_{h=1}^{H} \sum_{i \epsilon s_h} \frac{Y_{hi}}{b_h p_{hi}}, \tag{4.34}$$

onde s_h é o conjunto das unidades populacionais pertencentes a amostra do estrato h e Y_{hi} indica o valor da variável Y de interesse para a i-ésima unidade populacional do estrato h. O estimador em 4.34 é não-viciado e a sua variância é dada por:

$$Var\left(\hat{\tau_{HH}}\right) = \sum_{h=1}^{H} \sum_{i=1}^{N^h} \frac{p_{hi}}{b_h} \left(\frac{Y_{hi}}{p_{hi}} - \tau^h\right)^2, \tag{4.35}$$

onde N^h é o número de moradores do estrato h e τ^h é o total populacional da variável Y para o estrato h. O estimador não-viciado de $Var\left(\hat{\tau_{HH}}\right)$ é dado por:

$$\hat{Var}\left(\hat{\tau_{HH}}\right) = \sum_{h=1}^{H} \frac{1}{b_h(b_h - 1)} \sum_{i \epsilon s_h} \left(\frac{Y_{hi}}{p_{hi}} - \hat{\tau}^h\right)^2, \tag{4.36}$$

onde $\hat{\tau}^h = \sum_{i \epsilon s_h} \frac{Y_{hi}}{b_h p_{hi}}$.

Matematicamente, as probabilidade de seleção para a AC apresentadas nessa seção estão cor-

retas, porém é evidente que a suposição do modelo **GRH** para o caso da **AC** é muito mais forte do que para o caso da **APC**, pois utiliza-se o modelo mais simples definido em 4.4, supondo-se que todas as pessoas de uma mesma cota têm a mesma probabilidade de resposta, diferentemente da **APC**, onde essa suposição só é feita para pessoas que residem no mesmo conglomerado.

4.2.3 Número de voltas esperadas para completar as entrevistas na APC

É muito difícil calcular o tempo esperado no caso **APC**, assim aproximações terão que ser utilizadas. Um dos problemas ao comparar o tempo de execução da **APC** e da **APV** é que elas são calculadas em diferentes unidades de tempo. No caso da APC, a medida de tempo utilizada foi a quantidade de voltas completas dadas pelo entrevistador, em cada conglomerado, para terminar a coleta de dados. Já na **APV**, utiliza-se como medida de tempo o número de pessoas e domicílios abordados. Esses resultados estão descritos na Seção 4.2.5.

Definição 4.1 (Distribuição Binomial Negativa - BN(n,p)) *Seja X uma variável aleatória que represente o número de tentativas até se obter n sucessos ($n = \{1, 2, 3, ...\}$) , onde a probabilidade de sucesso é p. Então X segue uma distribuição binomial-negativa, com parâmetros n e p, com :*

$$E(X) = \frac{n}{p}. \tag{4.37}$$

Para calcular o número de voltas completas esperadas é necessário utilizar uma aproximação, pois é muito difícil encontrar a distribuição exata dessa quantidade. Se supormos que somente uma entrevista pode ser realizada por volta completa, da definição 4.1, temos que o número de voltas realizadas no estrato h no conglomerado k segue uma distribuição binomial negativa

$$N_{h,k}^{voltas} \sim BN\left(b_h, 1 - \left(1 - p_k^h\right)^{N_k^h}\right), \tag{4.38}$$

e consequentemente o número esperado de voltas realizadas até completar as b_h entrevistas do estrato h do conglomerado k é dado por

$$E_(N_{h,k}^{voltas}) = \frac{b_h}{1 - \left(1 - p_k^h\right)^{N_k^h}}. \tag{4.39}$$

Essa aproximação usualmente não é muito boa, pois podemos ver em 4.39 que basta N_k^h ser grande e/ou a probabilidade de resposta p_k^h ser pequena, para que o número de voltas esperadas seja aproximadamente b_h. Por outro lado, sabe-se que essa aproximação super-estima o número de voltas esperadas, assim dificilmente o entrevistador terá que completar mais do que b_h voltas para

terminar a coleta de dados dentro do estrato h do conglomerado k. Utilizando essa aproximação, o total de voltas esperadas por conglomerado é dado por:

$$E_(N_k^{voltas}) = \sum_{h=1}^{H} \frac{b_h}{1-\left(1-p_k^h\right)^{N_k^h}}. \tag{4.40}$$

Esse resultado supõe que a busca em cada estrato h será feita de forma independente, mas geralmente, a busca em todos os estratos h é realizada em paralelo, ou seja, o entrevistador procura completar todas as H cotas simultaneamente ao percorrer o trajeto. Modelando o tempo para concluir a coleta de dados dessa forma, temos que o número de voltas realizadas segue uma distribuição binomial negativa

$$N_k^{voltas.simul} \sim BN\left(b, 1-\prod_{h=1}^{H}\left(\left(1-p_k^h\right)^{N_k^h}\right)\right), \tag{4.41}$$

e consequentemente o número esperado de voltas realizadas até completar as b entrevistas do conglomerado k é dado por

$$E_(N_k^{voltas.simul}) = \frac{b}{1-\prod_{h=1}^{H}\left(\left(1-p_k^h\right)^{N_k^h}\right)}. \tag{4.42}$$

4.2.4 Amostragem Probabilística com Voltas com Não-Resposta (APV)

Nesta seção, o objetivo é derivar um desenho amostral totalmente probabilístico (**APV**) que leve em conta a não-resposta. O intuito é que esse desenho possa ser considerado equivalente com a APC apresentada na Seção 4.2.2, para que as inferências realizadas a partir desses dois tipos de desenhos amostrais possam ser comparadas.

Primeiramente, esta amostragem deve ser estratificada, onde os estratos devem coincidir com as cotas definidas para **APC**, de maneira a respeitar as suposições do model **GRH** definido em 4.5. Esse tipo de amostragem usualmente não é utilizada na prática, pois complica bastante a coleta dos dados. Também, de 4.33, fica evidente que o desenho amostral de **APV** deve ter três estágios, diferente da **APC** que tinha somente dois estágios. A seleção do domicílio e a seleção do entrevistado na **APV** são dois estágios diferentes. Assim os estágios da **APV** podem ser descritos como:

1. **Seleção do Conglomerado**: seleção dos conglomerados (ou áreas geográficas) com probabilidade π_k.

2. **Seleção do Domicílio**: seleção dos domicílios com probabilidades uniformes, devida a falta

de informação, com $P_{D_{j,k}}(selec) = \frac{1}{D_k}$.

3. **Seleção da Pessoa**: seleção do entrevistado com probabilidades uniformes dentre aqueles moradores que pertencem ao mesmo estrato, com $P_i^h(selec) = \frac{1}{N^h_{j[i],k}}$.

Utilizando esse desenho proposto com as probabilidades acima, obtemos facilmente que:

$$p_{hi/k} = \frac{1}{D_k}\frac{1}{N^h_{j[i],k}}, \tag{4.43}$$

que é igual ao resultado obtido em 4.33, ou seja, os dois desenhos propostos, **APC** e **APV** são idênticos se não levarmos em conta a probabilidade de resposta.

Na **APV**, o estágio de seleção dos domicílios pode ser realizado pelo estatístico no escritório, com probabilidades uniformes. Porém, o último estágio, da seleção do respondente, tem que ser realizado pelo entrevistador, durante a coleta dos dados, pois não existe informação no nível do domicílio. Ou seja, não sabe-se quantos moradores residem no domicílio selecionado, assim o entrevistador seleciona o respondente com probabilidades uniformes. Para conseguir selecionar um residente do domicílio selecionado, o entrevistador terá que, após descobrir o número de moradores do domicílio, fazer contato com algum morador do domicílio (de qualquer estrato). Feito isso, pergunta ao residente contactado quantas pessoas residem nesse domicílio. Com essas informações, o próprio entrevistador seleciona probabilisticamente a pessoa que será entrevistada.

Por esses motivos, a probabilidade de resposta se faz presente nos dois últimos estágios da **APV** Estratificada: quando o entrevistador tenta fazer contato com o domicílio selecionado e com a pessoa selecionada. Para formalizar as probabilidades de resposta nesse dois casos, foram utilizados dois parâmetros importantes, que devem ser definidos pelo pesquisador antes de iniciar a coleta de dados:

1. $\kappa 1$ **- Máximo de tentativas de contactar o domicílio selecionado:** Esse será o número máximo de vezes que o entrevistador tentará fazer contato com algum morador do domicílio selecionado. Se não for possível falar com nenhum morador, outro domicílio será selecionado.

2. $\kappa 2$ **- Máximo de tentativas de contactar o residente selecionado:** Esse será o número máximo de vezes que o entrevistador tentará fazer contato com o morador selecionado. Se não for possível falar com esse morador, outro domicílio será selecionado.

Esses dois parâmetros influenciam diretamente a qualidade da pesquisa e também no tempo de coleta de dados. A probabilidade de inclusão $p_{hi/k}$, levando em conta a probabilidade de resposta, é dada por:

$$p_{hi/k} = P_{D_{j[i],k}}(selec) * P_{D_{j[i],k}}(resp) * P_i^h(selec) * P_i^h(resp), \tag{4.44}$$

onde $P_{D_{j[i],k}}(selec)$ é a probabilidade do domicílio $D_{j[i],k}$ ser selecionado, $P_{D_{j[i],k}}(resp)$ é a probabilidade de algum morador do domicílio $j[i]$ responder dado que o domicílio $D_{j[i]}$ foi selecionado, $P_i^h(selec)$ é a probabilidade do morador i do estrato h do domicílio $j[i]$ ser selecionado dado que o domicílio $D_{j[i]}$ foi selecionado e algum morador respondeu e $P_i^h(resp)$ é a probabilidade do morador i do domicílio $j[i]$ pertencente ao estrato h ser entrevistado dado que ele foi selecionado e que o domicílio $D_{j[i]}$ foi selecionado e algum morador respondeu. Apesar de todas as probabilidades em 4.44 serem condicionais, a notação utilizada não explicitará esse condicionamento.

Para formalizar a probabilidade $P_{D_{j,k}}(resp)$ do entrevistador conseguir fazer contato com alguém do domicílio, basta notar que essa probabilidade é simplesmente dada pela probabilidade de pelo menos uma pessoa do domicílio responder em até κ_1 tentativas:

$$\begin{aligned} P_{D_{j,k}}(resp) &= \left(1-\prod_{h=1}^{H}(1-p_k^h)^{N_{j,k}^h}\right)\left(\sum_{i=0}^{\kappa_1-1}\left(\prod_{h=1}^{H}(1-p_k^h)^{N_{j,k}^h}\right)^i\right) \\ &= \left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j,k}^h}\right) \end{aligned} \tag{4.45}$$

É interessante notar que a probabilidade em 4.45 depende da quantidade p_k^h de todos os H estratos. Usando exatamente o mesmo raciocínio, a probabilidade $P_i^h(resp)$ do entrevistador conseguir entrevistar o morador selecionado é dada por:

$$\begin{aligned} P_i^h(resp) &= p_k^h\left(\sum_{i=0}^{\kappa_2-1}\left(1-p_k^h\right)^i\right) \\ &= \left(1-\left(1-p_k^h\right)^{\kappa_2}\right) \end{aligned} \tag{4.46}$$

Substituindo as probabilidades 4.45 e 4.46 em 4.44 obtemos que:

$$p_{hi/k} = \frac{\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j[i],k}^h}\right)}{D_k}\frac{\left(1-\left(1-p_k^h\right)^{\kappa_2}\right)}{N_{j[i],k}^h}, \tag{4.47}$$

Novamente, se desconsiderarmos a probabilidade de resposta fazendo $p_k^h = 1$ obtemos que $p_{hi/k}$ em 4.47 se reduz ao caso teórico, apresentado em 4.43.

Apesar da **APV** ter sido desenhada para ser comparável com a **APC**, existe uma diferença fundamental entre elas: a **APV**, como definida até aqui, tem um tamanho de amostra aleatório. Isso ocorre porque, do jeito que as probabilidades em 4.44 foram calculadas, quando o entrevistador não consegue fazer contato com um domicílio ou um morador selecionado, o tamanho efetivo da APV decresce de uma unidade. O interesse em modificar a **APV** para que passe a ter tamanho amostral fixo também é relevante do ponto de vista prático, pois usualmente as pesquisas contratadas têm

tamanho amostral pré-determinado.

Amostragem Probabilística com Voltas com Não-Resposta - Tamanho da amostra fixo

Para encontrar as probabilidades de seleção para o caso da **APV** com tamanho de amostra fixo ($p_{hi/k}^{n_{fixo}}$), o primeiro passo é calcular qual é a probabilidade de não se entrevistar alguma pessoa em uma única busca. Aqui, a busca deve ser interpretada como selecionar-se um domicílio/pessoa e realizar todo o esforço necessário para tentar entrevistá-la, ou seja, tentar fazer contato com o domicílio até κ_1 vezes e com a pessoa até κ_2 vezes.

É mais fácil calcular essa probabilidade como sendo o complemento da probabilidade de se entrevistar alguma pessoa em uma única busca, denotada por $p_{entrev}^{h,k}$. Assim temos que, no estrato h do conglomerado k essa probabilidade é dada por:

$$\begin{aligned} p_{entrev}^{h,k} &= \sum_{i=1}^{N_k^h} p_{hi/k} \\ &= \sum_{j=1}^{D_k} N_{j,k}^h * p_{hi/k} \\ &= \left(1-\left(1-p_k^h\right)^{\kappa_2}\right) \sum_{j=1}^{D_k} \frac{\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j,k}^h}\right)}{D_k} \end{aligned} \tag{4.48}$$

Cada pessoa que pertence a amostra pode ser efetivamente entrevistada na primeira busca, ou na segunda busca e assim por diante. Ou seja, o número de buscas que foram realizadas para entrevistar uma pessoa forma uma partição, permitindo que possamos calcular as probabilidades de seleção $p_{hi/k}$ para o caso da APV com tamanho de amostra fixo como sendo:

$$\begin{aligned} p_{hi/k}^{n_{fixo}} &= \sum_{m=0}^{\infty} p_{hi/k}\left(1-p_{entrev}^{h,k}\right)^m = \frac{p_{hi/k}}{p_{entrev}^{h,k}} \\ &= \frac{1}{N_{j[i],k}^h} \frac{\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j[i],k}^h}\right)}{\sum_{a=1}^{D_k}\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{a,k}^h}\right)} \end{aligned} \tag{4.49}$$

Ou seja, as probabilidades de seleção da **APV** com tamanho de amostra fixo são uma re-parametrização das probabilidades de seleção da APV em 4.47, de formar a garantir que a soma dessas probabilidades seja um. Fica evidente em 4.49 que as probabilidades de seleção para **APV** com tamanho de amostra fixo não dependem do parâmetro κ_2. Apesar disso ficará claro nas próximas seções que o parâmetro κ_2 continua sendo relevante.

Um problema com as probabilidades em 4.49 é que essas probabilidades de seleção dependem de todas as quantidades $N_{j,k}^h$, que são desconhecidas. Essa questão será discutida na Seção 4.4.

4.2.5 Número de Contatos esperados para completar as entrevistas na APV

Existe o interesse em calcular quanto tempo leva, ao utilizar a **APV** com tamanho fixo, para se completar a coleta de dados, principalmente porque um dos principais fatores utilizados para justificar **APC** é que o seu tempo de coleta de dados é menor do que o da **APV**. Nesse contexto, três quantidades são importantes:

Domicílios Buscados é o total de domicílios nos quais se tentou fazer contato, inclui tanto aqueles onde se conseguiu fazer contato quanto aqueles onde não foi possível fazer contato.

Pessoas Buscadas é o total de pessoas que foram procuradas para pertencer a amostra, inclui tanto aquelas que foram efetivamente entrevistadas quanto aquelas que não responderam. Só considerá-se uma pessoa como procurada se o entrevistador conseguiu fazer contato ao menos com o domcílio onde ela reside.

Contatos Realizados é o total de vezes que o entrevistador tentou fazer contato com os domicílios e as pessoas selecionadas, tendo conseguido ou não fazer uma entrevista.

O número de domicílios buscados sempre é maior do que o número de pessoas buscadas, pois sempre para se buscar uma pessoa é preciso ter feito contato com o domicílio, já para fazer contato com um domicílio não é necessário ter feito contato com uma pessoa. Quando não se consegue contato com um domicílio ou com uma pessoa, o próximo passo é selecionar outro domicílio e tentar fazer contato novamente, por isso utilizamos para calcular o número de domicílios buscados a probabilidade de se entrevistar alguma pessoa em uma única busca, definida em 4.48, pois ela considera a probabilidade de fazer contato com o domicílio e com a pessoa.

Da definição 4.1, temos que o número de domicílios abordados no estrato h segue uma distribuição binomial negativa

$$N_{h,k}^{dom} \sim BN\left(b_h, p_{entrev}^{h,k}\right), \tag{4.50}$$

e consequentemente o número esperado de domicílios abordados do estrato h do conglomerado k é dado por

$$E_(N_{h,k}^{dom}) = \frac{b_h}{p_{entrev}^{h,k}} = \frac{b_h}{\left(1-\left(1-p_k^h\right)^{\kappa_2}\right)\sum_{j=1}^{D_k}\frac{\left(1-\prod_{h=1}^{H}\left(1-p_k^h\right)^{\kappa_1 N_{j,k}^h}\right)}{D_k}}. \tag{4.51}$$

De maneira semelhante, também conseguimos modelar o número de pessoas abordadas no estrato h com uma distribuição binomial negativa

$$N_{h,k}^{pes} \sim BN\left(b_h, 1-\left(1-p_k^h\right)^{\kappa_2}\right), \tag{4.52}$$

e consequentemente o número esperado de pessoas abordadas do estrato h do conglomerado k é dado por

$$E_(N_{h,k}^{pes}) = \frac{b_h}{1-\left(1-p_k^h\right)^{\kappa_2}}. \tag{4.53}$$

De 4.50 e de 4.52 é facil ver que o número esperado de domicílios do estrato h do conglomerado k onde o entrevistador não consegue fazer contato é dado por:

$$\begin{aligned} E_(\tilde{N}_{h,k}^{dom}) &= E(N_{h,k}^{dom}) - E_(N_{h,k}^{pes}) \\ &= \frac{b_h}{1-\left(1-p_k^h\right)^{\kappa_2}} \left(\frac{1}{\frac{\sum_{j=1}^{D_k}\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j,k}^h}\right)}{D_k}} - 1 \right) \\ &= \frac{b_h}{1-\left(1-p_k^h\right)^{\kappa_2}} \left(\frac{\sum_{j=1}^{D_k}\left(\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j,k}^h}\right)}{\sum_{j=1}^{D_k}\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j,k}^h}\right)} \right) \end{aligned} \tag{4.54}$$

Para calcular o número de contatos, é importante separar os contatos em dois momentos distintos, ao tentar contactar o domicílio, e ao tentar contactar o morador selecionado. Podemos modelar ambos os casos como tendo uma distribuição geométrica truncada, pois existe um limite máximo de contatos, dados por κ_1 e κ_2, respectivamente. Para mais detalhes sobre essa distribuição, consulte Colwell and Gillett [1989] e Thomasson and Kapadia [1975].

Definição 4.2 (Distribuição Geométrica Truncada - GT(κ,p)) *Seja X uma variável aleatória que represente o número de tentativas até se obter o primeiro sucesso, onde a probabilidade de sucesso é p e serão realizados no máximo κ tentativas. Então X segue uma distribuição geométrica truncada, com parâmetros κ e p, com a seguinte média:*

$$E(X) = \frac{1-(\kappa+1)(1-p)^{\kappa}+\kappa(1-p)^{\kappa+1}}{(1-(1-p)^{\kappa})p}. \tag{4.55}$$

Utilizando a distribuição definida em 4.2, o número de contatos realizados até conseguir falar com alguém do domicílio j do conglomerado k, $C_{j,k}^{dom}$ segue a distribuição:

$$C_{j,k}^{dom} \sim GT\left(\kappa_1, 1 - \prod_{h=1}^{H}(1-p_k^h)^{N_{j,k}^h}\right), \tag{4.56}$$

assim temos que

$$E(C_{j,k}^{dom}) = \frac{1-(\kappa_1+1)\left(\prod_{h=1}^{H}(1-p_k^h)^{N_{j,k}^h}\right)^{\kappa_1} + \kappa_1\left(\prod_{h=1}^{H}(1-p_k^h)^{N_{j,k}^h}\right)^{\kappa_1+1}}{\left(1-\left(\prod_{h=1}^{H}(1-p_k^h)^{N_{j,k}^h}\right)^{\kappa_1}\right)\left(1-\prod_{h=1}^{H}(1-p_k^h)^{N_{j,k}^h}\right)}. \tag{4.57}$$

Note que somente domicílios nos quais o entrevistador efetivamente conseguiu falar com alguma pessoa podem ser modelados assim. Para todos os outros domicílios, o número de contatos é dado por κ_1 com probabilidade 1. Já o número de contatos realizados até conseguir falar com uma pessoa selecionada do estrato h do conglomerado k, $C_{h,k}^{pes}$ segue uma distribuição

$$C_{h,k}^{pes} \sim GT\left(\kappa_2, p_k^h\right), \tag{4.58}$$

assim temos que

$$E(C_{h,k}^{pes}) = \frac{1-(\kappa_2+1)(1-p_k^h)^{\kappa_2} + \kappa_2(1-p_k^h)^{\kappa_2+1}}{(1-(1-p_k^h)^{\kappa_2})p_k^h}. \tag{4.59}$$

Novamente, somente as pessoas com as quais o entrevistador efetivamente conseguiu falar com alguma pessoa podem ser modelados assim. Para todas as outras pessoas, o número de contatos é dado por κ_2 com probabilidade 1.

Na tabela 4.1, resumimos o resultados 4.51, 4.53, 4.57 e 4.59 para o estrato h no conglomerado k. Note que o número de pessoas entrevistadas b_h é fixo, então o número de pessoas que foram buscadas porém não entrevistadas é obtido subtraindo essa quantidade do total de pessoas buscadas, ou seja, é dado por $E_(N_{h,k}^{pes}) - b_h$. Também foi utilizado o número médio de contatos esperados por domicílio, como forma de simplificar as contas e remover a dependência nos domicílios efetivamente abordados.

Se considerarmos somente as pessoas buscadas, temos que o total de contatos originados somente pela busca das pessoas no estrato h do conglomerado k pode ser escrito como:

Tabela 4.1: Número médio de Contatos Esperados

Origem	Tipo	Contatos Esperados	Unidades Buscadas Esperadas	Total Contatos Esperados
Domicílio	Fez Contato	$\sum_{j=1}^{D_k} \frac{E(C_{j,k}^{dom})}{D_k}$	$E_(N_{h,k}^{dom})$	$E_(N_{h,k}^{dom}) \sum_{j=1}^{D_k} \frac{E(C_{j,k}^{dom})}{D_k}$
Domicílio	Não Fez Contato	κ_1	$E_(\tilde{N}_{h,k}^{dom})$	$E_(\tilde{N}_{h,k}^{dom})\kappa_1$
Pessoa	Fez Contato	$E(C_{h,k}^{pes})$	b_h	$b_h E(C_{h,k}^{pes})$
Pessoa	Não Fez Contato	κ_2	$E_(N_{h,k}^{pes}) - b_h$	$\left(E_(N_{h,k}^{pes}) - b_h\right)\kappa_2$

$$\begin{aligned} E(Cont_{h,k}^{Pes}) &= \left(E_(N_{h,k}^{pes}) - b_h\right)\kappa_2 + b_h E(C_{h,k}^{pes}) \\ &= b_h \frac{\kappa_2 \left(1-p_k^h\right)^{\kappa_2} p_k^h + 1 - (\kappa_2+1)(1-p_k^h)^{\kappa_2} + \kappa_2(1-p_k^h)^{\kappa_2+1}}{(1-(1-p_k^h)^{\kappa_2})p_k^h} \\ &= \frac{b_h}{p_k^h}, \end{aligned} \tag{4.60}$$

ou seja, o número de contatos esperados com pessoas, denotado por $E(Cont_{h,k}^{Pes})$, é independente de κ_2. O que fica evidente é que apesar de κ_2 não controlar o total de contatos com pessoas, ele regula o que o pesquisador deseja: κ_2 **pequeno** - mais pessoas buscadas com menos voltas ou κ_2 **grande** - menos pessoas buscadas com mais voltas. Se uma opção for mais barata e/ou interessante que a outra, basta utilizar o κ_2 adequado. Por exemplo, a coleta de dados pode ser mais rápida se o pesquisador evitar totalmente as voltas, que são demoradas, e utilizar diversos entrevistadores ao mesmo tempo no conglomerado, já considerando a logística de coleta de dados o número de pessoas que serão abordadas e não-entrevistadas. Para isso, basta utilizar $\kappa_2 = 1$. Note que sempre existe um compromisso, pois ao reduzir o tamanho de κ_2 a qualidade da estimativa de p_k^h (que será discutida na Seção 4.3) poderá ser pior.

Resumindo, a média de contatos esperados para o estrato h no conglomerado k é dado por:

$$E(Cont_{h,k}) = E_(N_{h,k}^{dom}) \sum_{j=1}^{D_k} \frac{E(C_{j,k}^{dom})}{D_k} + E_(\tilde{N}_{h,k}^{dom})\kappa_1 + \frac{b_h}{p_k^h}, \tag{4.61}$$

e a média de contatos esperados para o conglomerado k é dado por:

$$E(Cont_k) = \sum_{h=1}^{H} E(Cont_{h,k}). \tag{4.62}$$

4.2.6 APV com Não-Resposta ignorando o modelo GRH (APVS)

O desenho amostral **APV** apresentado na Seção 4.2.4 foi desenvolvido para ser comparável com o desenho amostral **APC** apresentado na Seção 4.2.2. Apesar disso, na prática, usualmente utiliza-se um desenho amostral totalmente probabilístico mais simples, o qual ignora completamente o modelo **GRH** em 4.5, ou seja, supõem-se que as probabilidades de resposta são iguais para todas as pessoas do mesmo conglomerado. Nesta seção será desenvolvida a teoria para esse desenho amostral, que será denominado de **APV** Simples (**APVS**).

Ao invês de considerarmos H grupos, cada um com probabilidade de resposta p_k^h, esses grupos (estratos) na **APVS** serão ignorados, e vamos modelar todas as pessoas do conglomerado k com a mesma probabilidade de resposta, denotada por $\overline{p}_k$. Assim, as probabilidades utilizadas nessa seção não terão o índice h.

Seguindo o mesmo raciocínio da Seção 4.2.4, podemos escrever:

$$p_{i/k} = P_{D_{j[i],k}}(selec) * P_{D_{j[i],k}}(resp) * P_i(selec) * P_i(resp), \tag{4.63}$$

onde $P_{D_{j[i],k}}(selec)$ é a probabilidade de algum morador do domicílio $j[i]$ ser selecionado, $P_{D_{j[i],k}}(resp)$ é a probabilidade de algum morador do domicílio $j[i]$ responder dado que o domicílio $D_{j[i]}$ foi selecionado, $P_i(selec)$ é a probabilidade do morador i do domicílio $j[i]$ ser selecionado dado que o domicílio $D_{j[i]}$ foi selecionado e algum morador respondeu e $P_i(resp)$ é a probabilidade do morador i do domicílio $j[i]$ ser responder dado que esse morador foi selecionado e que o domicílio $D_{j[i]}$ foi selecionado e algum morador respondeu.

No caso **APVS**, temos que a probabilidade $P_{D_{j,k}}(resp)$ do entrevistador conseguir fazer contato com alguém do domicílio $j[i]$ do conglomerado k é dada por:

$$\begin{aligned} P_{D_{j,k}}(resp) &= \left(1-(1-\overline{p}_k)^{N_{j,k}}\right)\left(\sum_{i=0}^{\kappa_1-1}\left((1-\overline{p}_k)^{N_{j,k}}\right)^i\right) \\ &= \left(1-(1-\overline{p}_k)^{\kappa_1 N_{j,k}}\right) \end{aligned} \tag{4.64}$$

Usando exatamente o mesmo raciocínio, a probabilidade $P_i(resp)$ do entrevistador conseguir entrevistar o morador selecionado é dada por:

$$\begin{aligned} P_i(resp) &= \overline{p}_k\left(\sum_{i=0}^{\kappa_2-1}(1-\overline{p}_k)^i\right) \\ &= (1-(1-\overline{p}_k)^{\kappa_2}) \end{aligned} \tag{4.65}$$

Substituindo as probabilidades 4.64 e 4.65 em 4.63 obtemos que:

$$p_{i/k} = \frac{\left(1-(1-\overline{p}_k)^{\kappa_1 N_{j[i],k}}\right)}{D_k}\frac{(1-(1-\overline{p}_k)^{\kappa_2})}{N_{j[i],k}}, \tag{4.66}$$

Se desconsiderarmos a probabilidade de resposta fazendo $\overline{p}_k = 1$ obtemos que $p_{i/k}$ em 4.66 se reduz ao caso teórico em 4.67, bastante similar ao apresentado em 4.43, porém ignorando os grupos do **GRH**. Ou seja:

$$p_{i/k} = \frac{1}{D_k}\frac{1}{N_{j[i],k}}. \tag{4.67}$$

APVS com Não-Resposta - Tamanho da amostra fixo

Utilizando o mesmo raciocínio apresentado na Seção 4.2.4, precisamos inicialmente encontrar a probabilidade de se entrevistar alguma pessoa em uma única busca. Assim temos que, no conglomerado k essa probabilidade é dada por:

$$\begin{aligned} p^k_{entrev} &= \sum_{i=1}^{N_k} p_{i(j)/k} = \sum_{j=1}^{D_k} N_{j,k} * p_{i(j)/k} \\ &= (1-(1-\overline{p}_k)^{\kappa_2})\sum_{j=1}^{D_k}\frac{\left(1-(1-\overline{p}_k)^{\kappa_1 N_{j,k}}\right)}{D_k} \\ &= (1-(1-\overline{p}_k)^{\kappa_2})\frac{\left(D_k - \sum_{j=1}^{D_k}(1-\overline{p}_k)^{\kappa_1 N_{j,k}}\right)}{D_k}, \end{aligned} \tag{4.68}$$

onde $i(j)$ indica qualquer morador i que resida no domicílio j. Reparametrizando as probabilidades de seleção em 4.66 da mesma forma como em 4.49 obtemos as probabilidades de seleção $p_{i/k}$ para o caso da APVS com tamanho de amostra fixo:

$$\begin{aligned} p^{n_{fixo}}_{i/k} &= \frac{p_{i/k}}{p^k_{entrev}} \\ &= \frac{1}{N_{j[i],k}}\frac{\left(1-(1-\overline{p}_k)^{\kappa_1 N_{j[i],k}}\right)}{\left(D_k - \sum_{j=1}^{D_k}(1-\overline{p}_k)^{\kappa_1 N_{j,k}}\right)} \end{aligned} \tag{4.69}$$

APVS quando o modelo GRH está correto

Quando o modelo **GRH** está correto, ou seja, a população dentro do conglomerado realmente está dividida em H grupos, cada um com uma probabilidade de resposta diferente, as probabilidades

de seleção reais da **APVS** não são como em 4.69, pois ao calculá-las ignorou-se o modelo **GRH**. Além disso, como a **APVS** não é estratificada, não é possível utilizar as probabilidades de seleção em 4.49.

Nesse contexto, a probabilidade real de seleção é muito similar a probabilidade em 4.47, sendo dada por:

$$p_{i/k}^{REAL} = \frac{\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j[i],k}^h}\right)}{D_k}\frac{\left(1-\left(1-p_k^{h[i]}\right)^{\kappa_2}\right)}{N_{j[i],k}}, \tag{4.70}$$

onde $h[i]$ indica o estrato h ao qual a i-ésima unidade amostral realmente pertence. Reparametrizando as probabilidades em 4.70 como em 4.49 para encontrar as probabilidades de seleção reais para o caso da **APVS** com tamanho de amostra fixo, obtemos:

$$p_{i/k}^{REAL(n_{fixo})} = \frac{1}{N_{j[i],k}}\frac{\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j[i],k}^h}\right)\left(1-\left(1-p_k^{h[i]}\right)^{\kappa_2}\right)}{\sum_{a=1}^{N_{Ak}}\frac{\left(1-\prod_{h=1}^{H}(1-p_k^h)^{\kappa_1 N_{j[a],k}^h}\right)\left(1-\left(1-p_k^{h[a]}\right)^{\kappa_2}\right)}{N_{j[a],k}}}. \tag{4.71}$$

As probabilidades reais quando o modelo **GRH** está correto definidas em 4.71 são claramente diferentes daquelas utilizadas pela **APVS** em 4.69. Dessa forma, se utilizarmos o estimador $\hat{\tau}_k$ definido em 4.11 com as probabilidades "erradas", ele será viciado e terá a variância como segue:

$$E\left(\hat{\tau}_k\right) = \sum_{i=1}^{N_k}\frac{p_{i/k}^{REAL(n_{fixo})}}{p_{i/k}^{n_{fixo}}}Y_{ki} \qquad e \tag{4.72}$$

$$Var\left(\hat{\tau}_k\right) = \frac{1}{b}\left(\sum_{i=1}^{N_k}\frac{p_{i/k}^{REAL(n_{fixo})}}{\left(p_{i/k}^{n_{fixo}}\right)^2}Y_{ki}^2 - \left(\sum_{i=1}^{N}\frac{p_{i/k}^{REAL(n_{fixo})}}{p_{i/k}^{n_{fixo}}}Y_{ki}\right)^2\right). \tag{4.73}$$

Uma característica importante das probabilidades de seleção da **APVS** é que

$$\lim_{\substack{\kappa_1\to\infty\\ \kappa_2\to\infty}} p_{i/k}^{n_{fixo}} = \lim_{\substack{\kappa_1\to\infty\\ \kappa_2\to\infty}} p_{i/k}^{REAL(n_{fixo})}. \tag{4.74}$$

Ou seja, se o pesquisador optar por utilizar o desenho amostral APVS, uma forma dele se previnir contra a má-especificação do modelo de resposta (o modelo **GRH** estar correto) é aumentar o valor dos parâmetros κ_1 e κ_2.

Número de Contatos esperados para completar as entrevistas na APVS

Para calcular o número de contatos esperados no caso da **APVS**, utilizaremos o mesmo raciocínio

da Seção 4.2.5. Assim, é fácil ver que o número de domicílios abordados segue uma distribuição binomial negativa

$$N_k^{dom} \sim BN\left(b, p_{entrev}^k\right), \tag{4.75}$$

e consequentemente o número esperado de domicílios abordados do conglomerado k é dado por

$$E_(N_k^{dom}) = \frac{b}{p_{entrev}^k} = \frac{b}{(1-(1-\overline{p}_k)^{\kappa_2})\frac{\left(D_k - \sum_{j=1}^{D_k}(1-\overline{p}_k)^{\kappa_1 N_{j,k}}\right)}{D_k}}. \tag{4.76}$$

De maneira semelhante, também conseguimos modelar o número de pessoas abordadas com uma distribuição binomial negativa

$$N_k^{pes} \sim BN\left(b, 1-(1-\overline{p}_k)^{\kappa_2}\right), \tag{4.77}$$

e consequentemente o número esperado de pessoas abordadas do conglomerado k é dado por

$$E_(N_k^{pes}) = \frac{b}{1-(1-\overline{p}_k)^{\kappa_2}}. \tag{4.78}$$

De 4.75 e de 4.77 é facil ver que o número esperado de domicílios do conglomerado k onde o entrevistador não consegue fazer contato é dado por:

$$\begin{aligned} E_(\tilde{N}_k^{dom}) &= E(N_k^{dom}) - E_(N_k^{pes}) \\ &= \frac{b}{1-(1-\overline{p}_k)^{\kappa_2}}\left(\frac{1}{\frac{\left(D_k - \sum_{j=1}^{D_k}(1-\overline{p}_k)^{\kappa_1 N_{j,k}}\right)}{D_k}} - 1\right) \\ &= \frac{b}{1-(1-\overline{p}_k)^{\kappa_2}}\left(\frac{\left(\sum_{j=1}^{D_k}(1-\overline{p}_k)^{\kappa_1 N_{j,k}}\right)}{\left(D_k - (1-\overline{p}_k)^{\kappa_1 N_k}\right)}\right) \end{aligned} \tag{4.79}$$

Utilizando a distribuição definida em 4.2, o número de contatos realizados até conseguir falar com alguém do domicílio j do conglomerado k, $C_{j,k}^{dom}$ segue a distribuição:

$$C_{j,k}^{dom} \sim GT\left(\kappa_1, 1-(1-\overline{p}_k)^{N_{j,k}}\right), \tag{4.80}$$

assim temos que

Tabela 4.2: Número médio de Contatos Esperados

Origem	Tipo	Contatos Esperados	Unidades Buscadas Esperadas	Total Contatos Esperados
Domicílio	Fez Contato	$\sum_{j=1}^{D_k} \frac{E(C_{j,k}^{dom})}{D_k}$	$E_(N_k^{dom})$	$E_(N_k^{dom}) \sum_{j=1}^{D_k} \frac{E(C_{j,k}^{dom})}{D_k}$
Domicílio	Não Fez Contato	κ_1	$E_(\tilde{N}_k^{dom})$	$E_(\tilde{N}_k^{dom})\kappa_1$
Pessoa	Fez Contato	$E(C_k^{pes})$	b	$bE(C_k^{pes})$
Pessoa	Não Fez Contato	κ_2	$E_(N_k^{pes}) - b$	$\left(E_(N_k^{pes}) - b\right)\kappa_2$

$$E(C_{j,k}^{dom}) = \frac{1 - (\kappa_1 + 1)\left((1-\overline{p}_k)^{N_{j,k}}\right)^{\kappa_1} + \kappa_1\left((1-\overline{p}_k)^{N_{j,k}}\right)^{\kappa_1+1}}{\left(1 - \left((1-\overline{p}_k)^{N_{j,k}}\right)^{\kappa_1}\right)\left(1-(1-\overline{p}_k)^{N_{j,k}}\right)}. \tag{4.81}$$

Note que somente domicílios nos quais o entrevistador efetivamente conseguiu falar com alguma pessoa podem ser modelados assim. Para todos os outros domicílios, o número de contatos é dado por κ_1 com probabilidade 1. Já número de contatos realizados até conseguir falar com uma pessoa selecionada conglomerado k, C_k^{pes} segue uma distribuição

$$C_k^{pes} \sim GT\left(\kappa_2, \overline{p}_k\right), \tag{4.82}$$

assim temos que

$$E(C_k^{pes}) = \frac{1 - (\kappa_2 + 1)(1-\overline{p}_k)^{\kappa_2} + \kappa_2(1-\overline{p}_k)^{\kappa_2+1}}{(1-(1-\overline{p}_k)^{\kappa_2})\pi_k}. \tag{4.83}$$

Novamente, somente as pessoas com as quais o entrevistador efetivamente conseguiu falar com alguma pessoa podem ser modelados assim. Para todos as outras pessoas, o número de contatos é dado por κ_2 com probabilidade 1.

Na tabela 4.2, resumimos o resultados 4.76, 4.78, 4.81 e 4.83 para o conglomerado k. Note que o número de pessoas entrevistadas b é fixo, então o número de pessoas que foram buscadas porém não entrevistadas é obtido subtraindo-se essa quantidade do total de pessoas buscadas, ou seja, é dado por $E_(N_k^{pes}) - b$. Também foi utilizado o número médio de contatos por domicílio, de forma a simplificar as contas e remover a dependência do número de contatos total nos domicílios efetivamente abordados.

Se considerarmos somente as pessoas buscadas, temos que o total de contatos originados somente pela busca das pessoas no estrato h do conglomerado k pode ser escrito como:

$$\begin{aligned} E(Cont_k^{Pes}) &= \left(E_(N_k^{pes}) - b\right)\kappa_2 + bE(C_k^{pes}) \\ &= b\frac{\kappa_2\left(1-\overline{p}_k\right)^{\kappa_2}\overline{p}_k + 1 - (\kappa_2+1)(1-\overline{p}_k)^{\kappa_2} + \kappa_2(1-\overline{p}_k)^{\kappa_2+1}}{(1-(1-\overline{p}_k)^{\kappa_2})\overline{p}_k} \\ &= \frac{b}{\overline{p}_k}, \end{aligned} \tag{4.84}$$

ou seja, o número de contatos esperados com pessoas $E(Cont_k^{Pes})$ é independente de κ_2. Resumindo, a média de contatos esperados no conglomerado k é dado por:

$$E(Cont_k) = E_(N_k^{dom})\sum_{j=1}^{D_k}\frac{E(C_{j,k}^{dom})}{D_k} + E_(\tilde{N}_k^{dom})\kappa_1 + \frac{b}{\overline{p}_k}. \tag{4.85}$$

4.3 Inferência Incondicional - Estimando p_k^h

Na Seção 4.2 as probabilidades de resposta $\overline{p}_k$ e p_k^h foram tratadas como conhecidas. Ou seja, para utilizar os resultados da Seção 4.2 o estatísitico que quiser fazer inferência com os dados da pesquisa tem que conhecer essas quantidades. No geral, esse não é o caso, e essas quantidades são desconhecidas. Nessa seção discutiremos como estimar essas quantidades utilizando informações que podem ser facilmente obtidas durante a coleta de dados.

4.3.1 Impacto de estimar a probabilidade de resposta

Quando as probabilidades de resposta $\overline{p}_k$ e p_k^h não são conhecidas, o estimador utilizado para fazer inferência é um pouco diferente daquele definido em 4.8, pois além de estimar a quantidade populacional de interesse (nesse caso τ_y), também é necessário estimar as probabilidades de seleção $p_{hi/k}$, pois elas são uma função dessas probabilidades desconhecidas.

Nessa seção iremos trabalhar apenas com o estimador $\hat{\tau}_k^h = \sum_{i=1}^{N_k^h}\frac{Y_{khi}}{b_h p_{hi}}$ de τ_k^h, o total populacional da variável Y do estrato h do conglomerado k, definido em 4.11, pois essa é a única parte do estimador em 4.8 que depende das quantidades $\overline{p}_k$ ou p_k^h. Nessa seção, denotaremos as probabilidades $p_{hi/k}$ por p_{hi} para que as equações fiquem mais enxutas. Assim, estamos interessados nas quantidades:

$$\tau_k^h = \sum_{i=1}^{N_k^h} Y_{khi}, \qquad e \qquad V(\hat{\tau}_k^h) = V_{hk}^I = \sum_{i=1}^{N_k^h}\frac{p_{hi}}{b_h}\left(\frac{Y_{khi}}{p_{hi}} - \tau_k^h\right)^2 \tag{4.86}$$

e em seus estimadores, dados por:

$$\hat{\tau}_k^h = \sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h \hat{p}_{hi}}, \qquad e \qquad \hat{V}(\hat{\tau}_k^h) = \hat{V_{hk}^I} = \frac{1}{b_h(b_h-1)} \sum_{i \epsilon s_k^h} \left(\frac{Y_{khi}}{\hat{p}_{hi}} - \hat{\tau}_k^h \right)^2 \tag{4.87}$$

onde $\hat{p}_{hi} = g_{[hi]}\left(\hat{p}_k^h\right)$ para a **APV** e **APC** com as respectivas g's definidas por 4.49 e 4.29. Para o caso da **APVS** ignoram-se os estratos h em 4.87 e $\hat{p}_i = g_{[i]}\left(\hat{\overline{p}}_k\right)$ com g definida em 4.69. Não trabalharemos explicitamente o caso da **APVS** aqui, porém não é difícil obter resultados similares para **APVS** utilizando os resultados obtidos para a **APV**.

Utilizando as igualdades $E(X) = E_1\left(E_2(X/Y)\right)$ e $V(X) = E_1\left(V_2(X/Y)\right) + V_1\left(E_2(X/Y)\right)$ é possível calcular a esperança e a variância do estimador em 4.87, onde o índice 2 indica esperança e variância condicional a uma particular amostra e o índice 1 indica esperança e variância de todas as possíveis amostras. Assim obtemos que:

$$E(\hat{\tau}_k^h) = E_1\left(E_2\left(\sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h \hat{p}_{hi}}\right)\right) = E_1\left(\sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h} E_2\left(\frac{1}{\hat{p}_{hi}}\right)\right) = \sum_{i=1}^{N_k^h} Y_{khi} p_{hi} E_2\left(\frac{1}{\hat{p}_{hi}}\right) \tag{4.88}$$

Se o estimador de $\frac{1}{p_{hi}}$ for não-viciado, ou seja, $E_2\left(\frac{1}{\hat{p}_{hi}}\right) = \frac{1}{p_{hi}}$, então temos que $E(\hat{\tau}_k^h) = \tau_k^h$, e consequentemente também obtemos $E(\hat{\tau}_{HH}) = \tau_y$. Ou seja, mesmo com as probabilidades de resposta $\overline{p}_k$ e p_k^h desconhecidas, se o estimador utilizado for não-viciado para $\left(\frac{1}{\hat{p}_{hi}}\right)$ então a esperança do estimador $\tau_{\hat{HH}}$ se mantém a mesma.

Já para a variância do estimador obtemos o seguinte resultado:

$$\begin{aligned}
V(\hat{\tau}_k^h) &= E_1\left(V_2\left(\sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h \hat{p}_{hi}}\right)\right) + V_1\left(E_2\left(\sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h \hat{p}_{hi}}\right)\right) \\
&= E_1\left(\sum_{i \epsilon s_k^h} \frac{Y_{khi}^2}{b_h^2} V_2\left(\frac{1}{\hat{p}_{hi}}\right) + \sum_{i \epsilon s_k^h} \sum_{\substack{j \epsilon s_k^h \\ j \neq i}} \frac{Y_{khi} Y_{khj}}{b_h^2} Cov_2\left(\frac{1}{\hat{p}_{hi}}, \frac{1}{\hat{p}_{hj}}\right)\right) + V_1\left(\sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h} E_2\left(\frac{1}{\hat{p}_{hi}}\right)\right) \\
&= \sum_{i=1}^{N_k^h} \frac{Y_{khi}^2 p_{hi}}{b_h} V_2\left(\frac{1}{\hat{p}_{hi}}\right) - \sum_{i=1}^{N_k^h} \sum_{\substack{j=1 \\ j \neq i}}^{N_k^h} \frac{Y_{khi} Y_{khj} p_{hi} p_{hj}}{b_h} Cov_2\left(\frac{1}{\hat{p}_{hi}}, \frac{1}{\hat{p}_{hj}}\right) \\
&+ \sum_{i=1}^{N_k^h} \frac{Y_{khi}^2 p_{hi}(1-p_{hi})}{b_h} \left(E_2\left(\frac{1}{\hat{p}_{hi}}\right)\right)^2 - \sum_{i=1}^{N_k^h} \sum_{\substack{j=1 \\ j \neq i}}^{N_k^h} \frac{Y_{khi} Y_{khj} p_{hi} p_{hj}}{b_h} E_2\left(\frac{1}{\hat{p}_{hi}}\right) E_2\left(\frac{1}{\hat{p}_{hj}}\right) \\
&= I_{11} - I_{12} + I_{21} - I_{22} = I_1 + I_2.
\end{aligned} \tag{4.89}$$

Novamente, se o estimador de $\frac{1}{p_{hi}}$ for não-viciado, então temos que a parte I_2 (4.89) da variância

de $\hat{\tau}_k^h$ é igual a

$$I_2 = \sum_{i=1}^{N_k^h} \frac{Y_{khi}^2 p_{hi}(1-p_{hi})}{b_h} \left(E_2 \left(\frac{1}{\hat{p}_{hi}} \right) \right)^2 - \sum_{i=1}^{N_k^h} \sum_{\substack{j=1 \\ j \neq i}}^{N_k^h} \frac{Y_{khi} Y_{khj} p_{hi} p_{hj}}{b_h} E_2 \left(\frac{1}{\hat{p}_{hi}} \right) E_2 \left(\frac{1}{\hat{p}_{hj}} \right)$$

$$= \sum_{i=1}^{N_k^h} \frac{Y_{khi}^2 (1-p_{hi})}{b_h p_{hi}} - \sum_{i=1}^{N_k^h} \sum_{\substack{j=1 \\ j \neq i}}^{N_k^h} \frac{Y_{khi} Y_{khj}}{b_h} = \sum_{i=1}^{N_k^h} \frac{p_{hi}}{b_h} \left(\frac{Y_{khi}}{p_{hi}} - \tau_k^h \right)^2 = V(\hat{\tau}_k^h) = V_{hk}^I, \quad (4.90)$$

ou seja, é idêntica a variância V_{hk}^I obtida em 4.9 condicionada ao conhecimento das probabilidades de seleção. Ou seja, a parcela I_2 pode ser estimada utilizando o estimador $\hat{V}_{hk}^2$ apresentado em 4.87.

Para simplicar a parcela I_1, que está relacionada com a variância do estimador da quantidade p_{hi}, é necessário especificar qual tipo de estimador será utilizado. Na Seção 4.3.2 apresentaremos diferentes estimadores $\hat{p}_{hi}$, para os casos da **APC**, da **APV** e da **APVS**. Todos os estimadores que serão apresentados são estimadores de máxima-verossimilhança (EMV). Os estimadores pertencentes a essa classe têm propriedades importantes, uma das quais será apresentada sem prova, a seguir. As condições de regularidade que garantem esses resultados podem ser encontradas em Lehmann and Casella [1998].

Definição 4.3 (Distribuição Assintótica do EMV de θ) *Se a densidade $f(x/\theta)$ satisfaz certas condições de regularidade e se $\hat{\theta}_n$ é o estimador de máxima-verossimilhança de $\tau(\theta)$ para uma amostra aleatória simples de tamanho n de $f(x/\theta)$, então:*

$$\hat{\theta}_n \sim \mathcal{N}\left(\tau(\theta), \frac{(\tau'(\theta))^2}{-nE\left[I(\theta)\right]} \right), \quad (4.91)$$

para um n suficientemente grande, onde $Z \sim \mathcal{N}(a,b)$ indica que a variável aleatória Z têm uma distribuição normal com média a e variância b e $I(\theta) = \frac{\partial^2}{\partial \theta^2} log f(X/\theta)$ é a matriz de informação de Fisher.

No contexto desse capítulo, estamos interessados em estimar a função $g(\theta)$. Apresentaremos abaixo um resultado importante, conhecido como método Delta, que permite encontrar a distribuição assintótica de $g(\hat{\theta}_n)$ se a função $g(\theta)$ satisfizer algumas condições de regularidade que também podem ser encontradas em Lehmann and Casella [1998].

Definição 4.4 (Método Delta) *Se o estimador $\hat{\theta}_n$ de θ tem distribuição assintótica dada por $Z \sim \mathcal{N}(\theta, \sigma^2)$, e seja $g(\theta)$ uma função que satisfaz certas condições de regularidade, então:*

$$g(\hat{\theta}_n) \sim \mathcal{N}\left(g(\theta), \left(\frac{\partial g}{\partial \theta}\right)^2 \sigma^2\right). \tag{4.92}$$

O método Delta também pode ser enunciado para o caso multivariado. Nesta seção estamos interessados somente na $Cov_2(\frac{1}{\hat{p}_{hi}}, \frac{1}{\hat{p}_{hj}})$, encontrada em 4.89, apresentaremos a seguir um resultdo somente da covariância.

Definição 4.5 (Covariância do Método Delta Multivariado) *Seja o estimador $\hat{\theta}_n$ de θ com distribuição assintótica dada por $Z \sim \mathcal{N}(\mu, \sigma^2)$, e sejam $g_1(\theta)$ e $g_2(\theta)$ funções que satisfazem certas condições de regularidade, então:*

$$Cov\left(g_1(\hat{\theta}_n), g_2(\hat{\theta}_n)\right) \approx \frac{\partial g_1}{\partial \theta} \sigma^2 \frac{\partial g_2}{\partial \theta}. \tag{4.93}$$

Assim, utilizando os resultados dados nas definições 4.4 e 4.5, supondo um b_h suficientemente grande, se $\hat{p}_k^h$ for um estimador de máxima-verossmilhança de p_k^h então temos que:

$$\begin{aligned}
E_2\left(\frac{1}{\hat{p}_{hi}}\right) &= E_2\left(g_{[hi]}\left(\hat{p}_k^h\right)\right) \approx g_{[hi]}\left(p_k^h\right) = \frac{1}{p_{hi}}, \\
V_2\left(\frac{1}{\hat{p}_{hi}}\right) &= V_2\left(g_{[hi]}\left(\hat{p}_k^h\right)\right) \approx \left(\frac{\partial g_{[hi]}}{\partial p_k^h}\right)^2 V_2\left(\hat{p}_k^h\right), \\
Cov_2\left(\frac{1}{\hat{p}_{hi}}, \frac{1}{\hat{p}_{hj}}\right) &= Cov_2\left(g_{[hi]}\left(\hat{p}_k^h\right), g_{[hj]}\left(\hat{p}_k^h\right)\right) \approx \left(\frac{\partial g_{[hi]}}{\partial p_k^h}\right) V_2\left(\hat{p}_k^h\right) \left(\frac{\partial g_{[hj]}}{\partial p_k^h}\right),
\end{aligned} \tag{4.94}$$

onde $g_{[hi]}(\hat{p}_k^h)$ indica a função $g(.)$ referente a unidade populacional i do estrato h e $V_2\left(\hat{p}_k^h\right)$ é a variância do estimador de p_k^h que será calculada na Seção 4.3.2.

Utilizando os resultados em 4.94 é possível simplificar o termo I_1 em 4.89, lembrando que estamos supondo que b_h seja suficientemente grande e que as condições de refgularidade das funções $g_{[hi]}$ sejam satisfeitas. Assim, substituindo esses resultados em 4.89 obtemos:

$$
\begin{aligned}
I_1 &= \sum_{i=1}^{N_k^h} \frac{Y_{khi}^2 p_{hi}}{b_h} V_2\left(\frac{1}{\hat{p}_{hi}}\right) - \sum_{i=1}^{N_k^h} \sum_{\substack{j=1 \\ j \neq i}}^{N_k^h} \frac{Y_{khi} Y_{khj} p_{hi} p_{hj}}{b_h} Cov_2\left(\frac{1}{\hat{p}_{hi}}, \frac{1}{\hat{p}_{hj}}\right) \\
&= \sum_{i=1}^{N_k^h} \frac{Y_{khi}^2 p_{hi}}{b_h} \left(\frac{\partial g_{[hi]}}{\partial p_k^h}\right)^2 V_2\left(\hat{p}_k^h\right) \\
&- \sum_{i=1}^{N_k^h} \sum_{\substack{j=1 \\ j \neq i}}^{N_k^h} \frac{Y_{khi} Y_{khj} p_{hi} p_{hj}}{b_h} \left(\frac{\partial g_{[hi]}}{\partial p_k^h}\right) V_2\left(\hat{p}_k^h\right) \left(\frac{\partial g_{[hj]}}{\partial p_k^h}\right) \\
&= \frac{V_2\left(\hat{p}_k^h\right)}{b_h} \left(\sum_{i=1}^{N_k^h} Y_{khi}^2 p_{hi}(1+p_{hi}) \left(\frac{\partial g_{[hi]}}{\partial p_k^h}\right)^2 - \left(\sum_{i=1}^{N_k^h} Y_{khi} p_{hi} \frac{\partial g_{[hi]}}{\partial p_k^h} \right)^2 \right) \\
&= \frac{V_2\left(\hat{p}_k^h\right)}{b_h} \left(T_1^{h,k} - \left(T_2^{h,k}\right)^2 \right). \qquad (4.95)
\end{aligned}
$$

Fica bastante evidente com a simplificação de I_1 dada em 4.95 que podemos interpretar a parcela I_1 como sendo o acréscimo na variância de $\hat{\tau}_k^h$ pela necessidade de estimar as probabilidades de seleção. Assim, podemos re-escrever a variância de $\hat{\tau}_{HH}$ em 4.9 como sendo:

$$
\begin{aligned}
Var\left(\hat{\tau}_{HH}\right) &= \sum_{k=1}^{A} \sum_{k'=1}^{A} V_{kk'}^E + \sum_{k=1}^{A} \sum_{h=1}^{H} \frac{V(\hat{\tau}_k^h)}{\pi_k} \\
&= \sum_{k=1}^{A} \sum_{k'=1}^{A} V_{kk'}^E + \sum_{k=1}^{A} \sum_{h=1}^{H} \frac{V_{hk}^I + \frac{V_2\left(\hat{p}_k^h\right)}{b_h} \left(T_1^{h,k} - \left(T_2^{h,k}\right)^2 \right)}{\pi_k} \\
&= \sum_{k=1}^{A} \sum_{k'=1}^{A} V_{kk'}^E + \sum_{k=1}^{A} \sum_{h=1}^{H} \frac{V_{hk}^I}{\pi_k} + \sum_{k=1}^{A} \sum_{h=1}^{H} \frac{V_2\left(\hat{p}_k^h\right) \left(T_1^{h,k} - \left(T_2^{h,k}\right)^2 \right)}{b_h \pi_k} \qquad (4.96)
\end{aligned}
$$

É importante encontrar um estimador não-viciado para estimar a variância de $\hat{\tau}_{HH}$. Utilizando o estimador apresentado em 4.15, ainda precisamos obter estimadores para as quantidades $V_2\left(\hat{p}_k^h\right)$, $T_1^{h,k}$ e $T_2^{h,k}$. Diferentes estimadores para a quantidade $V_2\left(\hat{p}_k^h\right)$ serão apresentados na Seção 4.3.2. Para os termos $T_1^{h,k}$ e $T_2^{h,k}$, os seguintes estimadores são considerados:

$$
\hat{T}_1^{h,k} = \sum_{i \epsilon s_k^h} \frac{Y_{khi}^2}{b_h} (1 + \hat{p_k^h}) \left(\left. \frac{\partial g_{[hi]}}{\partial p_k^h} \right|_{p_k^h = \hat{p_k^h}} \right)^2 \quad \text{e} \quad \hat{T}_2^{h,k} = \sum_{i \epsilon s_k^h} \frac{Y_{khi}}{b_h} \left. \frac{\partial g_{[hi]}}{\partial p_k^h} \right|_{p_k^h = \hat{p_k^h}}. \qquad (4.97)
$$

Novamente é importante ressaltar as quantidades $\frac{\partial g_{[hi]}}{\partial p_k^h}$ dependem de todas as quantidades $N_{j,k}^h$, que são desconhecidas, essa questão será discutida na Seção 4.4. Um estimador para a variância de $\hat{\tau}_{HH}$ é dado por:

$$\begin{aligned} \hat{Var}\left(\hat{\tau}_{HH}\right) &= \sum_{k\epsilon s_I}\sum_{k'\epsilon s_I}\left(\frac{\pi_{kk'}-\pi_k\pi_{k'}}{\pi_{kk'}}\right)\frac{\hat{\tau}_k}{\pi_k}\frac{\hat{\tau}_{k'}}{\pi_{k'}} + \sum_{k\epsilon s_I}\sum_{h=1}^{H}\frac{\hat{V_{hk}^I}}{\pi_k} \\ &+ \sum_{k\epsilon s_I}\sum_{h=1}^{H}\frac{\hat{V}\left(\hat{p}_k^h\right)\left(\hat{T}_1^{h,k}-\left(\hat{T}_2^{h,k}\right)^2\right)}{b_h\pi_k}. \end{aligned} \tag{4.98}$$

4.3.2 Estimando a probabilidade de resposta

Nesta seção, o objetivo é encontrar estimadores da probabilidade de resposta $\hat{p}_k^h$ e da sua variância $V_2\left(\hat{p}_k^h\right)$, para os desenhos amostras **APC**, **APV** e **APVS**.

Esses estimadores não dependerão dos valores da variável Y das pessoas pertencentes a amostra, e sim de quantidades que serão controladas durante o processo de coleta de dados. É importante que essas quantidades sejam facilmente registradas pelo entrevistador, caso contrário essas estimativas podem se tornar bem caras e demoradas de serem obtidas.

Os estimadores desenvolvidos nessa seção serão funções das quantidades abaixo, onde o estimador da **APC** utilizará a quantidade Dom_k^h e os estimadores da **APV** e **APVS** utilizarão as quantidades $Cont_k^h$ e $Pess_k^h$:

Contagem de Contatos Contagem de quantos contatos foram realizados até o entrevistador completar todas as entrevistas (incluindo o contato realizado com quem foi entrevistado) no conglomerado k no estrato h, somente considerando os contatos realizados para entrevistar a pessoa selecionada, e não o domicílio. Essa quantidade será denotada por $Cont_k^h$.

Contagem de Pessoas Contagem de quantas pessoas foram contactadas até o entrevistador completar todas as entrevistas (incluindo a pessoa que foi entrevistada) no conglomerado k no estrato h, somente considerando as pessoas contactadas para entrevistar a pessoa selecionada, e não o domicílio. Essa quantidade será denotada por $Pess_k^h$.

Contagem de Domicílios Contagem de quantos domicílios nos quais o entrevistador tentou fazer contato **porém não conseguiu realizar um entrevista** até completar todas as entrevistas, no conglomerado k no estrato h. Essa quantidade será denotada por Dom_k^h.

Estimando a probabilidade de resposta na APC

A maior dificuldade em estimar $\hat{p}_k^h$ na **APC** ocorre pela falta de informação, pois não conhecemos $N_{j,k}^h$ para os domicílios não pertencentes a amostra. Se essas quantidades fossem conhecidas poderíamos escrever:

$$L\left(p_k^h/Dom_k^h\right) = \prod_{i=1}^{b_h}\left(1-\left(1-p_k^h\right)^{N_{j[i],k}^h}\right)\prod_{a=1}^{Dom_k^{h,i}}\left(1-p_k^h\right)^{N_{j[i]-a,k}^h}, \quad (4.99)$$

onde $j[i]$ indica o índice do domicílio onde a i-ésima pessoa reside e $Dom_k^{h,i}$ indica quantos domicílios o entrevistador não conseguiu entrevistar até conseguir entrevistar a i-ésima pessoa do estrato h do conglomerado k. Essa verossimilhança pode ser simplificada, não sendo necessário conhecer $Dom_k^{h,i}$ para todas as pessoas pertencentes a amostra:

$$L\left(p_k^h/Dom_k^h, n_z\right) = \prod_{i=1}^{b_h}\left(1-\left(1-p_k^h\right)^{N_{j[i],k}^h}\right)\prod_{z=0}^{n_{max}}\left(1-p_k^h\right)^{zn_z}, \quad (4.100)$$

onde n_z é o número de domicílios nos quais o entrevistador tentou mas não conseguiu entrevistador um morador do estrato h nos quais residiam z moradores do estrato h, com $\sum_{z=0}^{n_{max}} n_z = Dom_k^h$ e n_{max} é o número máximo de moradores do do estrato h em um único domicílio.

Procedendo da maneira usual para se obter um estimador de máxima-verossimilhança, derivando $\log L\left(p_k^h/Dom_k^h, n_z\right)$ com relação a p_k^h e igualando a zero, obtemos:

$$\frac{\partial l}{\partial p_k^h} = \sum_{i=1}^{b_h}\frac{N_{j[i],k}^h\left(1-p_k^h\right)^{N_{j[i],k}^h-1}}{1-\left(1-p_k^h\right)^{N_{j[i],k}^h}} - \frac{\sum_{z=0}^{n_{max}} zn_z}{1-p_k^h} = 0, \quad (4.101)$$

de onde o estimador $\hat{p}_k^h$ é obtido implicitamente através de:

$$\hat{p}_k^h = \arg\min_{0<p_k^h<1}\left|\sum_{i=1}^{b_h}\frac{N_{j[i],k}^h\left(1-p_k^h\right)^{N_{j[i],k}^h-1}}{1-\left(1-p_k^h\right)^{N_{j[i],k}^h}} - \frac{\sum_{z=0}^{n_{max}} zn_z}{1-p_k^h}\right|. \quad (4.102)$$

O problema é que as quantidades n_z são desconhecidas, ou seja, temos um caso de dados faltantes. Usualmente nesse cenário, utiliza-se um algoritmo iterativo conhecido como EM (Expectation-Maximization) para se estimar p_k^h. Esse método recebe esse nome pois ele é composto de dois passos básicos, E e M, que são repetidos até a convegência do algoritmo :

E - Expectation Calcula-se a esperança com relação a $Z/\theta^{(i)}$ da log-verossimilhança conjunta de Y e Z dada por $Q(\theta/\theta^{(i)}) = E_{Z/\theta^{(i)}}(\log L(\theta/Y,Z))$.

M - Maximization Encontra-se a estimativa $\theta^{(i+1)}$ que maximiza $Q(\theta/\theta^{(i)})$, usualmente obtida resolvendo $\left.\frac{\partial Q(\theta/\theta^{(i)})}{\partial\theta}\right|_{\theta=\theta^{(i+1)}} = 0$,

onde θ é o parâmetro de interesse, $\theta^{(i)}$ é a estimativa do parâmetro θ na iteração i, Y representa os dados originais (nesse caso Dom_k^h) e Z representa variáveis latentes (não-observadas - nesse caso

as variáveis n_z) que auxiliam na estimativa da quantidade de interesse. Mais detalhes sobre o algoritmo EM podem ser encontrados em Dempster et al. [1977] e Tanner [1996].

Para utilizar o algoritmo EM, é preciso especificar uma distribuição para as quantidades n_z. Uma distribuição bastante razoável para n_z é a distribuição Binomial, pois percebe-se que quanto menor um domicílio, maior a sua probabilidade de não pertencer a amostra:

$$n_z \sim Bin\left(Dom_k^h; \frac{\left(1-p_k^h\right)^z}{\sum_{a=0}^{n_{max}}\left(1-p_k^h\right)^a}\right) \qquad com \qquad E(n_z) = Dom_k^h \frac{\left(1-p_k^h\right)^z}{\sum_{a=0}^{n_{max}}\left(1-p_k^h\right)^a}. \quad (4.103)$$

Utilizando a distribuição das quantidades n_z, é possível mostrar, no contexto dessa seção, que a etapa E é dada por

$$\begin{aligned}
Q(p_k^h/p_k^{h(i)}) &= E_{n_z/p_k^{h(i)}}\left(L\left(p_k^h/Dom_k^h, n_z\right)\right) \\
&= \sum_{i=1}^{b_h} \log\left(1-\left(1-p_k^h\right)^{N_{j[i],k}^h}\right) + \sum_{z=0}^{n_{max}} zE_{n_z/p_k^{h(i)}}(n_z)\log\left(1-p_k^h\right) \\
&= \sum_{i=1}^{b_h} \log\left(1-\left(1-p_k^h\right)^{N_{j[i],k}^h}\right) + Dom_k^h \log\left(1-p_k^h\right)\frac{\sum_{z=0}^{n_{max}} z\left(1-p_k^{h(i)}\right)^z}{\sum_{a=0}^{n_{max}}\left(1-p_k^{h(i)}\right)^a} \\
&= \sum_{i=1}^{b_h} \log\left(1-\left(1-p_k^h\right)^{N_{j[i],k}^h}\right) + \log\left(1-p_k^h\right)\sum_{z=0}^{n_{max}} zn_z^{(i)},
\end{aligned}$$

onde $n_z^{(i)} = Dom_k^h \frac{\left(1-p_k^{h(i)}\right)^z}{\sum_{a=0}^{n_{max}}\left(1-p_k^{h(i)}\right)^a}$. Também podemos mostrar, utilizando o resultado em 4.102, que a etapa M é dada por:

$$p_k^{h(i+1)} = \arg\min_{0<p_k^h<1}\left|\sum_{i=1}^{b_h}\frac{N_{j[i],k}^h\left(1-p_k^h\right)^{N_{j[i],k}^h-1}}{1-\left(1-p_k^h\right)^{N_{j[i],k}^h}} - \frac{\sum_{z=0}^{n_{max}} zn_z^{(i)}}{1-p_k^h}\right|, \quad (4.104)$$

Assim, como o algoritmo EM é iterativo, o estimador $\hat{p}_k^h$ é obtido quando o algoritmo convergir. O critério de convergência mais usual é assumir que o algoritmo convergiu na iteração i^* se $|p_k^{h(i^*)} - p_k^{h(i^*-1)}| < \varepsilon$, para um ε pequeno. Nesse caso, temos que $\hat{p}_k^h = p_k^{h(i^*)}$.

Note que o estimador $\hat{p}_k^h$ obtido pelo algoritmo EM é um EMV, porém para encontrar a sua variância não basta utilizar o resultado apresentado na definição 4.3, pois é necessário levar em consideração que parte dos dados utilizados nesse estimador provém dos dados faltantes (variáveis n_z). Para obter um estimador de $V_2(\hat{p}_k^h)$, iremos utilizar o método de Louis. Mais detalhes podem ser encontrados em Tanner [1996] e Louis [1982].

O estimador $\hat{p}_k^h = p_k^{h(i^*)}$ foi encontrado utilizando a verossimilhança $L\left(p_k^h/Dom_k^h, n_z\right)$ que inclue

os dados faltantes n_z. De maneira mais geral, podemos dizer que o estimador foi obtido utlizando a verossimilhança $L(\theta/Y,Z)$, que também é conhecida como a verossimilhança dos dados completados, porém o interesse está em obter a variância do estimador derivado da verossimilhança observada $L(\theta/Y)$, ou seja, $L\left(p_k^h/Dom_k^h\right)$. Em Louis [1982] o autor mostrou que a variância do estimador $\hat{\theta}$ pode ser obtida da seguinte igualdade:

$$\begin{aligned} V(\hat{\theta}) &= -\frac{\partial^2}{\partial\theta^2}\log L(\theta/Y)\Big|_{\theta=\hat{\theta}} \\ &= -E_Z\left(\frac{\partial^2}{\partial\theta^2}\log L(\theta/Y,Z)\Big|_{\theta=\hat{\theta}}\right) - V_Z\left(\frac{\partial}{\partial\theta}\log L(\theta/Y,Z)\Big|_{\theta=\hat{\theta}}\right). \end{aligned} \tag{4.105}$$

Assim, no caso discutido aqui, podemos estimar $\hat{Var}_2(\hat{p}_k^h)$ utilizando:

$$\begin{aligned} \hat{Var}_2(\hat{p}_k^h) = &- E_{n_z/p_k^{h(i)}}\left(\frac{\partial^2}{\partial\left(p_k^h\right)^2}\log L\left(p_k^h/Dom_k^h,n_z\right)\Big|_{p_k^h=\hat{p}_k^h}\right) \\ &- V_{n_z/p_k^{h(i)}}\left(\frac{\partial}{\partial p_k^h}\log L\left(p_k^h/Dom_k^h,n_z\right)\Big|_{p_k^h=\hat{p}_k^h}\right) \\ &= -E_{EM} - V_{EM} \end{aligned} \tag{4.106}$$

onde as quantidades E_{EM} e V_{EM} são dadas por:

$$\begin{aligned} E_{EM} &= \sum_{i=1}^{b_h}\frac{N_{j[i],k}^h\left(N_{j[i],k}^h-1\right)\left(1-\hat{p}_k^h\right)^{N_{j[i],k}^h-2}}{1-\left(1-\hat{p}_k^h\right)^{N_{j[i],k}^h}} + \sum_{i=1}^{b_h}\frac{\left(N_{j[i],k}^h\left(1-\hat{p}_k^h\right)^{N_{j[i],k}^h-1}\right)^2}{\left(1-\left(1-\hat{p}_k^h\right)^{N_{j[i],k}^h}\right)^2} \\ &- \sum_{z=0}^{n_{max}}\frac{z}{\left(1-\hat{p}_k^h\right)^2}\frac{Dom_k^h\left(1-\hat{p}_k^h\right)^z}{\sum_{a=0}^{n_{max}}\left(1-\hat{p}_k^h\right)^a} \end{aligned} \tag{4.107}$$

$$\begin{aligned} V_{EM} &= \sum_{z=0}^{n_{max}}\frac{z^2Dom_k^h}{\left(1-\hat{p}_k^h\right)^2}\frac{\left(1-\hat{p}_k^h\right)^z\left(1-\left(1-\hat{p}_k^h\right)^z\right)}{\left(\sum_{a=0}^{n_{max}}\left(1-\hat{p}_k^h\right)^a\right)^2} \\ &- \sum_{z=0}^{n_{max}}\sum_{\substack{z'=0\\z'\neq z}}^{n_{max}}\frac{zz'Dom_k^h}{\left(1-\hat{p}_k^h\right)^2}\frac{\left(1-\hat{p}_k^h\right)^z\left(1-\hat{p}_k^h\right)^{z'}}{\left(\sum_{a=0}^{n_{max}}\left(1-\hat{p}_k^h\right)^a\right)^2} \end{aligned} \tag{4.108}$$

Estimando a probabilidade de resposta na APV e APVS

No caso da amostragem probabilística, é mais fácil encontrar diferentes estimadores para as quantidades de interesse, se ignorarmos os contatos realizados antes de se fazer contato com o

domicílio selecionado. Isso ocorre pois, nesse contexto, todas as pessoas do mesmo conglomerado e do mesmo estrato têm a mesma probabilidade de resposta, tornando a amostra que será utilizada para estimar p_k^h uma Amostra Aleatória Simples (**AAS**), a qual foi apresentada na Seção 1.2.1. Aqui não existe a necesssidade de diferenciar amostragem com e sem reposição.

Foram derivados os seguintes estimadores para a amostragem probabilística:

BN - Binomial Negativa Estimador de Máxima-Verossimilhança obtido supondo que o número de pessoas contactadas até se obter b_h entrevistas (sucessos) segue uma distribuição Binomial Negativa.

GTS - Geométrica Truncada Simplificada Estimador simplificado do Método dos Momentos, obtido supondo que o número de contatos necessários até se obter uma entrevista segue uma distribuição Geométrica Truncada.

GT - Geométrica Truncada Estimador de Máxima-Verossimilhança, obtido supondo que o número de contatos necessários até se obter uma entrevista segue uma distribuição Geométrica Truncada.

C - Combinado Estimador de Máxima-Verossimilhança, obtido combinando o número de contatos realizados e o número de pessoas contactadas até se completar b_h entrevistas.

Os estimadores BN, GTS e GT já foram estudados na literatura, supondo funções de verossimilhança conhecidas, então apenas apresentaremos aqui os respectivos estimadores $\hat{p}_k^h$ e $\hat{V}\left(\hat{p}_k^h\right)$. Já o estimador C (Combinado) será derivado a seguir. Note que usualmente o tamanho da amostra é pequeno dentro da estrato h do conglomerado k, assim uma característica desejável para $\hat{p}_k^h$ seria que, de alguma forma, ela inflaciona-se o tamanho da amostra. O estimador BN utiliza somente informações de pessoas contactadas ($Pess_{i,k}^h$), enquanto os estimadores GTS e GT utilizam somente as informações de contatos realizados ($Cont_{i,k}^h$). Na Seção 4.2.5 vimos que se escolhermos κ_2 **pequeno** - mais pessoas serão contactadas com menos contatos ou κ_2 **grande** - menos pessoas serão contactadas com mais contatos, assim dependendo do κ_2 escolhido, a performance desses estimadores será afetada, pois esse parâmetro determina diretamente o tamanho da amostra utilizada por cada um desses estimadores.

O estimador C foi derivado justamente para "inflar"o tamanho da amostra, levando em consideração tanto as pessoas contactadas e quanto os contatos realizados. Para derivar o estimador C, foi necessário criar uma função de verossimilhança $L\left(p_k^h/.\right)$ onde ambas as quantidades $Cont_k^h$ e $Pess_k^h$ estivessem presentes. Pensando no processo de obtenção da amostra, podemos escrever que para cada pessoa i entrevistada, o entrevistador tentou entrevistar $(Pess_{i,k}^h - 1)$ pessoas sem sucesso, e também que para efetivamente entrevistar a pessoa i, ele tentou fazer $(Cont_{i,k}^h - 1)$ contatos com o entrevistado. Pensando dessa forma, podemos escrever a verossimilhança como:

Tabela 4.3: Estimadores de p_k^h e $V_2\left(p_k^h\right)$ para **APV** e **APVS**

Tipo	Verossimilhança	$\hat{p}_k^h$	$\hat{V}\left(\hat{p}_k^h\right)$
BN	$BN\left(b_h, 1-\left(1-p_k^h\right)^{\kappa_2}\right)$	$1-\left(\left(1-\frac{b_h}{pess}\right)^{\frac{1}{\kappa_2}}\right)$	$\frac{\frac{b_h}{\hat{p}}\left(\frac{1}{\hat{p}}-\frac{1}{(1-\hat{p})^2}\right)}{\kappa_2\left(1-\hat{p}_k^h\right)^{2(\kappa_2-1)}}$
GTS	$GT\left(\kappa_2, p_k^h\right)$	$\frac{\left((\kappa_2+1)b_h-2b_h\overline{cont}\right)}{\left(\kappa_2 b_h\overline{cont}-\sum_{i=1}^{b_h} cont_i^2\right)}$	ver Thomasson and Kapadia
GT	$GT\left(\kappa_2, p_k^h\right)$	$\arg\min_{p_k^h} \frac{1-(\kappa_2+1)(1-p_k^h)^{\kappa_2}+\kappa_2(1-p_k^h)^{\kappa_2+1}}{1-(1-p_k^h)-(1-p_k^h)^{\kappa_2}+(1-p_k^h)^{\kappa_2+1}}$	ver Thomasson and Kapadia
C	$L\left(p_k^h/Cont_k^h, Pess_k^h\right)$	$\frac{b_h}{\kappa_2(pess-b_h)+cont}$	$\left(\frac{b_h}{(\hat{p}_k^h)^2}-\frac{\kappa_2(pess-b_h)+cont-b_h}{(1-\hat{p}_k^h)^2}\right)^{-1}$

$$
\begin{aligned}
L\left(p_k^h/Cont_k^h, Pess_k^h\right) &= \prod_{i=1}^{b_h}\left[\left(1-p_k^h\right)^{\kappa_2}\right]^{Pess_{i,k}^h-1} p_k^h\left(1-p_k^h\right)^{Cont_{i,k}^h-1} \\
&= \left(1-p_k^h\right)^{\kappa_2\left(Pess_k^h-b_h\right)}\left(p_k^h\right)^{b_h}\left(1-p_k^h\right)^{Cont_k^h-b_h}. \qquad (4.109)
\end{aligned}
$$

Derivando $\log L\left(p_k^h/Cont_k^h, Pess_k^h\right)$ com relação a p_k^h e igualando a zero, é fácil mostrar que $\hat{p}_k^h = \frac{b_h}{\kappa_2(Pess_k^h-b_h)+Cont_k^h}$. Para encontrar um estimativa para a variância do estimador C, é preciso calcular $\frac{\partial^2 l}{\partial\left(p_k^h\right)^2}$, a segunda derivada de $\log L\left(p_k^h/Cont_k^h, Pess_k^h\right)$:

$$
\frac{\partial^2 l}{\partial\left(p_k^h\right)^2} = \frac{\kappa_2(Pess_k^h-b_h)+Cont_k^h-b_h}{(1-p_k^h)^2}-\frac{b_h}{(p_k^h)^2}. \qquad (4.110)
$$

Existem maneiras diferentes de se encontrar um estimador para a variância de um estimador de máxima-verossimilhança, uma é utilizando a informação de Fisher esperada e a outra utilizando a informação de Fisher observada. Em Efron and Hinkley [1978], os autores recomendam utilizar a observada, motivo pelo qual essa será a versão do estimador da variância que utilizaremos aqui. Assim temos que:

$$
\hat{V}\left(\hat{p}_k^h\right) = -\left(\left.\frac{\partial^2 l}{\partial\left(p_k^h\right)^2}\right|_{p_k^h=\hat{p_k^h}}\right)^{-1} = \left(\frac{b_h}{(\hat{p}_k^h)^2}-\frac{\kappa_2(Pess_k^h-b_h)+Cont_k^h-b_h}{(1-\hat{p}_k^h)^2}\right)^{-1}. \qquad (4.111)
$$

Na tabela 4.3 todos estimadores são apresentados de forma resumida, onde $\hat{p} = \left(1-\hat{p}_k^h\right)^{\kappa_2}$, $cont = Cont_k^h$, $cont_i = Cont_k^{h,i}$ e $pess = Pess_k^h$. É importante ressaltar que esses estimadores são não-viciados para a **APV** se o modelo GRH está correto. Já para **APVS**, se modelo **GRH** estiver

correto, esses estimadores são não-viciados para estimar probabilidade de resposta média daquele conglomerado dada por $\overline{p}_k$, porém são viciados para estimar cada p_k^h.

Estimando a probabilidade de resposta quando o tamanho amostral é pequeno

Nessa seção discutimos como obter estimativas para as probabilidades de resposta p_k^h e das variâncias associadas. Os estimadores apresentados são estimadores de máxima-verossimilhança, ou seja, possuem uma distribuição assintótica conhecida, descrita em 4.3. O conhecimento dessa distribuição foi essêncial para derivar alguns dos resultados apresentados. Um problema em potencial é que o resultado em 4.3 só vale para um tamanho de amostra b_h suficientemente grande, e nos desenhos amostrais **APC** e APV apresentados aqui, usualmente b_h é menor do que 10. Ou seja, dificilmente um resultado assintótico poderia ser utilizado para estimar as probabilidades de resposta e a variância do estimador.

Nesse cenário problemático, onde o tamanho das amostras b_h no segundo estágio são muito pequenas, um alternativa simples para aumentar o tamanho da amostra é supor que conglomerados próximos, ou pertencentes a uma mesma região geográfica, possuem a mesma probabilidade de resposta. Fazendo isso, para todos esses conglomerados estima-se conjuntamente as probabilidades de resposta utilizando as quantidades $\sum_{k\epsilon G} Cont_k^h$, $\sum_{k\epsilon G} Pess_k^h$ e $\sum_{k\epsilon G} Dom_k^h$, onde G indica o conjunto dos índices dos conglomerados considerados iguais.

Outra alternativa, mais sofisticada e flexível, é utilizar modelos hierárquicos para modelar as probabilidades de resposta, de forma que os estimadores de uma particular cota utilizem também informação de conglomerados viznhos ou pertencentes a um mesmo bairro ou cidade. Para mais detalhes sobre esse tipo de modelo, veja Gelman and Hill [2007]. Nesse contexto, supondo que o estimador da probabilidade de resposta p_k^h seja dado por $\hat{p}_k^h$ e que $\hat{p}_G^h$ é a estimativa da probabilidade de resposta do estrato h para todos os conglomerados pertencentes a mesma região geográfica sendo considerada $(k\epsilon G)$, obtemos que:

$$\hat{p}_k^{h(Hier)} = \frac{V(\hat{p}_k^h)^{-1}\hat{p}_k^h + V(\hat{p}_G^h)^{-1}\hat{p}_G^h}{V(\hat{p}_k^h)^{-1} + V(\hat{p}_G^h)^{-1}}. \tag{4.112}$$

4.4 Estimando todos os $N_{j,k}^h$

O grande problema em se calcular/estimar as probabilidades $p_{hi/k}$ da unidade i do estrato h do conglomerado k pertencer a amostra, no caso da **APC** e da **APV**, é a dependência delas em todos os $N_{j[i],k}^h$, que usualmente são quantidades desconhecidas para as unidades populacionais não pertencentes a amostra.

Uma solução simples seria supor que as quantidades são independentes, ou seja, a quantidade de moradores do domicílio j não tem influência sobre o número de moradores do domicílio i, para todo para i,j, com $i \neq j$. Nesse caso podemos estimar todas essas quantidades $N_{j,k}^h$ por

$\hat{\overline{N}}^h = \sum_{k\epsilon s_I} \sum_{i\epsilon s_k^h} \frac{N_{j[i],k}^h}{\pi_k\left(1-(1-\hat{p}_k^h)^{N_{j[i],k}^h}\right)}$ ou por $\hat{\overline{N}}_k^h = \sum_{i\epsilon s_k^h} \frac{N_{j[i],k}^h}{\left(1-(1-\hat{p}_k^h)^{N_{j[i],k}^h}\right)}$, ou seja, utilizando uma estimativa conjunta para todos os conglomerados, ou estimando essas quantidades separadamente para cada conglomerado.

Essa não é a solução ideal, pois domicílios que são precedidos por domicílios grandes têm uma probabilidade menor de serem selecionados, e domicílios precedidos por domicílios pequenos têm uma probabilidade maior de serem selecionados. Mesmo supondo independência entre os tamanhos domiciliares, esse fato não é alterado, e essa forma de estimar essas quantidades desconsidera totalmente essa característica da seleção da amostra, particularmente no caso da **APC**.

Uma outra forma de abordar o problema, é considerando todas as possíveis configurações marginais (tamanho de cada um dos D_k domicílios no conglomerado k) para o caso da **APVS** e de todas as possíveis configurações de cada estrato (número de moradores de cada estrato h de cada um dos D_k domicílios no conglomerado k) para o caso da **APV** e da **APC**, que coincidam com a amostra obsevada. As configurações marginais serão denotadas por Σ e as configurações de cada estrato h serão denotadas por Σ_h. Supondo o número máximo de moradores em um único domicílio como sendo n_{max}, temos que existem no máximo $n_{max}^{D_k-b}$ configurações marginais possíveis, pois para os b domicílios pertencentes a amostra do conglomerado k, vamos supor que o número de moradores é conhecido. O número máximo de configurações para cada estrato pode ser aproximado por $n_{max}^{D_k-b}$, porém para cada domicílio, a condição $\sum_{h=1}^H N_{j,k}^h = N_{j,k}$ tem que ser satisfeita.

O objetivo é calcular a probabilidade da amostra observada ser obtida condicionado a cada um das configurações possíveis, e estimar o número de moradores de cada estrato h de cada um dos D_k domicílios no conglomerado k como sendo a configuração que torna a amostra observada mais provável. Ou seja, estamos obtendo o estimador de máxima verossimilhança para o parâmetro Σ ou Σ_h. Para o caso estratificado, a verossimilhança é dada por:

$$L(\Sigma_h/\mathbf{d}) = \prod_{j\in s_k^h} P(D_{j,k}/\Sigma_h), \tag{4.113}$$

onde $P(D_{j,k}/\Sigma_h)$ é a probabilidade do domicílio j do conglomerado k ser selecionado dado a configuração Σ_h dos moradores do estrato h. A probabilidade $P(D_{j,k}/\Sigma_h)$, no caso da **APC**, é dada por $N_{j[i],k}^h p_{hi/k}$, onde $p_{hi/k}$ é definido em 4.29. Já no caso da **APV**, essa probabilidade é dada por $N_{j[i],k}^h p_{hi/k}^{n_{fixo}}$, onde $p_{hi/k}^{n_{fixo}}$ é dado em 4.49. Em ambos os casos, trata-se a configuração Σ_h como se fosse a real quantidade de moradores do estrato h em cada domicílio, e utiliza-se a estimativa $\hat{p}_k^h$ no lugar do parâmetro p_k^h quando ele for desconhecido.

Note que a tendência do estimador de máxima verossimilhança, nesse caso, será estimar o número de moradores em domicílios que estão localizados logo antes (no pecurso) dos domicílios com moradores entrevistados como tendo poucos ou nenhum morador, dessa forma é preciso evitar que essas configurações sejam consideradas. Isso pode ser feito de pelo menos duas formas, obrigando que todas as configurações consideradas tenham o mesmo número de domicílios com cada

possível quantidade de moradores que a população (quando essa quantidade for conhecida) ou desconsiderando configurações "improváveis". Uma forma mais direta de fazer esse controle pode ser obtida utilizando-se **IBM**, considerando uma probabilidade a priori $\pi(\Sigma_h)$ para cada configuração, dessa forma obtendo:

$$\pi(\Sigma_h/\mathbf{d}) \propto \prod_{j \in s_k^h} P(D_{j,k}/\Sigma_h)\pi(\Sigma_h), \tag{4.114}$$

e nesse caso, escolhemos a configuração com a maior probabilidade a posteriori $\pi(\Sigma_h/\mathbf{d})$.

Ao estimar todas as quantidades $N_{j,k}^h$, não estamos considerando na variância do estimador do total populacional a variabilidade acrescentada por estimar essas quantidades. Uma possibilidade é acrescentar a essa variância uma medida de quanto as probabilidades de inclusão variam nas diferentes possíveis configurações, como por exemplo $\frac{1}{\#\Sigma_h}\sum_{g\in\Sigma_h}\left(L(g/\mathbf{d}) - \overline{L}\right)^2$, onde $\overline{L} = \frac{1}{\#\Sigma_h}\sum_{g\in\Sigma_h} L(g/\mathbf{d})$ e $\#\Sigma_h$ é a cardinalidade do conjunto Σ_h, porém ainda não foi encontrada uma solução mais interessante para essa questão.

Finalmente, é possível evitar totalmente essa questão se ao invês de seguir um único trajeto fixo, se o trajeto a ser seguido pelo entrevistador for selecionado aleatoriamente. Ou seja, quando o entrevistador terminar de realizar a entrevista ou de tentar fazer contato com algum morador, seleciona-se aleatoriamente qual será o próximo domicílio que ele tentará contato. O problema com essa solução é de ordem prática, pois dessa forma o entrevistador pode levar mais tempo apenas se locomovendo do que efetivamente tentando completar entrevistas.

Capítulo 5

Simulação e Dados Reais

Nesse capítulo, temos dois objetivos principais: comparar atráves de simulações a performance dos diferentes desenhos amostrais (**APC**, **APV** e **APVS**) combinados com diferentes estimadores do total populacional, e também comparar a performance de 898 pesquisas eleitorais realizadas no Brasil, entre os anos de 1989 e 2004.

5.1 Simulação comparativa entre APC, APV e APVS

Nesta seção iremos descrever o estudo de simulação realizado com o objetivo de comparar os desenhos amostrais **APC**, **APV** e **APVS**, apresentados no Capítulo 4. Comparar esses desenhos amostrais do ponto de vista teórico é muito difícil, e em muitos casos, a comparação tem que ser feita com relação a uma população específica. Assim o estudo de simulação permite um entendimento maior sobre a performance dos diferentes desenhos amostrais.

Consideraremos aqui somente desenhos amostrais com reposição, em um único estágio. Para continuar utilizando as fórmulas dos estimadores e das variâncias apresentadas no Capítulo 4, basta supor que existe um único conglomerado ($A = 1$), o qual tem probabilidade de inclusão $\pi_1 = 1$. Nesse contexto, não é necessário utilizar a notação $/k$, assim as probabilidades de seleção $p_{hi/k}$ serão denotadas, nesse capítulo, como p_{hi} para o caso da **APC** e da **APV**, e p_i para o caso da **APVS**. Também será retirado da notação todos os índices k, que fazem referência ao conglomerado k. *Nesse estudo, a preocupação de restringir a simulação somente a um conglomerado ocorre porque é somente dentro de cada conglomerado que existem diferenças nos desenhos* ***APC***, ***APV*** *e* ***APVS****. Note que a performance do estimador do total populacional depende da performance dos estimadores dos totais populacionais de cada conglomerado selecionado no primeiro estágio. Como em cada conglomerado a performance desses estimadores é sempre igual, dependendo apenas do população do próprio conglomerado, não há a necessidade de fazer uma simulação levando em consideração os dois estágios desses desenhos amostrais, basta entender o comportamento desse estimador em função da distribuição da variável de interesse na populaçao de um único conglomerado.* Ou seja, podemos limitar o estudo de simulação a apenas um conglomerado sem perda de generalidade.

Outro objetivo dessa simulação é comparar diferentes estimadores, além do estimador $\tau_{\hat{H}H}$ apresentado em 4.1. A seguir serão brevemente apresentados os diferentes estimadores utilizados na simulação, os quais serão discutidos com mais detalhes nas seções 5.1.1, 5.1.2 e 5.1.3. Nessa seção, os resultados apresentados não levarão em conta a necessidade de estimar p_h^1 e nem o impacto dessa

estimativa no variância total. O intuito é apenas comparar os diferentes estimadores teoricamente, a fim de compreender quando um estimador pode ser mais eficiente que o outro.

HH Denotado por $\hat{\tau_{HH}} = \sum_{i\epsilon s} \frac{Y_i}{np_i^{Selec}}$, é o estimador usualmente recomendado, tem propriedades teóricas bem conhecidas. É um estimador não-viciado.

Razão Denotado por $\hat{\tau_{raz}} = N\frac{\sum_{i\epsilon s} \frac{Y_i}{np_i^{Selec}}}{\sum_{i\epsilon s} \frac{1}{np_i^{Selec}}}$, é um estimador também bastante recomendado, existem apenas aproximações das suas propriedades teóricas, porém empiricamente sabe-se que ele usualmente tem uma performance melhor do que o $\hat{\tau_{HH}}$. É um estimador viciado, porém o seu vício geralmente é pequeno.

Simples Denotado por $\hat{\tau_{simples}} = \frac{N}{n}\sum_{i\epsilon s} Y_i$, esse estimador não é recomendado quando as probabilidades de seleção das unidades são diferentes pois nesse contexto ele usualmente é viciado, porém é o estimador mais utilizado pelos institutos de pesquisa.

O critério que será utilizado para comparar a performance das diferentes combinações de **desenho amostral** e **estimador** será o Erro Quadrático Médio (EQM). A vantagem de se usar o EQM é que ele leva em consideração tanto a variância do estimador quanto o vício, como podemos ver na definição 1.1. Como alguns dos estimadores utilizados nessa seção são viciados, o EQM se torna uma medida de performance interessante para compará-los.

5.1.1 Propriedades teóricas dos estimadores do tipo HH

O estimador do tipo HH foi amplamente discutido na Seção 1.2.6 e no Capítulo 4, assim aqui iremos apenas re-escrever suas propriedades para o caso de amostragem em um único estágio. Esse estimador é dado por $\hat{\tau}_{HH} = \sum_{h=1}^{H}\sum_{i\epsilon s_h} \frac{Y_{hi}}{b_h p_{hi}}$, onde s_h representa o conjunto dos índices dos elementos populacionais que estão na amostra. Como esse estimador é não-viciado, seu EQM é igual a sua variância. A fórmula da variância é apresentada a seguir:

$$\begin{aligned} EQM\left(\hat{\tau}_{HH}\right) = Var\left(\hat{\tau}_{HH}\right) &= \sum_{h=1}^{H}\sum_{i=1}^{N_h} \frac{p_{hi}}{b_h}\left(\frac{Y_{hi}}{p_{hi}} - \tau_h\right)^2 \\ &= \sum_{h=1}^{H}\frac{1}{b_h}\left(\sum_{i=1}^{N_h}\frac{Y_{hi}^2}{p_{hi}} - \left(\sum_{i=1}^{N_h} Y_{hi}\right)^2\right), \end{aligned} \tag{5.1}$$

onde τ_h representa o total populacional do estrato h do conglomerado A_1 e N_h representa o número de unidades populacionais do estrato h do conglomerado A_1. Usualmente, o estimador $\hat{\tau}_{HH}$ é recomendado pela simplicidade de derivar suas propriedades teóricas e por levar em consideração o desenho amostral, pois ele é baseado nas probabilidades de seleção p_{hi}, as quais são geradas pelo desenho amostral utilizado.

5.1.2 Propriedades teóricas dos estimadores do tipo Razão

Nem sempre o fato de um estimador ser não-viciado quer dizer que ele seja um estimador eficiente. Um bom exemplo disso é o estimador não-viciado da média populacional usualmente utilizado na amostragem de Bernoulli. Nesse tipo de amostragem, todas as unidades populacionais têm uma probabilidade conhecida p de serem selecionados para a amostra, onde para cada unidade gera-se uma variável aleatória $Bern(p)$, se o resultado for 1, aquela unidade populacional pertencerá a amostra. Procedendo dessa forma, esse desenho amostral tem um tamanho amostral $n(s)$ aleatório, sendo que o tamanho esperado é dado por $E(n(s)) = Np$.

Um estimador para a média populacional bastante utilizado nesse caso, justamente por ser não-viciado, é $\hat{\mu}_{Bern} = \sum_{i\epsilon s} \frac{Y_i}{Np}$. É fácil mostrar que esse estimador é realmente não-viciado:

$$E(\hat{\mu}_{Bern}) = E\left(\sum_{i\epsilon s} \frac{Y_i}{Np}\right) = \sum_{i=1}^{N} \frac{Y_i p}{Np} = \sum_{i=1}^{N} \frac{Y_i}{N} = \mu. \tag{5.2}$$

Esse estimador, apesar de ser não-viciado e de suas propriedades teóricas serem facilmente obtidas, é muito contra-intuitivo. Dividir a quantidade $\sum_{i\epsilon s} Y_i$ por $E(n(s))$ ao invés de $n(s)$ é muito difícil de ser justificado, pois sempre que $E(n(s)) \neq n(s)$ estamos acrescentando desnecessariamente variabilidade ao estimador. Utiliza-se esse estimador porque ele é não-viciado e tem propriedades teóricas mais fáceis de serem obtidas.

Um estimador alternativo é a razão de duas quantidades aleatórias, obtido dividindo $\sum_{i\epsilon s} Y_i$ pelo tamanho amostral aleatório $n(s)$. Ele tem propriedades teóricas muito mais complicadas de serem obtidas, geralmente sendo só possível obter aproximações. Outra desvantagem desse estimador é que ele é viciado, porém o vício usualmente é pequeno.

No artigo Strand [1979], o autor mostrou que o estimador $\sum_{i\epsilon s} \frac{Y_i}{n(s)}$ realmente é mais eficiente do que $\sum_{i\epsilon s} \frac{Y_i}{Np}$, ou seja, seu EQM é sempre menor ou igual. Assim, apesar do estimador alternativo ser viciado, ele é melhor do que o estimador $\hat{\mu}_{Bern}$. Esse exemplo foi utilizado para mostrar que, ser não-viciado é apenas uma propriedade de um estimador, não quer dizer que esse é o melhor estimador. O estimador de razão do total populacional que será apresentado nessa seção é a razão do dois estimadores, é viciado, porém usualmente é melhor do que o estimador $\hat{\tau}_{HH}$ não-viciado apresentado na Seção 5.1.1. Mais detalhes sobre a comparação desses dois estimadores podem ser obtidos em Särndal et al. [1992]. O estimador de razão, no contexto desse capítulo, é definido como

$$\hat{\tau}_{raz} = \sum_{h=1}^{H} N_h \frac{\sum_{i\epsilon s_h} \frac{Y_{hi}}{b_h p_{hi}}}{\sum_{i\epsilon s_h} \frac{1}{b_h p_{hi}}} = \sum_{h=1}^{H} N_h \frac{\hat{\tau}^h_{HH}}{\hat{N}_h} = \sum_{h=1}^{H} \hat{\tau}^h_{raz}, \tag{5.3}$$

onde $\hat{\tau}^h_{HH} = \sum_{i\epsilon s_h} \frac{Y_{hi}}{b_h p_{hi}}$ é o estimador HH de τ^h, o total populacional do estrato h, e $\hat{N}_h = \sum_{i\epsilon s_h} \frac{1}{b_h p_{hi}}$ é o estimador HH de N_h, o tamanho do estrato h.

O estimador $\hat{\tau}_{raz}$ é um pouco contra-intuitivo, pois parece estranho utilizar um estimador

"desnecessário"como $\hat{N}_h$ pois N_h é conhecido, porém é justamente isso que faz esse estimador ter uma performance melhor. Dividir por um estimador é importante pois é uma maneira de corrigir distorções, pois as mesmas probabilidades p_{hi} são utilizadas no numerador e no denominador do estimador, e se por acaso algumas dessas probabilidades forem muito grandes ou muito pequenas, esse efeito é cancelado ao se utilizar o estimador de razão. A lógica por trás é similar ao estimador da média populacional para a amostragem de Bernoulli, onde dividir pelo tamanho da amostra $n(s)$ efetivamente observada evita oscilações desnecessárias que ocorrem ao se utilizar o tamanho esperado Np.

Apesar de ser difícil derivar as propriedades teóricas desse estimador, pode-se mostrar que um limite superior para o vício do estimador $\hat{\tau}_{raz}$ é dado por:

$$Vicio(\hat{\tau}_{raz}) \leq \sqrt{\sum_{h=1}^{H} \frac{V(\hat{\tau}^h_{raz})V(\hat{N}^h_{HH})}{N_h}}, \tag{5.4}$$

onde a $V(\hat{N}^h_{HH})$ é igual a $Var\left(\hat{\tau}^h_{HH}\right)$ porém substituindo todos os Y_{hi} por 1.

Para calcular a variância do estimador $\hat{\tau}_{raz}$ é necessário utilizar a técnica de linearização de Taylor para estimação de variâncias, a qual é descrita com bastante detalhes em Wolter [1985]. Ou seja, é possível obter a seguinte aproximação para essa variância, que será denotada por $AV(\hat{\tau}_{raz})$:

$$AV(\hat{\tau}_{raz}) = \sum_{h=1}^{H} AV(\hat{\tau}^h_{raz}) = \sum_{h=1}^{H} \left(\frac{\tau_h}{N_h}\right)^2 \left(\frac{V(\hat{\tau}^h_{HH})}{(N_h)^2} + \frac{V(\hat{N}^h_{HH})}{(\tau_h)^2} - 2\frac{Cov(\hat{\tau}^h_{HH}, \hat{N}^h_{HH})}{N_h\tau_h}\right), \tag{5.5}$$

onde $Cov(\hat{\tau}^h_{HH}, \hat{N}^h_{HH}) = \frac{1}{b_h}\left(\sum_{i=1}^{N_h} \frac{Y_{hi}}{p_{hi}} - N_h\tau_h\right)$.

5.1.3 Propriedades teóricas dos estimadores do tipo Simples

Nessa seção estudaremos o comportamento teórico do estimador $\hat{\tau_{simples}} = \sum_{h=1}^{H} N_h \frac{\sum_{i\epsilon s_h} Y_{hi}}{b_h}$, que também é conhecido como estimador de expansão, pois a média amostral do estrato h é expandida pelo fator N_h. Esse estimador usualmente não é considerado como estimador do total populacional quando as probabilidades p_{hi} são desiguais, justamente porque essas mesmas probabilidades não são utilizadas no cálculo desse estimador e por ele ser viciado. É fácil mostrar que o vício do estimador é dado por:

$$Vicio(\hat{\tau}_{simples}) = \sum_{h=1}^{H}\sum_{i=1}^{N_h} Y_{hi}\left(1 - N_h p_{hi}\right), \tag{5.6}$$

ou seja, somente quando todo $p_{hi} = \frac{1}{N_h}$ esse estimador é não-viciado. Como foi discutido na Sessão 5.1.2, o fato de um estimador ser viciado não quer dizer que ele não seja eficiente, assim é

necessário calcular o EQM desse estimador para avaliar a eficiência do mesmo. Assim, temos que:

$$Var(\hat{\tau}_{simples}) = \sum_{h=1}^{H} \frac{(N_h)^2}{b_h} \left(\sum_{i=1}^{N_h} Y_{hi}^2 p_{hi} - \left(\sum_{i=1}^{N_h} Y_{hi} p_{hi} \right)^2 \right), \tag{5.7}$$

e dessa forma o EQM é dado por:

$$\begin{aligned} EQM(\hat{\tau}_{simples}) &= Vicio^2(\hat{\tau}_{simples}) + Var(\hat{\tau}_{simples}) \\ &= \sum_{h=1}^{H} \frac{(N_h)^2}{b_h} \left(\sum_{i=1}^{N_h} Y_{hi}^2 p_{hi} - \left(\sum_{i=1}^{N_h} Y_{hi} p_{hi} \right)^2 \right) + \left(\sum_{i=1}^{N_h} Y_{hi} \left(1 - N_h p_{hi}\right) \right)^2 \end{aligned} \tag{5.8}$$

Comparando as variâncias dos estimadores $\hat{\tau}_{simples}$ e $\hat{\tau}_{HH}$ não é possível chegar a uma conclusão geral sobre qual estimador é melhor. O que fica evidente é que o primeiro termo $\sum_{i=1}^{N_h} Y_{hi}^2 p_{hi}$ do estimador $\hat{\tau}_{simples}$ é menor ou igual ao termo $\sum_{i=1}^{N_h} \frac{Y_{hi}^2}{p_{hi}}$ de $\hat{\tau}_{HH}$, e em compensação, o termo ao quadrado $\left(\sum_{i=1}^{N_h} Y_{hi} p_{hi}\right)^2$ de $\hat{\tau}_{simples}$ também é menor ou igual ao termo $\left(\sum_{i=1}^{N_h} Y_{hi}\right)^2$ de $\hat{\tau}_{HH}$. Ou seja, soma-se na variância de $\hat{\tau}_{simples}$ uma quantidade menor porém também subtrai-se uma quantidade menor. Assim, um resultado geral não é possível, pois depende da relação entre todos p_{hi} e os Y_{hi}, ou seja, para diferentes populações o estimador mais eficiente pode ser diferente, sem esquecer que é preciso incluir o vício de $\hat{\tau}_{simples}$ nessa comparação.

Um exemplo famoso onde o $\hat{\tau}_{simples}$ é intuitivamente melhor do que o $\hat{\tau}_{HH}$ foi apresentado em Basu [1971], em forma de um estória que será reproduzida aqui. Nesse exemplo, conhecido como **Elefantes de Basu**, o dono de um circo quer saber o peso total dos 50 elefantes do circo. Como é muito trabalhoso pesar os elefantes, ele considera que a melhor forma de estimar o peso total seria escolher um elefante "médio", e multiplicar o seu peso por 50. Olhando os registros de peso de todos os elefantes do circo, realizado 3 anos antes, ele descobre que o elefante Sambo é quem tinha o peso mais parecido com o peso médio dos elefantes. Para ter certeza que esse fato ainda é válido, o dono do circo questiona o treinador se ele acha que Sambo ainda é um bom representante do elefante de peso médio do circo, e o treinador confirma que sim. Assim, o dono decide utilizar para estimar o total populacional $50y$, onde y é o peso de Sambo. Note que esse é o estimador $\hat{\tau}_{simples}$, com $n = 1$ e $N = 50$.

Para ter certeza de que estava procedendo de maneira adequada, o dono resolve falar com o estatístico do circo, e esse, horrorizado, diz que não se pode escolher o elefante Sambo, pois a única maneira de conseguir uma estimativa não viciada do peso total dos elefantes é fazendo uma amostra probabilística, onde para cada elefante i será associada uma probabilidade de seleção p_i, e depois utilizar o estimador HH. O dono e o estatístico chegam a um acordo, e resolvem dar ao elefante Sambo uma probabilidade bem alta de seleção, fazendo $p_{sambo} = \frac{99}{100}$, e para todos os outros elefantes dão a mesma probabilidade $p_{outros} = \frac{1}{4900}$.

Ao realizar o sorteio, como esperado, o elefante Sambo foi selecionado, e o dono do circo fica feliz, pois sabia que ele era o elefante médio. Então o dono pergunta ao estatístico como eles deveriam estimar o total populacional. E a resposta foi "Utilizando o estimador de HH, assim a estimativa do total é $\frac{y}{p_{sambo}} = \frac{100y}{99}$". Ou seja, a estimativa do total populacional seria, aproximadamente, o peso de Sambo, e claramente o peso de um único elefante é muito menor do que o peso de todos os elefantes do circo.

O dono, incrédulo, pergunta como ele teria estimado o total populacional se outro elefante, o Jumbo, tivesse sido sorteado. A resposta do estatístico foi "A estimativa seria $\frac{y}{p_{outros}} = 4900y$, onde y é o peso de Jumbo". Novamente o dono ficou incorfomado, pois agora a estimativa do total seria 4900 vezes o peso de um elefante, o que claramente será muito maior do que o peso de todos os 50 elefantes do circo. E essa é a história de como o estatístico perdeu o emprego. A moral da estória dos elefantes de Basu é: apesar do estimador de HH ser um estimador não-viciado, ele pode fornecer estimativas muito ruins em todas as amostras efetivamente observadas.

Distribuição assintótica da média amostral

Nessa seção, apresentaremos brevemente um resultado novo muito interessante para a média amostral (estimador simples) no contexto de amostragem com probabilidades desiguais com reposição, o qual permite avaliar quando esse estimador é viciado e entender o comportamento do mesmo em função da covariância entre as probabilidades de seleção p_i (não há necessidade de considerar amostragem estratificada nessa sessão, por isso não utilizamos o índice h) e as quantidades populacionais Y_i. Também discutiremos como estimar a variância da média simples nesse contexto. Aqui não há necessidade de considerar o estimador estratificado como foi utilizado na Seção 5.1.3, e estaremos interessados em estimar a média populacional. Re-apresentando os resultados nesse contexto, temos então:

$$Vicio(\overline{Y}_n) = \sum_{i=1}^{N} Y_i \left(p_i - \frac{1}{N}\right) = NCov(Y,p), \tag{5.9}$$

onde $\overline{Y}_n$ é a média amostral, e $Cov(Y,p)$ é a covariância entre as probabilidades de seleção p_i e as quantidades populacionais Y_i, a qual é definida como:

$$\begin{aligned} Cov(Y,p) &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(p_i - \overline{p}\right) \\ &= \frac{\sum_{i=1}^{N} Y_i p_i}{N} - \overline{Y}\overline{p} - \frac{\overline{Y}}{N} + \overline{Y}\overline{p} \\ &= \frac{\sum_{i=1}^{N} Y_i p_i}{N} - \frac{\overline{Y}}{N} = \frac{1}{N}\sum_{i=1}^{N} Y_i\left(p_i - \frac{1}{N}\right), \end{aligned} \tag{5.10}$$

onde $\overline{p} = \frac{\sum_{i=1}^{N} p_i}{N} = \frac{1}{N}$. Nesse contexto, é interessante encontrar um limite superior para o vício da

média simples. Podemos facilmente encontrar um limite superior grosseiro considerando que:

$$\begin{aligned} Cov(Y,p) &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)(p_i-\overline{p}) \leq \frac{1}{N}\sum_{i=1}^{N}\left|Y_i-\overline{Y}\right||p_i-\overline{p}| \\ &\leq \frac{1}{N}\sum_{i=1}^{N}\left|Y_i-\overline{Y}\right| \leq \frac{1}{N}\sum_{i=1}^{N}|Y_i|+\left|\overline{Y}\right|, \end{aligned} \tag{5.11}$$

pois $|p_i-\overline{p}|$ é sempre menor ou igual a 1. Se as quantidades populacionais Y_i forem todas positivas, ou seja, se $Y_i \geq 0\ \forall i$, então de 5.11 podemos obter que:

$$Cov(Y,p) \leq \frac{1}{N}\sum_{i=1}^{N}|Y_i|+\left|\overline{Y}\right| = \frac{1}{N}\sum_{i=1}^{N}Y_i+\overline{Y} = 2\overline{Y}, \tag{5.12}$$

ou seja, sob essas condições, o vício do estimador de média simples é no máximo de $2\tau_y$. Já para o caso da variância, temos de 5.7 que a variância da média amostral é dada por:

$$Var(\overline{Y}_n) = \frac{1}{n}\left(\sum_{i=1}^{N}Y_i^2p_i-\left(\sum_{i=1}^{N}Y_ip_i\right)^2\right). \tag{5.13}$$

Assim, utilizando o teorema obtido em Rosén [1972a] e Rosén [1972b], o qual mostra que a média amostral no contexto de amostragem com probabilidades desiguais sem reposição têm distribuição assintótica normal (o mesmo sendo válido para o caso com reposição), e supondo que as suas condições são válidas, temos que:

$$\overline{Y}_n \approx \mathcal{N}\left(\overline{Y}+NCov(Y,p); \frac{1}{n}\left(\sum_{i=1}^{N}Y_i^2p_i-\left(\sum_{i=1}^{N}Y_ip_i\right)^2\right)\right). \tag{5.14}$$

Do resultado em 5.14, vemos que quando $Cov(Y,p)=0$, a média amostral é um estimador não-viciado da média populacional. Ou seja, com esse resultado, é possível avaliar o impacto de ignorar as probabilidade de seleção p_i para estimar a média populacional. Do ponto de vista mais prático, também é possível estimar a covariância $Cov(Y,p)$ da própria amostra e dessa forma estimar o vício da média simples. Ainda é necessário avaliar o impacto da omissão dessas probabilidades ao estimar $Var(\overline{Y}_n)$. Iremos primeiramente obter um estimador não-viciado para essa variância, de forma que esse estimador possa ser comparado com o estimador da AAS $\frac{s^2}{n}$. Para isso, é importante observar que, conforme apresentado na Seção 1.2.1, no caso da amostragem com probabilidades desiguais com reposição, a esperança da variável aleatória f_{ij} é dada por $E(f_{ij}) = Cov(f_i,f_j)+E(f_i)E(f_j) = n(n-1)p_ip_j$. Utilizando esse resultado, obtemos que nesse desenho amostral a esperança do estimador

da variância $\frac{s^2}{n}$, o qual não depende das probabilidades de inclusão, é dada por:

$$\begin{aligned} E(\frac{s^2}{n}) &= E\left(\frac{\sum_{i\in s}(Y_i-\overline{Y}_n)^2}{n(n-1)}\right) = \frac{E\left(\sum_{i\in s}Y_i^2\right) - nE(\overline{Y}_n^2)}{n(n-1)} \\ &= \frac{1}{n}\left(\sum_{i=1}^{N}Y_i^2p_i - \left(\sum_{i=1}^{N}Y_ip_i\right)^2 + \sum_{i=1}^{N}Y_i^2p_i^2\right). \end{aligned} \tag{5.15}$$

Assim, comparando a esperança do estimador em 5.15 com a variância do estimador simples dada em 5.13, podemos ver que o vício é dado por $\frac{\sum_{i=1}^{N}Y_i^2p_i^2}{n}$. É importante notar que, quanto maior for o tamanho da amostra, menor é esse vício, pois ele depende de $\frac{1}{n}$. Além disso, como as probabilidades $p_i \in (0,1)$, temos que $\frac{\sum_{i=1}^{N}Y_i^2p_i^2}{n} \leq \frac{\sum_{i=1}^{N}Y_i^2}{n}$, ou seja, existe um limite superior para esse vício, o qual também diminue conforme aumentamos o tamanho da amostra.

Para obter um estimador não-viciado para a quantidade $Var(\overline{Y}_n)$, é necessário obter um estimador para esse vício. É fácil mostrar que:

$$E\left(\frac{\sum_{i\in s}Y_i^2p_i}{n^2}\right) = \frac{\sum_{i=1}^{N}Y_i^2p_i^2}{n}. \tag{5.16}$$

Utilizando esse resultado, obtemos que um estimador não-viciado para $Var(\overline{Y}_n)$ é dado por:

$$\hat{Var}(\overline{Y}_n) = \left(\frac{\sum_{i\in s}(Y_i-\overline{Y}_n)^2}{n(n-1)} - \frac{\sum_{i\in s}Y_i^2p_i}{n^2}\right) = \frac{1}{n}\left(s^2 - \frac{\sum_{i\in s}Y_i^2p_i}{n}\right). \tag{5.17}$$

Ou seja, o estimador não-viciado é dado pelo estimador usual da AAS, porém corrigido por um fator que depende das probabilidades de seleção, o qual sempre é negativo. Assim, ao utilizar apenas o s^2 para estimar a variância de $Var(\overline{Y}_n)$, estamos sendo conservadores, pois o estimador s^2 é sempre maior ou igual ao estimador não-viciado $\hat{Var}(\overline{Y}_n)$. O interesse nos resultados dessa seção, é que fazendo algumas suposições sobre os p_i e sobre $Cov(Y,p)$, é possível avaliar o impacto de ignorar essas probabilidades de seleção ao se estimar a média populacional.

5.1.4 Comparação empírica dos estimadores do HH, Razão e Simples

Para comparar a performance dos estimatores HH, Razão e Simples em diferentes cenários, foi feito um pequeno estudo de simulação. Como foi visto na Seção 1.2.6, a performance dos estimador HH depende da relação existente entre as probabilidades de seleção p_i e dos valores populacionais Y_i, o mesmo vale para para os estimadores de Razão e Simples. O intuíto dessa seção é apenas mostrar como nenhum desses estimadores é sempre melhor, e entender melhor como é o comportamento desses estimadores. Os resultados dessas simulações são importantes para ter entendimento mais completo dos resultados da comparação da perfomance dos diferentes estimadores combinados com

os desenhos amostrais **APC**, **APV** e **APVS**.

Os estimadores foram comparados considerando diferentes correlações entre as probabilidades de seleção e os valores populacionais, e além disso estão sendo utilizados para estimar a média populacional, ao invês do total populacional. Diferentes tamanhos amostrais também foram considerados. É importante destacar que no caso do estimador HH, se a correlação linear entre as probabilidades de seleção e os valores populacionais for 1 não implica que a variância desse estimador seja zero, como ocorre quando fazemos $p_i = \frac{Y_i}{\tau_y}$, onde o estimador de HH assume o valor τ_y em todas as possíveis amostras. Para ver isso, vamos calcular as probabilidades $p_i = \frac{Z_i}{\tau_z}$ em função da variável auxiliar z, obtendo o estimador de HH:

$$\hat{\tau}_{HH} = \sum_{i \in s} \frac{Y_i}{np_i} = \tau_z \sum_{i \in s} \frac{Y_i}{nZ_i}, \tag{5.18}$$

onde o valor que esse estimador assumirá em cada possível amostral depende da relação entre Z e Y. Nesse caso, a variância do estimador é dada por:

$$Var(\hat{\tau}_{HH}) = \tau_z \sum_{i=1}^{N} \frac{Z_i}{n} \left(\frac{Y_i}{Z_i} - \frac{\tau_y}{\tau_z} \right)^2. \tag{5.19}$$

Se a correlação linear entre Y e Z for igual a 1, isso implica que $Z_i = a + bY_i$ e consequentemente que $\tau_z = Na + b\tau_y$, para algum a e b. Ou seja, nesse caso obtemos de 5.18 que:

$$\hat{\tau}_{HH} = \frac{Na + b\tau_y}{n} \sum_{i \in s} \frac{Y_i}{a + bY_i}, \tag{5.20}$$

o qual assumirá valores que dependem da unidades populacionais pertencentes a amostra e das quantidades a e b, implicando que a variância do mesmo é maior que zero se $a > 0$. Ou seja, mesmo no caso onde a correlação linear entre Y e as probabilidades de seleção é 1, usualmente a variância do estimador de HH não será nula.

Nas simulações apenas consideramos populações com distribuição Normal. Na figura 5.1, são apresentados os resultados das simulações. Foram desenhados o vício, a variância e o EQM para cada estimador da média separadamente, com $n = 100$ e valores populacionais com uma distribuição $\mathcal{N}(100, 20)$. Dos resultados é possível perceber que a melhor performance do estimador simples ocorre quando a correlação entre as probabilidades de seleção e os valores populacionais é próxima de zero, aumentando quando a correlação se aproxima de 1 e de -1. Também, a variância desse estimador se mantém aproximadamente constante, para qualquer correlação. Já no caso do estimador de HH, sua variância é máxima quando a correlação é -1, e decresce uniformemente conforme a correlação se aproxima de 1, porém como mencionado anteriormente, não é necessário que ela seja nula quando a correlação for 1. Já no caso do estimador de Razão, o vício é usualmente

pequeno, e tanto o vício quanto a variância são constantes ao longo das diferentes correlações.

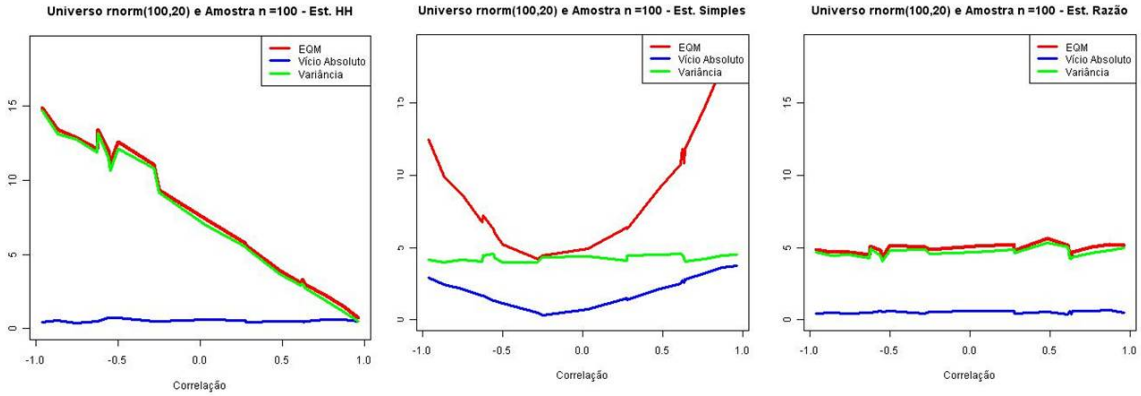

Figura 5.1: Vício, variância e EQM dos estimadores HH, Razão e Simples

Na figura 5.2, são apresentados os resultados das simulações para diferentes parâmetros populacionais e tamanhos de amostra. Foram desenhados o EQM para cada estimador conjuntamente. Desses resultados é evidente que, nos cenários simulados, quando a correlação está próxima de zero, o estimador com a melhor performance usualmente é o Simples, e o estimador de HH é o pior. Conforme a correlação se aproxima de 1, o estimador de HH passa a ser o melhor, e o Simples o Pior. Nas outra correlações, usualmente o melhor estimador é o de Razão, seguido pelo Simples se a correlação for negativa, e pelo de HH se a correlação for positiva.

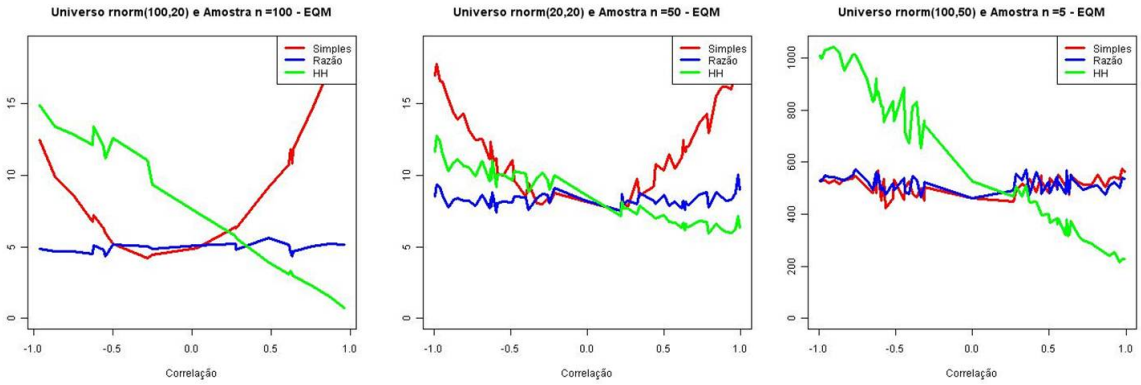

Figura 5.2: EQM dos estimadores HH, Razão e Simples para diferentes tamanhos de amostra

Claramente, a performance dos estimadores são bastante afetadas pelos valores populacionais e pelo tamanho da amostra. É necessário realizar um estudo mais detalhado para compreender melhor quando e como as performances desses estimadores são influenciadas por essas quantidades. Desses resultados, fica evidente que não é possível afirmar que um estimador é uniformemente melhor que o outro, monstrando mais uma vez que ser não-viciado é apenas uma propriedade do estimador, a qual sozinha não garante que o estimador tenha melhor performance que estimadores viciados.

A importância desse resultado se ele for válido no geral, ou seja, quando as probabilidades de seleção e os valores Y_i forem aproximadamente independentes a média simples é o estimador com o menor EQM, é que essa é uma justificativa do ponto de vista de inferência baseada no Desenho para se ignorar o desenho amostral. Ou seja, obtemos um resultado na ID que está de acordo com o princípio da verossimilhança, apresentado em 1.4.2, porém sem a necessidade de supor um modelo super-população para as quantidades Y_i, pelo menos no caso de estimativas pontuais. Mais que isso, esse resultado mostra não só que as probabilidades de seleção podem ser desconsideradas, mas que em alguns casos elas devem ser desconisderadas.

Já no caso de estimativas intervalares, essas probabilidades de seleção ainda são necessárias para se obter estimadores não-viciados, e nesse caso o princípio da verossimilhança seria violado. Para avaliar o impacto do uso de estimadores viciados para a variância na cobertura dos intervalos de confiança, outro estudo de simulação foi realizado.

Como foi visto na Seção 5.1.3, ao utilizar o estimador s^2, baseado na AAS (o qual ignora as probabilidades de seleção p_i) , na verdade estamos sendo mais conservadores do que utilizando o estimador não-viciado para o caso de amostragem com probabilidades desiguais, assim mesmo não considerando as probabilidades de seleção, a cobertura declarada deve ser atingida. O estudo de simulação para avaliar a cobertura dos intervalos de confiança terá dois objetivos principais: comparar a cobertura dos intervalos obtidos com estimadores não-viciados e viciados para a variância, e avaliar o impacto na cobertura dos intervalos causado pelo vício do estimador Simples da média populacional. Os resultados dessa simulação são apresentados na figure 5.3. Nessa simulação consideramos somente o caso onde a população tem distribuição Normal(100,20), onde o tamanho da amostra é 100 e a confiança dos intervalos é de 95%. Para o estimador do tipo Razão, como não existe um estimador não-viciado, nos limitamos a avaliar a cobertura utilizando somente o estimador baseado na estatística s^2.

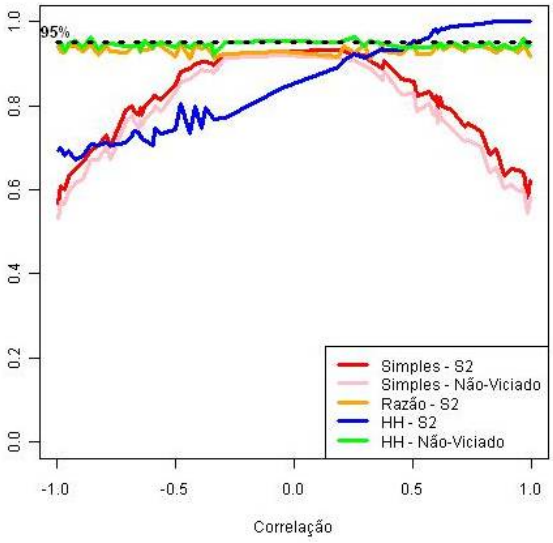

Figura 5.3: Comparação da cobertura dos IC de 95%

Dos resultados, é aparente que não há muita diferença, pelo menos nesse caso, em utilizar o estimador viciado ou não-viciado da variância para o estimador Simples, e como esperado, o estimador viciado tem uma cobertura um pouco melhor do que o estimador não-viciado, pois ele sempre é maior. É evidente que não é possível obter a cobertura esperada de 95% no geral, isso ocorre por causa do vício no estimador da média do tipo Simples, e apenas nos casos onde as probabilidades de seleção e as quantidades populacionais Y_i são aproximadamente independentes é que a cobertura dos intervalos se aproxima dos valores esperados, pois nesse caso o vício do estimador Simples é aproximadamente nulo. Já no caso do estimador de HH, a cobertura obtida utilizando o estimador viciado da variância aumenta conforme a correlação entre as probabilidades de seleção e Y aumenta. A cobertura começa bem baixa, próxima de 70%, e chega inclusive a ter cobertura maior do que a esperada, quando a correlação se aproxima de 1. Isso ocorre porque nesse caso, o estimador HH da média passa a ter uma variância bem pequena, fato que não é refletido pelo estimador baseado no s^2. Como esperado, a cobertura obtida utilizando o estimador não-viciado da variância está sempre próxima da cobertura esperada, o mesmo ocorrendo para o estimador do tipo Razão.

Como os intervalos de confiança são baseados nos quantis da distribuição normal, é importante avaliar a validade do teorema central do limite quando utilizamos o estimador da média do tipo Simples no contexto de amostragem com probabilidades desiguais. Na figura 5.4 apresentamos os resultados da simulação, novamente somente para o caso onde a população tem distribuição Normal(100,20) e o tamanho da amostra é 100. Fica evidente que, conforme discutido na Seção 5.1.3, o TCL é válido sob certas condições, ou seja, a distribuição amostral é normal, porém não necessariamente é centrada na média populacional, sua posição dependendo da correlação entre Y e as probabilidades de seleção. Na figura 5.4, desenhamos a distribuição amostral para os casos com correlação -0.99,-0.03 e 0.99.

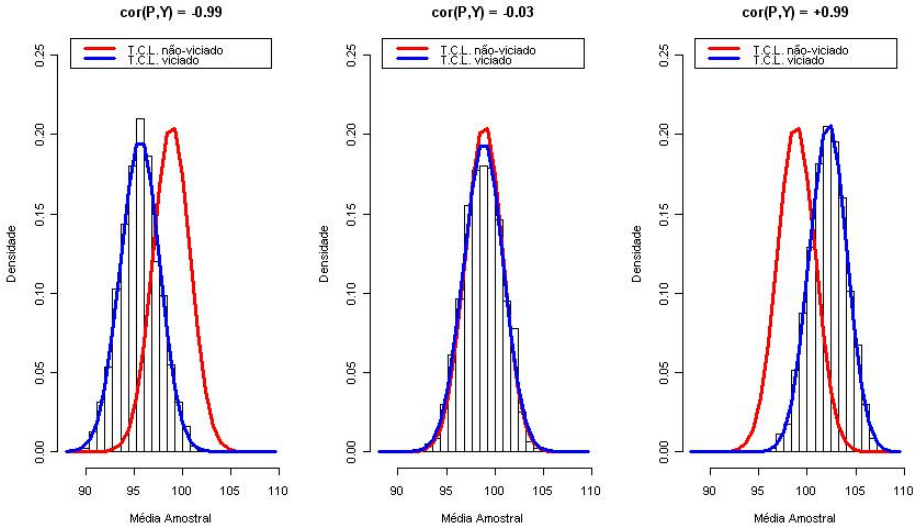

Figura 5.4: Distribuição amostral do estimador da média do tipo Simples

Desses resultados, é evidente que quando a correlação entre as quantidades populacionais de interesse e as probabilidades de seleção forem aproximadamente independentes, o estimador do tipo Simples é usualmente melhor, pois possue um EQM menor, é mais fácil de ser calculado pois não há a necessidade de se conhecer essas probabilidades e a cobertura dos intervalos de confiança se aproxima do valor esperado. Porém, quando mais maior for a correlação, pior é a sua performance, além da cobertura dos intervalos se afastar rapidamente da quantidade esperada.

Um último detalhe importante, é que muitas vezes o tamanho de um conglomerado está relacionado positivamente com o total populacional, ou seja, quanto maior o conglomerado, maior é o total populacional naquele conglomerado. Nesses casos, quando seleciona-se uma amostra com probabilidades proporcionais ao tamanho dos conglomerados, as probabilidades de seleção podem ser correlacionadas positivamente com a quantidade de interesse τ_y, sendo difícil dizer qual estimador seria mais eficiente, pois dependerá do gráu de correlação linear entre essas quantidades. E a correlação, por sua vez, dependerá da relação entre os totais populacionais e os tamanhos de todos os conglomerados. Por exemplo, no caso de populações humanas, existe muita desigualdade social e usualmente pessoas mais parecidas entre si moram nas mesmas áreas (vimos evidência disso na Seção 1.2.4, quando afirmamos que o coeficiente de correlação intra-classe usualmente é positivo). Assim, apesar de quanto maior o tamanho do conglomerado sendo considerado (bairro, setor-censitário, quarteirão, etc...) maior o total populacional de interesse, é comum que na periferia o tamanho dos conglomerado seja maior, porém o total populacional não seja tão grande, principalmente em variáveis que têm alguma relação com a renda das pessoas. Já nas áreas mais nobres de uma cidade, o contrário é verdade, a concentração populacional é menor porém a concentração de riqueza é maior. Essa relação entre os tamanhos populacionais e os totais populacionais usualmente diminuirá consideravelmente a correlação entre essas quantidades.

5.1.5 Universos para simulação

Todas as simulações foram feitas utilizando um universo de mesmo tamanho. Esse universo foi gerado artificialmente, onde fixou-se o número de domicílios em $D = 300$, pois esse é aproximadamente o número de domicílios existentes em um setor censitário, que é o conglomerado usualmente utilizado na prática. Em todas as simulações, considerou-se a existência de 4 estratos/cotas. O número de moradores de cada estrato de cada domicílio N_j^h foi gerado utilizando uma distribuição $Bin(2, 0.3)$, independente dos valores Y_{hi} e das probabilidades de resposta p^h. **Na simulação, supomos que as quantidades N_j^h são conhecidas, permitindo assim calcular diretamente as probabilidades p^h, com exceção do cálculo do estimador EM da APC, onde apenas supõem-se conhecidos os tamanhos dos domicílios que foram entrevistados**. Os tamanhos de amostra considerados foram $b = 8$ e $b = 40$. Também, foram utilizados para os parâmetros (κ_1, κ_2) dos desenhos amostrais **APV** e **APVS** os valores $\{(1, 1), (1, 3), (10, 10)\}$. Nas tabelas com os resultados no apêndice B, o parâmetro κ_1 não será explicitado para economizar espaço, porém conhecendo o κ_2, o κ_1 também é conhecido. **No caso da APC, nas simulações a coleta de dados foi conforme o caso 1, onde o domicílio onde o entrevistador continuará o trajeto após entrevistar uma pessoa é selecionado com probabilidades uniformes.**

Foram realizados diversos cenários de simulações, e em cada uma deles foram realizadas 10.000 replicações. Algumas características importantes tanto dos desenhos amostrais considerados quanto do universo foram alteradas, permitindo a comparação dos diferentes estimadores e desenhos amostrais sob diversos cenários. As características utilizadas nos diferentes universos foram:

Distribuição dos valores Y_{hi} Duas distribuições de probabilidade foram utilizadas para simular os valores das unidades populacionais: Bernoulli e Normal. Foram considerados dois cenários: o **Homogêneo**, onde a média das distribuições é igual para os diferentes estratos, e o **Heterogêneo**, onde foram utilizadas diferentes médias para cada estrato.

Probabilidades de resposta p^h Dois cenários foram considerados: **Homogêneo**, onde a probabilidade de resposta é igual em todos os estratos e **Heterogêneo**, onde a probabilidade de resposta é diferente em cada estrato.

Os diferentes cenários para a distribuição dos valores Y_{hi} e para as probabilidades de resposta p^h são muito importantes, pois permitem avaliar o impacto de duas questões bastante relevantes na teoria de amostragem: desenho amostral ignorável e não-resposta ignorável.

Para um desenho amostral ser classificado como ignorável, informalmente, isso quer dizer que as probabilidades de seleção das unidades populacionais não estão relacionadas com variável de interesse Y. Na definição 1.3, as condições necessárias para que um desenho amostral seja considerado ignorável são apresentadas. Mais detalhes sobre a definição de amostragem ignorável podem ser obtidos em Sugden and Smith [1984]. Uma referência mais resumida, porém abordando uma gama maior de assuntos relacionados a amostragem é Moura [2008]. O significado de não-resposta ignorável é similar ao de amostragem ignorável, ou seja, a probabilidade de resposta não está relacionada com a variável de interesse. Uma diferença importante é que usualmente o processo que define a não-resposta não está sob controle do pesquisador, diferentemente do processo de seleção da amostra. Um exemplo de não-resposta não-ignorável, suponha que quanto maior a renda de uma pessoa, menor a probabilidade dessa pessoa responder a uma pesquisa. Nesse caso, a não-resposta não é ignorável, e claramente esta pesquisa subestimará a renda média da população. As condições para que o mecanismo de resposta seja considerado ignorável podem ser encontradas em Little [1982] e Rubin [1976]. Denominaremos o processo de seleção da amostra como ignorável quando tanto o desenho amostral quanto o mecanismo de resposta forem ignoráveis, caso contrário será classificado como não-ignorável.

Assim, cada universo é composto pela combinação dos cenários homogêneos e heterogêneos tanto da distribuição dos valores Y_{hi} quanto da probabilidade de resposta. Pensando na distribuição dos valores Y_{hi}, no cenário homogêneo foram utilizadas para o caso Bernoulli a distribuição $Bern(0.50)$ e para o caso Normal a distribuição $\mathcal{N}(10000, 10)$, e no cenário heterogêneo, foram utilizadas para o caso Bernoulli as distribuições $Bern(1 - h * 0.20)$, e para o caso Normal as distribuições $\mathcal{N}(h * 10000 + N^h_{j[i]} * 5000, 10)$. Já no caso da probabilidade de resposta, no cenário homogêneo as probabilidades de resposta foram fixadas em $p^h = 0.5$ para todos estratos, e no cenário heterogêneo, utilizaram-se as probabilidades de resposta $p^h = h * 0.2$ para cada estrato.

Tabela 5.1: Resumo dos Universos Simulados

Distribuição da População	Valores Y_{hi}	Não-Resposta p^h	Processo de Seleção	Vantagem em Estratificar
Normal	Heterogêneos	Heterogênea	Não-Ignorável	Sim
		Homogênea	Ignorável	Sim
	Homogenêos	Heterogênea	Ignorável	Não
		Homogênea	Ignorável	Não
Bernoulli	Heterogêneos	Heterogênea	Não-Ignorável	Sim
		Homogênea	Ignorável	Sim
	Homogenêos	Heterogênea	Ignorável	Não
		Homogênea	Ignorável	Não

Cada universo pode ser classificado dependendo da combinação de valores. Existem 4 possíveis combinações, descritas a seguir e resumidas na tabela 5.1: 1) Quando os valores de Y_{hi} são heterogêneos e as probabilidades de resposta também são, o processo de seleção da amostra não é ignorável; 2) Quando os valores dos Y_{hi} e das probabilidades de resposta são homogêneas, o processo de seleção da amostra é ignorável; 3) Quando os valores de Y_{hi} são heterogêneos e as probabilidades de resposta são homogêneas, onde o processo de seleção da amostra é ignorável porém uma amostra estratificada deve ter uma performance melhor do que uma amostra sem estratificação e 4) Quando os valores de Y_{hi} são homogêneos e as probabilidades de resposta são heterogêneas, onde o processo de seleção da amostra é ignorável e uma amostra estratificada não deve ter uma performance melhor do que uma amostra sem estratificação.

O interesse em analisar esses diferentes universos existe pois espera-se que quando a amostragem não for ignorável, o estimador $\hat{\tau}_{simples}$ e os estimadores baseados no desenho amostral **APVS** tenham uma performance pior do que nos casos onde a amostragem é ignorável. No total foram simulados 8 universos diferentes, resumidos na tabela 5.1.

5.1.6 Resultados da Simulação - Condicionado ao conhecimento de p^h

Os resultados das simulações, nos quais as probabilidades de resposta eram conhecidas (ou seja, não precisavam ser estimadas) estão resumidas no apêndice B , nas tabelas B.5 e B.6. Incialmente, discutiremos os vícios relativos dos diferentes pares de estimadores e desenhos amostrais. O vício relativo é definido como:

$$VR(\hat{\tau}) = \frac{|Vicio(\hat{\tau})|}{\tau_y}. \tag{5.21}$$

Na tabela B.5 os vícios relativos são apresentados. Eles são maiores quando a distribuição de Y segue uma distribuição Bernoulli, isso ocorre pois nesse caso o total populacional é pequeno. Já no caso da distribuição Normal, o total populacional é muito grande, fazendo com que os vícios

relativos sejam menores.

Como esperado, os estimadores HH estratificados (**APV** e **APC**) não são viciados. Também, no geral, o vício dos estimadores do tipo Razão são menores do que os do tipo Simples. Além disso, os vícios dos estimadores do tipo Razão diminuem se o tamanho da amostra aumenta.

Claramente os maiores vícios ocorrem quando a distribuição de Y e da Não-Resposta é heterogênea, afetando principalmente os estimadores do tipo Simples para qualquer tipo de desenho amostral, porém os maiores vícios ocorrem em todos os estimadores do desenho amostral do tipo **APVS**.

Quando o Y é homogêneo, os efeitos da Não-Resposta são minimizados, com os vício relativos diminuindo consideravelmente. Ou seja, se os valores populacionais são homogêneos, a não-resposta passa a não ter tanta relevância. Essa redução é tão evidente, que tanto os estimadores da **APVS** e quanto os do tipo Simples passam a ter vícios relativos menores de 1%, com exceção aos estimadores Simples da **APC**, que continuam viciados.

Quando o Y e a Nâo-Resposta são heterogêneos, ou seja, naquele caso mais grave onde estamos ignorando o modelo **GRH**, os vícios relativos dos estimadores da **APVS** diminuem consideravelmente ao aumentar os valores de κ_2, confirmando o que foi discutido na Seção 4.2.6, onde vimos que aumentar κ_2 é uma forma de se prevenir contra a má-especificação do modelo de resposta.

Os EQM's para o universo Bernoulli estão dispostos na tabela B.3 e para o universo Normal na tabela B.4. Para facilitar a comparação dos estimadores nos diferentes cenários, foram calculados os rankings dos EQM's na tabela B.6. Analisando os rankings dos EQM dos pares de estimadores e desenhos amostrais, fica evidente que os estimadores da **APVS** sofrem muito quando o Y é heterogêneo. No caso do ranking do EQM, aumentar o parâmetros κ_2 não parece melhorar muito a performance desses estimadores. A tabela 5.2 ordenada dos rankings facilita a análise dos resultados.

Fica evidente que os estimadores do tipo Simples, apesar de viciados, têm uma performance melhor do que os outros. Em segundo lugar, estão os estimadores do tipo Razão. Ou seja, os estimadores HH usualmente recomendados têm a pior performance de todos os estimadores estudados nessa simulação, repetindo os resultados obtidos na Seção 5.1.4 quando a correlação entre as probabilidades de seleção e os valores populacionais são independentes, como ocorre nessa simulação dentro dos estratos. Ao aumentar o tamanho da amostra de 8 para 40, o estimador de Razão da **APV** passa a ter melhor performance do que o estimador Simples da **APVS**, provavelmente porque o vício do estimador de Razão diminue bastante com o aumento da amostra.

Uma possível explicação para a performance do estimador do tipo Simples é que, **a validade do modelo GRH implica, no geral, que a covariância entre as probabilidades de seleção p_{hi} e a variável de interesse Y_{hi} dentro de cada estrato conforme foi discutido na Seção 5.1.4, pois essas probabilidades são muito parecidas entre si, a única diferença entre elas sendo a ordem e o tamanho dos domicílios**. Talvez se fosse considerado um cenário onde a os tamanhos dos domicílios fossem altamente correlacionados com a variável Y, poderíamos ver uma queda na performance dos estimador Simples e melhora das performance dos outros estimadores.

Tabela 5.2: Ranking Médio do EQM dos estimadores de τ_y

b	Tipo do Estimador	Desenho Amostral	Ranking EQM Médio
8	SIMPLES	**APV**	1.71
		APC	1.75
		APVS	3.88
	RAZÃO	**APV**	3.88
		APC	5.17
		APVS	5.63
	HH	**APV**	6.79
		APC	8.08
		APVS	8.13
40	SIMPLES	**APV**	1.63
		APC	1.79
	RAZÃO	**APV**	3.92
	SIMPLES	**APVS**	4.25
	RAZÃO	**APC**	5.21
		APVS	5.88
	HH	**APV**	6.46
		APC	7.92
		APVS	7.96

Com relação aos desenhos amostrais, fica claro que a **APV** e a **APC** são os melhores e têm performance muito parecidas quando o estimador Simples é utilizado. Nos casos do estimadores de Razão e HH, a **APV** têm performance melhor, e nesses casos, a performance para os desenhos **APC** e **APVS** se aproxima, porém a **APC** sempre é melhor do que a **APVS**.

Nessa simulação também foram registrados o número de contatos com pessoas, o número de pessoas abordadas e o número de domicílios contactados como forma de avaliar o tempo de execução para os desenhos amostrais **APV**, **APVS** e **APC**. Esses resultados foram resumidos no apêndice B, na tabela B.1.

No caso da **APC**, foram registrados o número de domicílios contactados, e não o número de voltas realizadas pelo entrevistador, conforme a teoria apresentada na Seção 4.2.3. Essa diferença ocorre pois é muito difícil calcular o número esperado de domicílios contactados, porém durante a simulação foi possível fazê-lo.

Para o caso dos desenhos amostrais **APV** e **APVS**, fica claro para o caso do número de contatos que o resultado das simulação confirma o resultado teórico, pois o número de contatos é constante, independente de κ_2, e também porque o número de pessoas abordadas diminue conforme aumentamos esse parâmetro. Também podemos perceber que o número de contatos e de pessoas abordadas da **APVS** é uma média ponderada pelo número de moradores das probabilidades de resposta dos diferentes estratos da **APV**. Também, conforme aumentamos o κ_2, o número de pessoas contactadas por entrevista deve se aproximar de 1, ou seja, toda pessoa selecionada será

entrevistada se esse parâmetro for grande o suficiente, tanto para a **APV** quanto para a **APVS**.

É difícil comparar a **APC** e as **APV** e **APVS**, pois as unidades são diferentes, mas é importante destacar que apesar de na **APC** o número de contatos ser maior, o tempo de execução não é muito afetado, pois o que realmente leva tempo é a busca de pessoas específicas que não estão no domicílio ou não querem responder, fato que ocorre tanto na APV quanto na **APVS**.

5.1.7 Resultados da Simulação - Estimando p^h

Os resultados das simulações, nos quais as probabilidades de resposta foram estimadas (ou seja, eram desconhecidas) estão resumidas no apêndice B, nas tabelas B.9 e B.10. Incialmente, discutiremos os EQM's dos pares de estimadores e desenhos amostrais para a probabilidade de resposta p^h. Para o caso da **APVS** e da **APV** somente o estimador do tipo C, definido na Seção 4.3.2 foi considerado, e para a **APC** o estimador EM, definido na Seção 4.3.2 foi o considerado.

Na tabela B.2, a performance dos diferentes desenhos amostrais e estimadores foi resumida. Quando a não-resposta é Homogênea, o estimador do tipo C da **APVS** sempre é melhor do que os outros, isso ocorre porque, nesse caso, o tamanho da amostra para estimar a probabilidade de resposta é maior visto que nesse desenho amostral não se utiliza estratificação e as probabilidades de resposta nos diferentes estratos/cotas são iguais. Já no caso da não-resposta Heterogênea, usualmente o estimador do tipo C da **APV** é um pouco melhor, porém a performance do EM da **APC** é muito parecida no estrato 1 onde a probabilidade de resposta é menor (0.2). Algumas vezes o estimador do tipo EM da **APC** consegue ter uma performance melhor.

Analisando o vício relativo dos estimadores de τ_y, comparando a tabela B.9 com a tabela B.5 onde não havia a necessidade de estimar o p^h, percebe-se que ambas são muito parecidas, sendo a principal diferença que ao estimar o p^h, até os estimadores de HH são viciados, mesmo nos casos onde a dsitribuição de Y e a Não-Resposta são homogêneas. Isso ocorre pois os estimadores de p^h são apenas assintoticamente não-viciados, e nesse caso estamos trabalhando amostras muito pequenas. Na mesma tabela, é evidente que ao aumentar κ_2 (para **APV** ou **APVS**) ou o tamanho da amostra b_h (para todos), o vício dos estimadores HH reduz consideravelmente.

Os EQM's para o universo Bernoulli estão dispostos na tabela B.7 e para o universo Normal na tabela B.8. Para facilitar a comparação dos estimadores nos diferentes cenários, foram calculados os rankings dos EQM's na tabela B.10. Ao comparar os resultados dos rankings médios dos EQM dos estimadores/desenhos amostrais com e sem estimar o p^h, esses resultados também são muito parecidos, com poucos estimadores mudando de lugar no ranking. Na tabela 5.3 a comparação dos dois rankings é apresentada, também incluindo o aumento percentual médio do **EQM** ao estimar o p^h. Em alguns poucos casos, esse percentual é negativo, isso ocorre por causa da simulação, e não quer dizer que realmente a performance do estimador melhorou por estimarmos o p^h.

Os estimadores de HH são os que mais sofrem por estimar o p^h, chegando a aumentar o EQM em até 28%, enquanto os estimadores do tipo razão aumentam até 7%. Porém, ao aumentar um pouco o tamanho da amostra, esse impacto reduz consideravelmente.

No geral, os melhores estimadores são, na ordem, Simples, Razão e HH, independente do desenho amostral utilizado. Com relação aos desenhos amostrais, os melhores são a **APV**, em segundo lugar

a **APC** com performance bastante similar, seguidos pela **APVS**, que tem a pior performance.

Tabela 5.3: Comparação do Ranking Médio do EQM dos estimadores de τ_y

b	Tipo do Estimador	Desenho Amostral	Ranking Médio do EQM sem estimar p^h	Ranking Médio do EQM estimando p^h	Aumento (%) do EQM
8	Simples	**APC**	1.75	1.50	-0.6
		APV	1.71	1.88	0.3
		APVS	3.88	3.71	0.5
	Razão	**APV**	3.88	4.04	4.1
		APC	5.17	5.21	6.6
		APVS	5.63	5.25	0.9
	HH	**APV**	6.79	7.21	22.2
		APVS	8.08	7.71	14.2
		APC	8.13	8.50	27.6
40	Simples	**APV**	1.63	1.63	0.0
		APC	1.79	1.92	0.1
		APVS	4.25	4.08	0.0
	Razão	**APV**	3.92	4.08	2.0
		APC	5.21	5.04	-0.8
		APVS	5.88	5.83	0.1
	HH	**APV**	6.46	6.67	4.7
		APC	7.92	7.75	-2.6
		APVS	7.96	8.00	1.3

5.2 Avaliação empírica das pesquisas eleitorais no Brasil (1989-2004)

No Capítulo 4 foi apresentada a teoria para amostragem onde é levada em consideração a probabilidade de resposta. Nesse contexto, foram calculadas as probabilidades de seleção p_i^{Selec} para o desenho **APC**, que é um dos desenhos amostrais mais utilizados na prática, principalmente no contexto de pesquisas eleitorais, que usualmente têm que ser feitas com bastante rapidez. A importância desse resultado é que a **APC** pode ser considerada uma amostragem probabilística, e ela passa a ter o respaldo teórico que apenas a **APV** possuia anteriormente, condicionado ao modelo **GRH** estar correto.

Em contraste com a Seção 5.1, onde comparamos a performance dos diferentes estimadores onde todo o universo é conhecido para os diferentes cenários e sabia-se se o modelo GRH era válido, o objetivo dessa seção é avaliar a performance das pesquisas eleitorais na prática, onde não se sabe se o modelo **GRH** é valido e onde existem diversos erros não-amostrais, conforme discutido na Seção 1.3, os quais também podem afetar a performance dos estimadores.

As pesquisas eleitorais são dos poucos casos de pesquisas de opinão pública onde é possível

avaliar se as pesquisas realmente conseguiram estimar corretamente o total populacional, pois sabe-se o resultado das eleições. As pesquisas analisadas nessa seção foram adquiridas no Banco de Dados de Pesquisas por amostragem do **Centro de Estudos de Opinião Pública (CESOP) - UNICAMP**[1]. Somente pesquisas realizadas até 40 **dias antes das eleições** foram incluídas nessa análise.

No banco de dados do CESOP só estão disponíveis pesquisas dos institutos Ibope e DataFolha. Apesar de existirem outros institutos grandes no Brasil, como CNT-Sensus e VOX POPULI, esses dois são os maiores, que realizam mais pesquisas e têm maior exposição na mídia. Seria interessante incluir pesquisas de outros institutos, mas até o momento não foi possível.

Por questões de ordem prática, geralmente associadas as logística da coleta de dados, os maiores institutos de pesquisa do Brasil utilizam usualmente uma metodologia de amostragem que seleciona os respondentes através de cotas, como é possível verificar nos sites dos próprios institutos. Alguns exemplos são:

- **IBOPE - *www.ibope.com.br***

  "Seleção dos entrevistados por meio de cotas proporcionais de sexo, idade, grau de instrução e setor de dependência econômica, dentro dos setores censitários sorteados previamente. As cotas servem para evitar vieses decorrentes da não existência de cadastros dos eleitores dentro dos setores censitários e da impossibilidade do levantamento de tal informação durante o processo da pesquisa."

- **DATAFOLHA - *www.datafolha.folha.uol.com.br***

  "A pesquisa do Datafolha é um levantamento por amostragem com abordagem pessoal com cotas de sexo e idade e sorteio aleatório dos entrevistados."

- **CNT-SENSUS - *www.sensus.com.br***

  "Amostras estratificadas para 5 Regiões e 24 Estados, com o sorteio aleatório de 136 Municípios pelo método da Probabilidade Proporcional ao Tamanho - PPT. Probabilística sistemática até o Setor Censitário para Urbano e Rural, com cotas para Sexo, Idade, Escolaridade e Renda no Setor Censitário"

- **VOX POPULI - *www.voxpopuli.com.br***

  "Foi adotada uma amostra estratificada por cotas , com o total de 2000 entrevistas, distribuídas proporcionalmente entre regiões, de acordo com o número de eleitores. As cotas

[1]Todas essas pesquisas podem ser encontradas no site *www.cesop.unicamp.br/busca/CESOP/pesquisa_usuario*.

utilizadas foram gênero, idade, escolaridade, renda familiar e situação perante o trabalho, sendo calculadas proporcionalmente a cada estrato de acordo com os dados do IBGE, censo de 2000 e TSE"

Assim, o interesse nessa seção é avaliar a performance das pesquisas eleitorais na prática, as quais quase sempre são amostragem por cotas (**AC**) ou amostragem probabilística com cotas (**APC**). Nesse estudo serão avaliadas 898 pesquisas eleitorais realizadas pelos institutos de pesquisa DataFolha e Ibope, entre os anos de 1989 e 2004. Dessas pesquisas 397 (44%) pesquisas foram realizadas pelo DataFolha e 501 (56%) pelo Ibope. A relação de todas as pesquisas analisadas está na tabela D.1, no apêndice D. Uma avaliação desse tipo já foi feita pela Associação Brasileira de Empresas de Pesquisa (ABEP), considerando 469 pesquisas realizadas entre 1982 e 1998. Porém não existe um relatório com os resultados dessa análise, apenas uma menção aos resultados da mesma, a qual pode ser encontrada na página 19 do "Guia ABEP para Divulgação de Pesquisas Eleitorais"[2], onde utiliza-se o critério do erro absoluto por categoria, obtendo-se um erro médio de 3% no geral, e de 2% considerando somente os votos válidos.

Pela grande quantidade de pesquisas analisadas nessa seção, é muito dificil avaliar a metodologia de amostragem de cada pesquisa, assim nesse estudo não faremos distinção entre pesquisas domiciliares, telefônicas ou em pontos de fluxo, entre pesquisas com controles geográficos rígidos, ou até mesmo entre pesquisas com mais de uma etapa de seleção ou pesquisas com probabilidades de seleção diferentes. Ou seja, todos os desenhos amostrais serão considerados iguais, com exceção do tamanho da amostra. É evidente que essas características das pesquisas influenciam os resultados das mesmas, por isso as conclusões apresentadas aqui serão cautelosas.

Outra simplificação que será necessária nas análises diz respeito ao desenho amostral em si, pois não conhecemos para as pesquisas estudas as probabilidades de inclusão/seleção de cada unidade amostral, o que implica que não podemos calcular nem o estimador HH nem a variância do mesmo. Por causa disso, a única alternativa possível é utilizar o estimador Simples, apresentado na Seção 5.1.3, pois ele não depende dessas probabilidades. Utilizar esse estimador não é um grande problema, pois no estudo de simulação foi verificado que apesar de ser viciado, esse é o estimador que têm o menor EQM para todos os desenhos amostrais, e também porque esse é o estimador usualmente utilizado pelos institutos de pesquisa. Porém do ponto de vista de estimar a variância, será necessário recorrer a outra simplificação, pois a variância do estimador simples também depende das probabilidades de inclusão/seleção de cada unidade amostral, como foi mostrado em 5.7.

Além das características do proprio desenho amostral, outros fatores importantes que também podem ser bastante importantes para avaliar a qualidade de uma pesquisa, pois podem aumentar o erro não-amostral. No artigo Desart and Holbrook [2003], os autores apresentram alguns dos fatores que mais influenciam na precisão das pesquisas eleitorais. Na medida do possível, consideraremos esses fatores na análise dos resultados. São eles:

1. **Tamanho da Amostra:** Por definição, o tamanho do erro amostral diminue conforme

[2] www.datasurvey.com.br/manuais/pdfs/GuiaAbep_DivulgacaoPesquisasEleitorais.pdf

o tamanho da amostra aumenta. Assim esperamos que pesquisas com amostras maiores produzam resultados mais precisos, se os outros fatores forem constantes.

2. **Dias antes da Eleição:** Em teoria, as pesquisas eleitorais produzem uma estimativa sobre o parâmetro populacional no instante de tempo quando a pesquisa foi realizada. Se o parâmetro populacional muda com o tempo, as pesquisas realizadas anteriormente produziriam resultados menos precisos. Alguns autores acreditam que esse seja o fator mais importante.

3. **Indecisos:** Em quase todas as pesquisas existem eleitores indecisos. Podemos supor que uma pesquisa com um grande número de leitores indecisos teria resultados menos precisos, pois na eleição, os eleitores tem que votar em algum candidadto, ou então branco ou nulo, dessa forma, os indecisos na pesquisa terão que se decidir na eleição.

4. **Dias de Campo:** O número de dias em que foram realizadas entrevistas parece ter influência na precisão das estimativas de uma pesquisa. O argumento aqui é que quanto mais dias forem utilizados para a coleta dos dados, melhor serão as informações coletadas e consequentemente o erro não-amostral será reduzido.

5. **Fins-de-Semana:** Pesquisas nas quais as entrevistas foram realizadas somente em dias úteis têm precisão menor do que aquelas nas quais a coleta de dados se realiza em fins de semana também. O motivo é evidente: nos fins de semana é mais fácil encontrar alguns perfis de eleitores, conforme foi discutido ns Seção 1.3.2, fazendo com que essas pesquisas representam melhor o universo de eleitores.

6. **Efeitos específicos do Ano:** Pesquisas realizadas em diferentes anos, podem ter efeitos específicos daquele ano ou daquela eleicção, os quais podem melhorar ou piorar a precisão das estimativas das pesquisas.

Existem também dificuldades conceituais para se avaliar a qualidade das pesquisas eleitorais. Talvez a mais importante delas seja determinar quando consideramos que uma pesquisa eleitoral estimou corretamente o parâmetro populacional de interesse. Em Souza [1990], o autor avalia duas características que podem ser consideradas para afirmar que uma pesquisa estimou corretamente o parâmetro populacional. São elas:

Ranking dos Candidatos Se a pesquisa estimou corretamente o ranking dos candidatos na eleição. Isso inclui o critério mais ingênuo, que equivale a dizer se as pesquisas acertaram o nome do candidatos vencedor da eleição.

Diferença entre a estimativa e o parâmetro Se a diferença entre a estimativa e o parâmetro populacional é menor do que o erro amostral declarado, ou se é considerada pequena.

Nesta tese, trabalharemos principalmente com a diferença entre a estimativa e o parâmetro, que **só pode ser interpretada como erro amostral se ignorarmos a existência dos erros não-amostrais**. Na Seção 5.2.1, apresentaremos alguns critérios de erro que podem ser utilizados para

avaliar a performance das pesquisas eleitorais, na Seção 5.2.2 é realizada uma análise descritiva da performance das pesquisas eleitorais ena Seção 5.2.3 um modelo de regressão múltipla é ajustado aos dados das pesquisas eleitorais.

5.2.1 Critérios de Erro

Apresentaremos quatro classes diferentes de critérios de erro que podem ser utilizadas para avaliar a qualidades das pesquisas eleitorais, cada uma com objetivos distintos. A primeira fundamentada na teoria de Amostragem Aleatória Simples (**AAS**) a qual considera somente os erros amostrais, a segunda, mais descritiva, baseada em critérios criados pelo Social Science Research Council (SSRC), que é uma organização internacional que busca avançar as pesquisas de ciências sociais no mundo, a terceira, também descritiva, porém considerando o ranking dos candidatos, e a quarta, utilizando o conceito de distância entre dois pontos.

O interesse em apresentar tantas classes diferentes de critérios de erros, é que estamos interessados em considerar diversos tipos de erros, e não somente o erro amostral.

Critérios de Erro baseados na AAS

Antes de discutir com maior profundidade os critérios de erro utilizados nessa seção, é importante adaptar a notação apresentada para o estimador do tipo Simples na Seção 5.1.3 para se estimar proporcões populacionais, ao invês de totais populacionais. Supondo que existam C categorias (número de candidatos disputando um eleição - estamos desconsiderando outras categorias como não sabe, não respondeu, etc...), estamos interessados em estimar as quantidades populacionais:

$$P_c = \frac{\sum_{i=1}^{N} Y_i^{(c)}}{N},$$

onde $Y_i^{(c)} = 1$ se a i-ésima unidade populacional votar no candidato c, e $Y_i^{(c)} = 0$ caso contrário. É importante salientar que $\sum_{c=1}^{C} P_c = 1$. O estimador de P_c, para o caso de amostragem com reposição, é dado por:

$$\hat{P}_c = \frac{\sum_{i \epsilon s} Y_i^{(c)}}{n},$$

onde n é o tamanho da amostra. A esperança de $\hat{P}_c$ é dada por:

$$E(\hat{P}_c) = \sum_{i=1}^{N} Y_i^{(c)} p_i,$$

e a variância por:

$$\begin{aligned} Var(\hat{P}_c) &= \frac{1}{n}\left(\sum_{i=1}^{N}\left(Y_i^{(c)}\right)^2 p_i - \left(\sum_{i=1}^{N} Y_i^{(c)} p_i\right)^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{N} Y_i^{(c)} p_i - \left(\sum_{i=1}^{N} Y_i^{(c)} p_i\right)^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{N} Y_i^{(c)} p_i \left(1 - \sum_{i=1}^{N} Y_i^{(c)} p_i\right)\right) \\ &= \frac{1}{n}\left(E(\hat{P}_c)\left(1 - E(\hat{P}_c)\right)\right), \end{aligned} \tag{5.22}$$

ou seja, a $Var(\hat{P}_c)$ depende das probabilidades p_i que nesse estudo são consideradas desconhecidas. Nessa seção, vamos supor que todas as pesquisas foram feitas com AAS com reposição.

Agora, o enfoque é um pouco diferente, pois não repetimos o mesmo procedimento de seleção. Na verdade, temos uma aplicação de 898 procedimentos de seleção diferentes, para populações diferentes. Nesse contexto, o interesse está em avaliar se a distribuição dos erros de todas essas diferentes pesquisas e para todos os diferentes candidatos têm o comportamento teoricamente esperado. Para fazer isso, é necessário conhecer a distribuição amostral do estimador $\hat{P}_c$ para cada pesquisa, e se for possível, padronizá-las, de forma que todos os estimadores considerados tenham a mesma distribuição de referência. O interesse é verificar se essa distribuição padronizada teórica de referência se aproxima da distribuição empírica, indicando que as pesquisas eleitorais estão tendo o comportamento teórico esperado. Duas versões diferentes dessa distribuição padronizada serão comparadas, uma que leva em consideração cada categoria separadamente e outra que considera todas as categorias conjuntamente, denominadas respectivamente de, binomial e multinomial.

Para populações infinitas no contexto de AAS, não é difícil mostrar que a média amostral tem assintoticamente uma distribuição normal, ou seja, quanto maior for o tamanho da amostra n, melhor a distribuição amostral é aproximada pela distribuição normal. Esse resultado é conhecido como o Teorema Central do Limite (TCL). São necessárias mais condições, além da usual de que $n \to \infty$, para encontrar um resultado similar ao TCL para **AAS** de populações finitas. Detalhes desse resultado e das condições necessárias podem ser encontradas em Hájek [1960]. Existem também resultados similares para o estimador Simples no contexto de amostragem com probabilidades desiguais sem reposição, como pode ser visto em Rosén [1972a] e Rosén [1972b], porém as condições para validade do teorema não são as mesmas do caso com probabilidades iguais. Nessa seção, iremos supor que essas condíções são satisfeitas.

No contexto de pesquisas eleitorais, onde usualmente estamos interessados em estimar proporções, utilizaremos o resultado em 1.16 pois esse é um dos raros casos onde a quantidade σ^2 é conhecida. Adaptando a notação, podemos re-escrever 1.16 como sendo

$$\frac{\hat{P}_c - E(\hat{P}_c)}{\sqrt{Var(\hat{P}_c)}} = \frac{\hat{P}_c - \sum_{i=1}^{N} Y_i^{(c)} p_i}{\sqrt{\frac{1}{n}\left(\sum_{i=1}^{N} Y_i^{(c)} p_i \left(1 - \sum_{i=1}^{N} Y_i^{(c)} p_i\right)\right)}} \approx \mathcal{N}(0,1), \tag{5.23}$$

porém, como as probabilidades de seleção p_i são desconhecidas, substituiremos $\sum_{i=1}^{N} Y_i^{(c)} p_i$ por P_c e também $Var(\hat{P}_c)$ por σ^2/n, os quais não dependem das probabilidades p_i, obtendo:

$$\frac{\hat{P}_c - P_c}{\sqrt{\frac{P_c(1-P_c)}{n}}} \approx \mathcal{N}(0,1), \tag{5.24}$$

onde nesse contexto, $\overline{Y}_n = \hat{P}_c$, $\mu = P_c$ e $\sigma^2 = P_c(1 - P_c)$. Fica evidente desse resultado que estamos utilizando um estimador viciado para a estimar P_c e que estamos simplificando a variância do estimador $\hat{P}_c$, utilizando a variância da **AAS** com Reposição ao invês da variância gerada pelo desenho amostral. Conforme foi discutido nas Seção 5.1.4, se as probabilidades de seleção p_i e as quantidades populacionais Y_i tiverem correlação aproximadamente nula, não há problema com essas simplificações.

Distribuição Binomial

Nessa seção, criaremos um indicador de erro para cada categoria separadamente. Esse indicador será baseado no resultado apresentado em 5.24. Supondo que esse resultado esteja correto para as pesquisas sendo analisadas, existe o interesse em avaliar se as pesquisas têm o comportamento esperado.

Para simplificar a análise, a intenção é criar um indicador de forma que possa resumir a performance das pesquisas eleitorais na prática, sem a necessidade de avaliar a distribuição empírica dos resultados. O primeiro passo para obter tal indicador é calcular qual erro amostral foi divulgado para cada pesquisa. Essa informação não existe explicitamente, porém usualmente fixa-se o nível de confiança $1 - \alpha$ em 95%. Supondo esse nível de confiança e utilizando o tamanho da amostra, que é conhecido, obtemos de 1.24 que:

$$d_p = z_{(1-\frac{\alpha}{2})}\sqrt{\frac{P_{c,p}(1-P_{c,p})}{n_p}} = 1,96\sqrt{\frac{P_{c,p}(1-P_{c,p})}{n_p}} <= \frac{1,96}{2\sqrt{n_p}}, \tag{5.25}$$

onde d_p é o erro amostral da pesquisa p, $z_{(1-\frac{\alpha}{2})}$ é o quantil de $\mathcal{N}(0,1)$, tal que $P(\mathcal{N}(0,1) \leq z_{(1-\frac{\alpha}{2})}) = 1 - \frac{\alpha}{2}$, n_p é o tamanho da amostral da pesquisa p e $P_{c,p}$ é a proporção de votos da categoria c da pesquisa p. Ou seja, iremos supor que o erro amostral máximo, divulgado para uma particular pesquisa p, foi $\frac{1,96}{2\sqrt{n_p}}$.

O erro observado da pesquisa p para a categoria c é dado pela diferença $|\hat{P}_{c,p} - P_{c,p}|$. Da definção

de erro amostral, temos que:

$$P\left(|\hat{P}_{c,p} - P_{c,p}| \leq d_p\right) \geq 1 - \alpha, \tag{5.26}$$

ou seja, sabemos que o erro amostral deveria ser menor ou igual a $\frac{1{,}96}{2\sqrt{n_p}}$ em pelo menos $(1-\alpha)\%$ das categorias/pesquisas [3]. O indicador I^{α}_{BIN} criado para avaliar a qualidade das pesquisas eleitorais baseado em cada categoria separadamente, obtido desse fato, é dado por:

$$I^{\alpha}_{BIN} = \frac{\sum_{p=1}^{N_p}\sum_{c=1}^{C_p}\mathbf{1}_{\{|\hat{P}_{c,p}-P_{c,p}|\leq\frac{z_{(1-\frac{\alpha}{2})}}{2\sqrt{n_p}}\}}}{\sum_{p=1}^{N_p} C_p}, \tag{5.27}$$

onde N_p é o número de pesquisas avaliadas, C_p é o número de categorias da pesquisa p e $\mathbf{1}_{\{a\leq b\}}$ é uma função indicadora, que assume o valor 1 se $a \leq b$ e o valor 0 caso contrário. Se todas as suposições mencionadas forem satisfeitas, então temos que

$$E\left(I^{\alpha}_{BIN}\right) \geq 1 - \alpha. \tag{5.28}$$

Distribuição Multinomial

Um problema com o indicador I^{α}_{BIN} é que ele desconsidera que as C_p categorias de cada pesquisa não são independentes. Isso ocorre porque existe a restrição de que a soma $\sum_{c=1}^{C_p} P_{c,p} = 1$, assim se uma categoria aumenta percentualmente, alguma(s) das outras têm que diminuir. Pode-se mostrar que a covariância entre os estimadores das categorias c_i e c_j, no caso geral, é dada por:

$$Cov(\hat{P}_{c_i}, \hat{P}_{c_i}) = -\frac{1}{n}\sum_{a=1}^{N} Y_a^{(c_i)} p_a \sum_{a=1}^{N} Y_a^{(c_j)} p_a = -\frac{1}{n} E(\hat{P}_{c_i}) E(\hat{P}_{c_j}). \tag{5.29}$$

Enunciando matematicamente nosso objetivo, queremos encontrar um indicador baseado na probabilidade

[3] Note que também é possível considerar a margem de erro exata, ou seja, $d_p = z_{(1-\frac{\alpha}{2})}\sqrt{\frac{P_{c,p}(1-P_{c,p})}{n_p}}$ pois nesse contexto conhecemos os parâmetros populacionais, ou então considerarmos uma estimativa dessa quantidade, dada por $\hat{d_p} = z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{P}_{c,p}(1-\hat{P}_{c,p})}{n_p}}$, quantidade essa que poderia ser calculada e utilizada pelos próprios institutos de pesquisas, porém usualmente eles optam for inflacinoar a margem de erro, substituindo a quantidade $P_{c,p}(1-P_{c,p})$ pelo valor máximo que ela pode assumir, dado por $0,25$.

$$P\left(\bigcap_{c=1}^{C_p}\left\{|\hat{P}_{c,p}-P_{c,p}|\leq d_p\right\}\right)\geq 1-\alpha_G, \tag{5.30}$$

ao invês da probabilidade em 5.26, pois as empresas de pesquisa divulgam um único erro amostral d_p para todas as categorias da pesquisa p. Claramente, a probabilidade em 5.30 é muito menor do que a probabilidade em 5.26, além de descrever de uma maneira muito mais rigorosa o que é o erro amostral de uma pesquisa quando existem mais de 2 categorias sendo avaliadas. Adaptando os resultados apresentados em 1.36 para esse contexto, e utilizando o mesmo d_p da seção anterior, sabemos então que em aproximadamente $(1-2\alpha)\%$ das pesquisas, todos erros amostrais observados em cada uma das categorias de uma pesquisa p deveriam ser menores do que $\frac{1,96}{2\sqrt{n_p}}$. Ou seja, o indicador I^{α}_{MULT} criado para avaliar a qualidade das pesquisas eleitorais baseado em todas as categorias simultaneamente, obtido desse fato, é dado por:

$$I^{\alpha}_{MULT}=\frac{\sum_{p=1}^{N_p}\prod_{c=1}^{C_p}\mathbf{1}_{\{|\hat{P}_{c,p}-P_{c,p}|\leq\frac{z_{(1-\frac{\alpha}{2})}}{2\sqrt{n_p}}\}}}{N_p}. \tag{5.31}$$

Se todas as suposições mencionadas forem satisfetias, então temos que

$$E\left(I^{\alpha}_{MULT}\right)\geq 1-2\alpha. \tag{5.32}$$

Critérios de Erro Descritivos (SSRC)

No artigo Mitofsky [1998], o autor rebate algumas críticas que foram feitas as pesquisas eleitorais americanas em 1996. Para verificar a veracidade das críticas recebidas, ele discute alguns critérios criados pelo Social Science Research Council (SSRC) [4] para avaliar a precisão das estimativas das pesquisas eleitorais. O interesse em replicar alguns desses critérios é que dessa forma podemos comparar os resultados das pesquisas brasileiras com pesquisas de outros paises, como Estados Unidos e Reino Unido.

Outra questão importante é que os critérios definidos nessa seção não levam em consideração os erros-amostrais declarados. Os critérios dessa seção apenas quantificam o tamanho das diferenças observadas, porém não existe um referêncial teórico com o qual essas diferenças podem ser comparadas.

De todos os critérios apresentados em Mitofsky [1998], utilizaremos somente dois deles, que são avaliados pelo autor como os critérios mais interessantes e que foram utilizados para avaliar outras pesquisas internacionais. Esses critérios tem aspectos positivos e negativos, assim faremos

[4] Site: www.ssrc.org/

um breve discussão deles aqui. Será mantida a mesma nomenclatura utilizada no artigo.

1. **Critério 3 - Média das diferenças absolutas dos percentuais de votos válidos entre as pesquisas e o resultado das eleições para cada candidato:** O interesse desse critério é que ele avalia todos os candidatos. Para alguns autores, essa é a real medida de precisão das pesquisas. Porém é necessário cautela, pois se o número de candidatos for muito grande, esse critério terá valores muito baixos, pois usualmente temos muitos candidatos pouco expressivos com um baixo percentual de votos. É recomendável limitar o número de candidatos utilizados no cálculo desse critério. Alguns autores recomendam contabilizar somente candidatos com pelos menos 15% de intenção de voto. Nesse estudo utilizaremos duas variações desse critério, uma com todos os candidatos, e outra com somente aqueles candidatos que têm pelo menos 15% das intenções de voto. Esses critérios serão denotados, respectivamente, por I^{C3}_{SSRC} e $I^{C3(15\%)}_{SSRC}$.

2. **Critério 5 - Diferença entre duas diferenças, sendo a primeira a diferença entre as estimativas de intenção de voto dos 2 principais candidatos e a segunda é a diferença entre o resultado da eleição para os mesmos candidatos:** esse critério é interessante pois ele avalia uma das estatísticas mais divulgada nas pesquisas eleitorais, que é a diferença percentual entre os dois principais candidatos. A desvantagem é a relativa dificuldade de explicar esse critério, por ele ser definido como uma diferença entre duas diferenças. Esse critério será denotado de I^{C5}_{SSRC}.

O critério 3, denotado por I^{C3}_{SSRC}, matematicamente é definido como:

$$I^{C3}_{SSRC} = \frac{1}{N_p}\sum_{p=1}^{N_p}\frac{\sum_{c=1}^{C_p}|\hat{P}_{c,p} - P_{c,p}|}{C_p}, \tag{5.33}$$

e a sua variação, apenas considerando candidatos com pelo menos 15% dos votos, denotada por $I^{C3(15\%)}_{SSRC}$, é definida como:

$$I^{C3(15\%)}_{SSRC} = \frac{1}{N_p}\sum_{p=1}^{N_p}\frac{\sum_{c=1}^{C_p}\left(|\hat{P}_{c,p} - P_{c,p}|\right)\mathbf{1}_{\{P_{c,p}\geq 0{,}15\}}}{\sum_{c=1}^{C_p}\mathbf{1}_{\{P_{c,p}\geq 0{,}15\}}}, \tag{5.34}$$

onde $\mathbf{1}_{\{P_{c,p}\geq 0{,}15\}}$ assume valor 1 se $P_{c,p} > 0,15$ e 0 caso contrário. Já o critério 5, denotado por I^{C5}_{SSRC}, é definido matematicamente como:

$$I^{C5}_{SSRC} = \frac{1}{N_p}\sum_{p=1}^{N_p}\left((\hat{P}_{c_1,p} - \hat{P}_{c_2,p}) - (P_{c_1,p} - P_{c_2,p})\right), \tag{5.35}$$

onde o índice c_1 indica o candidato que obteve o maior percentual de votos na eleição sendo avaliada

pela pesquisa p e c_2 indica o candidato que obteve o segundo maior percentual de votos na eleição sendo avaliada pela pesquisa p. Também consideramos uma varição do I^{C5}_{SSRC}, de forma a tornar o critério positivo:

$$I^{|C5|}_{SSRC} = \frac{1}{N_p}\sum_{p=1}^{N_p}\left||\hat{P}_{c_1,p} - \hat{P}_{c_2,p}| - |P_{c_1,p} - P_{c_2,p}|\right|. \tag{5.36}$$

Critérios de Erro Descritivos (Rankings)

Também iremos considerar três critérios de erro descritivos não utilizados pelo SSRC, porém que avaliam se a pesquisa previu corretamente os rankings dos candidatos. O primeiro critério, denotado por $I^{Vencedor}_{Ranking}$, avalia se as pesquisas previram corretamente o candidato que venceu a eleição. Ele é definido como:

$$I^{Vencedor}_{Ranking} = \frac{1}{N_p}\sum_{p=1}^{N_p}\mathbf{1}_{\left\{\arg\max\left\{\hat{P}_{1,p},...,\hat{P}_{C_p,p}\right\}=\arg\max\left\{P_{1,p},...,P_{C_p,p}\right\}\right\}}, \tag{5.37}$$

onde $\mathbf{1}_{\{a=b\}}$ assume valor 1 se $a = b$ e 0 caso contrário.

Os outros dois critérios, denotados por $I^{Todos}_{Ranking}$ e $I^{\%Acerto}_{Ranking}$ avaliam, respectivamente, se as pesquisas acertaram corretamente o ranking de todos os candidatos e o percentual médio de acerto dos rankings de cada pesquisa. O crítério $I^{Todos}_{Ranking}$ é definido por:

$$I^{Todos}_{Ranking} = \frac{1}{N_p}\sum_{p=1}^{N_p}\prod_{c=1}^{C_p}\mathbf{1}_{\left\{\hat{P}_{(c),p}=P_{(c),p}\right\}}, \tag{5.38}$$

onde $\hat{P}_{(c),p}$ indica o candidado c-ésimo classificado segundo a pesquisa p, $P_{(c),p}$ indica o candidado c-ésimo classificado na eleição correspondente e $\mathbf{1}_{\left\{\hat{P}_{(c),p}=P_{(c),p}\right\}}$ assume valor 1 se os candidatos $\hat{P}_{(c),p}$ e $P_{(c),p}$ são os mesmos, caso contrário assume 0. Analogamente, o crítério $I^{\%Acerto}_{Ranking}$ é definido por:

$$I^{\%Acerto}_{Ranking} = \frac{1}{N_p}\sum_{p=1}^{N_p}\frac{\sum_{c=1}^{C_p}\mathbf{1}_{\left\{\hat{P}_{(c),p}=P_{(c),p}\right\}}}{C_p}. \tag{5.39}$$

Critérios de Erro baseados em Distâncias

Serão consideradas nessa seção dois critérios de erro baseados em distâncias. Nesse contexto, o resultado da pesquisa eleitoral p, definido pelo vetor $\hat{P}_p = (\hat{P}_{1,p}, ..., \hat{P}_{C_p,p})'$, é visto como um ponto no espaço $\mathbb{R}^{C_p}$. O interesse está em avaliar a distância desse ponto para o resultado das eleições,

também interpretado como um ponto e definido pelo vetor $P_p = (P_{1,p}, ..., P_{C_p,p})'$.

A distância entre dois pontos do mesmo espaço pode ser definida como uma função $d(.)$ que a cada par de pontos $P^{'}$ e $P^{''}$ associa um número real positivo, $d(P^{'}, P^{''})$, com as seguintes propriedades:

Positividade - $0 \leq d(P^{'}, P^{''})$ e $d(P^{'}, P^{''}) = 0$ se e somente se $P^{'} = P^{''}$.

Simetria - $d(P^{'}, P^{''}) = d(P^{''}, P^{'})$.

Desigualdade Triangular - $d(P^{'}, P^{''}) \leq d(P^{'}, P^{'''}) + d(P^{'''}, P^{''})$, onde $P^{'''}$ é um ponto qualquer do espaço.

A diferença entre essas duas distâncias consideradas nessa seção será explicada a seguir.

Distância de Mahalanobis Essa distância entre dois pontos, introduzida em Mahalanobis [1936], é similar a distância Euclidiana, porém leva em consideração as covariância das variáveis que compõem esse ponto.

Distância Composicional de Aitchison Essa distância é específica para dados composicionais. Esse tipo de dados representam pontos que pertencem a um sub-espaço do $\mathbb{R}^C$, no qual a soma das componentes do vetor é restrita a um valor fixo. Essa distância foi introduzida em Aitchison [1982].

A distância de Mahalanobis é bastante conhecida, sendo chamada por alguns autores de distância estatística. Ela recebe esse nome pois leva em consideração a variância e as covariâncias do vetor $\hat{P}_p$. Antes de discutir a vantagem dessa distância, vamos definí-la a seguir.

Definição 5.1 (Distância de Mahalanobis) *Seja $P = (P_1, ..., P_k)'$ um vetor onde cada componente p_i é uma variável aleatória, $p = (p_1, ..., p_k)'$ é o vetor observado de P e $q = (q_1, ..., q_k)'$ um vetor onde cada componente é uma constante definida por $q_i = E(P_i)$. A distância de Mahalanobis é definida como:*

$$D_{Maha}(p, q) = \sqrt{(p-q)'\Sigma^{-1}(p-q)}, \tag{5.40}$$

onde Σ é a matriz de covariâncias de P.

Para o caso bi-dimensional, considerando apenas as variáveis aleatórias X e Y, onde $E(X) = \mu_x$, $E(Y) = \mu_y$, $V(X) = \sigma_x^2$, $V(Y) = \sigma_y^2$ e ρ é o coeficiente de correlação de X e Y e também que x é o valor observado de X e y é o valor observado de Y, pode-se mostrar que:

$$D_{Maha}\left((x, y), (\mu_x, \mu_y)\right) = \sqrt{\frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]}. \tag{5.41}$$

Assim, por exemplo, podemos ver que a distância referente a componente x, dada por $x - \mu_x$ é ponderada pelo inverso da variância de X, assim se a variável X têm uma variância muito grande, isso fará com que grandes diferenças observadas $(x - \mu_x)$ tenham o seu peso diminuido na distância $D_{Maha}(x, y)$.

No contexto dessa seção, iremos supor que $E(\hat{P}_p) = P_p$ e que $V(\hat{P}_p) = \frac{P_p(1-P_p)}{n_p}$ conforme foi discutido em 5.24. O critério de erro baseado na distância de Mahalanobis, denotado por I^{Maha}_{Dist}, será definido como:

$$I^{Maha}_{Dist} = \sum_{p=1}^{N_p} \frac{\sqrt{\left(\hat{P}_p - P_p\right)' \Sigma_p^{-1} \left(\hat{P}_p - P_p\right)}}{N_p}, \tag{5.42}$$

onde Σ_p é a matriz de covariâncias do vetor $\hat{P}_p$ da pesquisa p, onde as componentes dessa matriz são definidas por $\sigma_{ij} = -\frac{P_{i,p}P_{j,p}}{n_p}$ se $i \neq j$ e $\sigma_{ii} = \frac{P_{i,p}(1-P_{i,p})}{n_p}$ caso contrário. Lembrando que também no caso da covariância σ_{ij} estamos supondo que foi utilizada **AAS** com reposição, caso contrário essa covariância seria dada por $-\frac{1}{n_p}E(\hat{P}_{i,p})E(\hat{P}_{j,p})$ para amostragem com reposição.

Note, que para cada pesquisa p, se $\hat{P}_p$ segue uma distribuição normal multivariada $\mathcal{N}_{C_{p-1}}(P_p, \Sigma_p)$, então $\left(\hat{P}_p - P_p\right)' \Sigma_p^{-1} \left(\hat{P}_p - P_p\right)$ segue uma distribuição $\chi^2_{C_{p-1}}$. Desse forma obtemos que:

$$E\left(\sum_{p=1}^{N_p} \frac{\left(\hat{P}_p - P_p\right) \Sigma_p^{-1} \left(\hat{P}_p - P_p\right)}{N_p}\right) = \sum_{p=1}^{N_p} \frac{C_{p-1}}{N_p}.$$

A distância composicional de Aitchison têm a importante característica de levar em consideração que a soma das componentes do vetor é restrita ao valor 1. Em Aitchison [1986], o autor discute diferentes interpretações e inferências que podem ser obtidas de dados desse tipo. Antes de discutir a vantagem em se utilizar a distância de Aitchison, iremos definí-la a seguir.

Definição 5.2 (Distância Composicional de Aitchison) *Sejam $P = (p_1, ..., p_k)'$ e $Q = (q_1, ..., q_k)'$ dois vetores composicionais k-dimensionais, ou seja, dois vetores tais que $\sum_{i=1}^k p_i = C$ e $\sum_{i=1}^k q_i = C$, onde C é uma constante. Fazendo $r_i = \ln\left(\frac{p_i}{q_i}\right)$ para $i = 1, ..., k$, a distância de Aitchison é definida como:*

$$D_{Aitch}(P, Q) = \sqrt{\sum_{i=1}^{k} (r_i - \overline{r})^2}, \tag{5.43}$$

onde $\overline{r} = \frac{1}{k}\sum_{i=1}^k r_i$.

A melhor forma de explicar a vantagem de se utilizar essa distância é através de um exemplo. O exemplo apresentado a seguir é uma variação de um exemplo encontrado em Fossaluza et al. [2009].

Vamos supor que foram realizadas duas pesquisas diferentes (1 e 2), para avaliar duas eleições diferentes (1 e 2), ambas com 3 candidatos. Os resultados das eleições 1 e 2 foram, respectivamente $e_1 = (0.2, 0.4, 0.4)$ e $e_2 = (0.1, 0.2, 0.7)$, e os resultados das pesquisas 1 e 2 foram, respectivamente $p_1 = (0.4, 0.2, 0.4)$ e $p_2 = (0.2, 0.1, 0.7)$.

Note que o erro observado pelas duas pesquisas é equivalente, no sentido de que a pesquisa 1 sub-estimou o percentual de votos do candidato 1 em $\frac{0.2}{0.4} = 50\%$, que é o mesmo erro cometido pela pesquisa 2 para o candidato 1 ($\frac{0.1}{0.2} = 50\%$). As duas pesquisas também cometem o mesmo erro ao estimar o percentual de votos no candidato 2, super-estimando o percentual do candidato em $\frac{0.4}{0.2} = \frac{0.2}{0.1} = 200\%$ e ambas as pesquisas não cometeram erros ao estimar o percentual de votos no candidato 3.

Conforme destacado acima, as duas pesquisas cometem os mesmos erros relativos, assim seria interessante nesse contexto que a distância utilizada $d(,)$, indicasse que as duas pesquisas cometeram o mesmo erro, ou seja, $d(e_1, p_1) = d(e_2, p_2)$. Se for utilizada a distância euclidiana, obtemos que $d(e_1, p_1) = 0,2828$ e $d(e_2, p_2) = 0,1414$, ou seja, essa distância considera que o erro cometido pela pesquisa 2 foi menor do que o erro cometido pela pesquisa 1. Porém, se utilizarmos a distância composicional de Aitchison, obtemos que $d(e_1, p_1) = d(e_2, p_2) = 0,9802$, resultado que faz mais sentindo no contexto de dados composicionais.

No contexto dessa seção, o critério de erro baseado na distância composicional de Aitchison, denotado por I_{Dist}^{Aitch}, será definido como:

$$I_{Dist}^{Aitch} = \frac{1}{N_p}\sum_{p=1}^{N_p} D_{Aitch}(\hat{P}_p, P_p) = \frac{1}{N_p}\sum_{p=1}^{N_p}\sqrt{\sum_{c=1}^{C_p}(r_{c,p} - \overline{r_p})^2}, \tag{5.44}$$

onde $r_{c,p} = \ln\left(\frac{\hat{P}_{c,p}}{P_{c,p}}\right)$ e $\overline{r_p} = \frac{1}{C_p}\sum_{c=1}^{C_p} r_{c,p}$.

5.2.2 Análise Descritiva dos Resultados

Os resultados oficiais das eleições foram obtidos no site do Tribunal Superior Eleitoral (TSE)[5]. Nesse site também foram obtidas as datas da realização de todas as eleições entre 1989 e 2004, dispostas na tabela 5.4.

As informações obtidas do TSE e do CESOP foram compatibilizadas, e delas foram criados dois conjuntos de dados, uma para cada tipo de análise que será realizada nessa seção: um onde as categorias foram analisadas separadamente, e outra onde as categorias de cada pesquisa são analisadas conjuntamente. A motivação para duas análises distintas é que os institutos de pesquisa calculam e divulgam o erro amostral para cada candidato da pesquisa separadamente, porém a maneira teoricamente mais correta é considerar conjuntamente todos os candidatos, conforme discutido na Seção 5.2.1. Fazendo essas duas análises, podemos avaliar se as pesquisas acertam as estimativas segundo os critérios utilizados pelos institutos, e também avaliar a sua performance

[5] www.tse.gov.br

segundo critérios mais rigorosos.

Tabela 5.4: Data de realização das Eleições - 1989 - 2004

CARGOS ELETIVOS - ELEIÇÕES DIRETAS			
ANO	PRESIDENTE	GOVERNADOR	PREFEITO
1989	15 de Novembro(1^{o}Turno) 17 de Dezembro(2^{o}Turno)		
1990		3 de Outubro(1^{o}Turno) 25 de Novembro(2^{o}Turno)	
1992			3 de Outubro(1^{o}Turno) 15 de Novembro(2^{o}Turno)
1994	3 de Outubro(1^{o}Turno)	3 de Outubro (1^{o}Turno) 15 de Novembro (2^{o}Turno)	
1996			3 de Outubro (1^{o}Turno) 15 de Novembro(2^{o}Turno)
1998	4 de Outubro(1^{o}Turno)	4 de Outubro(1^{o} Turno) 25 de Outubro(2^{o} Turno)	
2000			1 de Outubro (1^{o}Turno) 29 de Outubro(2^{o}Turno)
2002	6 de Outubro(1^{o}Turno) 27 de Outubro(2^{o}Turno)	6 de Outubro (1^{o}Turno) 27 de Outubro(2^{o}Turno)	
2004			3 de Outubro(1^{o}Turno) 31 de Outubro (2^{o}Turno)

Em ambos os casos, foram retiradas da análise todos os candidatos com a menor quantidade de votos em cada eleição. Esse ajuste foi realizado para evitar que os erros observados fossem super-estimados devido a restrição de que em todas as pesquisas, a soma do percentual de votos válidos dos candidatos têm que ser 1. Para exemplificar melhor como o comportamento dos erros amostrais são correlacionados por causa dessa restrição, calculamos a probabilidade em 1.23 para o caso onde existem somente 2 candidatos, denotando o percentual de votos na eleição por P_1 e P_2 e as respectivas estimativas por $\hat{P}_1$ e $\hat{P}_2$ (a mesma idéia vale para um número qualquer de candidatos):

$$\begin{aligned} P\left(\left\{|\hat{P}_1 - P_1| > \varepsilon\right\} / \left\{|\hat{P}_2 - P_2| > \varepsilon\right\}\right) &= P\left(\left\{|(1-\hat{P}_2) - (1-P_2)| > \varepsilon\right\} / \left\{|\hat{P}_2 - P_2| > \varepsilon\right\}\right) \\ &= P\left(\left\{|\hat{P}_2 - P_2| > \varepsilon\right\} / \left\{|\hat{P}_2 - P_2| > \varepsilon\right\}\right) \\ &= 1. \end{aligned} \tag{5.45}$$

De 5.45, podemos ver que para uma eleição com 2 candidatos, se a pesquisa eleitoral cometer um erro maior do que ε para um candidato, a probabilidade de cometer o mesmo erro para o outro candidato é 1, ou seja, nesse cenário a pesquisa comete o mesmo erro para ambos os candidatos. Por esse motivo, retiramos de cada pesquisa analisada o candidato com o menor número de votos válidos, para não contabilizarmos erros redundantes. Na tabela C.1 , no apêndice C, estão resumidos o total de pesquisas e o total de categorias análisadas. Foram analisadas um total de 898 pesquisas considerando as categorias conjuntamente, e 3870 categorias separadamente. Essas informações também estão cruzadas por fatores de interesse.

Os resultados dos critérios de erro apresentados nesse capítulo tanto para o geral quanto con-

siderando os fatores de interesse estão resumidos na tabela C.2, no apêndice C. Os critérios de erro apresentado nessa seção possuem diferentes características, resumidas na tabela 5.5.

As características descritas na tabela 5.5 indicam, respectivamente, se o critério leva em consideração a distribuição amostral, se avalia as categorias de cada pesquisa separadamente (binomial) ou conjuntamente (multinomial), se leva em consideração a variância do estimador e suas covariâncias, e finalmente se leva em conta que as estimativas das pesquisas são dados composicionais.

Tabela 5.5: Características do critérios de erro considerados

Tipo	Critério		Distrib.	Categorias		Variância	Covariância	Dados Com-
	Nome	Notação	Amostral	Separ/e	Conj/e	do Est.	das Categ.	posicionais
AAS	Binomial	$I^{0,05}_{BIN}$	X	X		X		
	Multinomial	$I^{0,05}_{MULT}$	X		X	X	X	
Descritivo	C3	I^{C3}_{SSRC}			X			
	C3 15%	$I^{C3(15\%)}_{SSRC}$			X			
	C5	I^{C5}_{SSRC}			X			
	C5 Abs	$\lvert I^{C5}_{SSRC}\rvert$			X			
Rankings	Acerto Venc.	$I^{Vencedor}_{Ranking}$		X				
	Acerto Todos	$I^{Todos}_{Ranking}$			X			
	% de Acerto	$I^{\%Acerto}_{Vencedor}$			X			
Distância	Mahalanobis	I^{Maha}_{Dist}			X	X	X	
	Aitchison	I^{Aitch}_{Dist}			X			X

Os fatores de interesse considerados na análise foram:

Cargo Indica qual tipo de eleição estava sendo avaliada: Prefeito, Governador ou Presidente.

Final de Semana Indica se pelo menos uma parte da coleta dos dados foi realizada durante o fim de semana.

Dias de Campo Indica quantos dias durou a coleta de dados.

Turno Indica qual turno da eleição estava sendo avaliada.

Votos Não-válidos Indica o percentual de votos nulos, em branco e/ou de pessoas indecisas na pesquisa.

Dias antes da eleição Indica quantos dias antes da eleção a pesquisa foi realizada. Foi utilizado como referência o último dia da coleta de dados.

Candidatos Indica a quantidade de candidatos disputando a eleição sendo avaliada.

Tamanho Amostral Indica o tamanho da pesquisa realizada.

Classes de Variância Indica a qual grupo de variância a eleição pesquisada pertence. Para o caso de eleições com 2 candidatos, quanto mais próximo de 0.5 for o percentual de voto nos candidatos, maior é a variância. Já para o caso com 3 ou mais candidatos, foi utilizada a distância de Aitchison para avaliar a distância do vetor P_p para um vetor onde todas as C_p categorias são equi-prováveis $(1/C_p)$. Em Thompson [1987], o autor mostra que para o caso multinomial, o vetor P_p que implica a maior variância dos estimadores é dado por $(1/m)$ para m categorias, e 0 nas outras $C_p - m$ categorias, onde $m \leq C_p$. Aqui, supomos que $m = C_p$. Foram criadas 4 classes de variância, cada uma com o mesmo número de pesquisas.

Complexidade Amostral Indica se no banco de dados da pesquisa existia um fator de ponderação.

Instituto de Pesquisa Indica qual instituto de pesquisa realizou cada pesquisa.

Muitos fatores discutidos aqui são correlacionados. Por exemplo todas as pesquisas feitas no segundo turno só têm 2 candidatos, as pesquisas presidencias do primeiro turno têm muitos candidatos e as classes de variância baixa têm em sua maioria, pesquisas com maior número de candidatos, como pode ser visto na tabela 5.6.

Tabela 5.6: Número de Candidatos por Classe de Variância

Número de Candidatos	Classes de Variância			
	Variância Maior(Col%)	Variância Grande (Col%)	Variância Pequena (Col%)	Variância Menor(Col%)
2 Candidatos	73.2	37.9	7.1	0.9
3 a 5 Candidatos	21.4	33.0	40.9	20.1
6 a 10 Candidatos	0.9	20.5	42.7	61.6
11 ou mais Candidatos	4.5	8.5	9.3	17.4

Analisando as categorias separadamente

Na prática, as diferenças $\hat{P}_{c,p} - P_{c,p}$ devem ser chamadas erros observados e não de erro-amostral observado, pois muitos fatores não-amostrais influenciam nesses erros. A média geral absoluta desses erros, definida por $\frac{\sum_{p=1}^{N_p}\sum_{c=1}^{C_p}|\hat{P}_{c,p}-P_{c,p}|}{\sum_{p=1}^{N_p} C_p}$, foi de $2,8\%$, ou seja, esse é o erro médio cometido para cada categoria avaliada nas pesquisas eleitorais analisadas.

Nos gráficos 5.5 e 5.6, os erros observados $\hat{P}_{c,p} - P_{c,p}$ de cada categoria, foram desenhados. Na figura 5.5, percebe-se visualmente que $\hat{P}_{c,p}$ e $P_{c,p}$ são muito correlacionados. A correlação de Pearson entre as estimativas e os parâmetros de interesse é de 0.977.

No histograma em 5.6, percebe-se que há uma grande concentração de pequenos erros observados, provavelmente porque é comum nas eleições a existência de muitos candidatos sem expressão eleitoral, os quais provavelmente têm um erro observado pequeno. Os gráficos 5.5 e 5.6 apenas

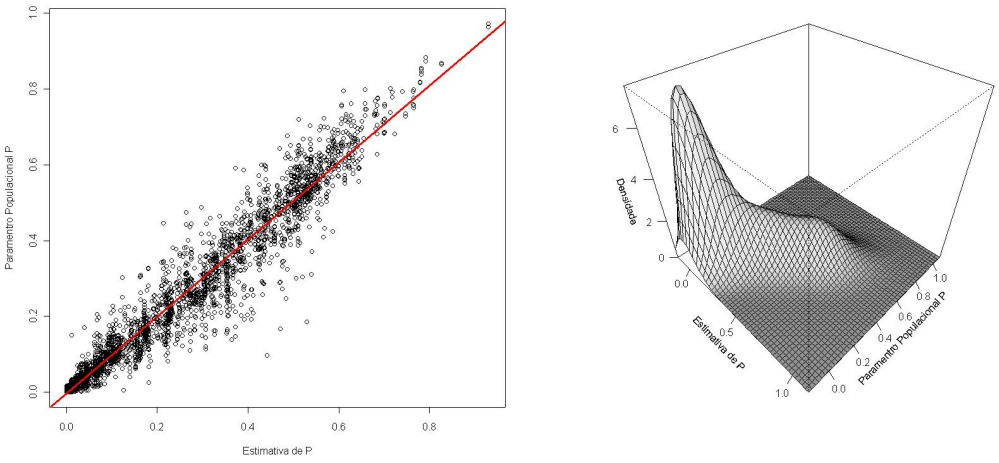

Figura 5.5: Gráfico de Dispersão dos Erros Observados

fornecem uma visão geral dos erros observados, porém eles não levam em consideração nem o desenho amostral nem os fatores não-amostrais que podem influenciar os resultados.

O critério de erro binomial é baseado na metodologia de cálculo de erro amostral utilizada pelos institutos de pesquisa, considerando cada categoria separadamente. Se todas as suposições sobre a **AAS** e o estimador Simples estiverem corretas e se não houvesse outros tipos de erro além do não-amostral, como discutimos na Seção 1.3.2, espera-se que esse valor seja 95%, porém o valor observado foi de $I_{BIN}^{0,05} = 72,7\%$. Ou seja, as pesquisas eleitorais erram $5,46$ vezes mais do que seria teoricamente esperado. No entanto, não é possível afirmar que isso é culpa do tipo de amostragem utilizada, pois não sabemos qual percentual desses erros provém de erros não-amostrais[6].

Analisando o indicador $I_{BIN}^{0,05}$ cruzado com alguns fatores de interesse, fica evidente que o erro não-amostral têm grande influência nos resultados de uma pesquisa. Quando as pesquisas foram realizadas no dia da eleição, o indicador aponta um acerto de $87,8\%$, já quando a pesquisa é realizada 20 dias ou mais antes da eleição, o percentual de acerto cai para $67,6$. Além disso, quanto maior o percentual de indecisos, de votos nulos e brancos na pesquisa, menos as pesquisas acertam, variando de $73,4\%$ a $55,6\%$. Ou seja, por causa de fatores que estão claramente relacionados a erros não-amostrais, o acerto das pesquisas chega a diminuir em 20%.

Analisando as categorias simultaneamente

Nessa seção iremos analisar a performance das pesquisas eleitorais considerando todas as categorias de uma pesquisa simultaneamente. O erro médio absoluto observado nas pesquisas, avaliado pelo indicador I_{SSRC}^{C3}, foi de $3,5\%$. Como foi mencinado anteriormente, esse indicador diminue o

[6]Se considerarmos a margem de erro exata, ou seja, $d_p = z_{(1-\frac{\alpha}{2})}\sqrt{\frac{P_{c,p}(1-P_{c,p})}{n_p}}$, o resultado é $I_{BIN}^{0,05} = 56,7\%$, muito mais intervalos de confiança deixariam de conter o parâmetro populacional. Utilizando a estimativa $\hat{d}_p = z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{P}_{c,p}(1-\hat{P}_{c,p})}{n_p}}$, o critério diminue ainda mais, com $I_{BIN}^{0,05} = 56,3\%$, resultado esse esperado, pois estamos utilizando uma estimativa ao invês do parâmetro populacional.

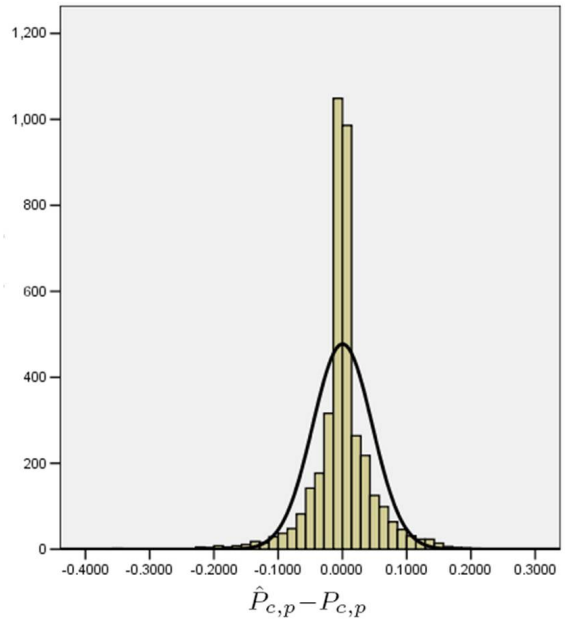

Figura 5.6: Histograma dos Erros Observados

tamanho do erro por causa da grande quantidade de candidatos inexpressivos, que acabam contribuindo somente no denominador dessa média. Já analisando o indicador $I_{SSRC}^{C3(15\%)}$ que somente considera candidatos que obtiveram pelo menos 15% dos votos na eleição, esse erro médio absoluto sobe para $4,7\%$.

Em um estudo apresentado em Mitofsky [1998], as pesquisas americanas realizadas entre os anos de 1956 e 1996 tiveram o indicador $I_{SSRC}^{C3} = 1,9\%$. Vale ressaltar que esse estudo somente considerou eleições presidenciais as quais usualmente têm somente 2 candidatos. As pesquisas presidenciais brasileiras analisadas aqui, as quais usualmente têm muitos candidatos no primeiro turno, obtêm $I_{SSRC}^{C3} = 1,8\%$, ou seja, são um pouco melhores do que as pesquisas eleitorais americanas. Fazendo uma comparação mais justa, analisando apenas as pesquisas presidencias com 2 candidatos, obtemos que $I_{SSRC}^{C3} = 3,7\%$, ou seja, as pesquisas brasileiras são melhores por causa do cenário eleitoral, e não pela qualidade das pesquisas. É importante ressaltar que outros fatores estão confundidos com esse erro observado, como por exemplo o tamanho amostral.

No histograma 5.7 podemos observar a distribuição empírica desses erros e percebe-se que ela tem caudas pesadas, ou seja, algumas pesquisas cometem erros absolutos grandes em algumas categorias.

O indicador $I_{MULT}^{0,05}$ é análogo ao indicador $I_{BIN}^{0,05}$ utilizado pelos institutos de pesquisa, porém considera todas as categorias simultaneamente. Justamente por isso, esse é um indicador muito sensível e rigoroso. Novamente, se todas as suposições sobre a **AAS** e o estimador Simples estivessem corretas e se não houvesse erro não-amostral, espera-se que esse valor seja de 90%, porém o valor observado foi de apenas 35%. Ou seja, por esse critério, as pesquisas eleitorais erram $6,6$ vezes mais do que o esperado teoricamente. Vale ressaltar novamente que esse não é o valor di-

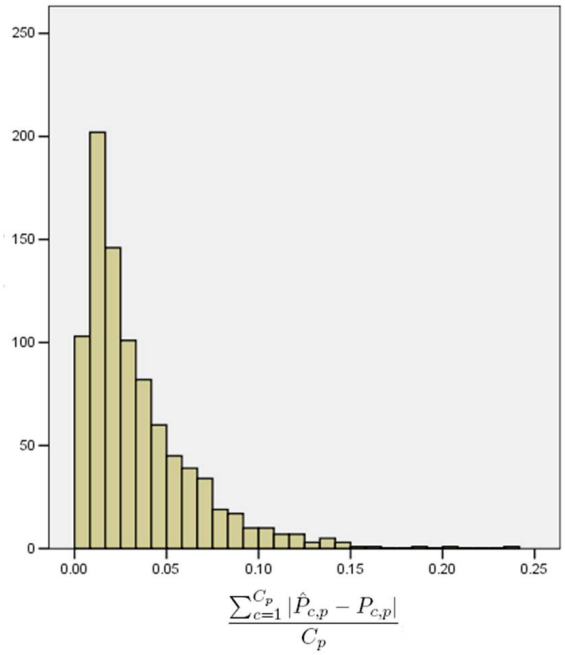

Figura 5.7: Histograma dos Erros Absolutos Médios por pesquisa

vulgado pelas empresas de pesquisa, ou seja, elas não prometem um resultado com a confiança de $I^{0,05}_{MULT}$.

Em oposição aos indicadres I^{C3}_{SSRC} e $I^{0,05}_{BIN}$, utilizando o critério $I^{0,05}_{MULT}$, quanto maior o número de candidatos na eleição, pior performance as pesquisas eleitorais têm. Outro fato interessante desse indicador é que as pesquisas parecem ter uma performance pior quando as variâncias dos estimadores diminuem. Apesar de um problema com as classes de variância ser explicado pela tabela 5.6, ou seja, que as classes de variância baixa têm em sua maioria, pesquisas com maior número de candidatos, é evidente que a tentativa de se classificar as pesquisas em classes de variância derivadas utilizando a distância de Aitchison não foi muito bem sucedida. Essa dificuldade de classificar pesquisas com mais de duas categorias será discutida novamente na Seção 5.2.3.

Os indicadores baseados em distâncias, I^{Maha}_{Dist} e I^{Aitch}_{Dist}, parecem concordar na maioria das vezes com relação aos fatores de interesse, como por exemplo, indicando que quanto mais dias antes da eleição que as pesquisas foram realizadas, pior a performance das mesmas. Não foi detectada nenhuma grande divergência entre os dois critérios, porém mesmo assim a correlação de Pearson dos dois indicadores é de apenas $0,140$.

Com relação aos indicadores de ranking, analisando o indicador $I^{Vencedor}_{Ranking}$ percebe-se que as pesquisas acertam o nome do candidato vencedor, no geral, $87,9\%$ das vezes, e considerando apenas as pesquisas relizadas no dia da eleição esse percentual sobe para $97,3\%$. Já no caso do indicador $I^{Todos}_{Ranking}$, percebe-se que as pesquisas acertam o ranking de todos os candidatos, no geral, apenas $62,8\%$ das vezes, já considerando apenas as pesquisas realizadas no dia da eleição esse percentual sobe para $78,4\%$.

De uma maneira geral, os fatores coleta no final de semana e dias de campo, não parecem influenciar a performance das pesquisas, já os fatores Cargo, Turno, votos inválidos, dias antes da eleição, tamanho amostral, classe de variância e complexidade amostral aparentam ter grande influência no resultados das pesquisas eleitorais.

5.2.3 Modelo Linear dos erros observados

Na análise descritiva realizada na Seção 5.2.2, foi possível observar que algumas fatores de interesse claramente estão correlacionados com os erros observados nas pesquisas eleitorais. A intenção dessa seção é quantificar de maneira mais precisa o impacto de cada fator de interesse nos resultados das pesquisas eleitorais. Para isso, utilizaremos um modelo linear para relacionar a variável dependente, no caso o erro observado nas pesquisas, denotada por Y, com as variáveis independentes, nesse caso os fatores de interesse, denotados por $\mathbf{X} = (X_1, ..., X_k)$, sendo que k variáveis independentes serão consideradas. Um modelo linear relaciona a variável dependente Y com as variáveis independentes $\mathbf{X}$ da seguinte forma:

$$Y_i = \alpha + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i \qquad \forall i = 1, .., N_p, \tag{5.46}$$

onde o parâmetro β_i pode ser interpretado como o aumento que ocorre na variável dependente Y ao se aumentar a variável independente X_i em uma unidade, mantendo todas as outras variáveis constantes, o parâmetro α é o intercepto da equação linear definida em 5.46, N_p é o número de pesquisas sendo analisadas e ε_i é o erro cometido pelo modelo em 5.46 para prever o valor de Y para elemento i, os quais usualmente supõem-se que são independentes e têm distribuição $\mathcal{N}(0, \sigma^2)$. O mesmo modelo pode ser escrito de forma mais enxuta, utilizando notação matricial, como $Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$. Maiores detalhes sobre modelos lineares podem ser obtidos em Graybill [1976].

Esse tipo de modelo já foi utilizado na literatura para tentar explicar os erros cometidos pelas pesquisas eleitorais americanas. Em Lau [1994], o autor apresenta o modelo ajustado considerando cada categoria da pesquisa separadamente, considerando os fatores de interesse apresentados na Seção 5.2.2 como variáveis independentes e utilizando como variável dependente o valor absoluto da diferença entre o percentual de votos válidos obtidos pelo candidato e o valor previsto pela pesquisa. O modelo obteve um R^2 ajustado de 0.24. O ponto de maior interesse nesse modelo é que o tamanho amostral não teve um efeito significativo nos erros absolutos observados, levando o autor a concluir que **"a prática comum de reportar as margens de erro baseadas somente no tamanho da amostra devem ser abandonadas porque elas nos dão uma falsa sensação de segurança. Um padrão novo e mais defensável empriricamente deve ser desenvolvido."**. Resultados similares aos obtidos em Lau [1994] também foram obtidos por Crespi [1988], no que diz respeito ao efeito do tamanho da amostra.

Já no artigo Desart and Holbrook [2003], o autor obtêm evidências de que o tamanho da

amostra têm um efeito significativo. O modelo linear utilizado considera as categorias da pesquisa simultaneamente, usando variáveis independentes similares aquelas utilizadas em Lau [1994], porém utilizando como variável dependente o crtitério de erro $|I^{C5}_{SSRC}|$, obtendo R^2 ajustado de 0.15. Ou seja, o autor apenas considerou as duas categorias mais frequentes.

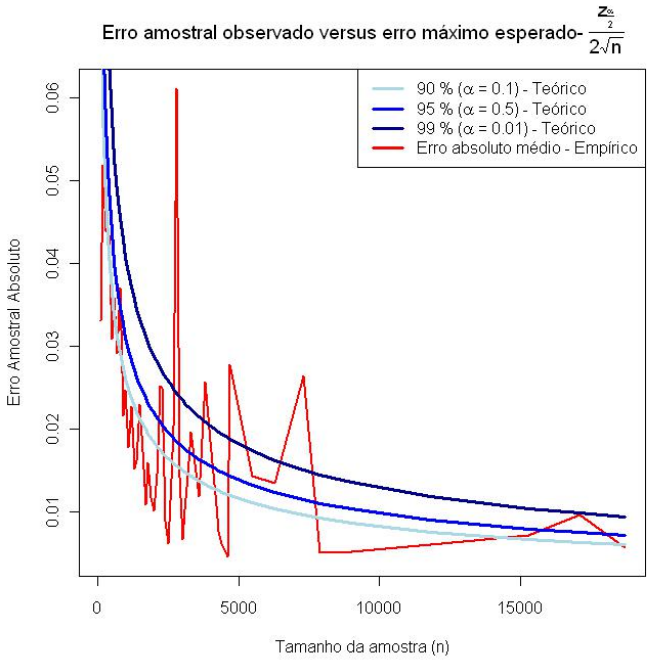

Figura 5.8: Comparação do comportamento teórico (em azul) e empírico (e vermelho) dos erros absolutos observados, segundo tamanho da amostra.

Devido a esses resultados contraditórios, haverá um interesse especial em avaliar o efeito do tamanho da amostral nos resultados das pesquisas eleitorais. No casos das 898 analisadas aqui, o compormento dos erros absolutos observados segundo o tamanho da amostra está razoavelmente alinhado com o resultado teórico em 1.24, como pode ser visualmente verificado na figura 5.8. Sabe-se, de 1.24, que o erro amostral máximo esperado ao estimar proporções, com uma amostra de tamanho n e com confiança de $(1-\alpha)\%$ é dado por $z_{\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}$. Nesse gráfico, compara-se o comportamento teórico esperado, para $\alpha \in (0.1, 0.05, 0.01)$, com o empírico observado.

Apesar disso, se ajustarmos um modelo equivalente ao utilizado em Lau [1994], o efeito do tamanho da amostra não é significativo. Isso ocorre porque o modelo linear considera que a relação entre a variável dependente e as variáveis independentes é linear, que não é o caso para o tamanho da amostra, pois como mencionado no parágrafo anterior, o erro amostral é proporcional a $\frac{1}{\sqrt{n}}$. Na tabela 5.7, o coeficiente de correlação de Pearson entre erro amostral teórico $\left(\frac{1}{\sqrt{n}}\right)$ e o tamanho da amostra foi calculado. A correlação também foi calculada para as transformações $\log n$ e $\sqrt{n}$. Além disso, foram destacados o maior tamanho de amostra (n) utilizado no cálculo das correlações.

Tabela 5.7: Correlação linear entre o erro amostral teórico $\left(\frac{1}{\sqrt{n}}\right)$ e o tamanho da amostra

Correlação Linear	Tamanho de amostra máx.		
	2000	10000	20000
n	-0.38	-0.25	-0.22
$\sqrt{n}$	-0.52	-0.35	-0.30
$\log n$	-0.85	-0.68	-0.60

É fácil perceber que quanto maior o tamanho da amostra considerada, menor a correlação. Como a relação entre n e o erro amostral não é linear, quanto maior o n mais distante as variáveis estão de uma relação linear, pois a inclinação da curva que representa o erro amostral máximo dimuinue consideravelmente, tornando-se quase paralela ao eixo horizontal, que representa o tamanho da amostra. É evidente que incluindo a variável $\frac{1}{\sqrt{n}}$ no modelo, os resultados serão mais satisfatórios, principalmente para avaliar o impacto do aumento do tamanho amostral na redução dos erros observados.

É importante notar que o valor do R^2 ajustado em todos os modelos citados nessa seção é baixo, nenhum sendo maior do que 0.24. Para melhorar o ajuste do modelo linear, serão incluídas algumas variávéis independentes que não foram incluídas nos modelos citados. São elas:

Desvio-Padrão Populacional sob AAS - $\sqrt{PQ}$ Quanto maior for a variância populacional, menos precisas serão as estimativas.

Desvio-Padrão do estimador sob AAS - $\sqrt{\frac{PQ}{n}}$ Quanto maior for a variância do estimador, menos precisas serão as estimativas. Note que essa fator é muito similar ao desvio-padrão populacional, porém também incorpora o efeito do tamanho amostral.

Tamanho amostral - $\frac{1}{\sqrt{n}}$ Como discutido anteriormente, essa variável deve ser mais importante ao modelo do que o tamanho amostral em si.

Note que as três variáveis apresentadas aqui não serão significativas simultaneamente no modelo, pois elas são muito correlacionadas. Usualmente, espera-se que sejam incluídos no modelo o par $\sqrt{PQ}$ e $\sqrt{\frac{PQ}{n}}$, ou então o par $\sqrt{PQ}$ e $\frac{1}{\sqrt{n}}$. Nessa seção, optaremos pelo segundo, pois dessa forma podemos explicitar a importância do tamanho amostral no erro observado.

Outro ponto importante sobre essas variáveis, é a dificuldade de encontrar um medida similar para a variância populacional no caso multinomial. Na Seção 5.2 apresentamos um forma de tentar classificar a variabilidade populacional de uma variável multinomial considerando a distância de Aitchison, porém ela não parece ter sido bem sucedida. Pela importância que a variância populacional têm para explicar o erro amostral e pela dificuldade de de encontrar um medida resumo para a mesma no caso multinomial, um modelo para o caso multinomial não será ajustado.

Outra questão relevante é a variável dependente considerada. No modelo descrito em Lau [1994] foi utilizado o erro absoluto observado. Existe um problema importante relacionado a essa variável, pois ela está restrita apenas ao intervalo $[0, 1]$, porém o modelo linear, como apresentado

em 5.46, pode assumir valores em $(-\infty, \infty)$. Além disso, a distribuição empírica dos erros absolutos observados está muito distante da distribuição normal, como pode ser visto na figura 5.9. Para corrigir parcialmente esse problema, a variável dependente considerada será $-log(Erro)$, que pode assumir valores em $[0, \infty)$ e têm uma distribuição empírica mais similar a distribuição normal apesar da dsitribuição ainda ser claramente assimétrica, como pode ser visto na figura 5.9. Essa transformação ainda permite que valores fora do intervalo $[0, 1]$ sejam obtidos na escala original, assim não solucionando completamente esse problema, porém nos dados sendo analisados não houve nenhuma ocorrência desse tipo. Uma outra tranformação que poderia ser utilizada é a $-log(\frac{Erro}{1-Erro})$, similar a função logito, utilizada em regressão logística, com a diferença de que no contexto de regressão logística utiliza-se o logito para modelar o parâmetro não-observado p, e não a variável dependente diretamente, como estamos fazendo aqui. A vantagem dessa tranformação é que ela assume valores em $(-\infty, \infty)$, sendo compatível, nesse aspecto, com o modelo linear geral. A desvantagem é que a interpretação dos efeitos das variáveis independentes nos erros absolutos observados na escala original é mais complicada. Na figura 5.9, podemos ver que a distribuição dos dados com essa tranformação é muito parecida com a distribuição obtida da tranformação $-log(Erro)$.

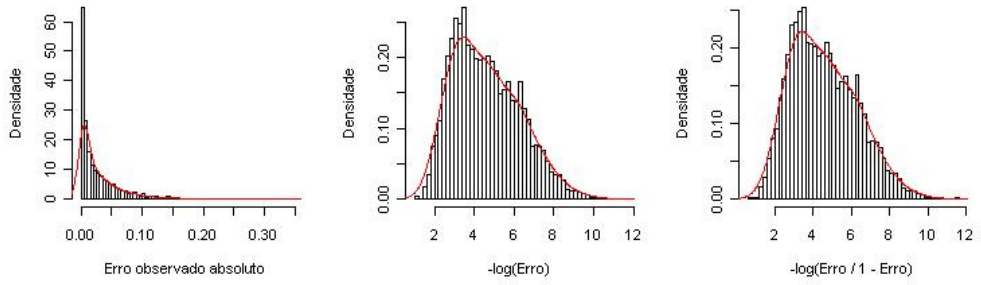

Figura 5.9: Histogramas dos erros observados absolutos e de suas trasnsformações.

Nessa seção, optamos por utilizar a tranformação $-log(Erro)$, pois apesar do modelo permitir valores incompatíveis com a variável dependente transformada, a interpretação dos efeitos é mais fácil, sendo que o R^2-ajustado, os p-valores e os sinais dos betas são iguais em ambos os modelos. Além disso, em ambos os modelos, a análise de resíduos realizada para verificar se as suposições do modelo linear são satisfeitas têm resultados similares e satisfitórios. O modelo na escala original, nesse contexto, pode ser escrito como:

$$Y_i = e^{-\alpha} e^{-\beta_1 X_{1i}} \dots e^{\beta_k X_{ki}} \qquad \forall i = 1, .., N_p, \tag{5.47}$$

Assim, para interpretar o efeito de cada variável na escala original dos erros observados absolutos, de 5.47 podemos interpretar $\left(e^{-\beta_k}\right)^x$, como o valor médio do erro amostral quando a variável $X_k = x$, mantendo-se todas as outras variáveis constantes. Apesar do efeito não ser linear, para

qualquer valor x que a variável independente assuma, se $e^{-\beta} > 1$, ao se aumentar o valor da variável independente, os erros observados aumentam, e se $e^{-\beta} < 1$, os erros diminuem. Esse será o enfoque principal na apresentação dos resultados nessa seção. Note que quando a variável independente assume apenas os valores 0 ou 1, como é o caso das variáveis indicadoras, podemos interpretar $\left(e^{-\beta_k}\right)$ como o aumento médio que ocorre no erro amostral quando a variável $X_k = 1$, mantendo-se todas as outras variáveis constantes. Esse efeito é multiplicativo, assim, se $e^{-\beta} = 1,2$ quer dizer que o aumento é de 20%, já no caso onde $e^{-\beta} = 0,8$ quer dizer há uma redução de 20% no erro absoluto observado quando a variável independente correspondente assume valor 1.

Para facilitar a interpretação de todas as variáveis independetes utilizadas no modelo, as variáveis baseadas em percentuais, como percentual de indecisos e percentual de dias de campo realizados no fim de semana, foram multiplicadas por 100, assim podemos interpretar os coeficientes tranformados como o impacto ao se aumentar essas variáveis em um ponto percentual. A variável desvio-padrão populacional também foi multiplicada por 100, assim podemos também interpretar o seu coeficiente da mesma forma, com a exceção de que essa variável está limitada ao intervalo $[0, 50]$. No caso específico da variável $\frac{1}{\sqrt{n}}$ não é possível obter uma interpretação simplificada como no caso das outras variáveis quando o interesse é avaliar o efeito da variável original n. Nesse caso, temos que $e^{-\hat{\beta}_n \frac{1}{\sqrt{n}}}$, e o impacto na variável dependente Y de se aumentar n em uma unidade, de 100 para 101, por exemplo, é diferente do efeito quando aumentamos n de 1000 para 1001. Assim, para interpretar o efeito em Y de se aumentar a variável n em uma unidade, é necessário escolher um valor específico para n, digamos n_0, e o efeito é estimado calculando-se $e^{-\frac{\hat{\beta}_n}{\sqrt{n_0}}} - e^{-\frac{\hat{\beta}_n}{\sqrt{(n_0+1)}}}$.

Finalmente, a última questão que merece atenção é com relação a independência dos erros. Sabemos da Seção 5.2.1 que os estimadores $\hat{P}_{c',p}$ e $\hat{P}_{c,p}$ de uma mesma pesquisa, no caso da AAS, têm variância dada por $\frac{-P_{c,p}(1-P_{c,p})}{n_p}$ e covariância dada por $\frac{-P_{c',p}P_{c,p}}{n_p}$. Assim, ao invés de supor independência entre os erros, um modelo mais realista seria obtido considerando essa estrutura de correlação ao estimar os parâmetros do modelo. Idealmente, a estrutura de correlação do desenho amostral efetivamente utilizado deveria ser utilizada, mas como nesse caso a mesma é desconhecida, aproximá-la pela estrutura da AAS é uma suposição menos forte do que supor independência entre os erros. Nesse caso, supondo que a matriz de covariâncias Σ seja conhecida, podemos considerar o modelo $Y \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{\Sigma})$. A matriz Σ é dada por:

$$\Sigma = \begin{pmatrix} \Sigma_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_{N_p} \end{pmatrix}, \tag{5.48}$$

sendo que as sub-matrizes Σ_p são:

$$\Sigma_p = \begin{pmatrix} \frac{P_{1,p}(1-P_{1,p})}{n_p} & \frac{-P_{1,p}P_{2,p}}{n_p} & \cdots & \frac{-P_{1,p}P_{c_p-1,p}}{n_p} \\ \frac{-P_{2,p}P_{1,p}}{n_p} & \frac{P_{2,p}(1-P_{2,p})}{n_p} & \cdots & \frac{-P_{2,p}P_{c_p-1,p}}{n_p} \\ \vdots & & \ddots & \vdots \\ \frac{-P_{c_p-1,p}P_{1,p}}{n_p} & \cdots & & \frac{P_{c_p-1,p}(1-P_{c_p-1,p})}{n_p} \end{pmatrix}, \tag{5.49}$$

onde c_p é o número de categorias da pesquisa p, ou seja, cada sub-matriz quadrada Σ_p tem um dimensão diferente, dada por $c_p - 1$, pois estamos desconsiderando as categorias com a menor proporção populacional de cada pesquisa (na definição das matrizes em 5.49, estamos supondo que a menor proporção ocorre na última categoria c_p). Nesse caso, retirar a menor categoria de cada pesquisa como discutido no inicío da Seção 5.2.2 não é uma opção, e sim uma necessidade, pois mantendo-se todas as categorias de uma pesquisa, as sub-matrizes Σ_p não têm inversa. E consequentemente, a matriz Σ também não tem inversa, pois a matriz Σ que é bloco-diagonal, e consequentemente a sua inversa é dada por:

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_1^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_2^{-1} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_{N_p}^{-1} \end{pmatrix}. \tag{5.50}$$

Procedendo dessa forma, considerando a matriz Σ das covariâncias, a estimativa usual de mínimos quadrados dada por $\hat{\beta} = (X'X)^{-1}X'Y$ é substituída pela sua versão generalizada, dada por $\hat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$[7]. Para obter o estimador de mínimos quadrados generalizados, pode se utilizar o fato de que a matriz Σ é definida positiva, o que implica que existe uma matriz P, tal que $\Sigma = PP'$ e $\Sigma^{-1} = P'^{-1}P^{-1}$. Transformando a variável dependende Y do modelo correlacionado em $Z = P^{-1}Y$, obtemos que o modelo transformado $Z \sim \mathcal{N}(\mathbf{Q}\beta, \sigma^2\mathbf{I})$ é independente, onde $Q = P^{-1}X$. Utilizando o modelo transformado, pode-se calcular o estimador de β de mínimos quadrados do modelo transformado como sendo $(Q'Q)^{-1}Q'Z$, o qual é equivalente ao estimador de mínimos quadrados generalizados $(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$ no modelo original.

Para permitir a comparação dos dois modelos, as estimativas obtidas para o modelo com erros independentes $-log(Erro) \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$ e para o modelo com erros correlacionados $-log(Erro) \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{\Sigma})$ são apresentadas na tabela 5.8.

Dos resultados na tabela 5.8, percebe-se que no modelo com erros independentes, as únicas variáveis que não são significativas a um nível de 10% e que possuem um efeito muito pequeno, são o número de dias de campo e a proporção dos dias de campo que foram realizados no fim de

[7]É importante mencionar que diferentes pesquisas realizadas para prever o resultado de uma mesma eleição são independentes, pois as amostras para cada uma foram obtidas de forma independente (análogo a amostragem estratificada), apesar dessas pesquisas terem uma sub-matriz de covariância similar, diferindo somente no tamanho da amostra.

semana, ou seja, a forma como a coleta de dados é realizada não parece afetar a qualidade das pesquisas. Esse resultado pode ocorrer porque o número de dias de campo realmente não tem relevância, ou pode ser fruto de estarmos comparando apenas dois institutos, os quais podem ter uma metodologia de coleta de dados que é, na medida do possível, replicada em todas as pesquisas. Além disso, qualquer diferença nas metodologias de coleta de dados dos institutos é absorvida na variável identificadora do instituto, a qual têm efeito significativo a um nível de 1% no modelo independente. Outra explicação, talvez mais plausível, é que as pesquisas com menor erro observado são realizadas no dia da eleição, o que implica que elas têm que ser realizadas em apenas um dia de campo, e esse efeito de confundimento pode estar anulando o efeito do número de dias de campo.

Os fatores relacionados a erros não-amostrais têm resultados bastante interessantes, para x dias a mais antes da eleição, o erro observado aumenta em $1,02^x$ e para cada aumento de x pontos percentuais no número de indecisos, votos brancos e nulos, o erro observado aumenta em $1,03^x$. No caso do fator número de candidatos (categorias), para x candidatos na disputa, em média, o erro observado decresce em $0,99^x$. Já quando a eleição é do segundo turno, o erro observado reduz em média 45%, porém essa resultado conflitante está confundido com o número de categorias, pois toda eleição de segundo turno têm apenas dois candidatos. O que causa esse resultado inesperado é que apenas 13 pesquisas do primeiro turno têm somente 2 candidatos, e o erro cometido nessas pesquisas é, em média, de 6%, bem maior do que o erro cometido pelas pesquisa do segundo turno, que é de $3,6\%$.

Tabela 5.8: Estimativas dos parâmetros do Modelo Linear

Variáveis Independentes	Erros Independentes (R^2 ajustado de 56%)				Erros correlacionados (R^2 ajustado de 51%)			
	β	p-valor	$\exp(-\beta)$	Efeito %	β	p-valor	$\exp(-\beta)$	Efeito %
Intercepto	6,895	0,0000	0,00	− − −	10,907	0,000	0,00	− − −
Cargo - Prefeito	0,395	0,0000	0,67	-32,6	0,813	0,000	0,44	-55,64
Cargo - Presidente	0,328	0,0000	0,72	-27,9	0,060	0,035	0,94	-5,87
Instituto - I	0,149	0,0070	0,86	-13,9	0,219	0,000	0,80	-19,71
Dias de campo	-0,008	0,6668	1,01	0,8	-0,076	0,000	1,08	7,89
Dias antes da eleição	-0,018	0,0000	1,02	1,8	-0,016	0,000	1,02	1,64
Percentual de votos não-válidos	-0,027	0,0006	1,03	2,8	-0,059	0,000	1,06	6,11
Amostra Ponderada	0,116	0,0734	0,89	-11,0	0,041	0,125	0,96	-4,05
Número de categorias	0,014	0,0804	0,99	-1,4	-0,165	0,000	1,18	17,88
Segundo Turno	0,598	0,0000	0,55	-45,0	1,245	0,383	0,29	-71,21
Prop. de dias de campo no fim de semana	0,000	0,7221	1,00	0,0	-0,004	0,000	1,00	0,35
$\frac{1}{\sqrt{n}}$	-17,370	0,0000	− − −	$\mathbf{-0,353}^*$	-26,104	0,000	− − −	$\mathbf{-0,678}^*$
$\sqrt{PQ}$	-0,067	0,0000	1,07	6,9	-0,152	0,000	1,16	16,45

*** O efeito foi calculado como sendo $e^{-\frac{\hat{\beta}}{\sqrt{\overline{n}}}} - e^{-\frac{\hat{\beta}}{\sqrt{(\overline{n}+1)}}}$, onde $\overline{n} = 1246$.**

Analisando o tamanho da amostra, a interpretação do efeito do mesmo é diferente dos outros casos, pois no modelo foi incluída a tranformação $\frac{1}{\sqrt{n}}$. Tranformando esse fator de volta a escala original, obtemos que seu efeito é dado por $e^{-\frac{\hat{\beta}}{\sqrt{n}}} - e^{-\frac{\hat{\beta}}{\sqrt{(n+1)}}}$. Como essa função é decrescente em n, fica evidente que quanto maior for o tamanho da amostra, menor será o erro observado. Além disso, esse efeito não é linear em n, ou seja, o efeito causado no erro observado absoluto ao aumentar o tamanho da amostra em 1 unidade depende qual n está sendo usado como referência. Por exemplo, considerando o menor n observado, que foi 187, obtemos um efeito multiplicativo de $-1,02$ ao se aumentar em uma unidade o tamanho amostral, porém ao avaliar o efeito em um n próximo de

$\overline{n}$ (1246) obtemos um efeito bem menor, de $-0,353$. Finalmente, é evidente a importância que o desvio-padrão populacional têm no erro absoluto observado, aumentando o erro absoluto observado em $1,07^x$ quando a quantidade $100\sqrt{PQ}$ for igual a x, a qual pode assumir o valor máximo de 50. A variância populacional é uma das grandes responsáveis por obter um R^2-ajustado de 56%, ao retirá-la do modelo, o R^2-ajustado reduz para 21%. Em contraste, ao retirar o tamanho da amostra, por exemplo, R^2-ajustado reduz somente $0,5$%.

O resultado interessante na tabela 5.8 é a redução de 5% do R^2-ajustado do modelo com erros correlacionados (51%), se comparado com o do modelo de erros independentes (56%). Ou seja, ao considerar que os erros são correlacionados, o modelo perde um pouco do poder explicativo, porém continua tendo um R^2-ajustado maior do que os outros modelos da literatura.

Existem outras diferenças evidentes ao comparar os dois modelos, porém em menor grau. As diferenças que merecem destaque são, o número de dias de campo, que pelo modelo independente não tinha efeito significativo, passou a ser significativo e o seu efeito quando consideramos x dias passou de $1,01^x$ para $1,08^x$. O efeito dos diferentes cargos também foi alterado, aumentando bastante a diferença entre os cargos de Prefeito e de Presidente. O efeito do percentual de votos não-válidos aumentou no modelo com erros correlacionados, passando de $1,03^x$ para $1,06^x$. O efeito do número de categorias mudou o sinal do coeficiente, implicando que o efeito de $0,99^x$ no modelo independente passou para $1,18^x$ no modelo com erros correlacionados, sendo mais coerente com o efeito do segundo-turno (sinais trocados), porém implicando em uma grande diferença entre os dois modelos, talvez porque as pesquisas com muitas categorias sejam justamente onde o efeito da correlação entre as categorias é mais forte e não deve ser ignorado. Por último, o efeito do tamanho da amostra se torna-se mais forte e o do desvio-padrão populacional também. Ou seja, utilizando o modelo com erros correlacionados, apesar do poder de explicação do modelo diminuir um pouco, a efeito de variáveis com importância prática aumentou.

Capítulo 6

Conclusões

Para facilitar a discussão, apresentaremos novamente as quatro principais críticas que usualmente são feitas as pesquisas eleitorais:

1 - Seleção da Amostra Não-Probabilística A metodologia de seleção da amostra, usualmente utilizando-se cotas quando a **ID** supõe seleção probabilística.

2 - Inferência baseada na AAS Ao analisar os resultados da amostra, não se leva em conta o desenho amostral.

3 - Desconsiderar a Correlação entre Candidatos Ao analisar os resultados da amostra, não leva-se em conta a correlação entre categorias da multinomial, ou seja, em média, quanto maior o percentual de votos de um candidato, menor o percentual de votos nos outros candidatos.

4 - Empate Técnico entre Candidatos A ocorrência de empates técnicos entre candidatos, ou seja, depois de observada a amostra, não ser possível inferir qual candidato está na frente.

A **crítica** 1 com relação a seleção não-probabilística da amostra é apenas relevante quando se faz inferência do tipo **ID** e/ou quando o mecanismo de seleção da amostra não é ignorável. Porém, no Capítulo 4 mostramos que a *amostragem probabilística com cotas (**APC**) pode ser vista como uma amostragem probabilística com probabilidades desiguais, sob a suposição de que as probabilidades de resposta são constantes dentro de cada cota, respeitando o modelo **GRH***. Ou seja, para pesquisas eleitorais utilizando **APC**, essa crítica deixa de ser relevante se a suposição acima mencionada estiver correta. *Para que as suposições do modelo **GRH** sejam mais próximas da realidade, é importante que os institutos de pesquisas considerem as probabilidades de resposta ao definir as variáveis de cota que serão utilizadas nas pesquisas.*

Vimos na Seção 5.2 ao ajustar o modelo linear dos erros observados, que existe um efeito significativo para o tamanho da amostra, e que o fator mais importante para explicar os erros observados é a variância populacional. O interessante nesses resultados é que ambos esses fatores estão relacionados com a eficiência dos estimadores obtidos sob amostragem probabilística, ou seja, *apesar do tipo de amostragem utilizada pelos institutos de pesquisa não ser probabilística, os resutados evidenciam que o comportamento das pesquisas eleitorais têm propriedades similares aquelas da amostragem probabilística.*

Já a **crítica** 2 relacionada a suposição de que a amostragem é **AAS** usualmente não têm muito importância, pois isso apenas implica que os estimadores podem ser viciados e consequentemente os intervalos de confiança podem não ter a cobertura desejada, porém não é possível afirmar que o EQM será maior do que ao utilizar os estimadores não-viciados para o desenho amostral em questão. Existem evidências empíricas, obtidas na simulação da Seção 5.1, de que os estimadores simples da **AAS** podem ter o EQM menor. *Se as probabilidades de seleção forem aproximadamente independentes das quantidades populacionais de interesse, então os estimadores pontuais não são viciados, e a cobertura dos intervalos de confiança das estimativas será próxima da esperada, mesmo utilizando o estimador baseado na estatística s^2, a qual desconsidera as probabilidades de seleção*, como foi visto nas seções 5.1.4 e 5.1.3. É importante destacar que mesmo no caso onde as probabilidades de seleção são independentes das quantidades populacionais de interesse, levar em consideração se foi feita amostragem por conglomerados e estimar as variâncias dos diferentes estágios de seleção pode ter um grande impacto na cobertura dos intervalos de confiança, pois é uma forma de levar em conta os possíveis efeitos de conglomeração, conforme foi discutido em 1.2.4.

Analisando as pesquisas eleitorais realizadas no Brasil e considerando as categorias separadamente (como é feito pelos institutos de pesquisa), a cobertura dos intervalos baseados na AAS está próxima daquela teoricamente esperada, principalmente nas pesquisas realizadas no dia da eleição. Esses resultados são apresentados na Seção 5.2.2. Esse fato pode ser visto como evidência de que no caso das pesquisas eleitorais realizadas no Brasil, as probabilidades de seleção são, no geral, independentes das intenções de voto declaradas. Ou seja, *existe evidência de que os desenhos amostrais utilizados são ignoráveis, implicando que os estimadores de AAS devem ser utilizados pois possuem o menor EQM.*

Com relação a **crítica 3** relacionada a desconsiderar a correlação entre os candidatos, fica claro que essa é uma crítica relevante, pois os institutos de pesquisa fazem uma suposição sabidamente errada de independência entre os estimadores e ao fazer inferência dessa forma os institutos aumentam a incidência de empates técnicos. Além disso, é possível diminuir a incidência do mesmo se fossem utilizadas as estimativas das variâncias dos estimadores, ao invês dos valores máximos possíveis. O impacto efetivo dessa crítica não é muito alto do ponto de vista de confiança nos resultados das pesquisas, é apenas dispendioso, pois gasta-se muito dinheiro para não se chegar a uma conclusão. Por causa desses argumentos, os institutos deveriam utilizar estimadores que levem em consideração a correlação entre as categorias da multinomial, conforme discutido na Seção 1.2.2.

No caso da **crítica 4**, a existência de empates técnicos pode ser totalmente evitada, utilizando-se inferência Bayesiana baseada em Modelos (**IBM**). Já no caso da inferência baseada no desenho (**ID**) e da inferência baseada em modelos (**IM**), o empate técnico é inevitável, porém pode ser menos frequente se as correlações entre as categorias forem levadas em consideração.

De maneira geral, do ponto de vista empírico, apesar da grande quantidade de pesquisas avaliadas, é impossível afirmar categoricamente se as pesquisas eleitorais realizadas no Brasil respeitam as margens de erro divulgadas ou não. Essa dificuldade ocorre pois não é possível afirmar quanto do erro observado provém de erros amostrais e quanto de erros não-amostrais. Dos resultados

apresentados, é evidente a importância que fatores relacionados com erros não-amostrais, como dias antes da eleição e percentual de pessoas indecisas, têm na qualidade das pesquisas. Apesar disso, considerando as categorias de cada pesquisa separadamente, a performance das pesquisas é razoável quando se reduz ao máximo as fontes de erros não-amostrais. Também é de interesse mencionar que se fossem consideradas ao se fazer inferência, as variâncias reais dos estimadores e não o valor máximo que elas assumem, a performance das pesquisas seria bastante inferior, como pode ser visto na Seção 5.2.2.

Em contra-partida, é evidente que as margens de erro divulgadas pelos institutos não consideram características importantes, principalmente a correlação entre os estimadores da intenção de voto nos diferentes candidatos de uma mesma eleição. Levando em consideração todas as categorias simultaneamente, é muito difícil acreditar, mesmo sabendo da existência de erros não-amostrais, que as pesquisas respeitem as margens de erro teóricas, sempre lembrando que em todas as análises realizadas nesse texto, a distribuição amostral dos estimadores foi aproximada, pois as probabilidades de inclusão são desconhecidas.

Em conclusão, é possível calcular sob algumas suposições, as probabilidades de seleção da amostragem probabilística por Cotas, permitindo que inferência do tipo **ID**, utilizando o estimador de HH seja seja realizada. Porém, não é necessário conhecer as probabilidades de seleção/inclusão para se obter um estimador eficiente do ponto de vista de **ID**, basta que elas existam e sejam independentes da quantidade populacional de interesse. Quanto a existência, é importante utilizar um desenho amostral de forma que todas as unidades populacionais tenham chance de pertencer a amostra, ou pelos menos, que o fato de uma unidade populacional ter probabilidade nula de pertencer a amostra não esteja relacionado com o quantidade populacional de interesse dessa unidade. Ou seja, mesmo sem conhecer as probabilidades de seleção da maioria dos tipos de amostragem por cotas, é possível fazer inferência baseada no Desenho. Com relação a obter um estimador com boas propriedades teóricas (aproximadamente não-viciado e com variância pequena), as probabilidades de seleção/inclusão do desenho amostral não precisam ser utilizadas para se estimar a quantidade populacional de interesse, basta que elas sejam aproximadamente independentes da quantidade populacional de interesse. Quando essa independência não ocorrer, os intervalos de confiança dos parâmetros não terão a cobertura declarada por causa do vício do estimador. E finalmente, se existe o interesse em diminuir a ocorrência de empates técnicos, sempre deve-se considerar a correlação entre os estimadores sendo estudados, e se a intenção for eliminar a ocorrência dos mesmos, a única opção é utilizar inferência Bayesiana.

Apêndice A

Legislação das Pesquisas Eleitorais

TRIBUNAL SUPERIOR ELEITORAL

RESOLUÇÃO Nº 23.190

INSTRUÇÃO Nº 127 – CLASSE 19ª – BRASÍLIA – DISTRITO FEDERAL.

Relator: Ministro Arnaldo Versiani.
Interessado: Tribunal Superior Eleitoral.

Dispõe sobre pesquisas eleitorais (Eleições de 2010).

O Tribunal Superior Eleitoral, usando das atribuições que lhe conferem o art. 23, inciso IX, do Código Eleitoral e o art. 105 da Lei nº 9.504, de 30 de setembro de 1997, resolve expedir a seguinte instrução:

CAPÍTULO I

DISPOSIÇÕES PRELIMINARES

Art. 1º A partir de 1º de janeiro de 2010, as entidades e empresas que realizarem pesquisas de opinião pública relativas às eleições ou aos candidatos, para conhecimento público, são obrigadas, para cada pesquisa, a registrar no tribunal eleitoral ao qual compete fazer o registro dos candidatos, com no mínimo 5 dias de antecedência da divulgação, as seguintes informações (Lei nº 9.504/97, art. 33, I a VII, e § 1º):

I – quem contratou a pesquisa;

II – valor e origem dos recursos despendidos no trabalho;

III – metodologia e período de realização da pesquisa;

IV – plano amostral e ponderação quanto a sexo, idade, grau de instrução e nível econômico do entrevistado; área física de realização do trabalho, intervalo de confiança e margem de erro;

V – sistema interno de controle e verificação, conferência e fiscalização da coleta de dados e do trabalho de campo;

VI – questionário completo aplicado ou a ser aplicado;

VII – nome de quem pagou pela realização do trabalho;

VIII – contrato social, estatuto social ou inscrição como empresário, que comprove o regular registro da empresa, com a qualificação completa dos responsáveis legais, razão social ou denominação, número de inscrição no Cadastro Nacional da Pessoa Jurídica (CNPJ), endereço, número de fac-símile em que receberão notificações e comunicados da Justiça Eleitoral;

IX – nome do estatístico responsável pela pesquisa – e o número de seu registro no competente Conselho Regional de Estatística –, que assinará o plano amostral de que trata o inciso IV retro e rubricará todas as folhas (Decreto nº 62.497/68, art. 11);

X – número do registro da empresa responsável pela pesquisa no Conselho Regional de Estatística, caso o tenham.

§ 1º Até 24 horas contadas da divulgação do respectivo resultado, o pedido de registro será complementado pela entrega dos dados relativos aos municípios e bairros abrangidos pela pesquisa; na ausência de delimitação do bairro, será identificada a área em que foi realizada a pesquisa.

§ 2º O arquivamento da documentação a que se refere o inciso VIII deste artigo, na secretaria judiciária do tribunal eleitoral competente, dispensa a sua apresentação a cada pedido de registro de pesquisa, sendo, entretanto, obrigatória a informação de qualquer alteração superveniente.

§ 3º As entidades e empresas deverão informar, no ato do registro, o valor de mercado das pesquisas que realizarão por iniciativa própria.

Art. 2º A contagem do prazo de que cuida o *caput* do art. 1º desta resolução far-se-á excluindo o dia de começo e incluindo o do vencimento (Código de Processo Civil, art. 184).

Parágrafo único. Os pedidos de registro enviados após às 19 horas ou, no período eleitoral, após o horário de encerramento do protocolo geral do tribunal eleitoral competente, serão considerados como enviados no dia seguinte.

Art. 3º A partir de 5 de julho de 2010, das pesquisas realizadas mediante apresentação da relação de candidatos ao entrevistado, deverá constar o nome de todos aqueles que tenham solicitado registro de candidatura.

CAPÍTULO II

DO REGISTRO DAS PESQUISAS ELEITORAIS

Seção I

Do Sistema Informatizado de Registro de Pesquisas Eleitorais

Art. 4º Para o registro de que trata o art. 1º desta resolução, deverá ser utilizado o Sistema Informatizado de Registro de Pesquisas Eleitorais disponível nos sítios dos tribunais eleitorais.

§ 1º Para a utilização do sistema as entidades e empresas deverão cadastrar-se por meio eletrônico, não permitido mais de um registro por número de inscrição no Cadastro Nacional da Pessoa Jurídica (CNPJ), sendo elementos obrigatórios do cadastro o nome dos responsáveis legais, razão social ou denominação, número de inscrição no CNPJ, endereço e número de fac-símile em que poderão receber notificações.

§ 2º É de inteira responsabilidade da empresa ou entidade a manutenção de dados atualizados perante a Justiça Eleitoral.

§ 3º O sistema possibilitará o cadastro prévio dos dados pela entidade ou empresa e gerará o documento que deverá ser protocolado perante a Justiça Eleitoral.

§ 4º Para verificação de atendimento aos prazos estabelecidos nesta resolução, as secretarias judiciárias observarão, exclusivamente, a data e horário de protocolo da documentação entregue em meio impresso.

Art. 5º As informações e dados registrados no sistema serão colocados à disposição, pelo prazo de 30 dias, no sítio do respectivo tribunal (Lei nº 9.504/97, art. 33, § 2º).

Seção II

Do Processamento do Registro das Pesquisas Eleitorais

Art. 6º O pedido de registro de pesquisa deverá ser dirigido:

I – ao Tribunal Superior Eleitoral, na eleição presidencial;

II – aos tribunais regionais eleitorais, nas eleições federais e estaduais.

Art. 7º O pedido de registro, gerado pelo sistema informatizado de que trata o art. 4º desta resolução, poderá ser enviado por fac-símile, ficando dispensado o encaminhamento do documento original.

Parágrafo único. O envio do requerimento por fac-símile e sua tempestividade serão de inteira responsabilidade do remetente, correndo por sua conta e risco eventuais defeitos.

Art. 8º Apresentada a documentação a que se refere o art. 1º desta resolução, a secretaria judiciária do tribunal eleitoral competente receberá o pedido de registro como expediente, devidamente protocolado sob número, que será obrigatoriamente consignado na oportunidade da divulgação dos resultados da pesquisa.

Parágrafo único. Não deverão ser juntadas aos autos folhas de fac-símile impressas em papel térmico, devendo a secretaria judiciária, nessa hipótese, providenciar cópia para fins de juntada.

Art. 9º Caberá às secretarias judiciárias, no prazo de 24 horas contadas do recebimento, conferir toda a documentação e afixar, em local previamente reservado para este fim, bem como divulgar no sítio do tribunal eleitoral na internet, aviso comunicando o registro das informações apresentadas, colocando-as à disposição dos partidos políticos ou coligações com candidatos ao pleito, os quais a elas terão livre acesso pelo prazo de 30 dias (Lei nº 9.504/97, art. 33, § 2º).

§ 1º Constatada a ausência de quaisquer das informações exigidas no art. 1º desta resolução, a secretaria judiciária notificará o requerente para regularizar a respectiva documentação, em até 48 horas.

§ 2º Transcorrido o prazo de que trata o parágrafo anterior, sem que a entidade ou empresa regularize o pedido de registro, será a pesquisa declarada insubsistente.

Seção III

Da Divulgação dos Resultados

Art. 10. Na divulgação dos resultados de pesquisas, atuais ou não, serão obrigatoriamente informados:

I – o período de realização da coleta de dados;

II – a margem de erro;

III – o número de entrevistas;

IV – o nome da entidade ou empresa que a realizou, e, se for o caso, de quem a contratou;

V – o número do processo de registro da pesquisa.

Art. 11. As pesquisas realizadas em data anterior ao dia das eleições poderão ser divulgadas a qualquer momento, inclusive no dia das eleições (Constituição Federal, art. 220, § 1º).

Art. 12. A divulgação de levantamento de intenção de voto efetivado no dia das eleições far-se-á da seguinte forma:

a) nas eleições relativas à escolha de deputados estaduais e federais, senador e governador, uma vez encerrado o escrutínio na respectiva unidade da Federação;

b) na eleição para a Presidência da República, tão logo encerrado, em todo o território nacional, o pleito.

Art. 13. Mediante requerimento ao tribunal eleitoral competente, os partidos políticos poderão ter acesso ao sistema interno de controle, verificação e fiscalização da coleta de dados das entidades e das empresas que divulgaram pesquisas de opinião relativas aos candidatos e às eleições, incluídos os referentes à identificação dos entrevistadores e, por meio de escolha livre e aleatória de planilhas individuais, mapas ou equivalentes,

confrontar e conferir os dados publicados, preservada a identidade dos entrevistados (Lei nº 9.504/97, art. 34, § 1º).

Parágrafo único. Além dos dados de que trata o *caput*, poderá o interessado ter acesso ao relatório entregue ao solicitante da pesquisa e ao modelo do questionário aplicado para facilitar a conferência das informações divulgadas.

Art. 14. Na divulgação de pesquisas no horário eleitoral gratuito devem ser informados, com clareza, o período de sua realização e a margem de erro, não sendo obrigatória a menção aos concorrentes, desde que o modo de apresentação dos resultados não induza o eleitor a erro quanto ao desempenho do candidato em relação aos demais.

Seção IV

Das Impugnações

Art. 15. O Ministério Público Eleitoral, os candidatos e os partidos políticos ou coligações estão legitimados para impugnar o registro e/ou divulgação de pesquisas eleitorais perante o tribunal competente, quando não atendidas as exigências contidas nesta resolução e no art. 33 da Lei nº 9.504/97.

Art. 16. Havendo impugnação, o pedido de registro será autuado como representação e distribuído a um relator que notificará imediatamente o representado, por fac-símile, para apresentar defesa em 48 horas (Lei nº 9.504/97, art. 96, *caput* e § 5º).

Parágrafo único. Considerando a relevância do direito invocado e a possibilidade de prejuízo de difícil reparação, o relator poderá determinar a suspensão da divulgação dos resultados da pesquisa impugnada ou a inclusão de esclarecimento na divulgação de seus resultados.

CAPÍTULO III

DA PENALIDADE ADMINISTRATIVA

Art. 17. A divulgação de pesquisa sem o prévio registro das informações constantes do art. 1º desta resolução sujeita os responsáveis à

multa no valor de R$ 53.205,00 (cinqüenta e três mil duzentos e cinco reais) a R$ 106.410,00 (cento e seis mil quatrocentos e dez reais) (Lei nº 9.504/97, art. 33, § 3º).

CAPÍTULO IV

DAS DISPOSIÇÕES PENAIS

Art. 18. A divulgação de pesquisa fraudulenta constitui crime, punível com detenção de 6 meses a 1 ano e multa no valor de R$ 53.205,00 (cinqüenta e três mil duzentos e cinco reais) a R$ 106.410,00 (cento e seis mil quatrocentos e dez reais) (Lei nº 9.504/97, art. 33, § 4º).

Art. 19. O não cumprimento do disposto no art. 13 desta resolução ou qualquer ato que vise a retardar, impedir ou dificultar a ação fiscalizadora dos partidos políticos constitui crime, punível com detenção de 6 meses a 1 ano, com a alternativa de prestação de serviços à comunidade pelo mesmo prazo, e multa no valor de R$ 10.641,00 (dez mil seiscentos e quarenta e um reais) a R$ 21.282,00 (vinte e um mil duzentos e oitenta e dois reais) (Lei nº 9.504/97, art. 34, § 2º).

Parágrafo único. A comprovação de irregularidade nos dados publicados sujeita os responsáveis às penas mencionadas no *caput*, sem prejuízo da obrigatoriedade da veiculação dos dados corretos no mesmo espaço, local, horário, página, caracteres e outros elementos de destaque, de acordo com o veículo usado (Lei nº 9.504/97, art. 34, § 3º).

Art. 20. Pelos crimes definidos nos arts. 17 e 18 desta resolução, serão responsabilizados penalmente os representantes legais da empresa ou entidade de pesquisa e do órgão veiculador (Lei nº 9.504/97, art. 35).

CAPÍTULO V

DAS DISPOSIÇÕES FINAIS

Art. 21. Na divulgação dos resultados de enquetes ou sondagens, deverá ser informado não se tratar de pesquisa eleitoral, descrita

no art. 33 da Lei nº 9.504/97, mas de mero levantamento de opiniões, sem controle de amostra, o qual não utiliza método científico para sua realização, dependendo, apenas, da participação espontânea do interessado.

Parágrafo único. A divulgação de resultados de enquetes ou sondagens sem o esclarecimento previsto no *caput* será considerada divulgação de pesquisa eleitoral sem registro, autorizando a aplicação das sanções previstas nesta resolução.

Art. 22. Esta resolução entra em vigor na data de sua publicação.

Brasília, 16 de dezembro de 2009.

CARLOS AYRES BRITTO – PRESIDENTE

ARNALDO VERSIANI – RELATOR

RICARDO LEWANDOWSKI

CÁRMEN LÚCIA

FELIX FISCHER

FERNANDO GONÇALVES

MARCELO RIBEIRO

CERTIDÃO DE PUBLICAÇÃO

Certifico a publicação desta Resolução no Diário da Justiça eletrônico de 22/12/2009, pág. 3/5.

Eu, _____, lavrei a presente certidão.

Antonio Moreira Cunha
Coordenador Substituto
da COAD/SJD

Apêndice B

Resultados das Simulações

Tabela B.1: Média de Número de Contatos, Pessoas Contactadas e Domicílios Contactados por Entevista Completada

Desenho Amostral	Estrato / Cota	Unidade	Resp. Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
APVS	Não tem	Pessoas	1.893	1.215	1.028	2.001	1.143	1.000
		Contatos	1.893	2.120	2.527	2.001	2.000	1.997
APV	1	Pessoas	4.980	2.060	1.119	1.997	1.145	1.001
		Contatos	4.980	5.038	4.989	1.997	2.006	2.004
	2	Pessoas	2.510	1.274	1.006	1.999	1.142	1.001
		Contatos	2.510	2.494	2.509	1.999	1.999	1.998
	3	Pessoas	1.670	1.067	1.000	1.997	1.144	1.000
		Contatos	1.670	1.663	1.667	1.997	2.003	2.002
	4	Pessoas	1.249	1.007	1.000	1.996	1.141	1.000
		Contatos	1.249	1.247	1.251	1.996	1.991	1.998
APC	1	Domicílios	8.573	8.607	8.614	3.699	3.705	3.708
	2	Domicílios	4.462	4.464	4.461	3.609	3.618	3.612
	3	Domicílios	3.017	3.000	3.002	3.468	3.464	3.458
	4	Domicílios	2.348	2.341	2.346	3.461	3.434	3.445

Tabela B.2: EQM dos estimadores de p^h

Desenho Amostral	Tipo do Estimador	Estrato / Cota	Amostra $b = 8$						Amostra $b = 40$					
			Resp. Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
APVS	C	Não tem	0.285	0.282	0.308	0.071	0.073	0.072	0.217	0.216	0.252	0.013	0.012	0.012
APV	C	1	0.051	0.052	0.054	0.08	0.079	0.079	0.004	0.004	0.004	0.014	0.014	0.014
		2	0.08	0.081	0.079	0.079	0.079	0.079	0.014	0.013	0.013	0.014	0.014	0.013
		3	0.072	0.07	0.073	0.079	0.079	0.079	0.017	0.016	0.016	0.014	0.014	0.013
		4	0.041	0.041	0.041	0.079	0.079	0.079	0.012	0.011	0.011	0.013	0.013	0.013
APC	EM	1	0.05	0.054	0.053	0.091	0.09	0.09	0.003	0.004	0.004	0.017	0.017	0.017
		2	0.088	0.089	0.089	0.093	0.094	0.092	0.016	0.014	0.014	0.016	0.016	0.016
		3	0.085	0.085	0.084	0.09	0.09	0.089	0.022	0.02	0.02	0.016	0.016	0.016
		4	0.05	0.05	0.05	0.091	0.092	0.09	0.017	0.015	0.016	0.016	0.016	0.016

Tabela B.3: EQM dos Estimadores HH, Simples e Razão (dividido por 10^6)- Condicionado ao conhecimento de p^h

b	Desenho Amostral	Tipo do Estimador	Dist. Bernoulli											
			Y Heterogêneo						Y Homogêneo					
			Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	2.7E-2	2.4E-2	2.3E-2	1.9E-2	2.4E-2	1.9E-2	1.7E-2	2.2E-2	1.8E-2	1.9E-2	2.3E-2	1.9E-2
		Simples	2.4E-2	1.8E-2	2.0E-2	1.6E-2	1.6E-2	1.6E-2	1.5E-2	1.6E-2	1.6E-2	1.7E-2	1.7E-2	1.6E-2
		Razão	2.6E-2	2.1E-2	2.2E-2	1.7E-2	1.9E-2	1.7E-2	1.6E-2	1.8E-2	1.7E-2	1.8E-2	2.0E-2	1.8E-2
	APV	HH	2.2E-2	1.9E-2	2.2E-2	1.7E-2	1.8E-2	1.7E-2	2.0E-2	1.8E-2	2.0E-2	2.0E-2	2.0E-2	1.9E-2
		Simples	1.5E-2	1.4E-2	1.4E-2	1.4E-2	1.4E-2	1.4E-2	1.5E-2	1.6E-2	1.5E-2	1.7E-2	1.7E-2	1.6E-2
		Razão	1.7E-2	1.6E-2	1.6E-2	1.5E-2	1.6E-2	1.5E-2	1.7E-2	1.7E-2	1.7E-2	1.8E-2	1.8E-2	1.8E-2
	APC	HH	2.0E-2	2.0E-2	2.0E-2	2.1E-2	2.0E-2	2.1E-2	2.0E-2	2.0E-2	1.9E-2	2.3E-2	2.2E-2	2.1E-2
		Simples	1.4E-2	1.4E-2	1.4E-2	1.4E-2	1.3E-2	1.3E-2	1.6E-2	1.6E-2	1.6E-2	1.7E-2	1.7E-2	1.7E-2
		Razão	1.7E-2	1.7E-2	1.7E-2	1.6E-2	1.7E-2	1.6E-2	1.8E-2	1.8E-2	1.8E-2	2.0E-2	1.9E-2	1.9E-2
40	APVS	HH	1.1E-2	7.0E-3	5.0E-3	4.0E-3	3.9E-3	4.8E-3	3.4E-3	3.5E-3	4.4E-3	3.7E-3	3.8E-3	4.6E-3
		Simples	1.0E-2	6.2E-3	3.5E-3	3.4E-3	3.2E-3	3.3E-3	3.3E-3	3.2E-3	3.4E-3	3.3E-3	3.4E-3	3.4E-3
		Razão	1.1E-2	6.7E-3	4.4E-3	3.6E-3	3.5E-3	3.9E-3	3.4E-3	3.3E-3	3.8E-3	3.6E-3	3.7E-3	4.0E-3
	APV	HH	4.5E-3	4.4E-3	3.7E-3	3.5E-3	3.4E-3	3.6E-3	4.1E-3	4.1E-3	3.6E-3	3.9E-3	3.9E-3	3.9E-3
		Simples	3.0E-3	2.9E-3	2.9E-3	2.8E-3	2.8E-3	2.8E-3	3.1E-3	3.1E-3	3.1E-3	3.3E-3	3.3E-3	3.3E-3
		Razão	3.4E-3	3.4E-3	3.3E-3	3.2E-3	3.1E-3	3.2E-3	3.5E-3	3.6E-3	3.3E-3	3.6E-3	3.6E-3	3.7E-3
	APC	HH	4.0E-3	4.0E-3	4.1E-3	4.1E-3	4.1E-3	4.2E-3	4.0E-3	3.8E-3	3.9E-3	4.4E-3	4.4E-3	4.5E-3
		Simples	2.8E-3	2.8E-3	2.9E-3	2.7E-3	2.8E-3	2.7E-3	3.4E-3	3.3E-3	3.3E-3	3.3E-3	3.4E-3	3.5E-3
		Razão	3.5E-3	3.4E-3	3.5E-3	3.3E-3	3.4E-3	3.3E-3	3.7E-3	3.5E-3	3.6E-3	3.9E-3	4.0E-3	4.0E-3

Tabela B.4: EQM dos Estimadores HH, Simples e Razão (dividido por 10^6) - Condicionado ao conhecimento de p^h

b	Desenho Amostral	Tipo do Estimador	Dist. Normal											
			Y Heterogêneo						Y Homogêneo					
			Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	3.4E+7	3.5E+7	2.5E+7	2.5E+7	4.0E+7	2.4E+7	6.1E+1	1.6E+2	6.2E+1	5.9E+1	1.5E+2	6.0E+1
		Simples	1.6E+7	1.2E+7	9.7E+6	1.3E+7	1.5E+7	1.3E+7	6.1E+0	6.1E+0	6.1E+0	7.0E+0	6.9E+0	6.9E+0
		Razão	2.2E+7	1.2E+7	1.4E+7	1.2E+7	1.4E+7	1.2E+7	6.4E+0	7.1E+0	6.5E+0	7.2E+0	7.5E+0	7.0E+0
	APV	HH	8.8E+6	1.2E+7	9.0E+6	1.0E+7	1.4E+7	1.0E+7	1.2E+2	7.6E+1	1.3E+2	7.3E+1	7.8E+1	6.9E+1
		Simples	2.0E+6	2.2E+6	2.0E+6	2.2E+6	2.5E+6	2.2E+6	6.0E+0	6.2E+0	6.2E+0	6.9E+0	7.0E+0	6.9E+0
		Razão	5.0E+6	3.1E+6	5.0E+6	3.6E+6	3.5E+6	3.6E+6	6.8E+0	6.7E+0	7.0E+0	7.5E+0	7.3E+0	7.4E+0
	APC	HH	2.3E+7	2.3E+7	2.2E+7	2.1E+7	2.1E+7	2.1E+7	1.4E+2	1.4E+2	1.4E+2	1.5E+2	1.5E+2	1.4E+2
		Simples	2.1E+6	2.1E+6	2.0E+6	2.4E+6	2.4E+6	2.4E+6	6.1E+0	6.2E+0	6.2E+0	6.6E+0	6.5E+0	6.7E+0
		Razão	4.4E+6	4.4E+6	4.3E+6	4.9E+6	4.8E+6	4.9E+6	7.6E+0	7.6E+0	7.6E+0	7.9E+0	7.9E+0	8.1E+0
40	APVS	HH	1.7E+7	9.5E+6	7.2E+6	5.0E+6	4.9E+6	8.0E+6	1.3E+1	1.2E+1	3.1E+1	1.2E+1	1.2E+1	3.0E+1
		Simples	9.8E+6	3.7E+6	4.4E+6	3.8E+6	3.7E+6	6.5E+6	1.3E+0	1.2E+0	1.2E+0	1.4E+0	1.4E+0	1.4E+0
		Razão	1.6E+7	7.4E+6	2.4E+6	2.5E+6	2.5E+6	2.8E+6	1.3E+0	1.3E+0	1.4E+0	1.5E+0	1.5E+0	1.5E+0
	APV	HH	1.8E+6	1.8E+6	2.4E+6	2.1E+6	2.0E+6	3.0E+6	2.4E+1	2.5E+1	1.6E+1	1.4E+1	1.4E+1	1.5E+1
		Simples	4.6E+5	4.6E+5	5.2E+5	4.7E+5	4.9E+5	6.5E+5	1.2E+0	1.2E+0	1.2E+0	1.4E+0	1.4E+0	1.4E+0
		Razão	1.1E+6	1.1E+6	6.2E+5	7.5E+5	7.6E+5	7.5E+5	1.4E+0	1.4E+0	1.4E+0	1.5E+0	1.5E+0	1.5E+0
	APC	HH	4.6E+6	4.5E+6	4.5E+6	4.3E+6	4.2E+6	4.2E+6	2.9E+1	2.8E+1	2.8E+1	3.0E+1	3.0E+1	3.0E+1
		Simples	4.3E+5	4.5E+5	4.4E+5	5.4E+5	5.2E+5	5.4E+5	1.3E+0	1.3E+0	1.2E+0	1.3E+0	1.3E+0	1.3E+0
		Razão	9.1E+5	8.9E+5	8.9E+5	1.0E+6	1.0E+6	1.0E+6	1.6E+0	1.6E+0	1.6E+0	1.7E+0	1.6E+0	1.6E+0

Tabela B.5: Vício Relativo (%) dos Estimadores HH , Simples e Razão - Condicionado ao conhecimento de p^h

b	Desenho Amostral	Tipo do Estimador	Dist. Bernoulli												Dist. Normal											
			Y Heterogêneo						Y Homogêneo						Y Heterogêneo						Y Homogêneo					
			Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	23.6	14.7	3.2	1.4	0.6	1.8	0.9	1.7	2.4	0.7	0.9	0.6	17.5	10.7	1.9	0.2	0.0	0.0	1.1	0.1	1.0	0.4	0.3	0.2
		Simples	22.7	13.7	1.0	0.8	1.2	1.4	3.9	3.9	4.6	1.2	1.2	2.7	14.0	6.8	7.6	5.3	5.4	8.8	0.3	0.1	0.1	0.2	0.2	0.0
		Razão	22.9	14.1	3.1	0.9	0.6	1.1	2.0	2.1	2.1	0.5	0.5	0.2	18.0	10.5	0.3	0.6	0.7	1.1	0.1	0.1	0.2	0.3	0.3	0.3
	APV	HH	0.2	0.5	0.8	0.4	0.2	0.5	0.1	0.5	0.4	0.2	0.4	0.5	0.0	0.2	0.0	0.1	0.1	0.1	0.3	0.0	0.0	0.2	0.1	0.1
		Simples	1.4	1.5	0.1	0.1	0.0	0.3	0.3	0.7	0.9	0.2	0.2	1.5	1.4	1.4	1.6	0.9	0.9	1.8	0.0	0.0	0.1	0.0	0.1	0.1
		Razão	0.8	1.0	0.5	0.1	0.3	0.3	0.1	0.7	0.4	0.4	0.2	0.6	0.6	0.7	0.3	0.1	0.1	0.3	0.0	0.0	0.0	0.0	0.0	0.1
	APC	HH	0.7	0.8	0.7	0.3	0.3	0.3	1.0	0.2	0.7	0.7	0.1	1.2	0.1	0.3	0.2	0.0	0.1	0.2	0.0	0.0	0.1	0.3	0.6	0.4
		Simples	1.5	1.5	1.6	2.5	1.9	1.9	4.9	4.2	4.4	2.8	2.9	3.4	0.8	0.8	0.8	1.1	1.1	1.1	0.2	0.1	0.1	0.0	0.1	0.1
		Razão	0.7	0.8	0.7	0.5	0.1	0.2	1.5	0.9	1.4	0.8	0.5	1.4	0.7	0.6	0.7	0.1	0.2	0.1	0.1	0.0	0.1	0.1	0.0	0.0
40	APVS	HH	22.7	14.5	3.1	1.2	1.3	1.2	1.2	2.0	2.3	0.6	0.6	0.5	17.3	11.0	2.1	0.1	0.1	0.0	1.1	0.4	0.8	0.3	0.3	0.3
		Simples	21.7	13.5	0.9	0.9	0.7	1.7	4.2	4.1	4.5	1.4	1.3	2.9	13.8	7.0	7.5	5.2	5.2	8.8	0.3	0.2	0.1	0.2	0.2	0.0
		Razão	21.9	14.0	3.3	1.1	1.2	1.2	2.1	2.3	1.7	0.6	0.6	0.3	18.3	11.3	1.4	0.1	0.0	0.2	0.1	0.0	0.2	0.3	0.3	0.3
	APV	HH	0.2	0.1	0.2	0.2	0.2	0.1	0.3	0.2	0.0	0.2	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0
		Simples	1.2	1.2	0.7	0.2	0.1	0.4	0.1	0.1	0.8	0.0	0.2	1.2	1.4	1.4	1.7	0.8	0.9	1.8	0.0	0.0	0.1	0.1	0.0	0.0
		Razão	0.0	0.0	0.3	0.2	0.2	0.1	0.3	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
	APC	HH	0.3	0.0	0.3	0.2	0.3	0.3	0.9	0.8	0.7	0.5	0.3	0.9	0.3	0.2	0.3	0.0	0.2	0.1	0.1	0.1	0.0	0.6	0.4	0.3
		Simples	1.7	2.0	1.8	1.8	2.2	2.1	4.4	4.5	4.4	2.7	2.5	2.8	0.7	0.9	0.8	1.1	1.0	1.0	0.1	0.1	0.1	0.1	0.0	0.1
		Razão	0.2	0.0	0.5	0.1	0.4	0.3	0.8	0.8	0.8	0.4	0.2	0.7	0.1	0.2	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.1

Tabela B.6: Ranking do EQM dos Estimadores HH, Simples e Razão - Condicionado ao conhecimento de p^h

b	Desenho Amostral	Tipo do Estimador	Ranking Médio	Dist. Bernoulli												Dist. Normal											
				Y Heterogêneo						Y Homogêneo						Y Heterogêneo						Y Homogêneo					
				Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.		
				κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	8.08	9	9	9	8	9	8	5	9	7	6	9	6	9	9	9	9	9	9	7	9	7	7	9	7
		Simples	3.88	7	5	5	4	4	4	2	3	2	2	2	2	6	6	6	7	7	7	2	1	1	3	2	3
		Razão	5.63	8	8	7	6	7	6	4	6	4	5	7	5	7	5	7	6	5	6	4	5	4	4	5	4
	APV	HH	6.79	6	6	8	7	6	7	8	7	9	7	6	7	5	7	5	5	6	5	8	7	8	8	7	8
		Simples	1.71	2	2	2	2	2	2	1	1	1	1	3	1	1	2	1	1	2	1	1	2	3	2	3	2
		Razão	3.88	3	3	3	3	3	3	6	4	5	4	4	4	4	3	4	3	3	3	5	4	5	5	4	5
	APC	HH	8.13	5	7	6	9	8	9	9	8	8	9	8	9	8	8	8	8	8	8	9	8	9	9	8	9
		Simples	1.75	1	1	1	1	1	1	3	2	3	3	1	3	2	1	2	2	1	2	3	3	2	1	1	1
		Razão	5.17	4	4	4	5	5	5	7	5	6	8	5	8	3	4	3	4	4	4	6	6	6	6	6	6
40	APVS	HH	7.96	9	9	9	8	8	9	5	5	9	6	6	9	9	9	9	9	9	9	7	7	9	7	7	8
		Simples	4.25	7	7	5	5	4	4	2	2	4	1	3	2	7	6	7	7	7	8	2	2	1	3	3	3
		Razão	5.88	8	8	8	7	7	7	4	4	7	4	5	7	8	8	6	6	6	5	4	4	5	4	4	5
	APV	HH	6.46	6	6	6	6	5	6	9	9	5	8	7	5	5	5	5	5	5	6	8	8	7	8	8	7
		Simples	1.63	2	2	1	2	2	2	1	1	1	3	1	1	2	2	2	1	1	2	1	1	2	2	2	2
		Razão	3.92	3	3	3	3	3	3	6	7	2	5	4	4	4	4	3	3	3	3	5	5	4	5	5	4
	APC	HH	7.92	5	5	7	9	9	8	8	8	8	9	9	8	6	7	8	8	8	7	9	9	8	9	9	9
		Simples	1.79	1	1	2	1	1	1	3	3	3	2	2	3	1	1	1	2	2	1	3	3	3	1	1	1
		Razão	5.21	4	4	4	4	6	5	7	6	6	7	8	6	3	3	4	4	4	4	6	6	6	6	6	6

Tabela B.7: EQM dos Estimadores HH, Simples e Razão (dividido por 10^6) - Estimando p^h

b	Desenho Amostral	Tipo do Estimador	Dist. Bernoulli											
			Y Heterogêneo						Y Homogêneo					
			Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	2.7E-2	2.4E-2	2.3E-2	1.9E-2	2.4E-2	1.9E-2	1.7E-2	2.2E-2	1.8E-2	1.9E-2	2.3E-2	1.9E-2
		Simples	2.4E-2	1.8E-2	2.0E-2	1.6E-2	1.6E-2	1.6E-2	1.5E-2	1.6E-2	1.6E-2	1.7E-2	1.7E-2	1.6E-2
		Razão	2.6E-2	2.1E-2	2.2E-2	1.7E-2	1.9E-2	1.7E-2	1.6E-2	1.8E-2	1.7E-2	1.8E-2	2.0E-2	1.8E-2
	APV	HH	2.2E-2	1.9E-2	2.2E-2	1.7E-2	1.8E-2	1.7E-2	2.0E-2	1.8E-2	2.0E-2	2.0E-2	2.0E-2	1.9E-2
		Simples	1.5E-2	1.4E-2	1.4E-2	1.4E-2	1.4E-2	1.4E-2	1.5E-2	1.6E-2	1.5E-2	1.7E-2	1.7E-2	1.6E-2
		Razão	1.7E-2	1.6E-2	1.6E-2	1.5E-2	1.6E-2	1.5E-2	1.7E-2	1.7E-2	1.7E-2	1.8E-2	1.8E-2	1.8E-2
	APC	HH	2.0E-2	2.0E-2	2.0E-2	2.1E-2	2.0E-2	2.1E-2	2.0E-2	2.0E-2	1.9E-2	2.3E-2	2.2E-2	2.1E-2
		Simples	1.4E-2	1.4E-2	1.4E-2	1.4E-2	1.3E-2	1.3E-2	1.6E-2	1.6E-2	1.6E-2	1.7E-2	1.7E-2	1.7E-2
		Razão	1.7E-2	1.7E-2	1.7E-2	1.6E-2	1.7E-2	1.6E-2	1.8E-2	1.8E-2	1.8E-2	2.0E-2	1.9E-2	1.9E-2
40	APVS	HH	1.1E-2	7.0E-3	5.0E-3	4.0E-3	3.9E-3	4.8E-3	3.4E-3	3.5E-3	4.4E-3	3.7E-3	3.8E-3	4.6E-3
		Simples	1.0E-2	6.2E-3	3.5E-3	3.4E-3	3.2E-3	3.3E-3	3.3E-3	3.2E-3	3.4E-3	3.3E-3	3.4E-3	3.4E-3
		Razão	1.1E-2	6.7E-3	4.4E-3	3.6E-3	3.5E-3	3.9E-3	3.4E-3	3.3E-3	3.8E-3	3.6E-3	3.7E-3	4.0E-3
	APV	HH	4.5E-3	4.4E-3	3.7E-3	3.5E-3	3.4E-3	3.6E-3	4.1E-3	4.1E-3	3.6E-3	3.9E-3	3.9E-3	3.9E-3
		Simples	3.0E-3	2.9E-3	2.9E-3	2.8E-3	2.8E-3	2.8E-3	3.1E-3	3.1E-3	3.1E-3	3.3E-3	3.3E-3	3.3E-3
		Razão	3.4E-3	3.4E-3	3.3E-3	3.2E-3	3.1E-3	3.2E-3	3.5E-3	3.6E-3	3.3E-3	3.6E-3	3.6E-3	3.7E-3
	APC	HH	4.0E-3	4.0E-3	4.1E-3	4.1E-3	4.1E-3	4.2E-3	4.0E-3	3.8E-3	3.9E-3	4.4E-3	4.4E-3	4.5E-3
		Simples	2.8E-3	2.8E-3	2.9E-3	2.7E-3	2.8E-3	2.7E-3	3.4E-3	3.3E-3	3.3E-3	3.3E-3	3.4E-3	3.5E-3
		Razão	3.5E-3	3.4E-3	3.5E-3	3.3E-3	3.4E-3	3.3E-3	3.7E-3	3.5E-3	3.6E-3	3.9E-3	4.0E-3	4.0E-3

Tabela B.8: EQM dos Estimadores HH, Simples e Razão (dividido por 10^6) - Estimando p^h

b	Desenho Amostral	Tipo do Estimador	Dist. Normal											
			Y Heterogêneo						Y Homogêneo					
			Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	3.4E+7	3.5E+7	2.5E+7	2.5E+7	4.0E+7	2.4E+7	6.1E+1	1.6E+2	6.2E+1	5.9E+1	1.5E+2	6.0E+1
		Simples	1.6E+7	1.2E+7	9.7E+6	1.3E+7	1.5E+7	1.3E+7	6.1E+0	6.1E+0	6.1E+0	7.0E+0	6.9E+0	6.9E+0
		Razão	2.2E+7	1.2E+7	1.4E+7	1.2E+7	1.4E+7	1.2E+7	6.4E+0	7.1E+0	6.5E+0	7.2E+0	7.5E+0	7.0E+0
	APV	HH	8.8E+6	1.2E+7	9.0E+6	1.0E+7	1.4E+7	1.0E+7	1.2E+2	7.6E+1	1.3E+2	7.3E+1	7.8E+1	6.9E+1
		Simples	2.0E+6	2.2E+6	2.0E+6	2.2E+6	2.5E+6	2.2E+6	6.0E+0	6.2E+0	6.2E+0	6.9E+0	7.0E+0	6.9E+0
		Razão	5.0E+6	3.1E+6	5.0E+6	3.6E+6	3.5E+6	3.6E+6	6.8E+0	6.7E+0	7.0E+0	7.5E+0	7.3E+0	7.4E+0
	APC	HH	2.3E+7	2.3E+7	2.2E+7	2.1E+7	2.1E+7	2.1E+7	1.4E+2	1.4E+2	1.4E+2	1.5E+2	1.5E+2	1.4E+2
		Simples	2.1E+6	2.1E+6	2.0E+6	2.4E+6	2.4E+6	2.4E+6	6.1E+0	6.2E+0	6.2E+0	6.6E+0	6.5E+0	6.7E+0
		Razão	4.4E+6	4.4E+6	4.3E+6	4.9E+6	4.8E+6	4.9E+6	7.6E+0	7.6E+0	7.6E+0	7.9E+0	7.9E+0	8.1E+0
40	APVS	HH	1.7E+7	9.5E+6	7.2E+6	5.0E+6	4.9E+6	8.0E+6	1.3E+1	1.2E+1	3.1E+1	1.2E+1	1.2E+1	3.0E+1
		Simples	9.8E+6	3.7E+6	4.4E+6	3.8E+6	3.7E+6	6.5E+6	1.3E+0	1.2E+0	1.2E+0	1.4E+0	1.4E+0	1.4E+0
		Razão	1.6E+7	7.4E+6	2.4E+6	2.5E+6	2.5E+6	2.8E+6	1.3E+0	1.3E+0	1.4E+0	1.5E+0	1.5E+0	1.5E+0
	APV	HH	1.8E+6	1.8E+6	2.4E+6	2.1E+6	2.0E+6	3.0E+6	2.4E+1	2.5E+1	1.6E+1	1.4E+1	1.4E+1	1.5E+1
		Simples	4.6E+5	4.6E+5	5.2E+5	4.7E+5	4.9E+5	6.5E+5	1.2E+0	1.2E+0	1.2E+0	1.4E+0	1.4E+0	1.4E+0
		Razão	1.1E+6	1.1E+6	6.2E+5	7.5E+5	7.6E+5	7.5E+5	1.4E+0	1.4E+0	1.4E+0	1.5E+0	1.5E+0	1.5E+0
	APC	HH	4.6E+6	4.5E+6	4.5E+6	4.3E+6	4.2E+6	4.2E+6	2.9E+1	2.8E+1	2.8E+1	3.0E+1	3.0E+1	3.0E+1
		Simples	4.3E+5	4.5E+5	4.4E+5	5.4E+5	5.2E+5	5.4E+5	1.3E+0	1.3E+0	1.2E+0	1.3E+0	1.3E+0	1.3E+0
		Razão	9.1E+5	8.9E+5	8.9E+5	1.0E+6	1.0E+6	1.0E+6	1.6E+0	1.6E+0	1.6E+0	1.7E+0	1.6E+0	1.6E+0

Tabela B.9: Vício Relativo (%) dos Estimadores HH, Simples e Razão - Estimando p^h

b	Desenho Amostral	Tipo do Estimador	Dist. Bernoulli												Dist. Normal											
			Y Heterogêneo						Y Homogêneo						Y Heterogêneo						Y Homogêneo					
			Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.		
			κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	21.4	14.2	3.8	2.0	1.4	0.8	2.4	2.3	1.9	0.3	0.2	0.1	18.5	11.4	1.2	1.0	0.5	0.3	0.3	0.0	0.3	1.0	0.7	0.0
		Simples	21.3	13.7	1.5	1.4	1.2	1.9	4.4	3.9	4.6	1.5	0.9	3.4	13.6	6.8	7.7	5.3	5.5	9.0	0.3	0.1	0.1	0.3	0.1	0.0
		Razão	21.5	14.0	3.6	0.5	0.6	0.2	2.3	2.3	1.7	0.3	0.8	0.4	18.0	10.7	0.1	0.5	0.8	1.3	0.1	0.1	0.2	0.4	0.2	0.2
	APV	HH	2.6	2.5	0.2	2.8	2.8	0.5	1.1	1.9	0.1	2.2	1.9	0.2	2.1	1.6	0.1	2.4	2.4	0.2	2.1	2.3	0.0	2.3	2.0	0.3
		Simples	0.9	0.9	0.1	0.7	0.2	0.1	1.2	0.2	0.8	0.3	0.5	1.3	1.3	1.1	1.6	0.8	0.8	1.8	0.0	0.0	0.1	0.0	0.0	0.0
		Razão	0.1	0.3	0.2	0.5	0.2	0.4	0.9	0.2	0.4	0.1	0.5	0.0	0.9	0.6	0.5	0.5	0.3	0.4	0.0	0.0	0.1	0.0	0.1	0.0
	APC	HH	1.8	1.6	1.4	0.2	2.3	1.2	2.0	1.9	2.3	0.7	0.3	1.1	1.1	0.8	1.2	2.6	1.7	1.8	0.6	0.3	0.5	1.2	1.1	1.3
		Simples	1.1	1.5	1.7	2.6	1.5	2.1	4.8	4.7	4.8	2.3	2.1	2.7	0.8	0.8	0.7	1.0	1.1	1.1	0.2	0.2	0.2	0.1	0.1	0.1
		Razão	1.3	0.8	0.6	0.9	0.7	0.2	1.7	1.7	2.0	0.2	0.4	0.3	0.6	0.7	0.4	0.4	0.2	0.0	0.2	0.1	0.1	0.1	0.0	0.0
40	APVS	HH	22.1	14.9	3.1	1.4	0.7	1.3	1.5	1.1	2.3	0.1	0.2	0.7	18.2	10.6	1.9	0.2	0.3	0.1	0.6	0.6	0.7	0.4	0.4	0.2
		Simples	21.5	13.8	0.9	0.9	1.6	1.6	4.3	3.4	4.5	1.7	1.5	2.7	13.9	6.9	7.5	5.3	5.3	8.8	0.3	0.2	0.1	0.2	0.2	0.1
		Razão	21.7	14.1	3.2	1.1	0.5	1.2	2.0	1.8	1.9	0.3	0.3	0.5	18.7	11.1	1.4	0.1	0.1	0.2	0.1	0.0	0.2	0.3	0.3	0.3
	APV	HH	0.4	0.8	0.3	0.8	0.4	0.1	0.7	0.3	0.5	0.5	0.4	0.2	0.5	0.5	0.1	0.5	0.5	0.0	0.4	0.4	0.0	0.5	0.5	0.1
		Simples	1.2	1.1	0.8	0.5	0.2	0.6	0.1	0.4	0.1	0.0	0.2	1.0	1.4	1.4	1.6	0.9	0.8	1.7	0.0	0.0	0.1	0.1	0.0	0.0
		Razão	0.1	0.1	0.4	0.3	0.0	0.1	0.2	0.3	0.5	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0
	APC	HH	0.0	0.1	0.1	0.3	0.7	0.9	0.8	0.8	0.3	0.0	0.4	0.1	0.1	0.2	0.1	0.6	0.6	0.6	0.5	0.5	0.6	0.2	0.3	0.2
		Simples	1.7	1.7	1.7	1.7	1.8	2.2	4.7	4.9	4.3	2.9	2.3	2.7	0.8	0.8	0.8	1.1	1.1	1.1	0.1	0.2	0.2	0.0	0.1	0.1
		Razão	0.3	0.5	0.3	0.2	0.1	0.3	1.1	1.5	0.6	0.6	0.0	0.5	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.0	0.0

Tabela B.10: Ranking do EQM dos Estimadores HH, Simples e Razão - Estimando p^h

b	Desenho Amostral	Tipo do Estimador	Ranking Médio	Aumento Médio do EQM (%)	Dist. Bernoulli												Dist. Normal											
					Y Heterogêneo						Y Homogêneo						Y Heterogêneo						Y Homogêneo					
					Resp. Hetero.			Resp. Homo.			Resp.Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.			Resp. Hetero.			Resp. Homo.		
					κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10	κ_2 1	κ_2 3	κ_2 10
8	APVS	HH	7.71	14.2	8	7	9	8	8	9	6	6	9	6	6	8	9	8	9	8	8	9	7	7	8	7	7	8
		Simples	3.71	0.5	5	5	4	4	4	4	2	2	3	1	2	3	6	5	6	6	6	7	3	3	1	2	2	3
		Razão	5.25	0.9	7	6	7	5	5	7	4	4	7	4	4	6	7	7	5	5	5	5	4	4	5	4	4	5
	APV	HH	7.21	22.2	9	9	6	7	7	6	9	9	6	8	8	5	5	6	7	7	7	6	8	8	7	8	8	7
		Simples	1.88	0.3	2	2	2	2	2	2	3	1	2	3	1	2	1	1	2	1	1	2	1	2	2	3	3	2
		Razão	4.04	4.1	4	3	3	3	3	3	7	5	4	5	5	4	4	4	3	3	3	3	5	5	4	5	5	4
	APC	HH	8.50	27.6	6	8	8	9	9	8	8	8	8	9	9	9	8	9	8	9	9	8	9	9	9	9	9	9
		Simples	1.50	-0.6	1	1	1	1	1	1	1	3	1	2	3	1	2	2	1	2	2	1	2	1	3	1	1	1
		Razão	5.21	6.6	3	4	5	6	6	5	5	7	5	7	7	7	3	3	4	4	4	4	6	6	6	6	6	6
40	APVS	HH	8.00	1.3	9	9	9	8	9	9	5	5	9	5	6	9	9	9	9	9	9	9	7	7	9	7	7	9
		Simples	4.08	0.0	7	7	4	4	5	5	2	2	2	1	2	1	7	6	7	7	7	8	2	2	1	3	3	3
		Razão	5.83	0.1	8	8	8	7	7	8	3	4	7	4	5	6	8	8	5	6	6	6	4	4	5	4	4	5
	APV	HH	6.67	4.7	6	6	6	6	6	6	9	9	5	9	8	7	5	5	6	5	5	5	8	8	7	8	8	7
		Simples	1.63	0.0	1	1	2	2	2	2	1	1	1	3	1	2	2	2	2	1	1	2	1	1	2	2	2	2
		Razão	4.08	2.0	3	3	3	3	3	3	6	7	4	7	4	4	4	4	3	3	3	3	5	5	4	5	5	4
	APC	HH	7.75	-2.6	5	5	7	9	8	7	8	8	8	8	9	8	6	7	8	8	8	7	9	9	8	9	9	8
		Simples	1.92	0.1	2	2	1	1	1	1	4	3	3	2	3	3	1	1	1	2	2	1	3	3	3	1	1	1
		Razão	5.04	-0.8	4	4	5	5	4	4	7	6	6	6	7	5	3	3	4	4	4	4	6	6	6	6	6	6

Apêndice C

Avaliação das Pesquisas Eleitorais

Tabela C.1: Quantidade de Pesquisas Eleitorais e Número de Categorias

Características da Pesquisa		Categorias		Pesquisas	
		número	%	número	%
Cargo	Governador	665	17.2	143	15.9
	Prefeito	2551	65.9	661	73.6
	Presidente	654	16.9	94	10.5
Final de Semana	Sem Coleta	3542	91.5	819	91.2
	Com Coleta	328	8.5	79	8.8
Dias de Campo	1	2506	64.8	567	63.1
	2	396	10.2	75	8.4
	3	832	21.5	229	25.5
	4 ou mais	136	3.5	27	3.0
Turno	Primeiro Turno	3616	93.4	644	71.7
	Segundo Turno	254	6.6	254	28.3
Votos Brancos, Nulos e Indecisos	Até 1%	3591	94.1	852	95.6
	De 1% a 5%	119	3.1	20	2.2
	De 5% a 10%	36	0.9	5	0.6
	10% ou mais	72	1.9	14	1.6
Dias antes da eleição	Mesmo Dia	147	3.8	37	4.1
	De 1 a 5	703	18.2	153	17.0
	De 6 a 10	768	19.8	181	20.2
	De 11 a 15	451	11.7	130	14.5
	De 16 a 20	841	21.7	180	20.0
	20 ou mais	960	24.8	217	24.2
Candi-datos	2	267	6.9	267	29.7
	3 a 5	836	21.6	260	29.0
	6 a 10	1777	45.9	282	31.4
	11 ou mais	990	25.6	89	9.9
Tamanho Amostral	Até 500	765	19.8	227	25.3
	De 501 a 1000	1404	36.3	371	41.3
	De 1001 a 2000	1303	33.7	230	25.6
	2001 ou mais	398	10.3	70	7.8
Classe de Variância	Maior	395	10.2	224	25.0
	Grande	808	20.9	224	25.0
	Pequena	1160	30.0	225	25.1
	Menor	1504	38.9	224	25.0
Complex. Amostral	Simples	3005	77.6	762	84.9
	Complexa	865	22.4	136	15.1
Total		3870	100.0	898	100.0

Tabela C.2: Médias dos Indicadores de Erros das Pesquisas Eleitorais

Características da Pesquisa		Categorias Individualmente		Pesquisas Completas (Categorias Simultaneamente)										
		AAS		AAS		Descritivos				Rankings			Distâncias	
		Erro Obs. Abs. em %	$I^{0,05}_{BIN}$ (%)	Número de Erros	$I^{0,05}_{MULT}$ (%)	I^{C3}_{SSRC} (%)	$I^{C3(15\%)}_{SSRC}$ (%)	I^{C5}_{SSRC} (%)	$I^{\|C5\|}_{SSRC}$ (%)	$I^{Vencedor}_{Ranking}$ (%)	$I^{Todos}_{Ranking}$ (%)	$I^{\%Acerto}_{Ranking}$ (%)	I^{Maha}_{Dist}	I^{Aitch}_{Dist}
Cargo	Governador	3.2	62.4	1.75	20.3	3.6	5.5	2.3	5.3	90.9	51.7	73.6	4,98	0.50
	Prefeito	3.0	73.3	1.03	38.3	3.7	4.7	1.0	4.8	86.8	68.5	84.2	3,12	0.24
	Presidente	1.2	81.0	1.32	24.5	1.8	3.1	1.9	3.2	90.4	39.4	71.0	3,16	0.44
Final de Semana	Sem Coleta	2.8	72.6	1.19	34.1	3.5	4.7	1.3	4.7	88.5	62.9	81.2	3,45	0.30
	Com Coleta	2.9	74.4	1.06	32.9	3.3	4.7	1.1	4.7	81.0	62.0	80.2	3,08	0.28
Dias de Campo	1	2.7	72.4	1.22	33.7	3.4	4.6	1.2	4.6	88.7	61.2	80.5	3,58	0.31
	2	2.0	75.3	1.31	26.7	2.8	4.0	1.0	4.1	85.3	53.3	77.8	3,02	0.34
	3	3.5	71.5	1.03	37.1	3.9	5.1	1.6	5.1	85.6	69.4	83.8	3,13	0.26
	4 ou mais	2.5	77.9	1.11	33.3	2.9	4.5	2.2	4.3	96.3	66.7	81.5	3,67	0.38
Turno	Primeiro Turno	2.7	73.8	1.47	25.0	3.4	5.1	1.5	5.1	83.1	48.1	73.7	3,82	0.41
	Segundo Turno	3.6	56.7	0.43	56.7	3.6	3.6	0.8	3.6	100	100	100	2,40	0.03
Votos Brancos, Nulos e Indecisos	Até 1%	2.7	73.4	1.12	35.4	3.5	4.6	1.3	4.6	88.4	64.7	82.1	3,40	0.29
	De 1% a 5%	3.3	67.2	1.95	15.0	4.2	6.2	2.8	6.1	85.0	30.0	67.0	3,56	0.56
	De 5% a 10%	2.0	69.4	2.20	0.0	3.1	5.1	3.0	5.3	80.0	40.0	69.7	5,25	0.60
	10% ou mais	3.8	55.6	2.29	0.0	4.6	5.7	-2.4	6.4	57.1	35.7	61.9	3,51	0.48
Dias antes da eleição	Mesmo Dia	1.0	87.8	0.49	64.9	1.3	1.7	0.9	1.8	97.3	78.4	92.3	1,95	0.13
	De 1 a 5	1.7	77.2	1.05	32.7	2.2	3.0	1.1	3.1	90.2	64.7	83.6	3,08	0.25
	De 6 a 10	2.3	75.1	1.06	39.8	2.9	3.9	1.1	3.9	89.0	62.4	82.3	2,89	0.27
	De 11 a 15	3.0	72.5	0.95	42.3	3.5	4.4	0.7	4.5	89.2	73.1	86.0	2,87	0.24
	De 16 a 20	3.0	70.0	1.40	26.7	4.0	5.4	2.3	5.4	87.8	57.8	77.3	3,65	0.37
	20 ou mais	3.8	67.6	1.43	25.8	4.8	6.5	1.3	6.5	82.9	57.1	76.7	4,49	0.37
Candi-datos	2	3.7	56.9	0.43	56.9	3.7	3.7	1.0	3.7	100	100	100	2,42	0.04
	3 a 5	4.6	61.7	1.23	33.1	4.8	5.7	1.7	5.8	81.5	74.2	83.5	3,99	0.22
	6 a 10	2.6	74.5	1.61	19.9	2.7	4.9	1.3	4.9	82.3	34.8	70.1	3,54	0.46
	11 ou mais	1.3	83.1	1.88	12.4	1.3	3.5	1.2	3.3	87.6	6.7	52.3	4,36	0.89
Tamanho Amostral	Até 500	4.5	68.0	1.08	37.0	5.0	6.0	1.2	6.1	77.1	67.8	80.6	2,89	0.22
	De 501 a 1000	3.1	71.9	1.06	38.3	3.4	4.6	1.4	4.6	90.3	68.2	84.9	3,60	0.26
	De 1001 a 2000	1.9	75.2	1.40	24.8	2.5	4.1	1.5	4.0	90.9	49.6	75.7	3,74	0.45
	2001 ou mais	1.2	76.6	1.33	31.4	1.9	2.4	0.9	2.3	100	61.4	80.6	3,12	0.31
Classe de Variância	Maior	3.5	63.0	0.65	50.9	4.1	4.2	0.9	4.2	93.3	90.2	93.2	2,48	0.07
	Grande	2.7	74.4	0.92	42.0	3.6	4.1	1.5	4.3	93.3	68.3	83.5	4,03	0.31
	Pequena	2.7	74.9	1.29	29.8	3.3	4.5	1.9	4.5	90.2	60.0	80.0	3,40	0.33
	Menor	2.6	72.7	1.83	13.4	2.9	5.8	1.0	5.7	75.0	33.0	67.9	3,77	0.50
Complex. Amostral	Simples	3.1	71.3	1.13	35.0	3.8	4.9	1.3	4.9	86.9	66.4	82.7	3,42	0.28
	Complexa	1.5	77.7	1.42	27.9	1.9	3.1	1.5	3.1	93.4	42.6	72.2	3,44	0.45
Total		2.8	72.7	1.18	34.0	3.5	4.7	1.3	4.7	87.9	62.8	81.1	3,42	0.30

Apêndice D

Relação das Pesquisas Eleitorais

Tabela D.1: Listagem das pesquisas eleitorais analisadas

INSTITUTO	NÚMERO DA PESQUISA	NOME DA PESQUISA
IBOPE	IBO/BR89.OUT-00198	VOTO NACIONAL XIII
IBOPE	IBO/BR89.OUT-00195	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/BR89.NOV-00199	VOTO NACIONAL XIV
IBOPE	IBO/BR89.NOV-00200	VOTO NACIONAL XV
IBOPE	IBO/BR89.NOV-00201	VOTO NACIONAL XVI
IBOPE	IBO/BR89.NOV-00203	VOTO NACIONAL XVIII
IBOPE	IBO/BR89.NOV-00205	VOTO NACIONAL XX
IBOPE	IBO/BR89.NOV-00204	
IBOPE	IBO/BR89.DEZ-00206	VOTO NACIONAL XXI
IBOPE	IBO/BR89.DEZ-00207	VOTO NACIONAL XXII
IBOPE	IBO/BR89.DEZ-00208	VOTO NACIONAL XXIII
IBOPE	IBO/BR89.DEZ-00209	VOTO NACIONAL XXIV
IBOPE	IBO/MG94.OUT-00380	PESQ. DE O.P.
IBOPE	IBO/ES94.OUT-00381	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/SE94.OUT-00382	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/RG94.OUT-00384	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/RO94.OUT-00385	PESQ. DE O.P.
IBOPE	IBO/DF94.OUT-00383	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/SP94.OUT-00386	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/BR94.SET-00371	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/BR94.SET-00375	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/BR94.SET-00378	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/MT98.SET-01225	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/MS98.SET-01226	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/GO98.SET-01224	PROJETO GLOBO 98 - 2ª RODADA
IBOPE	IBO/RJ98.SET-01230	PROJETO GLOBO 98 - 5ª RODADA
IBOPE	IBO/PE98.SET-01231	PROJETO GLOBO 98 - 3ª RODADA
IBOPE	IBO/SC98.SET-01235	PROJETO GLOBO 98 - 5ª RODADA
IBOPE	IBO/RS98.SET-01234	PROJETO GLOBO 98 - 5ª RODADA
IBOPE	IBO/DF98.SET-01233	PROJETO GLOBO 98 - 3ª RODADA
IBOPE	IBO/AL98.SET-01232	PROJETO GLOBO 98 - 1ª RODADA
IBOPE	IBO/PI98.SET-01238	PROJETO GLOBO 98 - 3ª RODADA
IBOPE	IBO/BA98.SET-01239	PROJETO GLOBO 98 - 3ª RODADA
IBOPE	IBO/TO98.SET-01241	PROJETO GLOBO 98 - 3ª RODADA
IBOPE	IBO/RO98.SET-01245	PROJETO GLOBO 98 - 2ª RODADA
IBOPE	IBO/RJ98.SET-01240	PROJETO GLOBO 98 - 6ª RODADA
IBOPE	IBO/MT98.SET-01244	PROJETO GLOBO 98 - 9ª RODADA
IBOPE	IBO/MS98.SET-01243	PROJETO GLOBO 98 - 5ª RODADA
IBOPE	IBO/MG98.SET-01247	PESQ. COM ELEITORES
IBOPE	IBO/GO98.SET-01242	PROJETO GLOBO 98 - 3ª RODADA
IBOPE	IBO/SC98.SET-01249	PROJETO GLOBO 98 - 6ª RODADA
IBOPE	IBO/RS98.SET-01248	PROJETO GLOBO 98 - 6ª RODADA
IBOPE	IBO/RS98.SET-01250	PROJETO GLOBO 98 - 7ª RODADA
IBOPE	IBO/SC98.SET-01252	PROJETO GLOBO 98 - 7ª RODADA
IBOPE	IBO/RN98.SET-01258	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/RJ98.SET-01256	PROJETO GLOBO 98 - 7ª RODADA
IBOPE	IBO/ES98.SET-01254	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/DF98.SET-01255	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/CE98.SET-01253	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/SP98.SET-01257	PROJETO GLOBO 98 - 7ª RODADA

IBOPE	IBO/MA98.SET-01251	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/MG98.SET-01247	PESQ. COM ELEITORES
IBOPE	IBO/SC98.SET-01249	PROJETO GLOBO 98 - 6ª RODADA
IBOPE	IBO/RS98.SET-01248	PROJETO GLOBO 98 - 6ª RODADA
IBOPE	IBO/RS98.SET-01250	PROJETO GLOBO 98 - 7ª RODADA
IBOPE	IBO/SC98.SET-01252	PROJETO GLOBO 98 - 7ª RODADA
IBOPE	IBO/RN98.SET-01258	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/RJ98.SET-01256	PROJETO GLOBO 98 - 7ª RODADA
IBOPE	IBO/ES98.SET-01254	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/DF98.SET-01255	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/CE98.SET-01253	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/BR98.SET-01259	PROJETO GLOBO 98 - 12ª RODADA
IBOPE	IBO/SP98.SET-01257	PROJETO GLOBO 98 - 7ª RODADA
IBOPE	IBO/MA98.SET-01251	PROJETO GLOBO 98 - 4ª RODADA
IBOPE	IBO/BELEM00.SET-01371	PESQ. COM ELEITORES
IBOPE	IBO/TERESINA00.SET-01373	PESQ. COM ELEITORES
IBOPE	IBO/SALVADOR00.SET-01372	PESQ. COM ELEITORES
IBOPE	IBO/CUBATAO00.SET-01384	PESQ. COM ELEITORES
IBOPE	IBO/ARACAJU00.SET-01376	PESQ. COM ELEITORES
IBOPE	IBO/SAOVICENTE00.SET-01383	PESQ. COM ELEITORES
IBOPE	IBO/SAOCARLOS00.SET-01388	PESQ. COM ELEITORES
IBOPE	IBO/SAOBERNARDODOCAMPO00.SET-01389	PESQ. COM ELEITORES
IBOPE	IBO/PRAIAGRANDE00.SET-01382	PESQ. COM ELEITORES
IBOPE	IBO/MANAUS00.SET-01378	PESQ. COM ELEITORES
IBOPE	IBO/MACEIO00.SET-01377	PESQ. COM ELEITORES
IBOPE	IBO/LONDRINA00.SET-01386	PESQ. COM ELEITORES
IBOPE	IBO/CURITIBA00.SET-01379	PESQ. COM ELEITORES
IBOPE	IBO/CASCAVEL00.SET-01387	PESQ. COM ELEITORES
IBOPE	IBO/ARARAQUARA00.SET-01385	PESQ. COM ELEITORES
IBOPE	IBO/VITORIA00.SET-01410	PESQ. COM ELEITORES
IBOPE	IBO/UBERLANDIA00.SET-01409	PESQ. COM ELEITORES
IBOPE	IBO/SP00.SET-01390	PESQ. COM ELEITORES
IBOPE	IBO/SAOLUIS00.SET-01393	PESQ. COM ELEITORES
IBOPE	IBO/SAOJOSEDOSCAMPOS00.SET-01408	PESQ. COM ELEITORES
IBOPE	IBO/SANTOS00.SET-01400	PESQ. COM ELEITORES
IBOPE	IBO/RJ00.SET-01391	PESQ. COM ELEITORES
IBOPE	IBO/PORTOVELHO00.SET-01397	PESQ. COM ELEITORES
IBOPE	IBO/PORTOALEGRE00.SET-01401	PESQ. COM ELEITORES
IBOPE	IBO/PIRACICABA00.SET-01407	PESQ. COM ELEITORES
IBOPE	IBO/PALMAS00.SET-01381	PESQ. COM ELEITORES
IBOPE	IBO/OSASCO00.SET-01413	PESQ. COM ELEITORES
IBOPE	IBO/GUARULHOS00.SET-01414	PESQ. COM ELEITORES
IBOPE	IBO/GUARUJA00.SET-01399	PESQ. COM ELEITORES
IBOPE	IBO/GOIANIA00.SET-01396	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/FRANCA00.SET-01395	PESQ. COM ELEITORES
IBOPE	IBO/CHAPECO00.SET-01403	PESQ. COM ELEITORES
IBOPE	IBO/BOAVISTA00.SET-01398	PESQ. COM ELEITORES
IBOPE	IBO/BH00.SET-01405	PESQ. COM ELEITORES
IBOPE	IBO/BELEM00.SET-01394	PESQ. COM ELEITORES
IBOPE	IBO/VITORIADACONQUISTA00.SET-01417	PESQ. COM ELEITORES
IBOPE	IBO/PETROLINA00.SET-01418	PESQ. COM ELEITORES
IBOPE	IBO/MOGIDASCRUZES00.SET-01416	PESQ. COM ELEITORES
IBOPE	IBO/DIADEMA00.SET-01412	PESQ. COM ELEITORES
IBOPE	IBO/RIOBRANCO00.SET-01419	PESQ. COM ELEITORES
IBOPE	IBO/JOAOPESSOA00.SET-01420	PESQ. COM ELEITORES
IBOPE	IBO/IMPERATRIZ00.SET-01421	PESQ. COM ELEITORES
IBOPE	IBO/PELOTAS00.SET-01425	PESQ. COM ELEITORES
IBOPE	IBO/MACAPA00.SET-01424	PESQ. COM ELEITORES
IBOPE	IBO/FEIRADESANTANA00.SET-01422	PESQ. COM ELEITORES
IBOPE	IBO/SANTAMARIA00.SET-01427	PESQ. COM ELEITORES
IBOPE	IBO/PR00.SET-01428	PESQ. COM ELEITORES
IBOPE	IBO/FOZDOIGUAÇU00.SET-01429	PESQ. COM ELEITORES
IBOPE	IBO/RIOCLARO00.SET-01431	PESQ. COM ELEITORES
IBOPE	IBO/MARINGA00.SET-01432	PESQ. COM ELEITORES
IBOPE	IBO/LIMEIRA00.SET-01430	PESQ. COM ELEITORES
IBOPE	IBO/CAXIASDOSUL00.SET-01426	PESQ. COM ELEITORES
IBOPE	IBO/SOROCABA00.SET-01447	PESQ. COM ELEITORES
IBOPE	IBO/SP00.SET-01446	PESQ. COM ELEITORES
IBOPE	IBO/REGISTRO00.SET-01445	PESQ. COM ELEITORES
IBOPE	IBO/PRAIAGRANDE00.SET-01449	PESQ. COM ELEITORES
IBOPE	IBO/PONTAGROSSA00.SET-01442	PESQ. COM ELEITORES

IBOPE	IBO/JOINVILE00.SET-01438	PESQ. COM ELEITORES
IBOPE	IBO/ITABUNA00.SET-01437	PESQ. COM ELEITORES
IBOPE	IBO/FORTALEZA00.SET-01434	PESQ. COM ELEITORES
IBOPE	IBO/CARUARU00.SET-01433	PESQ. COM ELEITORES
IBOPE	IBO/CARUARU00.SET-01491	PESQ. COM ELEITORES
IBOPE	IBO/CAMPINAGRANDE00.SET-01435	PESQ. COM ELEITORES
IBOPE	IBO/BLUMENAU00.SET-01439	PESQ. COM ELEITORES
IBOPE	IBO/BAURU00.SET-01448	PESQ. COM ELEITORES
IBOPE	IBO/VITORIA00.SET-01462	PESQ. COM ELEITORES
IBOPE	IBO/SAOLUIS00.SET-01453	PESQ. COM ELEITORES
IBOPE	IBO/SAOJOSEDOSCAMPOS00.SET-01461	PESQ. COM ELEITORES
IBOPE	IBO/RIOCLARO00.SET-01460	PESQ. COM ELEITORES
IBOPE	IBO/FLORIANOPOLIS00.SET-01441	PESQ. COM ELEITORES
IBOPE	IBO/NATAL00.SET-01451	PESQ. COM ELEITORES
IBOPE	IBO/LONDRINA00.SET-01444	PESQ. COM ELEITORES
IBOPE	IBO/FLORIANOPOLIS00.SET-01440	PESQ. COM ELEITORES
IBOPE	IBO/CURITIBA00.SET-01443	PESQ. COM ELEITORES
IBOPE	IBO/CHAPECO00.SET-01474	PESQ. COM ELEITORES
IBOPE	IBO/CAXIASDOSUL00.SET-01535	PESQ. COM ELEITORES
IBOPE	IBO/BLUMENAU00.SET-01475	PESQ. COM ELEITORES
IBOPE	IBO/BELEM00.SET-01450	PESQ. COM ELEITORES
IBOPE	IBO/ARACAJU00.SET-01457	PESQ. COM ELEITORES
IBOPE	IBO/VITORIADACONQUISTA00.SET-01492	PESQ. COM ELEITORES
IBOPE	IBO/VARGINHA00.SET-01487	PESQ. COM ELEITORES
IBOPE	IBO/UBERLANDIA00.SET-01483	PESQ. COM ELEITORES
IBOPE	IBO/TERESINA00.SET-01497	PESQ. COM ELEITORES
IBOPE	IBO/SOROCABA00.SET-01468	PESQ. COM ELEITORES
IBOPE	IBO/SAOVICENTE00.SET-01454	PESQ. COM ELEITORES
IBOPE	IBO/SAOCARLOS00.SET-01456	PESQ. COM ELEITORES
IBOPE	IBO/SAOCAETANODOSUL00.SET-01463	PESQ. COM ELEITORES
IBOPE	IBO/SAOBERNARDODOCAMPO00.SET-01467	PESQ. COM ELEITORES
IBOPE	IBO/SANTOS00.SET-01472	PESQ. COM ELEITORES
IBOPE	IBO/RPR2000.SET-01494	PESQ. COM ELEITORES
IBOPE	IBO/PORTOVELHO00.SET-01495	PESQ. COM ELEITORES
IBOPE	IBO/PETROLINA00.SET-01493	PESQ. COM ELEITORES
IBOPE	IBO/OSASCO00.SET-01465	PESQ. COM ELEITORES
IBOPE	IBO/MONTESCLAROS00.SET-01482	PESQ. COM ELEITORES
IBOPE	IBO/MOGIDASCRUZES00.SET-01469	PESQ. COM ELEITORES
IBOPE	IBO/MANAUS00.SET-01488	PESQ. COM ELEITORES
IBOPE	IBO/MACAPA00.SET-01489	PESQ. COM ELEITORES
IBOPE	IBO/JUIZDEFORA00.SET-01485	PESQ. COM ELEITORES
IBOPE	IBO/GUARULHOS00.SET-01470	PESQ. COM ELEITORES
IBOPE	IBO/GUARAPUAVA00.SET-01478	PESQ. COM ELEITORES
IBOPE	IBO/GOVERNADORVALADARES00.SET-01486	PESQ. COM ELEITORES
IBOPE	IBO/FRANCA00.SET-01458	PESQ. COM ELEITORES
IBOPE	IBO/DIADEMA00.SET-01473	PESQ. COM ELEITORES
IBOPE	IBO/CASCAVEL00.SET-01459	PESQ. COM ELEITORES
IBOPE	IBO/CAMPINAS00.SET-01466	PESQ. COM ELEITORES
IBOPE	IBO/CAMPINAGRANDE00.SET-01490	PESQ. COM ELEITORES
IBOPE	IBO/BOAVISTA00.SET-01496	PESQ. COM ELEITORES
IBOPE	IBO/BAURU00.SET-01471	PESQ. COM ELEITORES
IBOPE	IBO/IMPERATRIZ00.SET-01498	PESQ. COM ELEITORES
IBOPE	IBO/ARARAQUARA00.SET-01499	PESQ. COM ELEITORES
IBOPE	IBO/RJ00.SET-01501	PESQ. COM ELEITORES
IBOPE	IBO/PELOTAS00.SET-01477	PESQ. COM ELEITORES
IBOPE	IBO/MARINGA00.SET-01479	PESQ. COM ELEITORES
IBOPE	IBO/JOINVILLE00.SET-01476	PESQ. COM ELEITORES
IBOPE	IBO/FOZDOIGUAÇU00.SET-01480	PESQ. COM ELEITORES
IBOPE	IBO/BH00.SET-01484	PESQ. COM ELEITORES
IBOPE	IBO/SP00.SET-01500	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/RIOCLARO00.SET-01502	PESQ. COM ELEITORES
IBOPE	IBO/MACEIO00.SET-01452	PESQ. COM ELEITORES
IBOPE	IBO/RECIFE00.SET-01503	PESQ. COM ELEITORES
IBOPE	IBO/ARARAQUARA00.SET-01504	PESQ. COM ELEITORES
IBOPE	IBO/REGISTRO00.SET-01369	PESQ. COM ELEITORES
IBOPE	IBO/PORTOALEGRE00.OUT-01505	PESQ. COM ELEITORES
IBOPE	IBO/RECIFE00.OUT-01506	PESQ. COM ELEITORES
IBOPE	IBO/SP00.OUT-01513	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/SANTOS00.OUT-01516	PESQ. COM ELEITORES
IBOPE	IBO/RJ00.OUT-01512	PESQ. COM ELEITORES
IBOPE	IBO/PELOTAS00.OUT-01510	PESQ. COM ELEITORES

IBOPE	IBO/GUARULHOS00.OUT-01515	PESQ. COM ELEITORES
IBOPE	IBO/GOIANIA00.OUT-01509	PESQ. COM ELEITORES
IBOPE	IBO/FORTALEZA00.OUT-01507	PESQ. COM ELEITORES
IBOPE	IBO/DIADEMA00.OUT-01514	PESQ. COM ELEITORES
IBOPE	IBO/BH00.OUT-01511	PESQ. COM ELEITORES
IBOPE	IBO/MOGIDASCRUZES00.OUT-01518	PESQ. COM ELEITORES
IBOPE	IBO/MANAUS00.OUT-01519	PESQ. COM ELEITORES
IBOPE	IBO/MACEIO00.OUT-01508	PESQ. COM ELEITORES
IBOPE	IBO/SP00.OUT-01522	PESQ. DE OPINIÃO POLÍTICA
IBOPE	IBO/MARINGA00.OUT-01521	PESQ. COM ELEITORES
IBOPE	IBO/CAMPINAS00.OUT-01520	PESQ. COM ELEITORES
IBOPE	IBO/RJ00.OUT-01529	PESQ. COM ELEITORES
IBOPE	IBO/RECIFE00.OUT-01527	PESQ. COM ELEITORES
IBOPE	IBO/PORTOALEGRE00.OUT-01524	PESQ. COM ELEITORES
IBOPE	IBO/PELOTAS00.OUT-01525	PESQ. COM ELEITORES
IBOPE	IBO/LONDRINA00.OUT-01526	PESQ. COM ELEITORES
IBOPE	IBO/GOIANIA00.OUT-01523	PESQ. COM ELEITORES
IBOPE	IBO/BH00.OUT-01528	PESQ. COM ELEITORES
IBOPE	IBO/CURITIBA00.OUT-01530	PESQ. COM ELEITORES
IBOPE	IBO/CAMPINAS00.OUT-01531	PESQ. COM ELEITORES
IBOPE	IBO/UBERLANDIA00.OUT-01533	PESQ. COM ELEITORES
IBOPE	IBO/MAUA00.OUT-01534	PESQ. COM ELEITORES
IBOPE	IBO/FORTALEZA00.OUT-01532	PESQ. COM ELEITORES
IBOPE	IBO/SP02.AGO-01759	PESQ. COM ELEITORES
IBOPE	IBO/TO02.AGO-01763	
IBOPE	IBO/RJ02.AGO-01762	
IBOPE	IBO/MA02.SET-01757	
IBOPE	IBO/AC02.AGO-01764	
IBOPE	IBO/MG02.AGO-01761	
IBOPE	IBO/AL02.SET-01765	
IBOPE	IBO/ES02.SET-01768	
IBOPE	IBO/DF02.SET-01767	
IBOPE	IBO/BA02.SET-01769	
IBOPE	IBO/MT02.SET-01771	
IBOPE	IBO/MS02.SET-01772	
IBOPE	IBO/RN02.SET-01774	
IBOPE	IBO/AP02.SET-01773	
IBOPE	IBO/SP02.SET-01778	PESQ. COM ELEITORES
IBOPE	IBO/AM02.SET-01775	
IBOPE	IBO/RJ02.SET-01781	
IBOPE	IBO/PR02.SET-01780	
IBOPE	IBO/MG02.SET-01777	
IBOPE	IBO/MA02.SET-01776	
IBOPE	IBO/TO02.SET-01783	
IBOPE	IBO/GO02.SET-01784	
IBOPE	IBO/SC02.SET-01785	
IBOPE	IBO/GO02.SETO-01766	
IBOPE	IBO/ES02.SET-01786	
IBOPE	IBO/CE02.SET-01782	
IBOPE	IBO/BA02.SET-01787	
IBOPE	IBO/RS02.SET-01790	
IBOPE	IBO/RJ02.SET-01793	GLOBO 6ª RODADA
IBOPE	IBO/SP02.SET-01794	PESQ. COM ELEITORES
IBOPE	IBO/PI02.SET-01788	
IBOPE	IBO/SC02.SET-01789	
IBOPE	IBO/RS02.SET-01792	
IBOPE	IBO/PR02.SET-01791	
IBOPE	IBO/RO02.SET-01795	
IBOPE	IBO/SE02.SET-01758	
IBOPE	IBO/SC02.OUT-01798	
IBOPE	IBO/RS02.OUT-01797	
IBOPE	IBO/SP02.OUT-01801	
IBOPE	IBO/PR02.OUT-01802	
IBOPE	IBO/RR02.O2-01806	
IBOPE	IBO/SE02.OUT-01804	
IBOPE	IBO/CE02.OUT-01803	
IBOPE	IBO/RN02.Out-01800	
IBOPE	IBO/SC02.OUT-01807	
IBOPE	IBO/PA02.OUT-01805	
IBOPE	IBO/PR02.OUT-01809	
IBOPE	IBO/DF02.OUT-01808	

IBOPE	IBO/CE02.OUT-01810	
IBOPE	IBO/BR02.SET-01796	
IBOPE	IBO/SC02.OUT-01798	
IBOPE	IBO/RS02.OUT-01797	
IBOPE	IBO/SP02.OUT-01801	
IBOPE	IBO/PR02.OUT-01802	
IBOPE	IBO/BR02.OUT-01799	
IBOPE	IBO/RR02.O2-01806	
IBOPE	IBO/SE02.OUT-01804	
IBOPE	IBO/CE02.OUT-01803	
IBOPE	IBO/RN02.Out-01800	
IBOPE	IBO/SC02.OUT-01807	
IBOPE	IBO/PA02.OUT-01805	
IBOPE	IBO/PR02.OUT-01809	
IBOPE	IBO/DF02.OUT-01808	
IBOPE	IBO/CE02.OUT-01810	
IBOPE	IBO/SBC.AGO/SET-02207	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NITERÓI.AGO/SET-02105	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/GUARULHOS.AGO/SET-02206	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NOVA IGUAÇU.AGO/SET-02106	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JOINVILLE.SET-02156	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SALVADOR.SET-01890	PESQ. ELEIT.
IBOPE	IBO/LONDRINA.SET-02046	PESQ. ELEIT.
IBOPE	IBO/FORTALEZA.SET-01912	PESQ. ELEIT.
IBOPE	IBO/CURITIBA.SET-02034	PESQ. ELEIT.
IBOPE	IBO/BELOHORIZONTE.SET-01970	PESQ. ELEIT.
IBOPE	IBO/BREJO SANTO.SET-01900	PESQ. ELEIT.
IBOPE	IBO/SÃO GONÇALO.SET-02110	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIO DAS OSTRAS.SET-02109	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/VILA VELHA.SET-01931	PESQ. ELEIT.
IBOPE	IBO/SAQUAREMA.SET-02108	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NILÓPOLIS.SET-02114	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JOAOPESSOA.SET-02021	PESQ. ELEIT.
IBOPE	IBO/CUBATÃO.SET-02209	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/ANANINDEUA.SET-02004	PESQ. ELEIT.
IBOPE	IBO/UMUARAMA.SET-02066	PESQ. ELEIT.
IBOPE	IBO/REGISTRO.SET-02212	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NATAL.SET-02134	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/DUQUE DE CAXIAS.SET-02111	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CAMPINAGRANDE.SET-02016	PESQ. ELEIT.
IBOPE	IBO/SÃO VICENTE.SET-02210	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/IPATINGA.SET-01976	PESQ. ELEIT.
IBOPE	IBO/GOVERNADORVALADARES.SET-01974	PESQ. ELEIT.
IBOPE	IBO/FOZDOIGUAÇU.SET-02040	PESQ. ELEIT.
IBOPE	IBO/SANTOS.SET-02211	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/FORTALEZA.SET-01909	PESQ. ELEIT.
IBOPE	IBO/ALTAMIRA.SET-02001	PESQ. ELEIT.
IBOPE	IBO/PRAIA GRANDE.SET-02217	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/ITAPEVI.SET-02213	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CABO FRIO.SET-02113	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/VOLTA REDONDA.SET-02112	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PONTAGROSSA.SET-02061	PESQ. ELEIT.
IBOPE	IBO/CARUARU.SET-02077	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SJRP.SET-02219	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIBEIRÃO PRETO.SET-02214	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PRIMAVERA DO LESTE.SET-01957	PESQ. ELEIT.
IBOPE	IBO/PETROLINA.SET-02078	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PATOBRANCO.SET-02059	PESQ. ELEIT.
IBOPE	IBO/LAVRAS DA MANGABEIRA.SET-01922	PESQ. ELEIT.
IBOPE	IBO/ITAPETININGA.SET-02215	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/GUARUJÁ.SET-02220	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CAMPINAS.SET-02216	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CAMPINAS.SET-02254	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/ACARAU.SET-01898	PESQ. ELEIT.
IBOPE	IBO/UBERLANDIA.SET-01992	PESQ. ELEIT.
IBOPE	IBO/SÃO PAULO.SET-02218	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIO DE JANEIRO.SET-02115	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PETRÓPOLIS.SET-02116	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PALMAS.SET-02274	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MOGI DAS CRUZES.SET-02221	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/FEIRA DE SANTANA.SET-01883	PESQ. ELEIT.

IBOPE	IBO/CASTANHAL.SET-02011	PESQ. ELEIT.
IBOPE	IBO/CASCAVEL.SET-02027	PESQ. ELEIT.
IBOPE	IBO/TERESINA.SET-02086	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SAO LUIS.SET-01949	PESQ. ELEIT.
IBOPE	IBO/SALVADOR.SET-01887	PESQ. ELEIT.
IBOPE	IBO/RIO CLARO.SET-02224	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RECIFE.SET-02079	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/QUIXADA.SET-01927	PESQ. ELEIT.
IBOPE	IBO/PRESIDENTE PRUDENTE.SET-02228	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/POCOSDECALDAS.SET-01989	PESQ. ELEIT.
IBOPE	IBO/BELOHORIZONTE.SET-01968	PESQ. ELEIT.
IBOPE	IBO/VARGINHA.SET-01999	PESQ. ELEIT.
IBOPE	IBO/QUIXERAMOBIM.SET-01928	PESQ. ELEIT.
IBOPE	IBO/IMPERATRIZ.SET-01946	PESQ. ELEIT.
IBOPE	IBO/ARARAQUARA.SET-02223	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/VILA VELHA.SET-01932	PESQ. ELEIT.
IBOPE	IBO/SÃO CARLOS.SET-02225	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/QUISSAMÃ.SET-02117	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MACEIO.SET-01873	PESQ. ELEIT.
IBOPE	IBO/FORTALEZA.SET-01910	PESQ. ELEIT.
IBOPE	IBO/CAXIAS DO SUL.SET-02148	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/BELEM.SET-02007	PESQ. ELEIT.
IBOPE	IBO/ANANINDEUA.SET-02003	PESQ. ELEIT.
IBOPE	IBO/ARMAÇÃO DE BÚZIOS.SET-02118	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/GURUPI.SET-02275	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/UNAI.SET-01997	PESQ. ELEIT.
IBOPE	IBO/SANTAREM.SET-02013	PESQ. ELEIT.
IBOPE	IBO/PIRACICABA.SET-02229	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JOAOPESSOA.SET-02022	PESQ. ELEIT.
IBOPE	IBO/CACERES.SET-01953	PESQ. ELEIT.
IBOPE	IBO/SALVADOR.SET-01888	PESQ. ELEIT.
IBOPE	IBO/IGUATU.SET-01918	PESQ. ELEIT.
IBOPE	IBO/CANINDE.SET-01901	PESQ. ELEIT.
IBOPE	IBO/CAMPINAGRANDE.SET-02017	PESQ. ELEIT.
IBOPE	IBO/ARACAJU.SET-02163	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/VITÓRIA DE SANTO ANTÃO.SET-02080	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MOGI DAS CRUZES.SET-02240	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MASSAPE.SET-01925	PESQ. ELEIT.
IBOPE	IBO/CABO FRIO.SET-02120	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/VILA VELHA.SET-01933	PESQ. ELEIT.
IBOPE	IBO/SÃO PEDRO DA ALDEIA.SET-02121	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SAO LUIS.SET-01950	PESQ. ELEIT.
IBOPE	IBO/RIO CLARO.SET-02234	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CURITIBA.SET-02033	PESQ. ELEIT.
IBOPE	IBO/ASSIS.SET-02231	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/ARAGUAÍNA.SET-02276	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SÃO VICENTE.SET-02232	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PORTO ALEGRE.SET-02149	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/FLORIANÓPOLIS.SET-02157	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/LIMOEIRO DO NORTE.SET-01924	PESQ. ELEIT.
IBOPE	IBO/GUARUJÁ.SET-02237	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CRATO.SET-01905	PESQ. ELEIT.
IBOPE	IBO/UBERLANDIA.SET-01993	PESQ. ELEIT.
IBOPE	IBO/TERESINA.SET-02087	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SJRP.SET-02239	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RECIFE.SET-02082	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PORTO VELHO.SET-02140	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NOVA IGUAÇU.SET-02122	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JUAZEIRO DO NORTE.SET-01920	PESQ. ELEIT.
IBOPE	IBO/ITAPETININGA.SET-02235	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CARUARU.SET-02081	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/VARZEAGRANDE.SET-01963	PESQ. ELEIT.
IBOPE	IBO/SÃO PAULO.SET-02233	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SBC.SET-02236	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIO DE JANEIRO.SET-02123	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIO CLARO.SET-02241	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PETROLINA.SET-02083	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PARANAVAI.SET-02057	PESQ. ELEIT.
IBOPE	IBO/MANAUS.SET-01881	PESQ. ELEIT.
IBOPE	IBO/IMPERATRIZ.SET-01947	PESQ. ELEIT.
IBOPE	IBO/CUIABA.SET-01956	PESQ. ELEIT.

IBOPE	IBO/CUBATÃO.SET-02248	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/BELOHORIZONTE.SET-01969	PESQ. ELEIT.
IBOPE	IBO/BAURU.SET-02249	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/UMUARAMA.SET-02067	PESQ. ELEIT.
IBOPE	IBO/TAUBATÉ.SET-02243	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SJC. SET-02242	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SÃO GONÇALO.SET-02125	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PRAIA GRANDE.SET-02253	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MOGI DAS CRUZES.SET-02250	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MACEIO.SET-01874	PESQ. ELEIT.
IBOPE	IBO/GUARULHOS.SET-02247	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/GUARAPUAVA.SET-02044	PESQ. ELEIT.
IBOPE	IBO/FRANCA.SET-02238	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/DUQUE DE CAXIAS.SET-02124	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CAUCAIA.SET-01903	PESQ. ELEIT.
IBOPE	IBO/BOA VISTA.SET-02143	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/BELEM.SET-02008	PESQ. ELEIT.
IBOPE	IBO/VITÓRIA.SET-02278	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SOROCABA.SET-02245	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SAO LUIS.SET-01951	PESQ. ELEIT.
IBOPE	IBO/SALVADOR.SET-01889	PESQ. ELEIT.
IBOPE	IBO/RIOBRANCO.SET-01871	PESQ. ELEIT.
IBOPE	IBO/PRESIDENTE PRUDENTE.SET-02251	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/POCOSDECALDAS.SET-01990	PESQ. ELEIT.
IBOPE	IBO/PIRACICABA.SET-02252	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PALMAS.SET-02277	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NATAL.SET-02135	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JUIZDEFORA.SET-01979	PESQ. ELEIT.
IBOPE	IBO/GOVERNADORVALADARES.SET-01975	PESQ. ELEIT.
IBOPE	IBO/FOZDOIGUAÇU.SET-02041	PESQ. ELEIT.
IBOPE	IBO/CAMPINAGRANDE.SET-02018	PESQ. ELEIT.
IBOPE	IBO/VARGINHA.SET-02000	PESQ. ELEIT.
IBOPE	IBO/SÃO CARLOS.SET-02256	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SANTOS.SET-02255	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIBEIRÃO PRETO.SET-02244	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NITERÓI.SET-02126	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/GOIANIA.SET-01941	PESQ. ELEIT.
IBOPE	IBO/FEIRA DE SANTANA.SET-01884	PESQ. ELEIT.
IBOPE	IBO/ARARAQUARA.SET-02246	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PONTAGROSSA.SET-02062	PESQ. ELEIT.
IBOPE	IBO/MARINGA.SET-02052	PESQ. ELEIT.
IBOPE	IBO/JOINVILLE.SET-02067	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CAXIAS DO SUL.SET-02150	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JOAOPESSOA.SET-02023	PESQ. ELEIT.
IBOPE	IBO/FORTALEZA.SET-01911	PESQ. ELEIT.
IBOPE	IBO/SÃO PAULO.OUT-02258	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIO DE JANEIRO.OUT-02127	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIBEIRÃO PRETO.OUT-02257	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RECIFE.OUT-02084	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PORTOALEGRE.OUT-02151	PESQ. ELEIT.
IBOPE	IBO/MOGI DAS CRUZES.OUT-02260	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/FLORIANÓPOLIS.OUT-02159	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SPCAP04.OUT-02302	PROJETO GLOBO - PROGNÓSTICO
IBOPE	IBO/SÃO JOÃO DE MERITI.OUT-02128	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PIRACICABA.OUT-02261	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NITERÓI.OUT-02130	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/RIBEIRÃO PRETO.OUT-02263	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MACEIO.OUT-01875	PESQ. ELEIT.
IBOPE	IBO/CAMPINAS.OUT-02262	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/VITORIA.OUT-01937	PESQ. ELEIT.
IBOPE	IBO/PORTO ALEGRE.OUT-02152	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JUIZDEFORA.OUT-01980	PESQ. ELEIT.
IBOPE	IBO/FLORIANÓPOLIS.OUT-02160	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CAMPINAGRANDE.OUT-02019	PESQ. ELEIT.
IBOPE	IBO/BELEM.OUT-02009	PESQ. ELEIT.
IBOPE	IBO/SJRP.OUT-02266	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/PONTAGROSSA.OUT-02063	PESQ. ELEIT.
IBOPE	IBO/BAURU.OUT-02265	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/UBERLANDIA.OUT-01994	PESQ. ELEIT.
IBOPE	IBO/TERESINA.OUT-02088	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SALVADOR.OUT-01892	PESQ. ELEIT.

IBOPE	IBO/NATAL.OUT-02136	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MARINGA.OUT-02053	PESQ. ELEIT.
IBOPE	IBO/DUQUE DE CAXIAS.OUT-02129	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/NITERÓI.OUT-02131	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MONTESCLAROS.OUT-01986	PESQ. ELEIT.
IBOPE	IBO/LONDRINA.OUT-02048	PESQ. ELEIT.
IBOPE	IBO/PIRACICABA.OUT-02267	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/JUIZDEFORA.OUT-01981	PESQ. ELEIT.
IBOPE	IBO/GOIANIA.OUT-01942	PESQ. ELEIT.
IBOPE	IBO/FLORIANÓPOLIS.OUT-02161	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/CURITIBA.OUT-02036	PESQ. ELEIT.
IBOPE	IBO/SÃO PAULO.OUT-02271	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/UBERLANDIA.OUT-01995	PESQ. ELEIT.
IBOPE	IBO/RIBEIRÃO PRETO.OUT-02270	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MACEIO.OUT-01876	PESQ. ELEIT.
IBOPE	IBO/CAMPINAS.OUT-02269	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SALVADOR.OUT-01893	PESQ. ELEIT.
IBOPE	IBO/NATAL.OUT-02137	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/SOROCABA.OUT-02271	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/MARINGA.OUT-02054	PESQ. ELEIT.
IBOPE	IBO/BAURU.OUT-02268	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
IBOPE	IBO/FORTALEZA.OUT-01914	PESQ. ELEIT.
IBOPE	IBO/FORTALEZA.OUT-01915	PESQ. ELEIT.
IBOPE	IBO/FORTALEZA.OUT-01916	PESQ. ELEIT.
IBOPE	IBO/DUQUE DE CAXIAS.OUT-02132	PESQ. DE O.P. SOBRE ASSUNTOS POLÍTICOS
DATAFOLHA	DAT/BR89.OUT-00196	INT. DE VOTO para pressidente VIII
DATAFOLHA	DAT/BR89.DEZ-00211	INT. DE VOTO PARA PRES. XIX - CEDEC II
DATAFOLHA	DAT/BR89.DEZ-00210	INT. DE VOTO PARA PRES. XVII - CEDEC II
DATAFOLHA	DAT/BR89.DEZ-00212	INT. DE VOTO para PRES. XX
DATAFOLHA	DAT/BR94.SET-00377	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/BR94.SET-00379	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP96.SET-00663	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SÃOLUIS96.SET-00674	INT. DE VOTO PARA PREF. DE SÃO LUIS
DATAFOLHA	DAT/SALVADOR96.SET-00671	INT. DE VOTO PARA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ96.SET-00664	INT. DE VOTO PARA PREF. DA CIDADE DO RIO DE
DATAFOLHA	DAT/RECIFE96.SET-00672	INT. DE VOTO PARA PREF. DE RECIFE
DATAFOLHA	DAT/POA96.SET-00666	INT. DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/GOIÂNIA96.SET-00669	INT. DE VOTO PARA PREF. DE GOIANIA
DATAFOLHA	DAT/FORTALEZA96.SET-00673	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/FLORIANÓPOLIS96.SET-00668	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/CUR96.SET-00667	INT. DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/CAMPOGRANDE96.SET-00670	INT. DE VOTO PARA PREF. DE CAMPO GRANDE
DATAFOLHA	DAT/BH96.SET-00665	INT. DE VOTO PARA PREF. DE BELO
DATAFOLHA	DAT/SP96.SET-00687	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SÃOLUÍS96.SET-00685	INTENÇÃO DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/SÃOJOSEDOSCAMPOS96.SET-00698	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/SÃOJOSÉDORIOPRETO96.SET-00692	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/SÃOCAETANODOSUL96.SET-00695	INT. DE VOTO PARA PREF. DE SÃO CAETANO
DATAFOLHA	DAT/SÃOBERNARDODOCAMPO96.SET-00694	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/SANTOANDRE96.SET-00693	INT. DE VOTO PARA PREF. DE SANTO ANDRÉ
DATAFOLHA	DAT/SALVADOR96.SET-00682	INT. DE VOTO PARA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ96.SET-00675	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO
DATAFOLHA	DAT/RECIFE96.SET-00683	INT. DE VOTO PARA PREF. DE RECIFE
DATAFOLHA	DAT/POA96.SET-00677	INT. DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/PIRACICABA96.SET-00689	INT. DE VOTO PARA PREF. DE PIRACICABA
DATAFOLHA	DAT/GOIANIA96.SET-00680	INT. DE VOTO PARA PREF. DE GOIÂNIA
DATAFOLHA	DAT/FRANCA96.SET-00691	INT. DE VOTO PARA PREF. DE FRANCA
DATAFOLHA	DAT/FORTALEZA96.SET-00684	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/FLORIANOPOLIS96.SET-00679	INT. DE VOTO PARA PREF. DE FLORIANÓPOLIS
DATAFOLHA	DAT/DIADEMA96.SET-00696	INT. DE VOTO PARA PREF. DE DIADEMA
DATAFOLHA	DAT/SANTOS96.SET-00697	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA		INT. DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/CAMPOGRANDE96.SET-00681	INT. DE VOTO PARA PREF. DE CAMPO GRANDE
DATAFOLHA		INT. DE VOTO PARA O PREF. DE CAMPINAS
DATAFOLHA	DAT/BH96.SET-00676	INT. DE VOTO PARA PREF. DE BELO
DATAFOLHA	DAT/SP96.SET-00700	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SOROCABA96.SET-00716	INT. DE VOTO PARA PREF. DE SOROCABA
DATAFOLHA	DAT/SP96.SET-00701	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SAOLUIS96.SET-00712	INT. DE VOTO PARA PREF. DE SÃO LUÍS
DATAFOLHA	DAT/SAOJOSEDORIOPRETO96.SET-00719	INT. DE VOTO PARA PREF. DE SÃO JOSÉ DO RIO
DATAFOLHA	DAT/SAOCAETANODOSUL96.SET-00722	INT. DE VOTO PARA PREF. DE SÃO CAETANO DO SUL

DATAFOLHA	DAT/SAOBERNARDO96.SET-00721	INT. DE VOTO PARA PREF. DE SÃO BERNARDO DO
DATAFOLHA	DAT/SANTOS96.SET-00724	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA	DAT/SANTOANDRE96.SET-00720	INT. DE VOTO PARA PREF. DE SANTO ANDRÉ
DATAFOLHA	DAT/SAL96.SET-00709	INT. DE VOTO PARA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ96.SET-00702	INT. DE VOTO PARA PREF. DO RIO DE
DATAFOLHA	DAT/RIBEIRÃOPRETO96.SET-00717	INT. DE VOTO PARA PREF. DE RIBEIRÃO PRETO
DATAFOLHA	DAT/REC96.SET-00710	INT. DE VOTO PARA PREF. DE RECIFE
DATAFOLHA	DAT/POA96.SET-00704	INT. DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/PIRACICABA96.SET-00715	INT. DE VOTO PARA O PREF. DE PIRACICABA
DATAFOLHA	DAT/JUNDIAI96.SET-00714	INT. DE VOTO PARA PREF. DE JUNDIAI
DATAFOLHA	DAT/GOI96.SET-00707	INT. DE VOTO PARA PREF. DE GOIÂNIA
DATAFOLHA	DAT/FRANCA96.SET-00718	INT. DE VOTO PARA PREF. DE FRANCA
DATAFOLHA	DAT/FOR96.SET-00711	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/FLORIANÓPOLIS96.SET-00706	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/DIADEMA96.SET-00723	INT. DE VOTO PARA PREF. DE DIADEMA
DATAFOLHA	DAT/CUR96.SET-00705	INT. DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/CAMPOGRANDE96.SET-00708	INT. DE VOTO PARA PREF. DE CAMPO GRANDE
DATAFOLHA	DAT/CAMPINAS96.SET-00713	INT. DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	DAT/BH96.SET-00703	INT. DE VOTO PARA PREF. DE BELO
DATAFOLHA	DAT/SAOJOSEDOSCAMPOS96.DEZ-00949	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SAOCAETANODOSUL96.DEZ-00946	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SAOBERNARDO96.DEZ-00945	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SANTOS96.DEZ-00948	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SANTOANDRE96.DEZ-00944	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/POA96.OUT-00740	INT. DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/FLORIANÓPOLIS96.OUT-00742	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/DIADEMA96.DEZ-00947	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/CURITIBA96.OUT-00741	INT. DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/SP96.OUT-00950	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SP96.OUT-00951	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SP96.OUT-00953	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SAOLUIS96.OUT-00959	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/RJ96.OUT-00954	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/GOIANIA96.OUT-00956	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/FLORIANOPOLIS96.OUT-00957	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/CAMPOGRANDE96.OUT-00958	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/CAMPINAS96.OUT-00960	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/BELOHORIZONTE96.OUT-00955	INT. DE VOTO PARA PREF.
DATAFOLHA	DAT/SOROCABA96.OUT-00743	INT. DE VOTO PARA PREF. DE SOROCABA
DATAFOLHA	DAT/SÃO PAULO96.OUT-00748	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SJC96.OUT-00749	INT. DE VOTO PARA PREF. DE SÃO JOSÉ DOS
DATAFOLHA	DAT/SBC96.OUT-00745	INT. DE VOTO PARA PREF. DE SÃO BENARDO
DATAFOLHA	DAT/SBC96.OUT-00757	INT. DE VOTO PARA PREF. DE SÃO BENARDO
DATAFOLHA	DAT/SANTOS96.OUT-00746	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA	DAT/RIBEIRÃO PRETO96.OUT-00744	INT. DE VOTO PARA PREF. DE RIBEIRÃO
DATAFOLHA	DAT/RIBEIRÃO PRETO96.OUT-00756	INT. DE VOTO PARA PREF. DE RIBEIRÃO
DATAFOLHA	DAT/GOIÂNIA96.OUT-00760	INT. DE VOTO PARA PREF. DE GOIÂNIA
DATAFOLHA	DAT/FLORIANÓPOLIS96.OUT-00752	INT. DE VOTO PARA PREF. DE FLORIANÓPOLIS
DATAFOLHA	DAT/BELO HORIZONTE96.OUT-00751	INT. DE VOTO PARA PREF. DE BELO
DATAFOLHA	DAT/SOROCABA96.NOV-00755	INT. DE VOTO PARA PREF. DE SOROCABA
DATAFOLHA	DAT/SOROCABA96.NOV-00764	INT. DE VOTO PARA PREF. DE SOROCABA
DATAFOLHA	DAT/SÃO PAULO96.NOV-00761	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SP96.NOV-00770	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SÃO PAULO96.NOV-00781	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SÃOLUIS96.NOV-00734	INT. DE VOTO PARA PREF. DE SÃO LUIS
DATAFOLHA	DAT/SÃOLUIS96.NOV-00775	INT. DE VOTO PARA PREF. DE SÃO LUIS
DATAFOLHA	DAT/SJC96.NOV-00767	INT. DE VOTO PARA PREF. DE SÃO JOSÉ DOS
DATAFOLHA	DAT/SJC96.NOV-00780	INT. DE VOTO PARA PREF. DE SÃO JOSÉ DOS
DATAFOLHA	DAT/SBC96.NOV-00766	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/SBC96.NOV-00778	INT. DE VOTO PARA PREF. DE SÃO BENARDO
DATAFOLHA	DAT/SANTOS96.NOV-00758	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA	DAT/SANTOS96.NOV-00779	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA	DAT/RIO DE JANEIRO96.NOV-00750	INT. DE VOTO PARA PREF. DO RIO DE
DATAFOLHA	DAT/RIO DE JANEIRO96.NOV-00762	INT. DE VOTO PARA PREF. DO RIO DE
DATAFOLHA	DAT/RJ96.NOV-00771	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO
DATAFOLHA	DAT/RIBEIRÃO PRETO96.NOV-00765	INT. DE VOTO PARA PREF. DE RIBEIRÃO
DATAFOLHA	DAT/LONDRINA96.NOV-00759	INT. DE VOTO PARA PREF. DE LONDRINA
DATAFOLHA	DAT/LONDRINA96.NOV-00768	INT. DE VOTO PARA PREF. DE LONDRINA
DATAFOLHA	DAT/GOI96.SET-00733	INT. DE VOTO PARA PREF. DE GOIÂNIA
DATAFOLHA	DAT/GOIÂNIA96.NOV-00774	INT. DE VOTO PARA PREF. DE GOIÂNIA
DATAFOLHA	DAT/FLORIANÓPOLIS96.NOV-00732	INT. DE VOTO PARA PREF. DE

DATAFOLHA	DAT/FLORIANÓPOLIS96.NOV-00772	INT. DE VOTO PARA PREF. DE
DATAFOLHA	DAT/CAMPO GRANDE96.NOV-00753	INT. DE VOTO PARA PREF. DE CAMPO GRANDE
DATAFOLHA	DAT/CAMPO GRANDE96.NOV-00769	INT. DE VOTO PARA PREF. DE CAMPO GRANDE
DATAFOLHA	DAT/CAMPINAS96.NOV-00754	INT. DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	DAT/CAMPINAS96.NOV-00763	INT. DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	DAT/CAMPINAS96.NOV-00776	INT. DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	DAT/BELOHORIZONTE96.NOV-00731	INT. DE VOTO PARA PREF. DE BELO
DATAFOLHA	DAT/BELOHORIZONTE96.NOV-00772	INT. DE VOTO PARA PREF. DE BELO HORIZONTE
DATAFOLHA	DAT/SP98.SET-00869	INT. DE VOTO GOV. DE SÃO PAULO
DATAFOLHA	DAT/SP98.SET-00895	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/SC98.SET-00899	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/RS98.SET-00901	AVAL. PRES. FERNANDO HENRIQUE
DATAFOLHA	DAT/RJ98.DEZ-00904	AVAL. DO PRES. FERNANDO HENRIQUE
DATAFOLHA	DAT/PR98.SET-00902	AVAL. PRES. FERNANDO HENRIQUE
DATAFOLHA	DAT/PE98.SET-00900	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/MG98.SET-00903	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/DF98.SET-00898	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/CE98.SET-00896	AVAL. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/BA98.SET-00897	AVAL. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/SP98.SET-00879	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/SC98.SET-00886	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/RS98.SET-00883	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/RJ98.SET-00880	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/PR98.SET-00882	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/PE98.SET-00884	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/MG98.SET-00881	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/DF98.SET-00885	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/CE98.SET-00888	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/BA98.SET-00887	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/SP98.SET-00905	INT. DE VOTO PARA GOV. DE SÃO PAULO
DATAFOLHA	DAT/CE98.DEZ-00908	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP98.OUT-00916	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SC.OUT-00909	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RS98.OUT-00912	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RJ98.OUT-00915	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/PR98.OUT-00913	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/PE98.OUT-00911	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG98.OUT-00914	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SDF98.OUT-00910	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/BA98.OUT-00907	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP98.OUT-00917	INT. DE VOTO PARA GOV. DE SÃO PAULO
DATAFOLHA	DAT/SP98.OUT-00922	INT. DE VOTO PARA GOV. DE SÃO PAULO
DATAFOLHA	DAT/RS98.OUT-00919	INT. DE VOTO PARA GOV. DO RIO GRANDE
DATAFOLHA	DAT/RJ98.OUT-00921	INT. DE VOTO PARA GOV. DO RIO DE
DATAFOLHA	DAT/MG98.OUT-00920	INT. DE VOTO PARA GOV. DE MINAS
DATAFOLHA	DAT/DF98.OUT-00918	INT. DE VOTO PARA GPVERNADOR DO DISTRITO
DATAFOLHA	DAT/SP98.OUT-00927	INT. DE VOTO PARA GOV. DE SÃO PAULO
DATAFOLHA	DAT/SP98.OUT-00928	INT. DE VOTO PARA GOV. DE SÃO PAULO
DATAFOLHA	DAT/RS98.DEZ-00924	INT. DE VOTO PARA GOV. DO RIO GRANDE
DATAFOLHA	DAT/RJ98.OUT-00926	INT. DE VOTO PARA GOV. DO RIO DE
DATAFOLHA	DAT/MG98.OUT-00925	INT. DE VOTO PARA GOV. DE MINAS
DATAFOLHA	DAT/DF98.OUT-00923	INT. DE VOTO PARA GOV. DO DISTRITO
DATAFOLHA	DAT/SP98.OUT-00933	INT. DE VOTO PARA GOV. DE SÃO PAULO
DATAFOLHA	DAT/RS98.OUT-00930	INT. DE VOTO PARA GOV. DO RIO GRANDE
DATAFOLHA	DAT/RJ98.OUT-00932	INT. DE VOTO PARA GOV. DO RIO DE
DATAFOLHA	DAT/MG98.OUT-00931	INT. DE VOTO PARA GOV. DE MINAS
DATAFOLHA	DAT/DF98.OUT-00929	INT. DE VOTO PARA GOV. DO DISTRITO
DATAFOLHA	DAT/BR98.SET-00868	AVAL. FHC
DATAFOLHA	DAT/SP98.SET-00869	INT. DE VOTO GOV. DE SÃO PAULO
DATAFOLHA	DAT/BR98.SET-00870	AVAL. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/SP98.SET-00895	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/SC98.SET-00899	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/RS98.SET-00901	AVAL. PRES. FERNANDO HENRIQUE
DATAFOLHA	DAT/RJ98.DEZ-00904	AVAL. DO PRES. FERNANDO HENRIQUE
DATAFOLHA	DAT/PR98.SET-00902	AVAL. PRES. FERNANDO HENRIQUE
DATAFOLHA	DAT/PE98.SET-00900	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/MG98.SET-00903	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/DF98.SET-00898	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/CE98.SET-00896	AVAL. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/BA98.SET-00897	AVAL. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/BR98.SET-00863	AVAL. FERNANDO HENRIQUE CARDOSO

DATAFOLHA	DAT/SP98.SET-00879	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/SC98.SET-00886	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/RS98.SET-00883	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/RJ98.SET-00880	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/PR98.SET-00882	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/PE98.SET-00884	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/MG98.SET-00881	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/DF98.SET-00885	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/CE98.SET-00888	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/BA98.SET-00887	AVAL. PRES. FERNANDO HENRIQUE CARDOSO
DATAFOLHA	DAT/CE98.DEZ-00908	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/BR98.OUT-00906	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP98.OUT-00916	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SC.OUT-00909	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RS98.OUT-00912	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RJ98.OUT-00915	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/PR98.OUT-00913	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/PE98.OUT-00911	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG98.OUT-00914	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SDF98.OUT-00910	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/BA98.OUT-00907	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SPcap00.SET-01137	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SJC00.SET-01123	INT. DE VOTO PARA PREF. DE SÃO JOSÉ DOS CAMPOS
DATAFOLHA	DAT/CAM00.SET-01124	INT. DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	IBO/SP00.SET-01139	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SLU00.SET-01147	INT. DE VOTO PARA PREF. DE SÃO LUIS
DATAFOLHA	DAT/SAL00.SET-01145	INT. DE VOTO PARA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ00.SET-01140	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO
DATAFOLHA	DAT/REC00.SET-01142	INT. DE VOTO PARA PREF. DE RECIFE
DATAFOLHA	DAT/POA00.SET-01148	INT. DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/MAC00.SET-01144	INT. DE VOTO PARA PREF. DE MACEIÓ
DATAFOLHA	DAT/FOR00.SET-01146	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/CUR00.SET-01143	INT. DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/BH00.SET-01141	INT. DE VOTO PARA PREF. DE BELO HORIZONTE
DATAFOLHA	DAT/SP00.SET-01553	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SAOLUIS00.SET-01565	INT. DE VOTO PARA PREF. DE SAO LUIS
DATAFOLHA	DAT/SAOJOSEDOSCAMPOS00.SET-01561	INT. DE VOTO PARA PREF. DE SAO JOSE DOS CAMPOS
DATAFOLHA	DAT/SAOJOSEDORIOPRETO00.SET-01568	INT. DE VOTO PARA PREF. DE SAO JOSE DO RIO
DATAFOLHA	DAT/SAOCAETANODOSUL00.SET-01560	INT. DE VOTO PARA PREF. DE SAO CAETANO DO SUL
DATAFOLHA	DAT/SAOBERNARDODOCAMPO00.SET-01556	INT. DE VOTO PARA PREF. DE SAO BERNARDO DO
DATAFOLHA	DAT/SANTOS00.SET-01572	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA	DAT/SANTOANDRE00.SET-01555	INT. DE VOTO PARA PREF. DE SANTO ANDRE
DATAFOLHA	DAT/SALVADOR00.SET-01569	INT. DE VOTO PARA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ00.SET-01554	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO
DATAFOLHA	DAT/RIBEIRAOPRETO00.SET-01567	INT. DE VOTO PARA PREF. DE RIBEIRAO PRETO
DATAFOLHA	DAT/RECIFE00.SET-01570	INT. DE VOTO PARA PREF. DE RECIFE
DATAFOLHA	DAT/PORTOALEGRE00.SET-01559	INT. DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/OSASCO00.SET-01563	INT. DE VOTO PARA PREF. DE OSASCO
DATAFOLHA	DAT/MACEIO00.SET-01571	INT. DE VOTO PARA PREF. DE MACEIO
DATAFOLHA	DAT/GUARULHOS00.SET-01573	INT. DE VOTO PARA PREF. DE GUARULHOS
DATAFOLHA	DAT/FORTALEZA00.SET-01566	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/DIADEMA00.SET-01557	INT. DE VOTO PARA PREF. DE DIADEMA
DATAFOLHA	DAT/CURITIBA00.SET-01564	INT. DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/CAMPINAS00.SET-01562	INT. DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	DAT/BH00.SET-01558	INT. DE VOTO PARA PREF. DE BELO HORIZONTE
DATAFOLHA	IBO/SP00.SET-01575	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	IBO/SP00.SET-01574	INTENÇÃO DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SAOLUIS00.SET-01587	INT. DE VOTO PARA PREF. DE SAO LUIS
DATAFOLHA	DAT/SAOJOSEDOSCAMPOS00.SET-01583	INTENÇÃO DE VOTO PARA PREF. DE SJC
DATAFOLHA	DAT/SAOJOSEDORIOPRETO00.SET-01590	INT. DE VOTO PARA PREF. DE SAO JOSE DO RIO
DATAFOLHA	DAT/SAOCAETANODOSUL00.SET-01582	INTENÇÃO DE VOTO PARA PREF. DE SAO CAETANO
DATAFOLHA	DAT/SAOBERNARDODOCAMPO00.SET-01578	INT. DE VOTO PARA PREF. DE SAO BERNARDO DO
DATAFOLHA	DAT/SANTOS00.SET-01594	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA	DAT/SANTOANDRE00.SET-01577	INT. DE VOTO PARA PREF. DE SANTO ANDRE
DATAFOLHA	DAT/SALVADOR00.SET-01591	INT. DE VOTO PARA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ00.SET-01576	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO
DATAFOLHA	DAT/RIBEIRAOPRETO00.SET-01589	INT. DE VOTO PARA PREF. DE RIBEIRAO PRETO
DATAFOLHA	DAT/RECIFE00.SET-01592	INT. DE VOTO PARA PREF. DE RECIFE
DATAFOLHA	DAT/PORTOALEGRE00.SET-01581	INTENÇÃO DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/OSASCO00.SET-01585	INTENÇÃO DE VOTO PARA PREF. DE OSASCO
DATAFOLHA	DAT/MACEIO00.SET-01593	INT. DE VOTO PARA PREF. DE MACEIO

DATAFOLHA	DAT/GUARULHOS00.SET-01595	INT. DE VOTO PARA PREF. DE GUARULHOS
DATAFOLHA	DAT/FORTALEZA00.SET-01588	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/DIADEMA00.SET-01579	INT. DE VOTO PARA PREF. DE DIADEMA
DATAFOLHA	DAT/CURITIBA00.SET-01586	INTENÇÃO DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/CAMPINAS00.SET-01584	INTENÇÃO DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	DAT/BH00.SET-01580	INT. DE VOTO PARA PREF. DE BELO HORIZONTE
DATAFOLHA	IBO/SP00.SET-01596	INTENÇÃO DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/RJ00.SET-01597	INT. DE VOTO PARA PREF. DO RIO DE
DATAFOLHA	DAT/FOR00.SET-01598	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	IBO/SP00.SET-01159	INT. DE VOTO PARA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SAOLUIS00.SET-01171	INT. DE VOTO PARA PREF. DE SÃO LUIS
DATAFOLHA	DAT/SAOJOSEDOSCAMPOS00.SET-01167	INT. DE VOTO PARA PREF. DE SÃO JOSÉ DOS CAMPOS
DATAFOLHA	DAT/SAOJOSEDORIOPRETO00.SET-01174	INT. DE VOTO PARA PREF. DE SÃO JOSÉ DO RIO
DATAFOLHA	DAT/SAOCAETANODOSUL00.SET-01166	INT. DE VOTO PARA PREF. DE SÃO CAETANO DO SUL
DATAFOLHA	DAT/SBC00.SET-01162	INT. DE VOTO PARA PREF. DE SÃO BERNARDO DO CAMPO
DATAFOLHA	DAT/SANTOS00.SET-01178	INT. DE VOTO PARA PREF. DE SANTOS
DATAFOLHA	DAT/STA00.SET-01161	INT. DE VOTO PARA PREF. DE SANTO ANDRÉ
DATAFOLHA	DAT/SALVADOR00.SET-01175	INT. DE VOTO PARA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ00.SET-01160	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO
DATAFOLHA	DAT/RIBEIRÃOPRETO00.SET-01173	INT. DE VOTO PARA PREF. DE RIBEIRÃO PRETO
DATAFOLHA	DAT/RECIFE00.SET-01176	INT. DE VOTO PARA PREF. DE RECIFE
DATAFOLHA	DAT/POA00.SET-01165	INT. DE VOTO PARA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/OSASCO00.SET-01169	INT. DE VOTO PARA PREF. DE OSASCO
DATAFOLHA	DAT/MACEIO00.SET-01177	INT. DE VOTO PARA PREF. DE MACEIÓ
DATAFOLHA	DAT/GUARULHOS00.SET-01179	INT. DE VOTO PARA PREF. DE GUARULHOS
DATAFOLHA	DAT/FORTALEZA00.SET-01172	INT. DE VOTO PARA PREF. DE FORTALEZA
DATAFOLHA	DAT/DIA00.SET-01163	INT. DE VOTO PARA PREF. DE DIADEMA
DATAFOLHA	DAT/CURITIBA00.SET-01170	INT. DE VOTO PARA PREF. DE CURITIBA
DATAFOLHA	DAT/CAMPINAS00.SET-01168	INT. DE VOTO PARA PREF. DE CAMPINAS
DATAFOLHA	DAT/BH00.SET-01164	INT. DE VOTO PARA PREF. DE BELO HORIZONTE
DATAFOLHA	DAT/SP00.AGO-01180	BOCA DE URNA PREF. DE SÃO PAULO
DATAFOLHA	DAT/SAOLUIS00.OUT-01544	BOCA DE URNA PREF. DE SAO LUIS
DATAFOLHA	DAT/SJC00.OUT-01186	BOCA DE URNA PREF. DE SÃO JOSÉ DOS CAMPOS
DATAFOLHA	DAT/SAOJOSEDORIOPRETO00.OUT-01547	BOCA DE URNA PREF. DE SAO JOSE DO RIO PRETO
DATAFOLHA	DAT/SBC00.OUT-01182	BOCA DE URNA PREF. DE SÃO BERNARDO DO CAMPO
DATAFOLHA	DAT/SANTOS00.OUT-01551	BOCA DE URNA PREF. DE SANTOS
DATAFOLHA	DAT/SALVADOR00.OUT-01548	BOCA DE URNA PREF. DE SALVADOR
DATAFOLHA	DAT/RJ00.AGO-01181	BOCA DE URNA PREF. DO RIO DE JANEIRO
DATAFOLHA	DAT/RIBEIRAOPRETO00.OUT-01546	BOCA DE URNA PREF. DE RIBEIRAO PRETO
DATAFOLHA	DAT/RECIFE00.OUT-01549	BOCA DE URNA PREF. DE RECIFE
DATAFOLHA	DAT/POA00.OUT-01185	BOCA DE URNA PREF. DE PORTO ALEGRE
DATAFOLHA	DAT/MACEIO00.OUT-01550	BOCA DE URNA PREF. DE MACEIO
DATAFOLHA	DAT/GUARULHOS00.OUT-01552	BOCA DE URNA PREF. DE GUARULHOS
DATAFOLHA	DAT/FORTALEZA00.OUT-01545	BOCA DE URNA PREF. DE FORTALEZA
DATAFOLHA	DAT/DIA00.OUT-01183	BOCA DE URNA PREF. DE DIADEMA
DATAFOLHA	DAT/CUR00.OUT-01188	BOCA DE URNA PREF. DE CURITIBA
DATAFOLHA	DAT/CURITIBA00.OUT-01543	BOCA DE URNA PREF. DE CURITIBA
DATAFOLHA	DAT/CAM00.OUT-01187	BOCA DE URNA PREF. DE CAMPINAS
DATAFOLHA	DAT/BH00.OUT-01184	BOCA DE URNA PREF. DE BELO HORIZONTE
DATAFOLHA	DAT/SRP00.OUT-01613	INT. DE VOTO PARA PREF. DE SJRP - 2º TURNO
DATAFOLHA	DAT/SAN00.OUT-01616	INT. DE VOTO PARA PREF. DE SANTOS - 2º TURNO
DATAFOLHA	DAT/REC00.OUT-01614	INT. DE VOTO PARA PREF. DO RECIFE - 2º TURNO
DATAFOLHA	DAT/POA00.POA-01609	INT. DE VOTO PARA PREF. DE PORTO ALEGRE - 2º TURNO
DATAFOLHA	DAT/MAC00.OUT-01615	INT. DE VOTO PARA PREF. DE MACEIÓ - 2º TURNO
DATAFOLHA	DAT/GUA00.OUT-01617	INT. DE VOTO PARA PREF. DE GUARULHOS - 2º TURNO
DATAFOLHA	DAT/FOR00.OUT-01612	INT. DE VOTO PARA PREF. DE FORTALEZA - 2º TURNO
DATAFOLHA	DAT/DIA00.OUT-01607	INT. DE VOTO PARA PREF. DE DIADEMA - 2º TURNO
DATAFOLHA	DAT/CUR00.OUT-01611	INT. DE VOTO PARA PREF. DE CURITIBA - 2º TURNO
DATAFOLHA	DAT/CAM00.OUT-01610	INT. DE VOTO PARA PREF. DE CAMPINAS - 2º TURNO
DATAFOLHA	DAT/BH00.OUT-01608	INT. DE VOTO PARA PREF. DE BELO HORIZONTE - 2º TURNO
DATAFOLHA	DAT/SP00.OUT-01605	INT. DE VOTO PARA PREF. DE SÃO PAULO - 2º TURNO
DATAFOLHA	DAT/RJ00.AGO-01606	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO
DATAFOLHA	IBO/BELEM00.OUT-01517	PESQ. DE O.P.
DATAFOLHA	DAT/SPcap00.OUT-01619	INTENÇÃO DE VOTO PARA PREF. DE SÃO PAULO - 2° TURNO
DATAFOLHA	DAT/RJ00.OUT-01620	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO - 2º TURNO
DATAFOLHA	DAT/SJR00.OUT-01628	INT. DE VOTO PARA PREF. DE SJRP - 2º TURNO
DATAFOLHA	DAT/SAN00.OUT-01631	INT. DE VOTO PARA PREF. DE SANTOS - 2º TURNO
DATAFOLHA	DAT/POA00.OUT-01624	INT. DE VOTO PARA PREF. DE PORTO ALEGRE - 2º TURNO
DATAFOLHA	DAT/MAC00.OUT-01630	INT. DE VOTO PARA PREF. DE MACEIÓ - 2º TURNO
DATAFOLHA	DAT/GUA00.OUT-01632	INT. DE VOTO PARA PREF. DE GUARULHOS - 2º TURNO
DATAFOLHA	DAT/FOR00.OUT-01627	INT. DE VOTO PARA PREF. DE FORTALEZA - 2º TURNO

DATAFOLHA	DAT/CAM00.OUT-01625	INT. DE VOTO PARA PREF. DE CAMPINAS - 2º TURNO
DATAFOLHA	DAT/BH00.OUT-01623	INT. DE VOTO PARA PREF. DE BELO HORIZONTE - 2º TURNO
DATAFOLHA	DAT/SPcap00.OUT-01659	INT. DE VOTO PARA PREF. DE SÃO PAULO - 2º TURNO
DATAFOLHA	DAT/RJ00.OUT-01621	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO - 2º TURNO
DATAFOLHA	DAT/REC00.OUT-01629	INT. DE VOTO PARA PREF. DE RECIFE - 2º TURNO
DATAFOLHA	DAT/DAI00.OUT-01622	INT. DE VOTO PARA PREF. DE DIADEMA - 2º TURNO
DATAFOLHA	DAT/CUR00.OUT-01626	INT. DE VOTO PARA PREF. DE CURITIBA - 2º TURNO
DATAFOLHA	DAT/RJ00.OUT-01660	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO - 2º TURNO
DATAFOLHA	DAT/REC00.OUT-01661	INT. DE VOTO PARA PREF. DE RECIFE- 2º TURNO
DATAFOLHA	DAT/FOR00.OUT-01663	INT. DE VOTO PARA PREF. DE FORTALEZA - 2º TURNO
DATAFOLHA	DAT/DIA00.OUT-01666	INT. DE VOTO PARA PREF. DE DIADEMA - 2º TURNO
DATAFOLHA	DAT/CUR00.OUT-01662	INT. DE VOTO PARA PREF. DE CURITIBA - 2º TURNO
DATAFOLHA	DAT/SPcap00.OUT-01664	INT. DE VOTO PARA PREF. DE SÃO PAULO- 2o. TURNO
DATAFOLHA	DAT/RJ00.OUT-01665	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO - 2º TURNO
DATAFOLHA	DAT/REC00.OUT-01672	INT. DE VOTO PARA PREF. DE RECIFE - 2º TURNO
DATAFOLHA	DAT/POA00.OUT-01668	INT. DE VOTO PARA PREF. DE PORTO ALEGRE - 2º TURNO
DATAFOLHA	DAT/MAC00.OUT-01673	INT. DE VOTO PARA PREF. DE MACEIÓ - 2º TURNO
DATAFOLHA	DAT/GUA00.OUT-01674	INT. DE VOTO PARA PREF. DE GUARULHOS - 2º TURNO
DATAFOLHA	DAT/FOR00.OUT-01671	INT. DE VOTO PARA PREF. DE FORTALEZA - 2º TURNO
DATAFOLHA	DAT/CAM00.OUT-01669	INT. DE VOTO PARA PREF. DE CAMPINAS - 2º TURNO
DATAFOLHA	DAT/CUR00.OUT-01670	INT. DE VOTO PARA PREF. DE CURITIBA - 2º TURNO
DATAFOLHA	DAT/BH00.OUT-01667	INT. DE VOTO PARA PREF. DE BELO HORIZONTE - 2º TURNO
DATAFOLHA	DAT/MAC00.OUT-01640	INT. DE VOTO PARA PREF. DE MACEIÓ - 2º TURNO
DATAFOLHA	DAT/GUA00.OUT-01642	INT. DE VOTO PARA PREF. DE GUARULHOS - 2º TURNO
DATAFOLHA	DAT/FOR00.OUT-01643	INT. DE VOTO PARA PREF. DE FORTALEZA - 2º TURNO
DATAFOLHA	DAT/DIA00.OUT-01635	INT. DE VOTO PARA PREF. DE DIADEMA - 2º TURNO
DATAFOLHA	DAT/SPcap00.OUT-01633	INTENÇÃO DE VOTO PARA PREF. DE SÃO PAULO - 2ºTURNO
DATAFOLHA	DAT/SRP00.OUT-01645	INT. DE VOTO PARA PREF. DE SJRP - 2º TURNO
DATAFOLHA	DAT/SAN00.OUT-01641	INT. DE VOTO PARA PREF. DE SANTOS - 2º TURNO
DATAFOLHA	IBO/BR00.AGO-01634	INT. DE VOTO PARA PREF. DO RIO DE JANEIRO - 2º TURNO
DATAFOLHA	DAT/REC00.OUT-01639	INT. DE VOTO PARA PREF. DE RECIFE - 2º TURNO
DATAFOLHA	DAT/POA00.OUT-01637	INT. DE VOTO PARA PREF. DE PORTO ALEGRE - 2º TURNO
DATAFOLHA	DAT/CUR00.OUT-01644	INT. DE VOTO PARA PREF. DE CURITIBA - 2º TURNO
DATAFOLHA	DAT/CAM00.OUT-01638	INT. DE VOTO PARA PREF. DE CAMPINAS - 2º TURNO
DATAFOLHA	DAT/BH00.OUT-01636	INT. DE VOTO PARA PREF. DE BELO HORIZONTE - 2º TURNO
DATAFOLHA	DAT/SP00.OUT-01646	BOCA DE URNA PREF. DE SÃO PAULO - 2º TURNO
DATAFOLHA	DAT/SJRP00.OUT-01658	BOCA DE URNA PREF. DE SJRP - 2º TURNO
DATAFOLHA	DAT/SAN00.OUT-01654	BOCA DE URNA PREF. DE SANTOS - 2º TURNO
DATAFOLHA	DAT/RJ00.OUT-01647	
DATAFOLHA	DAT/REC00.OUT-01652	BOCA DE URNA PREF. DE RECIFE - 2º TURNO
DATAFOLHA	DAT/POA00.OUT-01650	BOCA DE URNA PREF. DE PORTO ALEGRE - 2º TURNO
DATAFOLHA	DAT/MAC00.OUT-01653	BOCA DE URNA PREF. DE MACEIÓ - 2º TURNO
DATAFOLHA	DAT/GUA00.OUT-01655	BOCA DE URNA PREF. DE GUARULHOS - 2º TURNO
DATAFOLHA	DAT/FOR00.OUT-01656	BOCA DE URNA PREF. DE FORTALEZA - 2º TURNO
DATAFOLHA	DAT/DIA00.OUT-01648	BOCA DE URNA PREF. DE DIADEMA - 2º TURNO
DATAFOLHA	DAT/CUR00.OUT-01657	BOCA DE URNA PREF. DE CURITIBA - 2º TURNO
DATAFOLHA	DAT/CAM00.OUT-01651	BOCA DE URNA PREF. DE CAMPINAS - 2º TURNO
DATAFOLHA	DAT/BH00.OUT-01649	BOCA DE URNA PREF. DE BELO HORIZONTE - 2º TURNO
DATAFOLHA	DAT/RJ02.SET-01700	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG02.SET-01701	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP02.SET-01702	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP02.SET-01703	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RJ02.SET-01704	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG02.SET-01705	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP02.OUT-01831	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RJ02.OUT-01830	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG02.OUT-01829	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RJ02.SET-01700	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG02.SET-01701	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP02.SET-01702	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/BR02.SET-01692	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP02.SET-01703	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RJ02.SET-01704	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG02.SET-01705	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/SP02.OUT-01831	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/RJ02.OUT-01830	INT. DE VOTO PARA PRES.
DATAFOLHA	DAT/MG02.OUT-01829	INT. DE VOTO PARA PRES.

Referências Bibliográficas

J. Aitchison. The statistical analysis of compositinal data. *Journal of the Royal Statistical Society - Series B*, 44(1):139–177, 1982. 174

J. Aitchison. *The Statistical Analysis of Compositional Data*. Chapman and Hall - London, 1986. 175

A. C. Almeida. *Como são feitas as Pesquisas Eleitorais e de Opinião*. Editora FGV, 2002. 87

A. C. Almeida. *A cabeça do Eleitor: Estratégia, Pesquisa e Vitória Eleitoral*. Editora Record, 2008. 87, 95

A. C. Almeida. *Erros nas Pesquisas Eleitorais e de Opinião*. Editora Record, 2009. 93

D. R. Appleton, J. M. French, and M. P. J. Vanderpump. Ignoring a covariate: An example of simpson's paradox. *The American Statistician*, 50(4):340–341, 1996. 32

D. Basu. On sampling with and without replacement. *Sankhyã: The Indian Journal of Statistics*, 20:287–294, 1958. 13

D. Basu. An essay on the logical foundations of survey sampling, part 1 (with discussion). *Foundations of Statistical Inference - Editors V. P. Godambe and D. A. Sprott*, pages 203–242, 1971. 149

J. O. Berger and R. L. Wopert. *The Likelihood Principle (Second Edition)*. Lecture Notes - Monograph Series, 1988. 59

J. M. Bernardo. Monitoring the 1982 spanish socialist victory: a bayesian analysis. *Journal of the American Statistical Association*, 79(387):510–515, 1984. 62

J. M. Bernardo. The concept of exchangeability and its aplications. *Far East Journal of Mathematical Science*, 4:111–121, 1996. 60

H. Bolfarine and W. O. Bussab. *Elementos de Amostragem*. Editora Edgard Blücher, 2005. 7, 15, 18, 24, 31, 38, 39

J. F. Bromaghin. Sample size determination for interval estimation of multinomial probabilities. *The American Statistician*, 47(3):203–206, 1993. 25

J. F. Carvalho and C. Ferraz. A falsidade das margens de erro em pesquisas eleitorais baseadas em amotragem por quotas. *Boletim informativo da Associação Brasileira de Estatística (ABE)*, 64: 14–16, 2006. 76

W. G. Cochran. *Sampling Techniques - third edition*. John Wiley & Sons, 1977. 7, 29, 42, 52

W. G. Cochran, F. Mosteller, and J. W. Tukey. Principles of sampling. *Journal of the American Statistical Association*, 49:13–35, 1954. 66

D. J. Colwell and J. R. Gillett. A truncated geometric distribution. *The Mathematical Gazette*, 73 (466):332–333, 1989. 123

L. K. Cordani and S. Wechsler. Teaching independence and exchangeability. *Proceedings of 7th ICOTS*, 2:1–5, 2006. 60

I. Crespi. *Pre-election Polling: Sources of Accuracy and Error*. New York: Russel Sage, 1988. 183

B. de Finetti. *Theory of Probability - Volumes I and II*. John Wiley & Sons, 1974. 58

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977. 138

J. Desart and T. Holbrook. Campaigns, polls, and the states: Assessing the accuracy of statewide presidential trial-heat polls. *Political Research Quarterly*, 56(4):431–439, 2003. 165, 183

J.-C. Deville. A theory of quota surveys. *Survey Methodology*, 17(2):163–181, 1991. 85

B. Efron and D. V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–487, 1978. 141

W. A. Ericson. Subjective bayesian models in sampling finite populations. *Journal of the Royal Statistical Society - Series B (Methodological)*, 31(2):195–233, 1969. 60

I. P. Fellegi. Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, 58(301):183–201, 1963. 42

C. Ferraz. *Crítica Metodológica às Pesquisas Eleitorais no Brasil*. Dissertação de Mestrado - Instituto de Matemática, Estatística e Ciência da Computação - Universidade Estadual de Campinas, 1996. 63, 76, 77, 87, 89

S. Fitzpatrick and A. Scott. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82(399):875–878, 1987. 24

V. Fossaluza, J. B. Diniz, B. B. Pereira, E. C. Miguel, and C. A. B. Pereira. Sequential allocation of patients with balancing for prognostic factors. *Clinics*, 64(6):511–518, 2009. 175

D. Freedman, R. Pisani, and R. Purves. *Statistics*. W.W. Norton and Company, New York, 1978. 32

A. Gelman and J. Hill. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007. 142

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis - Second Edition*. Chapman and Hall/CRC, 2003. 57, 59

R. Z. Gold. Tests auxiliary to χ^2 tests in a markov chain. *Annals of Mathematical Statistics*, 34: 56–74, 1963. 22

L. A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2):247–254, 1965. 22, 23

B. I. Graubard and E. L. Korn. Inference for superpopulation parameters using sample surveys. *Statistical Science*, 17(1):73–96, 2002. 54

F. A. Graybill. *Theory and Application of the Linear Model*. Duxbury Press, 1976. 183

R. M. Groves. *Survey Errors and Survey Costs*. John Wiley and Sons, Inc., 1989. 45, 48, 49, 67

J. B. S. Haldane. On a method of estimating frequencies. *Biometrika*, 33(3):222–225, 1945. 42

M. H. Hansen and W. N. Hurwitz. On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14:333–362, 1943. 43

M. H. Hansen, W. G. Madow, and B. J. Tepping. An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78(384):776–793, 1983. 54

J. Hájek. Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungary Acad. Sci.*, 5:361–374, 1960. 15, 168

J. R. Hochstim and D. M. K. Smith. Area sampling or quota control? three sampling experiments. *The Public Opinion Quarterly*, 12(1):73–80, 1948. 76

D. Holt and T. M. F. Smith. Post stratification. *Journal of the Royal Statistical Society - Series A*, 142(1):33–46, 1979. 35

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47:663–685, 1952. 43

D. Huff. *How to lie with statistics*. W. W. Norton, 1954. 27

D. Izrael, D. C. Hoaglin, and M. P. Battaglia. *A SAS Macro for Balancing a Weighted Sample*. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, Paper 275, 2000. 35

R. J. Jessen. *Statistical Survey techniques*. Wiley, New York, 1978. 6, 62

J. B. Kadane and T. Seidenfeld. Randomization in a bayesian perspective. *Journal of Statistical Planning and Inference*, 25:329–345, 1990. 61

B. F. King. Surveys combining probability and quota methods of sampling. *Journal of the American Statistical Association*, 80(392):890–896, 1985. 84

L. Kish. *Survey Sampling*. John Wiley & Sons, 1965. 7, 44, 64

K. Kwong. On sample size and quick simultaneous confidence interval estimation for multinomial proportions. *The American Statistical Association*, 7(2):212–222, 1998. 24, 25

R. R. Lau. An analysis of the accuracy of 'trial heat' polls during the 1992 presidential election. *The Public Opinion Quarterly*, 58(1):2–20, 1994. 183, 184, 185

E. E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, 1998. 133

J. T. Lessler and W. D. Kalsbeek. *Nonsampling Error in Surveys*. John Wiley and Sons, Inc., 1992. 45

D. V. Lindley and M. R. Novick. The role of exchangeability in inference. *The Annals of Statistics*, 9:45–58, 1981. 32

R. J. A. Little. Models for nonreponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250, 1982. 51, 158

T. A. Louis. Finding observed information using the em algorithm. *Journal of the Royal Statistical Society. Series B*, 44(1):98–130, 1982. 138, 139

P. Lynn and R. Jowell. How might opinion polls be improved? the case for probability sampling. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(1):21–28, 1996. 76

W. G. Madow and L. G. Madow. On the theory of systematic sampling. *Annals of Mathematical Statistics*, 15:1–24, 1944. 39

P. C. Mahalanobis. On the generalized distance in statistics. *Procedings of the National Institute of Science of India*, 2(1):49–55, 1936. 174

N. C. Meier and C. J. Burke. Laboratory tests of sampling techniques. *The Public Opinion Quarterly*, 11(4):586–593, 1947. 76

I. G. S. F. Mendonça and H. S. Migon. Pesquisa eleitoral - uma análise bayesiana. *Revista Brasileira de Estatística*, 48(189):25–34, 1987. 62

R. G. Miller. *Simultaneous Statistical Inference - Second Edition*. Springer-Verlag, 1980. 21

W. J. Mitofsky. Review: Was 1996 a worse year for polls than 1948? *The Public Opinion Quarterly*, 62(2):230–249, 1998. 171, 181

C. A. Moser. Quota sampling. *Journal of the Royal Statistical Society. Series A (General)*, 115(3): 411–423, 1952. 75

C. A. Moser and A. Stuart. An experimental study of quota sampling. *Journal of the Royal Statistical Society. Series A (General)*, 116(4):349–405, 1953. 77, 79

F. A. S. Moura. *Estimação em pequenos domínios - 18º sinape*. ABE - Associação Brasileira de Estatística, 2008. 55, 158

W. A. Nascimento. *Amostragem de Conglomerados*. Centro de Serviços Gráficos - Instituto Brasileiro de Geografia e Estatística (IBGE), 1981. 40

D. G. C. Pessoa and P. L. N. Silva. *Análise de Dados Amostrais Complexos - 13º sinape*. ABE - Associação Brasileira de Estatística, 1998. 38, 99

D. R. Plane and K. R. Gordon. A simple proof of the nonapplicability of the central limit theorem to finite populations. *The American Statistician*, 36(3):175–176, 1982. 15

A. Politz and W. Simmons. An attempt to get the 'not at homes' into the sample without callbacks. *Journal of the American Statistical Association*, 44(245):9–16, 1949a. 52, 72

A. Politz and W. Simmons. Further theoretical considerations regarding the plan for eliminating callbacks. *Journal of the American Statistical Association*, 44(245):17–31, 1949b. 72

C. P. Quesenberry and D. C. Hurst. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6(2):191–195, 1964. 23

R. E. R. Ravines. *Inferência em Modelos Hierárquicos Generalizados sob Planos Amostrais Informativos*. Dissertação de Mestrado - Instituto de Matemática - Universidade Federal do Rio de Janeiro, 2003. 57

B. Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement 1. *The Annals of Mathematical Statistics*, 43(2):373–397, 1972a. 151, 168

B. Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement 2. *The Annals of Mathematical Statistics*, 43(3):748–776, 1972b. 151, 168

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. 51, 158

D. B. Rubin. The use of propensity socres in applied bayesian inference. *Bayesian Statistics 2: Procedings of the second Valencia international meeting - Elsevier Science Publishers*, pages 463–471, 1985. 62

M. Salehi and K.-C. Chang. Multiple inverse sampling in post-stratification with subpopulation sizes unknown: a solution for quota sampling. *Journal of Statistical Planning and Inference*, 131: 379–392, 2005. 85

A. J. Scott and G. A. F. Seber. Difference of proportions from the same survey. *The American Statistician*, 37(4):319–320, 1983. 28

P. L. N. Silva and F. A. S. Moura. Efeito da conglomeração da malha setorial do censo demográfico de 1980. *Textos para discusão - IBGE*, 1(32):001–114, 1990. 108

C. P. Sison and J. Glaz. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90(429):366–369, 1995. 25

T. M. F. Smith. On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General)*, 146(4):394–403, 1983. 84

T. M. F. Smith. Post-stratification. *The Statistician*, 40:315–323, 1991. 34

M. Sobel and M. Ebneshahrashoob. Quota sampling for multinomial via dirichilet. *Journal of Statistical Planning and Inference*, 33:157–164, 1992. 86

J. Souza. *Pesquisa Eleitoral: Críticas e Técnicas*. Editora do Senado - Brasília - DF, 1990. 76, 166

C. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer Series in Statistics, 1992. 7, 44, 45, 51, 52, 54, 103, 106, 147

F. F. Stephan and P. J. McCarthy. *Sampling Opinions - An Analysis of Survey Procedure*. John Wiley and Sons, Inc., 1958. 69, 77, 79, 84

C. B. Stephenson. Probability sampling with quotas: An experiment. *Public Opinion Quarterly*, 43(4):477–496, 1979. 78, 82, 107

M. M. Strand. Estimation of a population total under a 'bernoulli sampling' procedure. *The American Statistician*, 33(2):81–84, 1979. 147

S. Sudman. *Reducing the Cost of Surveys*. ALDINE Publishing Company, 1967. 64, 70, 74, 75, 79, 83, 107

R. A. Sugden. A bayesian view of ignorable designs in survey sampling inference. *Bayesian Statistics 2: Procedings of the second Valencia international meeting - Elsevier Science Publishers*, pages 751–754, 1985. 58

R. A. Sugden and T. M. F. Smith. Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3):495–506, 1984. 56, 158

M. A. Tanner. *Tools for Statistical Inference*. Springer, 1996. 138

R. L. Thomasson and C. H. Kapadia. On estimating the parameter of a truncated geometric distribution by the method of moments. *Annals of the Institute of Statistical Mathematics*, 27 (1):269–272, 1975. 123, 141

S. K. Thompson. Sample size for estimating multinomial proportions. *The American Statistician*, 41(1):42–46, 1987. 179

C. Y. Wada and D. F. Andrade. *Tamanho da amostra em ensaios clínicos e bioequivalência - 19º sinape*. ABE - Associação Brasileira de Estatística, 2010. 27

D. B. Wajsbrot. *Aleatorização e Ética Médica*. Dissertação de Mestrado - Instituto de Matemática e Estatística - Universidade de São Paulo, 1997. 62

K. M. Wolter. *Introduction to Variance Estimation*. Springer-Verlag, 1985. 39, 44, 99, 148

D. H. Young. Quota fulfilmet using unristricted random sampling. *Biometrika*, 48(3):333–342, 1961. 86

F. J. Zabala. *Desempate técnico*. Dissertação de Mestrado - Instituto de Matemática e Estatística - Universidade de São Paulo, 2009. 28, 100