Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

ANÁLISE DAS PROPRIEDADES DO ESTIMADOR HORVITZ-THOMPSON

THIAGO DANTAS BHERING DOMINONI 09/49345

Brasília

2012

THIAGO DANTAS BHERING DOMINONI 09/49345

ANÁLISE DAS PROPRIEDADES DO ESTIMADOR HORVITZ-THOMPSON

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

2012

A família e amigos.

Thiago Dantas Bhering Dominoni

Agradecimentos

Agradeço a todas as oportunidades que tive.

Aos meus pais que proporcionaram com muito esforço tudo o que um filho precisa.

A todos que me apoiaram nos momentos de dificuldades e me ajudaram a superar desafios, em especial Vanessa.

Ao professor Alan, por ter sido mais que um orientador no desenvolvimento deste trabalho.

Resumo

O estimador Horvitz-Thompson (HT) é um estimador não viesado do total populacional utilizado em amostragem sem reposição de um universo finito com probabilidades desiguais de seleção. No entanto, certas propriedades do estimador surgem como um problema, como as estimativas de variância negativa, a falta de um intervalo de confiança adequado para suas estimativas e as condições necessárias para que seja mais eficiente do que planos amostrais que fazem uso de probabilidades iguais de seleção.

Este trabalho apresenta uma investigação descritiva acerca das condições ideais do uso de uma variável auxiliar a fim de resultar em maior eficiência para o estimador HT. Além disso, são comparados os estimadores de variância considerando situações onde seu uso é adequado. Por fim, é realizada uma comparação entre os intervalos exatos das estimativas HT e as estimativas da aleatória simples como também é proposta uma distribuição para a estimativas HT a ser utilizada na construção de um intervalo de confiança adequado fazendo uso de dados amostrais.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

Introdução

O estimador Horvitz-Thompson é um estimador não tendencioso do total populacional que foi construído para tratar de amostras retiradas sem reposição de um universo finito com probabilidades desiguais de seleção. No entanto, este estimador tem aplicação em qualquer plano amostral, com ou sem reposição. Conhecida sua utilidade e aplicação deseja-se conhecer suas propriedades mais a fundo.

Segundo Horvitz and Thompson (1952), o uso apropriado de probabilidades desiguais para a seleção dos elementos amostrais permite obter ganhos em relação aos métodos com iguais probabilidades de seleção. No entanto, quando faz-se uso de uma variável suplementar para obter-se as probabilidades de seleção não se sabe que características deve possuir para resultar em maior eficiência frente a métodos como a aleatória simples (AAS), que utiliza probabilidades iguais de seleção. Esta discussão é detalhada na seção 3.2.

Os estimadores da variância do estimador Horvitz-Thompson (HT) são computacionalmente intensos e segundo Lohr (1999) apresentam estimativas negativas com frequência. Sabe-se de acordo com Brewer (1963) que com escolhas adequadas das probabilidades de inclusão as estimativas da variância do estimador HT tornam-se

mais simples de serem calculadas e existe um padrão de quando são positivas. No entanto, quando as probabilidades de inclusão são obtidas a partir de uma variável suplementar, não se sabe como as características dessa variável afetam as estimativas de variância. Esta discussão é detalhada na seção 3.3

Na literatura o estimador HT é comumente apresentado unicamente com sua estimativa pontual de total populacional. Pouco foi feito na tentativa de obter um intervalo de confiança para as suas estimativas. A seção 3.4 apresenta uma tentativa de obter-se um intervalo de confiança adequado para o estimar HT.

Este trabalho é organizado em 4 capítulos. O capítulo 1 apresenta um resumo sobre os problemas que certas propriedades do estimador HT possuem. O capítulo 2 apresenta conceitos básicos que serão utilizados ao longo do trabalho. A análise das propriedades de interesse do estimador HT são discutidas no capítulo 3. As conclusões e as propostas para trabalhos futuros são apresentados no capítulo 4.

1.1 Objetivos

O objetivo geral do trabalho é construir um intervalo de confiança adequado para o estimador Horvitz-Thompson (HT).

Os objetivos específicos são:

- identificar motivos que fazem o estimador de variância de HT ser negativo;
- comparar a eficiência do estimador HT frente a AAS;
- implementar, computacionalmente, uma forma de obter intervalos de confiança para as estimativas geradas pelo estimador HT;

Capítulo 2

Conceitos básicos

2.1 Intervalos de confiança

Intervalos de confiança (IC) são uma importante parte da inferência estatística. De acordo com Bickel and Doksum (2001), O IC se refere a obter expressões do tipo $S_{\mathbf{x}} = P(a(X_1, ...X_n) \leq \theta \leq b(X_1, ..., X_n)) = 1 - \alpha$, onde θ é o parâmetro de interesse e a e b são números calculados baseados em uma amostra iid $(X_1, ..., X_n)$. A probabilidade $1 - \alpha$ é chamada de coeficiente de confiança. Ao contrário da estimação pontual $\widehat{\theta}$, IC's nos dão um intervalo. Podemos interpretar o IC da seguinte maneira: em amostras repetidas de uma mesma população se calcularmos em cada uma o intervalo $S_{\mathbf{x}}$, em aproximadamente $100(1-\alpha)\%$ das vezes, $S_{\mathbf{x}}$ conterá o valor descolhecido de θ.

Geralmente, para se construir um IC devemos ter alguma informação sobre a distribuição dos dados. Caso tal informação não estiver disponível, ainda assim poderemos construir intervalos de confiança usando teorias assintóticas ou até mesmo métodos de simulação.

Um método muito útil na construção de IC's é o método pivotal. Este método

constrói primeiramente um IC para o elemento pivô e então transforma o intervalo para o parâmetro θ.

Definição O elemento pivô é uma função de $\theta, X_1, ..., X_n$ que não depende de θ.

Tipicamente, os elementos pivô escolhidos $g(\theta, X_1, ..., X_n)$ tem distribuição Normal, Qui-Quadrado, T ou F. Como essas distribuições são bem conhecidas é fácil obter intervalos de confiança para os elementos pivô:

$$P(a \leq g(\theta, X_1, ..., X_n) \leq b) = 1 - \alpha$$

Contruído um IC para o elemento pivô, o próximo passo é obter o IC para θ e isso é obtido resolvendo a inequação presente no intervalo de confiança de $g(\theta, X_1, ..., X_n)$ para θ.

Em inferência estatística clássica temos especial interesse em obter intervalos de confiança para a média e a variância de uma população $\mathcal{N}(\mu, \sigma^2)$. Podemos investigar as propriedades de intervalos de confiança a partir de como são distribuídos os elementos pivô utilizados na construção de um IC para θ. Conforme Bickel and Doksum (2001), considere os resultados abaixo:

Sejam $X_1, ..., X_n$ itens de uma amostra da $\mathcal{N}(\mu, \sigma^2)$.

Então:

1. $\overline{X} = \frac{X_1 + ... + X_n}{n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

2. $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$ então $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

3. $\overline{X}$ e s^2 são independentes.

2.1.1 Intervalos simétricos

Considere a situação descrita anteriormente com $X_1, ..., X_n$ itens de uma amostra da $\mathcal{N}(\mu, \sigma^2)$ e todos os resultados que se seguem. Temos:

$$t = \sqrt{n}\frac{\overline{X} - \mu}{s} = \frac{\sqrt{(n)}\frac{\overline{X}-\mu}{\sigma}}{\sqrt{\frac{(n-1)s^2}{(n-1)\sigma^2}}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Portanto, t segue uma distribuição *t de student* e será o elemento pivô para a determinação de um intervalo de confiança para μ com coeficiente de confiança $1-\alpha$.

Temos,

$$P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha$$

$$P\left(-t_{\alpha/2} \leq \frac{\sqrt{n}\,\overline{X} - \mu}{s} \leq t_{\alpha/2}\right) = 1 - \alpha$$

Resolvendo para μ:

$$P\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

E portanto um intervalo de confiança para μ é:

$$S_{\mathbf{x}} : \left[\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

2.1.2 Intervalos assimétricos

Novamente considere a situação descrita anteriormente com $X_1, ..., X_n$ itens de uma amostra da $\mathcal{N}(\mu, \sigma^2)$ e todos os resultados que se seguem. Para obtermos um intervalo de confiança para σ^2 podemos usar como elemento pivô:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

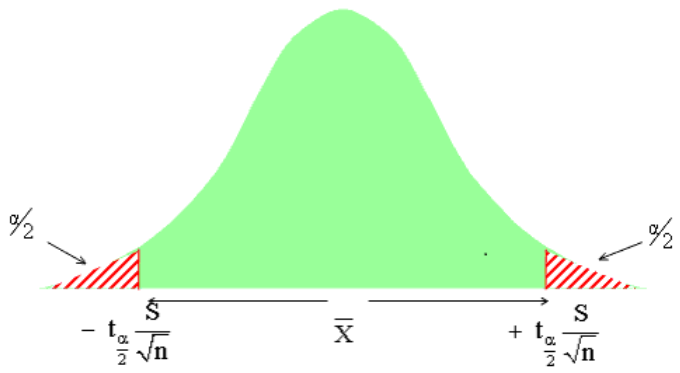

Figura 2.1: Intervalo de confiança para μ.

Temos,

$$P\left(\chi^2_{\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha$$
$$P\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha$$

E portanto um intervalo de confiança para σ^2 é:

$$S_{\mathbf{x}} : \left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right]$$

É importante ressaltar que $|\chi^2_{\alpha/2}| \neq |\chi^2_{1-\alpha/2}|$.

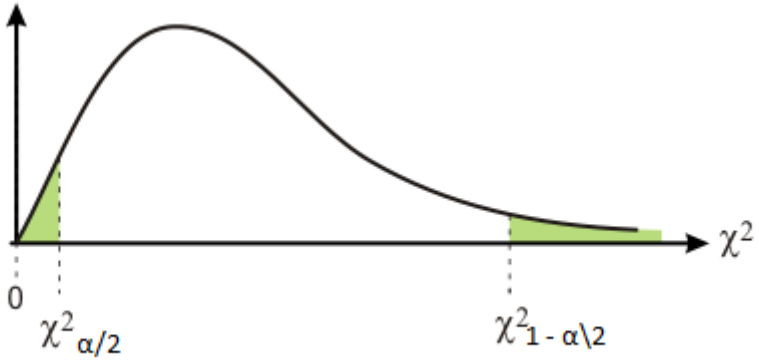

Figura 2.2: Assimetria da distribuição Qui-Quadrado.

É simples perceber a diferença entre intervalos simétricos e assimétricos a partir das Figuras 2.1 e 2.2. Se definirmos um intervalo de confiança para θ como $S_{\mathbf{x}} =$

$\left[\hat{\theta} + a \leq \theta \leq \hat{\theta} + b\right]$, então para o intervalo simétrico, $a = -|b|$ enquanto que para o intervalo assimétrico embora $a < b$, no entanto $a \neq -|b|$.

2.2 Estimador Horvitz-Thompson

O estimador Horvitz-Thompson (HT) é um estimador não tendencioso do total populacional que foi construído para tratar de amostras retiradas sem reposição de um universo finito com probabilidades desiguais de seleção. No entanto, este estimador tem aplicação em qualquer plano amostral, com ou sem reposição. Como visto em Nascimento (2011), o estimador é o caso geral dos principais planos amostrais e isso pode ser evidenciado analisando a fórmula de variância.

Conforme Horvitz and Thompson (1952), quando estudamos uma população finita onde somos capazes de identificar seus elementos individualmente, podemos atribuir um vetor qualquer de probabilidades de seleção a esses elementos. Fazendo uma escolha adequada desse vetor de probabilidades é possível reduzir a variância de estimativas não viesadas se compararmos com aquelas obtidas utilizando-se probabilidades iguais de seleção.

É exatamente isso que o estimador HT faz. Utilizando um vetor inicial de probabilidades de seleção, geralmente obtido a partir de uma variável relacionada com a variável de interesse, calculam-se as probabilidades de inclusão, que representam a probabilidade condicional de um certo elemento estar incluído na amostra em determinada escolha e também a probabilidade de inclusão na amostra de cada par de elementos. Dessa maneira, o estimador HT trata com especificidade cada elemento da população proporcionando confiança na utilização dos resultados obtidos.

O estimador não viesado para o total populacional é dado por:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} \tag{2.1}$$

onde π_i é a probabilidade de inclusão na amostra do i-ésimo elemento e y_i é a medida da variável de interesse do i-ésimo elemento.

Pode-ser obter a estimativa para a média populacional dividindo (2.1) por N. A variância associada ao estimador é dada por:

$$V(\hat{Y}_{HT}) = \sum_{i}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i}^{N} \sum_{i \neq j}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j \tag{2.2}$$

sendo π_i e π_j da forma descrita anteriormente e π_{ij} a probabilidade de que conjuntamente os elementos i e j estejam na amostra. O estimador de (2.2) foi dado por Horvitz and Thompson (1952):

$$v_1(\hat{Y}_{HT}) = \sum_{i}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i}^{n} \sum_{j>i}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j \tag{2.3}$$

Um outro estimador de 2.2 foi proposto por Yates and Grundy (1953) e por Sen (1953):

$$v_2(\hat{Y}_{HT}) = \sum_{i}^{n} \sum_{j>i}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{2.4}$$

No entanto, apesar de não viesados, os estimadores apresentados em (2.3) e (2.4) podem assumir valores negativos. Apesar disso, conforme Cochran (1977), com uma escolha adequada das probabilidades de seleção os estimadores de (2.2) podem tornar-se mais simples de se calcular e também mais estáveis.

2.3 Efeito de planejamento

Quando existem estimativas de dois ou mais planos amostrais distintos é importante saber qual é o "melhor". Para podermos comparar diferentes estimadores precisamos primeiramente de uma medida. Conforme Cochran (1977), temos que:

Definição Efeito de planejamento ou *design effect* (*deff*) é uma medida que compara a variância de um estimador qualquer com relação a outro considerado padrão.

Pode-se utilizar o efeito de planejamento para saber qual estimador é mais eficiente, desde que ambos estimadores sejam não viesados para um mesmo parâmetro populacional e admite-se que o custo de ambos seja igual. Portanto, o melhor estimador será aquele que apresentar menor variância. O *deff* é calculado por:

$$deff = Var(estimadorA)/Var(estimadorB) \tag{2.5}$$

Se $deff < 1$ tem-se a indicação que o estimador A é mais eficiente do que o estimador B.

Capítulo 3

Propriedades do estimador Horvitz-Thompson

3.1 Introdução

Conforme abordado nos objetivos deste trabalho, tem-se interesse na construção de um intervalo adequado para o estimador HT. Para isso é necessário implementar uma maneira de utilizar as estimativas de total geradas pelo estimador HT para a construção do intervalo. Tem-se também o interesse em investigar as razões que levam os estimadores da variância do estimador HT a apresentarem estimativas negativas. Por fim, deseja-se analisar a afirmação de Horvitz e Thompson de que o uso apropriado de probabilidades desiguais de seleção dos elementos amostrais pode trazer ganho de eficiência sobre sistemas que usam probabilidades iguais de seleção.

Este capítulo apresenta resultados empíricos que mostram como devem ser as probabilidades de seleção para garantir maior eficiência do estimador HT frente o estimador da AAS. É apresentado também uma investigação descritiva que tenta explicar as situações de estimativas negativas do estimador de variância do estimador HT, como também uma comparação entre os dois estimadores de variância. Por fim,

é apresentada a construção do intervalo de confiança das estimativas geradas por HT juntamente com uma comparação com os intervalos gerados pela AAS exata e estimada utilizando as distribuições normal e *t de student*.

3.2 Comparação das estimativas de variância do estimador HT e AAS

Usualmente, quando trabalha-se com amostragem sem reposição e com probabilidades desiguais de seleção, o vetor de probabilidades de seleção é obtido a partir da informação disponível em uma variável suplementar. Surge a dúvida de como devem ser essas probabilidades a fim de garantir variância mínima ao estimador HT.

Se analisarmos a Equação 2.1 e seus estimadores poderemos perceber que o que pode ser controlado pelo pesquisador são as probabilidades de inclusão π_i e π_{ij}. Se assumirmos que existe uma variável suplementar X associada a variável de interesse Y e que o pesquisador deseja utilizar a informação contida em X para designar as probabilidades de seleção de tal forma que as probabilidades de inclusão resultantes trarão uma redução na variância, surge o problema de como deve ser a relação entre Y e X para obter ganhos em eficiência.

Em Horvitz and Thompson (1952), os autores fazem uma breve discussão de como uma escolha de probabilidades de inclusão proporcionais a variável suplementar ($\pi_i = nXi/\sum_{i=1}^{N} X_i$) pode trazer redução na variância do estimador HT. No entanto, aparentemente a escolha dos $\pi_i's$ que levam à variância mínima depende da distribuição conjunta de Y e X. E isso complica o problema.

Neste trabalho, no entanto, ao invés de buscar-se condições de variância mínima

com base em uma variável suplementar, investigaremos como deve ser a relação entre Y e X para que o estimador HT seja mais eficiente do que a AAS.

Para compararmos os dois estimadores vamos utilizar o conceito de efeito de planejamento (*deff*) conforme definido anteriormente.

A comparação entre as estimativas de variância do estimador HT e da AAS foi realizada por diversas simulações onde o efeito de planejamento foi calculado comparando as estimativas de variância do estimador HT e da AAS obtidas para cada combinação de tamanho de população com tamanho de amostra possível. Gerou-se populações de tamanhos $N = 4$ a $N = 8$ e variáveis auxiliares de iguais tamanhos.

O primeiro ponto investigado foi se apenas uma forte correlação linear entre Y e X garantiria uma maior eficiência do estimador HT. Os resultados obtidos mostraram que esse fato sozinho não é suficiente para garantir maior eficiência do estimador HT frente a AAS.

Investigou-se então a natureza descritiva de X e Y para tentar obter um padrão nos efeitos de planejamento obtidos. Como resultado, descobriu-se que o coeficiente de variação de X e Y tem grande influência nas estimativas de variância.

As Tabelas abaixo mostram o resultado de uma das simulações onde a variabilidade das variáveis era controlada assim como a correlação entre elas.

Tabela 3.1: Medidas descritivas da população

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Média	434.45	222.17	622.27	318.16	228.37
Desvio Padrão	382.90	120.85	569.87	312.38	151.14
Coeficiente Variação	88.14	54.40	91.58	98.18	66.18
Desvio Interquantílico	554.85	157.35	269.77	246.24	267.36
Assimetria	1.30	0.53	1.74	2.05	0.70
Mínimo	147.40	91.55	51.28	86.05	91.75
Máximo	968.60	384.43	1711.52	985.60	460.81

Tabela 3.2: Medidas descritivas da variável auxíliar

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.77	0.94	0.83	0.97	0.94
Média	437.50	230.00	200.00	278.57	381.25
Desvio Padrão	311.92	152.48	130.38	328.96	277.67
Coeficiente Variação	71.30	66.30	65.19	118.08	72.83
Desvio Interquantílico	375.00	150.00	200.00	400.00	450.00
Mínimo	250.00	50.00	50.00	50.00	50.00
Máximo	900.00	450.00	350.00	950.00	800.00

Tabela 3.3: Efeitos de planejamento obtidos

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	17.33%	2.22%	12.97%	10.92%	22.44%
3	12.43%	3.82%	13.53%	11.33%	23.25%
4		9.26%	14.35%	11.01%	23.51%
5			15.26%	9.42%	21.65%
6				6.98%	15.64%
7					5.85%

A população foi gerada de uma distribuição qui-quadrado com 3 graus de liberdade e seus valores multiplicados por 100. A variável auxiliar foi gerada a partir de

uma distribuição geométrica de parâmetro 0.3 e seus resultados multiplicados por 50.

As Tabelas 3.1 e 3.2 apresentam medidas descritivas da população e da variável auxiliar, respectivamente. Já a Tabela 3.3 apresenta os *deff's* obtidos.

Podemos notar que todos os *deff's* foram menores do que 100%, o que indica que em todas as combinações possíveis de tamanho de população e amostra as estimativas de variância do estimador HT foram sempre menores do que as da AAS.

Se analisarmos as Tabelas 3.1 e 3.2 podemos notar que os coeficientes de variação das duas variáveis são altos, e são relativamente próximos. Esse é o fator que segundo as simulações realizadas parece ser mais determinante para um bom resultado do estimador HT. Além de escolhermos uma variável auxiliar que tenha uma alta correlação linear com a variável de interesse, o coeficiente de variação de ambas as variáveis devem ser próximos.

Nos apêndices são apresentados resultados utilizando diferentes distribuições para gerar diferentes variáveis auxiliares para uma mesma população, controlando os coeficientes de variação dessas variáveis auxiliares. É simples perceber que a partir do momento que o coeficiente de variação da variável auxiliar começa a ser muito diferente da variável de interesse o estimador HT vai piorando em relação à AAS.

A partir das simulações observou-se também que o estimador HT apresenta pior eficiência frente a AAS, quando temos uma população com baixo coeficiente de variação e utiliza-se como variável auxiliar uma variável com alto coeficiente de variação.

3.3 Estimadores da variância do estimador HT

A variância do estimador HT possui dois estimadores mais conhecidos, são eles *v1* e *v2*, conforme vimos em 2.3 e 2.4:

$$v_1(\hat{Y}_{HT}) = \sum_{i}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i}^{n} \sum_{j>i}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j$$

$$v_2(\hat{Y}_{HT}) = \sum_{i}^{n} \sum_{j>i}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

De acordo com o apresentado em Lohr (1999), os estimadores *v1* e *v2* podem resultar em estimativas negativas dependendo de como são as probabilidades de seleção e o cálculo dos mesmos é bastante problemático, exigindo muita intensidade computacional.

Temos especial interesse em tentar identificar padrões nos dados que geram estimativas negativas de *v1*. Para isso, foram gerados vetores Y de tamanhos $N = 4$ a $N = 7$, representando a população de interesse e também vetores X de mesmos tamanhos, representando a variável auxiliar.

O vetor Y foi gerado a partir de uma distribuição qui-quadrado com 3 graus de liberdade e e seus valores multiplicados por 100. O vetor X foi gerado a partir de uma distribuição geométrica de parâmetro 0.3 e seus resultados multiplicados por 50.

Uma vez com os dados, foram obtidas as estimativas de *v1* para cada combinação de tamanho de população e amostra possível. Por exemplo: para $N = 4$, temos dois tamanhos de amostra possíveis (2 e 3) já que com $n = 1$ não há variabilidade e não

faz sentido calcular *v1* e para $n = 4$ temos que $n = N$ e não precisamos estimar o parâmetro de interesse.

Obtidas as estimativas de *v1* para cada amostra possível dentro de cada população de $N = 4$ a $N = 7$, podemos selecionar aquelas que apresentaram *v1* < 0 e tentar identificar algo em comum nessas amostras. A Tabela abaixo apresenta a proporção de amostras que resultaram em estimativas negativas de *v1* em cada combinação de população e amostra.

Tabela 3.4: Proporção de amostras que apresentaram estimativas de $v1 < 0$

amostra	**pop4**	**pop5**	**pop6**	**pop7**
2	66.70%	60.00%	46.70%	66.70%
3	50.00%	70.00%	50.00%	68.60%
4		80.00%	46.70%	65.70%
5			0%	61.90%
6				71.40%

A partir da Tabela 3.4 pode-se notar a alta proporção de estimativas de *v1* negativas, com exceção das amostras obtidas para $N = 6$ de tamanho 5, onde nenhuma delas resultou em uma estimativa negativa. A alta frequência de estimativas negativas é uma clara evidência da preocupação em tentar entender esse motivo, uma vez que em uma situação real só teríamos uma amostra podendo obter uma estimativa negativa.

Para tentar identificar algum padrão em comum nas amostras que geraram estimativas negativas foram investigadas as seguintes hipóteses:

- A relação entre a média de Y populacional e a amostral influi em v1;

- A relação entre a variância de Y populacional e a amostral influi em v1;
- A relação entre a média de X populacional e a amostral influi em v1;
- A relação entre a variância de X populacional e a amostral influi em v1;

Como resultado, não obteve-se nenhum padrão identificável. Nenhuma relação em cada uma das hipóteses obedeceu a um padrão no fato de *v1* ser negativa.

Para tentarmos entender o que pode contribuir para que *v1* seja negativa, vamos analisar mais uma vez sua fórmula:

$$v_1(\hat{Y}_{HT}) = \underbrace{\sum_{i}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2}_{>0} + \underbrace{2\sum_{i}^{n}\sum_{j>i}^{n} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j\pi_{ij}} y_i y_j}_{<0 \text{ ou } >0}$$

Pode-se notar que para que *v1* seja negativo o segundo termo de sua equação precisa ser negativo e de maior intensidade que o primeiro, uma vez que $0 \leq \pi_i \leq 1$.

No entanto, há muitas variáveis influenciando nesse cálculo. A começar pelo número de termos da equação. A primeira parte da equação terá a quantidade de termos igual ao tamanho da amostra selecionada. Já a segunda metade terá $\binom{n}{2}$ termos, onde cada um desses termos pode ou não ser negativo. O cálculo envolve as observações da variável Y - cujo total deseja-se estimar - e também as probabilidades de inclusão de cada item da amostra, como também as probabilidades de inclusão conjunta 2 a 2 utilizando os itens da amostra.

Como identificar algum fator isolado que contribua para que *v1* seja negativo parece ser complicado, deseja-se analisar com que frequência o segundo termo da Equação 2.3 é negativo. É simples notar que o segundo termo da equação somente

será negativo quando $\pi_{ij} - \pi_i\pi_j$ for negativo, assumindo que estamos trabalhando com uma variável Y que assuma somente valores positivos.

Surpreendentemente, obtivemos que 100% das vezes o termo $\pi_{ij} - \pi_i\pi_j$ foi negativo, utilizando os dados que foram gerados para se trabalhar nesse problema.

Este é um resultado interessante, já que segundo Lohr (1999) *v2* também pode assumir valores negativos, mas se analisarmos a Equação 2.4, vemos que a fórmula de *v2* leva em consideração $\pi_i\pi_j - \pi_{ij}$ e portanto não poderia assumir valores negativos.

Temos também um indício de que *v1* tem sempre a possibilidade de assumir valores negativos, uma vez que o segundo termo sempre será negativo. Este estimador somente será positivo quando o primeiro termo tiver uma intensidade maior do que o segundo.

Os resultados anteriormente obtidos foram testados utilizando outras distribuições para gerar os dados. Foram utilizadas todas as combinações de população e amostra que estão presentes nos apêndices na seção "*Deff's*". O resultado foi consistente, e sempre foi obtido que $\pi_{ij} - \pi_i\pi_j < 0$.

Como não é necessário que as probabilidades de inclusão (π_i) sejam calculadas fazendo uso de uma variável auxiliar, é possível que o pesquisador que esteja trabalhando com o estimador HT consiga obter probabilidades de inclusão que façam *v2* ser negativa. No entanto, se essas probabilidades foram obtidas a partir de uma variável auxiliar, os resultados aqui obtidos indicam que *v2* será sempre positivo.

Sabe-se que tanto *v1* quanto *v2* são estimadores não viesados da verdadeira variância do estimador HT. Portanto, a esperança de ambos devem ser igual ao

mesmo valor, a variância do estimador HT. Vimos que *v1* pode assumir muitos valores negativos. A dúvida que surge é a respeito da qualidade das estimativas de *v1*, uma vez que suas estimativas positivas deverão compensar as negativas para que a esperança resulte em um valor positivo e igual a esperança de *v2*, cujas estimativas são sempre positivas da forma como foram obtidas as probabilidades de inclusão.

Para podermos analisar essa dúvida foram construídas as Figuras 3.1, 3.2 e 3.3. Essas figuras exemplificam resultados semelhantes obtidos com outras combinações de tamanho de população e amostra.

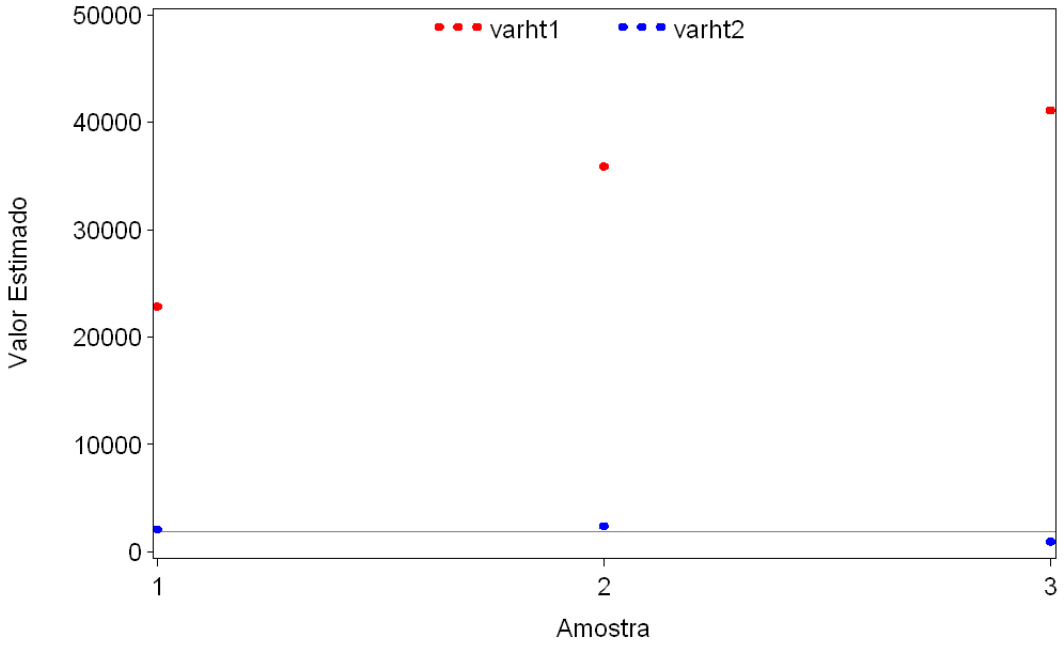

Figura 3.1: Comparação das estimativas de *v1* e *v2* ($N = 5$ e $n = 3$)

As Figuras 3.1, 3.2 e 3.3 foram obtidos selecionando somente aquelas amostras que resultaram em $v1 > 0$. Comparou-se então a estimativa obtida naquela amostra para *v1* e *v2* em relação a variância que estão estimando, representada pela linha contínua.

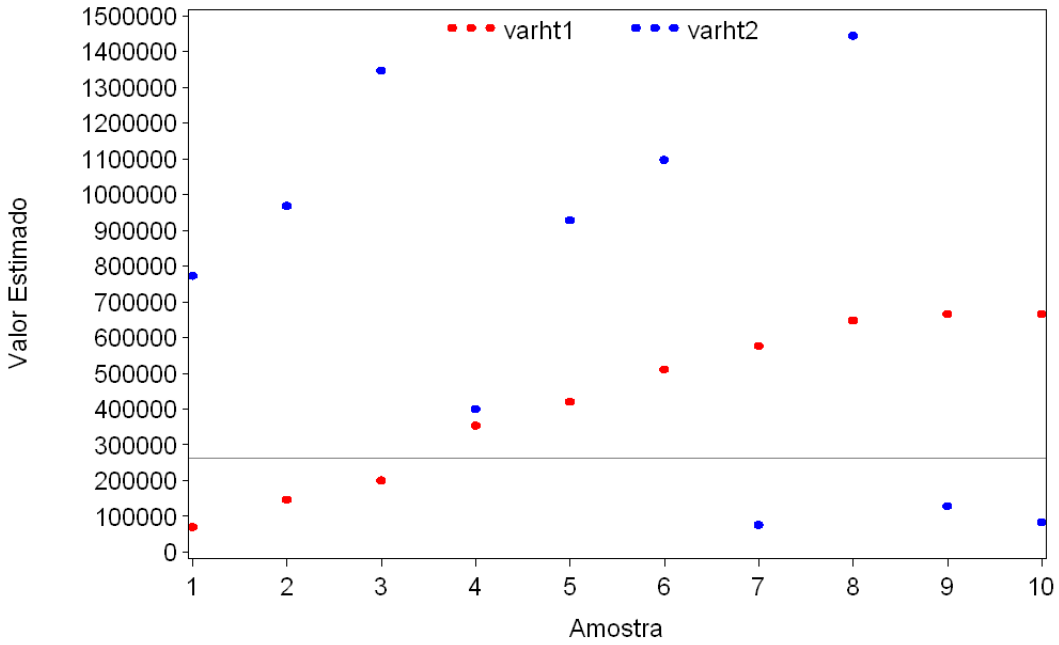

Figura 3.2: Comparação das estimativas de *v1* e *v2* ($N = 6$ e $n = 3$)

Ao contrário do que poderia se imaginar, as estimativas positivas de *v1* podem ser próximas do valor real, em muitos casos inclusive melhores do que as de *v2*. Observou-se que a medida que o tamanho da população cresce, as estimativas de *v1* melhoram. Pode-se notar na Figura 3.1 como as estimativas de *v1* foram distantes da variância real em comparação com *v2*, mas as estimativas foram ficando mais precisas quando o tamanho da população aumentou, como é demonstrado nas Figuras 3.2 e 3.3.

3.4 Intervalos de confiança das estimativas de total do estimador HT

A construção de um intervalo de confiança adequado para as estimativas do estimador Horvitz-Thompson (HT) é o principal objetivo deste trabalho. Para tal,

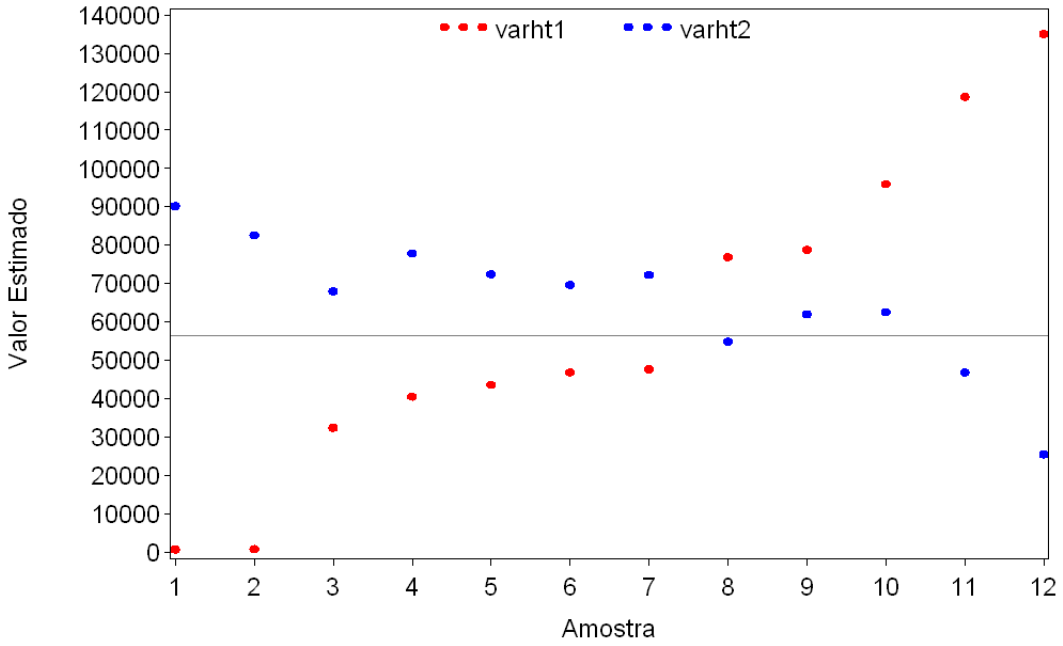

Figura 3.3: Comparação das estimativas de *v1* e *v2* ($N = 7$ e $n = 4$)

buscar-se-á uma distribuição adequada para as estimativas geradas.

Antes, no entanto, serão construídos intervalos de confiança exatos para as estimativas de total HT utilizando probabilidades de inclusão desiguais e probabilidades iguais, caso este em que o estimador HT é igual ao estimador de total da aleatória simples (AAS).

Para tal, faremos uso do fato de o estimador HT estar implementado no software estatístico *SAS* por meio da macro *%HTgeral* como mostrado em Nascimento (2011). Desta forma, podemos obter todas as estimativas possíveis de total e utilizá-las para a construção dos intervalos exatos.

O objetivo é poder visualizar o comportamento dos intervalos. Sabe-se que as estimativas de total dadas pela AAS seguem uma distribuição normal, e portanto, seu intervalo tende a ser simétrico. Por outro lado, não sabemos qual a distribuição

das estimativas de total das pelo estimador HT quando são utilizadas probabilidades desiguais de inclusão e portanto, não sabemos como se comportará o intervalo HT.

Para a construção dos intervalos foram utilizados dados simulados. Foram utilizados os mesmos dados apresentados na seção 3.2, onde a variável Y, representando a população cujo total deseja-se estimar, foi obtida a partir distribuição *qui-quadrado* com 3 graus de liberdade e seus valores multiplicados por 100 . Por outro lado, a variável X, representando a variável auxiliar, foi obtida a partir de uma distribuição *geométrica* com parâmetro 0.3 e seus valores multiplicados por 50.

Uma vez executada a macro *%HTgeral* obteve-se as estimativas de total HT e AAS com suas respectivas probabilidades. Esta macro foi executada para todas as combinações de tamanho de população e amostra possíveis, onde as populações tinham tamanhos entre N=4 e N=7.

De posse das estimativas de total HT e AAS com suas respectivas probabilidades, a construção de intervalos de confiança exatos de coeficiente de confiança $1 - \alpha$ é simples. Define-se o nível de confiança (α) desejado e acumula-se as probabilidades encontradas até $\alpha/2$ e $1 - \alpha/2$ e observa-se qual estimativa de total corresponde a essa probabilidade.

No entanto, temos uma limitação nos dados. O estimador HT é computacionalmente muito intenso e por isso indicado a populações finitas e pequenas e por este motivo foram utilizados apenas dados de $N = 4$ a $N = 7$. Teremos $\binom{N}{n}$ estimativas distintas de totais HT e AAS. Por muitas vezes, o total de estimativas é um número pequeno e as probabilidades individuais são altas. Se as probabilidades individuais

são maiores do que $\alpha/2$, não será obtido um intervalo com coeficiente de confiança $1-\alpha$.

Para contornar-se esse problema foram comparadas as probabilidades individuais e $\alpha = 0.05$, o nível de significância desejado. As estimativas de total HT e AAS foram ordenadas e comparou-se a probabilidade associada a primeira estimativa HT e AAS ordenadas com $\alpha/2$. A maior dessas probabilidades foi considerada como o $\alpha/2$ a ser utilizado. Dessa forma, poderemos acumular as probabilidades para obter o intervalo de confiança que desejamos.

Um problema surge no entanto, pois teremos intervalos de confiança com coeficiente de confiança definidos pelos dados. Assim, para cada combinação de tamanho de população com tamanho de amostra teremos um intervalo de confiança para os dados com coeficiente de confiança distinto, a menos que as probabilidades das estimativas de HT e AAS sejam pequenas o suficiente e nos permita usar $\alpha = 0.05$.

Além desses intervalos de confiança, deseja-se também obter os intervalos da AAS estimada. Assim poderemos construir um gráfico comparando todos os intervalos de confiança obtidos.

Para obter os intervalos da AAS estimada, utilizaremos a distribuição *normal* e a distribuição *t de student*. Essas escolhas se justificam, pois como aborda Cochran (1977) no caso da amostragem aleatória simples sem reposição, temos que:

$$\hat{T} \sim \mathcal{N}(N\overline{y}, N^2(1-f)\frac{S^2}{n}) \tag{3.1}$$

Onde $f = \frac{n}{N}$ e S^2 é a variância populacional considerada sobre $N-1$ ao invés

de N.

Como não será necessário estimar a variância em cada amostra, pois como os dados são simulados sabe-se qual a real variância do total, os intervalos de confiança estimados AAS normal e t serão, respectivamente:

1. $\text{srs} \pm z_{1-\alpha/2} N \sqrt{1-f} \frac{S}{\sqrt{n}}$

2. $\text{srs} \pm t_{(1-\alpha/2,n)} N \sqrt{1-f} \frac{S}{\sqrt{n}}$

Onde n é o tamanho da amostra, N o tamanho da população e *srs* é a estimativa de total.

Vemos que a única diferença entre os intervalos estimados de AAS serão os pontos da distribuição teórica que corresponde aos quantis desejados. Quando a amostra for muito pequena, espera-se que o intervalo utilizando a distribuição t tenha uma maior amplitude, devido a sua cauda mais pesada em relação a distribuição normal. Por outro lado, a partir do momento que tamanho da amostra for crescendo, os intervalos serão mais semelhantes.

A fim de obter um único intervalo AAS estimado normal e um intervalo AAS estimado t para compararmos com os intervalos AAS exato e HT exato, precisamos obter um intervalo que englobe os intervalos obtidos para cada estimativa de total. Teremos $k = \binom{N}{n}$ intervalos distintos utilizando a distribuição normal e outros k utilizando a distribuição t. Para obter um único intervalo nesses dois casos, basta buscarmos entre os k intervalos o menor limite inferior e o maior limite superior e juntar essas informações para construir um novo intervalo. Este novo intervalo englobará todos os outros.

As Figuras 3.4, 3.5 e 3.6 mostram os intervalos obtidos utilizando as estimativas obtidas quando $N = 6$ e $n = 3$, $N = 7$ e $n = 3$, $N = 7$ e $n = 4$. A linha vertical representa o verdadeiro valor do total populacional.

Pode-se notar que todos os intervalos contém o verdadeiro total populacional, portanto o coeficiente de confiança utilizado foi o bastante para que os intervalos obtidos conseguissem capturar o verdadeiro valor estimado.

Outra característica interessante é o fato dos intervalos baseados na AAS serem simétricos e em geral com o verdadeiro total populacional no centro do intervalo. O estimador HT apresenta uma assimetria em seus intervalos e nem sempre o verdadeiro valor populacional está no centro do intervalo.

Para a construção dos intervalos foram utilizados dados que resultavam em estimativas HT com menor variância do que as estimativas AAS, fazendo uso do que foi abordado na seção 3.2. Por este motivo os intervalos de HT possuem uma menor amplitude.

Obtidos os intervalos, deseja-se agora fazer com o intervalo HT o mesmo que foi feito com os intervalos AAS. Deseja-se obter intervalos estimados, utilizando uma distribuição adequada para as estimativas de total HT.

3.4.1 Intervalo de confiança HT estimado

Vimos anteriormente que a utilização das distribuições normal e t na construção dos intervalos AAS estimados foi justificada pelo apresentado na Equação 3.1. No entanto, não sabemos qual é a distribuição das estimativas de total dadas pelo estimador HT.

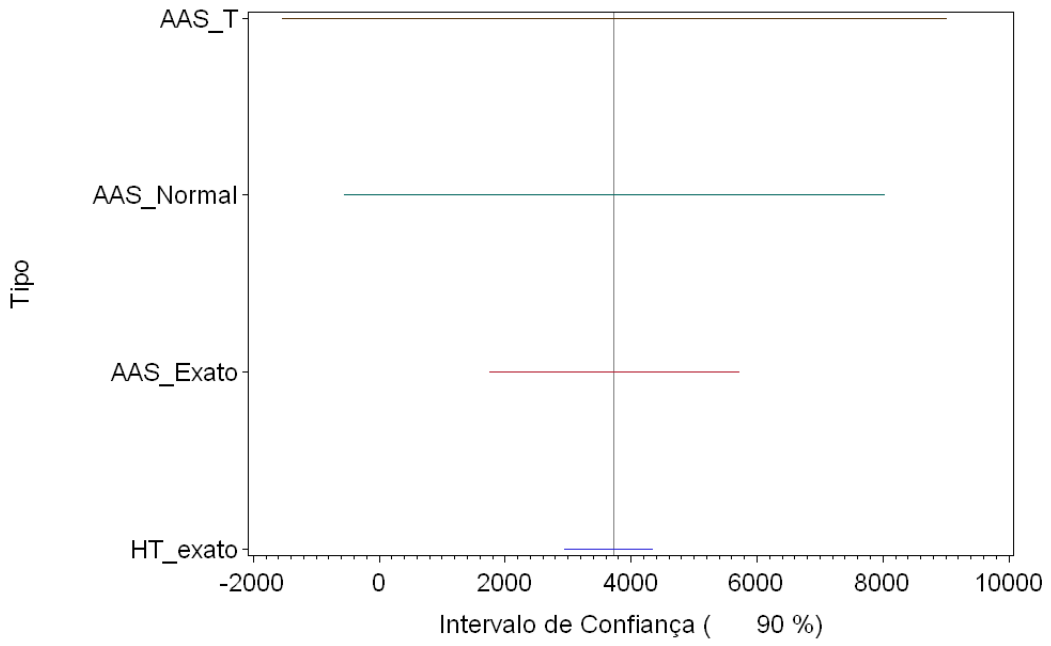

Figura 3.4: Intervalos de confiança (90%) para $N = 6$ e $n = 3$

Para termos uma idéia da distribuição das estimativas de total HT, foram construídos gráficos *box-plots* das estimativas ponderadas geradas utilizando os resultados das simulações que são apresentados nos apêndices. Foram criados *box-plots* para as estimativas de total geradas de populações de tamanho $N = 7$ com todos os tamanhos possíveis de amostra para que pudesse ser avaliado se as características descritivas da distribuição buscada variam de acordo com uma mudança no tamanho da amostra.

Como pode ser visto em Apêndices A.2, onde são apresentados os gráficos *box-plot* obtidos, buscamos uma distribuição que pode apresentar assimetria tanto a direita quanto a esquerda.

A Tabela 3.5 apresenta a assimetria das estimativas de HT considerando e não considerando suas probabilidades associadas. A utilização das probabilidades as-

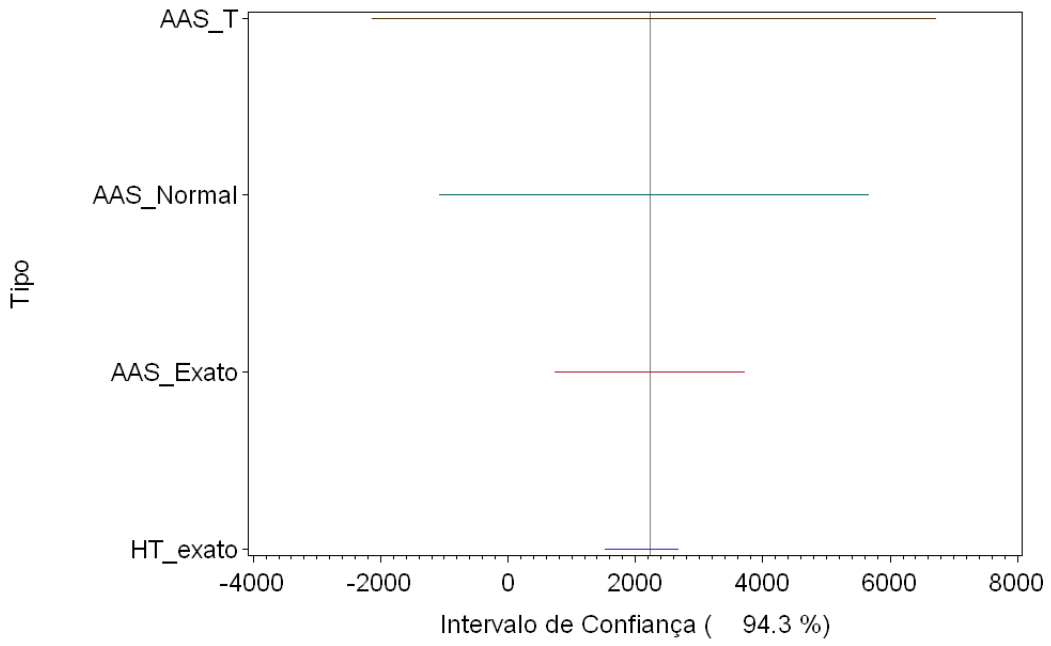

Figura 3.5: Intervalos de confiança (94.3%) para $N = 7$ e $n = 3$

sociadas a cada estimativa gera dúvidas, uma vez que em uma situação real não saberemos as probabilidade associadas a cada estimativa. As assimetrias foram obtidas analisando as estimativas de total HT geradas por dados simulados obtidos da mesma forma descrita na seção 3.3.

Apesar da presença de assimetrias, é importante recordar que o estimador HT é um estimador geral de total, ou seja, com uma escolha adequada do vetor de probabilidades de seleção, pode-se obter estimativas de amostragem com probabilidades iguais de seleção como as da aleatória simples. Sabe-se que as estimativas geradas pela AAS tendem a ser simétricas, pois como vimos estas seguem uma distribuição normal. Portanto, a distribuição buscada deverá ter a capacidade de se adequar dependendo das probabilidades de seleção utilizadas.

Uma distribuição que surge nesse contexto é a Normal Assimétrica (*Skew Nor-*

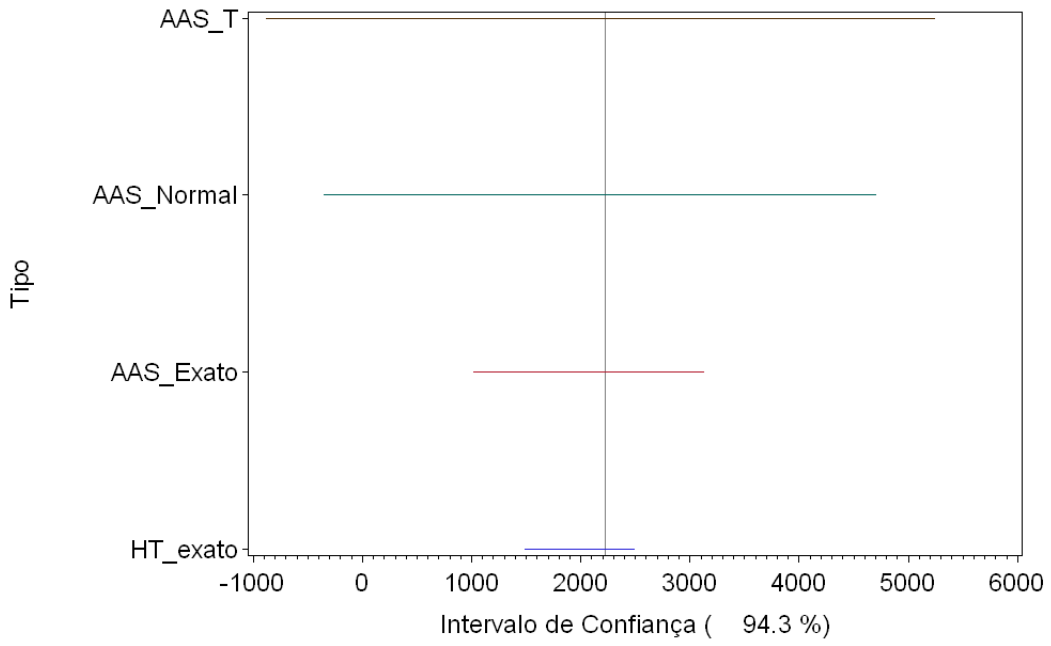

Figura 3.6: Intervalos de confiança (94.3%) para $N = 7$ e $n = 4$

mal) proposta por Azzalini (1985). De acordo com o autor, a distribuição *Skew Normal* é definida como:

Definição Se uma variável aleatória Z tem função densidade

$$\phi(z;\lambda) = 2\phi(z)\Phi(\lambda z) \qquad (-\infty < z < \infty) \tag{3.2}$$

onde ϕ e Φ são a densidade da normal padrão e a função de distribuição acumulada da normal padrão, respectivamente, então dizemos que Z tem distribuição Normal Assimétrica com parâmetro λ.

Quando $\lambda = 0$, Z tem distribuição Normal Padrão. Podem ser adicionados parâmetros de locação e escala por meio da transformação linear

$$Y = \xi + \omega Z$$

Tabela 3.5: Assimetria apresentada pelas estimativas de total HT

Sem pesos	Com pesos	Dados
0.00	-1.16	N=4, n=2
-1.08	-0.19	N=4, n=3
-0.15	0.18	N=5, n=2
-0.05	-0.23	N=5, n=3
-0.75	-1.76	N=5, n=4
-0.41	-0.32	N=6, n=2
0.00	-0.70	N=6, n=3
-0.32	-0.33	N=6, n=4
-1.60	0.91	N=6, n=5
0.44	-0.06	N=7, n=2
0.20	-0.46	N=7, n=3
-0.24	-0.58	N=7, n=4
-0.84	-0.84	N=7, n=5
-2.13	-1.45	N=7, n=6

De forma análoga a utilização da distribuição t no caso da AAS, utilizaremos também a distribuição t Assimétrica (*Skew t*) na construção de intervalos de confiança para as estimativas HT. Essa escolha se deve principalmente pelo fato de muitas vezes a quantidade de estimativas ser pequena. A forma de construção da distribuição *Skew t* é apresentada em Azzalini and Capitanio (2003).

As distribuições *Skew Normal* e *Skew t* parecem atender ao que precisamos. Elas tem a capacidade de ser simétrica ou assimétrica dependendo do ajuste dos dados.

Os intervalos de confiança das estimativas HT serão obtidos por

1. limite inferior = hts + $z_{\alpha/2}\sqrt{var(HT)}$

2. limite superior = hts + $z_{(1-\alpha/2)}\sqrt{var(HT)}$

no caso de um ajuste pela distribuição *Skew Normal* e

1. limite inferior = hts + $t_{(\alpha/2,n)}\sqrt{var(HT)}$

2. limite superior = hts + $t_{((1-\alpha/2),n)}\sqrt{var(HT)}$

no caso de um ajuste pela distribuição *Skew t*, onde *hts* é a estimativa de total dada pelo estimador HT e *var(HT)* é a variância conhecida do estimador.

Quando o parâmetro de assimetria, λ, for diferente de 0 teremos que $z_{\alpha/2} \neq -|z_{(1-\alpha/2)}|$, e portanto, teremos um intervalo assimétrico.

Portanto, como temos as estimativas de total (hts), os valores de α a serem utilizados serão os mesmos apresentados na seção 3.4 e conhecemos a variância do estimador HT, para obtermos estes intervalos precisamos apenas dos valores de z e t que correspondem aos quantis desejados.

Para obter tais valores, é necessário ajustar as estimativas de total HT às distribuições *Skew Normal* e *Skew t* para assim obter estimativas para o parâmetro λ.

Para realizar tal ajuste foi utilizado o software *R*, onde as distribuições aqui utilizadas estão implementadas no pacote *sn*. Obtidas as estimativas de λ, pode-se obter os pontos da distribuição que correspondem ao quantil desejado.

Para cada conjunto de dados, teremos $k = \binom{N}{n}$ intervalos distintos utilizando a distribuição *Skew Normal* e outros k utilizando a distribuição *Skew t*. A fim de se obter um único intervalo para cada distribuição para que possam ser comparados com os intervalos obtidos anteriormente, de forma análoga ao que foi feito no caso dos intervalos estimados da AAS, utilizaremos o menor limite inferior e o maior limite superior destes k distintos intervalos para assim termos um intervalo que englobe todos os outros.

As Figuras 3.7 e 3.8 mostram os intervalos obtidos para as estimativas de total HT quando $(N = 6, n = 3)$ e $(N = 7, n = 3)$, onde foram utilizadas as estimativas ponderadas de total para estimar os parâmetros de assimetria das distribuições assimétricas apresentadas. A linha vertical representa o verdadeiro valor do total populacional.

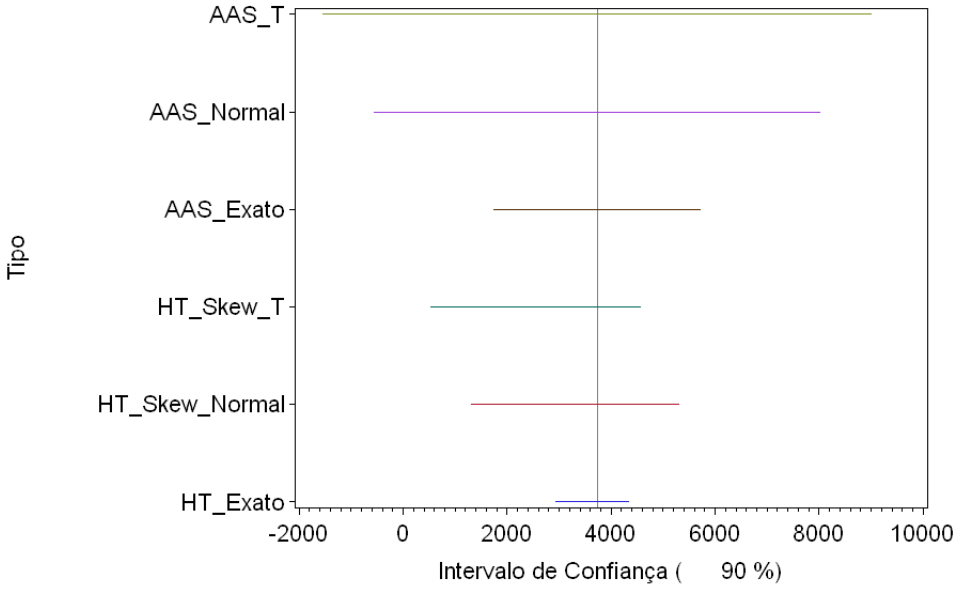

Figura 3.7: Intervalos de confiança (90%) para $N = 6$ e $n = 3$

Pode-se notar que aparentemente a assimetria do intervalo HT foi bem representada nos intervalos estimados. Todos os intervalos obtidos contêm o verdadeiro valor de total.

Para construção dos intervalos apresentados nas Figuras 3.7 e 3.8 foram utilizadas as estimativas ponderadas dos totais HT. As Figuras 3.9 e 3.10 mostram a diferença de intervalos obtidos quando os dados foram ajustados com os pesos e sem os pesos, respectivamente, no caso $(N = 4, n = 2)$.

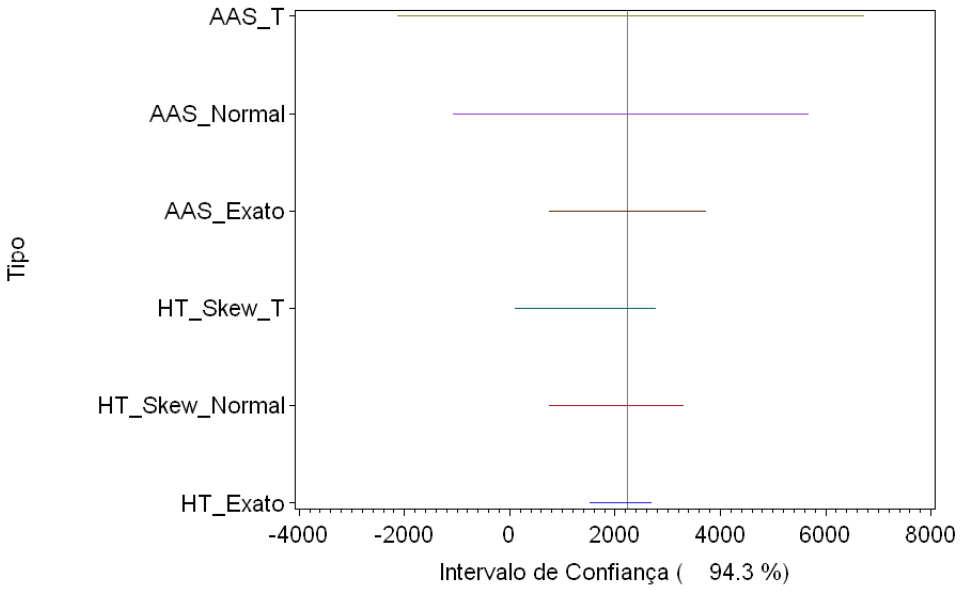

Figura 3.8: Intervalos de confiança (94.3%) para $N = 7$ e $n = 3$

A partir da Tabela 3.5 podemos ver que no caso das estimativas sem peso, a assimetria dos dados era 0 e a das estimativas com peso era de -1.163 para os dados que consideraram $(N = 4, n = 2)$. Apesar dessa diferença, os intervalos obtidos utilizando a distribuição *Skew Normal* foram próximos, enquanto a distribuição *Skew t* captou melhor a assimetria.

Em geral, esse resultado foi consistente: a utilização dos pesos na estimação dos parâmetros das distribuições assimétricas influenciou mais os resultados da distribuição *Skew t*.

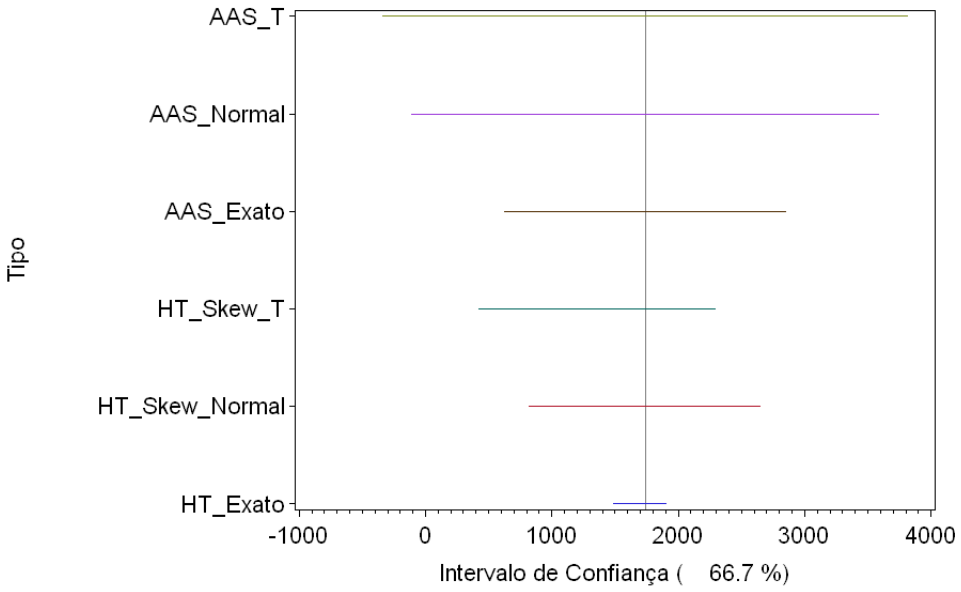

Figura 3.9: Intervalos de confiança (66.7%) para $N = 4$ e $n = 2$, utilizando dados ponderados na construção dos intervalos assimétricos

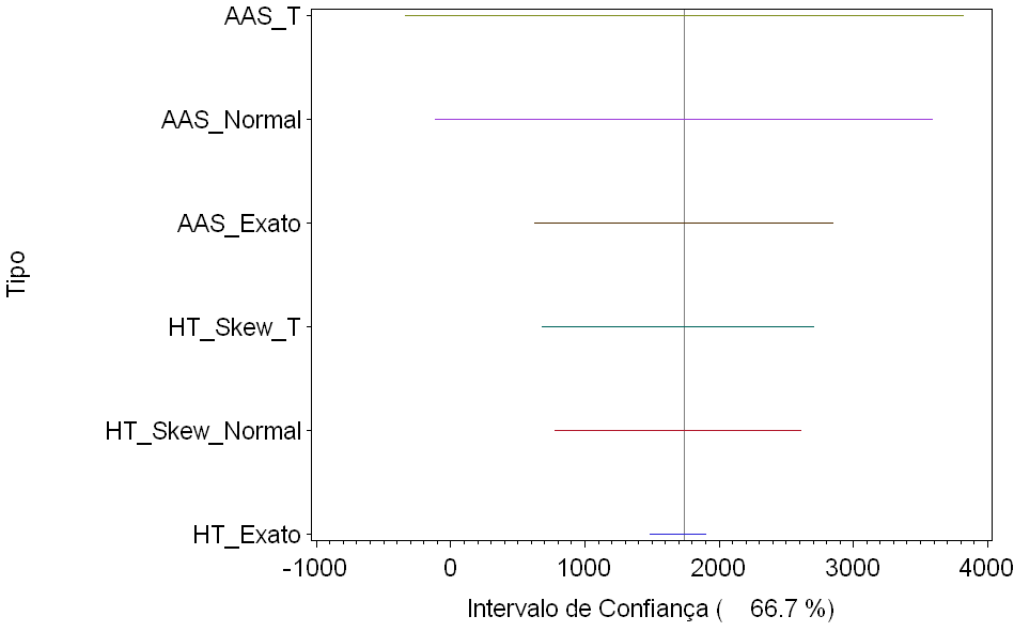

Figura 3.10: Intervalos de confiança (66.7%) para $N = 4$ e $n = 2$, utilizando dados sem ponderação na construção dos intervalos assimétrico

Capítulo 4

Conclusão

O estimador Horvitz-Thompson (HT) é um estimador com grande aplicabilidade devido ao fato de ser um estimador geral de total, onde são possíveis serem usadas diferentes probabilidades de seleção a fim de retratar da melhor maneira os dados disponíveis. Apesar de tamanho poder, o estimador apresenta alguns problemas como condições desconhecidas para que seja mais eficiente em relação aos métodos que fazem uso de probabilidades iguais de seleção, estimativas de variância negativas e ausência de um intervalo de confiança adequado para suas estimativas.

A respeito da eficiência do estimador HT frente ao estimador de total da Aleatória Simples (seção 3.2), foram mostradas evidências de que se for utilizada uma variável auxiliar X para obtenção das probabilidades de inclusão de cada observação de Y (variável cujo total deseja-se estimar), então para que o estimador HT seja mais eficiente do que o estimador AAS, Y e X devem ser fortemente correlacionados linearmente e além disso seus coeficientes de variação devem ser semelhantes. Por outro lado, o estimador HT tem seu pior resultado frente ao estimador AAS quando o coeficiente de variação de X é maior do que o coeficiente de variação de Y.

A tentativa de encontrar fatores individuais que influenciassem estimativas ne-

gativas de $v1$, estimador de variância do estimador HT (seção 3.3) não apresentou resultados expressivos. No entanto, observou-se que a utilização de uma variável auxiliar para a obtenção das probabilidades de inclusão resultou no fato de sempre o termo $\pi_{ij} - \pi_i \pi_j$ ser negativo. Este fato determina que $v1$ tem sempre a possibilidade de gerar estimativas negativas, enquanto $v2$ nunca poderá assumir valores negativos já que em sua fórmula as probabilidades de inclusão são inseridas na fórmula por meio de $\pi_i \pi_j - \pi_{ij}$.

Os intervalos das estimativas de total Horvitz-Thompson são construídos (seção 3.4) fazendo uso de simulações para obter todas as estimativas de total possíveis. Pode-se observar que a presença de assimetrias nos intervalos HT obtidos foram comuns, o que torna a utilização das distribuições normal e t duvidosas na construção de intervalos estimados. São propostas duas distribuições assimétricas (*Skew Normal* e *Skew t*) para a construção dos intervalos estimados, pois estas distribuições são capazes de se ajustar a assimetria dos dados, sendo assim distribuições mais gerais.

É necessário um estudo mais aprofundado sobre como a utilização de uma variável auxiliar na obtenção das probabilidades de inclusão pode garantir que $\pi_{ij} - \pi_i \pi_j < 0$, já que como os termos π_{ij}, π_i e π_j variam entre 0 e 1 existem possibilidades de valores que façam este termo ser positivo.

Um aspecto a ser explorado futuramente é como estimar o parâmetro de assimetria das distribuições assimétricas propostas utilizando apenas as informações disponíveis em uma situação real (probabilidades de inclusão e n observações de y_i, a variável cujo total deseja-se estimar).

A cobertura dos intervalos obtidos fazendo uso de distribuições assimétricas, como também a validade do ajuste das estimativas de total HT por estas distribuições precisam ser estudados.

Referências Bibliográficas

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 171–178.

Azzalini, A. & Capitanio, A. (2003). Distributions generated by pertubation of symmetry with emphasis on a multivariate skew t distribution. *J. Roy. Statist. Soc.*, pages 367–389.

Bickel, P. J. & Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice-Hall.

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

Brewer, K. W. R. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, pages 5–13.

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

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

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

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

Midha, C. (1988). The horvitz-thompson estimator and estimates of its variance. *Annual Meeting of the American Statistical Association*, (082).

Nascimento, I. F. (2011). Implementação e aplicação do estimador horvitz-thompson no software sas. Technical report, Departamento de Estatística - Universidade de Brasília. p. 15-18.

Sen, A. R. (1953). On the estimate of variance in sampling with varying probabilities. *Journal Indian Society of Agricultural Statistics*. p. 119-127.

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

Yates, F. & Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Royal Statistical Society.* p. 253-261.

Apêndice A

Simulações

A.1 Medidas descritivas e efeitos de planejamento obtidos

A.1.1 População qui-quadrado(3)*100

Tabela A.1: Medidas descritivas da população

Medida	pop4	pop5	pop6	pop7	pop8
Média	434.45	222.17	622.27	318.16	228.37
Desvio Padrão	382.90	120.85	569.87	312.38	151.14
Coeficiente Variação	88.14	54.40	91.58	98.18	66.18
Desvio Interquantílico	554.85	157.35	269.77	246.24	267.36
Assimetria	1.30	0.53	1.74	2.05	0.70
Mínimo	147.40	91.56	51.28	86.05	91.75
Máximo	968.60	384.44	1711.52	985.60	460.80

Tabela A.2: Medidas descritivas da variável auxiliar gerada pela distribuição geométrica(0,3)*50

Medida	pop4	pop5	pop6	pop7	pop8
Correlação	0.77	0.94	0.83	0.97	0.95
Média	437.50	230.00	200.00	278.57	381.25
Desvio Padrão	311.92	152.48	130.38	328.96	277.67
Coeficiente Variação	71.30	66.29	65.19	118.08	72.83

Tabela A.2: Medidas descritivas da variável auxiliar gerada pela distribuição geométrica(0,3)*50

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Desvio Interquantílico	375.00	150.00	200.00	400.00	450.00
Mínimo	250.00	50.00	50.00	50.00	50.00
Máximo	900.00	450.00	350.00	950.00	800.00

Tabela A.3: Efeitos de planejamento

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	17.33%	2.22%	12.97%	10.92%	22.44%
3	12.43%	3.82%	13.53%	11.33%	23.25%
4		9.26%	14.35%	11.01%	23.51%
5			15.26%	9.42%	21.65%
6				6.98%	15.64%
7					5.85%

Tabela A.4: Medidas descritivas da variável auxiliar gerada pela distribuição binomial negativa(0.4,9)*50

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.99	0.96	0.83	0.86	0.93
Média	875.00	700.00	733.33	642.86	700.00
Desvio Padrão	396.86	285.04	216.02	240.53	268.59
Coeficiente Variação	45.35	40.72	29.46	37.42	38.37
Desvio Interquantílico	600.00	350.00	150.00	450.00	450.00
Mínimo	500.00	300.00	350.00	350.00	300.00
Máximo	1400.00	1000.00	1000.00	1000.00	1050.00

Tabela A.5: Efeitos de planejamento

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	29.45%	15.67%	50.05%	37.12%	25.20%
3	31.14%	22.32%	51.37%	37.54%	28.45%
4		29.18%	52.65%	37.19%	31.77%
5			54.55%	35.31%	34.59%
6				30.15%	35.62%
7					32.36%

Tabela A.6: Medidas descritivas da variável auxiliar gerada pela distribuição binomial(300,0.54)

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.99	0.92	0.82	0.71	0.82
Média	163.50	159.80	167.00	160.28	162.12
Desvio Padrão	7.94	13.42	4.86	6.42	9.73
Coeficiente Variação	4.85	8.40	2.91	4.01	6.00
Desvio Interquantílico	12.00	16.00	7.00	12.00	6.50
Mínimo	156.00	146.00	160.00	150.00	145.00
Máximo	174.00	179.00	173.00	167.00	180.00

Tabela A.7: Efeitos de planejamento

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	87.99%	75.70%	93.60%	92.46%	84.69%
3	88.70%	79.43%	93.86%	92.65%	85.55%
4		83.81%	94.12%	92.80%	86.48%
5			94.28%	92.88%	87.46%
6				92.74%	88.49%
7					89.45%

A.1.2 População qui-quadrado(580)*10

Tabela A.8: Medidas descritivas da população

Medida	pop4	pop5	pop6	pop7	pop8
Média	5830.81	5728.25	5917.32	5853.17	5712.09
Desvio Padrão	307.17	342.48	392.35	340.89	333.70
Coeficiente Variação	5.27	5.98	6.63	5.82	5.84
Desvio Interquantílico	495.84	285.90	213.82	681.21	472.95
Assimetria	0.61	1.32	−0.97	0.45	1.14
Mínimo	5562.75	5409.22	5228.04	5423.20	5409.91
Máximo	6213.62	6281.67	6425.04	6370.75	6358.53

Tabela A.9: Medidas descritivas da variável auxiliar gerada pela distribuição geométrica(0.3)*50

Medida	pop4	pop5	pop6	pop7	pop8
Correlação	0.97	0.99	0.77	0.90	0.89
Média	225.00	200.00	150.00	114.29	250.00
Desvio Padrão	155.46	127.47	134.16	74.80	148.80
Coeficiente Variação	69.09	63.74	89.44	65.45	59.52
Desvio Interquantílico	250.00	150.00	150.00	100.00	250.00
Mínimo	50.00	100.00	50.00	50.00	50.00
Máximo	400.00	400.00	400.00	250.00	450.00

Tabela A.10: Efeitos de planejamento

amostra	pop4	pop5	pop6	pop7	pop8
2	30721.10%	5642.70%	9822.40%	9687.64%	25429.60%
3	21146.20%	2303.48%	5091.68%	7058.71%	24571.90%
4		399.79%	2308.52%	4276.73%	23248.20%
5			445.98%	1589.62%	20938.00%
6				223.59%	15913.60%
7					1297.51%

Tabela A.11: Medidas descritivas da variável auxiliar gerada pela distribuição binomial negativa(0.4,9)*50

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.99	0.90	0.99	0.92	0.89
Média	875.00	770.00	733.33	700.00	731.25
Desvio Padrão	396.86	189.08	216.02	322.75	249.19
Coeficiente Variação	45.36	24.56	29.46	46.11	34.08
Desvio Interquantílico	600.00	300.00	150.00	700.00	375.00
Mínimo	500.00	600.00	350.00	400.00	300.00
Máximo	1400.00	1000.00	1000.00	1100.00	1050.00

Tabela A.12: Efeitos de planejamento

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	3187.34%	620.77%	1662.25%	4205.33%	3676.51%
3	922.39%	265.25%	1378.94%	2981.57%	3312.08%
4		40.16%	1029.37%	1701.51%	2909.75%
5			527.58%	649.87%	2455.69%
6				89.59%	1916.70%
7					1182.36%

Tabela A.13: Medidas descritivas da variável auxiliar gerada pela distribuição binomial(300,0.54)

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.99	0.94	0.91	0.92	0.88
Média	163.50	162.00	167.00	162.00	159.75
Desvio Padrão	7.94	3.81	4.86	8.91	10.11
Coeficiente Variação	4.85	2.35	2.91	5.50	6.33
Desvio Interquantílico	12.00	4.00	7.00	21.00	7.00
Mínimo	156.00	157.00	160.00	150.00	145.00

Tabela A.13: Medidas descritivas da variável auxiliar gerada pela distribuição binomial(300,0.54)

Medida	pop4	pop5	pop6	pop7	pop8
Máximo	174.00	167.00	173.00	171.00	180.00

Tabela A.14: Efeitos de planejamento

amostra	pop4	pop5	pop6	pop7	pop8
2	9.93%	47.29%	45.29%	14.52%	23.42%
3	31.77%	56.64%	52.40%	16.90%	22.33%
4		70.37%	62.18%	22.22%	23.28%
5			76.98%	32.46%	27.10%
6				52.18%	35.25%
7					50.76%

A.1.3 População binomial negativa(0.7,9)*100

Tabela A.15: Medidas descritivas da população

Medida	pop4	pop5	pop6	pop7	pop8
Média	425.00	480.00	416.67	428.57	262.50
Desvio Padrão	287.23	216.79	271.42	249.76	226.38
Coeficiente Variação	67.58	45.17	65.14	58.28	86.24
Desvio Interquantílico	350.00	100.00	500.00	500.00	400.00
Assimetria	0.52	0.42	0.73	0.35	0.23
Mínimo	100.00	200.00	200.00	100.00	0.00
Máximo	800.00	800.00	800.00	800.00	600.00

Tabela A.16: Medidas descritivas da variável auxiliar gerada pela distribuição geométrica(0.18)*50

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Corr	0.95	0.94	0.95	0.88	0.97
Média	425.00	170.00	233.33	278.57	406.25
Desvio Padrão	332.92	144.05	177.95	328.96	312.18
Coeficiente Variação	78.33	84.73	76.26	118.08	76.84
Desvio Interquantílico	450.00	150.00	250.00	400.00	525.00
Mínimo	150.00	50.00	50.00	50.00	50.00
Máximo	900.00	400.00	500.00	950.00	850.00

Tabela A.17: Efeitos de planejamento

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	18.25%	103.38%	17.72%	77.32%	14.94%
3	29.55%	66.25%	14.58%	38.57%	17.92%
4		45.78%	14.36%	15.97%	21.97%
5			11.13%	15.53%	24.78%
6				20.29%	21.54%
7					13.39%

Tabela A.18: Medidas descritivas da variável auxiliar gerada pela distribuição binomial negativa(0.1,12)*50

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.94	0.89	0.86	0.96	0.96
Média	4725.00	3720.00	6375.00	5914.28	6381.25
Desvio Padrão	998.75	1620.42	1790.74	1687.88	2913.02
Coeficiente Variação	21.14	43.56	28.09	28.54	45.65
Desvio Interquantílico	1450.00	950.00	2300.00	3150.00	3925.00
Mínimo	3850.00	2600.00	3500.00	3100.00	3350.00
Máximo	6100.00	6500.00	8350.00	7900.00	11650.00

Tabela A.19: Efeitos de planejamento

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	58.19%	29.74%	43.63%	34.28%	37.96%
3	67.01%	38.13%	46.78%	39.52%	42.51%
4		48.96%	49.01%	46.01%	47.06%
5			48.42%	54.27%	50.79%
6				64.18%	52.35%
7					49.53%

Tabela A.20: Medidas descritivas da variável auxiliar gerada pela distribuição binomial (300,0.54)

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.90	0.93	0.68	0.94	0.94
Média	163.00	164.40	156.17	163.00	162.62
Desvio Padrão	6.68	8.62	11.96	8.91	4.27
Coeficiente Variação	4.10	5.24	7.66	5.46	2.63
Desvio Interquantílico	7.00	6.00	21.00	17.00	6.50
Mínimo	159.00	158.00	138.00	150.00	156.00
Máximo	173.00	179.00	165.00	174.00	168.00

Tabela A.21: Efeitos de planejamento

amostra	pop4	pop5	pop6	pop7	pop8
2	88.74%	80.34%	84.91%	84.69%	94.85%
3	89.93%	83.01%	86.31%	86.38%	95.35%
4		86.09%	87.97%	88.34%	95.91%
5			89.93%	90.69%	96.57%
6				93.64%	97.34%
7					98.30%

A.1.4 População binomial negativa(0.1,300)

Tabela A.22: Medidas descritivas da população

Medida	pop4	pop5	pop6	pop7	pop8
Média	3154.75	3106.20	3155.00	3106.57	3211.25
Desvio Padrão	239.75	251.50	187.12	213.74	184.47
Coeficiente Variação	7.60	8.10	5.93	6.88	5.74
Desvio Interquantílico	369.50	135.00	103.00	188.00	294.00
Assimetria	−0.53	1.78	−0.31	1.42	−0.17
Mínimo	2852.00	2904.00	2851.00	2876.00	2953.00
Máximo	3403.00	3537.00	3431.00	3528.00	3480.00

Tabela A.23: Medidas descritivas da variável auxiliar gerada pela distribuição geométrica(0.18)*50

Medida	pop4	pop5	pop6	pop7	pop8
Correlação	0.79	0.98	0.93	0.96	0.96
Média	437.50	230.00	200.00	278.57	381.25
Desvio Padrão	311.92	152.48	130.38	328.96	277.67
Coeficiente Variação	71.29	66.29	65.19	118.08	72.83
Desvio Interquantílico	375.00	150.00	200.00	400.00	450.00
Mínimo	250.00	50.00	50.00	50.00	50.00
Máximo	900.00	450.00	350.00	950.00	800.00

Tabela A.24: Efeitos de planejamento

amostra	pop4	pop5	pop6	pop7	pop8
2	81.36%	1547.47%	1826.70%	2990.25%	2954.02%
3	13.40%	1301.94%	1334.42%	1705.66%	2633.00%
4		1008.27%	564.04%	701.27%	2190.34%
5			188.64%	267.96%	1569.01%

Tabela A.24: Efeitos de planejamento

amostra	pop4	pop5	pop6	pop7	pop8
6				47.26%	1114.71%
7					770.87%

Tabela A.25: Medidas descritivas da variável auxiliar gerada pela distribuição binomial negativa(0.1,12)*50

Medida	pop4	pop5	pop6	pop7	pop8
Correlação	0.95	0.99	0.89	0.84	0.93
Média	4850.00	6060.00	6150.00	5257.14	6712.50
Desvio Padrão	1041.63	1502.66	1926.91	1477.45	2756.52
Coeficiente Variação	21.48	24.80	31.33	28.10	41.06
Desvio Interquantílico	1700.00	950.00	3600.00	2950.00	3525.00
Mínimo	3850.00	4650.00	3500.00	3400.00	3350.00
Máximo	6100.00	8550.00	8350.00	7000.00	11650.00

Tabela A.26: Efeitos de planejamento

amostra	pop4	pop5	pop6	pop7	pop8
2	150.65%	194.68%	1770.44%	1023.26%	3125.69%
3	16.56%	69.01%	1240.30%	778.12%	2385.35%
4		10.97%	697.37%	521.22%	1700.28%
5			190.29%	267.15%	1090.40%
6				65.45%	572.75%
7					175.67%

Tabela A.27: Medidas descritivas da variável auxiliar gerada pela distribuição binomial(300,0.54)

Medida	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
Correlação	0.89	0.82	0.96	0.95	0.99
Média	150.25	157.00	162.83	161.57	161.87
Desvio Padrão	8.22	5.34	12.22	11.80	4.45
Coeficiente Variação	5.47	3.40	7.50	7.30	2.75
Desvio Interquantílico	9.50	9.00	15.00	18.00	8.00
Mínimo	138.00	151.00	144.00	146.00	156.00
Máximo	155.00	163.00	179.00	181.00	168.00

Tabela A.28: Efeitos de planejamento

amostra	**pop4**	**pop5**	**pop6**	**pop7**	**pop8**
2	30.49%	52.61%	10.99%	9.56%	32.35%
3	49.78%	59.97%	8.37%	10.60%	37.34%
4		71.27%	13.47%	14.97%	43.45%
5			34.87%	24.47%	51.14%
6				42.81%	61.26%
7					75.68%

A.2 Box-Plots das estimativas de total HT

A.2.1 População qui-quadrado(3)*100

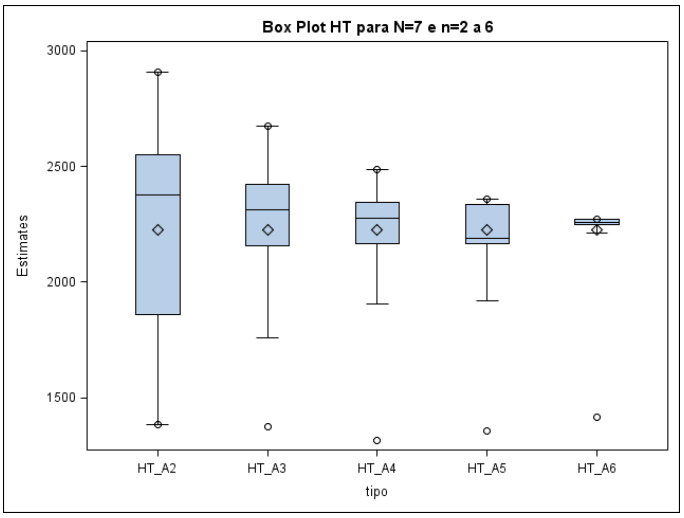

Figura A.1: Auxiliar: geométrica(0.3)*50

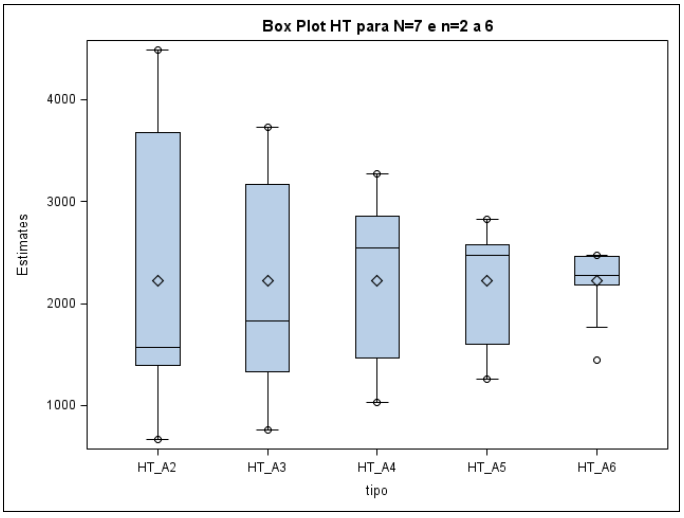

Figura A.2: Auxiliar: binomial(300,0.54)

A.2.2 População qui-quadrado(580)*10

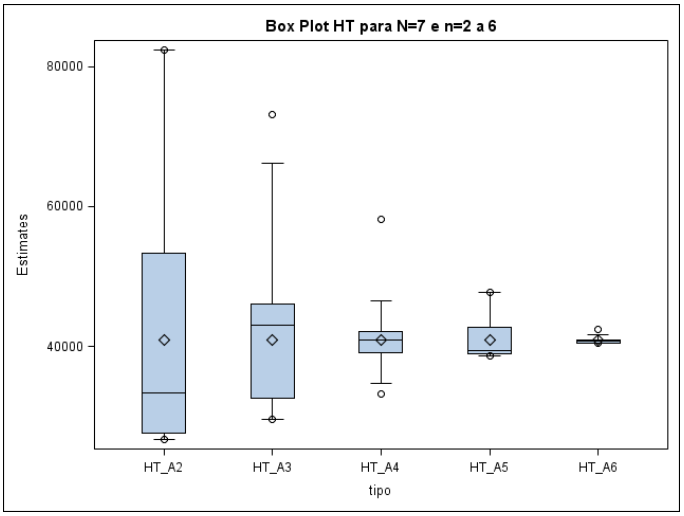

Figura A.3: Auxiliar: geométrica(0.3)*50

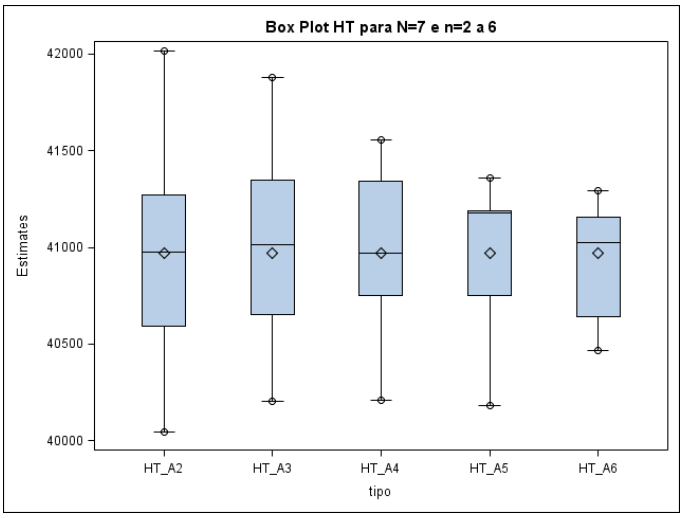

Figura A.4: Auxiliar: binomial(300,0.54)

A.2.3 População binomial negativa(0.7,9)*100

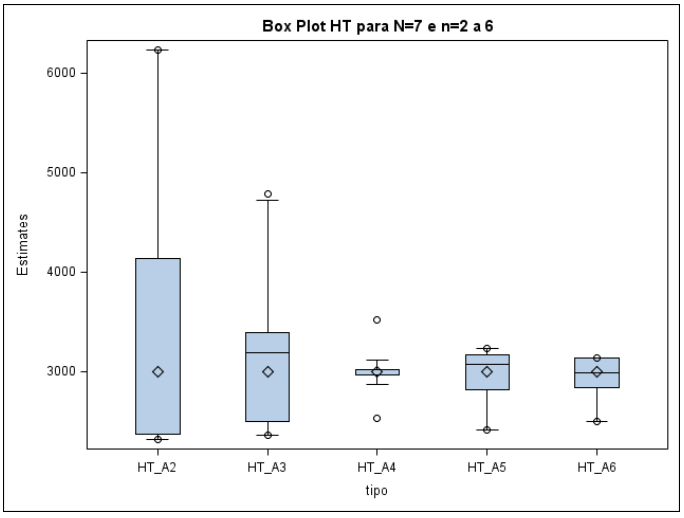

Figura A.5: Auxiliar: geométrica(0.18)*50

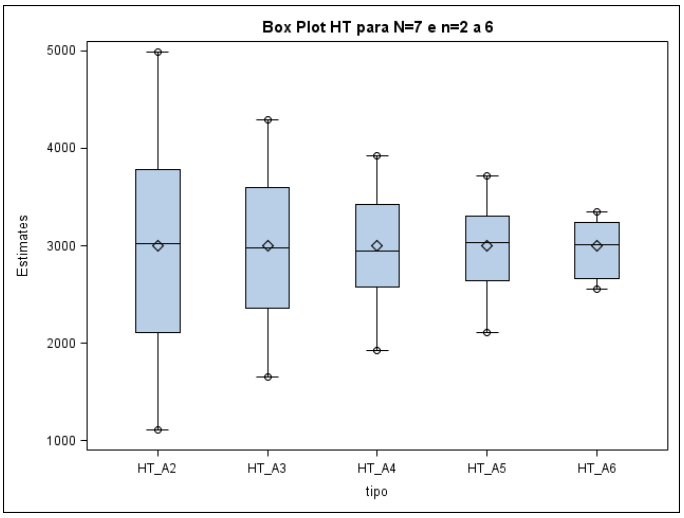

Figura A.6: Auxiliar: binomial(300,0.54)

A.2.4 População binomial negativa(0.1,350)

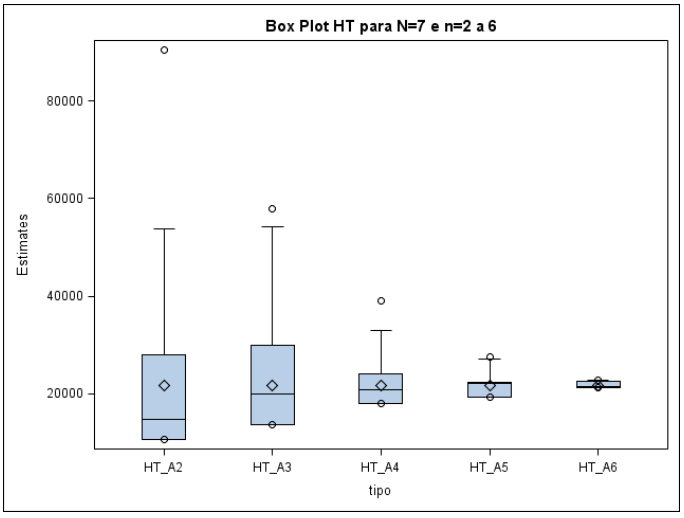

Figura A.7: Auxiliar: geométrica(0.18)*50

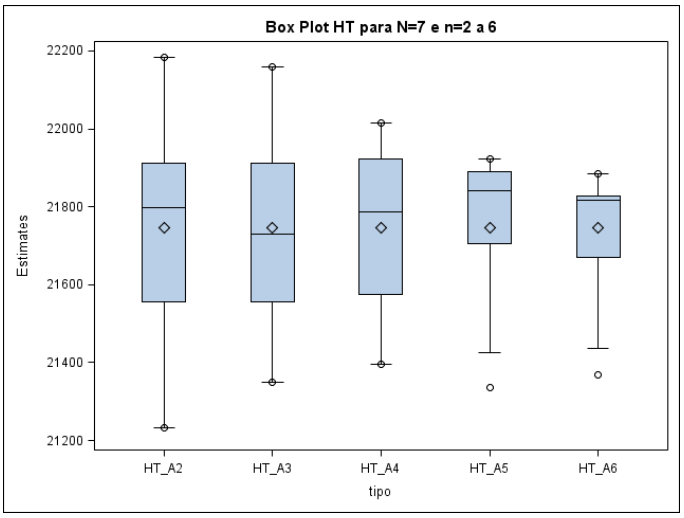

Figura A.8: Auxiliar: binomial(300,0.54)