AMOSTRAGEM

Unidade 11

Amostragem com Probabilidades Desiguais

©2012 - Pedro Luis do Nascimento Silva

Amostragem Com Probabilidades Desiguais

Porque?

- Unidades de amostragem têm variação de tamanho
- Ignorar variação de tamanho pode resultar em desenhos ineficientes

Quando?

- Variação dos tamanhos for grande
- Informação auxiliar precisa sobre tamanhos disponível
- Tamanho fortemente correlacionado com variáveis de interesse

©2011 - Pedro Luis do Nascimento Silva

Amostragem Com Probabilidades Desiguais

Como?

- ➢ Amostragem com probabilidades proporcionais ao tamanho

Outros casos (veremos mais adiante)

- o Amostragem estratificada com alocação desproporcional;
- o Seleção de um morador para ser entrevistado em cada domicílio;
- o Amostras de números telefônicos ("random digit dialling samples").

Amostragem com Probabilidades Proporcionais ao Tamanho (PPT)

População: $U = \{ 1; 2; ...; N \}$

Valores de uma variável auxiliar x_i, $i \in U$, são conhecidos para todos os elementos da população.

Se $x_i > 0 \; \forall i \in U$, então podemos usar esta variável como uma medida de tamanho das unidades populacionais.

Se x for correlacionada com a(s) variável(is) de estudo y, então podemos esperar aumentar a eficiência fazendo seleção com PPT comparada com AAS.

Amostragem PPT

Por enquanto, vamos assumir que é possível selecionar amostras de acordo com um plano amostral tal que:

$\pi_i \propto x_i$ para todo $i=1,\ldots,N$;

$\pi_{ij}>0$ para todo $i \neq j \in U$.

Mais tarde, discutiremos algoritmos para garantir que essas condições sejam cumpridas.

©2011 - Pedro Luis do Nascimento Silva

Teoria Básica

Sejam δ_i as variáveis indicadoras de inclusão na amostra s, para todo $i \in U$.

Para um plano amostral p(s) qualquer sabemos que:

$E_p(\delta_i) = \pi_i$, $\quad E_p(\delta_i \delta_j) = \pi_{ij}$

$V_p(\delta_i) = \pi_i (1 - \pi_i)$, $\quad COV_p(\delta_i ; \delta_j) = \pi_{ij} - \pi_i \pi_j = \Delta_{ij}$

Estimação linear do total populacional $Y = \sum_{i \in U} y_i$:

$\hat{Y} = \sum_{i \in s} \dfrac{y_i}{\pi_i} = \hat{Y}_{HT}$ ➔ **Estimador de Horvitz-Thompson**

©2011 - Pedro Luis do Nascimento Silva

Propriedades do Estimador HT

- Cada unidade da amostra tem um peso amostral igual ao inverso da respectiva probabilidade de inclusão na amostra:

$$w_i = \pi_i^{-1} \ \forall i \in U.$$

- O estimador HT do total é não viciado, isto é:

$$E_p\left(\hat{Y}_{HT}\right) = Y$$

e sua variância é dada por

$$V_p\left(\hat{Y}_{HT}\right) = \sum_{i\in U}\sum_{j\in U}\left(\pi_{ij} - \pi_i\pi_j\right)\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right)$$

Esta é a forma de Horvitz-Thompson da variância.

Propriedades do Estimador HT

Um estimador não viciado da variância do estimador HT é:

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

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

$$V_{SYG}\left(\hat{Y}_{HT}\right) = -\frac{1}{2}\sum_{i\in U}\sum_{j\in U}\left(\pi_{ij} - \pi_i\pi_j\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Propriedades do Estimador HT

Um estimador não viciado alternativo de variância obtido a partir da forma de Sen-Yates-Grundy é dado por:

$$\hat{V}_{SYG}\left(\hat{Y}_{HT}\right) = -\frac{1}{2}\sum_{i\in s}\sum_{j\in s}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2$$

Note que esta fórmula não coincide com o estimador de variância derivado a partir da expressão de Horvitz-Thompson.

Eficiência da amostragem PPT

Da forma Sen-Yates-Grundy da variância, podemos observar que a variância 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 $V_{SYG}\left(\hat{Y}_{HT}\right) = 0$.

Isto indica que se y e x forem aproximadamente proporcionais (logo, altamente correlacionadas), a variância do estimador HT do total será pequena.

Também se pode notar também que a variância deve ser pequena quando $\pi_{ij} \cong \pi_i\,\pi_j$ $\forall i \neq j \in U$.

Eficiência da amostragem PPT

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**’.

Entretanto, Amostragem de Poisson não é eficiente, como veremos adiante, devido à variabilidade do tamanho amostral.

Chave para eficiência da amostragem PPT é ter medidas de tamanho (x) altamente correlacionadas com respostas de interesse na pesquisa (y).

Comentários

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.

Estimação Não Viciada da Média Populacional

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

$$\bar{y}_{HT} = \hat{Y}_{HT}/N = \frac{1}{N}\sum_{i \in s}\frac{y_i}{\pi_i} = \sum_{i \in s} w_i^{HT} y_i$$

onde $w_i^{HT} = \pi_i^{-1}/N$.

As fórmulas de variância e estimador da variância seguem diretamente das anteriores mediante divisão por N^2.

Estimador Tipo Razão da Média Populacional

Mesmo quando o tamanho N da população é conhecido, ele pode ser estimado por

$$\hat{N}_{HT} = \sum_{i \in s}\frac{1}{\pi_i} = \sum_{i \in s} w_i^{HT}.$$

Portanto, um estimador tipo razão para a média é dado por

$$\bar{y}_R = \hat{Y}_{HT}/\hat{N}_{HT} = \frac{\sum_{i \in s} y_i/\pi_i}{\sum_{i \in s} 1/\pi_i} = \frac{\sum_{i \in s} w_i^{HT} y_i}{\sum_{i \in s} w_i^{HT}} = \sum_{i \in s} w_i^{R} y_i$$

onde $w_i^{R} = w_i^{HT}/\sum_{j \in s} w_j^{HT}$.

Estimador Tipo Razão da Média Populacional

Sua variância é dada por

$$V_p(\bar{y}_R) \cong \frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i - \bar{Y}}{\pi_i} \right) \left(\frac{y_j - \bar{Y}}{\pi_j} \right)$$

Um estimador aproximadamente não viciado para essa variância é dado por

$$\hat{V}_p(\bar{y}_R) \cong \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i - \bar{y}_R}{\pi_i} \right) \left(\frac{y_j - \bar{y}_R}{\pi_j} \right)$$

Comentários

Para alguns planos amostrais, os dois estimadores são equivalentes, isto é, $\bar{y}_R = \bar{y}_{HT}$ porque $w_i^R = w_i^{HT}$.

O **estimador de razão da média** é geralmente mais eficiente que o de HT.

O estimador tipo razão da média é invariante sob transformações de locação. Isto é, se tomarmos $z_i = y_i + A$, então $\bar{z}_R = \bar{y}_R + A$.

Exercício 11.1 – Verifique que o estimador de HT da média não possui esta propriedade.

Planos Amostrais Auto-ponderados

Em planos amostrais auto-ponderados, isto é, em que os π_i são constantes, os pesos w_i ficam todos iguais a 1/n para ambos os estimadores de média (HT e de Razão).

Esta é uma vantagem de planos deste tipo, pois a tarefa de estimação fica simplificada.

Maneiras de Selecionar Amostras com PPT

COM REPOSIÇÃO	SEM REPOSIÇÃO
Simplicidade da seleção	Alternativas de seleção + complexidade
Simplicidade da estimação	Dificuldade na estimação de precisão
Eficiência não é plena	Eficiência plena
Coletar unidade repetida?	Não tem esse problema

Amostragem PPT Com Reposição

Método dos Totais Cumulativos

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,...,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.

Método dos Totais Cumulativos

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
   $r \in (X_{(k-1)} ; X_{(k)}]$.

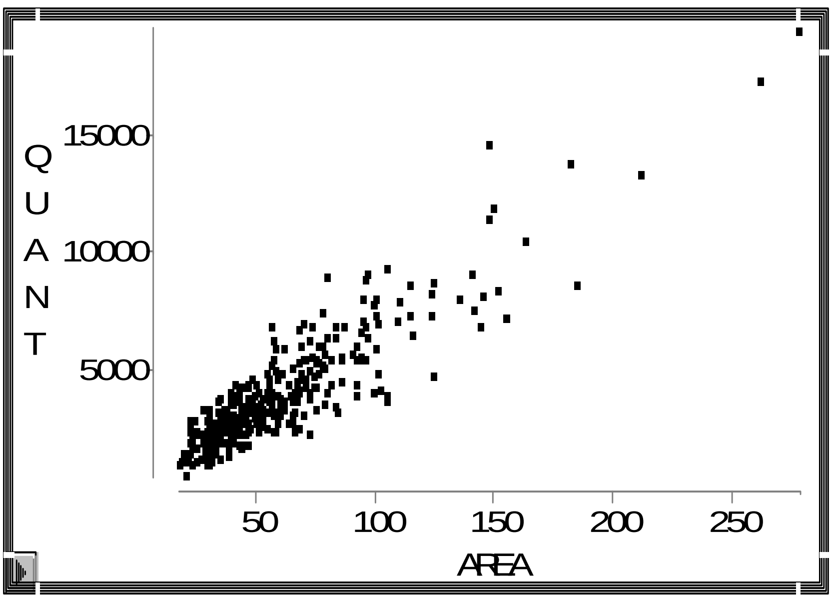
5. Repita os passos 3 e 4 tantas vezes quantas forem necessárias para obter a amostra do tamanho n desejado.

Exemplo 11.1 – População de N=6 Fazendas

Fazenda	Área	Tamanho Acumulado	Intervalo de Seleção	
			Limite Inferior	Limite Superior
1	50	50	0	50
2	1000	1050	51	1050
3	125	1175	1051	1175
4	300	1475	1176	1475
5	500	1975	1476	1975
6	25	2000	1976	2000

Extrair amostra de n=3 fazendas com PPT $\propto$ Área

Exemplo 11.2 – Diagrama de dispersão com dados de quantidade colhida e área plantada de cana de açúcar.



Estimação do Total Sob Amostragem PPT Com Reposição

$$\hat{Y}_{PPTC} = \frac{1}{n} \sum_{i \in s} \frac{y_i}{p_i} \text{ onde } p_i = x_i / X \text{ para } i = 1, 2, ..., N.$$

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

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

Amostragem PPT de Poisson

Passos

1. Para cada unidade populacional, determine o valor da probabilidade de inclusão $\pi_i = n\, x_i/X$.

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 $A_i \le \pi_i$.

Estimador Simples de Total Sob Amostragem De Poisson

$$\hat{Y}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$$

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

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

Amostragem PPT de Poisson - Cuidado

- Verifique se nenhuma unidade tem tamanho x_i maior que X/n. Se isto ocorrer, a 'probabilidade de inclusão' desta unidade seria maior que 1, o que é impossível.
- Caso alguma unidade j seja tão grande que $x_j > X/n$ inclua esta unidade com certeza (isto é, faça $\pi_j = 1$), e refaça os cálculos dos π_i 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.

Comentários

1. Amostragem PPT de Poisson é pouco usada na prática devido à variabilidade do tamanho da amostra.
2. Amostragem PPT de Poisson é menos eficiente que outros métodos de seleção PPT sem reposição.
3. É possível usar estimador de total mais eficiente do que o estimador simples.
4. Método moderno que corrige este defeito é "Amostragem Seqüencial de Poisson" (ASP) - veja Ohlsson(1998).

Amostragem Sequencial De Poisson (ASP)

Passos

1. Gerar número aleatório uniforme independente A_i para cada unidade i do cadastro.
2. Calcular medida de tamanho relativo p_i da unidade i.
3. Calcular 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

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

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

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

Amostragem PPT

A amostragem com probabilidades proporcionais ao tamanho é bastante utilizada em planos amostrais conglomerados.

Geralmente é o método escolhido para sorteio das Unidades Primárias de Amostragem.

Uma alternativa para planos amostrais de unidades elementares é a amostragem estratificada.

Outra alternativa é o emprego de AAS combinada com estimadores tipo razão ou regressão.