Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Analisando a Eficiência da Amostragem Dupla em Indicadores da Indústria Brasileira

Tomás Moura da Veiga 09/16374

Brasília

2013

Tomás Moura da Veiga **09/16374**

Analisando a Eficiência da Amostragem Dupla em Indicadores da Indústria Brasileira

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

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2013

Além da família e amigos, dedico à quem deseja o meu sucesso.

Tomás Moura da Veiga

Agradecimentos

Agradeço a todos que me ajudaram.

Aos meus pais que sempre me deram apoio, confiança e segurança.

Aos amigos e outros que, me ajudaram, ouvindo minhas reclamações e meus receios.

Ao professor Alan que teve paciência e que além de ser um excelente estatístico é, também, um excelente professor e orientador.

Resumo

Os indicadores da indústria brasileira, obtidos a partir das pesquisas RAIS, PIA e PINTEC, possuem a função de monitorar a atividade na indústria de transformação e, além disso, são instrumentos importantes para a análise da evolução de curto prazo da indústria brasileira. Tais pesquisas são realizadas por técnicas de amostragem, visto que o universo das empresas do setor industrial brasileiro é muito grande. A fim de capturar toda a heterogeneidade da área industrial, são necessárias técnicas mais robustas do que a amostragem aleatória simples, como a amostragem estratificada e estimadores tipo razão e regressão. No entanto, tais técnicas só podem ser aplicadas mediante o conhecimento de certas características populacionais. A amostragem dupla torna possível a utilização destas técnicas, uma vez que toma uma grande amostra com parte do recurso para estimação desses parâmetros, e faz as estimativas na amostra resultante do recurso ainda disponível.

Este trabalho apresenta um estudo sobre a eficiência da técnica de amostragem dupla, que consiste em utilizar uma pré-amostra para a obtenção dos parâmetros populacionais. Os dados deste trabalho foram criados a partir das estatísticas descritivas dos indicadores da indústria, sendo estes utilizados para a comparação das

curvas de variâncias da amostragem dupla e da amostragem aleatória simples.

Os resultados mostraram que a amostragem dupla não se mostrou eficiente para nenhuma das três técnicas utilizadas para os indicadores da indústria brasileira quando comparada com a amostragem aleatória simples.

Palavras-Chaves

Indicadores, Eficiência, Amostragem Dupla

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

Introdução

Examinando os meios de comunicação social é fácil concluir que um país desenvolvido é um país que tem como principal característica agregar valor as *commodities* e não apenas exportá-las. Aparece com grande destaque nas mídias a preocupação do governo em saber como está o desenvolvimento do seu país. Uma importante ferramenta para analisar isso são as pesquisas sobre as empresas, já que elas possuem a função de agregar valor a essas *commodities*. Uma dessas pesquisas é a Pesquisa de Indicadores Industriais feita mensalmente pela CNI (Confederação Nacional da Indústria) para monitorar a atividade na indústria de transformação. Além disso, esses indicadores são instrumentos importantes para a análise da evolução de curto prazo da indústria brasileira.

Essa pesquisa é realizada em parceria com 12 federações da indústria, sendo elaborados, em termos nacionais, indicadores de Faturamento, Horas Trabalhadas na Produção, Emprego, Massa Salarial, Rendimento Médio e Utilização da Capacidade Instalada (CNI, 2011). Para a criação desses indicadores é utilizada a PIA (Pesquisa Industrial Anual) feita pelo IBGE que reúne informações econômico-financeiras sobre o setor industrial brasileiro, abrangendo, entre outros aspectos, dados sobre Pessoal

Ocupado, Salários, Retiradas e Outras Remunerações, Receitas, Custos e Despesas, Consumo Intermediário, Valor da Produção e da Transformação Industrial referentes às empresas de extração mineral e transformação, segundo as categorias de atividades definidas na CNAE (Classificação Nacional de Atividades Econômicas) ou por detalhamento geográfico (IBGE, 2001).

A técnica de AD (amostragem dupla), que será o foco desse trabalho, consiste em trazer o máximo das técnicas de amostragem, a AE (amostragem estratificada), a AR (amostragem tipo razão) e a AReg (amostragem tipo regressão), visto que elas dependem de informações populacionais. A AE divide a população em subpopulações e essas subpopulações são chamadas de estratos. Nesses estratos é obtido uma subpopulação mais homogênea, o que faz com que a variância dessa técnica seja menor do que a variância da AAS (amostragem aleatória simples). A AR ou AReg é uma técnica de amostragem que utiliza uma covariável, geralmente correlacionada para estimar a média ou o total populacional.

Esse estudo tem como objetivo analisar a eficiência da amostragem dupla nos indicadores da indústria brasileira, além de comparar essa técnica com a amostragem aleatória simples. Para isso serão utilizadas informações da PINTEC (Pesquisa Industrial de Inovações Tecnológicas), realizada pelo IBGE, a PIA (Pesquisa Industrial Anual), do IBGE, e RAIS (Relação Anual de Indicadores Sociais), do MTE (Ministério do Trabalho e Emprego).

1.1 Objetivos

O objetivo geral do trabalho é medir a eficiência da Amostragem Dupla em Indicadores da Indústria Brasileira.

Os objetivos específicos são:

- medir até quando a Amostragem Dupla é melhor do que a Amostragem Aleatória Simples;
- observar até quanto se compensa financeiramente o uso da técnica de amostragem dupla.

Capítulo 2

Pesquisa Sobre a Indústria

2.1 Introdução

O IBGE (Instituto Brasileiro de Geografia e Estatística) é responsável por fazer pesquisas na área de estatística cujo objetivo principal é medir o crescimento e desenvolvimento do Brasil. Nesse trabalho serão utilizadas duas pesquisas feitas pelo IBGE para a obtenção dos dados: PINTEC (Pesquisa de Inovação Tecnológica) e a PIA (Pesquisa Industrial Anual), além da RAIS (Relação Anual de Informações Sociais), realizada pelo MTE. A PIA e a PINTEC são feitas com base no CEMPRE (Cadastro Central de Empresas) e adota a CNAE 2.0 como referência, além de ser atualizado anualmente com base na RAIS.

2.2 CNAE

Com o propósito de classificar as empresas nacionais pela sua principal atividade econômica, foi criada a CNAE (Classificação Nacional de Atividades Econômicas). Com essa classificação é possível agrupar as empresas com a mesma atividade em setores, sendo que esses setores possuem códigos de identificação.

Portanto a CNAE é um tipo de padronização que melhora a qualidade dos sis-

temas de informação que dão suporte às decisões e ações do Estado. Além de proporcionar uma forma para comparar as diversas pesquisas estatísticas existentes feitas pelas mais diferentes entidades de pesquisa, ampliando, assim, a análise dessas pesquisas, uma vez que é possível complementá-las.

A CNAE original foi criada em 1994 junto com o CONCLA (Comissão Nacional de Classificação). Essa comissão, composta por representantes de quinze diferentes ministérios e do IBGE, possui, justamente, o objetivo de estabelecer normas e padronizar as classificações, colocando, assim, cada empresa no seu devido setor (IBGE, 2007). Os setores onde as empresas são agrupadas possuem cinco níveis, sendo esses uma divisão hierárquica: seção, divisão, grupo, classe e subclasse (CONCLA, 2007).

Nesse trabalho será utilizada a CNAE 2.0 que entrou em vigor no ano de 2007 e é a terceira e mais nova versão da CNAE. A PIA e a PINTEC são pesquisas do IBGE feitas de acordo com a CNAE 2.0 e possuem toda a sua população-alvo nas seguintes divisões da CNAE 2.0: seções B e C, nas divisões 61, 62 e 72, no grupo 63.1 e na combinação de divisão e grupo 58+59.

A Tabela 2.1 apresenta a classificação das seções da CNAE 2.0 (CONCLA, 2007).

Tabela 2.1: Classificação da CNAE 2.0

Seção	Divisões	Descrição CNAE
A	01 .. 03	AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQÜICULTURA
B	05 .. 09	INDÚSTRIAS EXTRATIVAS
C	10 .. 33	INDÚSTRIAS DE TRANSFORMAÇÃO
D	35 .. 35	ELETRICIDADE E GÁS
E	36 .. 39	ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO
F	41 .. 43	CONSTRUÇÃO
G	45 .. 47	COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS
H	49 .. 53	TRANSPORTE, ARMAZENAGEM E CORREIO
I	55 .. 56	ALOJAMENTO E ALIMENTAÇÃO
J	58 .. 63	INFORMAÇÃO E COMUNICAÇÃO
K	64 .. 66	ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS
L	68 .. 68	ATIVIDADES IMOBILIÁRIAS
M	69 .. 75	ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS
N	77 .. 82	ATIVIDADES ADMINISTRATIVAS E SERVIÇOS COMPLEMENTARES
O	84 .. 84	ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL
P	85 .. 85	EDUCAÇÃO
Q	86 .. 88	SAÚDE HUMANA E SERVIÇOS SOCIAIS
R	90 .. 93	ARTES, CULTURA, ESPORTE E RECREAÇÃO
S	94 .. 96	OUTRAS ATIVIDADES DE SERVIÇOS
T	97 .. 97	SERVIÇOS DOMÉSTICOS
U	99 .. 99	ORGANISMOS INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS

2.3 PINTEC

Cada vez mais a estatística está sendo aplicado nas tomadas de decisões, o que faz reduzir as incertezas e complexidade das informações. Com isso e com o avanço da tecnologia, tem-se que a criação de um sistema de informações sobre as atividades de inovação tecnológica das empresas do Brasil tornou-se inevitável (IBGE, 2004).

A PINTEC foi criada pelo IBGE e visa conhecer as atividades inovativas desen-

volvidas nas empresas industriais, de telecomunicações, de atividades de informática e serviços relacionados e de pesquisa e desenvolvimento, de modo a acompanhar sua evolução no tempo.

As informações da PINTEC contribuem para ampliar o entendimento do processo de inovação tecnológica nas empresas brasileiras, sendo um componente crucial para o desenvolvimento econômico do país, e de fundamental importância para o desenho, implementação e avaliação de políticas públicas voltadas para tecnologia e na definição das estratégias privadas (IBGE, 2004).

Essa pesquisa foi realizada quatro vezes: 2000, 2003, 2005 e 2008, sendo que a fase de coleta da PINTEC para o ano de 2011 está prevista para iniciar em junho de 2012. A pesquisa terá como referência o período de 2009 a 2011. A divulgação dos resultados está prevista para meados de 2013 (IBGE, 2012). Nesse trabalho será utilizada a PINTEC de 2008 que tem como um período de referência de 2006 à 2008.

A PINTEC 2008 inclui as empresas que:

- estão em situação ativa no CEMPRE (Cadastro Central de Empresas), do IBGE, que cobre as entidades com registro no CNPJ (Cadastro Nacional da Pessoa Jurídica);
- estão identificadas no CEMPRE pela CNAE 2.0 nas seções B e C, nas divisões 61, 62 e 72, no grupo 63.1 e na combinação de divisão e grupo 58+59;
- estar sediada em qualquer parte do Território Nacional; e
- ter dez ou mais pessoas ocupadas em 31 de dezembro do ano de referência do cadastro básico de seleção da pesquisa.

Para atingir o enfoque da pesquisa, inovação tecnológica brasileira, a PINTEC 2008 fez o questionário com doze variáveis, são elas:

- Características das empresas;
- Produtos e processos novos ou substancialmente aprimorados;.
- Atividades inovativas;
- Fontes de financiamento;
- Atividades internas de P&D;
- Impactos das inovações;
- Fontes de informação;
- Relações de cooperação para inovação;
- Apoio do governo;
- Patentes e outros métodos de proteção;
- Problemas e obstáculos à inovação;
- Inovações organizacionais e de marketing;

Para esse trabalho, a variável "Atividades Inovativas" será a mais importante, pois é onde está alocada a informação sobre gastos com P&D (Pesquisa e Desenvolvimento). P&D é uma área da empresa onde são criados novos produtos ou serviços, sendo esta uma variável que será utilizada neste trabalho.

2.3.1 Aspectos da Amostragem

Como a PINTEC é uma pesquisa diferente das outras, pois tem o objetivo de medir o nível de inovação das empresas, seu desenho amostral também será. A maioria das empresas não desenvolve nenhum tipo de inovação tecnológica, tornando essa idéia mais difícil de ser notada nas unidades pesquisadas.

Desse modo, há a necessidade de, antes de obter a amostra, investigar quais empresas possuem a maior chance de terem implantado inovações no ano de 2008. Com esse intuito foram criados dois indicadores (IBGE, 2004):

1. Indicadores principais, que são as empresas com mais de um contrato de tecnologia nos anos de 2006, 2007 e 2008; as empresas que fizeram pagamentos de *royalties* em três anos consecutivos (2004 a 2006); as empresas que possuíram incentivos fiscais para a P&D e inovação tecnológica; as empresas com três ou mais patentes registradas; as empresas com mais de dois registros de programas de computador; as empresas graduadas em incubadoras; e as empresas com departamento formal de P&D na PINTEC 2003 e 2005.

2. Indicadores secundários, que são todas as outras empresas que estavam em um ou mais cadastros definidos pelos indicadores principais.

Com base nesses indicadores, foi feito uma amostra com três estratos: um estrato certo e dois estratos amostrados. No estrato certo, estão presentes todas as empresas com 500 ou mais pessoas ocupadas na indústria extrativa e na indústria de transformação e com 100 ou mais pessoas ocupadas nos serviços, além das empresas

que possuíam ao menos um indicador principal de atividade tecnológica e, também, as empresas com oito ou mais indicadores secundários. Esse estrato certo é um censo feito por todas essas empresas que possuem essas características. No segundo estrato estão as empresas que possuem entre zero e sete indicadores secundários e no terceiro estrato estão as empresas que não possuíam indicadores nenhum. Esses dois últimos estratos citados são amostrados e possuem um peso para cada empresa na pesquisa, cujo peso das empresas que estão no segundo estrato é maior que o peso das que estão no terceiro estrato, ordenados pela chance de possuírem inovações.

Tendo essa amostra definida, a pesquisa é feita através de entrevistas presenciais e por telefone com todas as empresas, já que o termo inovação tecnológica, possui diferentes formas de interpretação e poderia não haver uniformidade com as respostas dadas pelas diferentes empresas (IBGE, 2004).

2.4 PIA

A indústria brasileira precisa de constante análise e cuidado. Para isso foi criada a PIA (Pesquisa Industrial Anual) que tem justamente como ideia básica monitorar e levantar informações para haver dados que permita um amplo estudo sobre essa indústria.

A série da PIA começa em 1966, quando foi realizada a primeira pesquisa, e é realizada todo ano com exceção daqueles em que acontece o censo e com exceção dos anos de 1971 e 1991. A partir do ano de 1996 a PIA passou a ser feita, também, nos anos do censo (IBGE, 2009).

Com isso, essa pesquisa tem como objetivo apontar as principais características

da indústria brasileiras, assim como observar as transformações ocorridas no tempo dessas empresas que compõem essa indústria através dos levantamentos anuais. Além disso, essa pesquisa é de suma importância para pesquisas econômicas como a pesquisa dos Indicadores Industriais feita pela CNI que traça o perfil da indústria.

A unidade que é investigada serão as empresas, sendo que essa unidade de investigação para participar da seleção da amostra precisa estar dentro dos seguintes requisitos (IBGE, 2009):

- estar em situação ativa no CEMPRE (Cadastro Central de Empresas), do IBGE, que cobre as entidades com registro no CNPJ (Cadastro Nacional da Pessoa Jurídica);
- estar identificadas no CEMPRE pela CNAE 2.0 nas seções B e C, nas divisões 61, 62 e 72, no grupo 63.1 e na combinação de divisão e grupo 58+59;
- estar sediada em qualquer parte do Território Nacional; e
- ter dez ou mais pessoas ocupadas em 31 de dezembro do ano de referência do cadastro básico de seleção da pesquisa.

Para esse trabalho será utilizada a PIA referente ao ano de 2009, já que, até o momento, esse é o ano mais próximo dos dias de hoje em que o resultado da PIA foi divulgado.

As variáveis pesquisadas na PIA são (IBGE, 2009):

- Pessoal ocupado;

- Salários, retiradas e outras remunerações;
- Receita líquida de vendas;
- Demais receitas;
- Custos e despesas;
- Aquisições, melhorias e baixas de ativos tangíveis realizadas no ano;
- Variáveis derivadas das variáveis investigadas na empresa (inclusive valor da transformação industrial); e
- Variáveis investigadas e derivadas na unidade local.

Para esse trabalho serão utilizadas da PIA as variáveis VTI (Valor da Transformação Industrial), RLV (Receita Líquida de Vendas) e SM (Salário Médio).

2.4.1 Aspectos da Amostragem

A PIA tem como objetivo reportar o andamento da indústria brasileira, portanto a população-alvo da amostra retirada dessa pesquisa é justamente a indústria, sendo cada empresa a unidade de seleção. A amostra é obtida por amostragem estratificada sendo que, primeiramente, essa amostra é dividida em dois estratos: o estrato natural e o final. A partir do estrato final tem-se outros dois estratos, onde é realmente feita a divisão dessa amostra: o estrato certo e o estrato amostrado.

São dois estratos naturais, um estrato onde tem-se todas as empresas com uma a quatro pessoas ocupadas alocadas nas CNAEs, portanto é feito um censo dessas empresas, e o outro estrato natural são todas as classificações de atividade onde

todas as empresas do estado são alocadas, como já indica o nome, é um estrato que existe naturalmente.

Já o estrato final, quando realmente a amostra é estratificada, é definido pelo estrato certo e o estrato amostrado. O estrato certo é um censo de todas as empresas com trinta ou mais pessoas ocupadas ou são as empresas que possuíram uma renda bruta no ano de 2008 igual ou superior a oito milhões e oitocentos mil reais. Já para o estrato final amostrado, é retirada uma amostra feita por amostragem aleatória simples sem reposição para cada estrato, sendo que, as empresas que participam dessa seleção, possuíam trinta ou menos pessoas ocupadas, menos as empresas com uma a quatro pessoas ocupadas onde também é feito um censo dessas empresas (IBGE, 2009).

2.5 RAIS

O MTE (Ministério do Trabalho e Emprego) realiza, todo ano, a RAIS, que é um Registro Administrativo cuja finalidade é controlar as atividades trabalhistas do Brasil, sendo, ainda, um importante meio para a obtenção de informações sobre o mercado de trabalho, fornecendo dados para a elaboração de pesquisas estatísticas. Além de pesquisas estatísticas, esses dados são indispensáveis para, por exemplo, criação de políticas públicas de combate às desigualdades de emprego e renda, caracterização do mercado de trabalho formal, auxiliar a tomada de decisões dos mais diversos segmentos da sociedade, e, atualmente sua principal função, a identificação do trabalhador que possui o direito de utilizar o abono salarial (MTE, 2012).

As principais variáveis dessa pesquisa são:

- empregos em 31 de dezembro segundo gênero;
- faixa etária;
- grau de escolaridade;
- tempo de serviço e rendimentos; e
- desagregados em nível ocupacional, geográfico e setorial.

Em complemento das variáveis, esse Registro Administrativo também contêm informações sobre o número de empregos por tamanho de estabelecimento, massa salarial e nacionalidade do empregado (MTE, 2011).

Nesse trabalho, a RAIS será utilizada para a obtenção do número de PO (Pessoal Ocupado) e do número de ENG (engenheiros de cada estabelecimento), tendo em vista que o último ano que tem-se os resultados da RAIS é em 2009, esse ano será o escolhido para a obtenção dessa variável.

2.5.1 Aspectos da Amostragem

A RAIS é de declaração obrigatória e destinada a todos os estabelecimentos do País, mesmo aqueles que não possuem vínculo empregatício, tornando-se, assim, uma pesquisa censitária. Estima-se que cerca de 97% do universo do mercado formal brasileiro possui cobertura por essa pesquisa.

Nesse trabalho, será utilizada a RAIS 2009 que teve 7,4 milhões de estabelecimentos declarantes, sendo que cerca de 4,2 milhões desses não tiveram nenhum empregado no ano de 2009 (MTE, 2011).

Capítulo 3

Amostragem Dupla

3.1 Introdução

A técnica de AD (amostragem dupla) consiste em trazer o máximo das técnicas de: AE (amostragem estratificada), AR (amostragem tipo razão) e a AReg (amostragem tipo regressão).

A AE divide a população em classes, sendo que essas são chamadas de estratos. Nesses estratos são alocados indivíduos mais parecidos, fazendo com o que esses estratos sejam homogêneos diminuindo, assim, a variância dessa técnica quando comparada com a AAS (amostragem aleatória simples).

A AR ou AReg são técnicas de amostragem que utilizam uma covariável, ou variável auxiliar, geralmente correlacionada, para estimar a média ou o total populacional. Sendo que a diferença entre elas é que, com a AR, quando a variável auxiliar é zero, a variável de interesse também é zero, enquanto na AReg essa propriedade não é verdadeira.

Para estimar a variável desejada usando as técnicas de AR ou AReg é necessário que se conheça o verdadeiro total ou a média dessa variável, algo que não acontece em inúmeros casos, já que, nesses casos, não se sabe os valores da população. Surge,

assim, a técnica de amostragem dupla.

Essa técnica possui duas fases. A primeira fase consiste em estimar a média da covariável, usando uma amostra preliminar de tamanho grande, que, em alguns casos, possui um custo pequeno. O objetivo dessa grande amostra é possuir uma boa estimativa da média ou a distribuição de frequência da variável auxiliar (Cochran, 1977). Em seguida, será retirada uma amostra dessa amostra preliminar para estimar a variável desejada, essa segunda amostra é consideravelmente menor do que a amostra preliminar.

Portanto, essa técnica, apresentada pela primeira vez por Neyman em 1938, representa um proveito apenas se o ganho de precisão da estratificação ou dos estimadores tipo razão e regressão for maior do que a perda de precisão devido a necessidade da redução do tamanho da amostra principal, já que possuir duas grandes amostras possui um custo muito alto (Cochran, 1977).

3.2 Amostragem Dupla para Estratificação

A amostragem estratificada consiste em, primeiramente, dividir a população de tamanho N em classes (estratos) de $N_1, N_2, ..., N_L$. Todas essas classes juntas são a população total, portanto $N_1 + N_2 + ... + N_L = N$. Em seguida, retira-se amostras desses estratos que são chamados, respectivamente, de $n_1, n_2, ..., n_L$.

Essa técnica é utilizada para melhorar, principalmente, a precisão das estimativas, garantindo que os mais diferentes elementos da população estejam presentes na amostra e produzir estimativas para a população e os estratos.

A melhora das estimativas acontece já que uma população heterogênea é dividida

em estratos mais homogêneos possíveis. Sendo as unidades de cada estrato com pouca variação, já que são homogêneos, é possível obter uma estimativa precisa da média de cada estrato, por exemplo, com uma amostra menor. Possuindo cada média dos estratos, pode-se fazer uma média global com uma boa precisão e sendo a amostra menor (Cochran, 1977).

Para a utilização da técnica de amostragem dupla a população será estratificada em h estratos. Primeiramente, será retirada uma grande amostra utilizando amostragem aleatória simples, que será chamada de n' e tendo como objetivo estimar os pesos dos estratos. Portanto tem-se que a proporção da população que está no estrato é:

$$W_h = \frac{N_h}{N} \tag{3.1}$$

e a proporção da primeira amostra que está no estrato h é:

$$w_h = \frac{n'_h}{n'} \tag{3.2}$$

Conforme Cochran (1977) w_h é uma estimativa não viesada para W_h e será utilizada nos cálculos da amostragem estratificada.

Após a primeira amostra retirada, será feita uma AE, de tamanho n onde n_h unidades são selecionadas do estrato h e y_{hi} é medido onde i é o valor da i-ésima unidade. O objetivo dessa amostra é estimar a média dos estratos ($\bar{Y}_h$).

Portanto, tem-se que a média da população é:

$$\bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_h \tag{3.3}$$

e sua estimativa é dada por:

$$\bar{y}_{st} = \sum_{h=1}^{H} w_h \bar{y}_h \tag{3.4}$$

Calculando a esperança de $\bar{y}_{st}$, observa-se que esse estimador é não viesado, portanto sua esperança é igual a $\bar{Y}$. A fim de minimizar a variância desse estimador para um dado custo, assume-se que n'_h são sub-amostras aleatórias de n_h, sendo, assim, $n_h = v_h n'_h$, onde $0 < v_h \leq 1$ e os v_h são fixos, escolhidos antecipadamente. A variância de $\bar{y}_{st}$ é dada por:

$$Var(\bar{y}_{st}) = S^2 \left(\frac{1}{n'} - \frac{1}{N} \right) + \sum_{h=1}^{H} \frac{W_h S_h^2}{n'} \left(\frac{1}{v_h} - 1 \right) \tag{3.5}$$

Se uma proporção estiver sendo estimada na segunda amostra, tem-se que:

$$S^2 = \frac{NP(1-P)}{N-1} \tag{3.6}$$

e

$$S_h^2 = \frac{N_h P_h (1-P_h)}{N_h - 1} \tag{3.7}$$

onde P é a proporção estimada e $(\bar{Y}_h - \bar{Y})^2 = (P_h - P)^2$ (Cochran, 1977).

A seguir tem-se a variância de P_{st}, a variância estimada de $\bar{y}_{st}$ e a variância estima de p_{st}:

$$Var(p_{st}) = \sum_{h=1}^{H} \frac{W_h P_h (1-P_h)}{n' v_h} + \frac{N-n'}{n'(N-1)} \sum_{h=1}^{H} W_h (P_h - P_{st})^2 \tag{3.8}$$

$$\hat{Var}(\bar{y}_{st}) = \sum_{h=1}^{H} w_h s_h^2 \left(\frac{1}{n' v_h} - \frac{1}{N} \right) + \frac{g'}{n'} \sum_{h=1}^{H} w_h (\bar{y}_h - \bar{y}_{st})^2 \tag{3.9}$$

sendo $g' = \dfrac{N-n'}{N-1}$

$$\hat{Var}(p_{st}) = \sum_{h=1}^{H} \frac{w_h n_h p_h(1-p_h)}{n_h - 1}\left(\frac{1}{n' v_h} - \frac{1}{N}\right) + \frac{N-n'}{n'(N-1)} \sum_{h=1}^{H} w_h (p_h - p_s t)^2 \quad (3.10)$$

Uma forma de encontrar v_h que minimize a variância de 3.5 é dado por (Cochran, 1977)

$$v_h = v = \left[\frac{c'}{c}\frac{S_w^2}{(S^2 - S_w^2)}\right]^{\frac{1}{2}} = \left[\frac{c'}{c}\frac{1}{(\phi - 1)}\right]^{\frac{1}{2}} \quad (3.11)$$

Para comparar o resultado com a AE clássica, foram calculadas as estimativas e variâncias desta técnica. No caso, a estimativa da média é dado conforme (3.4) e a variância é dado por:

$$Var(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}(1 - f_h) \quad (3.12)$$

Quando utilizada a alocação proporcional a variância é dado por:

$$Var(\bar{y}_{st}) = \sum \frac{N_h}{N}\frac{S_h^2}{n}\left(\frac{N-n}{N}\right) = \frac{1-f}{n}\sum W_h S_h^2 \quad (3.13)$$

3.3 Amostragem Dupla para Estimador tipo Razão

No método do estimador tipo razão tem-se uma variável X correlacionada com Y, sendo X obtido para cada unidade da amostra. Essa variável correlacionada é utilizada para possuir uma estimativa melhor dos parâmetros de média e do total, fazendo com que a alta correlação entre X e Y seja vantajosa.

Portanto tem-se Y_i como sendo a variável de interesse, onde $i = 1, 2, ..., N$, sendo que deseja-se estimar seu total $\left(T_y = \sum_{i=1}^{N} Yi\right)$. Um estimador natural seria $\hat{T}_y = N\bar{y}$, porém, como não se conhece N é necessário encontrar outra solução (Cochran, 1977).

A solução do estimador tipo razão é achar uma variável X_i fortemente correlacionada com Y_i, para definir uma razão do tipo:

$$R = \frac{T_y}{T_x} = \frac{\mu_y}{\mu_x} \tag{3.14}$$

Portanto,

$$T_y = \frac{\mu_y}{\mu_x} T_x \tag{3.15}$$

onde T_x é o peso total da variável que é altamente correlacionada, logo $T_x = \sum_{i=1}^{N} x_i$.

Para encontrar uma estimativa para os parâmetros μ_y e μ_x, que são, respectivamente, as médias populacionais de Y e X, é retirada uma amostra da população de N elementos chamada de n e, tendo conhecimento do total da variável correlacionada, obtém-se o estimador tipo razão:

$$\hat{T}_{yr} = \frac{\bar{y}}{\bar{x}} T_x \tag{3.16}$$

Para o estimador tipo razão da média é utilizada a mesma idéia resultando em:

$$\bar{y}_r = \frac{\bar{y}}{\bar{x}} \bar{x'} \tag{3.17}$$

Sendo, no caso da técnica de amostragem dupla, $\bar{x'}$ e $\bar{x}$ as médias de x_i na primeira e na segunda amostra, respectivamente (Cochran, 1977).

Calculando a esperança de $\bar{y_r}$, conclui-se que esse é um estimador viesado porém, quando n grande, é, aproximadamente, não viesado. Entretanto, para pequenas amostras pode-se observar um viés com um valor considerável. A variância de $\bar{y_r}$ é:

$$Var(\bar{y_r}) = \frac{S_y^2 - 2RS_{x,y} - R^2 S_x^2}{n} + \frac{2RS_{x,y} - R^2 S_x^2}{n'} - \frac{S_y^2}{N} \tag{3.18}$$

onde $s^2_{x,y}$ que é o estimador de $S^2_{x,y}$ é:

$$s^2_{x,y} = \frac{1}{n-2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 - \left(\frac{\bar{y}}{\bar{x}}\right)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2\right) \tag{3.19}$$

Uma forma de comparar a técnica de amostragem apresentada nesta seção com a AR clássica é obtendo as estimativas e as variâncias destas duas técnicas. O estimador tipo razão da média é dada pela equação 3.17 e a variância da AR clássica é:

$$Var(\bar{y}_r) = \frac{1-f}{n}\left[\sum_{i=1}^{N}\frac{(y_i - Rx_i)^2}{N-1}\right] \tag{3.20}$$

onde $f = \dfrac{n}{N}$

3.4 Amostragem Dupla para Estimador tipo Regressão

Esse estimador, como o estimador tipo razão, também tem como objetivo aumentar a precisão da estimativa utilizando uma covariável muito correlacionada com a variável resposta. Quando esses dados altamente correlacionados sugerem uma regressão linear, é possível construir uma reta de regressão.

Na amostragem dupla, tem-se que a grande amostra, de tamanho n', foi medido somente x_i. Já na segunda amostra de tamanho n, com $n < n'$, foi retirada uma amostra aleatória dessa grande amostra de tamanho vn', e então é medido x_i e y_i.

Para uma amostra de tamanho n, tem-se a seguinte reta de regressão:

$$\bar{y}_{reg} = \bar{y} + b(\bar{x'} - \bar{x}) \tag{3.21}$$

sendo b o coeficiente de regressão estimado de y_i em x_i, calculado na segunda amostra e $\bar{x'}$, $\bar{x}$ as médias, repectivamente, da primeira e da segunda amostra.

A variância estimada de $\bar{y}_{reg}$ é:

$$\hat{Var}(\bar{y}_{reg}) = s^2_{x,y}\left(\frac{1}{n} + \frac{(\bar{x'} - \bar{x})^2}{\sum (xi - \bar{x})^2}\right) + \frac{s^2_y - s^2_{x,y}}{n'} - \frac{s^2_y}{N} \tag{3.22}$$

onde

$$s^2_{x,y} = \frac{1}{n-2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \tag{3.23}$$

e

$$s^2_y = \sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{n-1} \tag{3.24}$$

Para critérios de comparação com a AReg clássica, sabe-se que a reta de regressão desta técnica também é igual a 3.21, sendo sua variância:

$$Var(\bar{y}_{lr}) = \frac{1-f}{n}(S^2_y - 2bS_{yx} + b^2S^2_x) \tag{3.25}$$

onde $f = \dfrac{n}{N}$ e b o coeficiente estimado de y_i em x_i.

O estimador tipo razão é um caso particular do estimador tipo regressão, observando o $\bar{y}_{reg}$ e substituindo b da fórmula desse estimador por $\dfrac{\bar{y}}{\bar{x}}$, obtém-se, justamente, o $\bar{y}_r$. A diferença entre eles é que com o estimador tipo razão é necessário que a variável correlacionada seja zero quando a variável de interesse for zero, propriedade que não existe no estimador tipo regressão.

Capítulo 4

Material e Métodos

4.1 Introdução

Este capítulo mostrará o material e a metodologia aplicada afim de alcançar o objetivo proposto de comparar a Amostragem Dupla com a Amostragem Aleatória Simples. Para isso serão feitos ensaios com diferentes tamanhos de amostras das empresas dos Indicadores Industriais, para averiguar qual o ponto de interseção entre a curva da variância da técnica de AD e a reta da variância da técnica da AAS.

4.2 Material

Os dados foram simulados a partir das estatísticas descritivas dos dados reais, já que não foi possível ter acesso aos microdados devido a confidencialidade das pesquisas, que fazem com que esses dados sejam sigilosos. Para isso foram calibrados modelos de regressão a partir de dados reais da RAIS para estimar as variáveis da PIA e da PINTEC. A média e o desvio padrão (dp) gerados a partir dos modelos de regressão calibrados no IBGE estão nas Tabelas 4.1 e 4.2, respectivamente.

Tabela 4.1: Médias das variáveis analisadas

CNAE 2	PO	VTI	SM	RLV	ENG	P&D
10	221,71	22.941.101,58	964,52	66.043.164,70	0,75	1.543,13
11	142,73	57.651.837,05	1.216,65	102.141.604,64	0,89	1.237,56
12	299,25	153.666.468,29	1.210,94	329.243.989,11	0,78	13.100,33
13	141,39	7.250.492,79	992,73	17.287.114,75	0,24	663,27
14	73,89	2.150.374,78	700,31	4.140.555,48	0,02	215,52
15	143,87	4.698.983,89	742,83	9.474.056,47	0,07	822,77
16	77,12	3.869.175,65	831,16	8.318.966,41	0,23	132,67
17	136,23	20.484.202,07	1.267,28	43.596.102,92	1,02	1.929,09
18	73,44	8.170.247,93	1.350,35	15.701.756,78	0,15	337,98
19	678,34	455.316.166,57	1.709,79	669.240.627,71	47,64	71.414,89
20	108,92	31.062.102,68	1.971,25	94.514.545,76	2,23	1.444,88
21	158,25	62.626.038,07	2.216,19	98.871.740,53	1,47	3.600,07
22	98,57	8.278.353,89	1.201,73	20.410.467,70	0,51	699,13
23	85,35	8.278.703,95	942,67	17.189.672,37	0,68	446,65
24	205,45	41.910.026,69	1.495,46	117.893.052,49	5,12	3.871,54
25	79,36	6.299.484,06	1.344,17	13.873.261,77	0,69	455,10
26	134,46	21.372.885,03	1.663,33	64.810.144,86	4,73	3.300,85
27	172,00	18.718.241,33	1.456,70	48.687.656,02	4,69	2.515,83
28	98,08	11.326.170,60	1.833,90	26.865.878,86	2,49	786,81
29	294,77	56.247.896,54	1.511,07	148.836.282,30	9,10	13.238,15
30	277,20	42.308.290,45	1.548,74	116.304.266,69	16,92	29.392,51
31	81,09	3.861.344,80	933,69	9.483.740,10	0,10	462,94
32	81,97	6.358.057,24	1.074,00	10.666.941,34	0,38	387,20
33	103,78	6.770.578,71	1.545,53	13.003.235,74	1,33	924,85

A título de ilustração os modelos calibrados foram:

- $VTI_{CNAE_i} = \beta_0 + \beta_1 * PO$,

  onde -951.606.247,90 $\leq \beta_0 \leq$ 26.970.901,15 e -85.826,67 $\leq \beta_1 \leq$ 2.074.081,39

- $RLV_{CNAE_i} = \beta_0 + \beta_1 * PO$,

  onde -1.149.380.419,00 $\leq \beta_0 \leq$ 46.103.930,83 e -297.465,26 $\leq \beta_1 \leq$ 2.681.006,45

- $P\&D_{CNAE_i} = \beta_0 + \beta_1 * ENG$,

  onde -3.455,21 $\leq \beta_0 \leq$ 6.763,55 e -501,29 $\leq \beta_1 \leq$ 4.431,22

- $SM_{CNAE_i} = \beta_0 + \beta_1 * PO$, onde 697,49 $\leq \beta_0 \leq$ 6.585,32 e 0,01 $\leq \beta_1 \leq$ 5,29

Tabela 4.2: Desvios Padrões das variáveis analisadas

CNAE 2	PO	VTI	SM	RLV	ENG	P&D
10	1.154,19	151.582.257,44	475,02	421.301.496,06	6,70	21.078,99
11	492,24	401.266.850,24	1.051,62	639.773.222,69	4,84	2.275,52
12	527,99	582.722.192,88	648,22	921.370.683,60	1,99	23.585,74
13	451,54	23.098.913,02	425,91	53.393.901,14	1,55	1.914,58
14	341,26	13.432.462,09	247,39	21.965.921,03	0,32	1.710,42
15	867,64	34.798.179,84	223,26	54.320.088,15	0,62	8.399,08
16	202,41	26.473.112,23	299,79	49.900.764,71	3,09	2.700,82
17	350,41	126.379.137,07	683,71	222.635.974,43	6,98	7.823,66
18	190,08	36.990.282,08	683,32	79.362.433,69	1,95	365,93
19	2.545,58	6.095.048.209,20	1.234,09	7.880.334.648,41	681,89	580.672,71
20	276,94	165.269.007,82	1.609,27	507.260.881,37	17,40	24.683,67
21	228,27	153.519.286,45	1.838,05	231.929.735,32	4,20	17.207,30
22	241,30	40.445.877,43	575,71	97.936.697,73	5,10	13.058,54
23	230,33	58.459.041,17	531,22	115.156.932,93	5,15	3.549,72
24	800,75	262.413.387,94	758,54	686.950.548,69	38,02	11.744,40
25	185,18	26.447.818,79	629,48	56.809.325,15	4,57	3.702,43
26	338,51	92.028.614,70	1.174,97	299.307.556,74	19,33	22.888,04
27	811,76	106.437.059,61	857,46	272.578.073,70	33,77	17.971,98
28	239,21	39.779.071,63	1.103,32	114.920.469,42	12,36	6.327,77
29	1.108,65	391.550.578,25	750,64	1.073.202.087,28	75,27	102.177,62
30	1.023,11	202.898.073,52	1.030,42	730.105.718,22	189,94	135.135,00
31	124,62	9.847.872,27	339,77	25.404.133,88	0,53	2.498,02
32	139,73	19.007.091,74	505,63	30.047.096,18	1,95	2.008,97
33	332,40	26.523.737,05	984,73	94.216.334,72	4,90	2.595,48
Total	850,16	660.666.634,00	437.74	9.076.277.654,00	62,26	534.687,66

Observa-se na Tabela 4.2 que existe uma grande heterogenidade dentro das CNAEs, tendo casos em que o desvio padrão da CNAE (estrato) é maior que o desvio padrão geral.

Vale ressaltar que as variáveis explicativas PO (Pessoal Ocupado) e ENG (Número de Engenheiros da Empresa) foram retiradas da RAIS e serviram para a previsão do modelo. A base de dados conta com um total de 34.834 empresas.

4.3 Métodos

Para a realização desse trabalho serão feitos quatro estudos de caso, sendo um

estudo de caso para cada uma das seguintes variáveis: RLV (Receita Líquida de Venda), P&D (gastos com Pesquisa e Desenvolvimento), SM (Salário Médio) e VTI (Valor da Transformação Industrial). Para cada estudo de caso serão retiradas duas amostras: uma amostra maior que terá pelo menos 50% do total de empresas, ou seja, pelo menos 17.418 empresas, e uma amostra menor que dependerá do valor restante para a realização da pesquisa.

Será utilizado como base para o custo total, uma pesquisa realizada pela CNI, que orçou, para cada questionário, R$2,78. Este valor é uma estimativa, já que leva em consideração apenas o papel, as etiquetas, a postagem nos correios e a repografia, não levando em consideração o valor do trabalho dos funcionários e nem dos softwares utilizados, ou seja, este representa apenas o custo variável. Para este trabalho, o custo suposto para a realização da pesquisa foi um valor total de R$ 85.000, com isso, pode-se amostrar um total de 30.576 empresas.

Como visto no Capítulo 3 a AD é aplicada para melhorar as estimativas do estimador no caso estratificado e do tipo razão e regressão. Para a amostragem estratificada serão utilizadas as variáveis RLV e SM, tendo as CNAEs como estrato. Já para a amostragem tipo razão será utilizada a variável VTI e a covariável PO, visto que se não houver trabalhadores então a empresa não produzirá nada, no caso há a necessidade de pelo menos um para operar a máquina. Por fim, a amostragem tipo regressão, a variável utilizada será P&D, tendo como covariável ENG, visto que pode haver inovação tecnológica sem a presença de engenheiros. Para cada estudo serão feitas 100 repetições e em seguida será retirada a média e o máximo e o mínimo

para a construção dos intervalos de confiança.

A fim de verificar se a AD é melhor do que a AAS serão comparados os gráficos das variâncias dessas duas técnicas de amostragem, observando, assim, o ponto em que a reta da variância da AAS e a curva da variância da AD se interceptam.

Capítulo 5

Análise dos Resultados

5.1 Introdução

Neste capítulo será aplicado e analisado a metodologia descrita anteriormente. Dessa forma, será observado qual a técnica de amostragem é mais eficiente, a AD ou a AAS.

5.2 Amostragem Estratificada

A partir das Figuras 5.1 e 5.2 que representam as estimativas das CNAEs 10 e 12 (maior e menor frequência, repectivamente), observa-se que o w estimado, ou seja, o peso de cada estrato, se aproxima do peso real (W) a medida que aumenta a amostra inicial. Pode-se observar, também, que a média das amostras está muito próxima do valor real de W.

Nestes dois exemplos, tem-se que a única reta das Figuras 5.1 e 5.2 é o peso do estrato W real e a curva variando ao entorno desta é a média amostral. Já as curvas em vermelho que estão se aproximando do W real são o mínimo e o máximo obtidos das amostras. Com isso é possível observar o resultado esperado, ou seja, o w estimado se aproxima do W real a medida em que a amostra inicial aumenta.

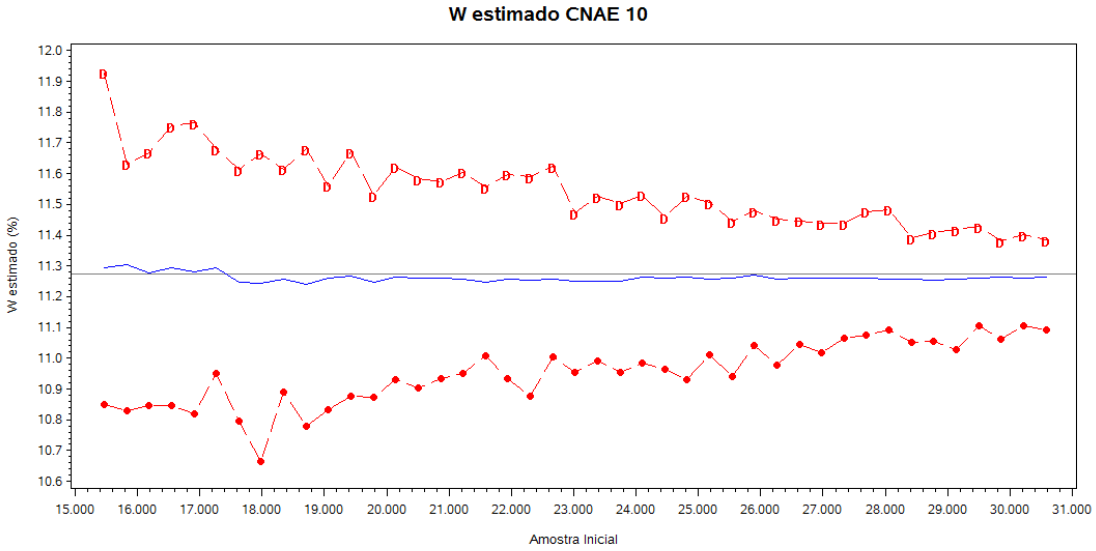

Figura 5.1: Comparação dos Ws para a CNAE 10

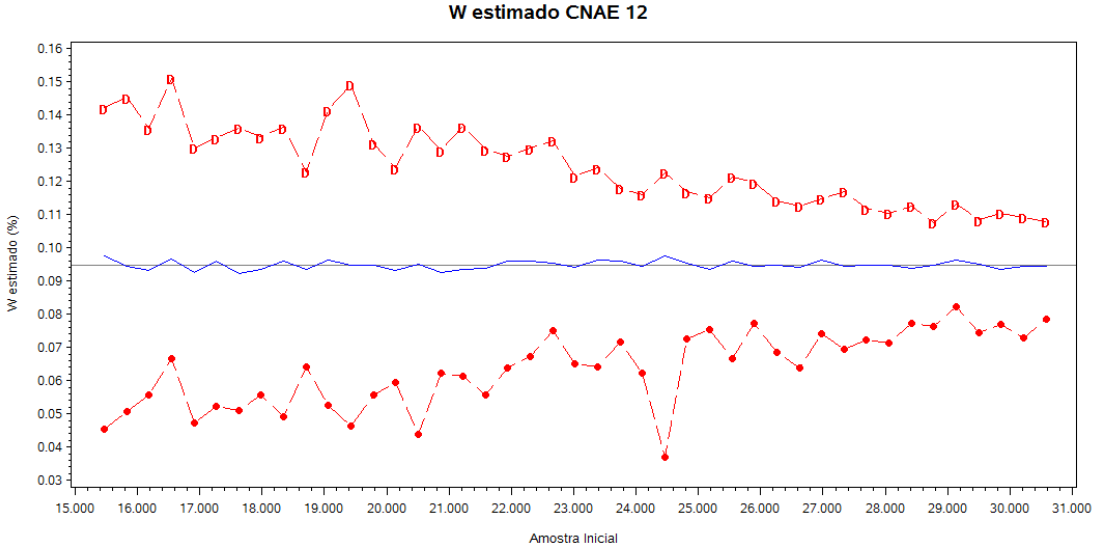

Figura 5.2: Comparação dos Ws para a CNAE 12

Ao se analisar as curvas de variâncias das variáveis RLV (Figura 5.3) e SM (Figura 5.4) é observado algumas características similares. Na variável RLV somente quando a AD utiliza até 75% do recurso para a estimação dos pesos dos estratos w a estratificação clássica com alocação ótima (3.12) produz melhores resultados do que

a AAS com todos os recursos disponíveis. Note que quando é feita a estratificação com a alocação proporcional (3.13) ou utilizada a variância da AD para estratificação (3.5) elas são sempre piores que a AAS quando utilizado todo o recurso.

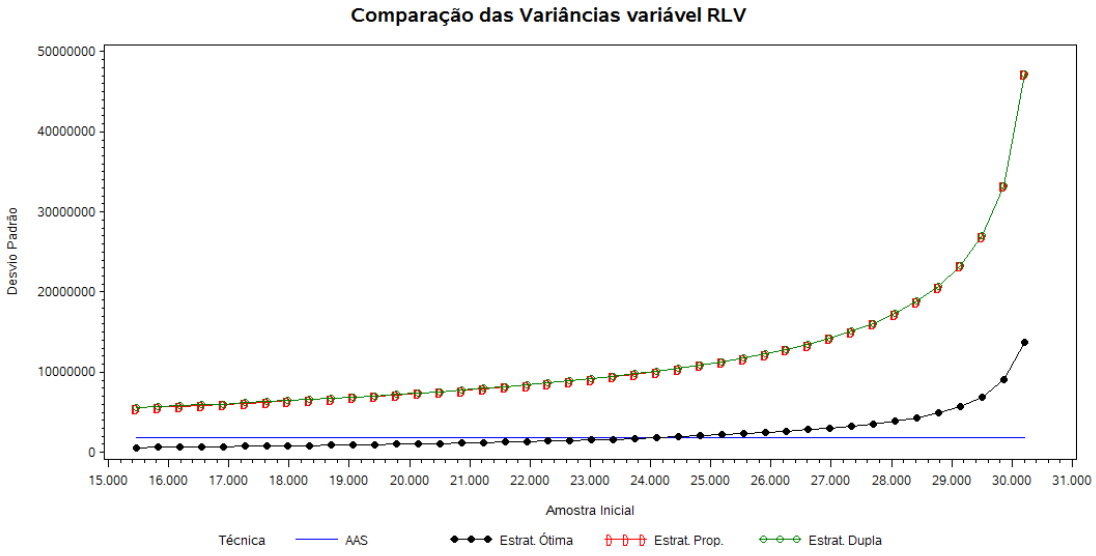

Figura 5.3: Comparação das Variâncias RLV

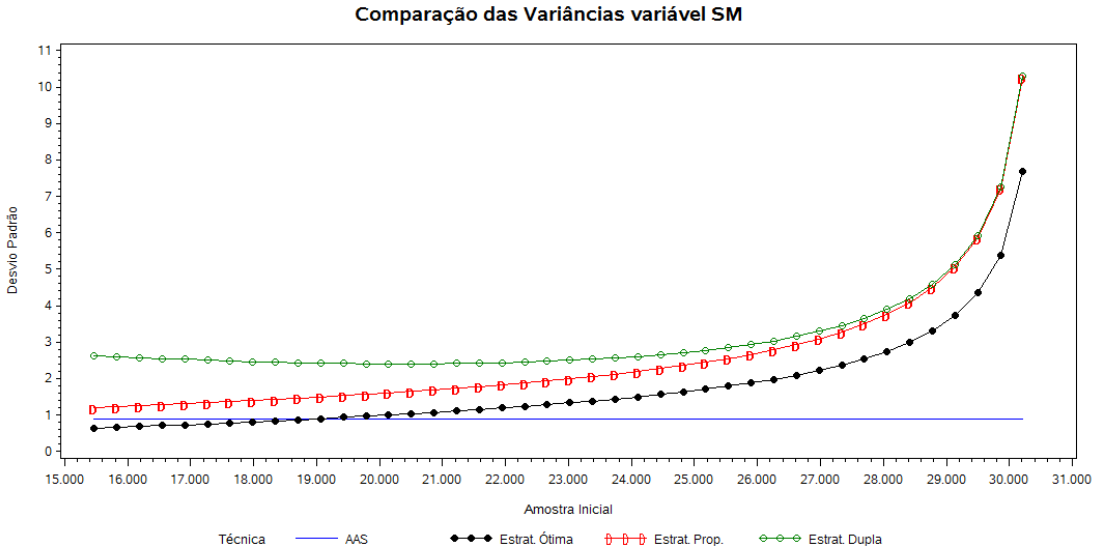

Figura 5.4: Comparação das Variâncias SM

Este comportamento também é visto com a variável SM só que com o percentual

bem menor, em torno de 60%.

Como não era esperado que a variância da AD fosse maior do que a variância da AAS foi feito um novo estudo de caso, só que agora com estratos que apresentam baixa variância, quando comparada com a variância geral, a fim de averiguar se o motivo desse resultado foi a grande heterogenidade dos estratos. Para este estudo foi utilizado a variável SM, já que seu desvio padrão geral não é tão grande quando comparado com os desvios padrões de dentro das CNAEs (Tabela 4.2) e quando comparado, também, com as outras variáveis. Foi pressuposto que o motivo da variância da AD ser maior do que a variância da AAS é a quantidade de estratos e as altas variâncias dentro destes estratos, sendo assim, foi feito um estudo somente com as 6 CNAEs que possuem uma variância menor (CNAEs 10, 13, 14, 15, 16, 31).

Para este estudo de caso foi presumido que o custo total da pesquisa seria de R$36.000 com um valor para cada questionário também de R$2,78. Com isso, a amostra inicial, que deve ser pelo menos 50% maior do que a amostra principal, começa com 6.835 empresas e a amostra principal com 6.116. Quando comparada a variância da AD com a AAS, mostrado na Figura 5.5, é visto que as curvas de variâncias possuem um comportamento parecido com os estudos de casos anteriores, dessa forma, conclui-se que somente quando a AD utiliza até 68% do recurso para a estimação dos pesos dos estratos w a estratificação clássica com alocação ótima produz melhores resultados do que a AAS com todos os recursos disponíveis.

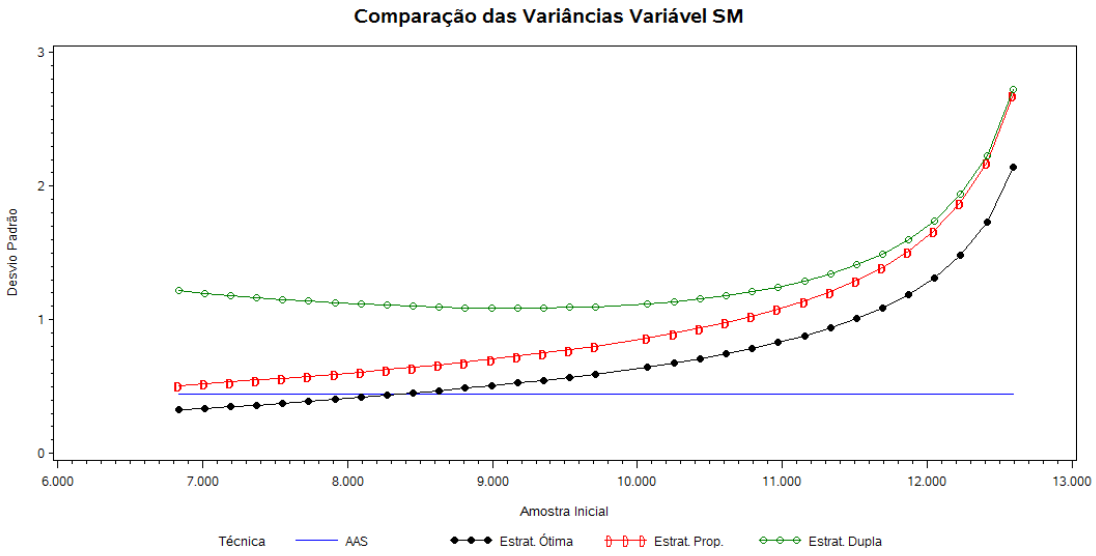

Figura 5.5: Comparação das Variânicas SM - Apenas 6 CNAEs

5.3 Estimador Tipo Razão

Para analisar a eficiência do estimador tipo razão foi utilizada a variável VTI. A Figura 5.6 compara a estimativa feita para a variável em análise junto com a sua média real.

As curvas em vermelho são os mínimos e os máximos das 100 repetições de cada amostra, sendo possível observar que quanto maior a amostra inicial, mais essas duas curvas se aproximam da média populacional. A estimativa da média está representada pela curva em azul que está ao redor da média populacional.

As curvas de variâncias, como é possível observar na Figura 5.7, mostram que a técnica de AD para o estimador tipo razão não se mostra vantajosa se comparada com a AAS para os indicadores da indústria brasileira.

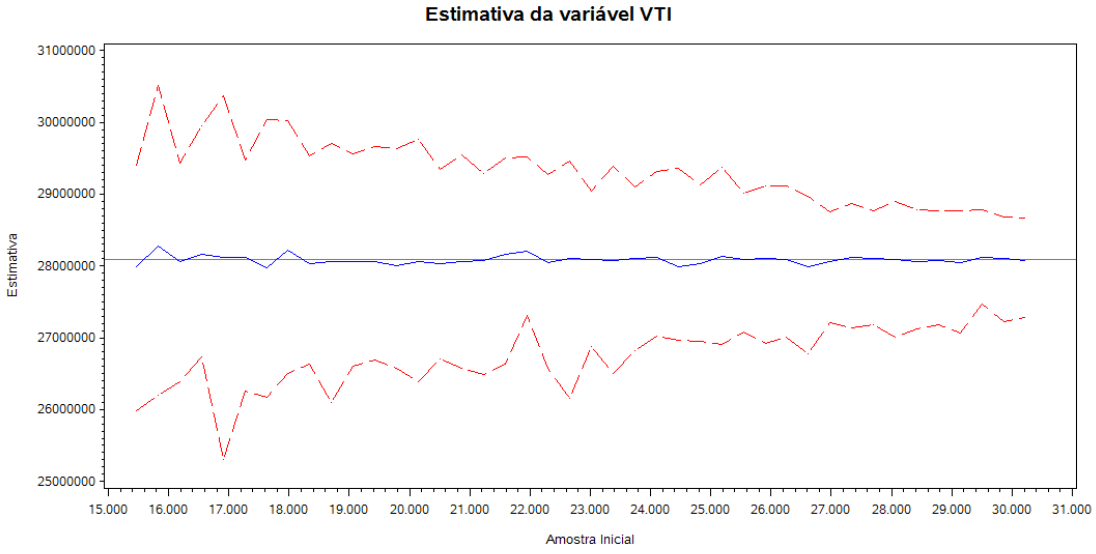

Figura 5.6: Estimativa da Variável VTI

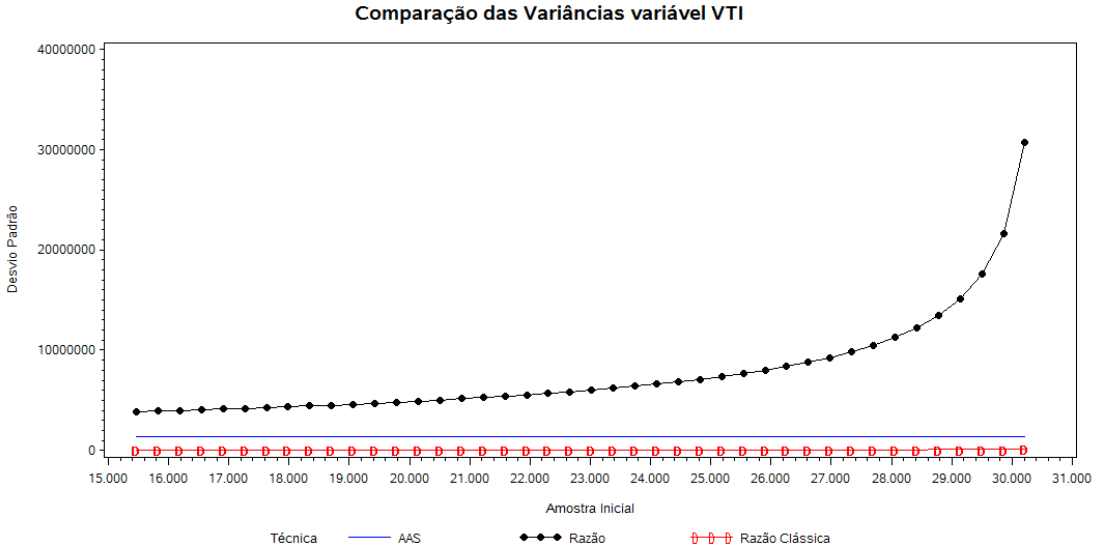

Figura 5.7: Comparação das Variâncias VTI

Quando analisada a técnica de AR clássica é possível concluir que esta técnica possui uma variância profundamente menor do que a técnica de AAS nos indicadores da indústria brasileira. É possível obter uma melhor visualização desta discrepância

na Figura 5.8.

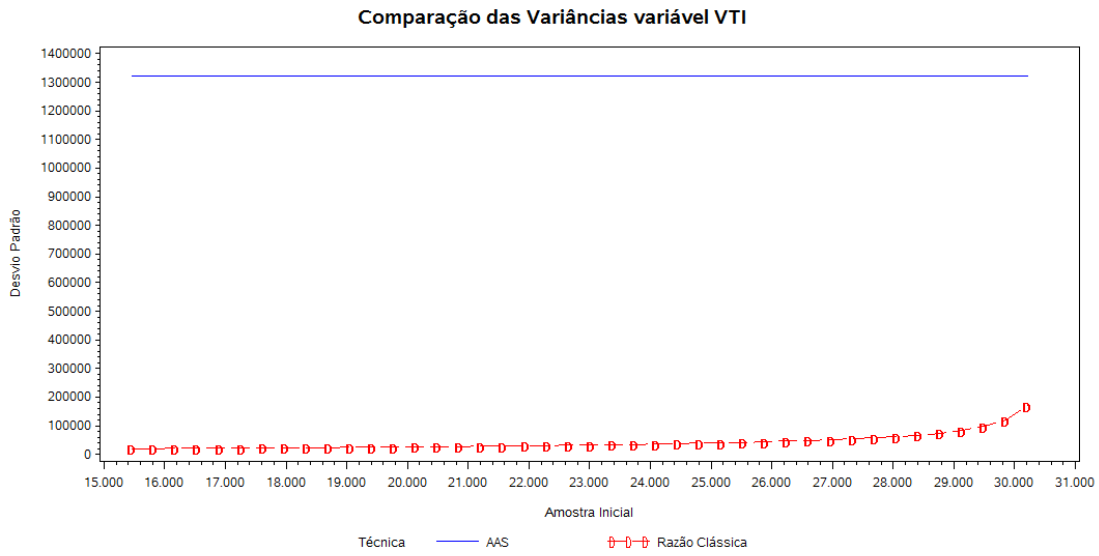

Figura 5.8: Comparação das Variâncias da AAS e da AR clássica

Não se sabe, ao certo, se é posto em prática a variância da AR clássica para a realização de pesquisas, já que a variância da AD para o estimador tipo razão é dada pela equação 3.18.

5.4 Estimador Tipo Regressão

A partir da variável ENG foi feita uma reta de regressão para estimar a variável P&D. A Figura 5.9 mostra que a estimativa da variável P&D, azul, está sempre ao entorno da verdadeira média e as menores e maiores estimativas, curvas em vermelho, estão se aproximando da verdadeira média a medida que a amostra inicial aumenta.

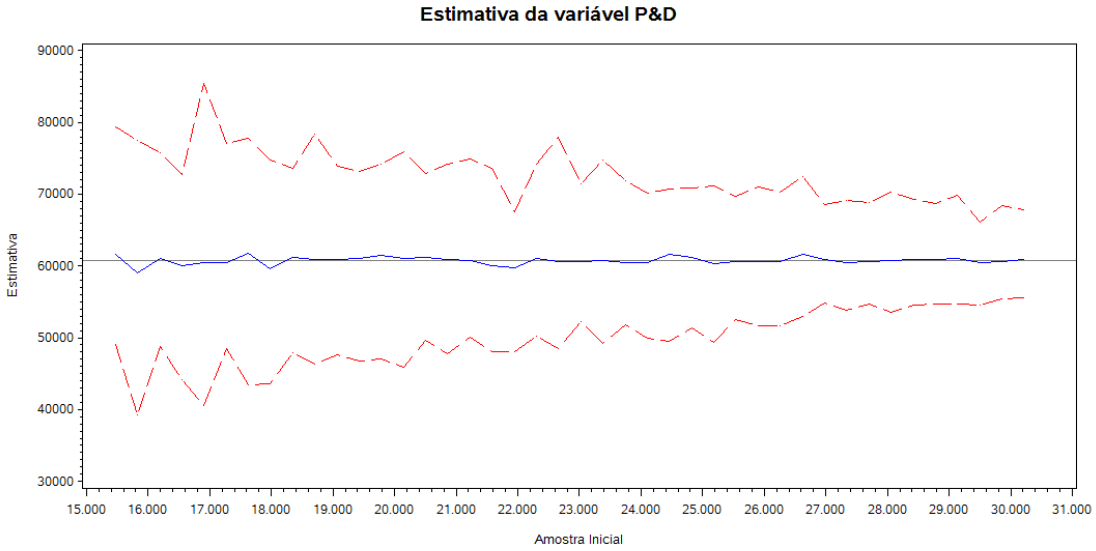

Figura 5.9: Estimativa da Variável P&D

Estas duas variáveis possuem uma correlação fraca, 16,53%, e um R^2 igual a 2,73%, isto posto, apenas 2,73% da variabilidade da variável P&D é explicada pela variável ENG. A baixa influência que a variável explicativa (ENG) possui em relação a variável P&D faz com que a reta de regressão não explique de forma convincente o essas duas variáveis. Além disso, a baixa correlação faz com que as curvas de variância das técnicas de AD para o estimador tipo regressão e AReg clássica sejam quase iguais como mostra a Figura 5.10.

Como nas outras técnicas, a AD com o estimador tipo regressão também não se mostrou vantajoso se comparado com a AAS e, neste caso, nem mesmo a técnica de AReg clássica se mostrou assim. Vale ressaltar que, como AE clássica e a AR clássica, não se sabe se é uma prática comum utilizar a variância da AReg clássica para a realização da AD, uma vez que a variância da AD para o estimador tipo regressão é dada pela equação 3.22.

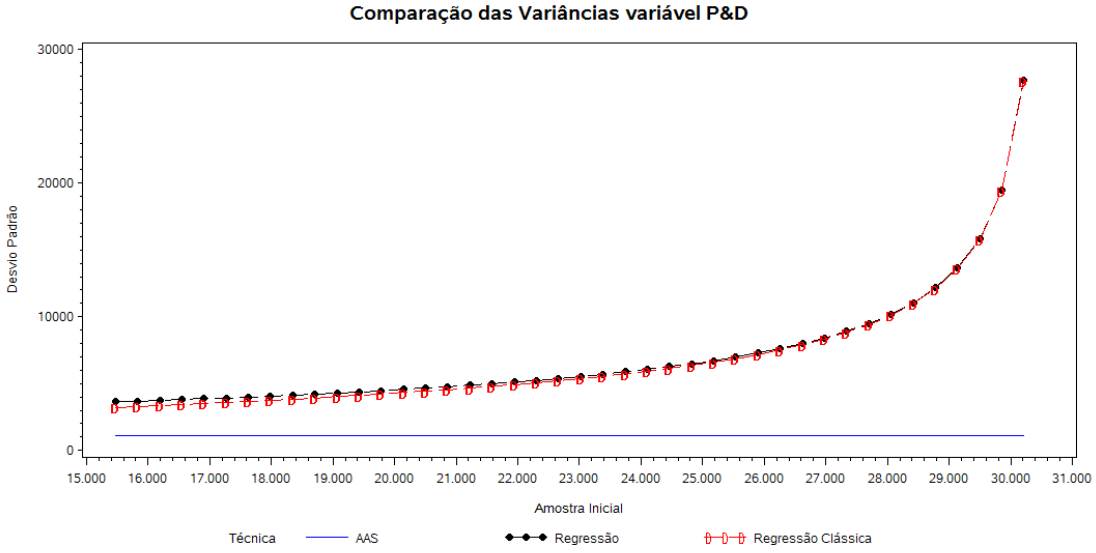

Figura 5.10: Comparação das Variâncias P&D

Capítulo 6

Conclusões

No presente trabalho foi analisado a eficiência da AD nos indicadores da indústria brasileira comparando a sua curva de variância com a reta de variância da AAS. Estes indicadores da indústria foram simulados a partir das estatísticas descritivas das pesquisas PINTEC e PIA, realizadas pelo IBGE, e obtidos pela pesquisa RAIS, realizada pelo MTE, que serviu de base para calibrar os modelos de regressão. Portanto, as variáveis explicativas PO e ENG, que foram retiradas da RAIS, foram utilizadas para a obtenção das variáveis VTI, SM, RLV e P&D, já que apenas as estatísticas descritivas destas variáveis eram conhecidas.

Como visto anteriormente, a técnica de AD consiste em selecionar a amostra base de n elementos não diretamente de N, mas de uma pré-seleção de uma grande amostra de n' permitindo, assim, maximizar a utilização de variáveis auxiliares ou variável de estratificação.

Nos indicadores da indústria brasileira, a AD para estratificação não se mostra vantajosa se comparada com a AAS na grande maioria da série (Figuras 5.3 e 5.4). Apenas quando utilizada a AE clássica com alocação ótima, é possível observar que a variância desta técnica para as variáveis RLV e SM é menor em aproximadamente

75% e 60%, respectivamente, na amostra inicial se comparada com a variância da AAS. Vale ressaltar que essa prática foi feita apenas para comparação e não se sabe se é correta, uma vez que a correta variância da AD para estratificação é dada pela Equação 3.5. Quando comparada as variâncias da AD para estratificação e da AE clássica com alocação proporcional com a reta de variância da AAS, nota-se que a AAS é sempre melhor.

Tinha-se uma expectativa que a AD para estratificação fosse mais eficiente do que a AAS, o que não aconteceu. Com isso foi feito mais um estudo de caso para averiguar se o resultado aconteceu devido ao grande número de estratos ou as altas variâncias dos estratos. Sendo assim, foi feito o mesmo estudo de caso com a variável SM, porém com apenas as 6 CNAEs que possuem a menor variância. O resultado se mostrou muito parecido com o anterior. Dessa forma, pode-se verificar que para os indicadores da indústria brasileira, a AAS é mais eficiente do que a AD para estratificação.

A AD para o estimador tipo razão (5.7) e regressão (5.10) também não se mostraram vantajosas quando comparada com a AAS para os indicadores da indústria brasileira. Apenas quando analisada a curva de variância da AR clássica, nota-se que esta é muito mais eficaz do que a AAS. Mas uma vez não se sabe se é uma prática comum utilizar a variância da AR clássica e da AReg clássica, uma vez que as variâncias da AD para o estimador tipo razão e AD para o estimador tipo regressão são, respectivemente, dadas pelas equações 3.18 e 3.22.

A partir do exposto acima, não se obteve indícios de que a AD para os indicadores

da indústria brasileira seja mais vantajosa do que a AAS utilizando todo rescurso disponível. Contudo, vale ressaltar que o sucesso da AD pode não ter ocorrido devido ao alto custo gasto na amostra preliminar, custo que foi igual ao da amostra principal. Com isso, a amostra principal foi prejudicada, o que deve ter influenciado o resultado do presente trabalho, já que, uma característica da técnica de AD é, justamente, o baixo custo da primeira amostra.

Referências Bibliográficas

Bolfarine, H. & Bussab, W. O. (2005). *Elementos de Amostragem*. ABE - Projeto Fisher.

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

CNI (2011). Metodologia dos indicadores industriais. Technical report, Confederação Nacional da Indústria.

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

CONCLA (2007). Classificação nacional de atividades econômicas - cnae versão 2.0. Technical report. URL http://www.ibge.gov.br/concla/revisao2007.php?l=6. acesso em 28 mai. 2012.

IBGE (2001). *Pesquisa Industrial Anual - Empresa (1996-2001)*. URL http://www.ibge.gov.br/home/estatistica/economia/industria/pia/empresas/defaultempresa.shtm. acesso em 25 abr. 2012.

IBGE (2004). Notas técnicas - pintec. Technical report, Instituto Brasileiro de Geografia e Estatística.

IBGE (2007). Introdução à classicação nacional de atividades econômicas - cnae versão 2.0. Technical report. URL http://www.ibge.gov.br/concla/pub/revisao2007/PropCNAE20/CNAE20_Introducao.pdf. acesso em 28 mai. 2012.

IBGE (2009). Notas técnicas - pia. Technical report, Instituto Brasileiro de Geografia e Estatística.

IBGE (2012). Início da coleta pintec. Technical report. URL http://www.pintec.ibge.gov.br/index.php?option=com_content&view=article&id=60:-inicio-da-coleta-da-pintec-2011&catid=7:noticias&Itemid=10. acesso em 06 mai. 2012.

Kish, L. (1965). *Survey Sampling*. Wiley.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

MTE (2011). Rais - relação anual de informações sociais. Technical report. URL http://www.mte.gov.br/pdet/o_pdet/reg_admin/rais/apres_rais.asp. acesso em 27 mai. 2012.

MTE (2012). O que é a rais. Technical report. URL http://www.rais.gov.br/rais_sitio/oque.asp. acesso em 27 mai. 2012.

SAS (2008). *SAS software, Versão 9.2 do SAS System para Windows. Licenciado para o Departamento de Estatística - Universidade de Brasília. Cary - Carolina do Norte.*

Scheaffer, R. L., Medenhall, W., & Ott, R. L. (1996). *Elementary Survey Sampling*, (5th ed.). Duxbury Press.

Sukhatme, P. V. (1954). *Sampling Theory of Surveys with Applications*. The Iowa State College Press.