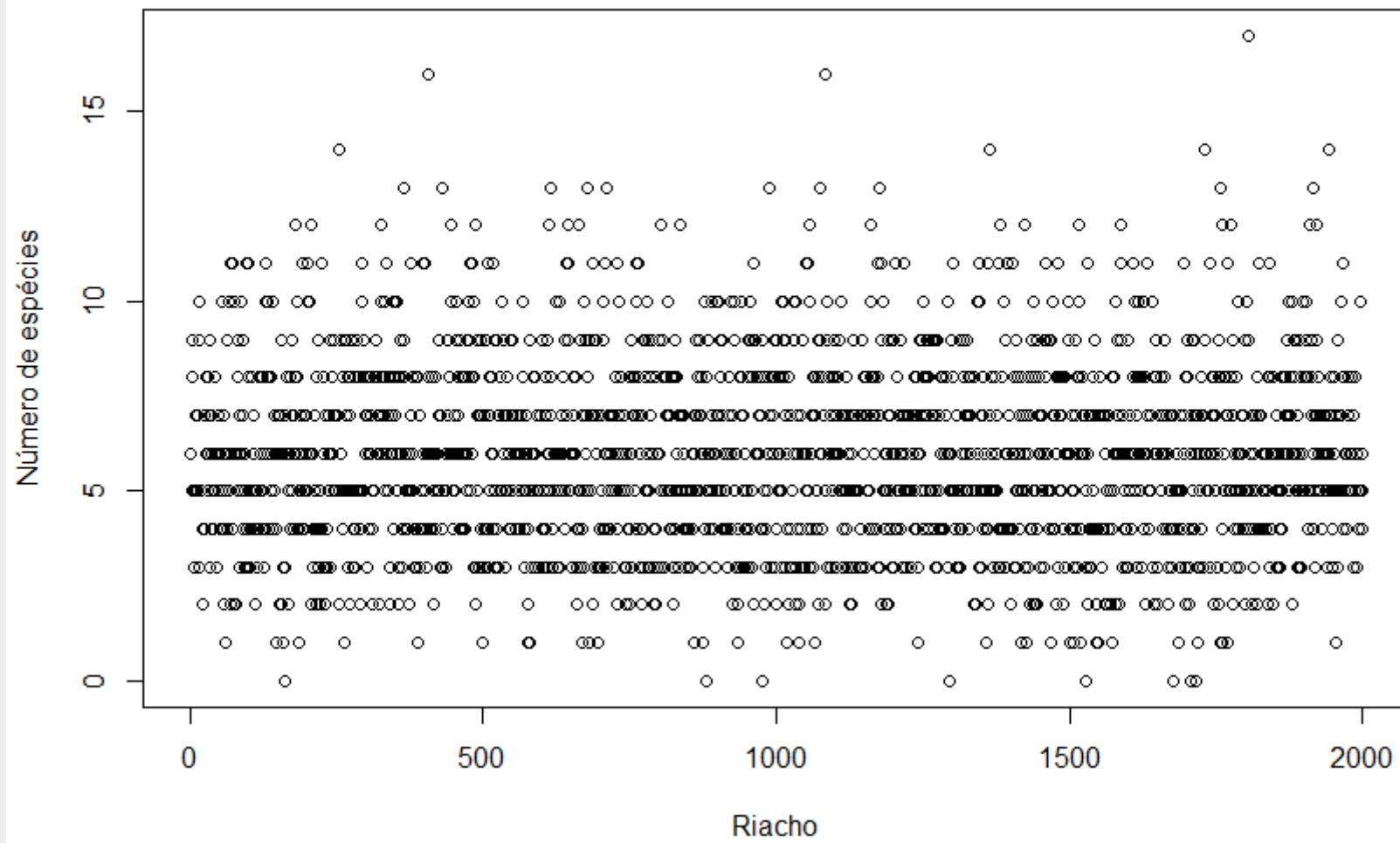


# Análise estatística por aleatorização, *bootstrap* e Monte Carlo

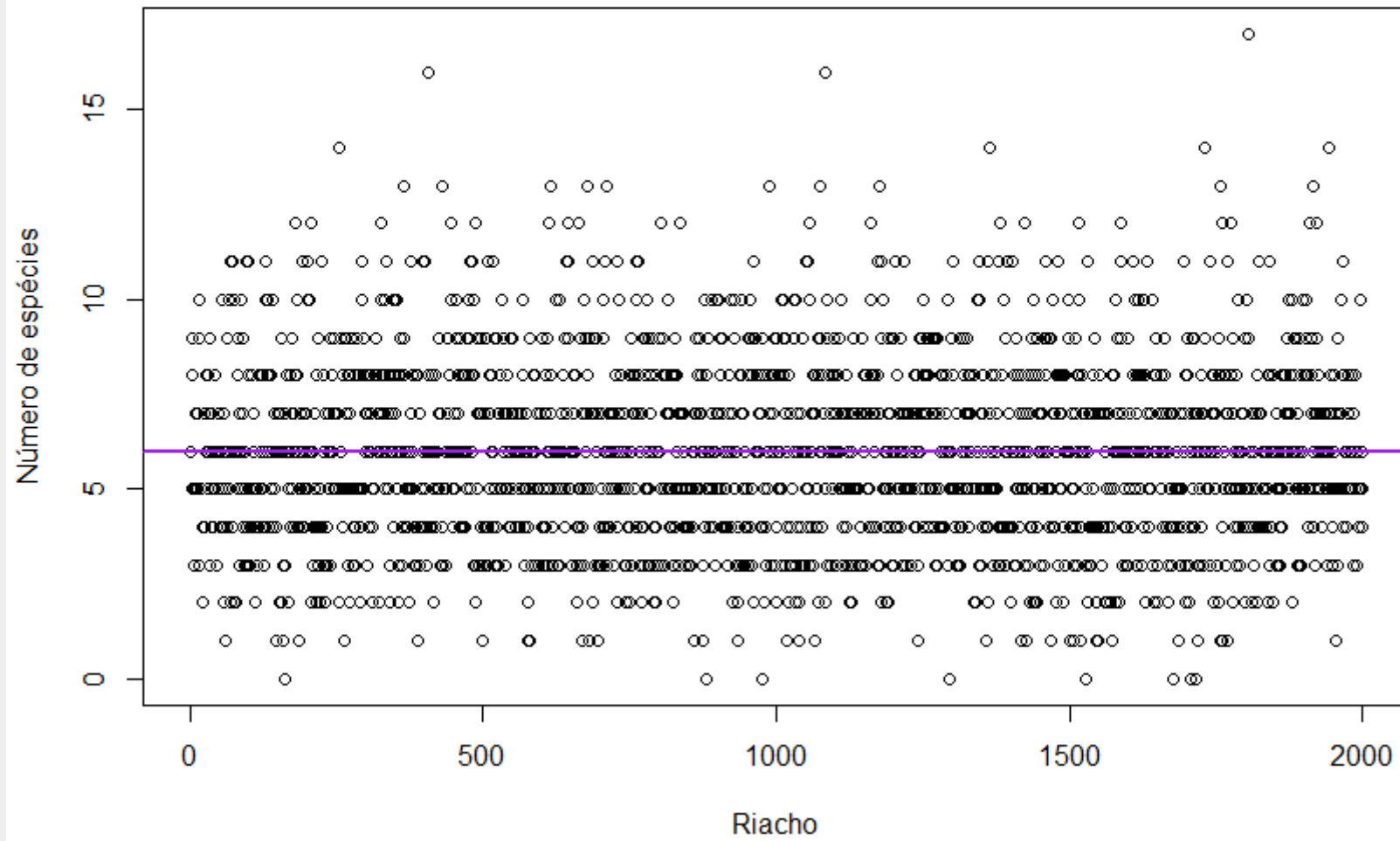
Pavel Dodonov  
pdodonov@gmail.com

Laboratório de Ecologia Aplicada à Conservação (LEAC)  
Universidade Estadual de Santa Cruz (UESC)  
Ilhéus - BA

# População simulada



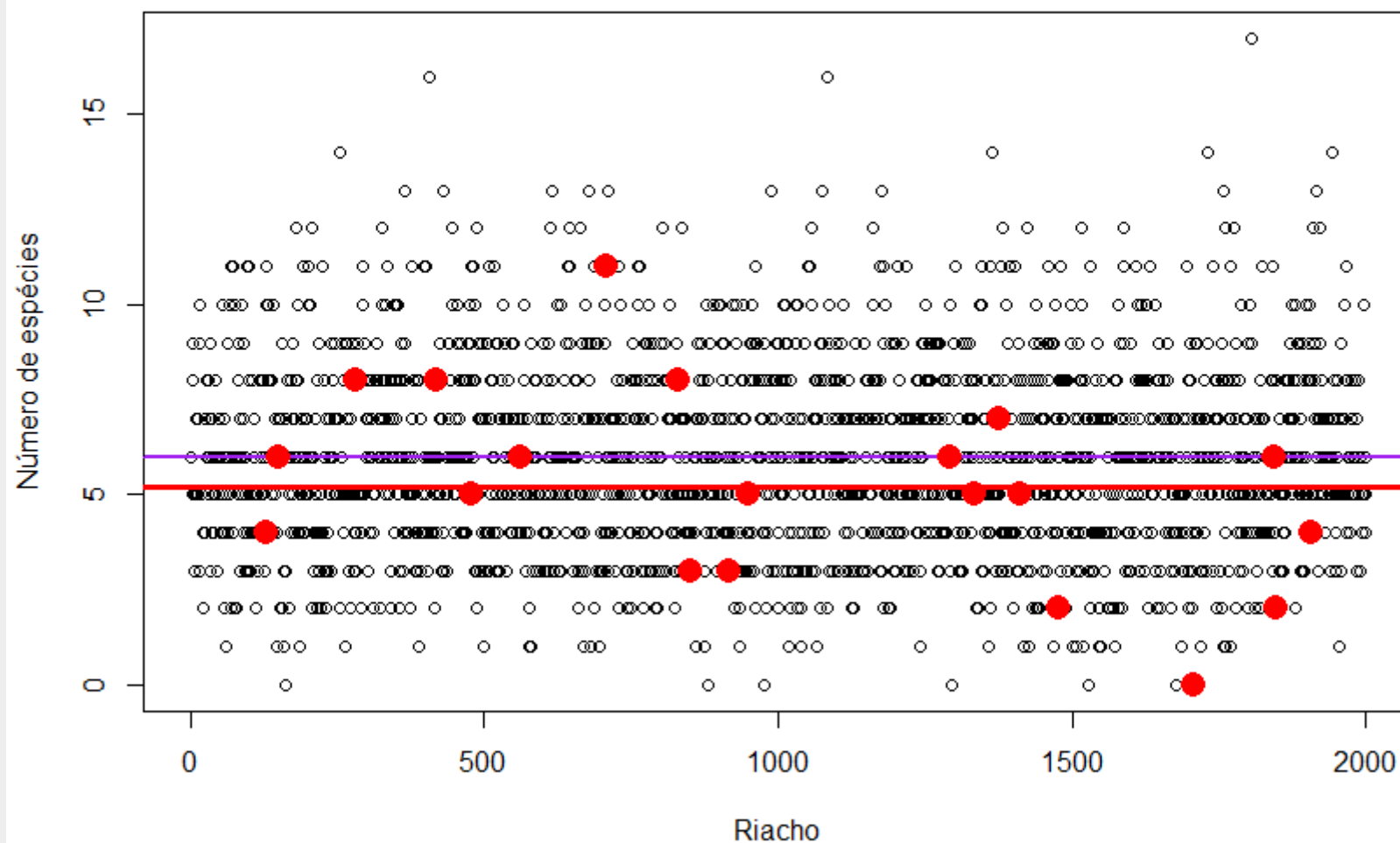
# População simulada



População simulada

Média = 5.97

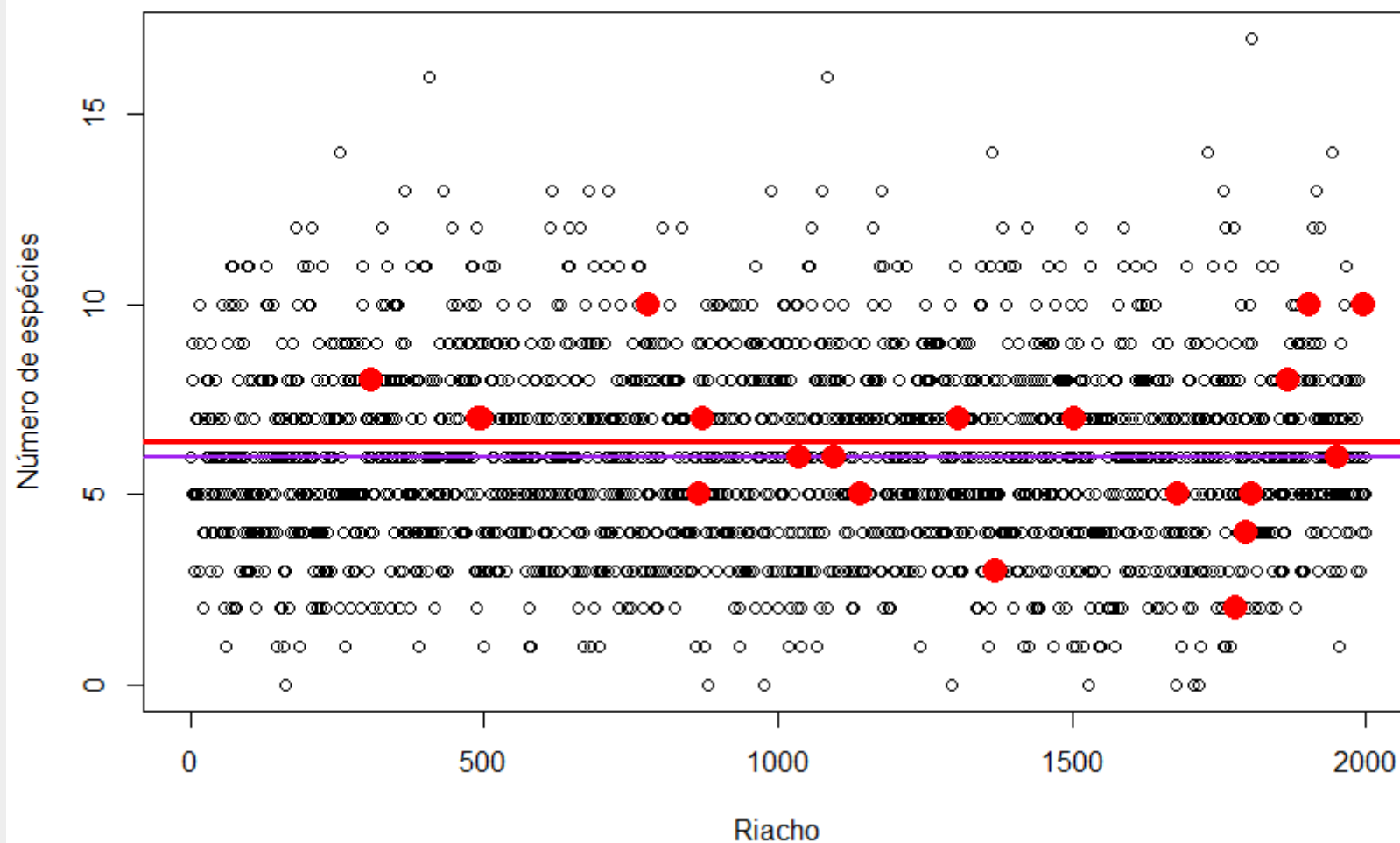
# Amostragem



Amostra 1

Média = 5.2

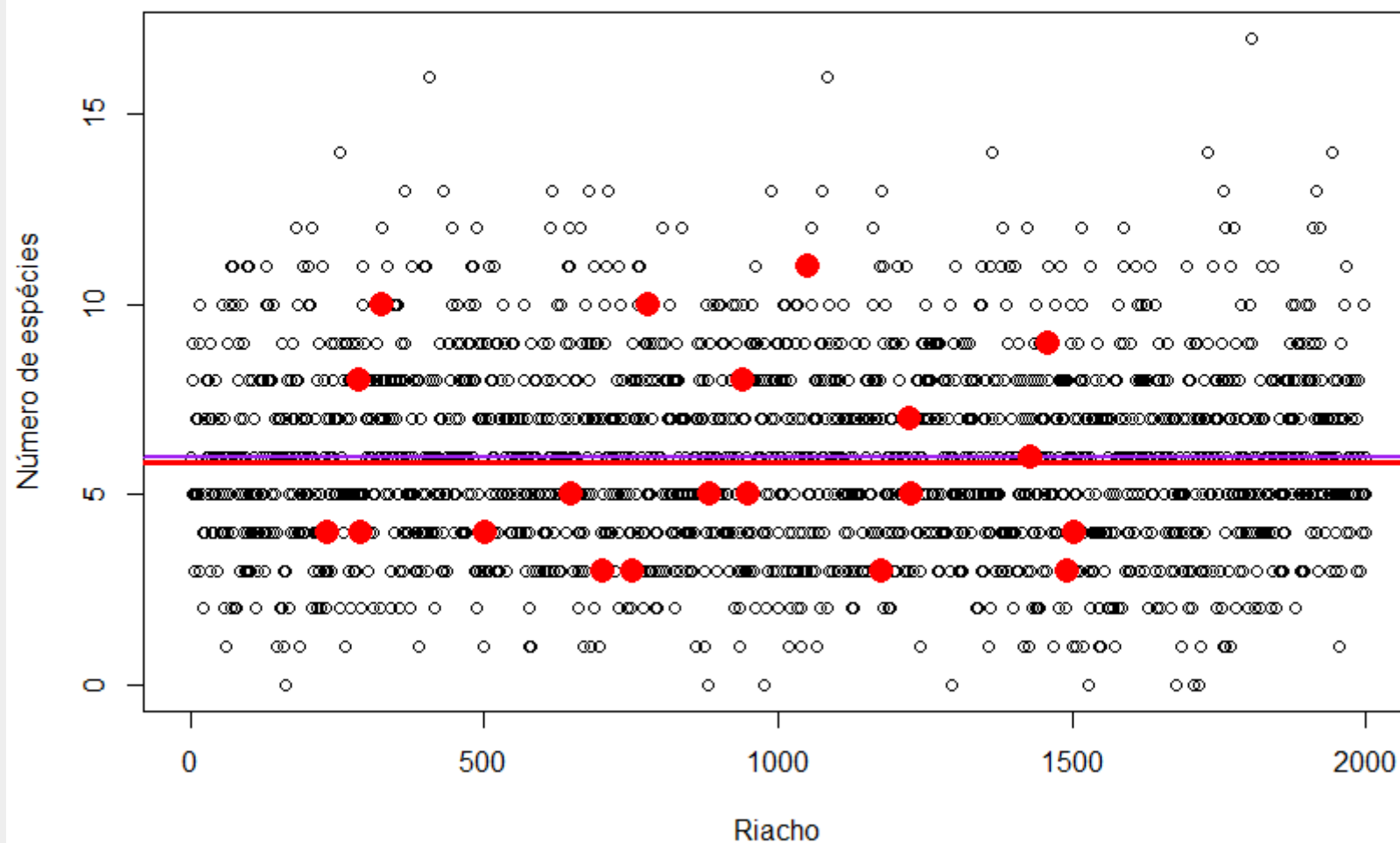
# Amostragem



Amostra 2

Média = 6.4

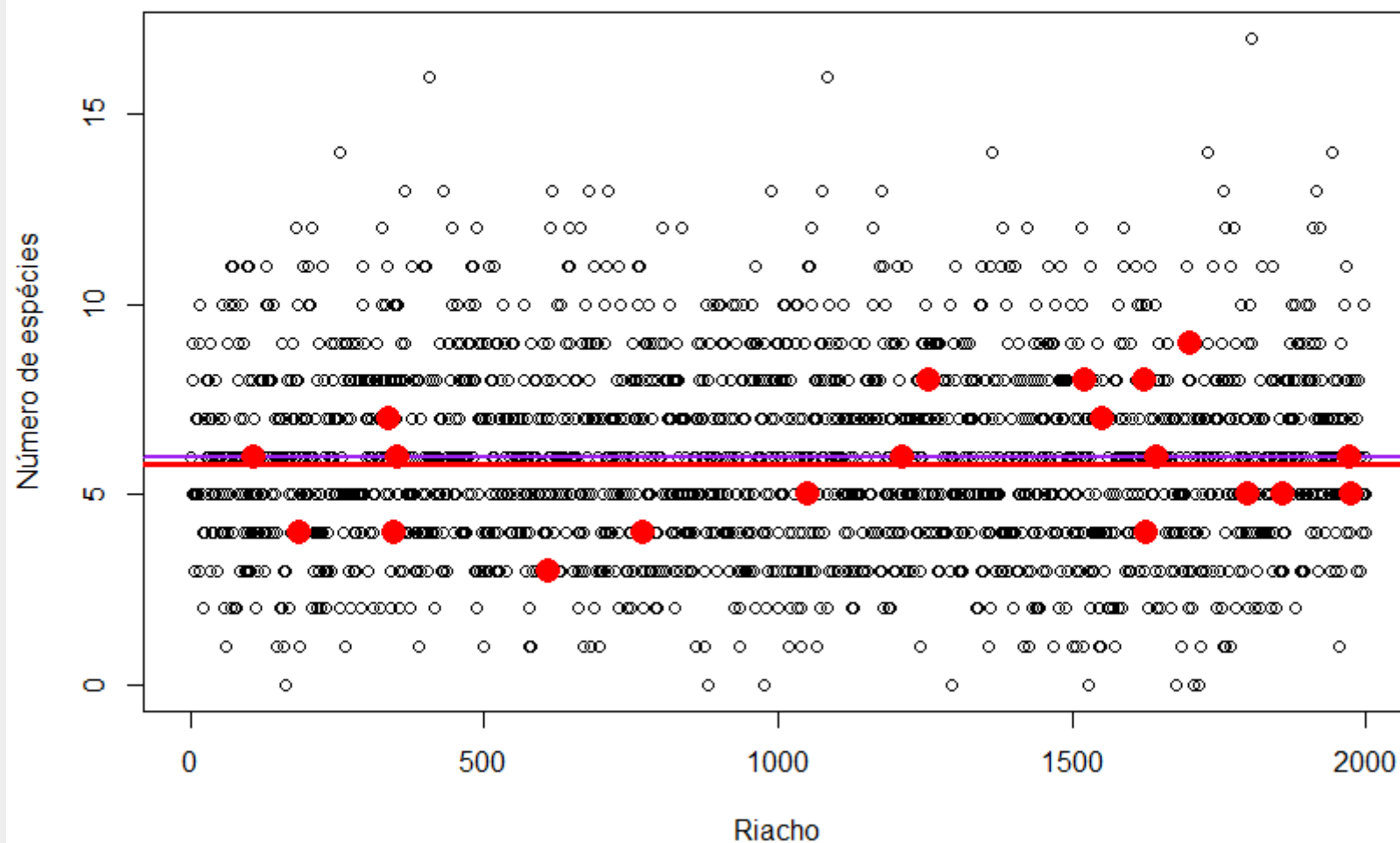
# Amostragem



Amostra 3

Média = 5.85

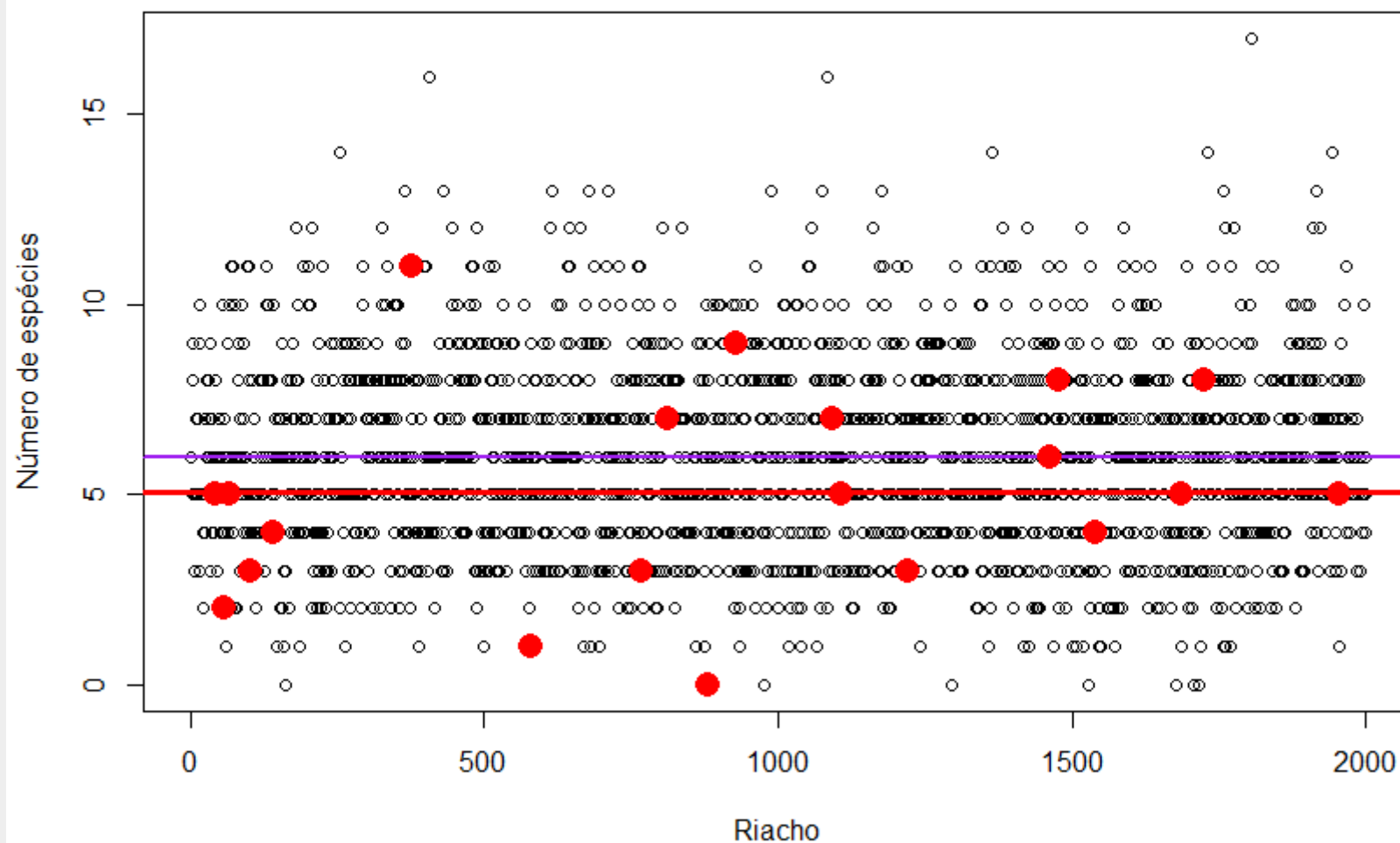
# Amostragem



Amostra 4

Média = 5.8

# Amostragem

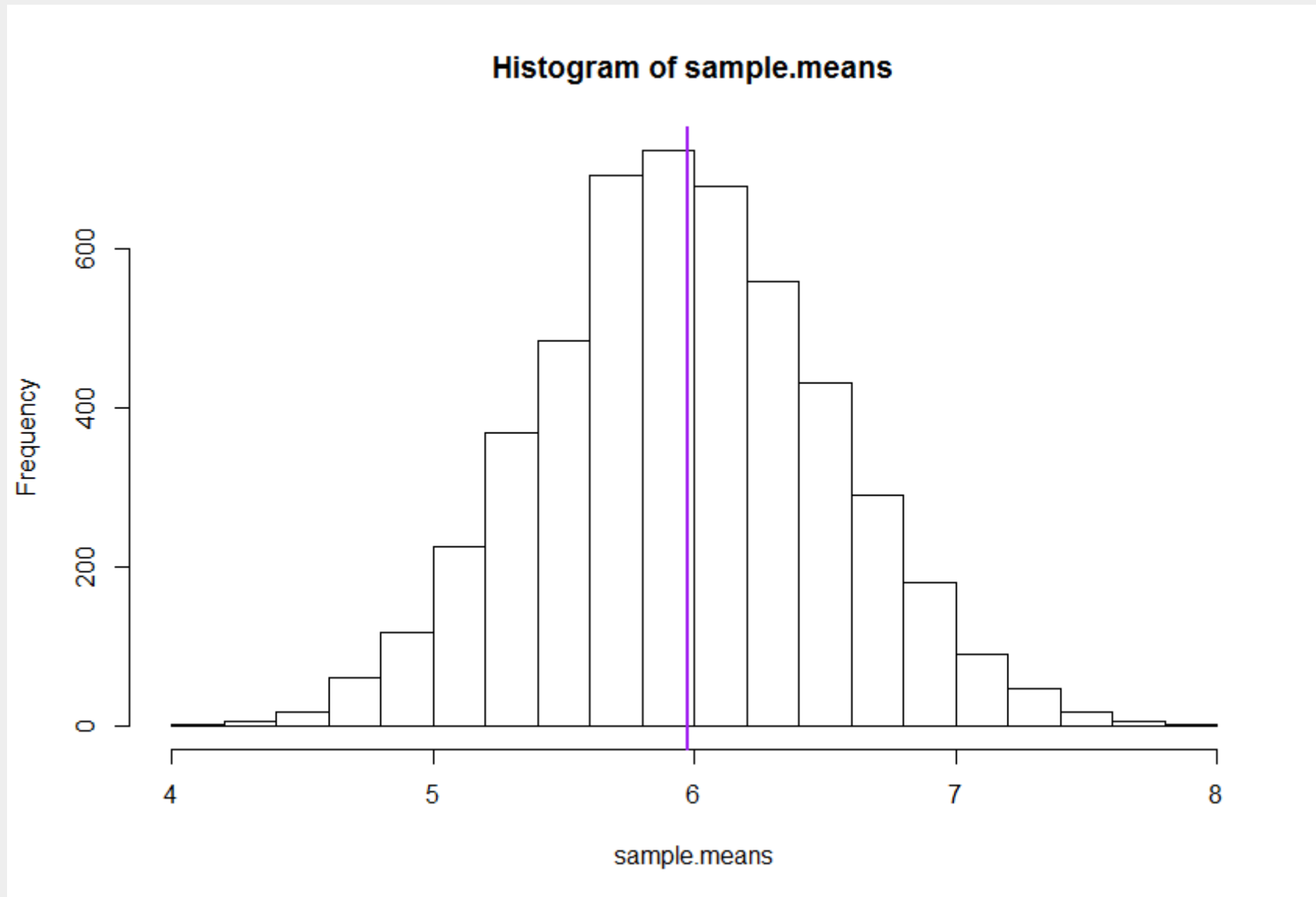


Amostra 5

Média = 5.05

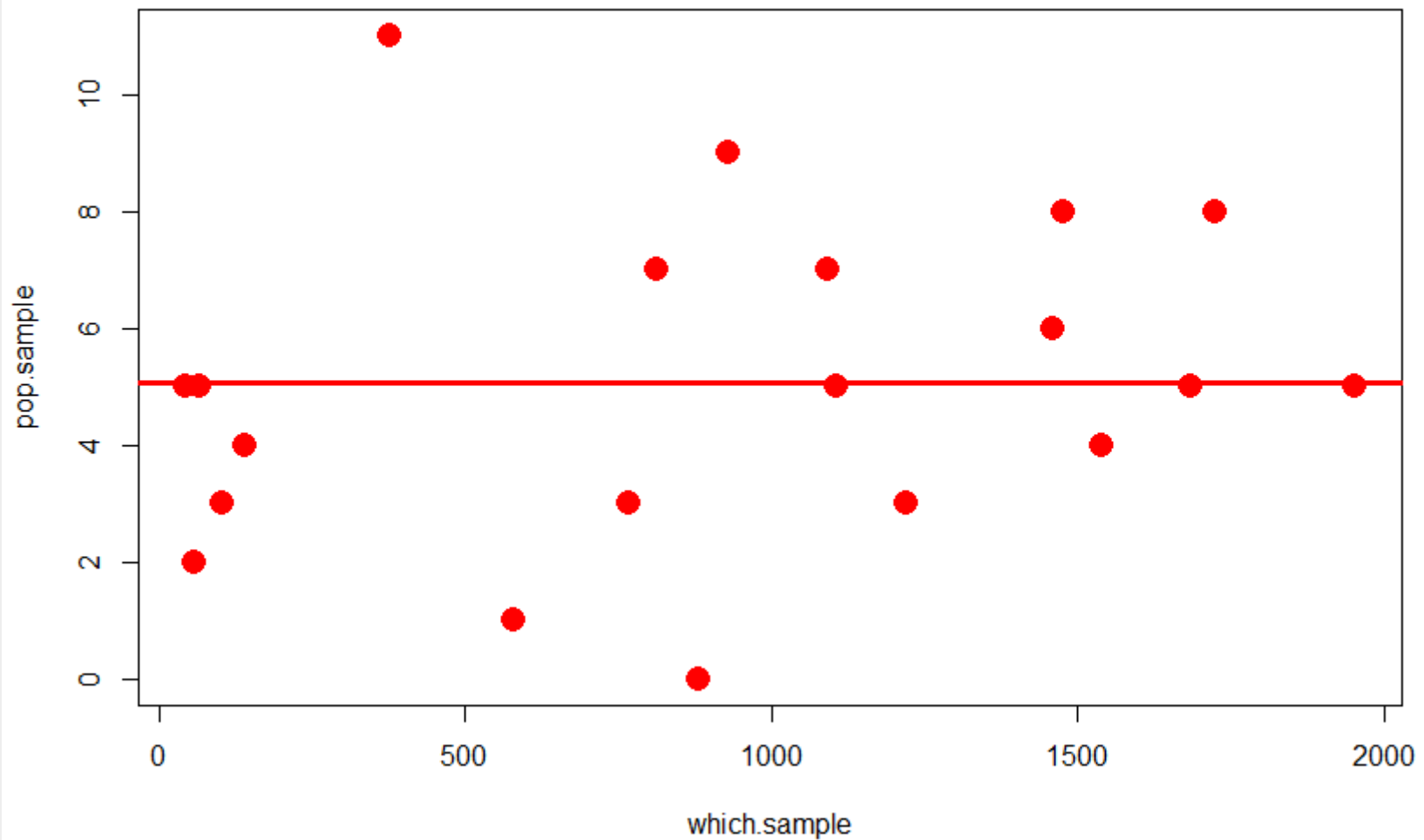


# Amostragem



# Amostra 5

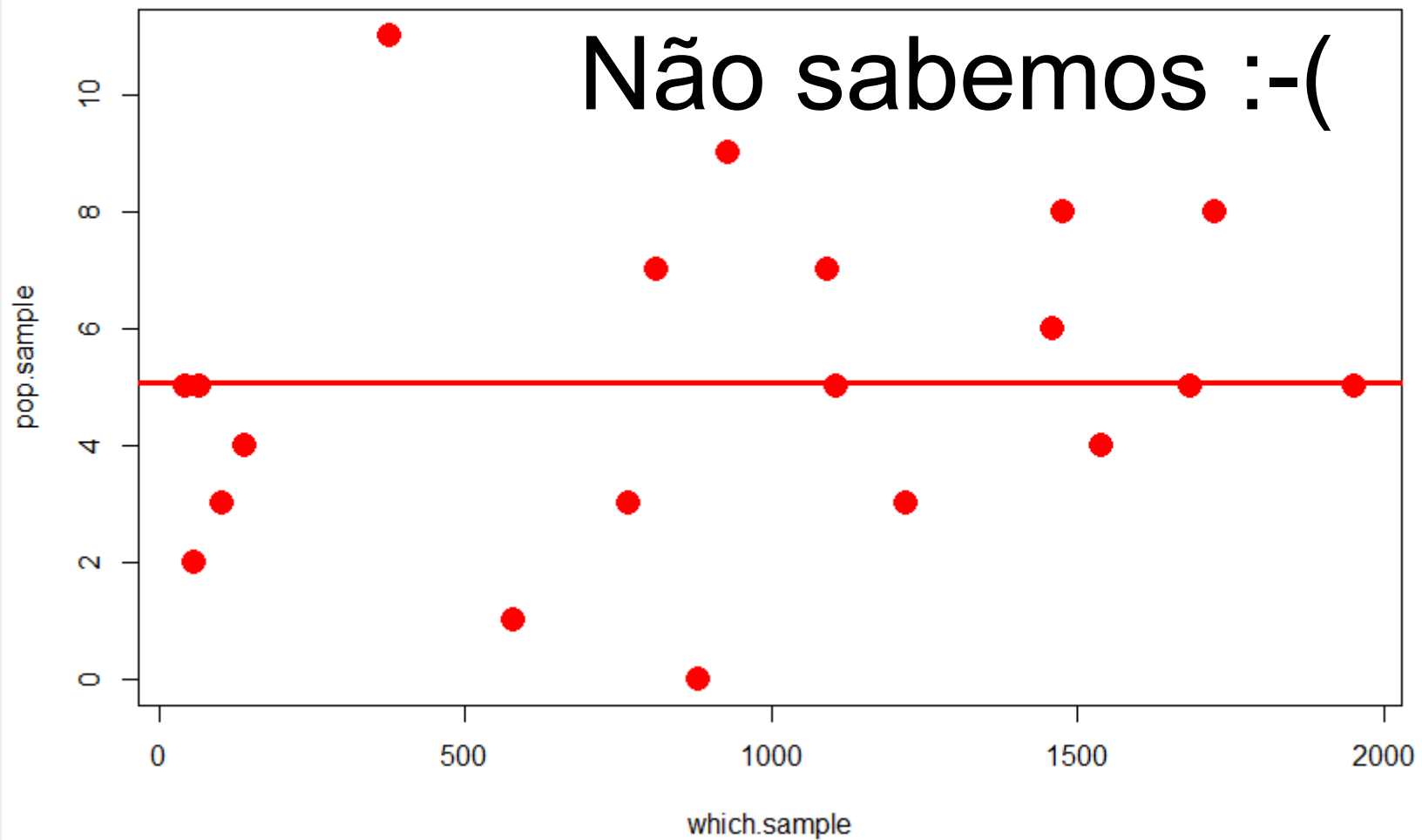
Média = 5.05



Onde estará a média da população?

# Amostra 5

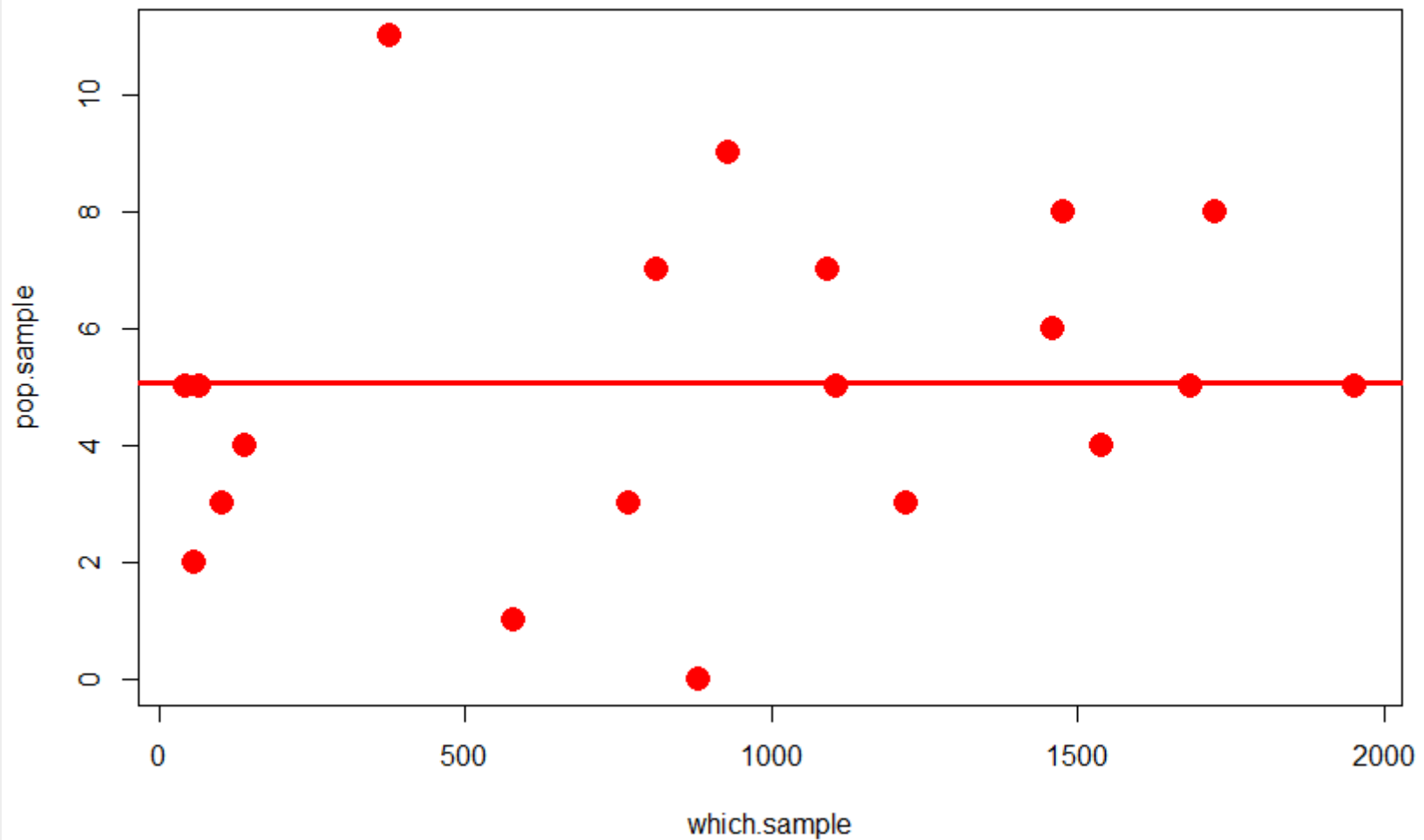
Média = 5.05



Onde estará a média da população?

# Amostra 5

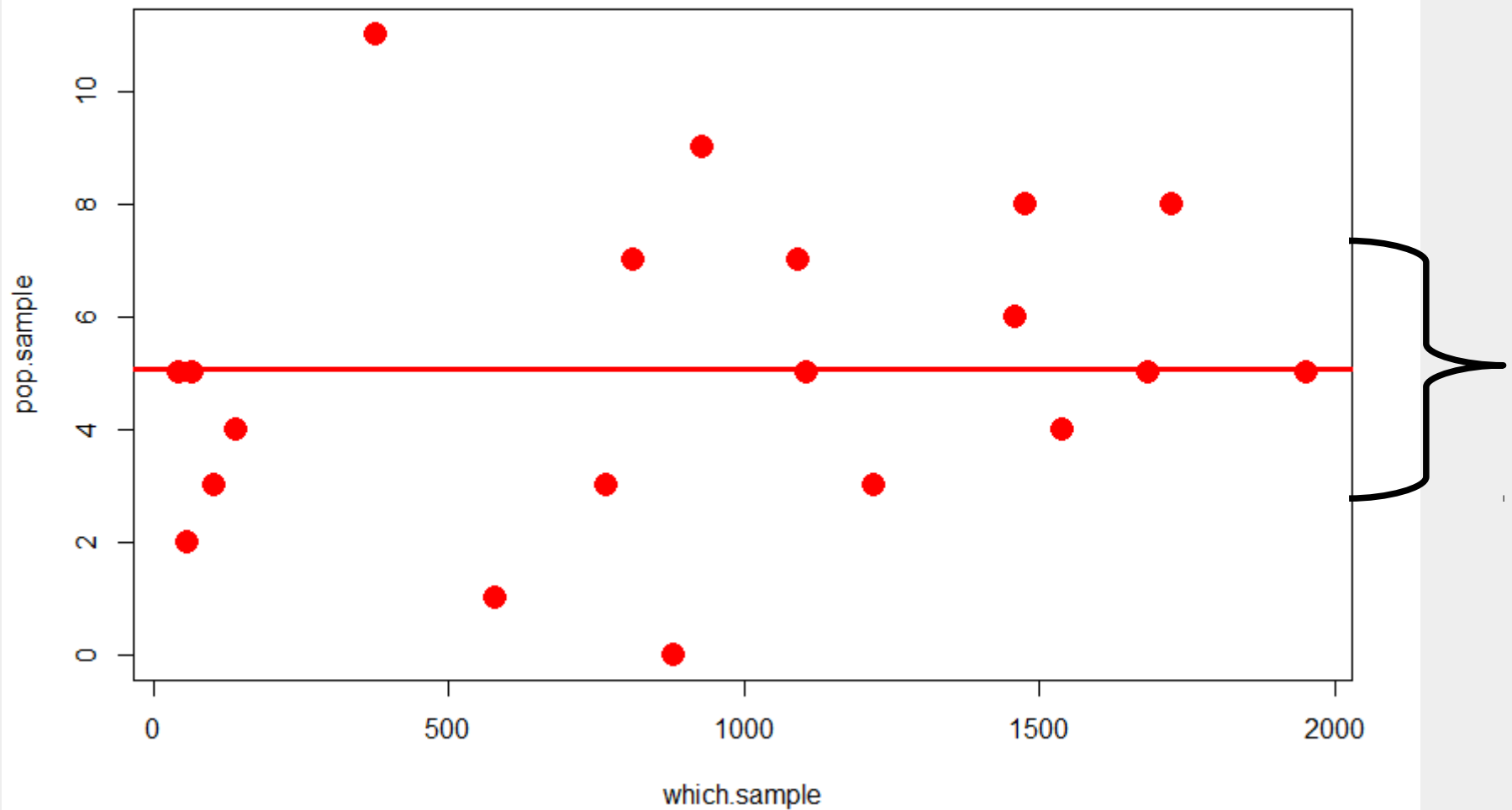
Média = 5.05



Onde poderia estar a média da população?

# Amostra 5

Média = 5.05



Onde *poderia* estar a média da população?

# Intervalo de confiança

**A probabilidade de que a  
média real está dentro  
deste intervalo = 95%**  
(Gotelli & Ellison 2004)

# Intervalo de confiança

~~A probabilidade de que a  
média real está dentro  
deste intervalo = 95%  
(Gotelli & Ellison 2004)~~

Ou está ou não está; o  
experimento já foi feito.  
(Vários autores...)

# Intervalo de confiança

Se o procedimento fosse **repetido em múltiplas amostras**, o **IC** calculado **englobaria a média** da população **95%** das vezes  
(Cox & Hinkley 1974)



# Intervalo de confiança

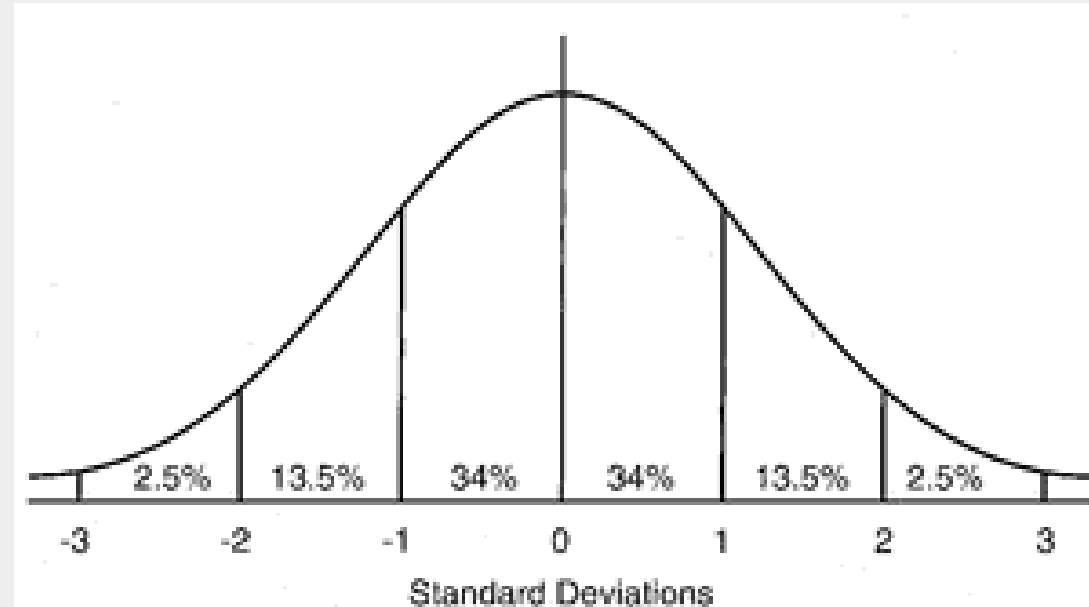
Calculado a partir da  
**variação nas médias**

**Erro-padrão da média:**  
**variação nas médias**  
com **repetições do**  
**experimento**

# Intervalo de confiança

- Pode ser calculado analiticamente

$$CI = \bar{X} \pm Z_{\alpha(2)} \sigma_{\bar{X}}$$
$$Z_{\alpha(2)} = 1.96$$

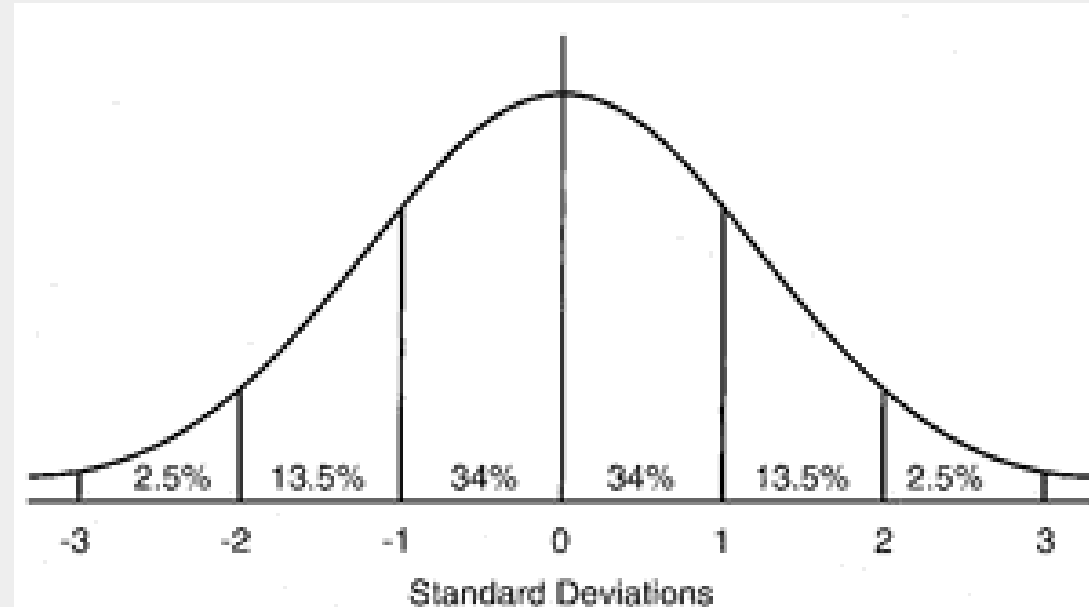


# Intervalo de confiança

- Pode ser calculado analiticamente

$$CI = \bar{X} \pm Z_{\alpha(2)} \sigma_{\bar{X}}$$

$$Z_{\alpha(2)} = 1.96$$



$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$$

Assumindo normalidade dos dados

# Bootstrap



# Bootstrap

**“Na ausência de qualquer outro conhecimento sobre uma população, a distribuição de valores encontrados em uma amostra aleatória de tamanho  $n$  da população é o melhor guia para a distribuição na população”**

(Manly 2007)

# Bootstrap

“Para **aproximar** o que aconteceria se a **população** fosse **reamostrada**, faz sentido **reamostrar a amostra.**”

(Manly 2007)

# Bootstrap

“Para **aproximar** o que aconteceria se a **população** fosse **reamostrada**, faz sentido **reamostrar a amostra.**”

(Manly 2007)

“A população infinita que consiste dos **n valores amostrados observados**, cada um com **probabilidade  $1/n$** , é usada para **modelar a população real desconhecida.**”

# Bootstrap

- Reamostragem com reposição

Dados originais



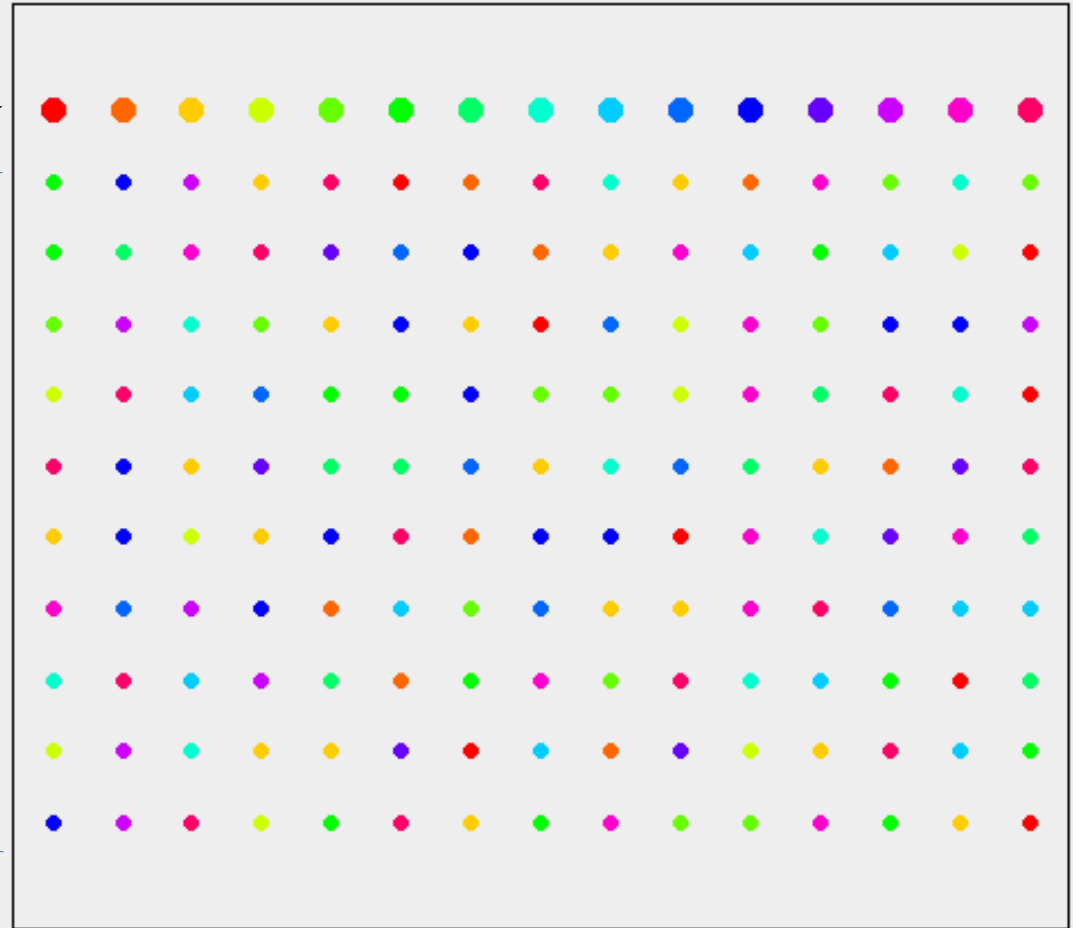


# Bootstrap

- Reamostragem com reposição

Dados originais

10 pseudoamostras aleatórias

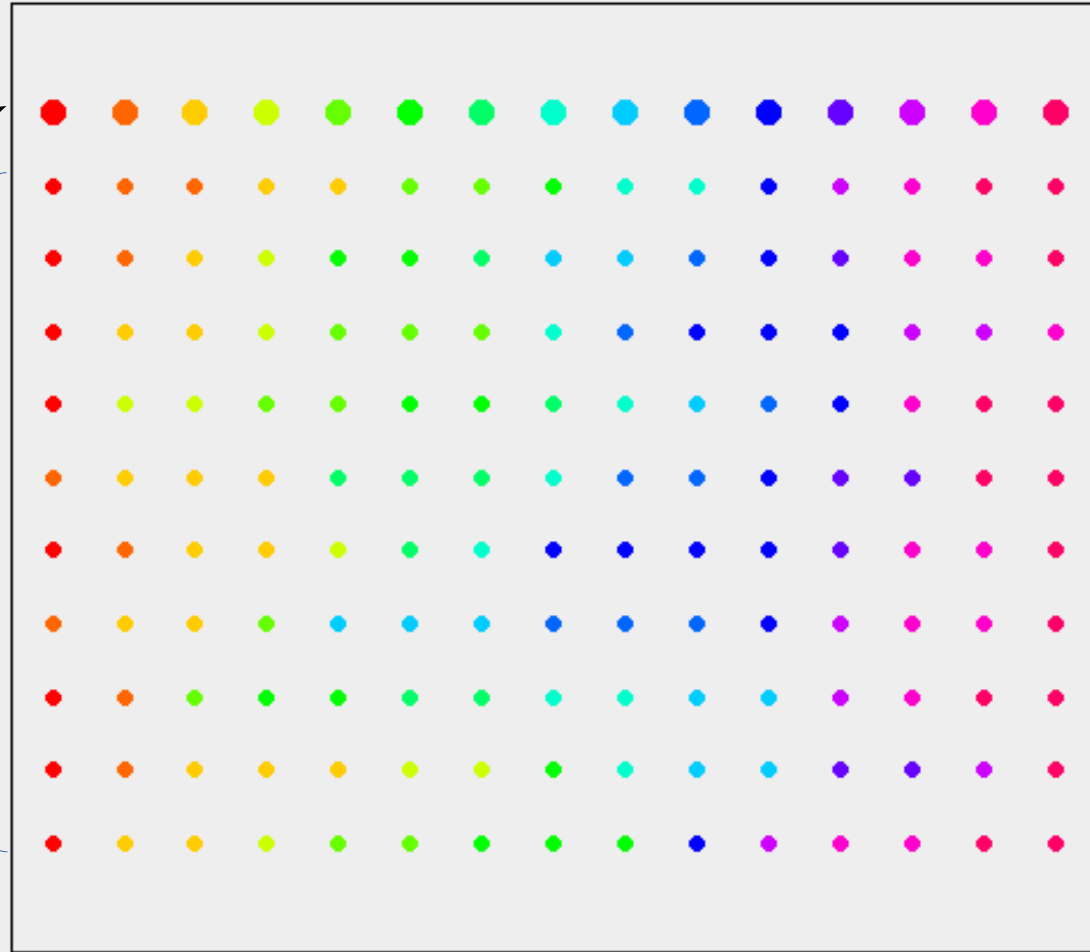


# Bootstrap

- Reamostragem com reposição

Dados originais

10 pseudoamostras aleatórias

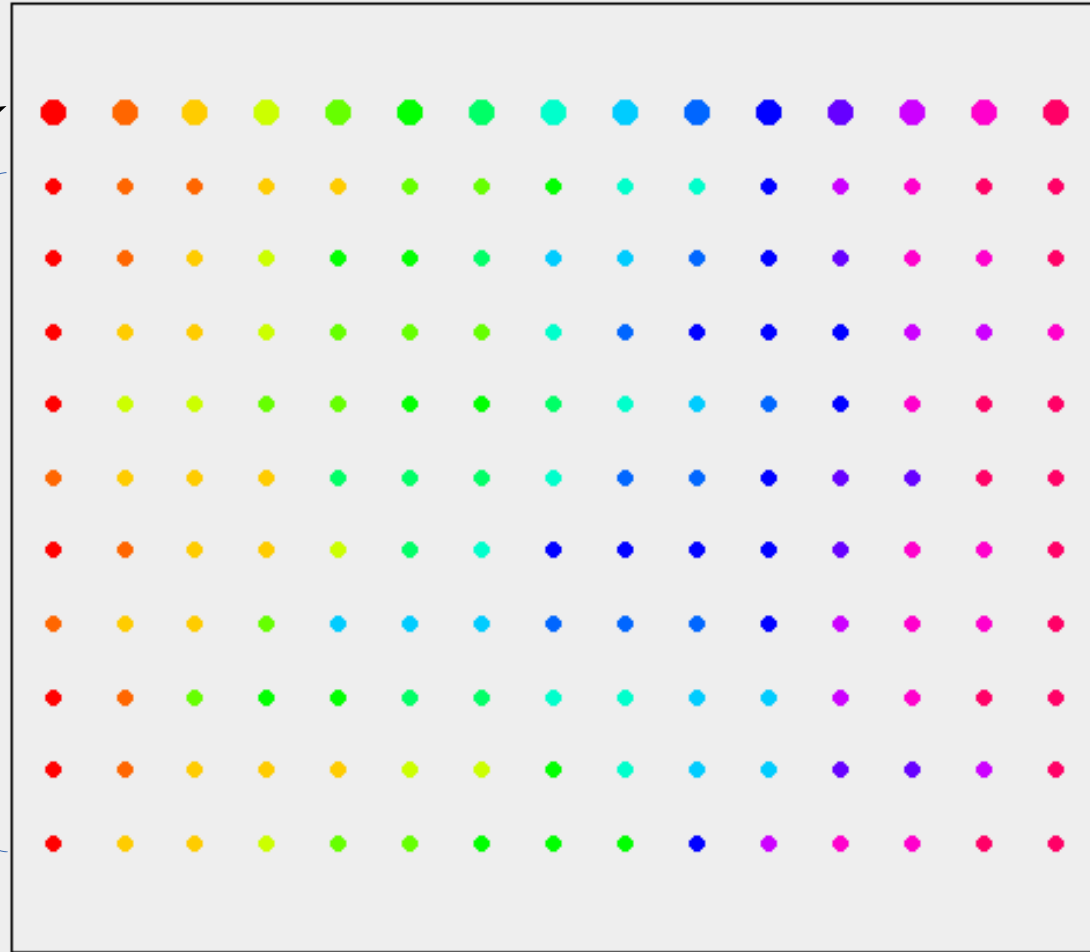


# Bootstrap

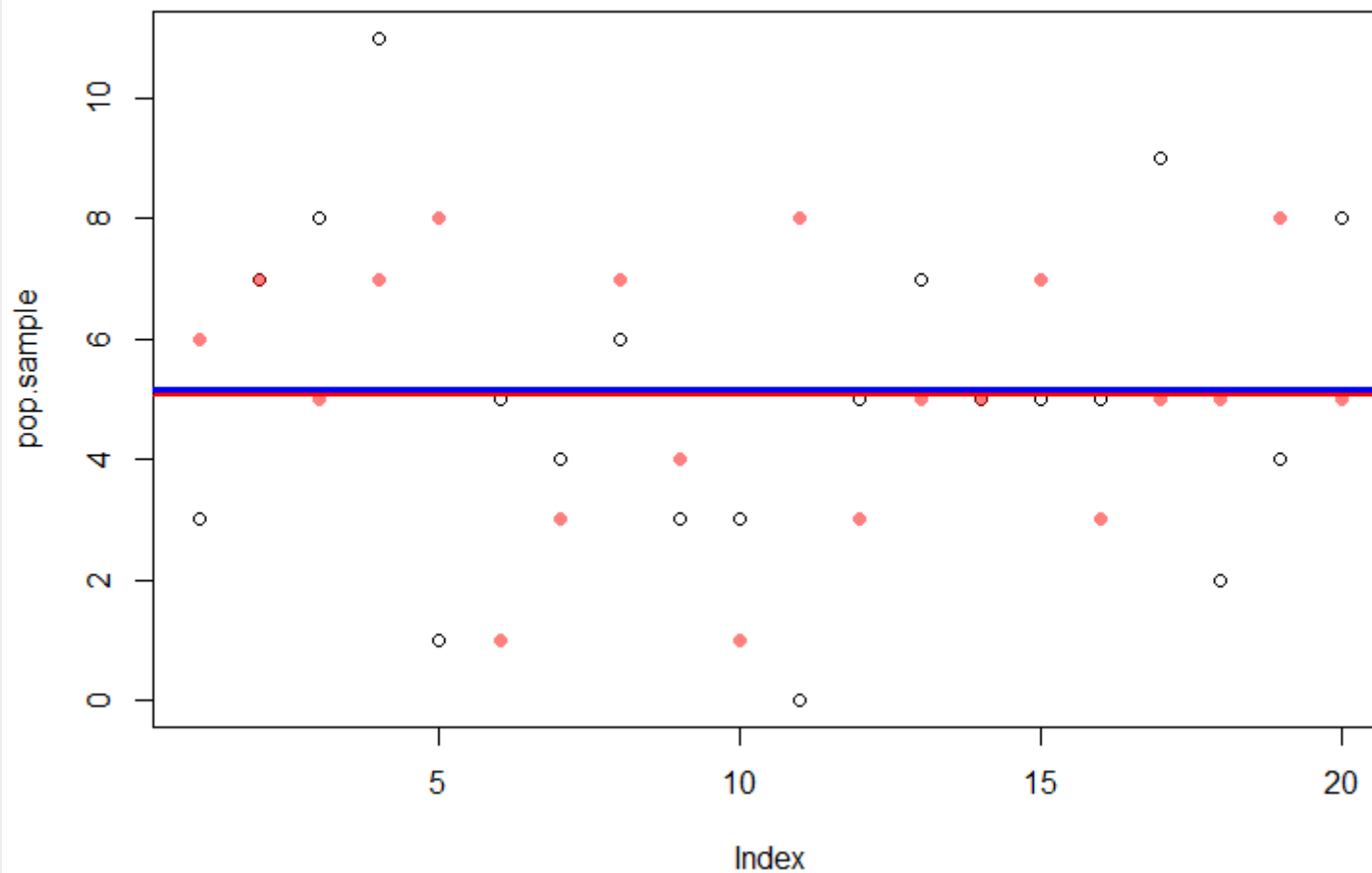
- Reamostragem com reposição

Dados originais  
Média = 8

10 pseudoamostras aleatórias  
Médias de 7.4, 8.2, 7.8,  
8.0, 8.3, 8.5, 9.1, 8.3, 6.9,  
8.1

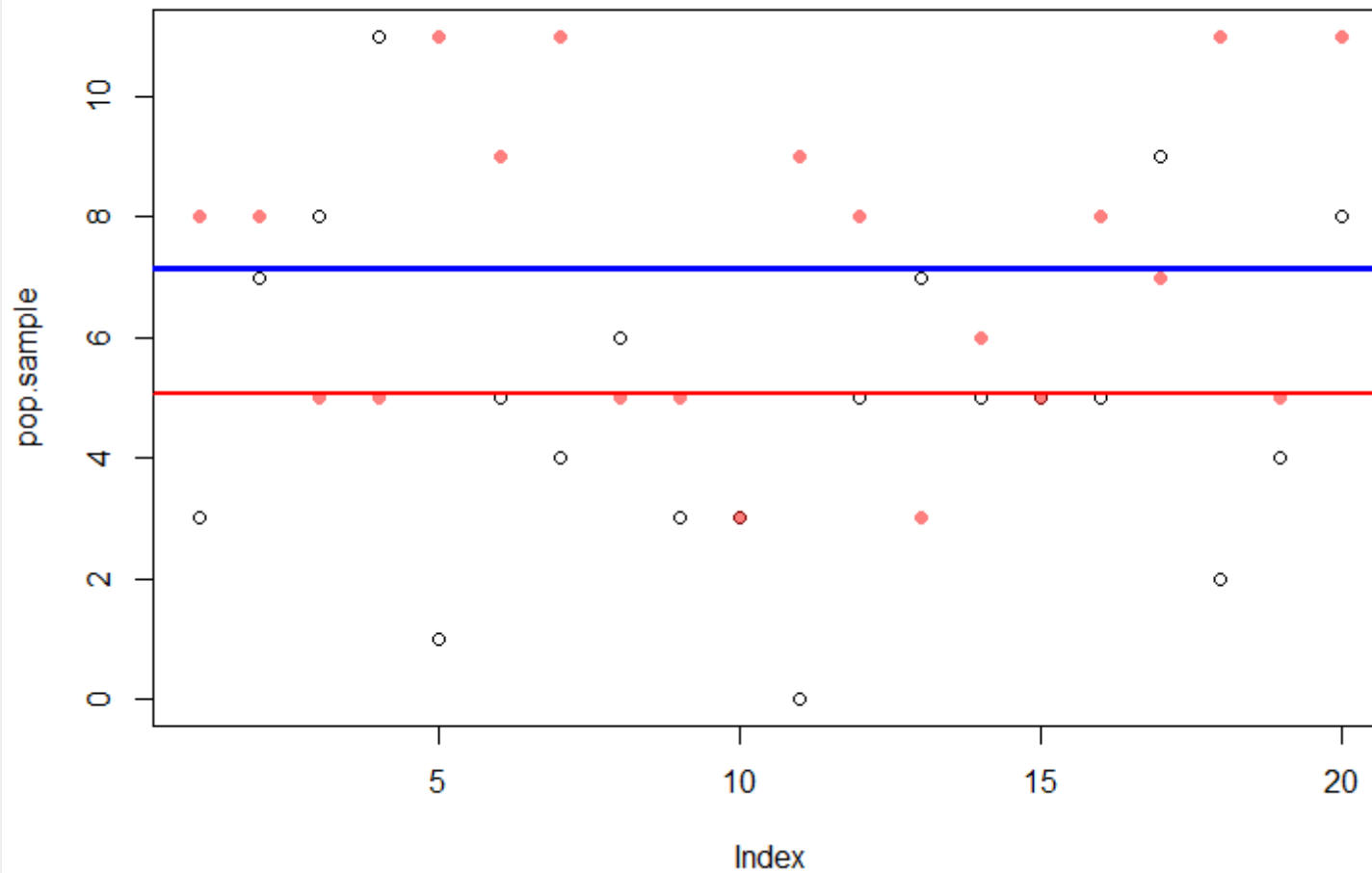


# Pseudoamostra 1



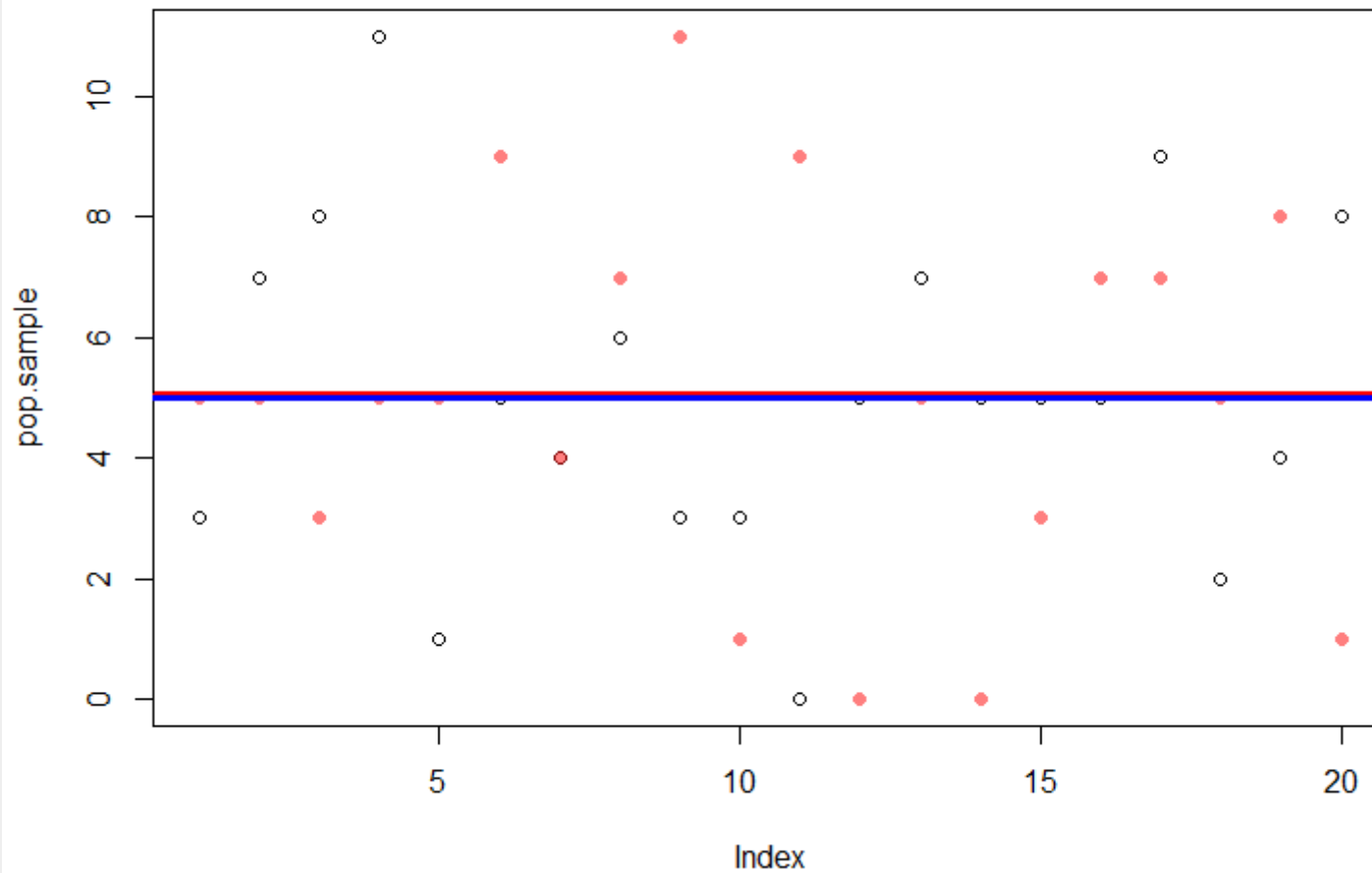
Média = 5.15

# Pseudoamostra 2



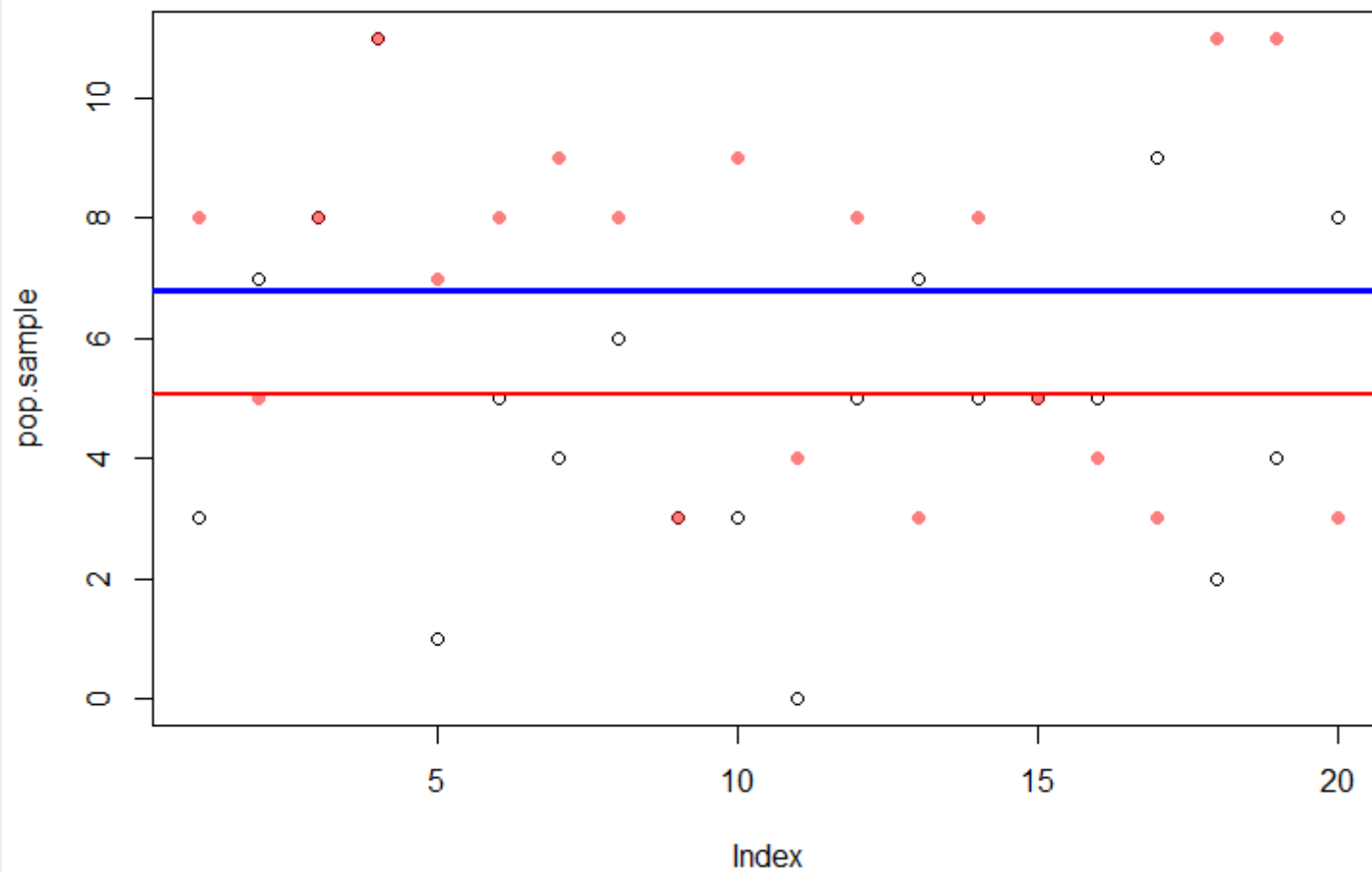
Média = 7.15

# Pseudoamostra 3



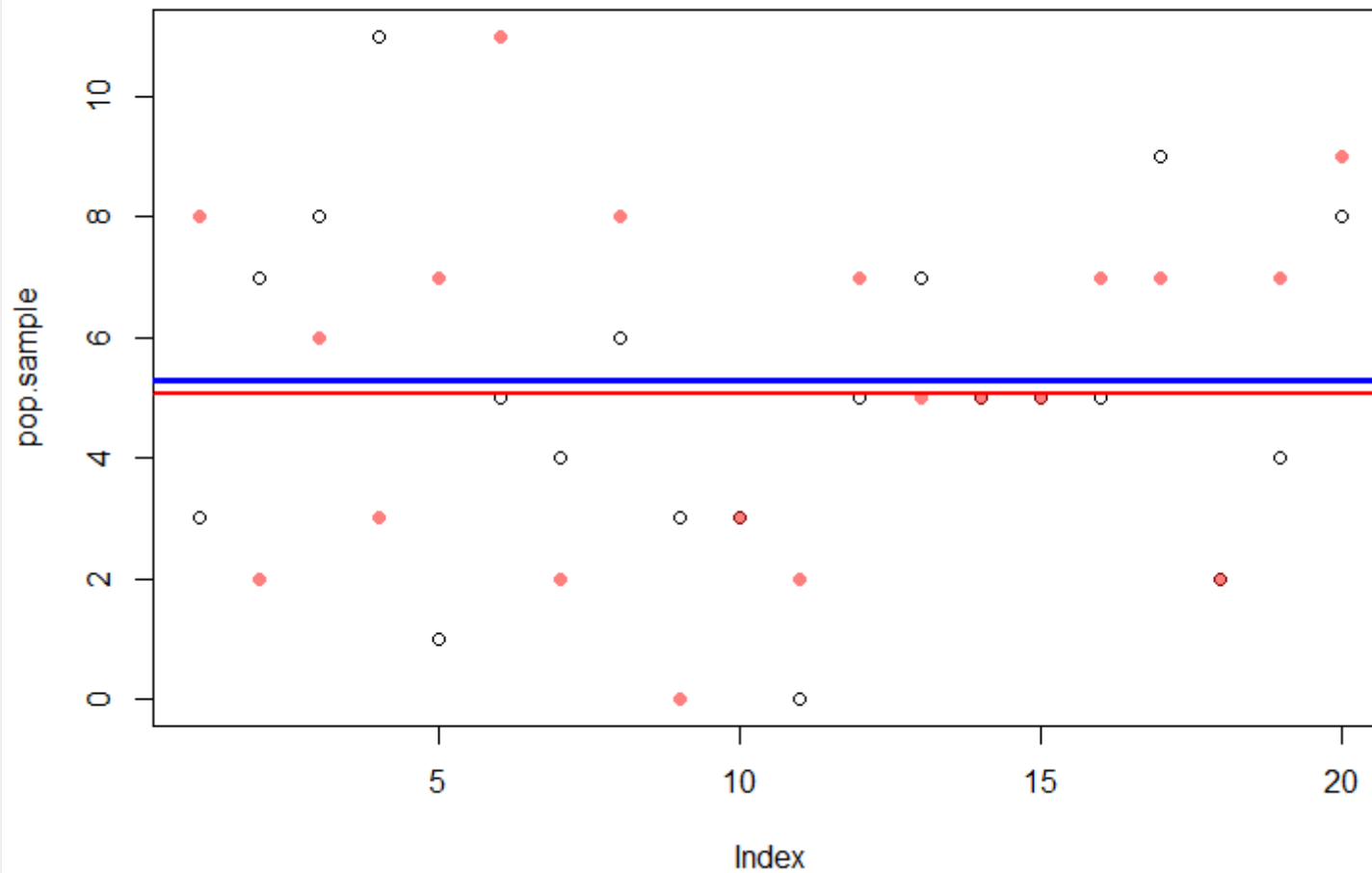
Média = 5.0

# Pseudoamostra 4



Média = 6.8

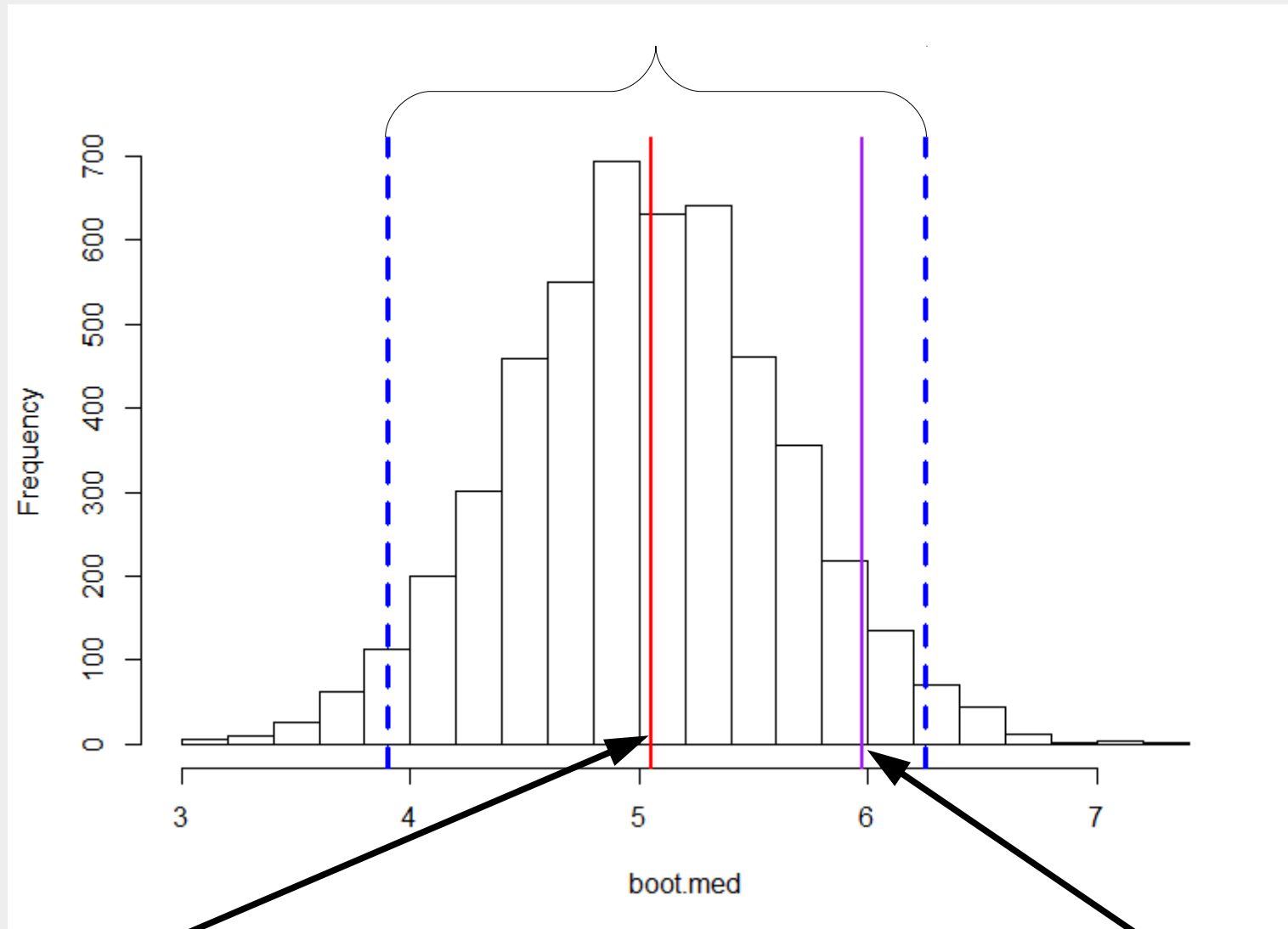
# Pseudoamostra 5



Média = 5.3



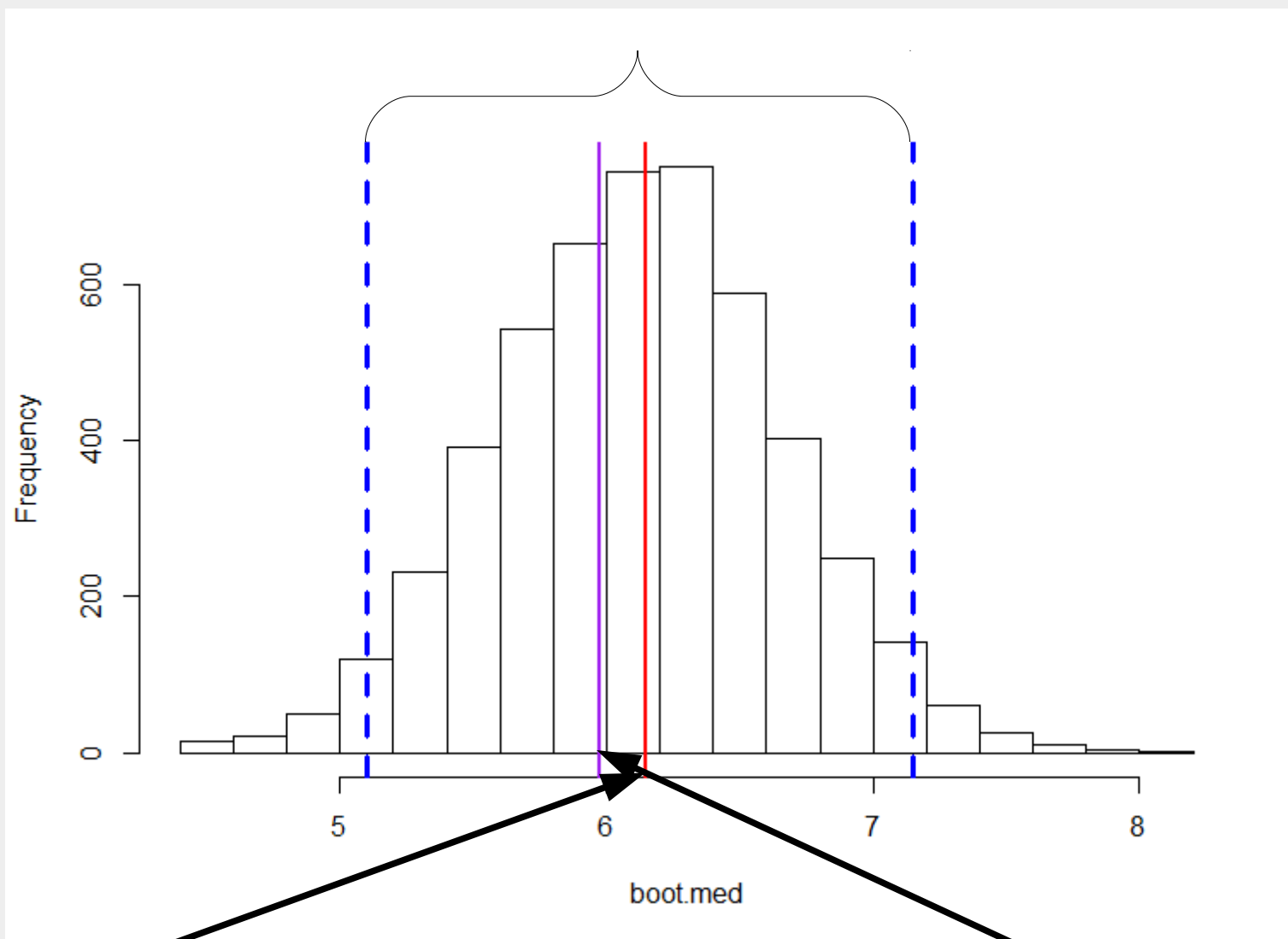
# Intervalo de confiança



Média da amostra

Média da população

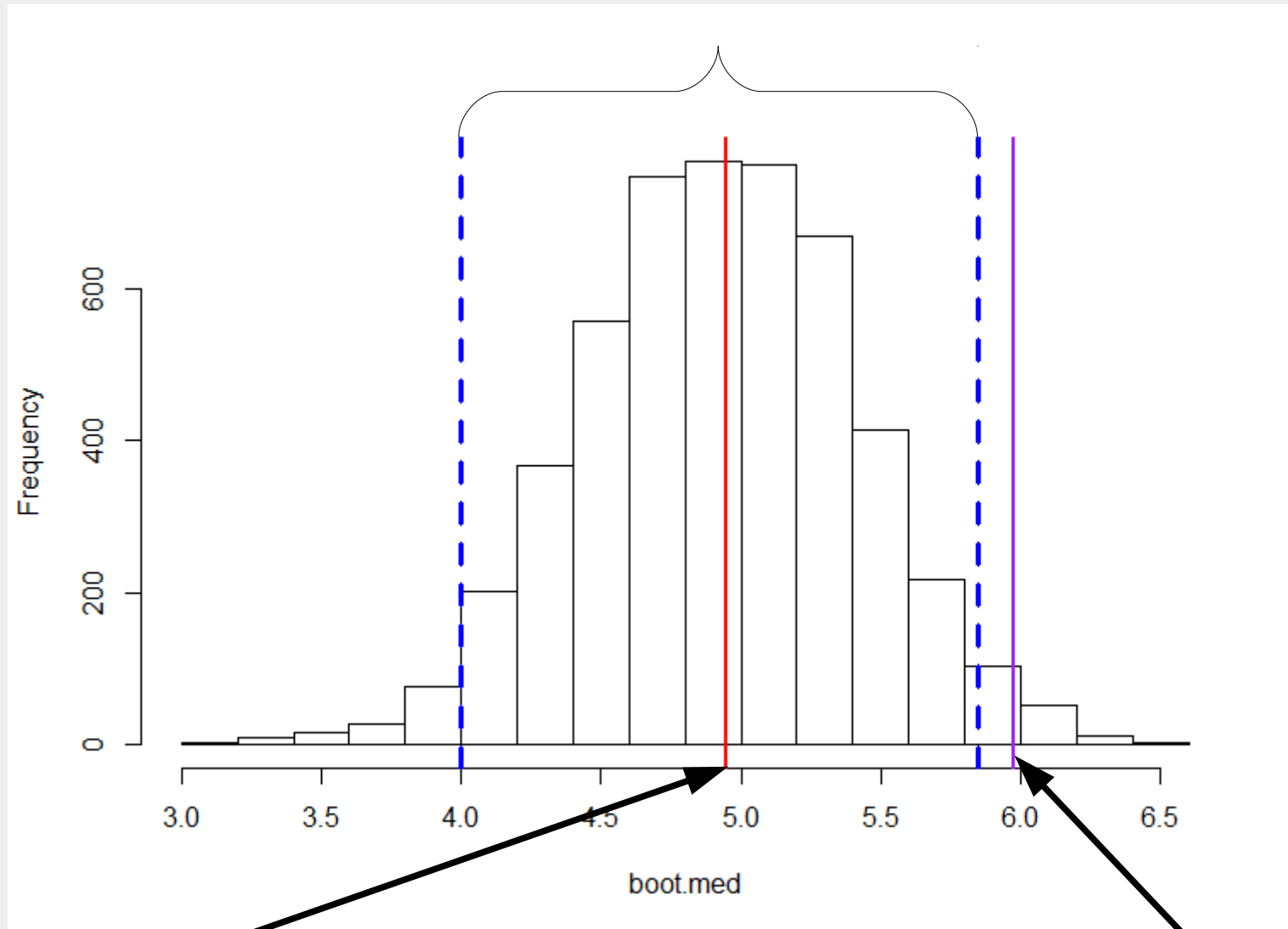
# Para outra amostra



Média da amostra

Média da população

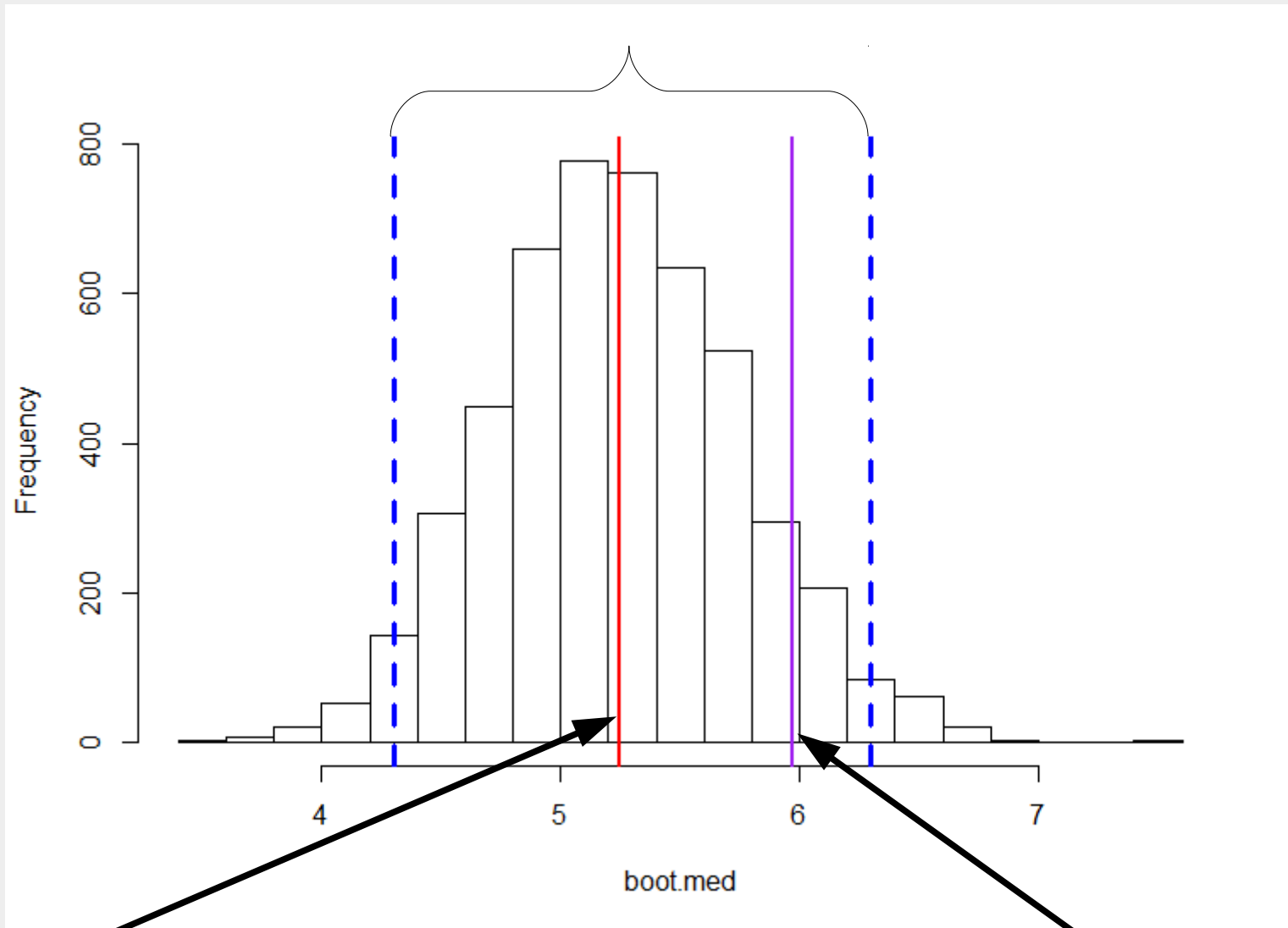
# Para outra amostra



Média da amostra

Média da população

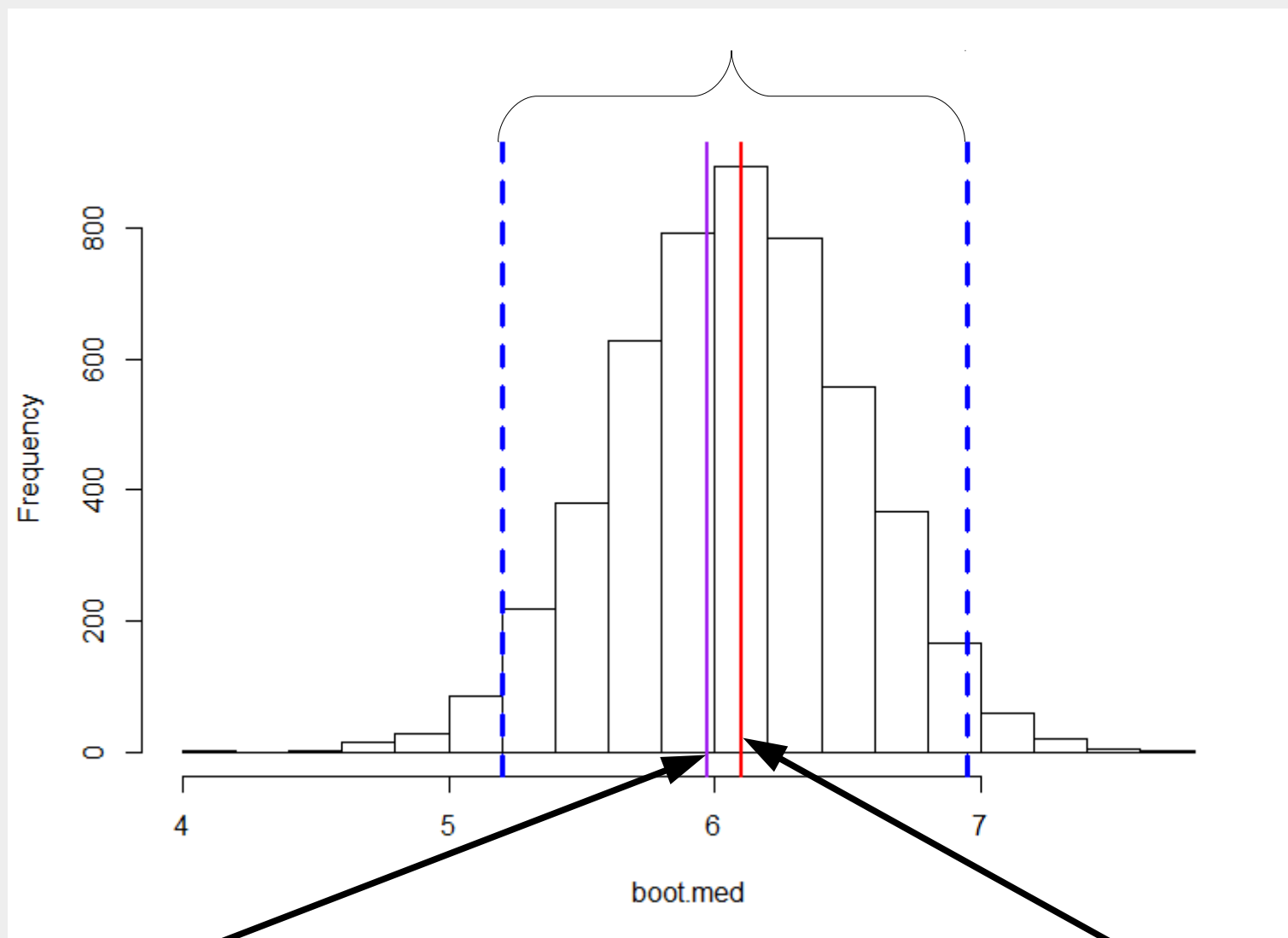
# Para outra amostra



Média da amostra

Média da população

# Para outra amostra



Média da amostra

Média da população

# Bootstrap padrão

1) Reamostra e calcula a média M vezes

2) Calcula o desvio padrão das médias

3) Usa a fórmula analítica

$$CI = \bar{X} \pm Z_{\alpha(2)} \sigma_{\bar{X}}$$

$$Z_{\alpha(2)} = 1.96$$

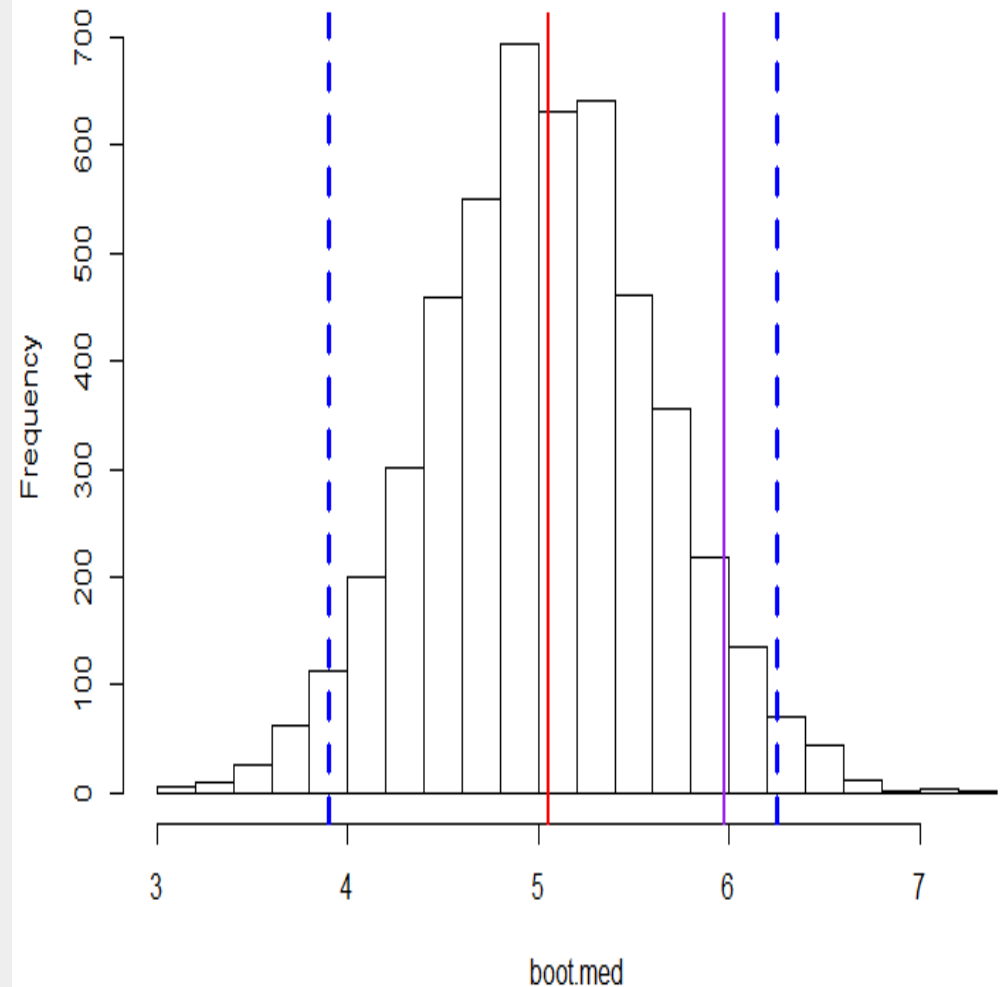
# Standard Bootstrap Confidence Limits

- Premissas
  - distribuição aproximadamente normal do  $\theta_{est}$
  - valor médio de  $\theta_{est} = \theta$
  - reamostragem por bootstrap dá uma boa aproximação de  $\sigma_{\theta}$ .



# Simple Percentile Confidence Limits - Efron

- Reamostragem com reposição
  - Os valores que contêm e.g. 95% das estimativas por bootstrap
    - Efron 1979

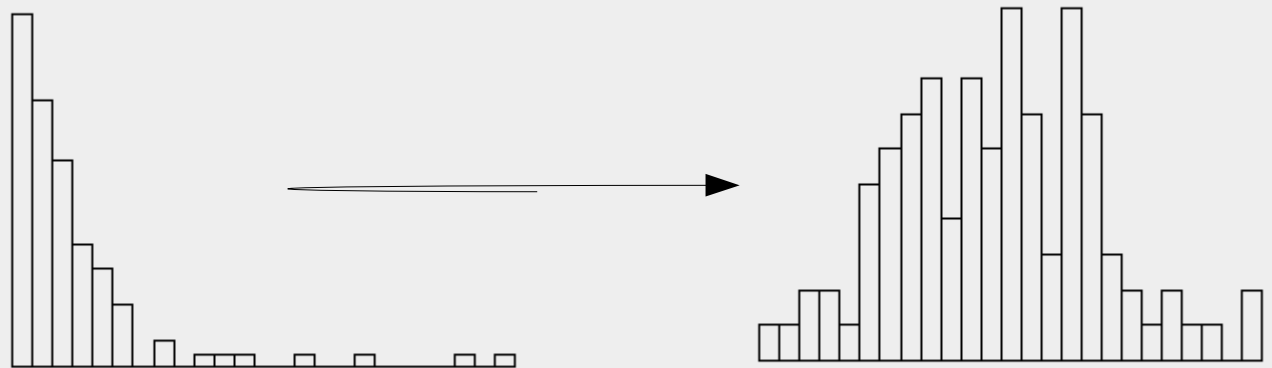




# Simple Percentile Confidence Limits - Efron

- Premissas

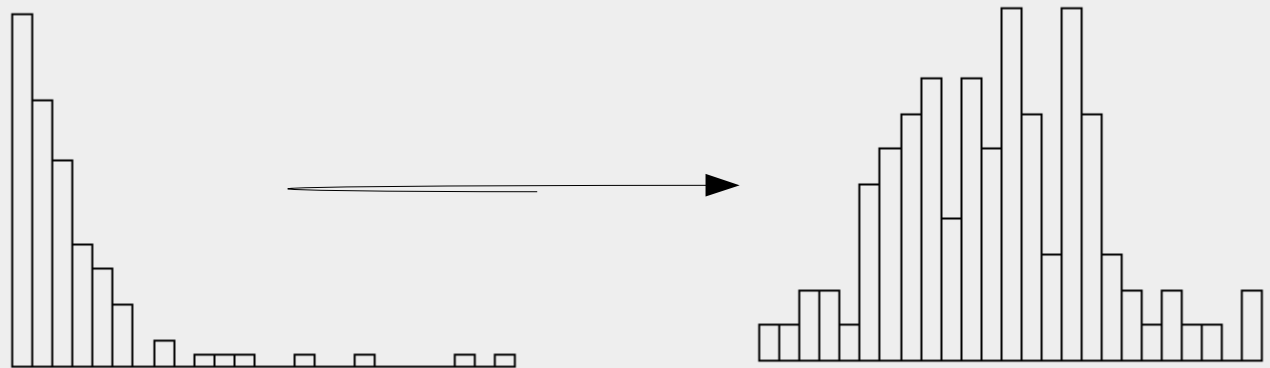
- Existe uma transformação que converteria a distribuição do estimador em consideração em uma distribuição normal



# Simple Percentile Confidence Limits - Efron

- Premissas

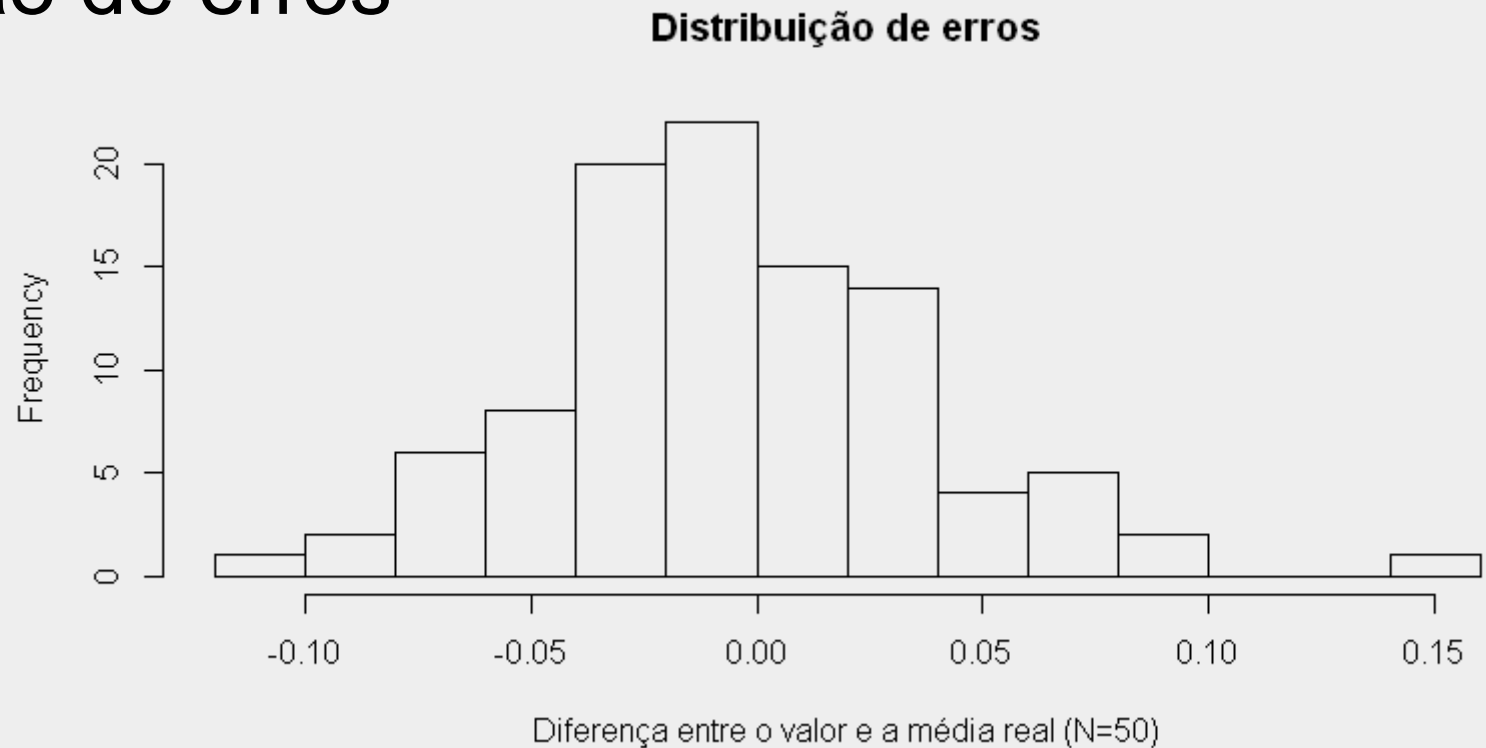
- Existe uma transformação que converteria a distribuição do estimador em consideração em uma distribuição normal
- "Olhar a tabela estatística errada de trás pra frente" (Hall 1992)



# Simple Percentile Confidence Limits - Hall

- Distribuição de erros

$$\varepsilon = \hat{\theta} - \theta$$
$$\varepsilon_B = \hat{\theta}_B - \hat{\theta}$$



$$Prob(\varepsilon_{inf} < \hat{\theta}_B - \hat{\theta} < \varepsilon_{sup}) = 1 - \alpha$$

$$Prob(\hat{\theta} - \varepsilon_{sup} < \theta < \hat{\theta} - \varepsilon_{inf}) = 1 - \alpha$$

# Bootstrap T

$$T = (\hat{\theta} - \theta) / \hat{SE}(\hat{\theta})$$

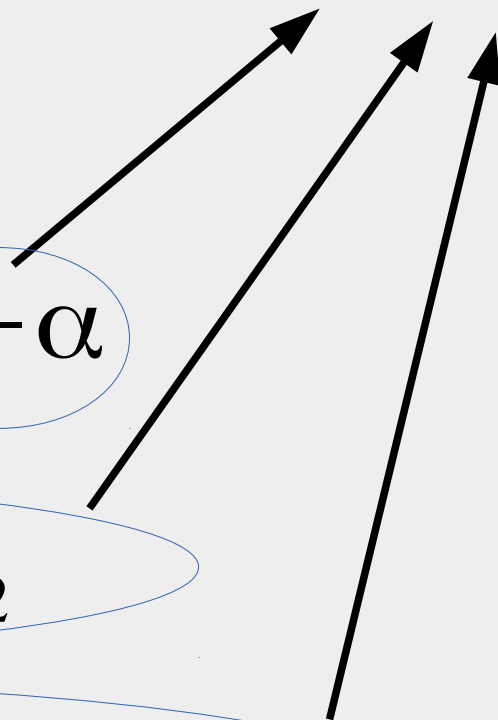
$$T_B = (\hat{\theta}_B - \hat{\theta}) / \hat{SE}(\hat{\theta}_B)$$

$$Prob(t_{1-\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$$

$$t_{1-\alpha/2} < (\hat{\theta} - \theta) / \hat{SE}(\hat{\theta}) < t_{\alpha/2}$$

$$\hat{\theta} - t_{\alpha/2} \hat{SE}(\hat{\theta}) < \theta < \hat{\theta} - t_{1-\alpha/2} \hat{SE}(\hat{\theta})$$

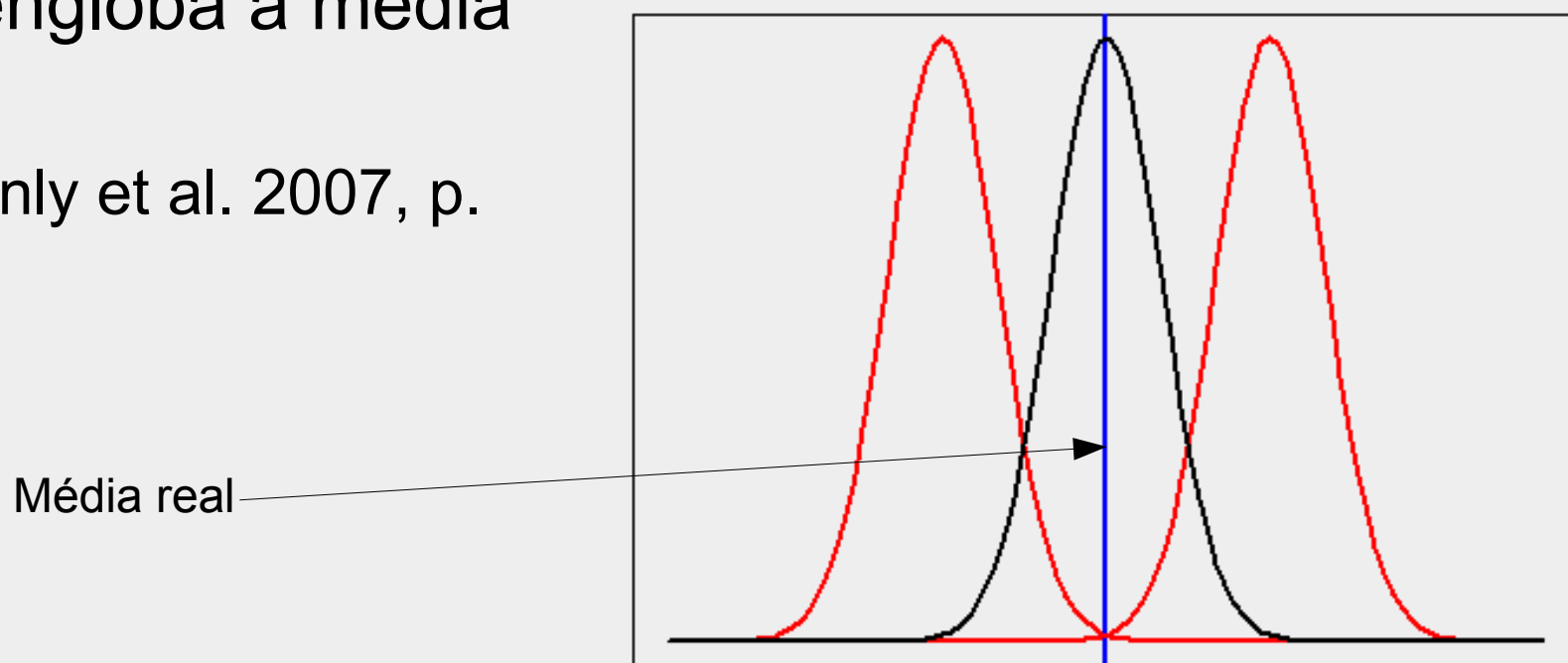
95%



# Comparando os tipos de bootstrap

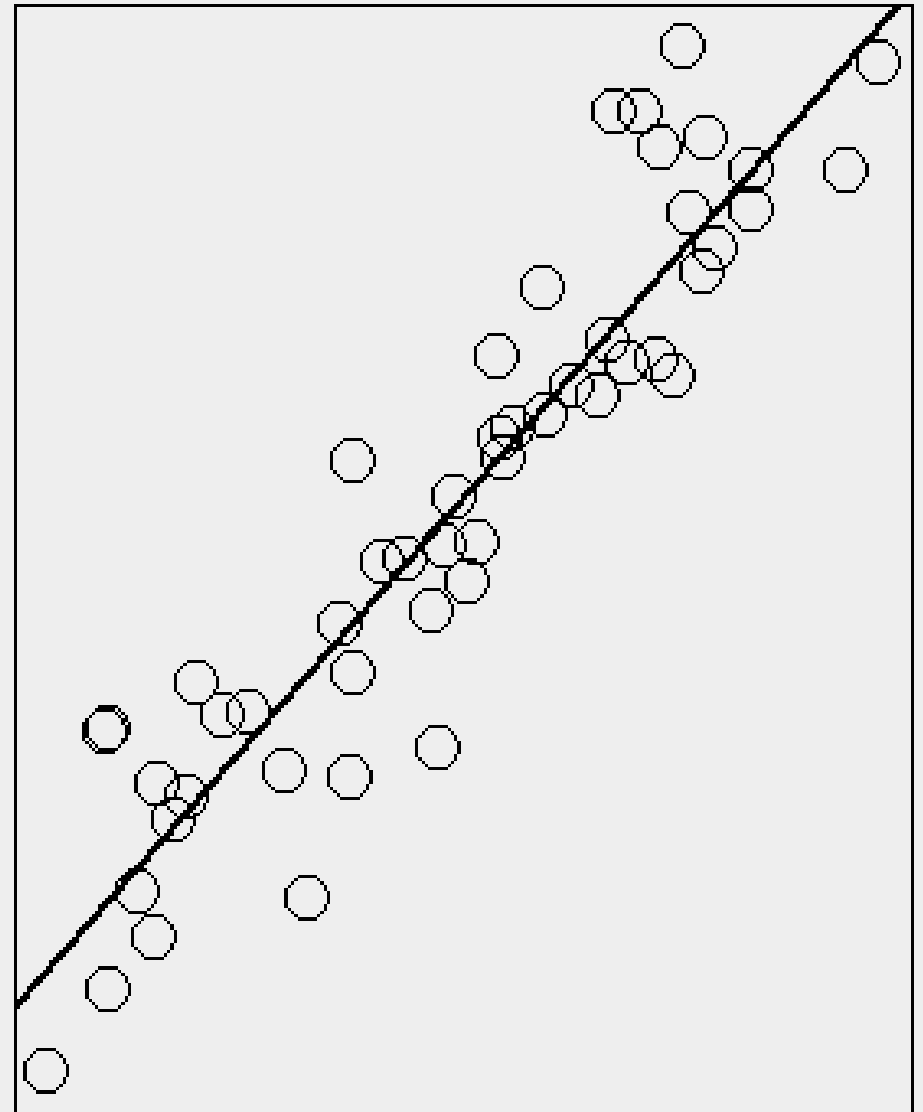
- Dados simulados (distribuição exponencial)
  - Porcentagem de vezes com que o IC não engloba a média real
    - Manly et al. 2007, p. 68

Tipo de bootstrap	% IC <sub>inf</sub> acima da média real	% IC <sub>sup</sub> abaixo da média real
Padrão	1.1	9.4
Efron	1.7	8.2
Hall	0.7	10.5
Bootstrat-T	1.4	3.4
<i>Desejado</i>	2.5	2.5



# Para outros parâmetros

- Nem só de médias é feita a vida!
  - Desvio padrão e erro padrão
  - Inclinação de uma reta
  - Resultado de uma análise qualquer...



# Seleção de modelos

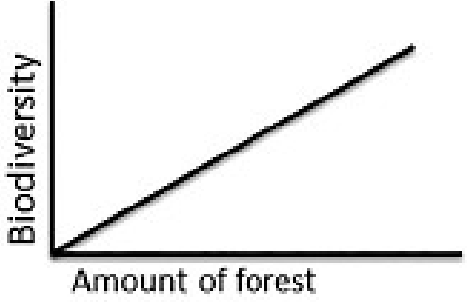
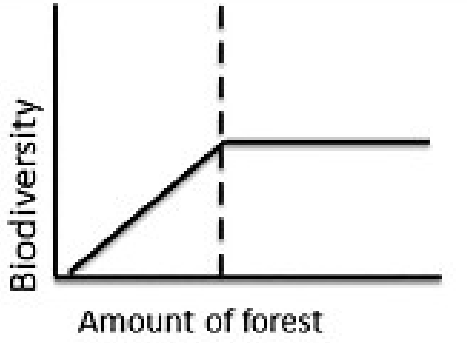
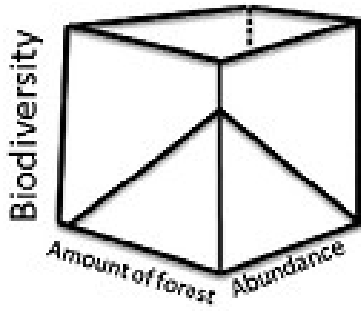
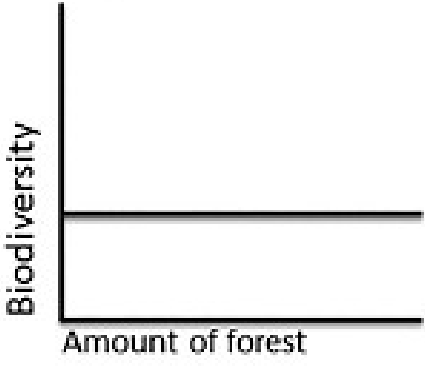
# Threshold effect of habitat loss on bat richness in cerrado-forest landscapes

RENATA L. MUYLAERT,<sup>1,3</sup> RICHARD D. STEVENS,<sup>2</sup> AND MILTON C. RIBEIRO<sup>1</sup>

<sup>1</sup>*Department of Ecology, Universidade Estadual Paulista (UNESP), 24A Av., 1515, 13506-900, Rio Claro, Brazil*

<sup>2</sup>*Department of Natural Resources Management, Museum of Texas Tech University, Lubbock, Texas 79409 USA*



Hypotheses	Description	Visual representation
Linear	Biodiversity increases on a linear trend in response to changes in total amount of forest of landscape.	
Fragmentation threshold	Biodiversity increases on a non linear trend in response to changes in total amount of forest in landscape, including a threshold point around 30%.	
Forest+abundance	Biodiversity increases in response to changes in total amount of forest in landscape, and to an increase in bat abundance.	
Null	Biodiversity does not vary in function of amount of forest.	

Species richness

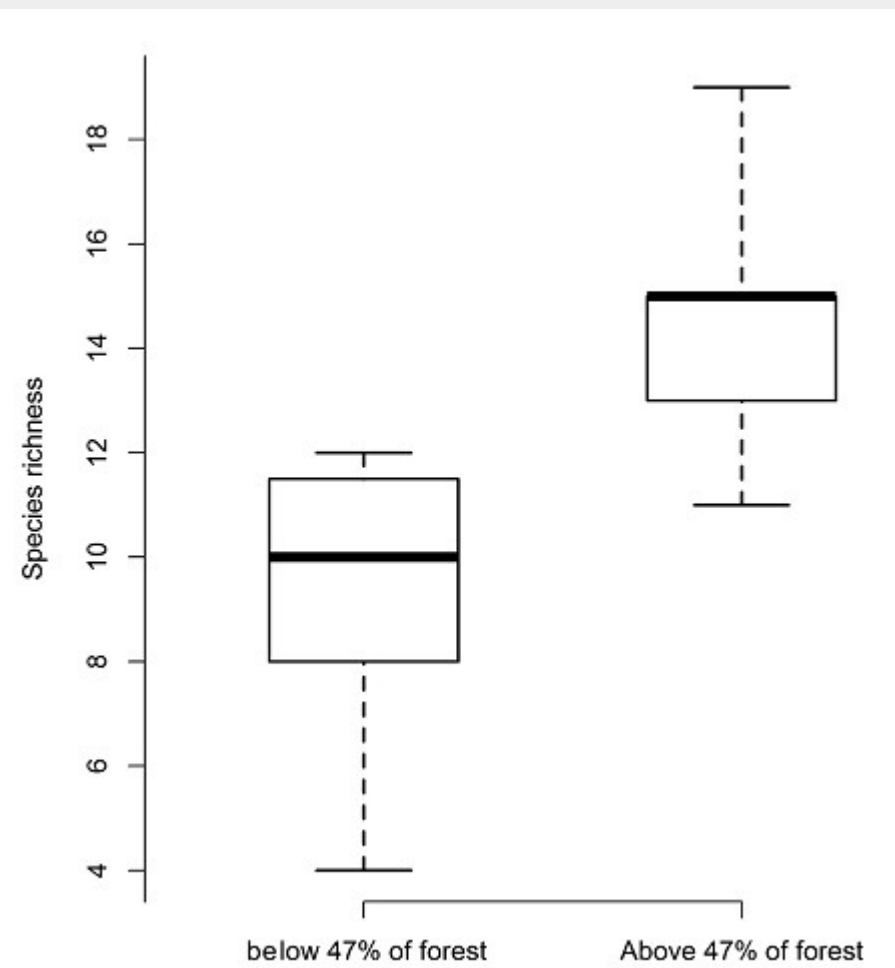
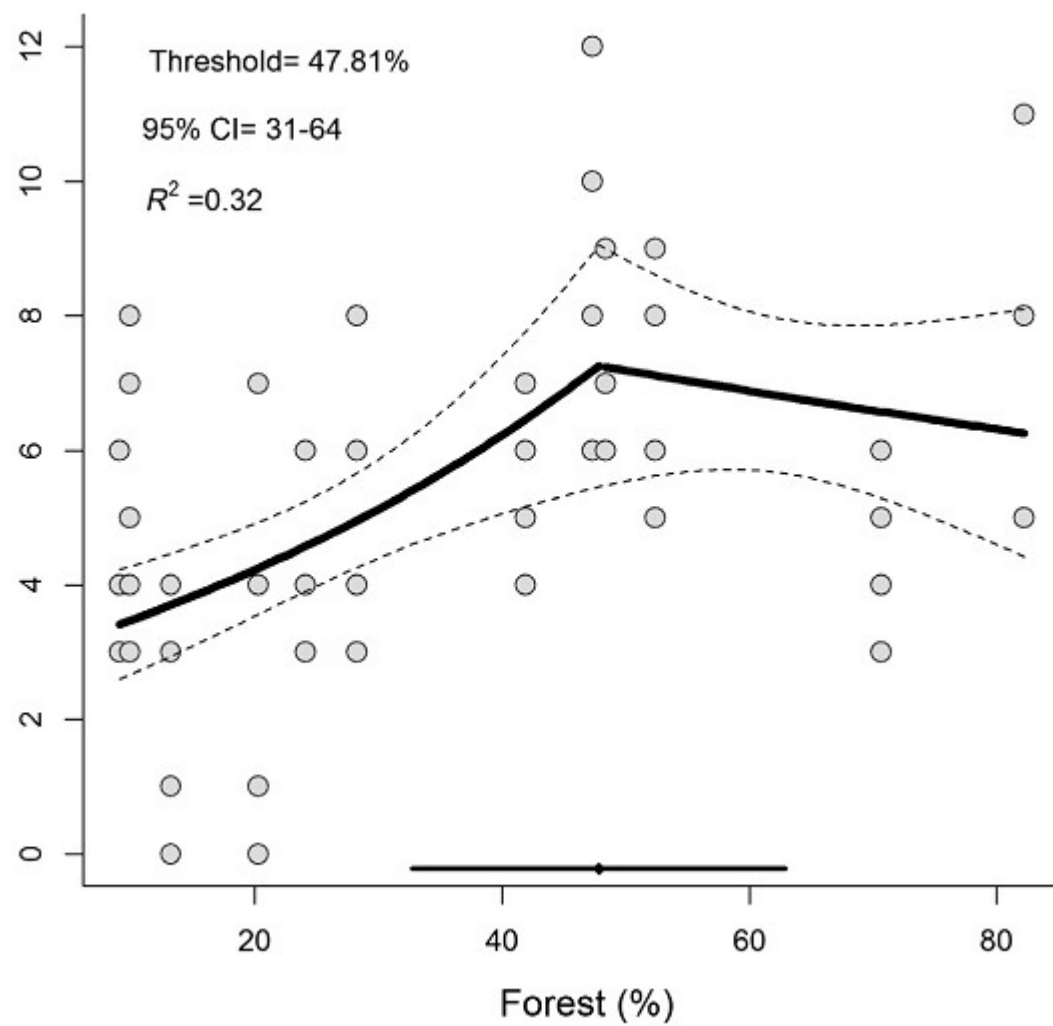
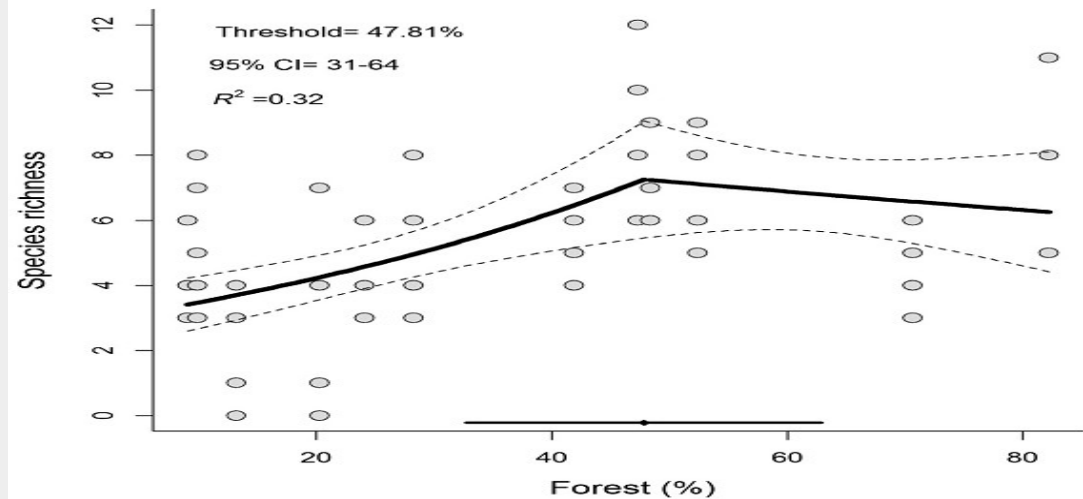
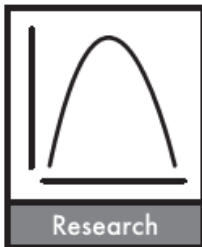


TABLE 2. Models describing the relationship between bat richness in response to amount of forest sampled within 12 landscapes of forest–savanna formations in southeastern Brazil.

Model	$\Delta\text{AIC}_c$	$K$	$w_i$	$R^2$	$\varpi_i$
<b>Piecewise</b>	<b>0</b>	<b>5</b>	<b>0.86</b>	<b>0.34</b>	<b>0.99</b>
Linear	6.7	4	0.08	0.21	0.01
Habitat + abundance	9	3	0.05	0.22	0
Null	17.7	2	<0.001	0.016	0

*Notes:* The best models ( $\text{AIC}_c = 0$ ) are in bold and the plausible models ( $\text{AIC}_c < 2$ ) are in bold and italics.  $K$  represents number of estimated parameters,  $\text{AIC}_c$  is the corrected Akaike Information Criterion,  $\Delta\text{AIC}_c$  is the Akaike difference,  $w_i$  is the Akaike weight, and  $\varpi_i$  is selection frequency (10000 bootstraps).





# Patch size, shape and edge distance influence seed predation on a palm species in the Atlantic forest

**Calebe P. Mendes, Milton C. Ribeiro and Mauro Galetti**

*C. P. Mendes (calebepm3@hotmail.com), M. C. Ribeiro and M. Galetti, Depto de Ecologia, Univ. Estadual Paulista (UNESP), 13506-900, Rio Claro, São Paulo, Brazil.*

Hypothesis	Description of the expected responses	Sketch of expected responses
<b>H1: Rodentation hypothesis</b>	A positive relationship exists between degradation and seed predation, due to the high presence of rodents in more degraded areas	
<b>H2: Intermediate hypothesis</b>	Seed predation is high in areas with intermediate levels of degradation, because rodents thrive in these areas	
<b>H3: Invertebrate control hypothesis</b>	Areas with high seed predation by rodents should have low seed predation by invertebrates, because rodents prey upon invertebrate larvae and destroy their reproduction sites	
<b>H4: Predator turnover hypothesis</b>	The overall seed predation is not affected by any of the tested variables, but there is a turnover between the groups of seed predators	
<b>H0: Null hypothesis</b>	Neither the overall seed predation nor the main groups of seed predators are affected by habitat degradation	

---

## Models

---

GAM0: SP ~ Mean (Null model)

GAM01: SP ~ Edge distance (m)

GAM02: SP ~ Fruit amount

GAM03: SP ~ Fragment area (ha)

GAM04: SP ~ Fragment shape (index)

GAM05: SP ~ Forest cover (only the best scale)

GAM06: SP ~ Fruit amount + Edge distance

GAM07: SP ~ Fragment area + Edge distance

GAM08: SP ~ Fragment area + Fragment shape

GAM09: SP ~ Forest cover + Edge distance

GAM10: SP ~ Rodent activity (proxy)

GAM11: SP ~ Edge distance + Rodent activity

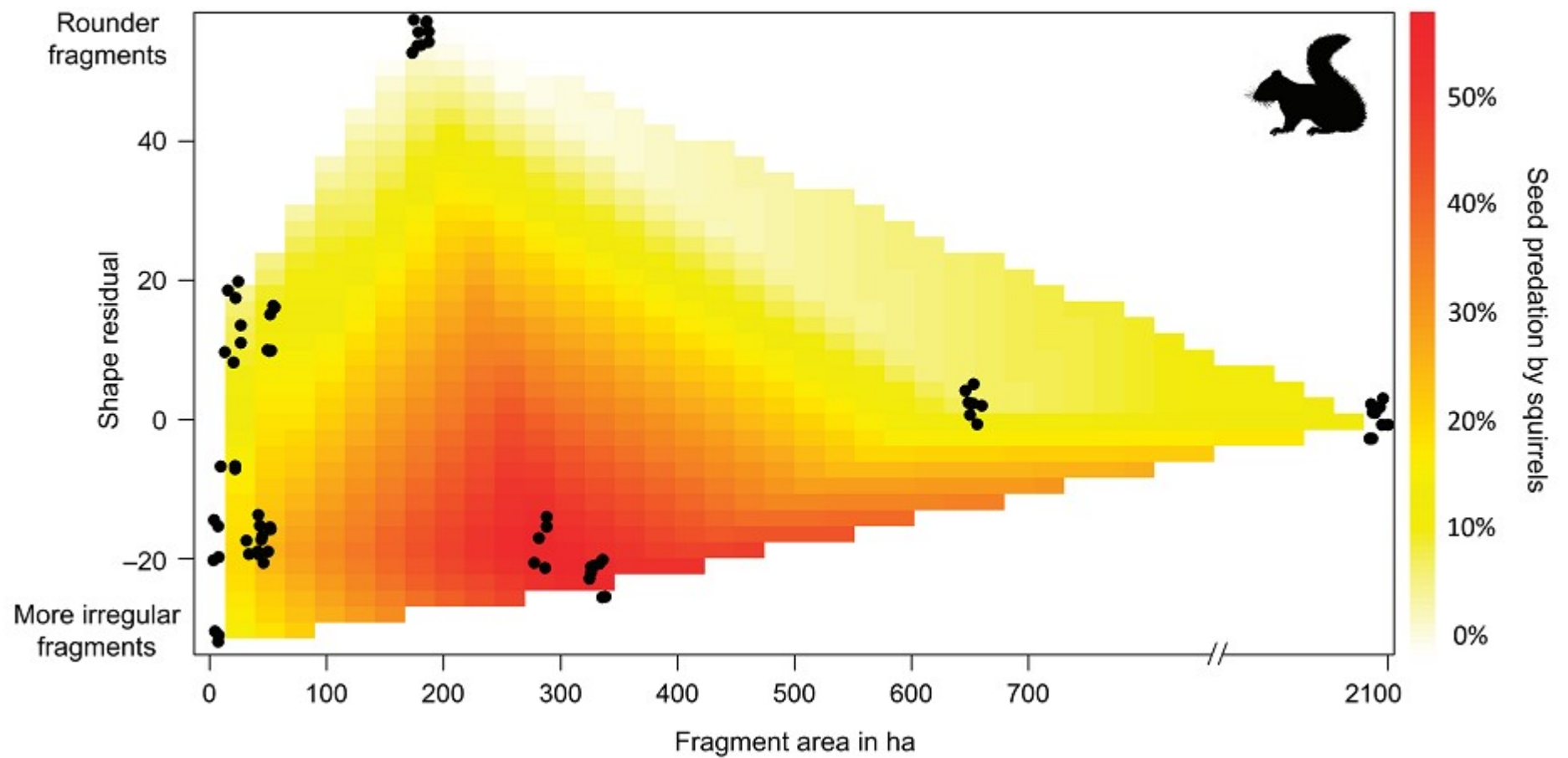
GAM12: SP ~ Fragment area + Rodent activity

GAM13: SP ~ Forest cover + Rodent activity

GAM14: SP ~ Fruit amount + Rodent activity

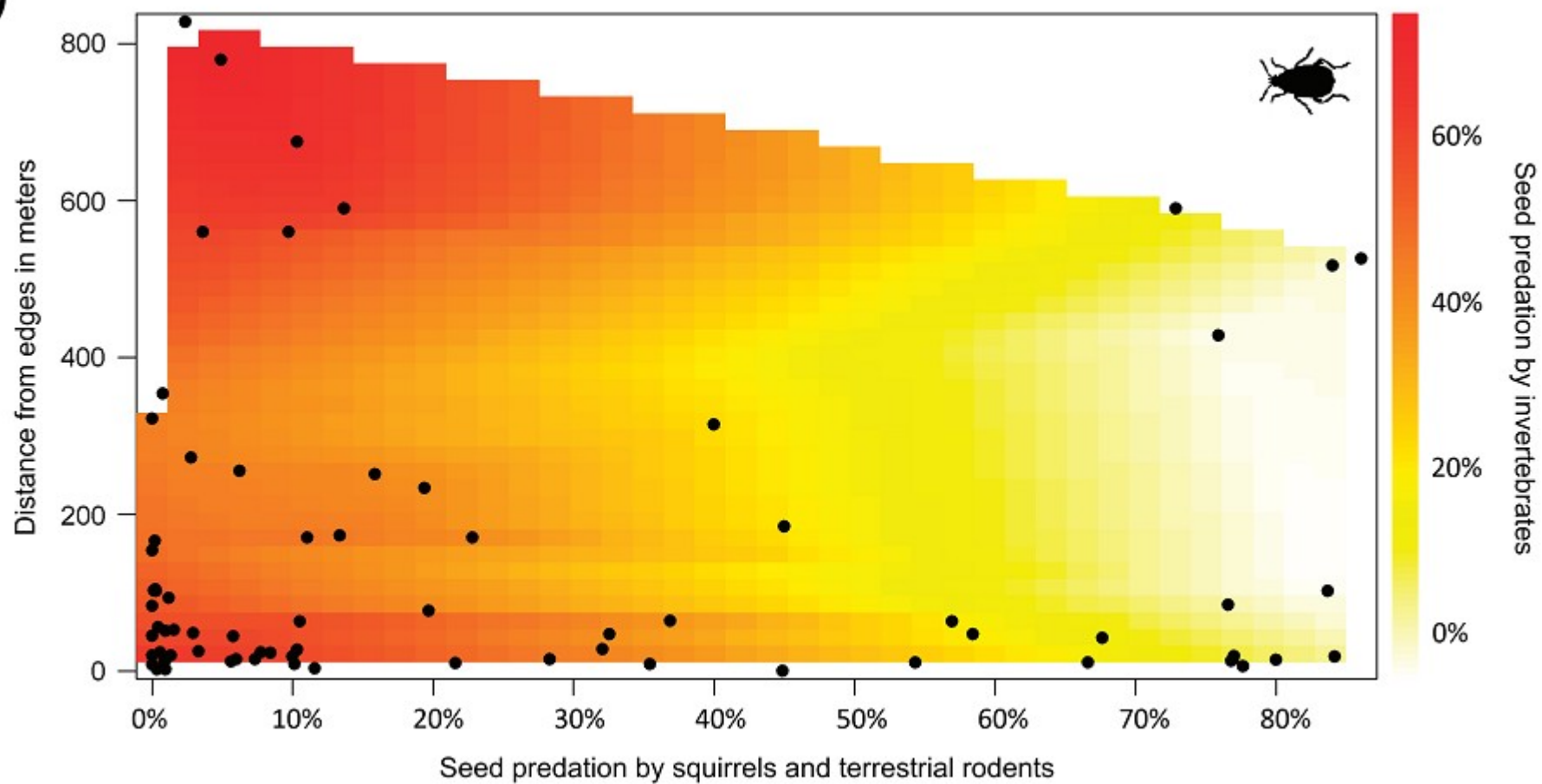
---

(A)



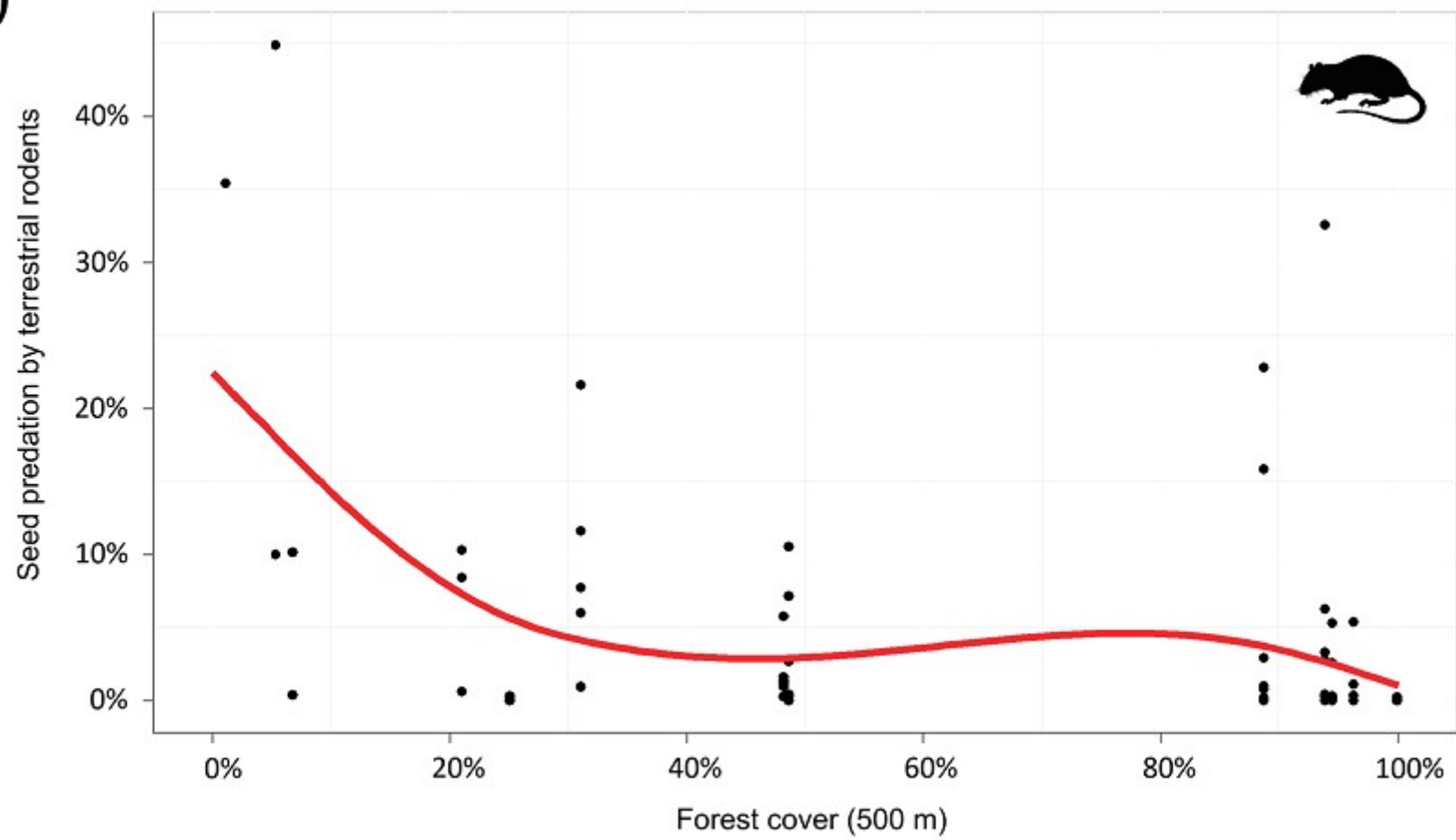


(B)





(C)



# *Bootstrap* e seleção de modelos

- 1) Calcula o AIC para os modelos
- 2) Seleciona o modelo com o menor AIC
- 3) Reamostra os dados (com reposição)  
(Mantendo as relações entre as variáveis)
- 4) Anota qual modelo teve o menor AIC
- 5) Repete passos 3 e 4 muitas (e.g. 5000) vezes
- 6) Calcula a frequência com que cada modelo foi selecionado