

---

# Estatística: Aplicação ao Sensoriamento Remoto

SER 204

Teoria da amostragem

Camilo Daleles Rennó

camilo.renno@inpe.br

acesso do conteúdo do curso em [Bibdigital do INPE](#) ou [GitHub](#)

# Algumas Considerações...

---

É importante ter consciência de que dominar as técnicas estatísticas não é suficiente para garantir o sucesso de uma análise, ou seja, conseguir chegar a conclusões “interessantes”.

De forma geral, para que as análises estatísticas sejam válidas, as amostras devem representar a população, ou seja, a menos que discrepâncias ocorram devido ao acaso, as amostras devem reproduzir as mesmas características da população considerando a variável estudada.

É fundamental que as amostras sejam obtidas por processos adequados de modo a evitar que erros grosseiros possam comprometer a análise dos dados.

# Algumas Considerações...

---

Em muitos casos, é bastante tentador que as observações mais convenientes sejam as selecionadas para compor uma amostra ou então aplicar algum tipo de critério (ou julgamento) no momento dessa seleção.

Nesses casos, pode-se introduzir algum tipo de tendência que poderá causar uma super ou subestimativa dos parâmetros de interesse. A identificação (e descrição) desta tendência pode ser difícil (ou impossível) de ser feita após a coleta dessas amostras.

Algumas vezes, há sim um propósito explícito na escolha da amostra com o objetivo de se escolher poucas amostras representativas da população.

No entanto, de modo geral, a seleção das amostras é feita através de algum processo aleatório, de modo que qualquer elemento da população tenha igual chance de ser escolhido para compor a amostra.

# Censo ou Amostragem?

---

Por que fazer **Censo**?

- a população é pequena ou amostragem "ideal" é quase tão grande quanto a população
- necessita-se de uma precisão completa (não é permitido nenhum erro)
- os dados de toda população já se encontram disponíveis

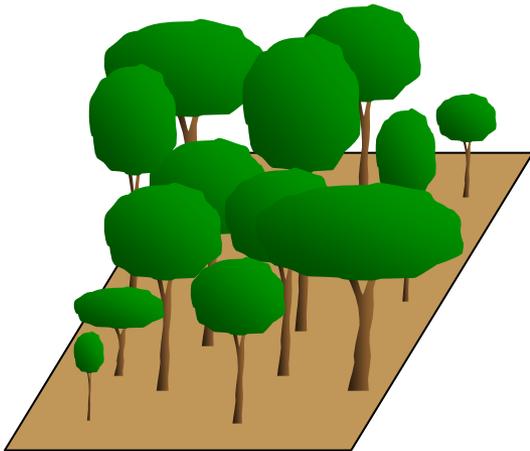
Por que fazer **Amostragem**?

- a população é infinita (ou muito grande)
- os custos de obtenção das medidas são elevados (análises muito caras)
- o tempo para caracterização da população é muito longo
- deseja-se aumentar a representatividade, amostrando-se diferentes populações
- necessita-se melhorar a precisão das medidas (mais cuidado na obtenção dos dados)
- a obtenção das medidas requer a destruição das amostras (p. ex: biomassa)

# Amostragem

---

Toda amostragem requer **planejamento**



- a) O que quero caracterizar neste estudo?  
algum parâmetro específico (média, variância, etc),  
distribuição espacial e/ou variação temporal é importante?
- b) Qual é a unidade amostral apropriada para o estudo?  
quem é o elemento da população (unidade amostral)?
- c) Como estas amostras devem ser coletadas?  
há variabilidade espacial e temporal?  
quais fatores podem influenciar nos resultados?
- d) Quantas amostras são necessárias?  
qual é a precisão exigida?  
quanto tempo e recurso disponho?

# Unidade Amostral

---

A unidade amostral representa a menor entidade identificada na população e é considerada o objeto de estudo.

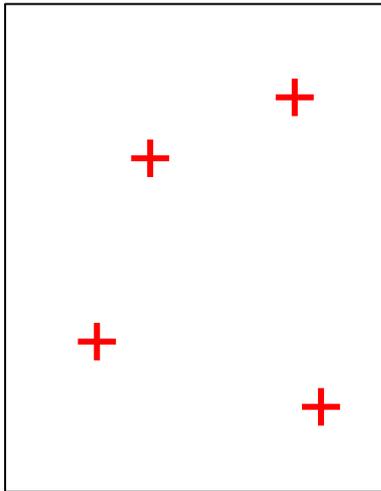
Ela constitui o elemento da população na qual são coletadas as medidas ou informações (qualitativas ou quantitativas) que serão analisadas.

Em estudos na área de Sensoriamento Remoto e Geoprocessamento, podem ser representados por:

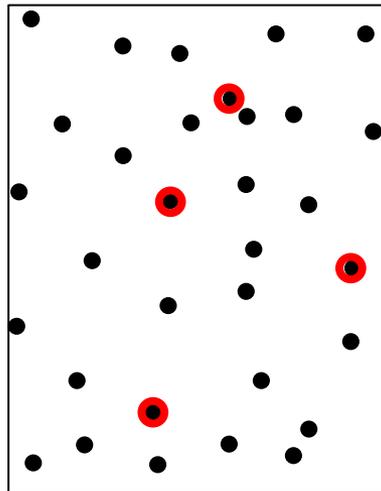
- pontos
- objetos (polígonos ou linhas)

# Unidade Amostral

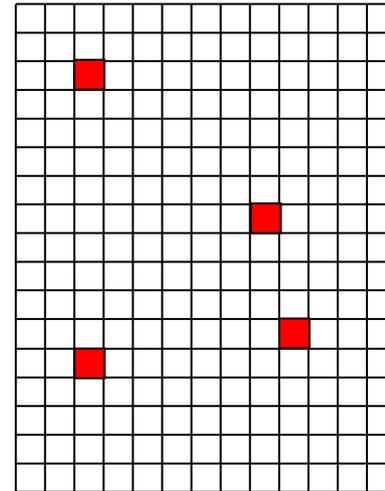
## Pontos



posição no espaço  
(p.ex. ponto num lago)



indivíduo da população  
(p.ex. árvore numa floresta)

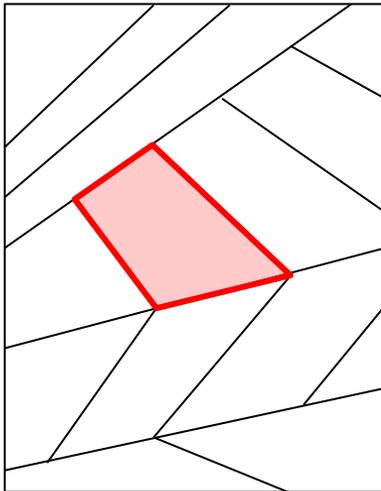


*pixel* da imagem ou grade

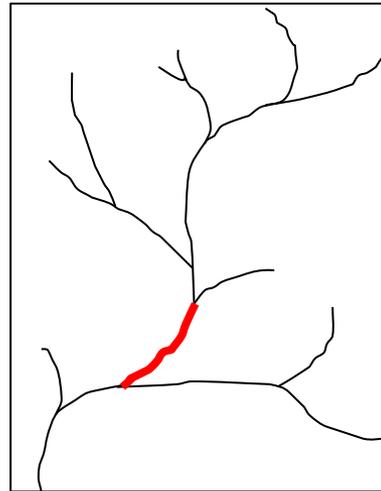
- sorteio aleatório é facilitado
- em coletas em campo, a localização precisa do ponto sorteado pode ser difícil
- pode induzir a erros em regiões heterogêneas

# Unidade Amostral

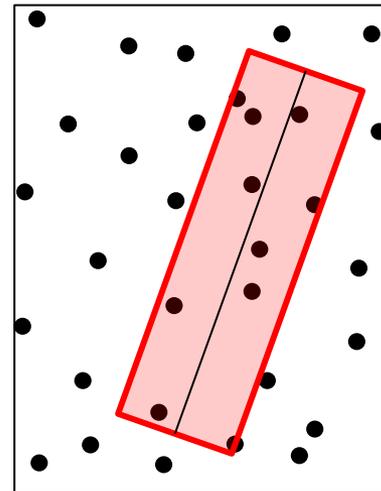
## Objetos



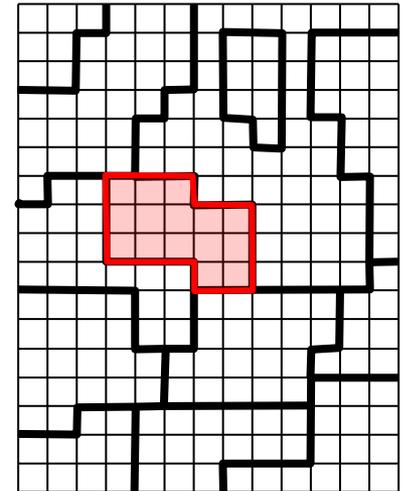
região no espaço  
(p.ex. talhão agrícola)



linha  
(p.ex. trecho de rio)



indivíduos da população  
(p.ex. transecto)



segmento

- deve representar áreas homogêneas (deve-se evitar áreas de transição)
- em coletas de campo, minimiza problemas de posicionamento quando informação contextual é considerada
- mesmo podendo conter muitos valores medidos, deve ser contabilizado como apenas uma observação e portanto deve-se adotar uma medida representativa (total, média, mediana, etc)

# Tipos de Amostragem

---

## Como amostrar?

amostragem probabilística X não probabilística

## Amostragem probabilística:

cada elemento da população tem uma probabilidade (não nula) de ser escolhido  
em geral, todo elemento tem a mesma probabilidade de ser escolhido



Neste tipo de amostragem, todos os elementos devem ser previamente identificados e a escolha é feita por sorteio realizado posteriormente e de forma independente

# Tipos de Amostragem

## Como amostrar?

amostragem probabilística X não probabilística

## Amostragem probabilística:

cada elemento da população tem uma probabilidade (não nula) de ser escolhido  
em geral, todo elemento tem a mesma probabilidade de ser escolhido



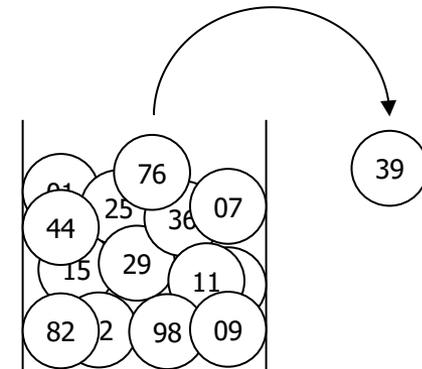
# Tipos de Amostragem

## Como amostrar?

amostragem probabilística X não probabilística

## Amostragem probabilística:

cada elemento da população tem uma probabilidade (não nula) de ser escolhido  
em geral, todo elemento tem a mesma probabilidade de ser escolhido



# Tipos de Amostragem

---

## Como amostrar?

amostragem probabilística X não probabilística

## Amostragem não probabilística:

escolha a esmo (rotulagem inviável ou impossível)

populações muito grandes (ex: estudo sobre a variabilidade no DAP em talhões de reflorestamento de eucalipto)

populações dinâmicas (ex: estudo sobre qualidade de água num rio)

amostragem restrita aos elementos que se tem acesso (ex: estudo sobre ocorrência de focos de dengue em casas de veraneio)

amostragem intencional ou por julgamento (ex: estudo sobre diversidade florística de plantas com DAP maior que 30cm dentro de um transecto)

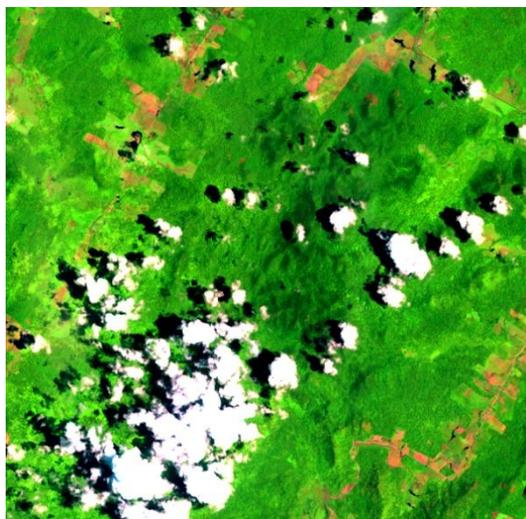
voluntários (ex: estudo sobre a eficácia de uma nova vacina contra febre amarela)

**OBS:** escolha a esmo é a abordagem que mais se assemelha à amostragem probabilística desde que se garanta que não haja nenhum tipo de influência na seleção das amostras

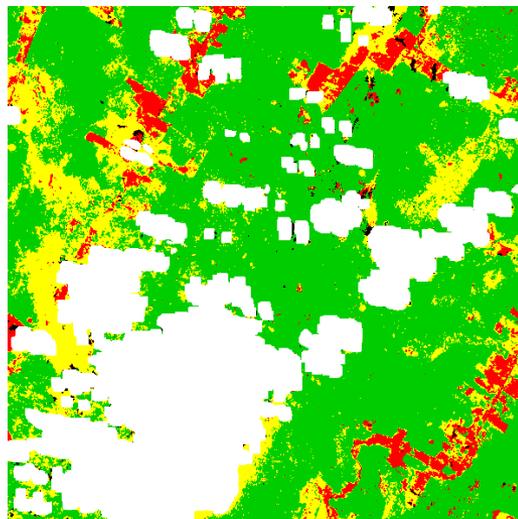
# Tipos de Amostragem

Do ponto de vista estatístico, a amostragem probabilística é a ideal

Sempre que uma abordagem não probabilística for adotada, deve-se explicitá-la no trabalho de pesquisa



OLI/Landsat R6G5B4



Classificação



Numa análise sobre a qualidade da classificação, deve-se explicitar que as regiões marcadas como “Não Classificado” e “Não Observado” não serão consideradas na avaliação

Nesse caso, a amostragem não é tipicamente probabilística pois os pixels pertencentes a essas classes não podem ser sorteados (probabilidade nula)

# Desenho amostral (*Sampling Design*)

---

O Desenho Amostral define como as amostras serão coletadas.

A escolha da melhor estratégia dependerá:

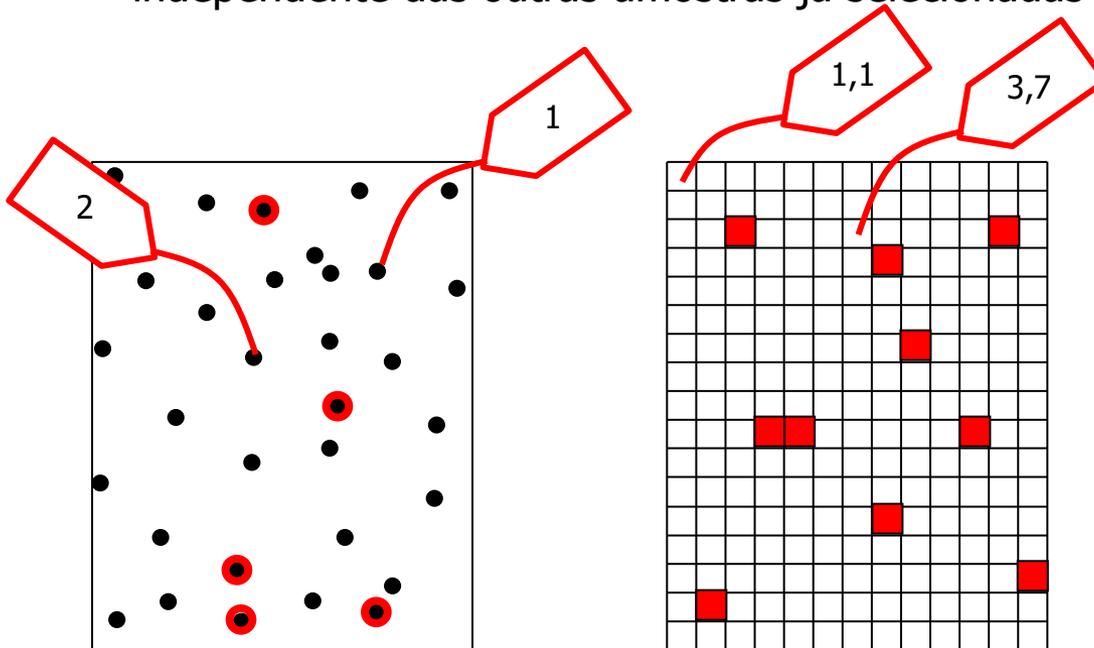
- da facilidade e praticidade de implementação
- dos custos para obtenção das amostras
- da heterogeneidade espacial dos dados (distribuição espacial)

Decisões chaves:

- usar abordagem simples ou sistemática?
- usar ou não uma amostragem estratificada?
- selecionar amostras isoladas ou em conglomerados (*clusters*)?

# Amostragem Aleatória Simples

Nesta abordagem, a escolha de uma amostra é feita de modo totalmente independente das outras amostras já selecionadas



etapas:

- rotular cada elemento com um código único
- sortear aleatoriamente  $n$  códigos (usando-se geradores de números aleatórios)
- identificar os elementos com os códigos selecionados

OBS: método simples

rotulação dos elementos pode ser dispendiosa

pressupõe população homogênea

não garante representatividade pois alguns

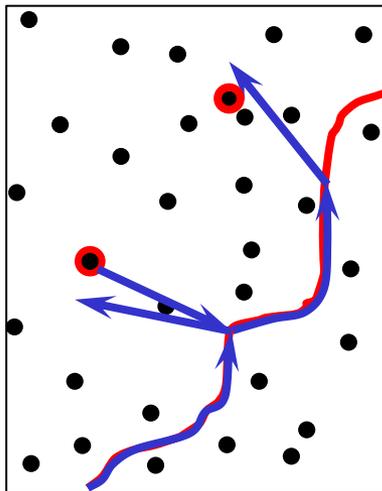
grupos (mais raros) podem não ser sorteados

# Amostragem Aleatória Simples

Em trabalhos de campo, muitas vezes não é possível fazer a identificação prévia dos elementos

Nesse caso, é usual fazer a **escolha a esmo** dos elementos amostrados usando artifícios que garantam a escolha imparcial

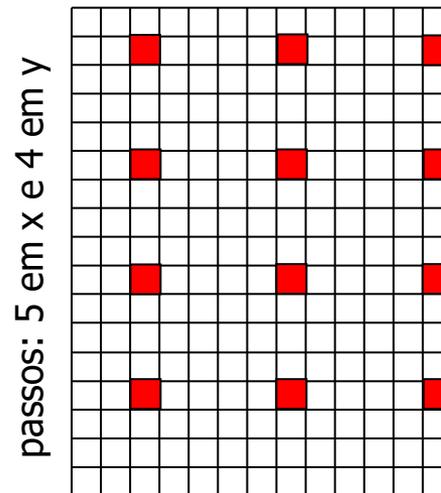
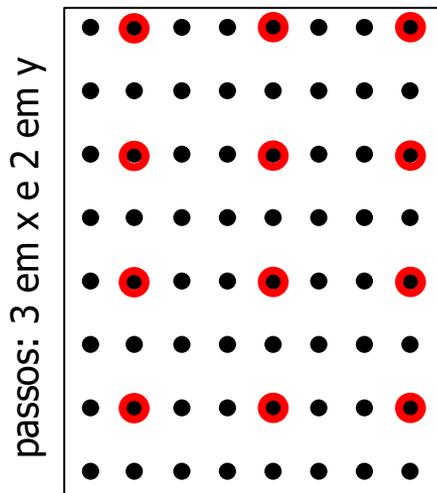
Exemplo: numa floresta, deseja-se amostrar 10 árvores



- numa trilha, caminha-se **x metros**
- caminha-se **y metros** numa determinada **direção**
- escolhe-se a árvore mais próxima
- faz-se as medições necessárias
- retorna-se ao ponto inicial
- repete-se o procedimento até selecionar-se as 10 árvores

# Amostragem Sistemática

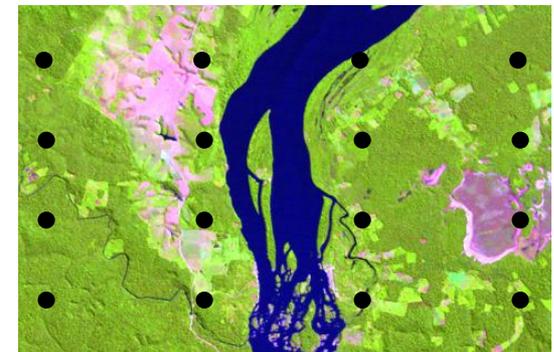
Se os elementos da população já se encontram ordenados segundo algum critério, pode-se selecionar um elemento qualquer e escolher um "passo" que definirá qual será o próximo elemento escolhido.



etapas:

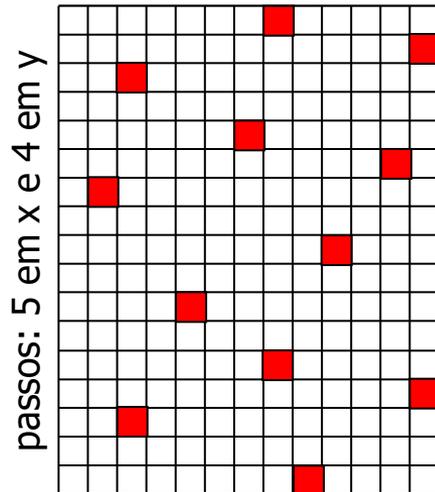
- definir o passo (ou os passos em x e em y)
- escolher aleatoriamente um elemento
- com base nesse elemento, identificar os demais elementos de acordo com o passo pré-definido

OBS: amostra-se uniformemente todo o espaço  
pode-se não conseguir o valor exato de amostras pretendidas  
desaconselhado para ordenações periódicas ou com feições dispostas na horizontal e/ou vertical



# Amostragem Sistemática Não Alinhada

A ideia é semelhante da amostragem sistemática mas, nesse caso, tenta-se aleatorizar os passos de modo a desalinhar as amostras sorteadas.

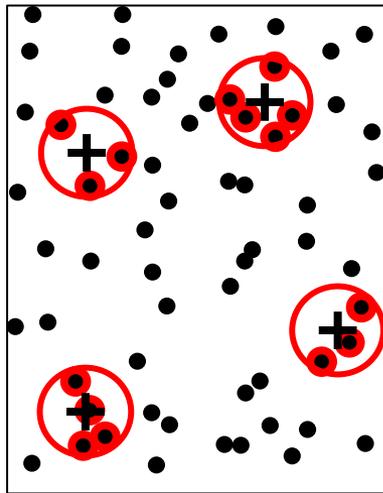


etapas:

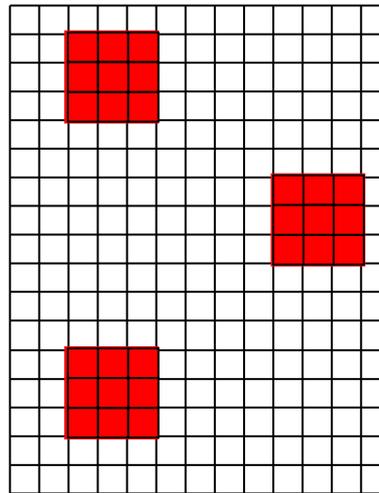
- definir o passo (ou os passos em x e em y)
- escolher aleatoriamente um elemento
- com base nesse elemento, identificar os elementos da mesma linha e mesma coluna de acordo com o passo pré-definido
- desalinhar aleatoriamente esses elementos
- utilizar esses novos posicionamentos para identificar os demais elementos

# Amostragem em Conglomerados (*Cluster*)

Nesta abordagem, a amostra é formada por um grupo de elementos próximos (cluster)



raio  $r$



janela 3x3

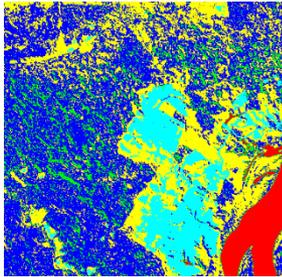
etapas:

- definir critério de proximidade (raio ou janela)
- sortear aleatoriamente  $n$  posições
- identificar os elementos que atendam o critério de proximidade

OBS: simplifica a coleta de dados das amostras  
cada elemento do conglomerado constitui uma unidade amostral  
diminui os custos da amostragem  
pode reduzir a precisão na estimação devido a autocorrelação espacial

# Amostragem Estratificada

A estratificação é a divisão da área de estudo em regiões segundo algum critério (mapas pré-existentes ou regiões geográficas)



Mas para que estratificar?

- os estratos representam regiões de interesse no estudo  
p.ex., estimar a área desmatada por Estado ou município
- deseja-se melhorar a precisão nas estimativas obtidas em cada estrato  
como quanto maior heterogeneidade, maior incerteza na estimativa. Pode-se  
assim concentrar a amostragens nos estratos com maior variabilidade
- deseja-se aumentar a representatividade da amostra coletada na área de estudo  
estratos raros podem não estar representados adequadamente numa  
amostragem totalmente aleatória

Dentro de cada estrato, pode-se adotar a Amostragem Aleatória Simples, Sistemática ou Sistemática Não-Alinhada. Além disso, pode-se inclusive selecionar elementos em conglomerados

# Tamanho da Amostra

---

## Quanto amostrar?

depende:

da variabilidade original dos dados (maior variância  $\Rightarrow$  maior  $n$ )

da precisão requerida no trabalho (maior precisão  $\Rightarrow$  maior  $n$ )

do tempo disponível (menor o tempo  $\Rightarrow$  menor  $n$ )

do custo da amostragem (maior o custo  $\Rightarrow$  menor  $n$ )

Em geral, é calculado com base no parâmetro que se deseja estimar e leva em

consideração as **incertezas inerentes** a esta estimação:

a) variação "natural" dos dados (variância populacional)

b) erros de estimativa

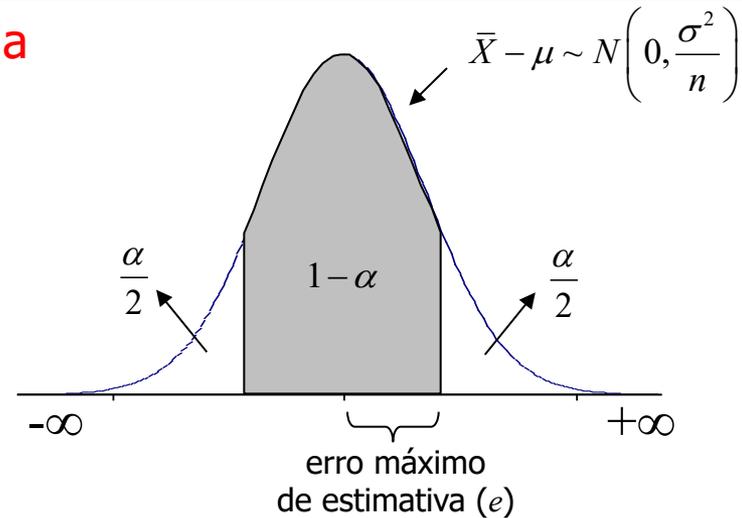
# Tamanho da Amostra

$$\bar{X} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

Média

$$P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Rightarrow n = \frac{(z_{\alpha/2})^2 \sigma^2}{e^2}$$



- A variância populacional deve ser conhecida
- Pode-se estimar a variância populacional através de uma "pré-amostragem" nesse caso, o tamanho da amostra é aquele que for viável (tempo, custo, etc) se o tamanho estimado for maior que o utilizado na pré-amostragem, complementa-se a amostragem, reestima-se a variância e recalcula-se o tamanho da amostra até atingir o tamanho ideal

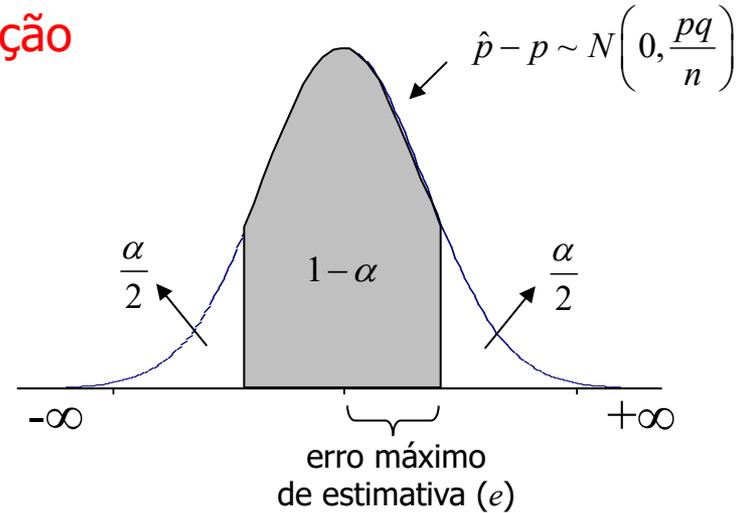
# Tamanho da Amostra

Proporção

$$\hat{p} - p \sim N\left(0, \frac{pq}{n}\right)$$

$$P\left(-z_{\alpha/2} \sqrt{\frac{pq}{n}} < \hat{p} - p < z_{\alpha/2} \sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

$$e = z_{\alpha/2} \sqrt{\frac{pq}{n}} \Rightarrow n = \frac{(z_{\alpha/2})^2 pq}{e^2}$$



- Necessita-se conhecer o parâmetro  $p$  que se quer estimar!
- Pode-se estimar  $p$  através de uma "pré-amostragem"
- Pode-se adotar o valor de  $p = 0,5$  que representa o "pior caso"

$$p = 0,5 \rightarrow \text{máx } Var(\hat{p}) \rightarrow \text{máx } n$$

# Tamanho da Amostra

---

Correção para populações finitas

(quando a amostra representa mais que 5% da população)

$$n' = \frac{n}{1 + \frac{n-1}{N}}$$

$n$  = tamanho de amostra sem correção

$N$  = tamanho da população

$n'$  = tamanho de amostra corrigido

Para média: 
$$n' = \frac{N\sigma^2(z_{\alpha/2})^2}{(N-1)e^2 + \sigma^2(z_{\alpha/2})^2}$$

Para proporção: 
$$n' = \frac{Npq(z_{\alpha/2})^2}{(N-1)e^2 + pq(z_{\alpha/2})^2}$$

# Tamanho da Amostra

Exemplo: Deseja-se estimar a exatidão de um mapa de modo que o valor estimado não ultrapasse em 8% a exatidão verdadeira (para mais ou para menos), utilizando-se um nível de confiança de 95%. Suponha que a exatidão verdadeira é de 80%.

$$n = \frac{(z_{\alpha/2})^2 pq}{e^2}$$

$$n = \frac{1,96^2 0,80 0,20}{0,08^2} = 96,04 \quad n = 96$$

No pior caso (maior variância), a exatidão verdadeira seria de 50%.

$$n = \frac{1,96^2 0,50 0,50}{0,08^2} = 150,06 \quad n = 150$$

# Tamanho da Amostra

Na **Amostragem Estratificada**, como distribuir as amostras em cada estrato?

Suponha que precisamos selecionar  $n$  amostras de uma população de tamanho  $N$  e que esta população está dividida em  $L$  estratos com  $N_1, N_2, \dots, N_L$  elementos.

$$n_i = \frac{n}{L}$$

todos iguais

$$n_i = n \frac{N_i}{N}$$

proporcionais a  $N_i$

$$n_i = n \frac{N_i s_i}{\sum_{i=1}^L N_i s_i}$$

tamanho ótimo  
(considera a variabilidade)

O modo como as amostras são distribuídas entre os estratos têm forte impacto na estimativa de parâmetros globais (que representam toda a população) mas permite maior controle da representatividade e precisão de cada estrato.

# Tamanho da Amostra

O tamanho da amostra ( $n$ ) também pode considerar o erro  $\beta$  (tipo II)

Exemplo para **proporção**

Hipóteses

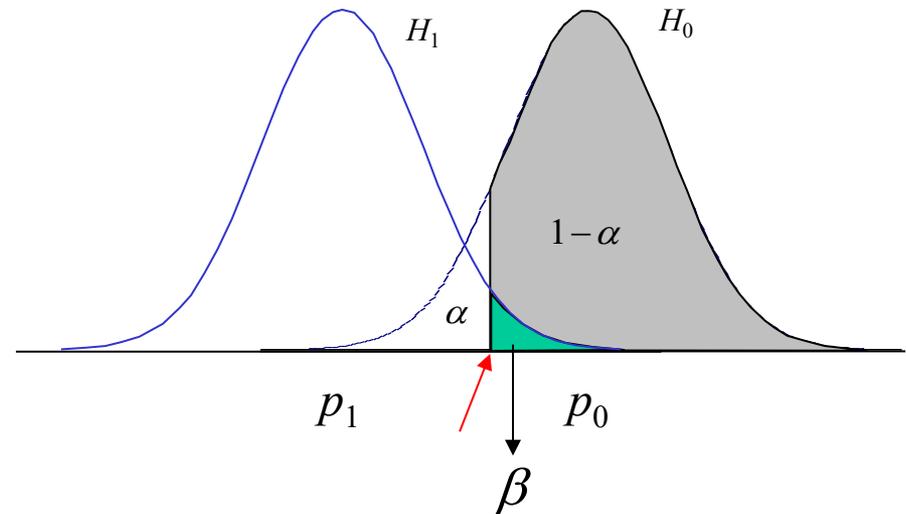
$$H_0 : p = p_0$$

$$H_1 : p < p_0$$

$$P\left(\hat{p} > p_0 - z_\alpha \sqrt{\frac{p_0 q_0}{n}}\right) = 1 - \alpha$$

Considerando  $H_1$  verdadeira ( $p = p_1$ )

$$P\left(\frac{\hat{p} - p_1}{\sqrt{\frac{p_1 q_1}{n}}} > \frac{p_0 - z_\alpha \sqrt{\frac{p_0 q_0}{n}} - p_1}{\sqrt{\frac{p_1 q_1}{n}}}\right) = \beta$$

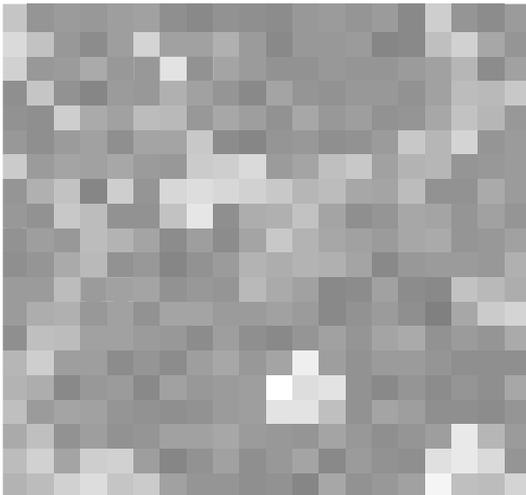


$$P\left(z_\beta \sqrt{\frac{p_1 q_1}{n}} + z_\alpha \sqrt{\frac{p_0 q_0}{n}} > p_0 - p_1\right) = \beta$$

$$n = \frac{\left(z_\beta \sqrt{p_1 q_1} + z_\alpha \sqrt{p_0 q_0}\right)^2}{(p_0 - p_1)^2}$$

# Qual impacto da amostragem estratificada?

Suponha que queremos determinar a média da região abaixo considerando-se que não teríamos como acessar todos os valores mas somente uma amostra. Suponha ainda que tenhamos um mapa que poderia ser utilizado para estratificar as amostras. Qual a vantagem de se dividir as amostras entre os diferentes estratos?



Aleatória Simples

X

Aleatória Estratificada

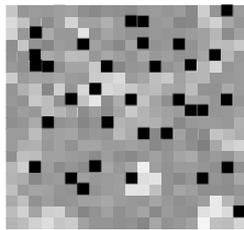
Tamanhos de amostras:

Total	11	20	50	100
	7	14	36	72
	2	4	9	18
	2	2	5	10

\*proporcional por estrato

# Qual impacto da amostragem estratificada?

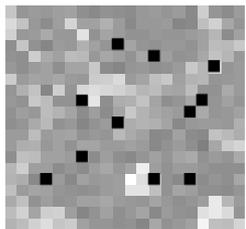
## Aleatória Simples



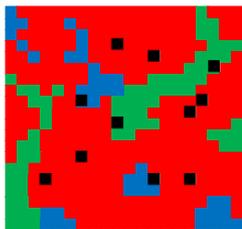
$\mu = 14,03$

#11  
 $\bar{X}_1 = 12,73$   
 $\bar{X}_2 = 15,18$   
 $\bar{X}_3 = 12,45$   
⋮

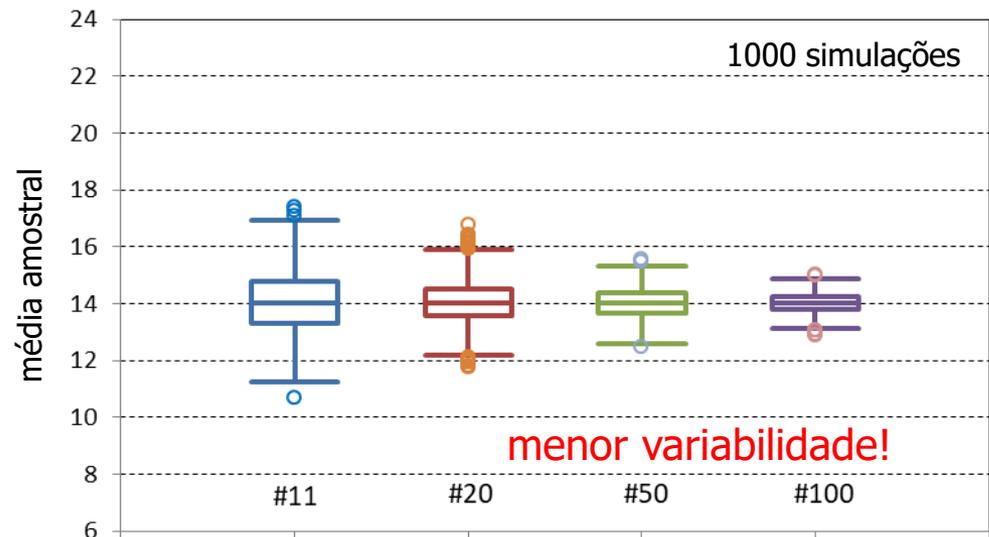
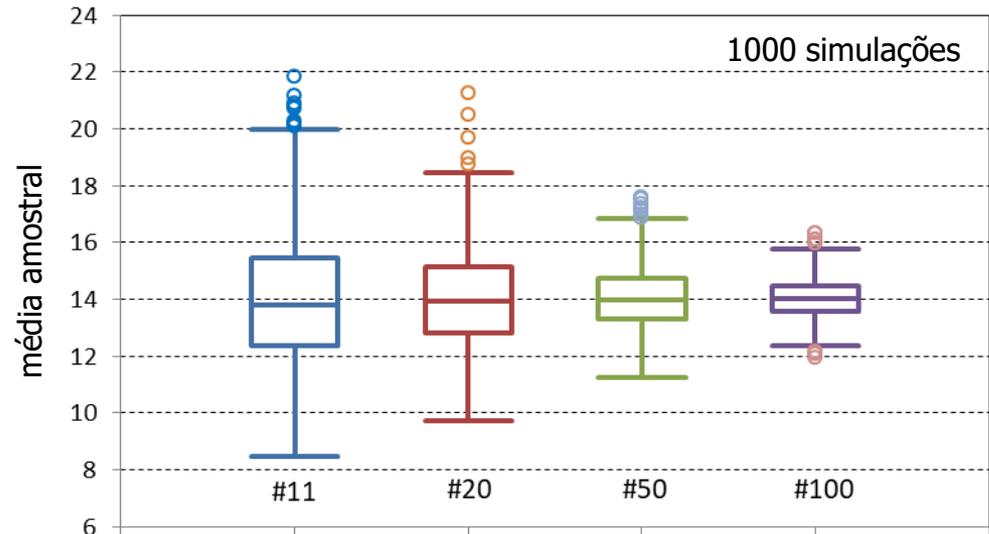
## Aleatória Estratificada



+



#11  
(#7 #2 #2)  
 $\bar{X}_1 = 16,45$   
⋮



# Amostras de Treinamento, Teste e Validação

Numa classificação ou numa modelagem em geral, as amostras são utilizadas para estimar os parâmetros ou para criar as regras usadas pelo classificador/modelo

Como estes ajustes visam minimizar erros, a utilização desse mesmo conjunto amostral para avaliar os resultados do classificador/modelo sempre resultarão numa superestimação dos índices de desempenho.

Dessa forma, é comum reservar parte das amostras de modo a avaliar os resultados de forma independente, gerando índices de desempenho não enviesados (nesse caso, superestimados).

Usualmente, o conjunto amostral total deve ser dividido em 3 partes excludentes:

- Treinamento
- Teste
- Validação

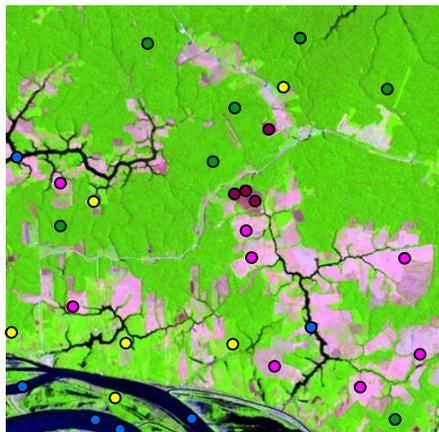
Os termos "teste" e "validação" podem ter seu significado trocado dependendo da literatura consultada ou então constituírem um único grupo

# Amostras de Treinamento, Teste e Validação

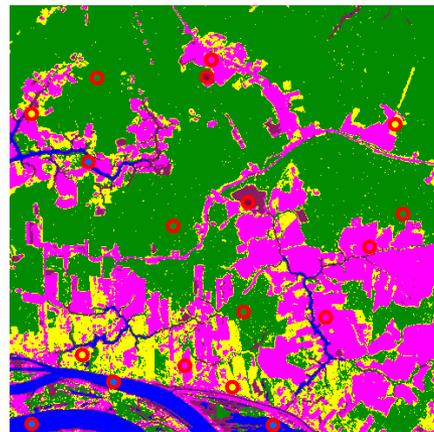
Na prática, a definição das amostras que formarão esses grupos pode ser feita a partir de um único conjunto amostral ou então podem ser coletadas em fases diferentes do processo de classificação/modelagem.

Por exemplo, se o objetivo for avaliar uma classificação única, pode-se

- dividir as amostras coletadas em treinamento e validação, ou
- usar todas as amostras para treinamento e após a obtenção da classificação final, coletar novos pontos que serão avaliados por terceiros de forma totalmente independente.



○ treinamento



○ validação

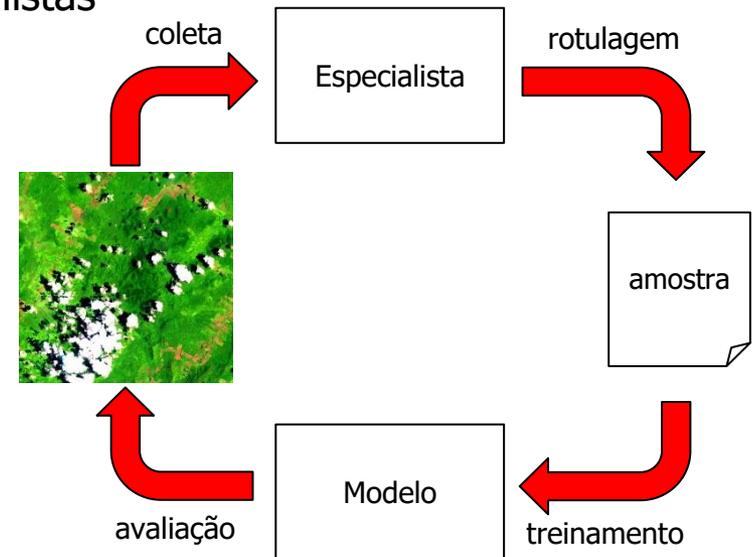
# Amostragem Iterativa

Em muitos casos, o sorteio aleatório de amostras de treinamento pode ser inviável e não garantir a representatividade dos dados analisados

Numa classificação de imagens, a escolha das amostras e posterior rotulação podem ser feitas através de interpretação visual realizada por especialistas

A seleção das amostras pode então ser feita por um processo iterativo:

- inicialmente a coleta de pontos é feita selecionando-se amostras representativas das classes de interesse
- as amostras são utilizadas para treinar o modelo
- o modelo é aplicado sobre os dados e novos pontos são escolhidos com base numa avaliação dos resultados
- repete-se o processo até que os resultados sejam considerados adequados

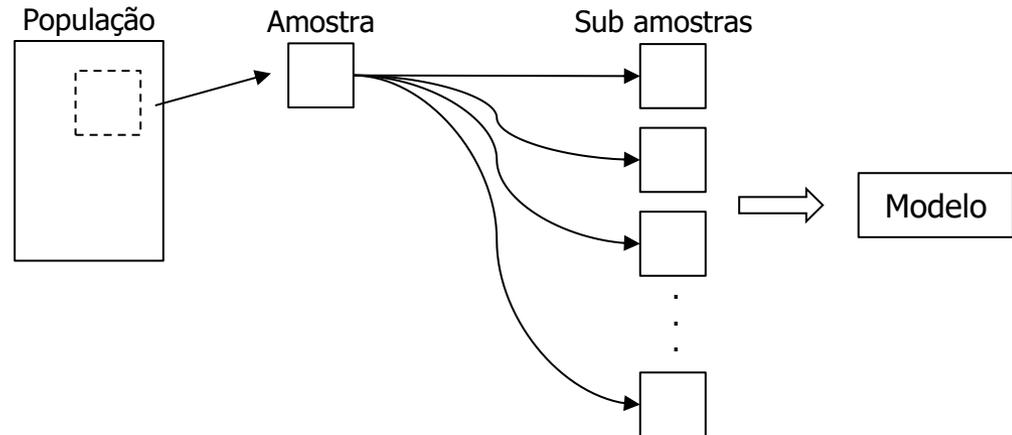


De modo geral, a escolha de novos pontos deve se basear em algum critério que avalie o erro e/ou o grau de incerteza de como um determinado ponto foi classificado/estimado

Uma discussão mais aprofundada sobre erros e incertezas será feita no tema "Avaliação de Classificação"

# Reamostragem

A **Reamostragem** é o nome que se dá a um conjunto de técnicas ou métodos que se baseiam em ajustar ou manipular um conjunto amostral de modo a gerar novas amostragens com o objetivo de melhorar ou avaliar o desempenho de modelos



Estas técnicas se propõem a avaliar as incertezas relacionadas a obtenção de estatísticas com distribuições amostrais desconhecidas ou gerar um rearranjo nos dados de entrada de modo que modelos determinísticos gerem diferentes resultados.

Duas técnicas são comumente empregadas nesse contexto: *Jackknife* e *Bootstrap*

Também podem ser utilizadas para avaliar a significância de testes cujas estatísticas básicas não têm suas propriedades bem estabelecidas ou cujas premissas não podem ser consideradas verdadeiras.

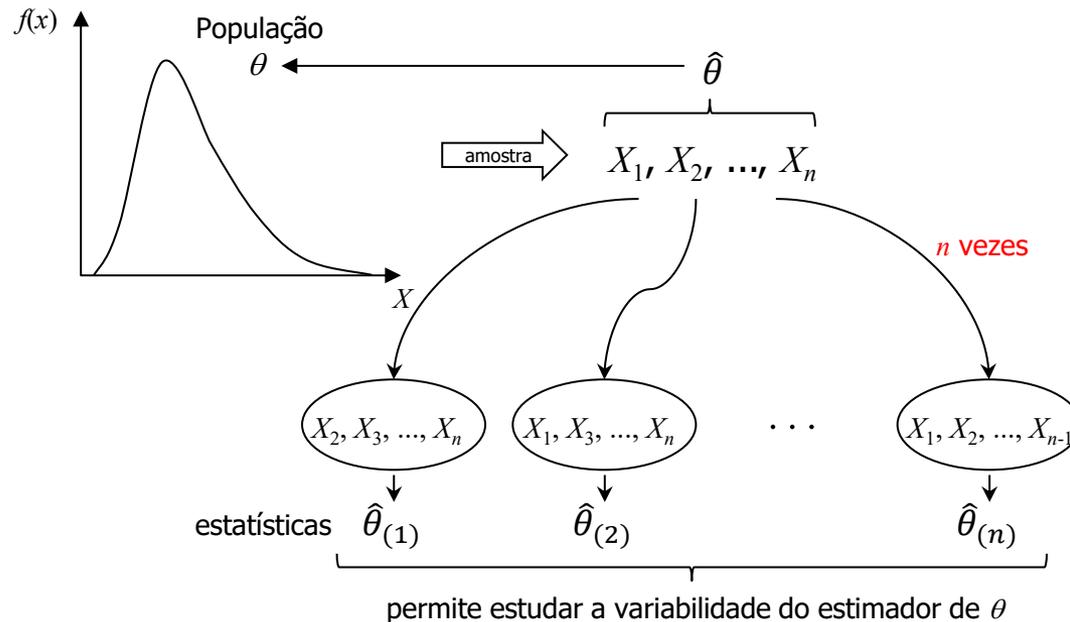
# Jackknife

Também chamado *Leave-One-Out Test*

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

Baseia-se na remoção de 1 amostra do conjunto total observado ( $n$ ), recalculando-se o estimador a partir dos valores restantes ( $n-1$ )

É de fácil implementação e possui número fixo de iterações (igual a  $n$ )



# Jackknife

Suponha que um determinado parâmetro  $\theta$  possa 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  $i$ :

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

Com base nessas  $n$  estimativas, pode-se calcular então:

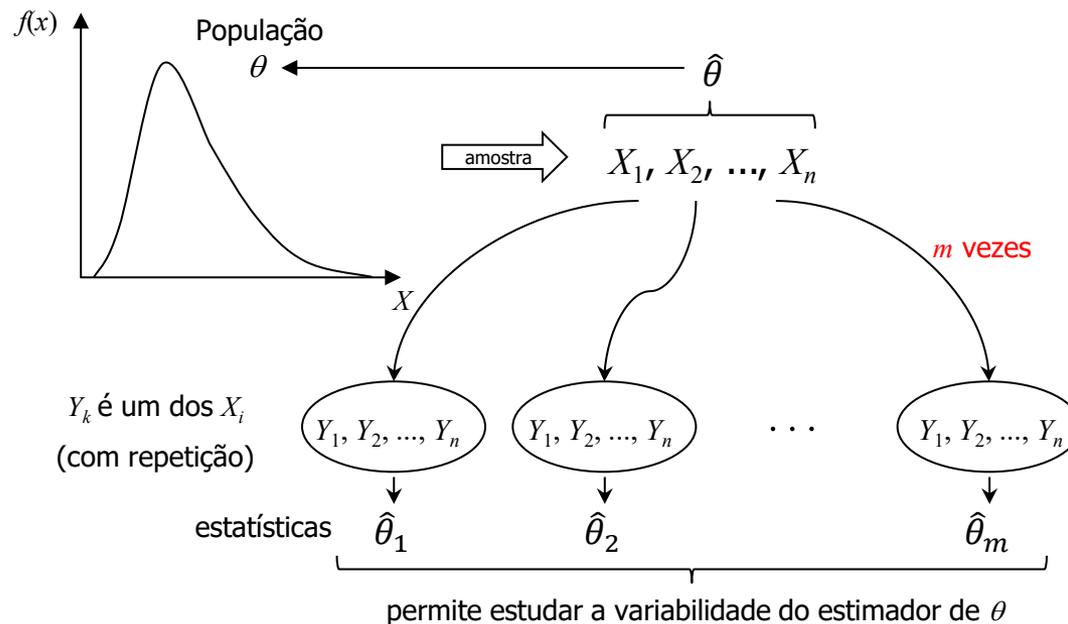
$$\hat{\theta}_{jk} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)} \quad \text{onde} \quad \hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}$$

$$Var_{jk}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \quad \frac{\hat{\theta}_{jk} - \theta}{\sqrt{Var_{jk}(\hat{\theta})}} \sim t_{n-1} \quad (n \text{ grande})$$

# 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

Suponha que um determinado parâmetro  $\theta$  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:

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

Com base nas estimativas de  $m$  iterações, pode-se calcular então:

$$\hat{\theta}_b = \frac{1}{m} \sum_{i=1}^m \hat{\theta}_i$$
$$\frac{\hat{\theta}_b - \theta}{\sqrt{\text{Var}_b(\hat{\theta})}} \sim t_n \quad (n \text{ grande})$$
$$\text{Var}_b(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^m (\hat{\theta}_i - \hat{\theta})^2$$

Recomenda-se que  $m \geq 200$ , mas pode ser necessário  $m \geq 2000$  caso o objetivo seja construir intervalos de confiança para o parâmetro  $\theta$

# Bootstrap - Exemplo

Suponha que se deseja saber qual é a média geométrica de uma população e para isso obteve-se uma amostra de 10 valores:

	$X$
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 é o valor da média geométrica desta amostra e qual a variância deste estimador?

$$mg = \sqrt[10]{2,2 \times 3,5 \times \dots \times 10,1} = 5,9844 \quad (\text{amostra completa})$$

$$Y_1 = \{3,4; 6,7; 8,2; 7,9; 10,1; 9,2; 7,9; 6,2; 3,5; 10,1\}$$

$$mg_1 = 6,8794$$

⋮

$$Y_{200} = \{7,9; 9,2; 9,0; 8,2; 10,1; 8,2; 6,2; 7,9; 9,2\}$$

$$mg_{200} = 8,3158$$

$$mg_b = \frac{1}{200} \sum_{i=1}^{200} mg_i = 6,0703$$

$$Var_b(mg) = \frac{1}{200} \sum_{i=1}^{200} (mg_i - 5,9844)^2 = 0,9611$$

$$s_b(mg) = \sqrt{0,9611} = 0,9804$$

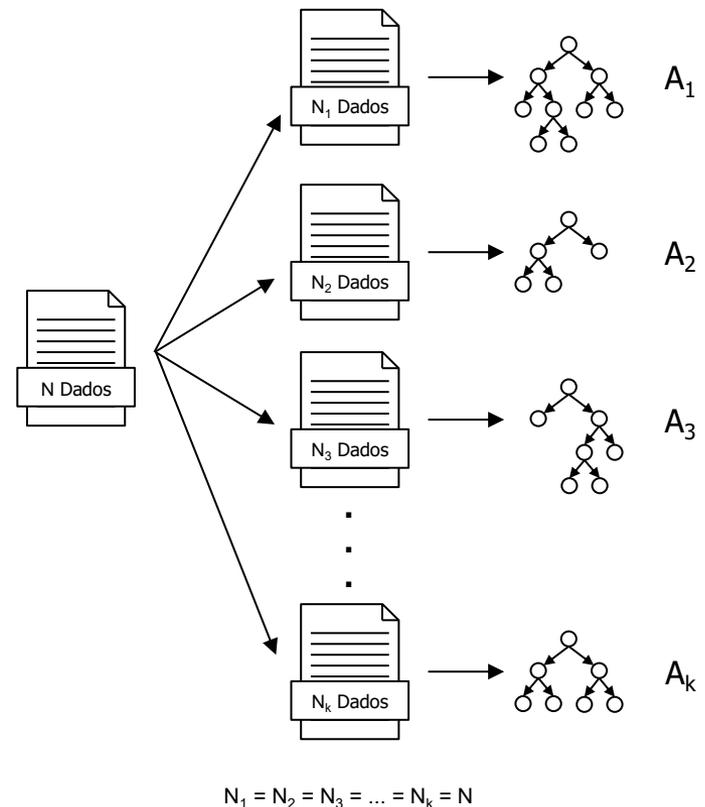
# Bootstrap

No contexto de classificação, o *Bootstrap* pode ser utilizado para avaliar o desempenho do classificador

No processo de sorteio com reposição, parte das amostras não são utilizadas. Em geral, cerca de 30% das amostras ficam de fora e podem ser usadas para avaliação do classificador

No classificador *Random Forest*, por exemplo, cada árvore de decisão é gerada a partir de uma amostra obtida por *Bootstrap*. As amostras não sorteadas, chamadas de OOB (*Out of Bag*), são utilizadas para avaliar o desempenho desse classificador de modo independente pois não foram efetivamente usadas na construção de cada árvore. Todo esse processo é denominado *Bagging* (*Bootstrap Aggregating*)

O objetivo principal é reduzir a influência de dados espúrios (*outliers*) e aumentar a estabilidade (diminuir a variância)



# 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

Uma aplicação típica é quando se deseja comparar dados pareados para avaliar se a diferença entre eles é significativamente grande para justificar que são, de fato, diferentes (por exemplo, pode-se verificar se uma determinada característica mudou de uma data para outra)

Outra aplicação é, num mapa de *Kernel*, identificar se em certas regiões há mesmo uma tendência de valores altos ou baixos, ou se o que se observa pode ter sido obtido casualmente

Nem sempre todos os rearranjos das amostras podem ser avaliados quando a amostragem é muito grande. Nesse caso, faz-se uma simulação de modo a testar o maior número possível de rearranjos (quanto maior o número de simulações melhor!)

# Testes de Aleatorização - Exemplo

Uma determinada característica foi medida 8 vezes em duas datas distintas

Há evidências de que o valor dessa característica aumentou?

Amostra	Valor medido		Dif
	antes	depois	
1	3,5	4,3	0,8
2	4,7	4,4	-0,3
3	5,3	5,9	0,6
4	10,3	11,3	1,0
5	3,8	5,7	1,9
6	6,6	6,4	-0,2
7	5,1	5,1	0,0
8	9,6	10,9	1,3
		média	0,6375

Se  $H_0$  é verdadeiro, então poderia trocar os valores entre os dois grupos

Dif média = 0,6375 ← estatística usada no teste t

Qual valor esperado caso não houvesse diferença na área corretamente classificada quando uma ou duas imagens forem utilizadas ( $H_0$ )?  
**zero**

Quão raro seria encontrar a média 0,6375 nesse caso? Ou seja, qual o valor-P associado a esta estatística?

Solução: calcular todos os valores possíveis de diferença média quando trocamos ou não os valores entre as 2 abordagens para cada amostra. Com isso, obtém-se a distribuição amostral desta estatística.

# Testes de Aleatorização - Exemplo

$H_0$ : não há diferença entre o valor medido antes e depois (Dif média = 0)

$H_1$ : o valor medido depois é maior (Dif média > 0)

Se  $H_0$  é verdadeira, então haverá  $2^8$  possibilidades de trocas, gerando 256 resultados diferentes

Amostra	Valor medido		Dif
	antes	depois	
1	3,5	4,3	0,8
2	4,4	4,7	0,3
3	5,3	5,9	0,6
4	10,3	11,3	1,0
5	3,8	5,7	1,9
6	6,4	6,6	0,2
7	5,1	5,1	0,0
8	9,6	10,9	1,3
		média	0,7625

...

Dif
0,8
-0,3
0,6
1,0
1,9
-0,2
0,0
1,3
0,6375

...

Amostra	Valor medido		Dif
	antes	depois	
1	4,3	3,5	-0,8
2	4,7	4,4	-0,3
3	5,9	5,3	-0,6
4	11,3	10,3	-1,0
5	5,7	3,8	-1,9
6	6,6	6,4	-0,2
7	5,1	5,1	-0,0
8	10,9	9,6	-1,3
		média	-0,7625

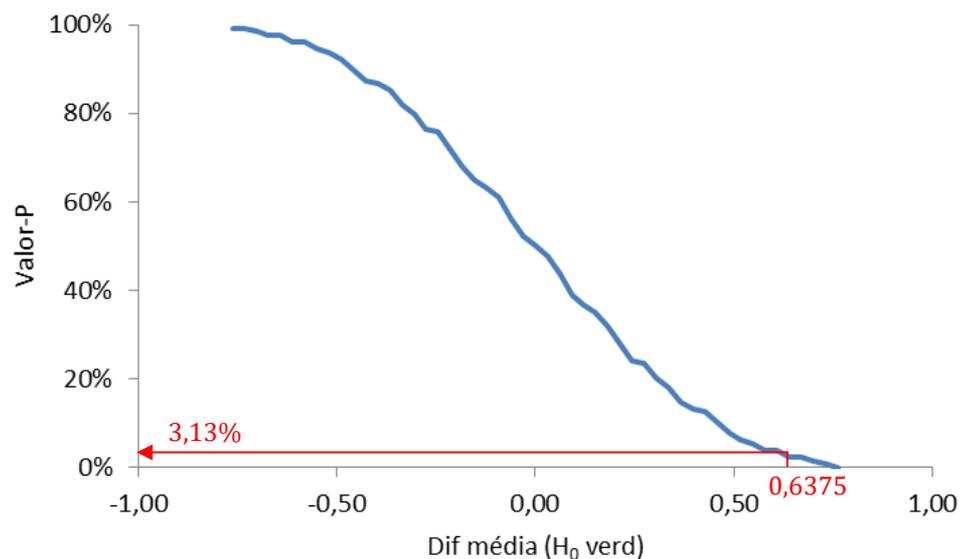
# Testes de Aleatorização - Exemplo

$H_0$ : não há diferença entre o valor medido antes e depois (Dif média = 0)

$H_1$ : o valor medido depois é maior (Dif média > 0)

Se  $H_0$  é verdadeira, então haverá  $2^8$  possibilidades de trocas, gerando 256 resultados diferentes

Amostra	Valor medido		Dif
	antes	depois	
1	3,5	4,3	0,8
2	4,7	4,4	-0,3
3	5,3	5,9	0,6
4	10,3	11,3	1,0
5	3,8	5,7	1,9
6	6,6	6,4	-0,2
7	5,1	5,1	0,0
8	9,6	10,9	1,3
		média	0,6375



Valor-P =  $P(\text{Dif média } H_0 \text{ verdadeiro} \geq \text{Dif média observada}) = 3,13\%$

**Conclusão:** rejeita-se  $H_0$  a 5% de significância, ou seja, há evidências de que o valor, em média, aumenta na última medição

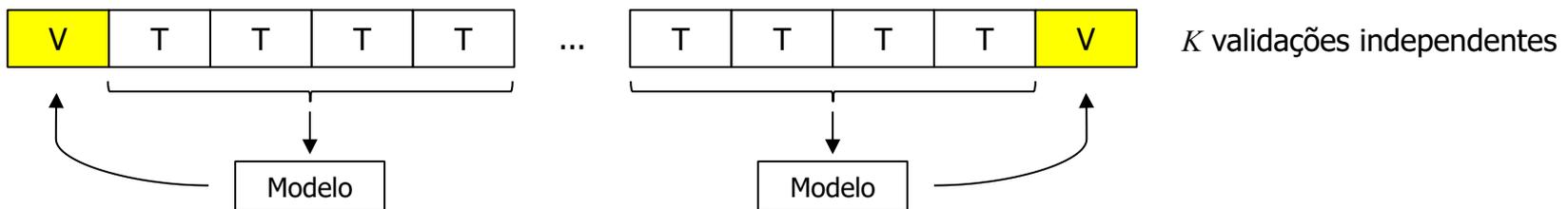
# Validação Cruzada

Como dito anteriormente, é sempre melhor usar um conjunto de amostras independentes para avaliar/validar os resultados de uma modelagem ou classificação

Tipicamente, na validação cruzada, a amostra é particionada aleatoriamente em  $K$  subconjuntos disjuntos (*K-Folds*)

O valor de  $K$  é em geral 5 ou 10. Quando  $K$  é igual ao número de amostras, então esse método é conhecido como Validação Cruzada LOO (*Leave One Out Cross Validation*)

O método *K-Folds* consiste em usar cada uma das partições como amostras de validação e as demais como treinamento. Assim consegue-se realizar  $K$  validações independentes



Em geral, os resultados das validações são sintetizados em uma medida de tendência central (média, mediana, etc) ou então através de um intervalo de credibilidade ou um boxplot

# Boosting

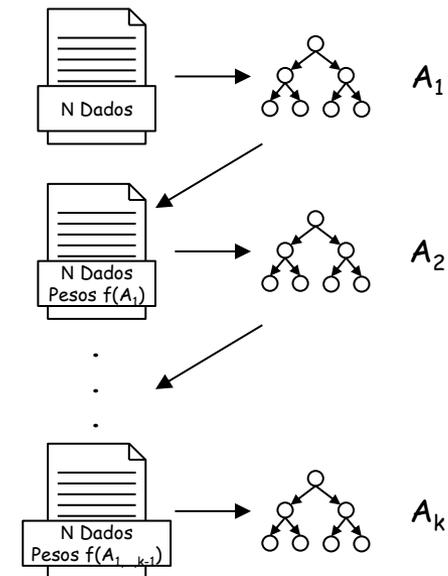
Não se trata de técnica tradicional de reamostragem como *Bootstrap* e *Jackknife*

Consiste em alterar a importância de cada amostra, através de pesos, de modo a favorecer algumas amostras em detrimento de outras

O *AdaBoost*, por exemplo, ajusta o peso de cada amostra, aumentando a importância de amostras mal classificadas/estimadas de modo a diminuir os erros de classificação/estimação

Num classificador baseado em árvores de decisão, por exemplo, a primeira árvore é construída com base em todas as amostras. Essa árvore é usada para avaliar cada amostra, aumentando-se o peso daquelas mal classificadas. Uma nova árvore é gerada. Esse processo é repetido um determinado número de vezes

O objetivo principal desse método é reduzir a tendência (ou viés), tentando-se minimizar os erros



# Balanceamento de Classes/Estratos

---

Muitas técnicas estatísticas são fortemente influenciadas pelo desbalanceamento das classes ou estratos usados no treinamento. Em geral, classes minoritárias são preteridas ou negligenciadas durante o processo de treinamento de modelos supervisionados

Algumas técnicas podem ser aplicadas para mitigar esse tipo de problema:

- *Oversampling* (Superamostragem)

Aumenta-se o número de amostras da classe minoritária: superamostragem aleatória (duplicação de amostras), SMOTE, ADASYN

- *Undersampling* (Subamostragem)

Reduz-se o número de amostras da classe majoritária: subamostragem aleatória (eliminação de amostras), remoção de amostras redundantes ou pouco informativas (usando critérios de proximidade)

# Perturbação aleatória

---

Em alguns casos (amostragens pequenas, modelos determinísticos, por exemplo), a introdução de perturbações nos dados de entrada podem ajudar o processo de modelagem

- Estimação de distribuições de resultados em modelos determinísticos

O método de Monte Carlo pode ser empregado para estimar a distribuição de resultados

- Simulação de dados reais

A perturbação permite avaliar a incerteza “natural” esperada em dados reais

- Robustez do modelo

A adição de ruídos aos dados ou aos parâmetros do modelo permitem avaliar sua estabilidade e sensibilidade

- Otimização

Em algoritmos de otimização, como algoritmos genéticos, a perturbação aleatória pode favorecer a busca por soluções globais por evitar que o mesmo fique “preso” a soluções locais

# Aprendizado Ativo – *Active Learning*

---

É uma técnica na qual o algoritmo de aprendizado busca especificamente os dados que são mais informativos para o modelo em vez de ser treinado por todo o conjunto de dados disponível

Em vez de sortear um subconjunto de amostras aleatoriamente (por alguma técnica de reamostragem), o modelo seleciona as amostras que considera mais informativas ou atribui pesos às amostras, de acordo com algum critério:

- a) Maior incerteza: o algoritmo privilegia as amostras que tem maior dificuldade para classificar ou maior erro ao estimar seu valor
- b) Maior acerto: o algoritmo escolhe amostras que tenham maior impacto no desempenho final do modelo