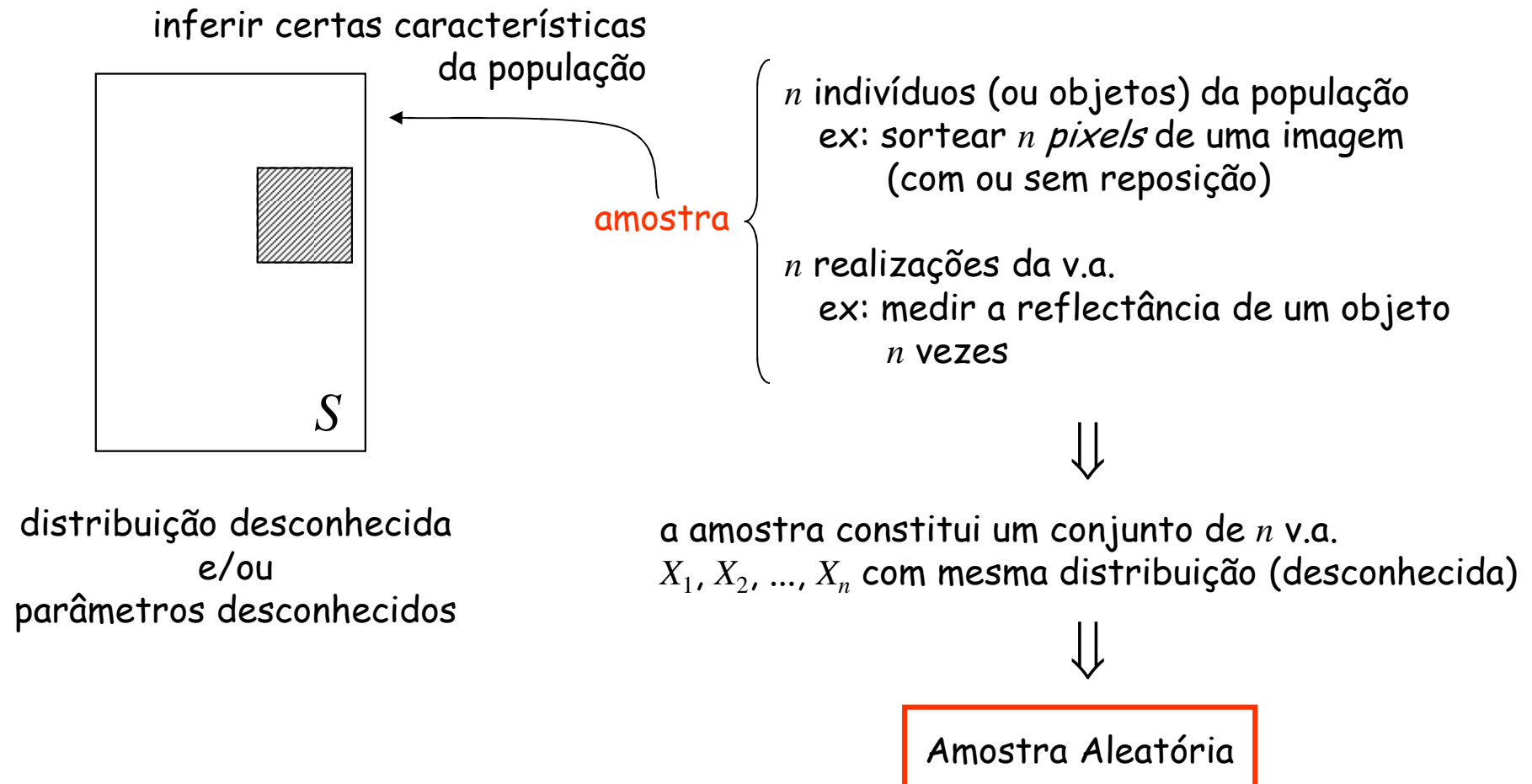

Jackknife, Bootstrap e outros métodos de reamostragem



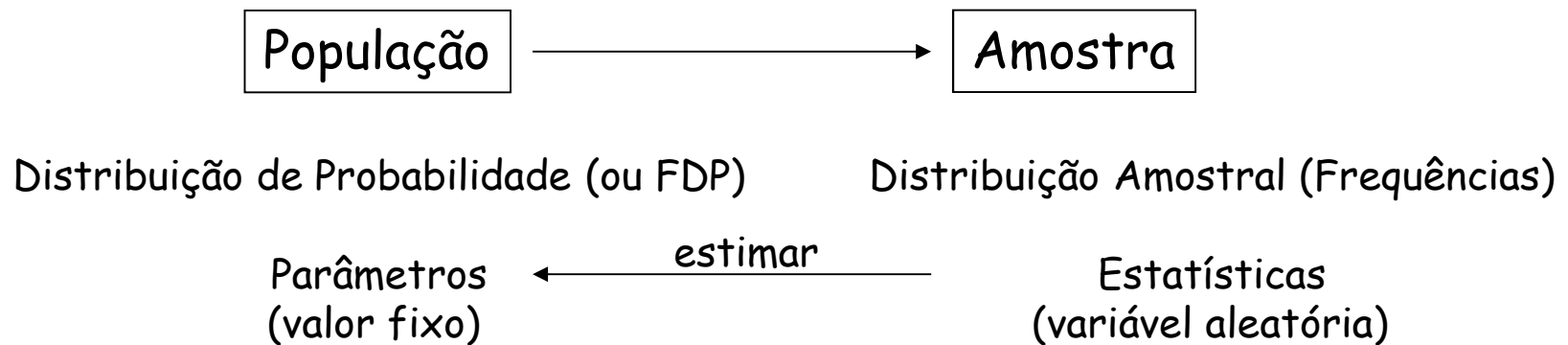
Camilo Daleles Rennó
camilo@dpi.inpe.br

Referata Biodiversa (<http://www.dpi.inpe.br/referata/index.html>)
São José dos Campos, 8 de dezembro de 2011

Inferência Estatística



Estimação de Parâmetros



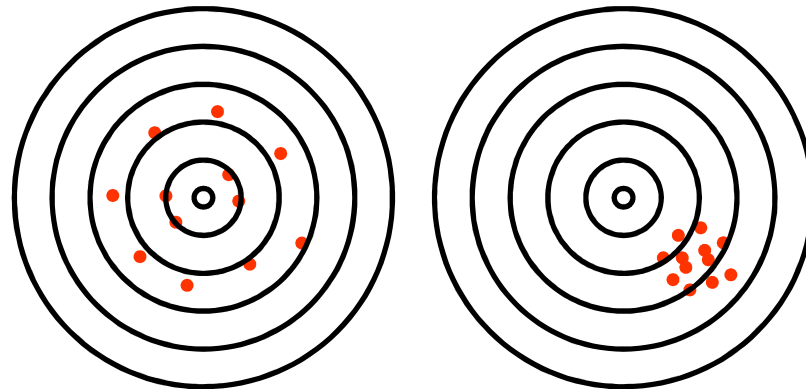
Estimação {
pontual (*estatísticas*)
por intervalo (*intervalos de confiança*)

Como as *estatísticas* são usadas para estimar o parâmetro, elas também são chamadas *estimadores*

Estimação Pontual

Característica ideal de um estimador

- não tendencioso \leftrightarrow exatidão/acurácia
- variância mínima \leftrightarrow precisão/incerteza



Exato
Impreciso

Inexato
Preciso

Tiro ao alvo

Avaliação da Incerteza de um Estimador

Exemplo: Seja X uma v.a. com distribuição desconhecida, com a média (μ) e a variância (σ^2) também desconhecidas. Retira-se uma amostra de tamanho n com a finalidade de se estimar μ .

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \underbrace{\sum_{j=1}^N x_j FR(X = x_j)}_{\text{dados agrupados}} \quad \text{Var}(\hat{\mu}) = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Como \bar{X} é uma v.a., qual sua distribuição?

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

se X tiver distribuição normal
ou
 n for grande (TLC)

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \quad (\text{Normal Padrão})$$

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \quad (\text{t de student})$$

desvio padrão amostral

Desvantagens da Estatística Clássica

- nem todos os estimadores têm sua distribuição amostral facilmente definida, mesmo quando se conhece a distribuição original da variável aleatória estudada

exemplo: mediana, coeficientes de regressões não lineares, etc

- quando a amostra é pequena, certas suposições podem não ser válidas, dificultando a obtenção da distribuição amostral de um estimador qualquer.

exemplo: média amostral pode não ter distribuição normal (amostra pequena = TLC inválido)

Suposições de algumas Estatísticas Clássicas

- ANOVA (comparação entre r médias)
r populações normalmente distribuídas com variâncias iguais

- regressão

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$Y_i = \beta_0 X_i^{\beta_1} \varepsilon_i \quad \log \varepsilon_i \sim N(0, \sigma^2)$$

- proporção

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n} \quad \text{amostras "grandes"}$$

- índice de concordância Kappa

$$\hat{\kappa} = \frac{\theta_1 - \theta_2}{1 - \theta_2} \quad \text{Var}(\hat{\kappa}) = \frac{1}{n} \left[\frac{\theta_1(1-\theta_1)}{(1-\theta_2)^2} + \frac{2(1-\theta_1)(2\theta_1\theta_2 - \theta_3)}{(1-\theta_2)^3} + \frac{(1-\theta_1)^2(\theta_4 - 4\theta_2^2)}{(1-\theta_2)^4} \right]$$

$$\frac{\hat{\kappa} - \kappa}{\sqrt{\text{Var}(\hat{\kappa})}} \sim N(0,1) \quad \frac{(\hat{\kappa}_1 - \hat{\kappa}_2) - (\kappa_1 - \kappa_2)}{\sqrt{\text{Var}(\hat{\kappa}_1) + \text{Var}(\hat{\kappa}_2)}} \sim N(0,1) \quad \text{amostras "grandes" e independentes}$$

Reamostragem

Testes paramétricos clássicos comparam estatísticas calculadas a partir de uma amostra à distribuições amostrais teóricas.

A reamostragem é o nome que se dá a um conjunto de técnicas ou métodos que se baseiam em calcular estimativas a partir de repetidas amostragens dentro da mesma amostra (única).

Tipos de reamostragem:

- Testes de Aleatorização (Testes de Permutação)
- Validação Cruzada
- Jackknife
- Bootstrap

Testes de Aleatorização

Testes de aleatorização (ou testes de permutação ou testes exatos) são típicos testes de significância onde a distribuição da estatística testada é obtida calculando-se todos os possíveis valores desta estatística rearranjando-se os valores da amostra considerando uma hipótese nula verdadeira.

Pode-se usar a simulação Monte Carlo quando número exato de permutações é muito grande.

Região	Área corretamente classificada		Dif
	1 imagem	2 imagens	
1	70	117	47
2	51	48	-3
3	60	63	3
4	57	90	33
5	43	41	-2
6	15	21	6
7	25	36	11
8	103	122	19

Dif média = 14,25

Qual valor esperado caso não houvesse diferença entre o número de imagens utilizadas?

Quão raro seria encontrar o valor 14,25 nesse caso?

(ver exemplos.xls)

Validação Cruzada

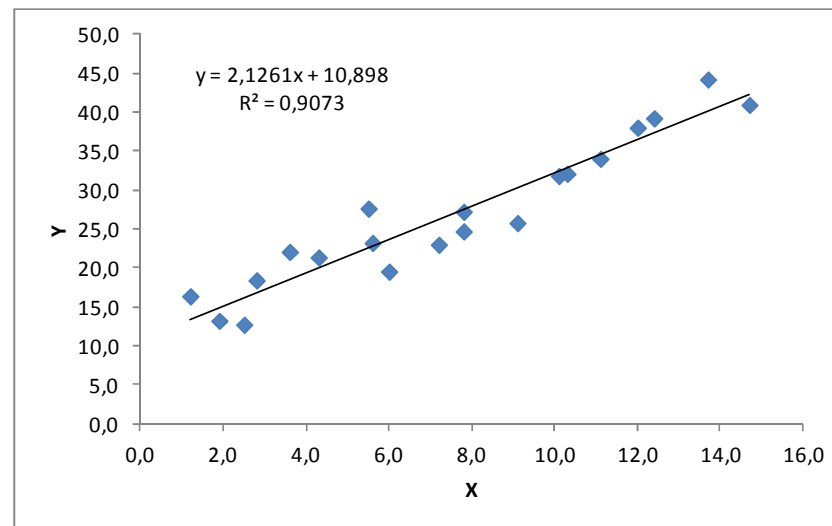
Tipicamente, na validação cruzada, a amostra é dividida aleatoriamente em dois subconjuntos: um de treinamento e outro de teste (validação).

Num estudo de regressão, por exemplo, um conjunto pode ser usado para calcular os coeficientes da equação e o outro para comparar com os valores estimados por esta regressão.

Esta análise pode ficar comprometida quando a amostra é muito pequena.

X	Y
1,2	16,4
1,9	13,3
2,8	18,4
4,3	21,4
5,5	27,7
6,0	19,6
7,2	23,0
7,8	27,2
9,1	25,8
10,3	32,1

X	Y
11,1	34,0
13,7	44,2
14,7	41,0
2,5	12,8
3,6	22,1
5,6	23,3
7,8	24,7
10,1	31,9
12,0	38,0
12,4	39,2



(ver exemplos.xls)

Jackknife

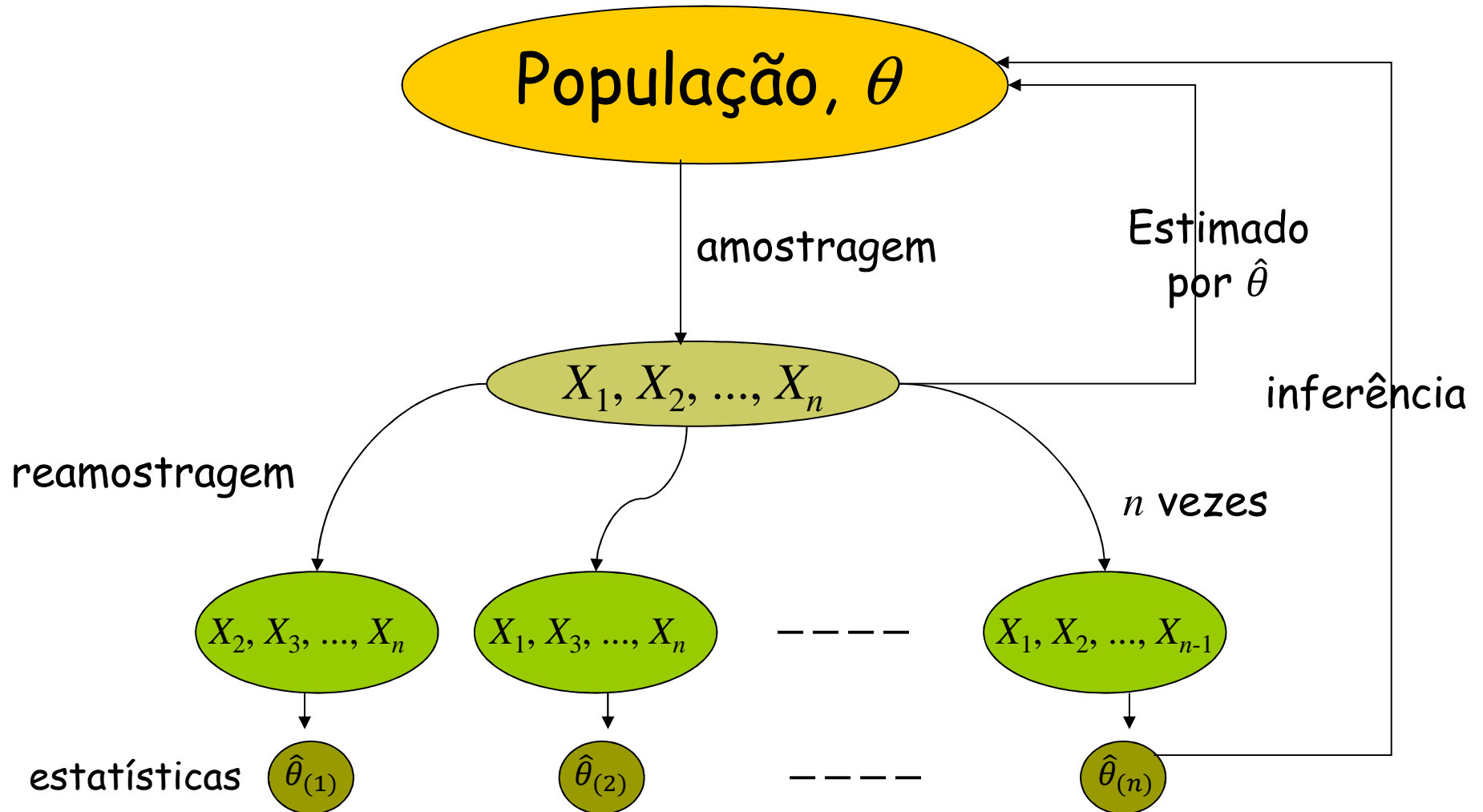
Também chamado "leave-one-out"

Usado para estimar a variância e a tendência de um estimador qualquer.

Baseia-se na remoção de 1 amostra (podendo ser mais) do conjunto total observado, recalculando-se o estimador a partir dos valores restantes.

É de fácil implementação e possui número fixo de iterações (n caso se retire apenas 1 amostra por vez).

Jackknife



Variância de Jackknife

Suponha que um determinado parâmetro θ pode ser estimado a partir de uma amostra de n valores, ou seja,

$$\hat{\theta} = f(x_1, x_2, \dots, x_n)$$

Então a i -ésima replicação Jackknife corresponde ao valor estimado sem a amostra i :

$$\hat{\theta}_{(i)} = f(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

Define-se o i -ésimo pseudovalor como:

$$x_{(i)}^* = N\hat{\theta} - (N-1)\hat{\theta}_{(i)}$$

Variância de Jackknife

Com base nos pseudovalores, pode-se calcular então:

$$\hat{\theta}_{jk} = \frac{1}{n} \sum_{i=1}^n x_{(i)}^* = N\hat{\theta} - (N-1)\hat{\theta}_{(\cdot)} \quad \text{onde} \quad \hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}$$

$$\text{Var}_{jk}(\hat{\theta}) \cong \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2$$

	X
1	2,2
2	3,5
3	3,4
4	6,7
5	6,2
6	8,2
7	9,2
8	7,9
9	9,0
10	10,1

Qual a média geométrica?

Qual a incerteza associada a esta estimativa?

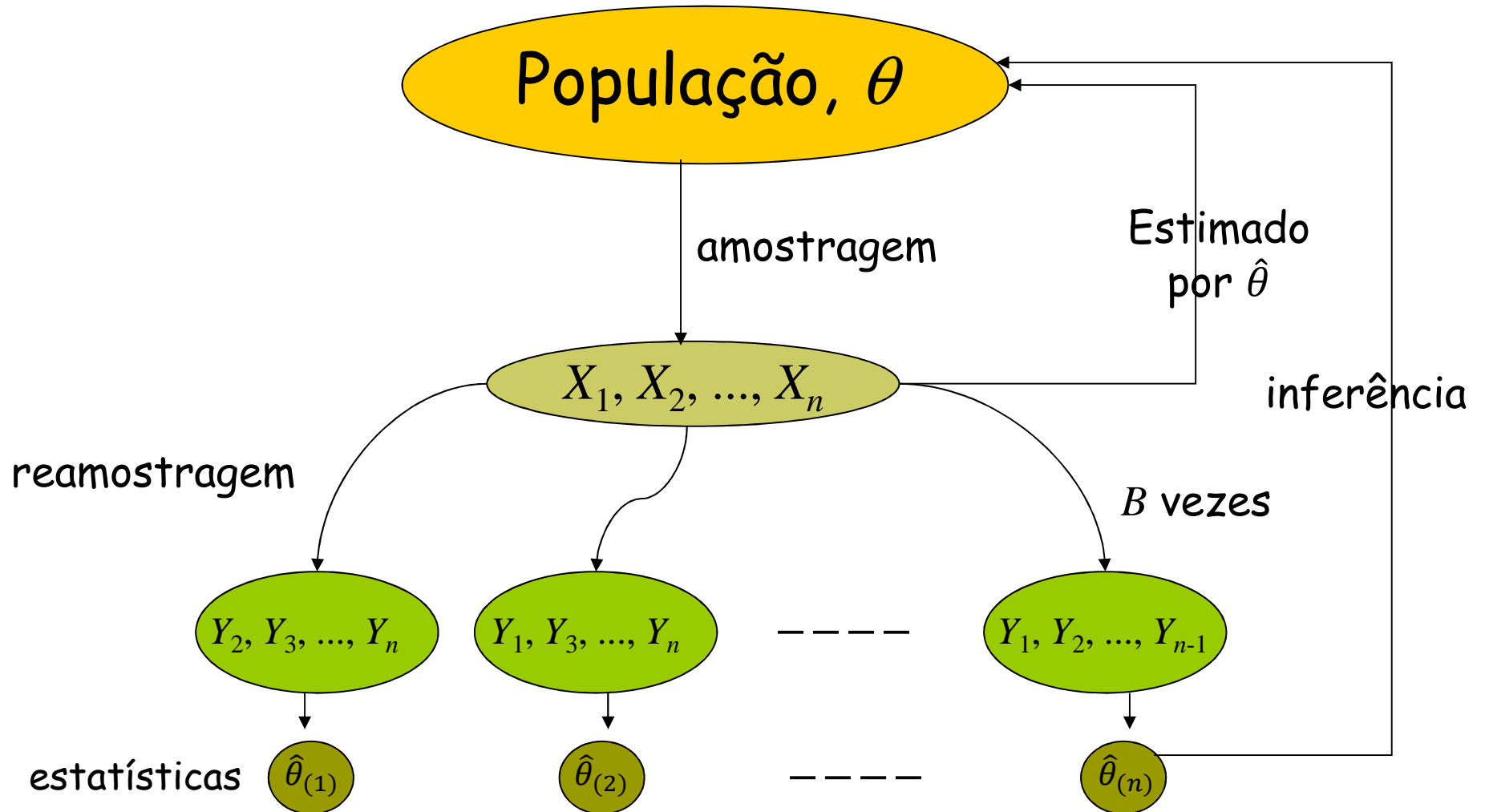
(ver exemplos.xls)

Bootstrap

Pode ser considerado uma estratégia mais abrangente que o Jackknife por permitir um maior número de replicações. Também é usado para estimar a variância e a tendência de um estimador qualquer.

Baseia-se na geração de uma nova amostra de mesmo tamanho da amostra original, a partir do sorteio aleatório **com reposição** de seus elementos.

Bootstrap



Variância de Bootstrap

Suponha que um determinado parâmetro θ pode ser estimado a partir de uma amostra de n valores, ou seja,

$$\hat{\theta} = f(x_1, x_2, \dots, x_n)$$

Então a cada iteração j o valor estimado a partir da amostra será:

$$\hat{\theta}_{(j)} = f(y_1, y_2, \dots, y_n) \quad \text{onde } y_i \text{ é um dos valores da amostra (com reposição)}$$

Variância de Bootstrap

Com base nas estimativas, pode-se calcular então:

$$\hat{\theta}_b = \frac{1}{n} \sum_{j=1}^n \theta_{(j)}$$

$$\text{Var}_b(\hat{\theta}) \cong \frac{1}{m-1} \sum_{j=1}^m (\hat{\theta}_{(j)} - \hat{\theta}_{(\cdot)})^2$$

	X
1	2,2
2	3,5
3	3,4
4	6,7
5	6,2
6	8,2
7	9,2
8	7,9
9	9,0
10	10,1

Qual a média geométrica?

Qual a incerteza associada a esta estimativa?

(ver exemplos.xls)