A Importância do Desenho Amostral

Donald Pianto
Departamento de Estatística
UnB

Objetivo dessa aula

- Explicar os tipos básicos de amostragem e a razão pelo uso de cada um
- Contemplar o uso simultaneo de mais que um tipo básico (desenhos complexos)
- Examinar estimadores de regressão para desenhos complexos fundamentados em:
  - modelos
  - o desenho amostral
- Verificar as consequências do uso ou não do desenho amostral em algumas análises empíricas

Técnicas de Amostragem

- Amostragem probabilística (aleatória) – onde a probabilidade de um elemento da população ser escolhida é conhecida
- Amostragem não probabilística (não-aleatória) – onde não se conhece, a priori, a probabilidade de um elemento da população pertencer à amostra
- Nós consideraremos técnicas de amostragem probabalística

Amostragem Aleatória Simples (AAS)

- Consideramos uma população com N elementos e uma amostra de tamanho n
- Cada elemento é selecionado independentemente dos outros
- Cada elemento é selecionado com probabilidade n/N
- Modelos podem ser estimados usando os procedimentos padrão dos programas
- Pode ser usado em populações completamente enumerados e de fácil acesso

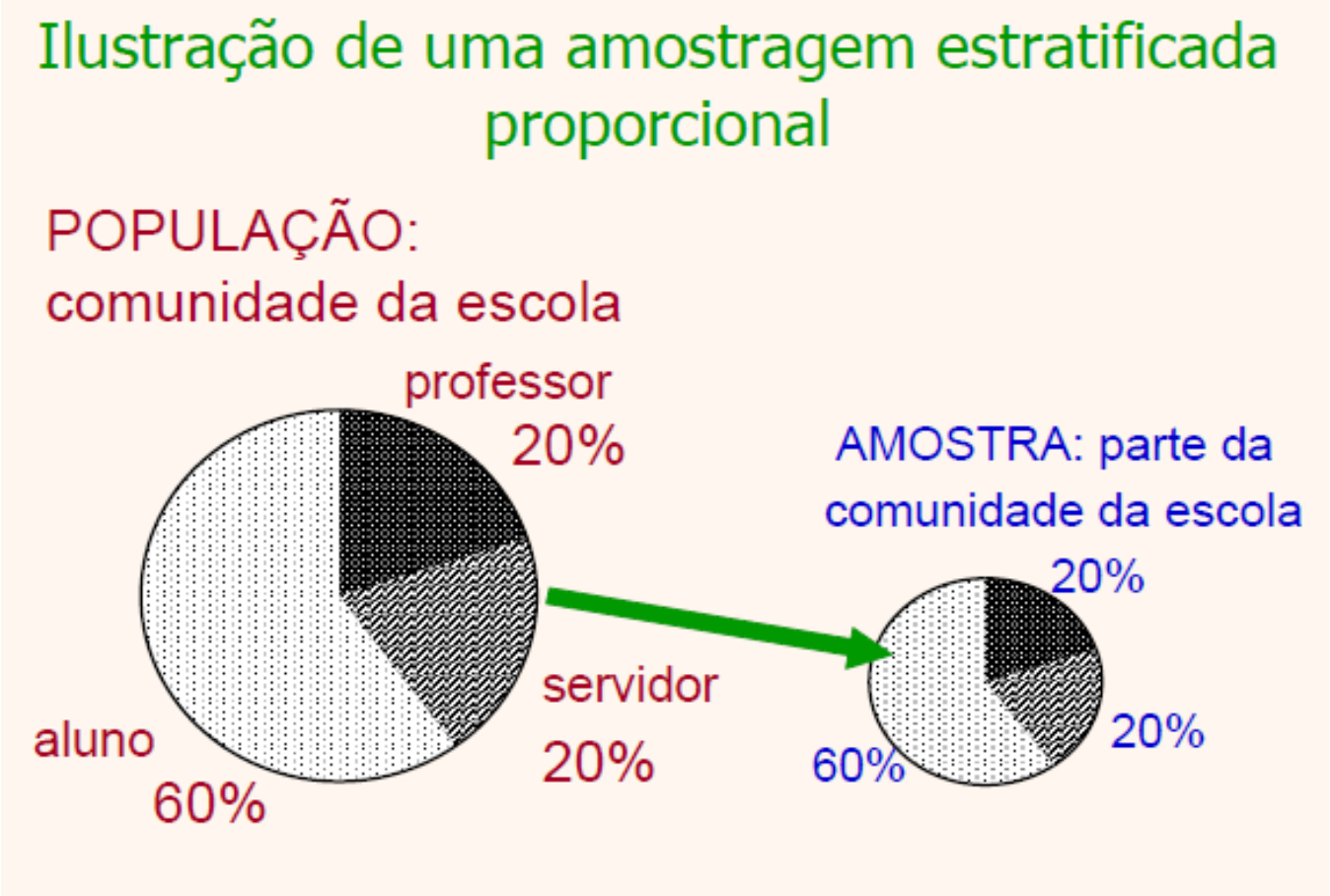
Amostragem Estratificada

- Se tiver grupos com valores médios diferentes em sub-populações diferentes
- Se quiser se proteger contra uma amostra "ruim" (90 homens e 10 mulheres)
- Se quiser fazer estimativas para sub-grupos
- Se faz sentido aplicar instrumentos de coleta diferentes para sub-grupos diferentes
- Teoricamente oferece mais precisão que AAS para o mesmo tamanho de amostra

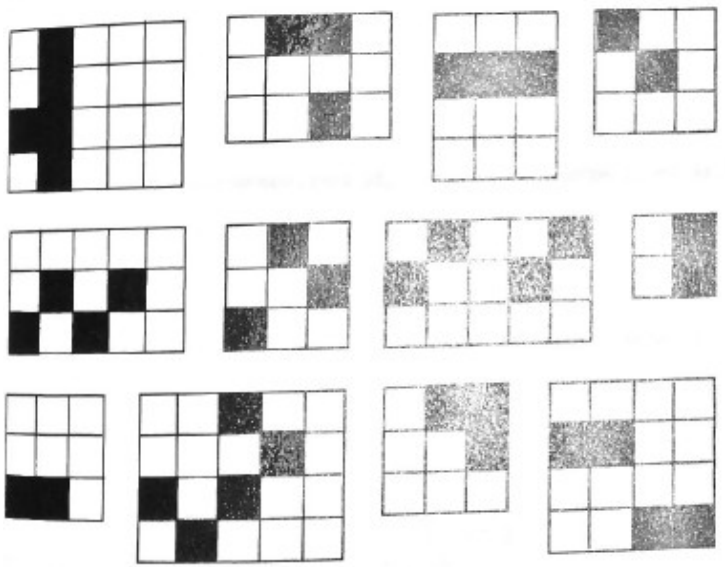
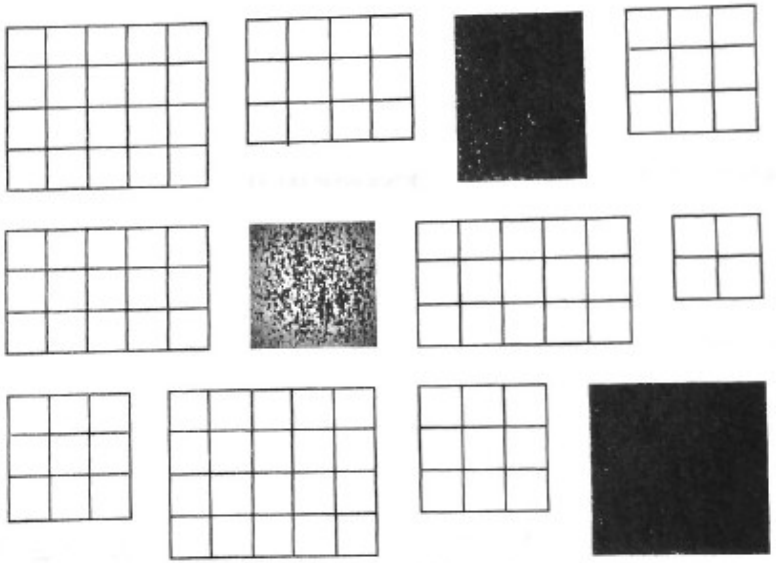
Amostragem Estratificada

- Se algum estrato for superamostrado, então a probabilidade de inclusão na amostra não é igual para todos os elementos
- Nesse caso seria necessário usar pesos para estimar médias populacionais
- Amostragem Estratificada não é Amostragem por Cotas (onde não sabemos a probabilidade de pertencer à amostra)

Amostragem Estratificada



Amostragem por Conglomerados

Take an SRS from *every* stratum:	Take an SRS of clusters; observe all elements within the clusters in the sample:
	
Variance of the estimate of $\bar{y}_U$ depends on the variability of values *within* strata.	The cluster is the sampling unit; the more clusters we sample, the smaller the variance. The variance of the estimate of $\bar{y}_U$ depends primarily on the variability *between* cluster means.
For greatest precision, individual elements within each stratum should have similar values, but stratum means should differ from each other as much as possible.	For greatest precision, individual elements within each cluster should be heterogeneous, and cluster means should be similar to one another.

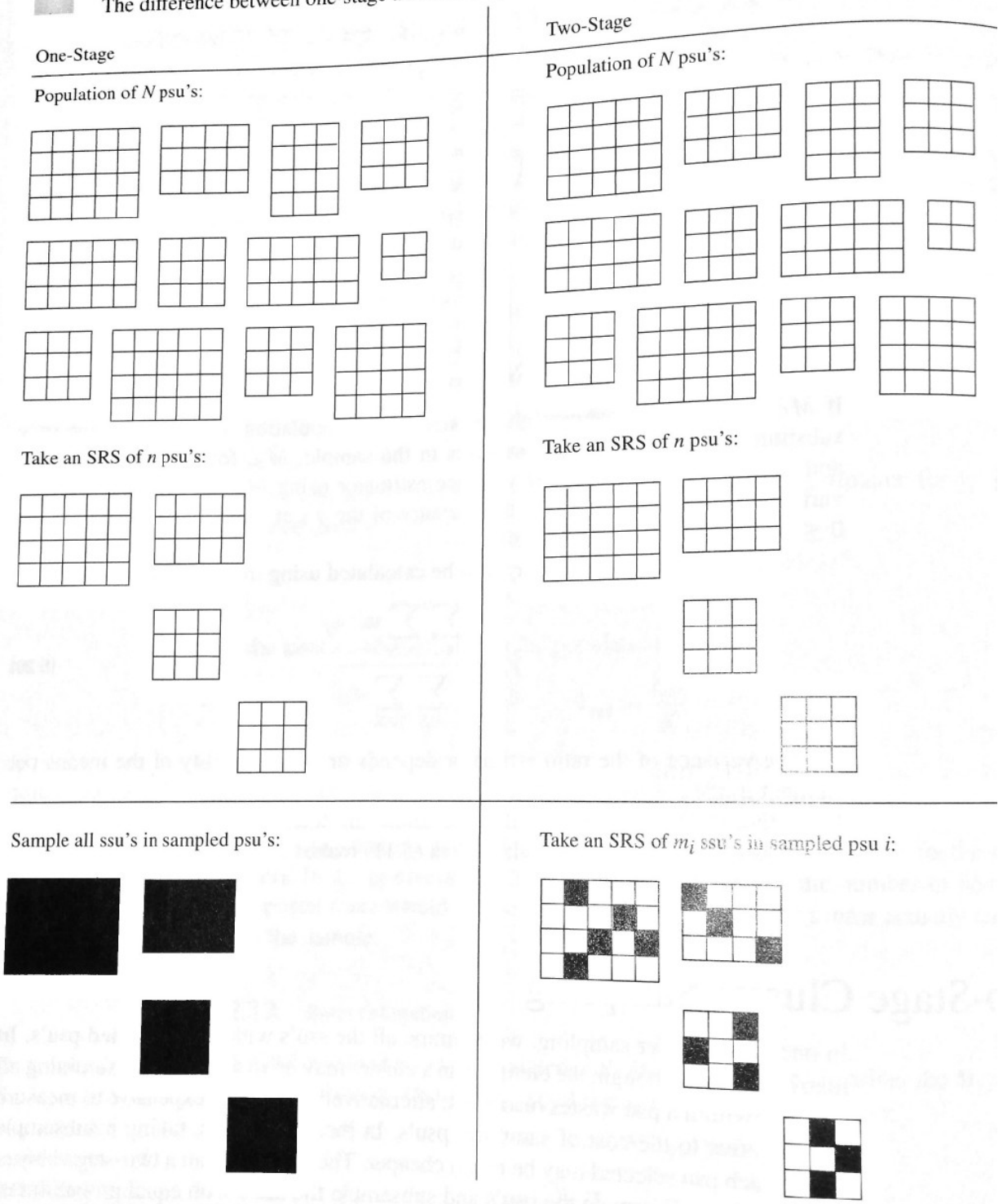
Amostragem por Conglomerados

- Se for difícil ou impossível listar todas os elementos da população (abelhas, árvores, pessoas)
- Se tiver uma distribuição geográfica muito grande ou naturamente ocorre em clusters (com AAS você pode visitar uma área isolada só para entrevistar uma única pessoa)

Amostragem por Conglomerados

- Elementos em conglomerados tendem a ser semelhantes; isso reduz a precisão de estimativas
- Mesmo se a probabilidade de inclusão de cada elemento for igual, as variâncias não podem ser calculadas como se as observações fossem independentes

FIGURE 5.2
The difference between one-stage and two-stage cluster sampling



Amostragem por Conglomerados em Dois Estágios

- Mesmo com probabilidades iguais de seleção o desenho amostral deve ser levado em conta para calcular as variâncias
- As fórmulas para as variâncias dependem de características de ambos os estágios (usar um comando que só leva em conta a correlação no último conglomerado vai subestimar a variância)

Amostragem com probabilidades desiguais

- Tanto amostragem estratificada quanto amostragem por conglomerados podem acomodar probabilidade desiguias
- Nesse caso as estimativas de médias podem ser calculadas usando pesos
- As estimativas de variâncias usam os pesos e o desenho amostral

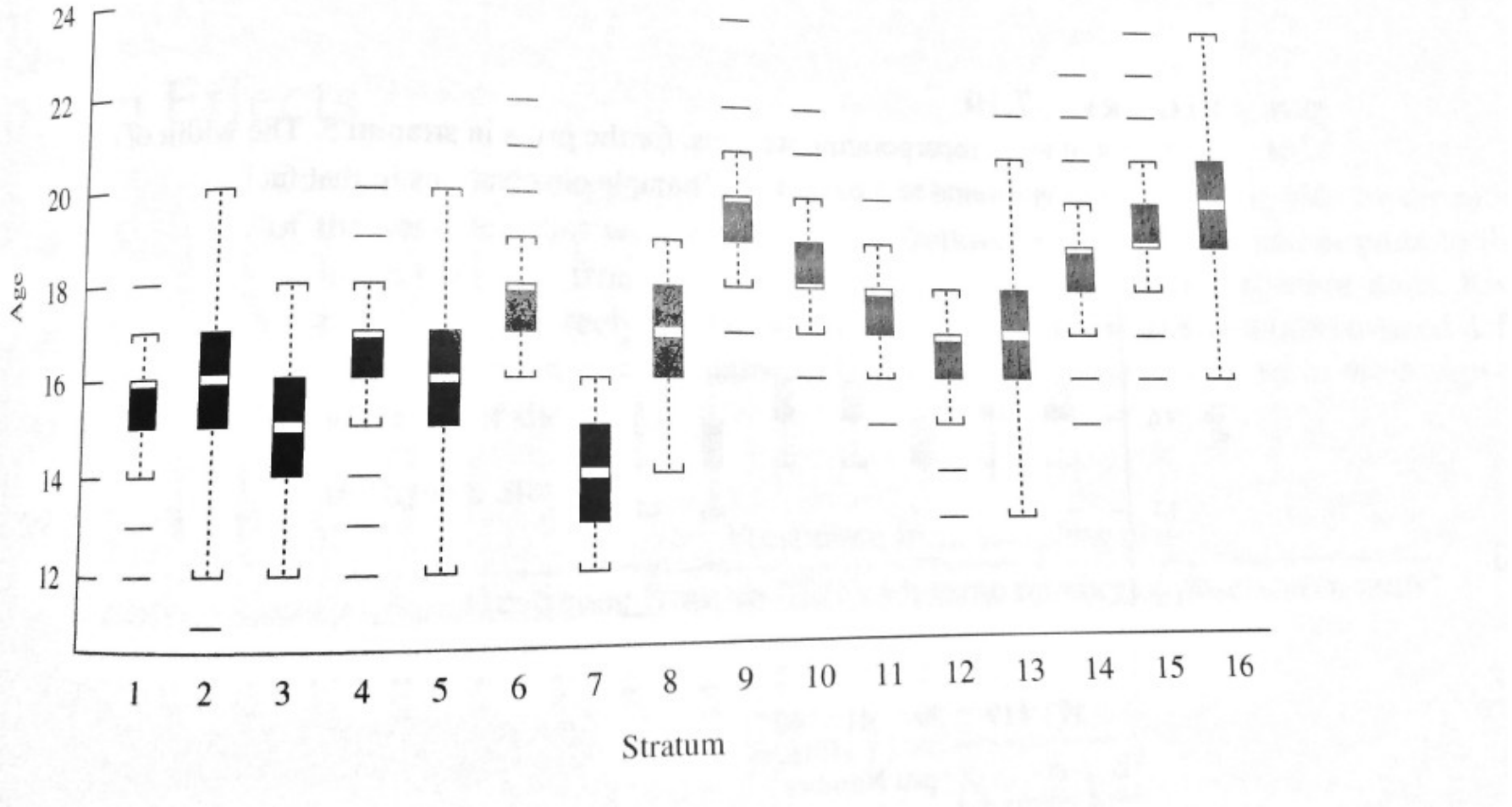
Amostras Complexas

- Amostragem complexa se refere a desenhos amostrais que misturam os elementos apresentados até agora
- Nesses casos tanto o desenho quanto os pesos precisam ser levados em conta para estimar as variâncias

Amostras Complexas

FIGURE 7.8

A boxplot of age distributions for each stratum, incorporating the weights. Note the wide variability from stratum to stratum.



Amostras Complexas

- Efeitos do Desenho (Efeito do Plano Amostral)

$$EPA(\hat{\theta}) = \frac{V_p(\hat{\theta})}{V_{AAS}(\hat{\theta})}$$

- Maior EPA indica maior variância por causa do plano amostral

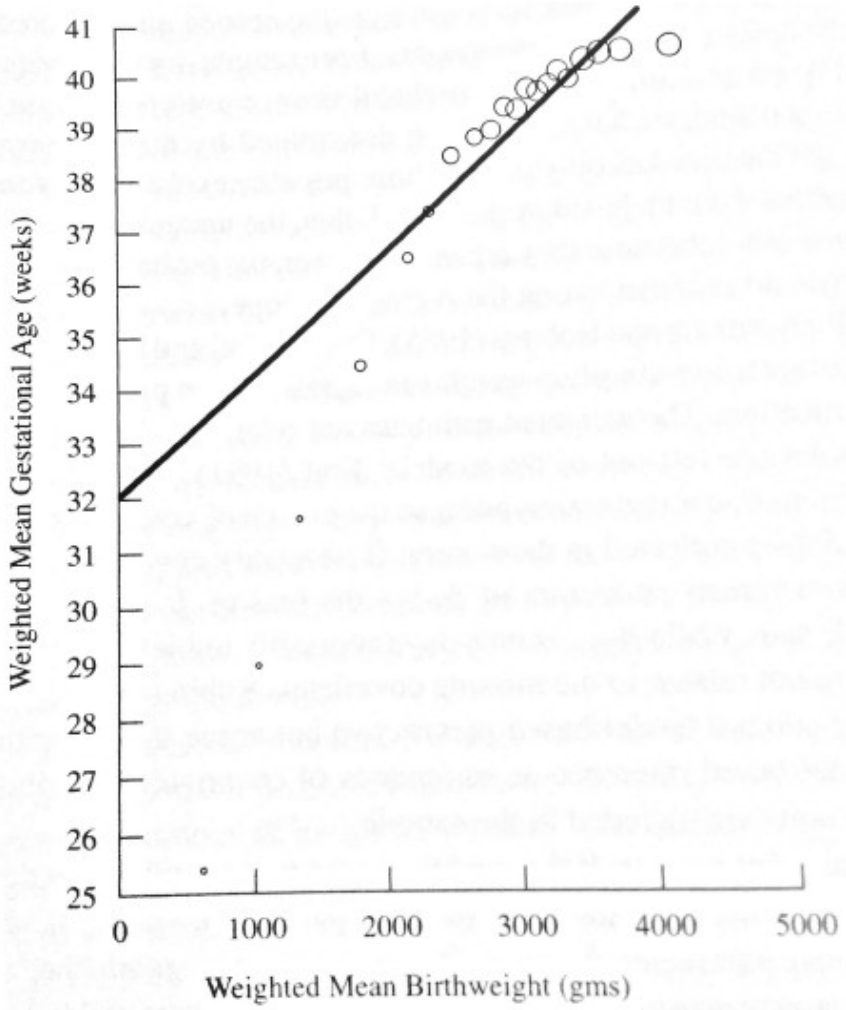
Amostras Complexas

- Mesmo quando consideramos somente tabelas de contingência é importante levar o desenho e os pesos em conta (no R tem svytable e svychisq)
- O efeito de não levar o desenho em conta (as correlações dentro dos conglomerados) é de rejeitar homogeneidade com mais frequência do que o nível descritivo do teste

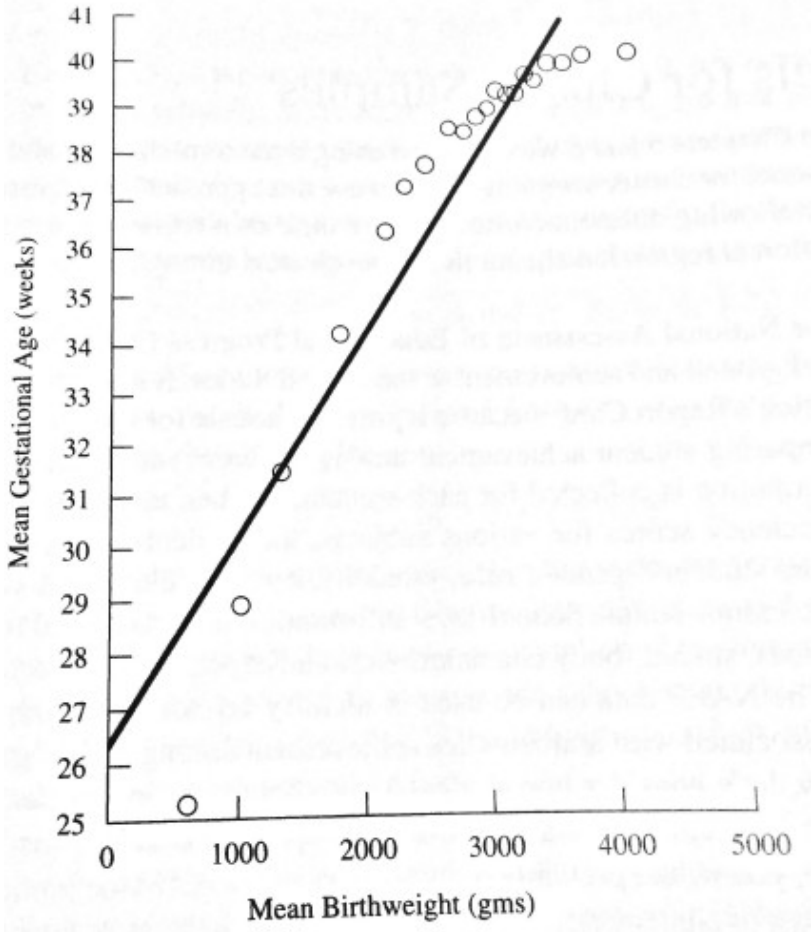
Regressão

- Em uma amostra complexa, devemos usar os pesos amostrais?
- Uma regressão pode:
  - Descrever a associação entre duas variáveis
  - Prever resultados futuros
  - Revelar modelos de causa e efeito
- Você quer saber o valor na amostra o que generalizar para a população?
- Se acreditar no modelo, pesos não importam (física)

Regressão

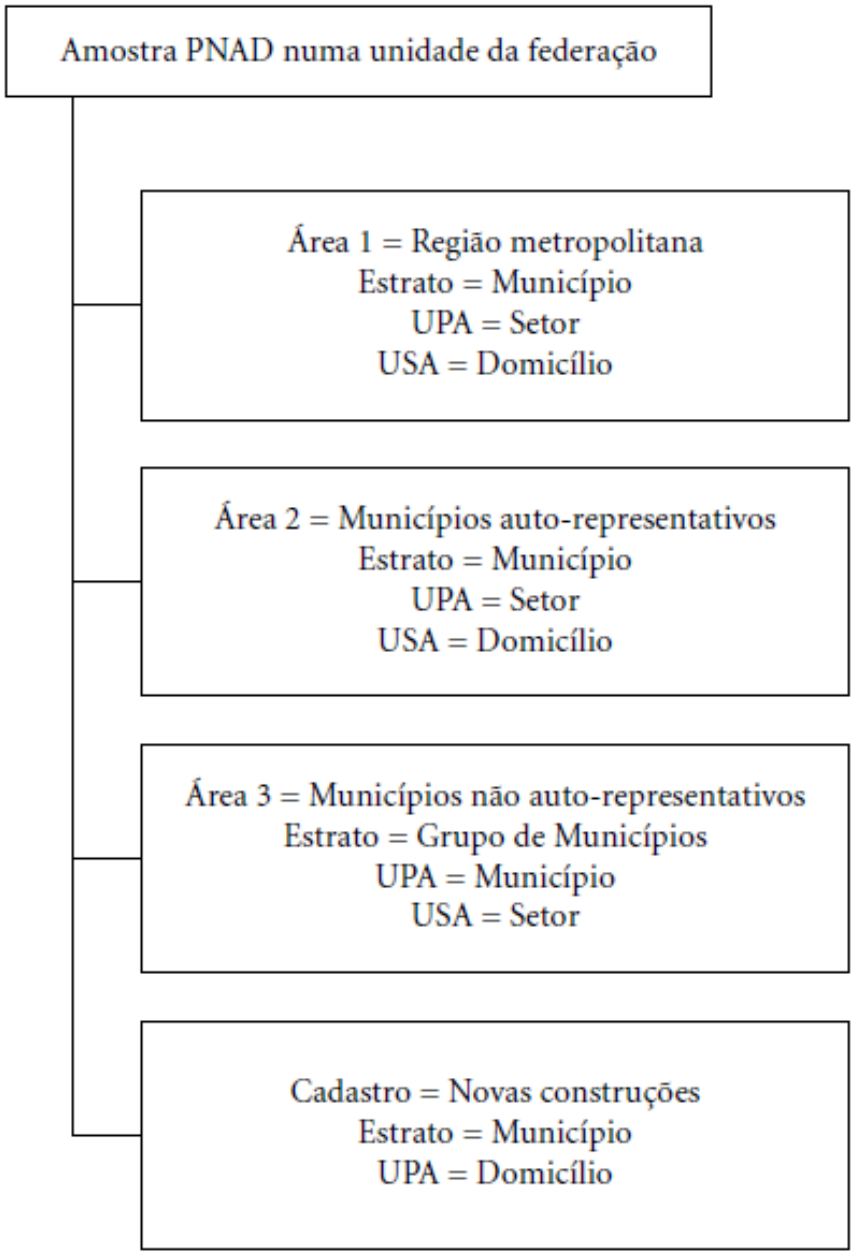


Com Pesos



Sem Pesos

Figura 1
Ilustração do plano amostral da PNAD durante a década de 1990.



PNAD 1998

Tabela 1
Estimativas, desvios padrão, coeficientes de variação e efeitos do plano amostral para variáveis de pessoas – PNAD – 1998.

Linha	Descrição da variável	Estimativa	Desvio padrão	CV(%)	EPA
1	Proporção de pessoas brancas	53,8%	0,3%	0,6	13,7
2	Proporção de pessoas negras ou pardas	45,4%	0,3%	0,7	13,7
3	Proporção de pessoas analfabetas	24,4%	0,2%	0,7	5,8
4	Proporção de pessoas que freqüentam escola	30,9%	0,1%	0,4	2,3
5	Proporção de pessoas exercendo trabalho infantil	2,8%	0,2%	5,2	2,6
6	Proporção de pessoas que trabalham	54,8%	0,2%	0,3	3,4
7	Proporção de pessoas empregadas	2,7%	0,1%	2,9	8,4
8	Proporção de pessoas conta própria	2,7%	0,1%	2,5	6,2
9	Proporção de pessoas empregadoras	0,3%	0,0%	5,3	3,0
10	Proporção de pessoas com auxílio-moradia	7,8%	0,2%	2,4	4,5
11	Proporção de pessoas com auxílio-alimentação	37,2%	0,3%	0,8	3,3
12	Proporção de pessoas com auxílio-transporte	34,2%	0,3%	0,9	3,7
13	Proporção de pessoas com auxílio-creche/educação	2,6%	0,1%	2,8	1,9
14	Proporção de pessoas com auxílio-saúde	16,5%	0,3%	1,6	4,8
15	Renda média do trabalho principal	512,8	5,8	1,1	5,4
16	Proporção de pessoas com previdência	44,2%	0,3%	0,7	5,6

Pesquisa de Padrões de Vida 96/97

Tabela 3.8 – Freqüências relativas, simples e com ponderação, segundo as variáveis categóricas consideradas no estudo. (Continua)

Variável	Categorias	Freqüência Relativa(%)	
		Simples	Com ponderação
Sexo			
	Masculino	61,37	63,04
	Feminino	38,63	36,96
Raça/Cor			
	Branco	47,24	57,19
	Não-branco	52,76	42,81

Os pesos são correlacionados com a variável dependente

Tabela 4.2 – Estimativas da média da variável resposta (logaritmo do salário-hora), simples e com ponderação, segundo as variáveis categóricas consideradas no estudo.

(Continua)

Variável	Categorias	Média da variável		
		Simples	Com ponderação	Diferença
Sexo				
	Masculino	0,5960	0,6735	0,0775
	Feminino	0,3806	0,4274	0,0469
Raça/Cor				
	Branco	0,8196	0,8566	0,0369
	Não-branco	0,2380	0,2165	-0,0214
Nível educacional				
	0 a 1 ano de estudo	-0,2278	-0,1533	0,0745
	2 a 4 anos de estudo	0,1817	0,2960	0,1143
	5 a 8 anos de estudo	0,3591	0,4485	0,0894
	9 a 11 anos de estudo	0,8816	0,9225	0,0409
	12 a 15 anos de estudo	1,7977	1,8546	0,0569
	16 ou mais anos de estudo	2,5012	2,5778	0,0766

Pesos mudam estimativas e o desenho muda o erro padrão

Tabela 5.1 – Estimativas pontuais dos efeitos principais, respectivos erros padrões e EPA's obtidos pelos métodos de estimação (continua)

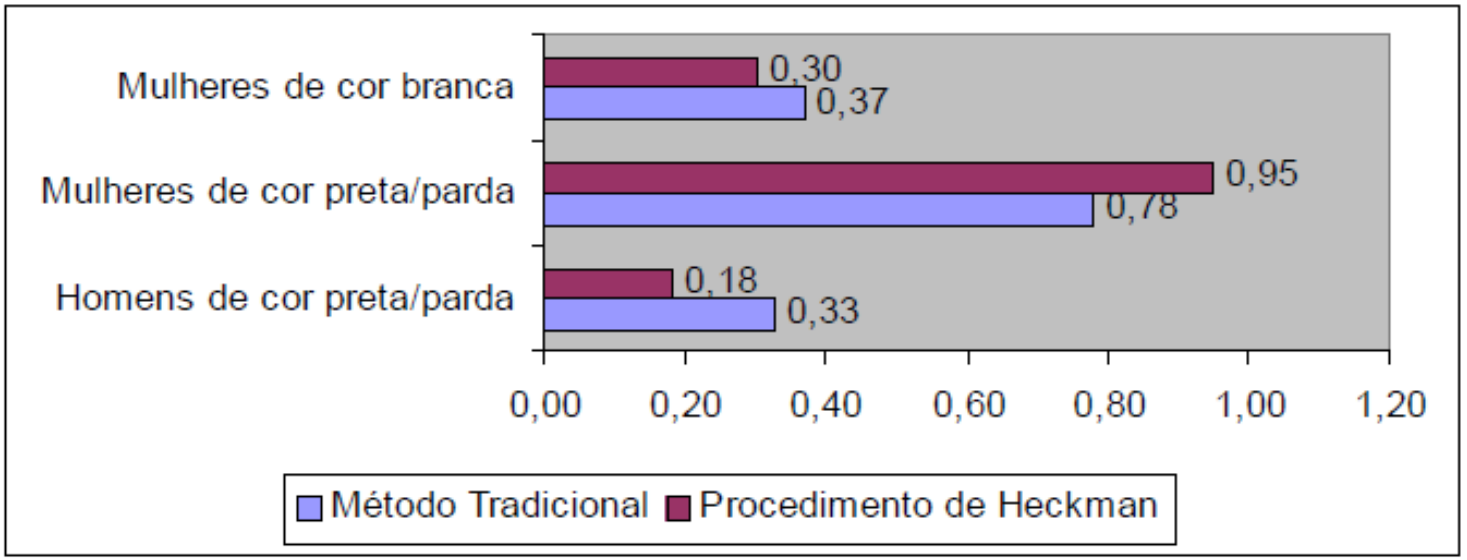
Var. indep.	Níveis de fatores		Estimativa Pontual		Desvio Padrão			EPA
			MQO	MQP e MPV	MQO	MQP	MPV	
Intercept			0,9631	1,0454	0,0991	0,1037	0,1449	1,95
SEX	MASCULINO	X_1	0,2291	0,2482	0,0237	0,0242	0,0309	1,63
RAC	BRANCO	X_2	0,1489	0,1467	0,0221	0,0227	0,0338	2,23
NIV	0 A 1 ANO	X_3	-1,9677	-1,9708	0,0707	0,0754	0,0971	1,66
	2 A 4 ANOS	X_4	-1,7237	-1,7208	0,0660	0,0706	0,0892	1,60
	5 A 8 ANOS	X_5	-1,5622	-1,5682	0,0655	0,0701	0,0916	1,71
	9 A 11 ANOS	X_6	-1,1795	-1,2163	0,0638	0,0687	0,0853	1,54
	12 A 15 ANOS	X_7	-0,5342	-0,5478	0,0683	0,0719	0,0902	1,57
EXF		X_8	0,0105	0,0091	0,0014	0,0014	0,0021	2,23
CPE	SIM	X_9	-0,1000	-0,0558	0,0271	0,0266	**0,0373**	1,97
EXP		X_{10}	0,0089	0,0095	0,0009	0,0009	0,0014	2,37
SIN	SINDICAL	X_{11}	0,2129	0,2378	0,0279	0,0283	0,0380	1,80
SAZ	IRREGULAR	X_{12}	**0,0129**	**-0,0549**	**0,0497**	**0,0490**	**0,0748**	2,33

* As estimativas para os parâmetros não significativos ao nível de 5% encontram-se em negrito.

** As linhas referentes às categorias-base foram excluídas, portanto foram excluídos os seguintes níveis de fatores: feminino (SEX), não-branco (RAC), 16 ou + anos (NIV), não (CPE), não sindicalizado (SIN) e regular (SAZ).

Compensar para o desenho e seleção amostral (Carvalho, Neri, Nascimento 2005)

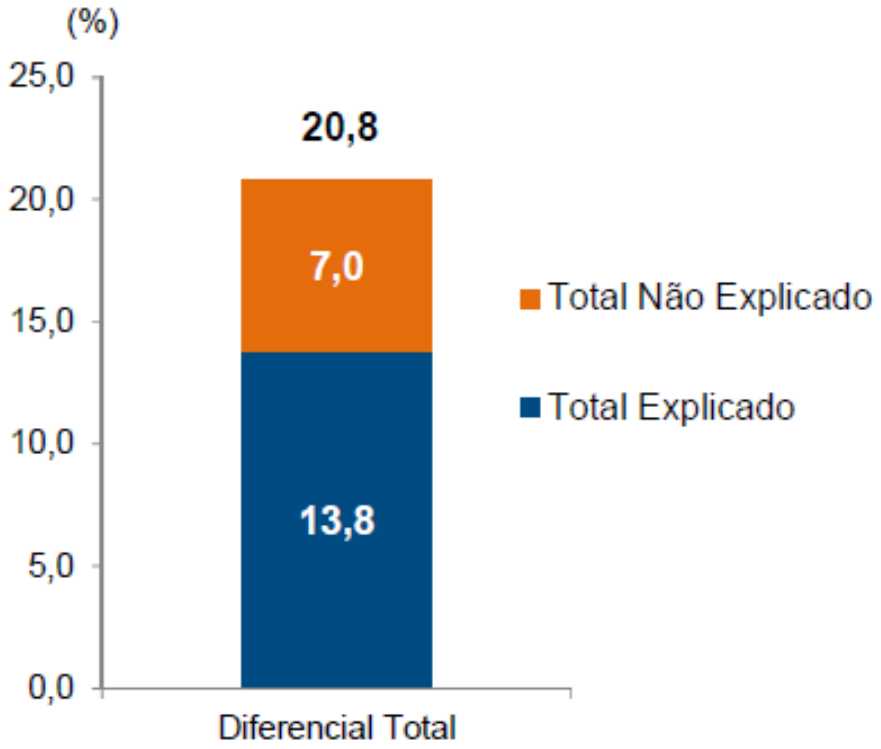
Gráfico 6.3 – Estimativa do Coeficiente de Discriminação – Método Tradicional vs Procedimento de Heckman



FEE-RS 2015 – PNAD 2013

Gráfico 5

Decomposição do diferencial de salários entre homens e mulheres no Brasil — 2013



FONTE DOS DADOS BRUTOS: IBGE/PNAD.

FEE-RS 2015 – PNAD 2013

- Esses autores não reportaram se usaram os pesos ou o desenho amostral
- Não usaram um modelo de correção de viés
- Seria interessante verificar se alguma conclusão muda.....

FEE-RS 2015 – PNAD 2013

- Esses autores não reportaram se usaram os pesos ou o desenho amostral
- Não usaram um modelo de correção de viés
- Seria interessante verificar se alguma conclusão muda.....

Algumas Referências

- T. Lumley (2014) "survey: analysis of complex survey samples". R package version 3.30.
- T. Lumley (2004) Analysis of complex survey samples. Journal of Statistical Software 9(1): 1-19
- S. L. Lohr (1999) Sampling: Design and Analysis. Duxbury Press.
- P. L. do Nascimento Silva, D. G. Carneiro Pessoa, M. Franca Lila (2002) "Análise estatística de dados da PNAD:incorporando a estrutura do plano amostral", Ciência & Saúde Coletiva, 7(4):659-670.
- S. de Castro Rodrigues (2003) Análise da estrutura salarial revelada pela PPV incorporando peso e plano amostral.
- A. P. de Carvalho (2005) Decomposição do Diferencial de Salários no Brasil em 2003.
- G. Stein, V. N. Sulzbach, M. Bartels (2015) Relatório sobre o mercado de trabalho do Rio Grande do Sul – 2001-13.