Capítulo 10 Amostragem com Probabilidades Proporcionais ao Tamanho

10.1 Justificativa para amostragem PPT

Como já indicado no capítulo 3, a *Amostragem Probabilística* contempla o emprego de métodos de amostragem que levam a ter probabilidades de inclusão na amostra desiguais, isto é, $\pi_i \neq \pi_j$ para algum par de unidades distintas $i \neq j \in U$. A teoria apresentada no capítulo 3 cobre o caso geral, mas não indica quando o recurso de usar probabilidades desiguais de inclusão na amostra seria vantajoso. Neste capítulo tratamos justamente dessa questão, indicando estratégias que se pode usar para tirar proveito do emprego de amostragem com probabilidades desiguais para aumentar a eficiência dos estimadores de totais e médias.

A ideia central já mencionada na seção **??** é fazer com que as probabilidades de inclusão fiquem proporcionais aos valores da(s) variável(is) de estudo y. É claro que não podemos conseguir isso de forma exata, já que os valores da variável y são desconhecidos antes da seleção da amostra. Mas em muitas situações práticas é possível contar com cadastros que contenham valores de uma variável auxiliar x indicativa do tamanho das unidades populacionais. Sempre que a variação dos tamanhos das unidades for grande, a informação auxiliar disponível sobre os tamanhos for precisa e o tamanho for positivamente correlacionado com as variáveis de interesse poderemos empregar métodos de sorteio que permitem aumentar a eficiência na estimação de totais e médias, em comparação com métodos de amostragem com probabilidades iguais.

10.1.1 Estimação do total populacional

Seja $U = \{1, 2, \ldots, N\}$ a população de pesquisa. Considere o caso em que os valores de uma variável auxiliar x_i, $i \in U$ são conhecidos para todos as unidades da população mediante um cadastro. Se $x_i > 0 \; \forall \, i \in U$, então podemos usar esta variável como uma *medida de tamanho* das unidades populacionais. Se x for positivamente correlacionada com a(s) variável(is) de estudo y, então podemos esperar aumentar a eficiência fazendo seleção com PPT quando comparada com AAS.

Por enquanto, vamos supor que é possível selecionar amostras de acordo com um plano amostral tal que: $\pi_i \propto x_i \ \forall \ i = 1, \ldots, N$. Mais tarde, apresentaremos diversos métodos para implementar essa ideia.

Lembrando a teoria já apresentada no capítulo 3, sabemos que o estimador de Horvitz-Thompson (HT) para estimar o total populacional

$$Y = \sum_{i \in U} y_i$$

é dado por:

$$\widehat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i \tag{10.1}$$

onde cada unidade da amostra tem um peso amostral igual ao inverso da respectiva probabilidade de inclusão, dado por $d_i = 1/\pi_i \, \forall \, i \in U$.

O estimador HT do total é não viciado, isto é: $E(\widehat{Y}_{HT}) = Y$ e sua variância na forma de Horvitz-Thompson é dada por:

$$\begin{aligned} V_{HT}\left(\widehat{Y}_{HT}\right) &= \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_i y_j \\ &= \sum_{i \in U} \sum_{j \in U} \left(\frac{d_i d_j}{d_{ij}} - 1 \right) y_i y_j \end{aligned} \tag{10.2}$$

onde $d_{ij} = 1/\pi_{ij} \, \forall \, i, j \in U$.

Um estimador não viciado da variância do estimador HT é dado por:

$$\widehat{V}_{HT}\left(\widehat{Y}_{HT}\right) = \sum_{i \in s} \sum_{j \in s} \left(d_i d_j - d_{ij} \right) y_i y_j \tag{10.3}$$

Uma expressão alternativa para a variância do estimador HT, válida para planos amostrais de tamanhos fixos, é a chamada forma SYG (Sen-Yates-Grundy) da variância, dada por:

$$V_{SYG}\left(\widehat{Y}_{HT}\right) = \sum_{i \in U} \sum_{j > i} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{10.4}$$

A partir da expressão de Sen-Yates-Grundy para a variância do estimador de total é possível obter um estimador não viciado alternativo desta variância, dado por:

$$\widehat{V}_{SYG}\left(\widehat{Y}_{HT}\right) = \sum_{i \in s} \sum_{j > i} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{10.5}$$

Note que este estimador não coincide com o estimador de variância derivado a partir da expressão de Horvitz-Thompson, podendo os dois levar a estimativas distintas da variância do estimador HT do total. Cabe registrar que ambos os estimadores de variância para o estimador de total podem tomar valores negativos. Evidências empíricas sugerem que isto ocorre mais raramente com o estimador de Sen-Yates-Grundy. Veja a respeito Särndal, Swensson, e Wretman (1992).

Da expressão (10.4) podemos observar que a variância do estimador de total seria nula caso $y_i/\pi_i = y_j/\pi_j$ para todo $i \neq j \in U$. Portanto, se $\pi_i \propto x_i$ e $y_i \propto x_i \; \forall i \in U$, então $V_{SYG}\left(\widehat{Y}_{HT}\right) = 0$. Isto sugere que se y e x forem aproximadamente proporcionais (logo, alta e positivamente correlacionadas), então a variância do estimador HT do total será pequena.

Também se pode notar também que a variância deve ser pequena quando $\pi_{ij} \doteq \pi_i\pi_j \; \forall \, i \neq j \in U$. Acontece que $\pi_{ij} = \pi_i\pi_j \; \forall \, i \neq j \in U$ implica em indicadores de inclusão das unidades i e j independentes. Um plano amostral satisfazendo essa propriedade é a *Amostragem de Poisson* que estudaremos adiante. Entretanto, *Amostragem de Poisson* não é eficiente, como veremos, devido à variabilidade do tamanho amostral efetivo.

Sendo assim, a chave para alcançar eficiência através da amostragem PPT é ter medidas de tamanho (x) alta e positivamente correlacionadas com a(s) resposta(s) de interesse na pesquisa (y). Essa situação é muitas vezes encontrada ao realizar pesquisas de estabelecimentos ou instituições, onde as principais variáveis de estudo da pesquisa são bem correlacionadas com medidas de tamanho frequentemente disponíveis nos cadastros empregados para seleção da amostra.

10.2 Estimação da média populacional

Quando o tamanho da população N é conhecido, o estimador “natural” da média populacional baseado no estimador HT do total seria:

$$\bar{y}_{HT} = \widehat{Y}_{HT}/N = \frac{1}{N}\sum_{i\in s} d_i y_i = \sum_{i\in s} w_i^{HT} y_i \qquad (10.6)$$

onde $w_i^{HT} = d_i/N$.

As expressões de variância e seu estimador não viciado seguem diretamente das anteriores mediante divisão por N^2, levando a:

$$
\begin{aligned}
V_{HT}\left(\overline{y}_{HT}\right) &= \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)y_i y_j \\
&= \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}\left(\frac{d_i d_j}{d_{ij}}-1\right)y_i y_j
\end{aligned}
\tag{10.7}
$$

e

$$
\widehat{V}_{HT}\left(\overline{y}_{HT}\right) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\left(d_i d_j - d_{ij}\right)y_i y_j \tag{10.8}
$$

Expressões na forma Sen-Yates-Grundy podem ser obtidas de forma análoga.

Mesmo quando o tamanho N da população for conhecido, ele pode ser estimado usando o estimador HT do total de uma variável de contagem tomando valor igual a 1 para todas as unidades da população, levando ao estimador:

$$
\widehat{N}_{HT} = \sum_{i\in s} d_i
$$

Usando esse estimador do tamanho da população no denominador, um estimador tipo razão para a média populacional é dado por:

$$
\overline{y}_R = \widehat{Y}_{HT}/\widehat{N}_{HT} = \frac{\sum_{i\in s} d_i y_i}{\sum_{i\in s} d_i} = \sum_{i\in s} w_i^R y_i \tag{10.9}
$$

onde $d_i^R = d_i/\sum_{j\in s} d_j$.

A variância desse estimador de média pode ser aproximada por:

$$
V_{PPT}(\overline{y}_R) \doteq \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i-\overline{Y}}{\pi_i}\right)\left(\frac{y_j-\overline{Y}}{\pi_j}\right) \tag{10.10}
$$

Um estimador aproximadamente não viciado para essa variância é dado por:

$$
\widehat{V}_{PPT}(\overline{y}_R) = \frac{1}{\widehat{N}_{HT}^2}\sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i-\overline{y}_R}{\pi_i}\right)\left(\frac{y_j-\overline{y}_R}{\pi_j}\right) \tag{10.11}
$$

Cabe registrar que para alguns planos amostrais, os dois estimadores são equivalentes, isto é, $\overline{y}_R = \overline{y}_{HT}$ porque $w_i^R = w_i^{HT}$. Porém, quando diferem, o *estimador de razão da média* é geralmente mais eficiente que o estimador HT. Uma outra propriedade atraente do estimador

tipo razão da média é que ele é invariante sob transformações de locação, isto é, se tomarmos $z_i = y_i + A$, então $\overline{z}_R = \overline{y}_R + A$. Esta propriedade não se verifica para o estimador HT.

Em planos amostrais equi-ponderados, isto é, em que as probabilidades de inclusão π_i são todas iguais, os pesos w_i para estimação de médias ficam todos iguais a $1/n$ para ambos os estimadores (HT e de Razão). Esta é uma vantagem de planos deste tipo, pois a tarefa de estimação fica simplificada.

A Tabela 10.1 apresenta um resumo da estimação de parâmetros média e total sob PPT.

Tabela 10.1: Estimadores dos parâmetros média e total sob PPT

Parâmetro	**Estimador PPT**
$Y = \sum_{i \in U} y_i$	$\widehat{Y}_{HT} = \sum_{i \in s} d_i y_i$
$\overline{Y} = Y/N = \sum_{i \in U} y_i/N$	$\overline{y}_{HT} = \widehat{Y}_{HT}/N = \sum_{i \in s} d_i y_i/N = \sum_{i \in s} w_i^{HT} y_i$ $\overline{y}_R = \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i} = \sum_{i \in s} w_i^R y_i$
$V_{PPT}(\widehat{Y}_{HT})$	$\widehat{V}_{PPT}(\widehat{Y}_{HT}) = \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \right)$
$V_{PPT}(\overline{y}_{HT}) = V_{PPT}(\widehat{Y}_{HT})/N^2$	$\widehat{V}_{PPT}(\overline{y}_{HT}) = \widehat{V}_{PPT}(\widehat{Y}_{HT})/N^2$
$V_{PPT}(\overline{y}_R)$	$\widehat{V}_{PPT}(\overline{y}_R) = \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i - \overline{y}_R}{\pi_i} \right) \left(\underline{y}$
$V_{SYG}(\widehat{Y}_{HT})$	$\widehat{V}_{SYG}(\widehat{Y}_{HT}) = \sum_{i \in s} \sum_{j > i} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$

A seleção de amostras com PPT pode ser feita com ou sem reposição. O sorteio de amostras com reposição é pouco usado na prática, devido à perda de eficiência em comparação com métodos de sorteio sem reposição e também ao problema prático do que fazer em caso de repetição de unidades selecionadas. Apesar disso, é importante conhecer este método e suas propriedades, pois com frequência é usado como aproximação para obter estimativas simplificadas de variância. Já na amostragem PPT sem reposição há vários métodos de seleção, trazendo mais complexidade e dificuldades na estimação da precisão. Porém a eficiência é maior e se justifica o emprego de métodos mais complexos.

Na sequência apresentamos os principais métodos de sorteio de amostras PPT, com e sem reposição. Para um tratamento mais completo dos métodos disponíveis, consultar Brewer e Hanif (1983).

10.3 Amostragem PPT com reposição

O algoritmo para seleção de amostra PPT com reposição é denominado *método dos totais cumulativos*. consiste nos seguintes passos:

1. Acumule as medidas de tamanho na população, isto é, faça $X_{(0)} = 0$ e calcule $X_{(K)} = \sum_{i=1}^{k} x_i$ para $k = 1, \ldots, N$.
2. Determine *intervalos de seleção* com base no tamanho de cada unidade. Assim, o intervalo de seleção para a unidade k será dado por $(X_{(k-1)}; X_{(k)}]$, sendo o limite superior incluído, para $k = 1, \ldots, N$.
3. Selecione um número aleatório r com distribuição uniforme entre 0 e $X_{(N)}$, a soma dos tamanhos na população.
4. Selecione a unidade correspondente ao intervalo no qual cai o número aleatório r, isto é, selecione k tal que $r \in (X_{(k-1)}; X_{(k)}]$.
5. Repita os passos 3 e 4 tantas vezes quantas forem necessárias para obter a amostra do tamanho n desejado.

A *Amostra selecionada* é constituída pelas unidades $i_1, \ldots, i_j, \ldots, i_n$ cujos rótulos foram sorteados nas várias iterações do passo 4.

Amostragem PPT Com Reposição é um método muito simples de implementar, mas que pode implicar seleção repetida da(s) mesma(s) unidade(s). O tamanho efetivo da amostra (número de unidades distintas na amostra) é aleatório, podendo ser menor que o tamanho total desejado (n). O exemplo a seguir ilustra o emprego do método com uma pequena população para ajudar a compreensão.

Exemplo 10.1 Considere a população de $N = 6$ Fazendas, com as respectivas áreas apresentadas na Tabela 10.2. Este exemplo mostra como extrair uma amostra de $n = 3$ fazendas usando PPT com reposição, tomando a variável Área como medida de tamanho, usando o *método dos totais cumulativos*.

Tabela 10.2: Informações das áreas de uma população de $N = 6$ fazendas

Fazenda	Área	Tamanho Acumulado	Limite inferior do intervalo	Limite superior do intervalo
1	50	50	0	50
2	1.000	1.000	51	1.050
3	125	1.175	1.051	1.175
4	300	1.475	1.176	1.475
5	500	1.975	1.476	1.975
6	25	2.000	1.976	2.000

Caso os 3 números aleatórios sorteados de forma independente e com distribuição Uniforme entre 0 e 2000 fossem 654, 1230 e 1555, então as fazendas selecionadas seriam as de números 2, 4 e 5. Caso os 3 números aleatórios entre 0 e 2000 fossem 122, 754 e 1980, então as fazendas 2 e 6 seriam as selecionadas, com a fazenda 2 sendo selecionada duas vezes.

Estimação do total sob amostragem PPT com reposição

Um estimador não viciado do total sob amostragem PPT com reposição é dado por:

$$\widehat{Y}_{PPTC} = \frac{1}{n} \sum_{i \in s} \frac{f_i y_i}{p_i} \tag{10.12}$$

onde f_i representa o número de vezes que a unidade selecionada i foi sorteada, e $p_i = x_i/X$ é o tamanho relativo da unidade $i \in U$. Note que o número de unidades distintas no conjunto s pode ser menor que n, e também que $\sum_{i \in s} f_i = n$.

A variância de $\widehat{Y}_{PPTC}$ e o seu respectivo estimador são dados por:

$$V_{PPTC}\left(\widehat{Y}_{PPTC}\right) = \frac{1}{n} \sum_{i \in U} \left(\frac{y_i}{p_i} - Y\right)^2 p_i \tag{10.13}$$

$$\widehat{V}_{PPTC}\left(\widehat{Y}_{PPTC}\right) = \frac{1}{n(n-1)} \sum_{i \in s} f_i \left(\frac{y_i}{p_i} - \widehat{Y}_{PPTC}\right)^2 \tag{10.14}$$

Note que este estimador não viciado não é o estimador tipo Horvitz-Thompson do total. O estimador HT também pode ser empregado com este plano amostral, mas requer o cálculo das probabilidades de inclusão de primeira ordem das unidades populacionais, dadas por:

$$\pi_i = 1 - (1 - p_i)^n$$

A principal vantagem do estimador aqui apresentado é a simplicidade referente à estimação de variâncias. Não há resultados genéricos indicando em que situações a variância do estimador HT seria menor que a do estimador aqui descrito.

Como já discutido anteriormente, métodos de sorteio com reposição raramente são empregados na prática, pois sempre é possível aplicar métodos sem reposição de maior eficiência para o mesmo custo. Por esse motivo, passaremos agora a discutir alguns dos muitos métodos de sorteio de amostras com PPT sem reposição disponíveis. Nossa seleção de métodos a apresentar se guiou fortemente pela relevância da aplicação destes métodos em pesquisas conduzidas no Brasil, que vamos citar como exemplos ao longo da discussão.

10.4 Amostragem PPT de Poisson

O *método de Poisson* para seleção de amostras com PPT sem reposição é implementado mediante a realização de uma prova de Bernoulli independente para cada unidade da população, que determina se a unidade é incluída (ou não) na amostra com uma probabilidade proporcional ao seu tamanho. Caso todas as probabilidades de inclusão na amostra sejam iguais, este método se reduz à *Amostragem Binomial*, e portanto, este método é uma generalização simples daquele método.

Um algoritmo baseado em processamento sequencial de lista para implementar o método para selecionar uma amostra de tamanho n da população U de tamanho N consiste dos seguintes passos.

1. Para cada unidade populacional i, determine o valor da probabilidade de inclusão $\pi_i = nx_i/X$.
2. Para cada unidade da população selecione, de forma independente, um número aleatório A_i com distribuição uniforme no intervalo [0;1].
3. Inclua a unidade i na amostra se $A_i \leq \pi_i$.

O conjunto s de unidades selecionadas por este algoritmo não terá unidades repetidas, e terá um tamanho efetivo aleatório, de valor esperado igual a n.

Alguns cuidados devem ser observados ao implementar a *Amostragem PPT de Poisson*. Em primeiro lugar, verifique se nenhuma unidade tem tamanho relativo x_i/X maior que $1/n$. Se isto ocorrer, a 'probabilidade de inclusão' desta unidade seria maior que 1, o que é impossível. Caso alguma unidade seja tão grande que $x_i/X > 1/n$, inclua esta unidade com

certeza (isto é, faça $\pi_i = 1$), e refaça os cálculos dos π_i com o tamanho desta unidade excluído do total X, e o tamanho de amostra diminuído de uma unidade. Repita a verificação até que nenhuma unidade tenha tamanho relativo maior que 1 sobre o tamanho residual da amostra.

A *Amostragem PPT de Poisson* é pouco usada na prática devido à variabilidade do tamanho efetivo da amostra. É um metodo menos eficiente que outros métodos de seleção PPT sem reposição. Um método moderno que corrige este defeito é *Amostragem Sequencial de Poisson (ASP)* - veja Ohlsson (1998).

** Estimação HT do total sob Amostragem de Poisson**

O estimador HT do total sob Amostragem de Poisson é dado por:

$$\widehat{Y}_{PO} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i \tag{10.15}$$

A variância de $\widehat{Y}_{PO}$ e um estimador não viciado desta variância são dados por:

$$V_{PO}\left(\widehat{Y}_{PO}\right) = \sum_{i \in U} \pi_i (1 - \pi_i) \left(\frac{y_i}{\pi_i}\right)^2 = \sum_{i \in U} \frac{(1 - \pi_i)}{\pi_i} y_i^2 \tag{10.16}$$

$$\widehat{V}_{PO}\left(\widehat{Y}_{PO}\right) = \sum_{i \in s} (1 - \pi_i) \left(\frac{y_i}{\pi_i}\right)^2 = \sum_{i \in s} \frac{(1 - \pi_i)}{\pi_i^2} y_i^2 \tag{10.17}$$

Como ocorre na *Amostragem Binominal*, devido ao tamanho efetivo da amostra ser variável, também é possível usar um estimador de total tipo razão sob *Amostragem de Poisson*. Tal estimador é mais eficiente do que o estimador HT. Este estimador tipo razão é dado por:

$$\widehat{Y}_{PO}^{R} = \frac{N}{\widehat{N}} \sum_{i \in s} d_i y_i = N \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i} \tag{10.15}$$

A variância aproximada de $\widehat{Y}_{PO}^{R}$ e um estimador consistente desta variância são dados por:

XXX Até aqui

10.5 Amostragem sequencial de Poisson (ASP)

O método de *Amostragem Sequencial de Poisson (ASP)*, proposto por Ohlsson (1998), é uma modificação do método de *Amostragem de Poisson* que elimina a variabilidade do tamanho efetivo da amostra. O custo dessa modificação é um procedimento de amostragem um pouco

mais complexo, e que requer uso de resultados aproximados para a estimação tanto do total como de sua variância.

Um algoritmo baseado em processamento sequencial de lista para implementar o método para selecionar uma amostra de tamanho n da população U de tamanho N consiste dos seguintes passos.

1. Gerar um número aleatório uniforme independente A_i para cada unidade i da população.
2. Calcular a medida de tamanho relativo $p_i = x_i/X$ para cada unidade i da população.
3. Calcular o *número aleatório modificado* $C_i = A_i/p_i$.
4. Ordenar as unidades crescentemente segundo valores dos números aleatórios modificados C_i.
5. Selecionar para a amostra as n unidades com os menores valores de C_i.

Estimação com Amostragem Sequencial de Poisson

O estimador tipo HT do total sob *Amostragem Sequencial de Poisson* é dado por:

$$\widehat{Y}_{ASP} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i} \tag{10.18}$$

XXX Até aqui

A variância de $\widehat{Y}_{ASP}$ e o seu respectivo estimador são apresentados a seguir:

$$V_{ASP}\left(\widehat{Y}_{ASP}\right) = \frac{1}{n}\frac{N}{N-1}\sum_{i\in U}\left(\frac{y_i}{p_i} - Y\right)^2(1-np_i)p_i \tag{10.19}$$

$$\widehat{V}_{ASP}\left(\widehat{Y}_{ASP}\right) = \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{y_i}{p_i} - \widehat{Y}_{ASP}\right)^2(1-np_i)p_i \tag{10.20}$$

10.6 Amostragem sistemática com PPT

O método de Amostragem Sistemática com PPT consiste nos seguintes passos:

1. Acumule as medidas de tamanho na população, isto é, e faça $X_{(0)} = 0$ e calcule $X_{(k)} = \sum_{i=1}^{k} x_i$ para $k = 1, \ldots, N$.

2. Determine “intervalos de seleção” com base no tamanho de cada unidade. Assim, o intervalo de seleção para a unidade k será dado por $(X_{(k-1)}; X_{(k)}]$, sendo o limite superior incluído.

3. Determine o intervalo de amostragem (salto) $K = \frac{X_{(N)}}{n} = \frac{\text{Total dos Tamanhos}}{\text{Tamanho da Amostra}}$.

4. Selecione um número aleatório r (ponto de partida) com distribuição uniforme entre 0 e K.

5. Selecione as unidades correspondentes aos intervalos nos quais caem os valores $r + (j-1) \times K$, para $j = 1, 2, \ldots, n$. Isto é, selecione toda unidade i tal que $r + (j-1) \times K \in (X_{(i-1)}; X_{(i)}]$, para $j = 1, 2, \ldots, n$.

Seguem alguns cuidados a serem tomados no uso da Amostragem Sistemática com PPT:

- Verifique se nenhuma unidade tem tamanho x_i maior que K. Se isto ocorrer, esta unidade seria incluída ‘com repetição’ na amostra, o que é indesejável.
- Caso alguma unidade seja tão grande que $x_i > K$ inclua esta unidade com certeza, e refaça os cálculos para K com o tamanho desta unidade excluído do total e o tamanho de amostra diminuído de uma unidade.
- Repita a verificação até que nenhuma unidade tenha tamanho maior que o intervalo de seleção.

10.7 Amostragem PPT sistemática com ordenação

A seleção com amostragem PPT sistemática com ordenação segue os seguintes passos:

1. Faça uma ordenação das unidades da população segundo uma (ou mais) variável(is) de interesse.

2. Selecione uma amostra sistemática com PPT seguindo o algoritmo anterior.

Esta forma de implementar a Amostragem Sistemática PPT confere um efeito de ‘estratificação implícita’ pela variável (ou variáveis) usada(s) na ordenação.

Cabe registrar os seguintes comentários:

1. Amostragem sistemática com PPT é muito usada na prática por sua simplicidade na seleção da amostra.

2. Amostragem sistemática PPT tem desvantagens quanto à estimação de variâncias. Não há estimador de variância não viciado sob este plano amostral.

3. Com ordenação prévia da população por alguma outra característica, amostragem sistemática PPT confere efeito de estratificação implícita da amostra, podendo aumentar sua eficiência.

10.8 Amostragem PPT de Pareto (AP)

A seleção com amostragem PPT de Pareto, de acordo com (Rosén 2000), segue os seguintes passos:

1. Gerar número aleatório uniforme independente A_i para cada unidade i da população.
2. Calcular a *probabilidade de inclusão desejável* da unidade i : $\lambda_i = nx_i/X$.
3. Calcular número aleatório modificado $C_i = \frac{A_i(1-\lambda_i)}{(1-A_i)\lambda_i}$.
4. Ordenar as unidades crescentemente segundo valores dos números aleatórios modificados C_i.
5. Selecionar para a amostra as n unidades com os menores valores de C_i.

Seguem algumas considerações sobre Amostragem PPT de Pareto (AP)

1. Este é o método empregado para sorteio de unidades primárias de amostragem na PNAD Contínua.
2. Para tamanhos de amostra iguais, é mais eficiente que ASP.
3. Mesmos cuidados com relação a unidades muito grandes que com outros métodos PPT sem reposição são necessários.
4. Probabilidades exatas de inclusão não são estritamente proporcionais ao tamanho, e são difíceis de calcular.
5. Resultados quanto a estimadores são assintóticos e vício pode ocorrer com pequenas amostras.

10.8.1 Estimação com amostragem PPT de Pareto (AP)

O estimador do total sob Amostragem PPT de Pareto (AP) é dado por:

$$\widehat{Y}_{AP} = \sum_{i \in s} \frac{y_i}{\lambda_i} = \frac{1}{n} \sum_{i \in s} \frac{y_i}{p_i} \tag{10.21}$$

Note que $E\left(\widehat{Y}_{AP}\right) \doteq Y$.

A variância de $\widehat{Y}_{AP}$ e o seu respectivo estimador são apresentados a seguir:

$$V_{AP}\left(\widehat{Y}_{AP}\right) \doteq \frac{N}{N-1} \sum_{i \in U} \left(\frac{y_i}{\lambda_i} - \frac{\sum_{k \in U} y_k(1-\lambda_k)}{\sum_{k \in U} \lambda_k(1-\lambda_k)} \right)^2 \lambda_i(1-\lambda_i) \tag{10.22}$$

$$\widehat{V}_{AP}\left(\widehat{Y}_{AP}\right) = \frac{n}{n-1} \sum_{i \in s} \left(\frac{y_i}{\lambda_i} - \frac{\sum_{k \in s} y_k(1-\lambda_k)}{\sum_{k \in s}(1-\lambda_k)} \right)^2 (1-\lambda_i) \tag{10.23}$$

(Veja Rosén (2000)).

10.9 Exercícios

Exercício 10.1 Verifique que o estimador de HT da média não é invariante sob transformações de locação. Isto é, se tomarmos $z_i = y_i + A$, então $\overline{z}_{HT} \neq \overline{y}_{HT} + A$

10.10 Sobras do texto

Trataremos nessao caso de amostragem com probabilidades proporcionais ao tamanho. Outros casos serão vistos mais adiante, tais como: amostragem estratificada com alocação desproporcional, seleção de um morador para ser entrevistado em cada domicílio, amostras de números telefônicos (“random digit dialling samples”).

$\pi_{ij} > 0 \; \forall \; i \neq j \in U$

definição apresentada no capítulo 3, δ_i são as variáveis indicadoras de inclusão na amostra s , para todo $i \in U$. Para um plano amostral $p(s)$ qualquer sabemos que:

$E(\delta_i) = \pi_i$,

$E(\delta_i \delta_j) = \pi_{ij}$,

$V(\delta_i) = \pi_i(1-\pi_i)$ e

$COV(\delta_i, \delta_j) = \pi_{ij} - \pi_i \pi_j = \Delta_{ij}$.

$$V_{PPT}\left(\widehat{Y}_{HT}\right) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right) \tag{10.2}$$

$$\widehat{V}_{PPT}(\widehat{Y}_{HT}) = \sum_{i\in s}\sum_{j\in s}\frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right) \tag{10.3}$$

Exemplo 10.2 Diagrama de dispersão com dados de quantidade colhida e área plantada de cana-de-açúcar, apresentado na Figura 10.1.

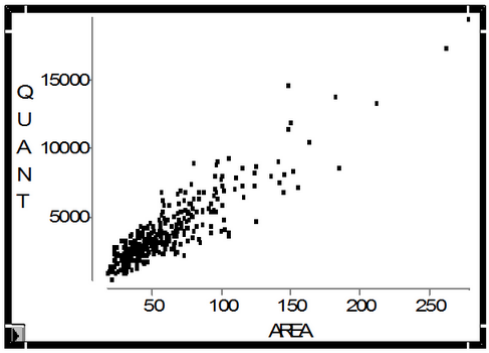

Figura 10.1: Diagrama de dispersão

1. Ordene as unidades da população U em ordem ascendente das medidas de tamanho x_i, $i \in U$. Denotar por $x_{(i)}$ a unidade posicionada no i-ésimo posto conforme a ordenação efetuada.

2. Calcule a soma das medidas de tamanho, dada por $X = \sum_{i\in U} x_i$.

3. Faça $i = N$ e verifique a condição $p_i < nx_{(i)}/X$. Caso a condição seja satisfeita, prossiga com os passos 4 e 5. Caso a condição não seja satisfeita, prossiga com o passo 3.

4. Calcule o total dos tamanhos excluída a unida