Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2023

Programa

- ➢ Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores
- ✓ Estrutura de Dados: variáveis (resposta, explicativa), unidades amostrais e experimentais

- ✓ 1.1. Comparação de 2 ou mais grupos: Testes Clássicos (teste t, Wilcoxon, ANOVA), Testes de Aleatorização, Comparações Múltiplas, Efeitos Genéticos
- ✓ 1.2. Análise de Tabelas de Contingência: Testes Qui-Quadrado, Regressão Logística.
- ✓ 2. Análise Multivariada de Dados: Componentes Principais, Coordenadas Principais, Análise de Correspondência, Análise Discriminante (MANOVA), Análise de Agrupamento, Correlação Canônica

3. **Reamostragem Bootstrap, Simulação de Monte Carlo**

Métodos Computacionalmente Intensivos

Lembre que, na Comparação de Duas Populações, introduzimos:

$\Rightarrow$ **Simulamos Intervalos de Confiança para μ**

$\Rightarrow$ Realizamos **Testes de Permutação** para calcular o valor-p, como alternativa a usar tabelas da “t” ou de Wilcoxon.

Intervalos de Confiança para o Parâmetro μ

Suposição sobre uma única População sob estudo:

Y: resposta de interesse, **$Y \sim N(\mu,\sigma^2)$**

Objetivo: Realizar inferência sobre a Média de Y, μ, com σ desconhecido

Amostra aleatória: ***n*** unidades amostrais são extraídas aleatoriamente da população sob estudo.

Contextualizamos esta situação para realizar inferências sobre a **Pulsação de Estudantes (variável P1)** na situação em repouso

$\Rightarrow$ **Dados**: Pulsação em repouso (batimentos/minuto) avaliada em n=92 estudantes

$\Rightarrow$ **Resultados amostrais**: Média = 72.86957 Desvio padrão = 11.00871

$$IC95\%(\mu) = (\text{Média} \pm t_{91,0.975} \times \text{Desvio Padrão}/\sqrt{92}) = (70.59 ; 75.15)$$

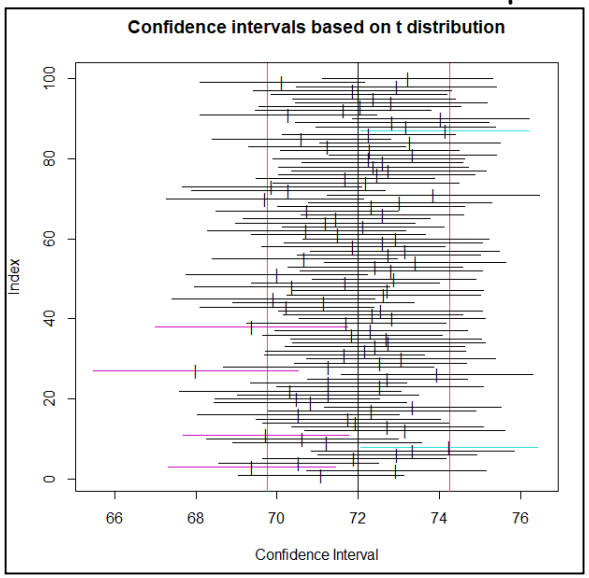
Interprete o IC95%(μ)

Intervalos de Confiança para o Parâmetro μ

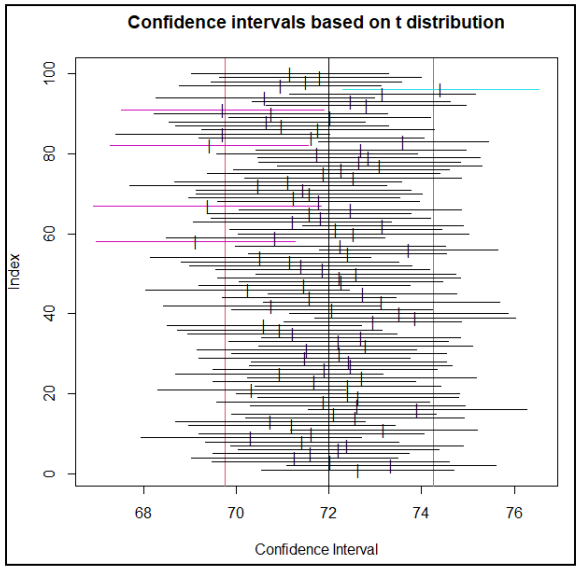
Estudo de simulação:

- ✓ Gerar aleatoriamente k=100 amostras de tamanho n=92 a partir de Y ~ N(μ=72,σ^2)
- ✓ Para cada amostra calcular o IC95% para μ (supondo σ=11)
- ✓ Repetir L=3 vezes o estudo de simulação.

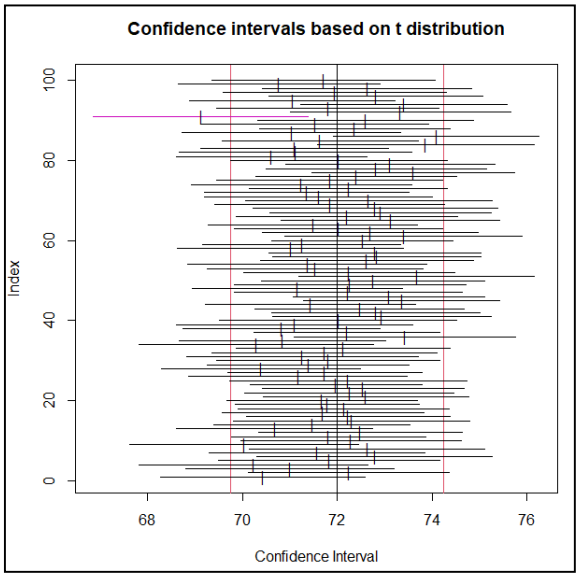
L = 1: 6 IC não contêm μ



L = 2: 5 IC não contêm μ



L = 3: 1 IC não contém μ



O pesquisador tem 95% de confiança que o intervalo calculado com seus dados contenha μ.

MAS

Se 100 experimentos forem realizados sob as mesmas condições e forem construídos os correspondentes 100 IC95%(μ) é esperado que 5 deles NÃO contenha o verdadeiro valor de μ.

Testes de Aleatorização ou Permutação

> **tomato.data do "R"**

cant	prod	fertiliz
1	29.9	A
2	11.4	A
3	26.6	B
4	23.7	B
5	25.3	A
6	28.5	B
7	14.2	B
8	17.9	B
9	16.5	A
10	21.1	A
11	24.3	B

Grupo	**A**	**B**	
Média	20.84	22.53	
Desvio padrão	7.25	5.43	⇒ sc = 6.30

Teste t unicaudal: variâncias homogêneas
t = -0.44368, df = 9, p-value = 0.3339

Suposição:

$$Y_{Ai} \overset{iid}{\sim} N\left(\mu_A;\sigma^2\right)$$

$$Y_{Bi} \overset{iid}{\sim} N\left(\mu_B;\sigma^2\right)$$

Teste: $H_0 : \mu_A = \mu_B$

$$t_{obs} = \frac{\bar{y}_A - \bar{y}_B}{s_c\sqrt{\frac{1}{5}+\frac{1}{6}}} \overset{H_0:\mu_A=\mu_B}{\sim} t_{11-2}$$

Alternativa: Teste de Permutação

A partir dos dados observados (11 respostas), obter todas as amostras possíveis que poderiam ter sido geradas por particionar aleatoriamente as 11 respostas em 5 para A e 6 para B ⇒ Existem 462 combinações possíveis (11!/5!6!)

Suposição: no plano experimental do estudo houve a aleatorização dos Tratamentos (A e B) às 11 unidades amostrais

Teste de Permutação (library BHH2, permtest) Teste bicaudal

N	t.obs	t-Dist:P(>t)	**PermDist:P(>t)**
462	-0.44368	0.6661273	**0.6666667**

Testes de Aleatorização: Comparação de 2 Populações

Dados: Tomato.data do R
Significância avaliada sob premissas clássicas

Suposição: amostras independentes de Normais homocedásticas

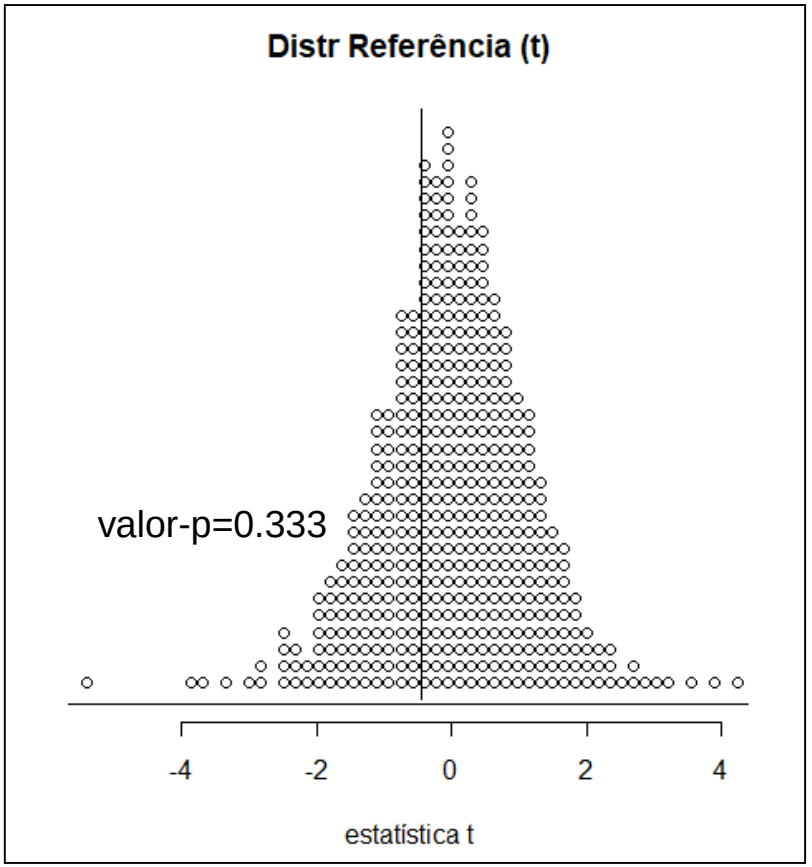
$$Y_{Ai} \overset{iid}{\sim} N\left(\mu_A;\sigma^2\right); \quad Y_{Bi} \overset{iid}{\sim} N\left(\mu_B;\sigma^2\right)$$

$$t_{obs} = \frac{\bar{y}_A - \bar{y}_B}{s_c\sqrt{\frac{1}{5}+\frac{1}{6}}} \overset{H_0:\mu_A=\mu_B}{\sim} t_{11-2}$$

Teste t unicaudal:

```
(variâncias homogêneas)
t = -0.44368, df = 9,
p-value = 0.3339
```

Dados: Tomato.data do R
Significância avaliada na Distribuição de Referência construída da análise das 462 amostras possíveis



t_{obs}=-0.44368 valor-p=154/462=0.333

Métodos Computacionalmente Intensivos

⇒ **Testes de Permutação** (Oehlert, 2010): uma distribuição de referência é construída (com base na permutação de dados amostrais) para avaliação da significância de uma estatística

⇒ **Reamostragem Bootstrap** (Efron, 1979): reamostrar com reposição um conjunto de dados (a amostra efetivamente observada) com o objetivo de estimar um parâmetro (baseado nos dados)

⇒ **Métodos de Monte Carlo** (Ross, 1997): o objetivo principal é realizar inferências sobre um sistema simulando ele por meio de amostras aleatórias.

Soluções Teóricas:	**x**	-Soluções baseadas nos dados efetivamente observados: significância calculada sob uma distribuição empírica -Soluções baseadas em simulação MC
-Significância calculada sob premissas clássicas -Média e variância de estimadores calculados analiticamente		

Métodos Computacionalmente Intensivos

$\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ Amostra de tamanho ***n***

Testes de Permutação
obter amostras permutadas de **Y**
n! = n(n-1)(n-2)...1

n=10 $\Rightarrow$ n!=3.628.800

Amostras Bootstrap
reamostrar com reposição **Y**
n^n = nnn...n

n=10 $\Rightarrow$ n^n=1e+10

Aproximação Monte Carlo
Soluções aproximadas baseadas em Simulação de Monte Carlo: obter um número de amostras suficientemente grande do Sistema sob estudo

Aproximação MC para a distribuição Bootstrap, para Intervalos de Confiança Bootstrap, para Testes de Permutação
Outros exemplos: Estimador Jackknife, Método Delta, Validação Cruzada
…

Simulação de Monte Carlo

⇒ A ideia principal dos métodos de MC é realizar inferências sobre um sistema **simulando o sistema** (ou características dele) por meio de um número suficientemente grande de **amostras aleatórias**.

⇒ Os métodos de MC são, em geral, formulados com base no comportamento médio (esperado) das amostras, isto é, envolvem valores esperados, estimativas de médias (cujos cálculos envolvem integrais).
No método de **MC simples** o objetivo é estimar uma média (esperança) populacional pela correspondente média amostral!

Soluções teóricas x Soluções via MC

Mooney (1997): método geral de simulação MC
Passo 1: Especifique uma pseudo-população **P**, tal que ela possa ser usada para gerar amostras do Sistema sob estudo
Passo 2: Obtenha uma amostra de tamanho **n** (**A1**) de **P**
Passo 3: "Estime" o parâmetro de interesse em **A1** e armazene em um vetor
Passo 4: Repita os passos 2 e 3 **L**-vezes, em que L é o número de ensaios
Passo 5: Construa a distribuição de frequência das estimativas resultants, a qual é uma estimativa MC da distribuição amostral do estimador de interesse

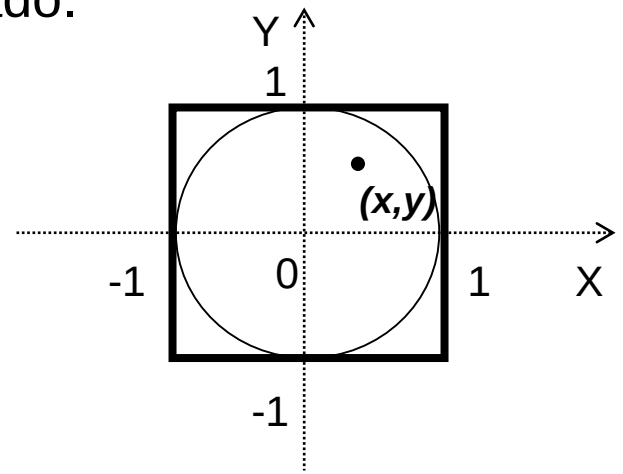
Aplicação

- Suponha que o vetor aleatório (X,Y) é distribuído uniformemente no quadrado de área 4 centrado na origem. Considere o cálculo da probabilidade de um ponto aleatório (x,y) nesse quadrado estar contido dentro do círculo de raio 1 inscrito no quadrado.

Solução teórica:

$$\Rightarrow P\left((x,y) \in circulo\right) = P\left(x^2 + y^2 \leq 1\right)$$

$$= \frac{\text{área do círculo}}{\text{área do quadrado}} = \frac{\pi r^2}{b \times h} = \frac{\pi}{4}$$



Aplicação de MC no cálculo do valor π

Solução por MC:

- Gerar aleatoriamente pontos no quadrado (um grande número de vezes)
- A proporção de pontos que caem dentro do círculo será, aproximadamente, $\pi/4$.

Aplicação

- **Estimação do Número π via Simulação MC no R**

```
> L<-1000 #número de amostras (ensaios) a serem gerados
> n<-2    #tamanho da amostra
> z<-numeric(0)
> for(i in 1:L){
    xy<-runif(n)  #amostra de tamanho n da U(0,1)
    z[i]<-xy[1]^2+xy[2]^2 #z calculado para cada amostra
 }
> vi<-ifelse(z<1,1,0) # variável indicadora
> sum(vi)/L
[1] 0.787   #estimativa MC de π/4
```

Simulação MC para estimar o valor de π:

Simulações	Valor aproximado	Erro
L=50	3.280000	0.13840735
L=100	3.120000	0.02159265
L=1.000	3.148000	0.00640735
L=1.000.000	3.141616	0.00002335

Aplicação

- **Formulação teórica da Estimação do Número π via Simulação MC**

Pode ser mostrado que:

$$X \sim U(-1;1) \quad Y \sim U(-1;1); \quad X \perp Y$$

A densidade conjunta de (X,Y) é dada por:

$$f_{XY}(x,y) = f_X(x)\, f_Y(y) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}; \quad -1 \le x, y \le 1$$

$$U \sim U(0;1); \quad 2U \sim U(0;2); \quad (2U-1) \sim U(-1;1) \quad \Rightarrow X = 2U_1 - 1, \quad Y = 2U_2 - 1$$

$$\Rightarrow I = \begin{cases} 1 & se \quad x^2 + y^2 \le 1 \\ 0 & cc \end{cases}$$

$$\Rightarrow E(I) = P(I=1) = P\left(X^2 + Y^2 \le 1\right) = \frac{\pi}{4}$$

Assim, π/4 pode ser estimado gerando aleatoriamente um grande número de pares de valores da Uniforme, U(0,1), digamos (u1, u2) e calculando a fração dos pares para os quais $(2u1 - 1)^2 + (2u2 - 1)^2 \le 1$.

Simulação Monte Carlo

Objetivo é estimar uma média populacional usando a média amostral

Resultado teórico (não precisa e especificar a forma da distribuição

$$Y \sim \left(\mu;\sigma^2\right); \quad \mu = E\left(Y\right) < \infty; \quad \left(Y_1,...,Y_n\right) \quad AAS_n \text{ de } Y; \quad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n$$

Lei Fraca dos Grandes Números: $\lim_{n\to\infty} P\left(\left|\hat{\mu}_n - \mu\right| \le \varepsilon\right) = 1$

Lei Forte dos Grandes Números: $P\left(\lim_{n\to\infty} \left|\hat{\mu}_n - \mu\right| = 0\right) = 1$

o método MC é não viciado pois $E\left(\hat{\mu}_L\right) = \mu$ e, eventualmente, produzirá um erro tão pequeno quanto nós desejarmos (fixando um ε)!!

$$Y \sim \left(\mu;\sigma^2\right); \quad \mu < \infty; \quad 0 < \sigma^2 < \infty;$$

$$\left(Y_1,...,Y_n\right) \quad AAS_n \text{ de } Y; \quad \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i; \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_n\right)^2; \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_n\right)^2$$

Teorema Limite Central: $\hat{\mu}_n - \mu \overset{n\to\infty}{\sim} N\left(0;\sigma^2/n\right)$

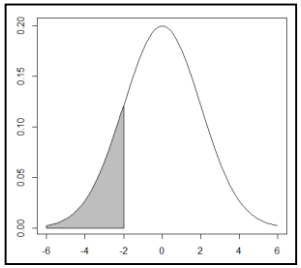
Erro Quadrático Médio do método MC: $\hat{E}\left(\hat{\mu}_L - \mu\right)^2 = \sigma^2 / L$

Intervalo de Confiança de Monte Carlo

Teorema Limite Central: $Y \sim (\mu; \sigma^2); \quad \mu < \infty; \quad 0 < \sigma^2 < \infty; \quad \hat{\mu}_n - \mu \overset{n\to\infty}{\sim} N(0; \sigma^2 / n)$

$$P\left(\frac{\hat{\mu}_n - \mu}{\sigma / \sqrt{n}} \le z\right) \overset{n\to\infty}{\to} \Phi(z)$$

Função de distribuição acumulada da Normal padrão



Para σ conhecido ou n grande: $\Rightarrow IC(\mu) \ a \ (1-\alpha)100\% = \left(\hat{\mu}_n - z_{\alpha/2}\frac{s}{\sqrt{n}}; \ \hat{\mu}_n + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$

Para σ desconhecido: $\Rightarrow \quad IC(\mu) \ a \ (1-\alpha)100\% = \left(\hat{\mu}_n - t_{n-1}^{1-\alpha/2}\frac{s}{\sqrt{n}}; \ \hat{\mu}_n + t_{n-1}^{1-\alpha/2}\frac{s}{\sqrt{n}}\right)$

erro padrão

Suponha que o erro padrão do estimador de μ não é conhecido ⇒ Obter o Intervalo de Confiança Bootstrap para μ:

$Y = (y_1, y_2, ..., y_n) \Rightarrow \hat{\mu}_n$

é amostra aleatória

$\Rightarrow ICMC(\mu) \ a \ 95\% = \hat{\mu}_n \pm 1{,}96 \ \tilde{s}_{\hat{\mu}_n\_L}$ **?**

Obter L amostras bootstrap de Y e calcular o desvio padrão das L estimativas $\hat{\mu}_L$

Intervalo de Confiança de Monte Carlo

Intervalo de Confiança Bootstrap (ou ICMC):

$Y = (y_1, y_2, ..., y_n) \Rightarrow \hat{\mu}_n$

é amostra aleatória

Algoritmo:

1. Obter uma **amostra Bootstrap de tamanho n de Y**
2. Obter a estatística média amostral: $\hat{\mu}_b$
3. Repetir L vezes os passos 1 e 2: $(\hat{\mu}_1, ,..., \hat{\mu}_L)$
4. Calcular o erro padrão da estatística média amostral:

$$\tilde{s}_{\hat{\mu}_n\_L} = desvio\ padrão(\hat{\mu}_1, ,..., \hat{\mu}_L)$$

$$\Rightarrow ICMC(\mu)\ a\ 95\% = \hat{\mu}_n \pm 1{,}96\, \tilde{s}_{\hat{\mu}_n\_L}$$

Outra Alternativa: Obter a distribuição amostral das estimativas $\hat{\mu}_b$, isto é, a distribuição da estatística de interesse nas L réplicas bootstrap, e então obter os quantis (2,5% e 97,5%) dessa distribuição.

Aplicação

- Muitos resultados teóricos em Estatística são assintóticos (valem para ***n*** grande) e, na prática, muitas vezes não temos tamanhos amostrais suficientemente grandes! Assim, podemos checar quão robusta a teoria é para tamanhos amostrais usuais (pequenos).

Uso de simulação MC para avaliar a validade do Teorema Limite Central no caso da aproximação ***t*** em diferentes tamanhos amostrais!

$$Y_{1i} \sim N\left(\mu_1;\sigma^2\right)$$
$$Y_{2i} \sim N\left(\mu_2;\sigma^2\right)$$
$$\rightarrow H_0 : \mu_1 - \mu_2 = 0; \quad \sigma \text{ desconhecido}$$

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{H_0}{\sim} t_{n_1+n_1-2}; \qquad s_c^2 = \frac{\left(n_1-1\right)s_1^2+\left(n_2-1\right)s_2^2}{n_1+n_2-2}$$

Resultado teórico para qualquer tamanho amostral!

Para tamanhos amostrais "grandes" vale o Teorema Limite Central:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \overset{H_0}{\sim} N\left(0;1\right)$$

Gerar Amostras aleatórias (n_1 e n_2) e verificar este resultado para diferentes "n"!

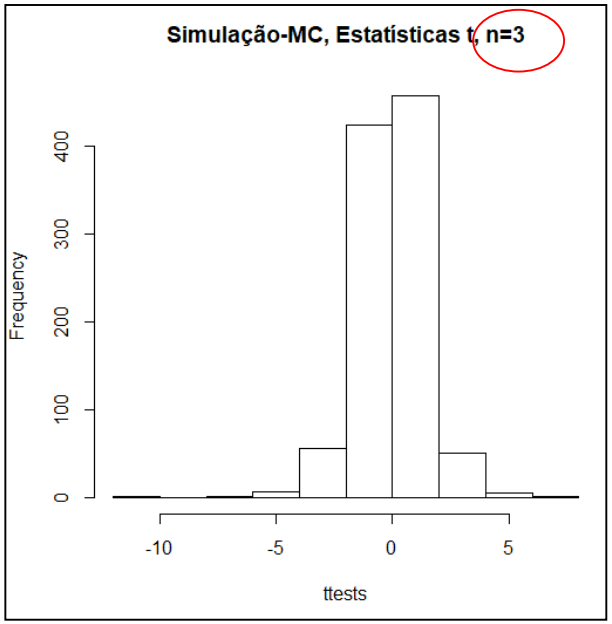
Simulação Monte Carlo

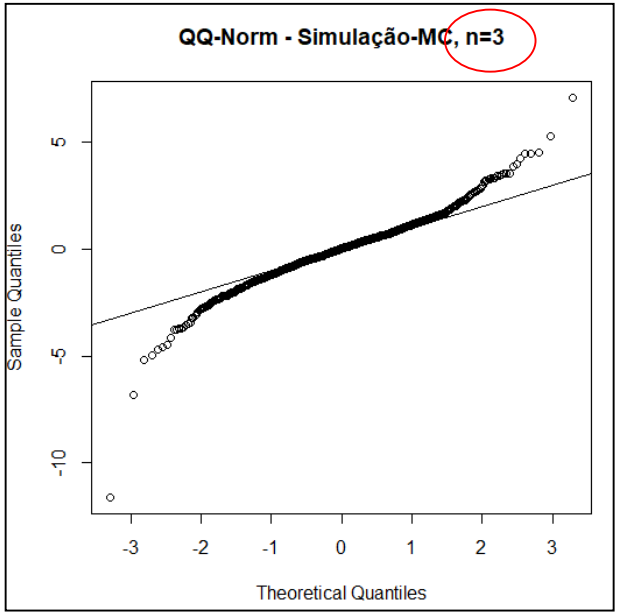
Validade do TLC:

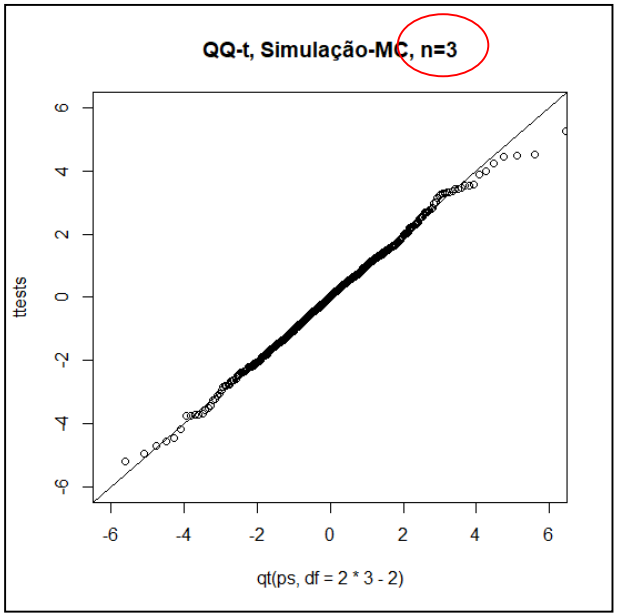
Simulação MC de amostras das distribuições Normais ($n_1 = n_2 = n$):

$$Y_{1i} \sim N\left(\mu_1;\sigma^2\right); Y_{2i} \sim N\left(\mu_2;\sigma^2\right)$$

- Adotar valores para os parâmetros e para n **(n=3)**
- Obter as amostras Bootstrap de cada distribuição (L=500)
- Cálcular a estatística *t* para cada amostra bootstrap
- Avaliar a distribuição dos valores *t* obtidos pela Normal e pela distribuição t





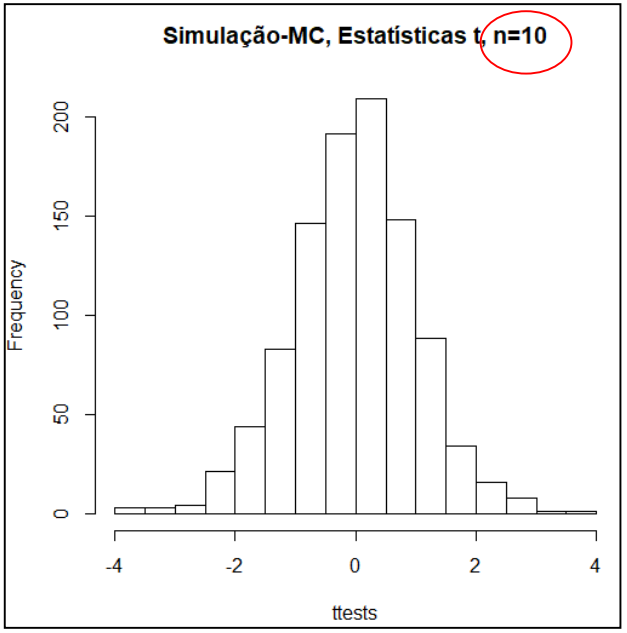


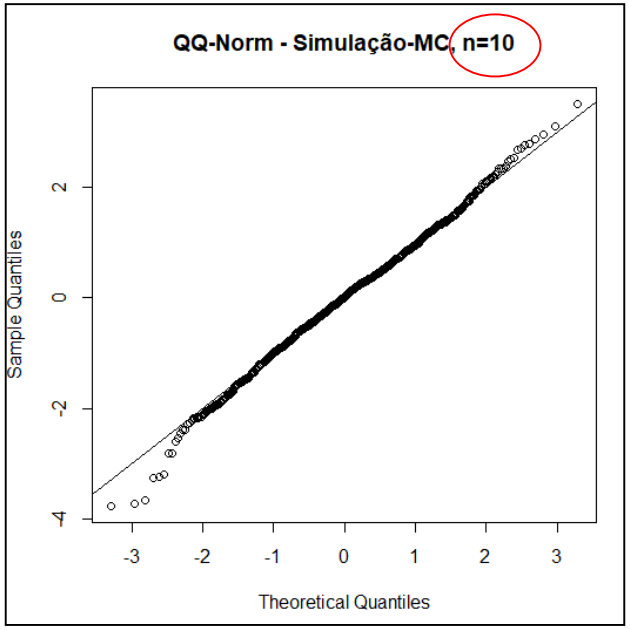
Simulação Monte Carlo

Validade do TLC:

Simulação MC de amostras das distribuições Normais ($n_1 = n_2 = n$):

- Adotar um valor para n **(n=10)**
- Obter as amostras Bootstrap (B=500)
- Cálcular a estatística *t* para cada amostra bootstrap
- Avaliar a distribuição dos valores obtidos

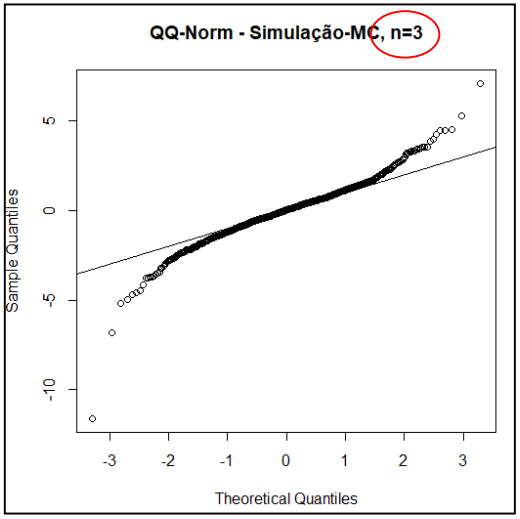


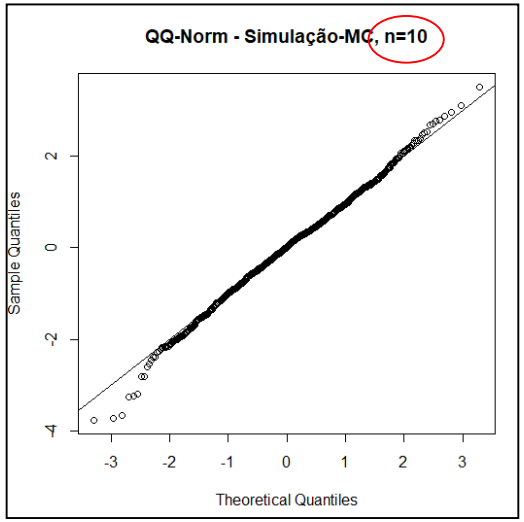


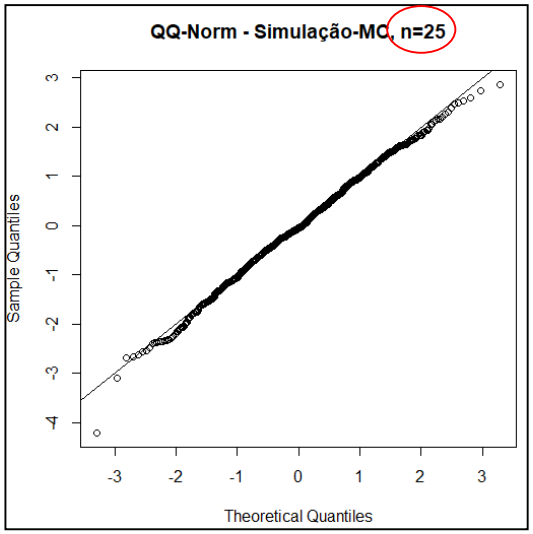
Simulação Monte Carlo

Validade do TLC:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{\substack{H_0 \\ n \to \infty}}{\sim} N(0;1)$$







Outras Motivações

$$\Rightarrow Y_i \sim N\left(\mu;\sigma^2\right) \rightarrow \sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{\sigma}\right)^2 \sim \chi^2_{n-1}$$

$$\Rightarrow t \sim t_{n-1} \rightarrow t^2 \sim F_{1,(n-1)}$$

Esses resultados podem ser mostrados via Simulação de Monte Carlo!

$$\Rightarrow U_1 \sim \chi^2_{n_1},\quad U_2 \sim \chi^2_{n_2} \rightarrow \frac{U_1/n_1}{U_2/n_2} \sim F_{n_1,n_2}$$

Distância de Mahalanobis:

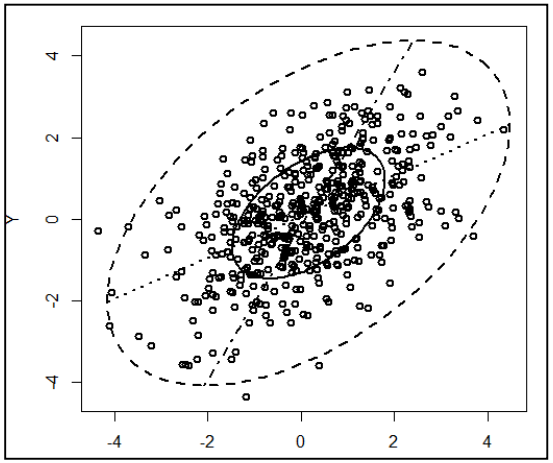
$$\Rightarrow Y_i \sim N_p\left(\mu_{p\times 1};\Sigma_{p\times p}\right) \rightarrow d_M^2 = \left(Y_i-\bar{Y}\right)S^{-1}\left(Y_i-\bar{Y}\right)' \sim \chi^2_p$$

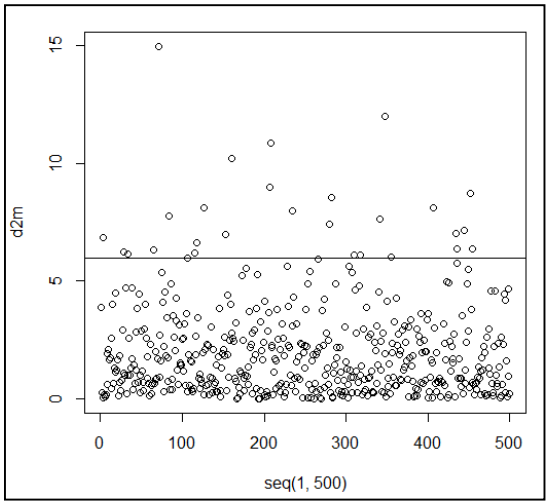
$$\rightarrow \frac{d_M^2}{p} \sim t_p^2 = F_{1,p}$$

Motivação

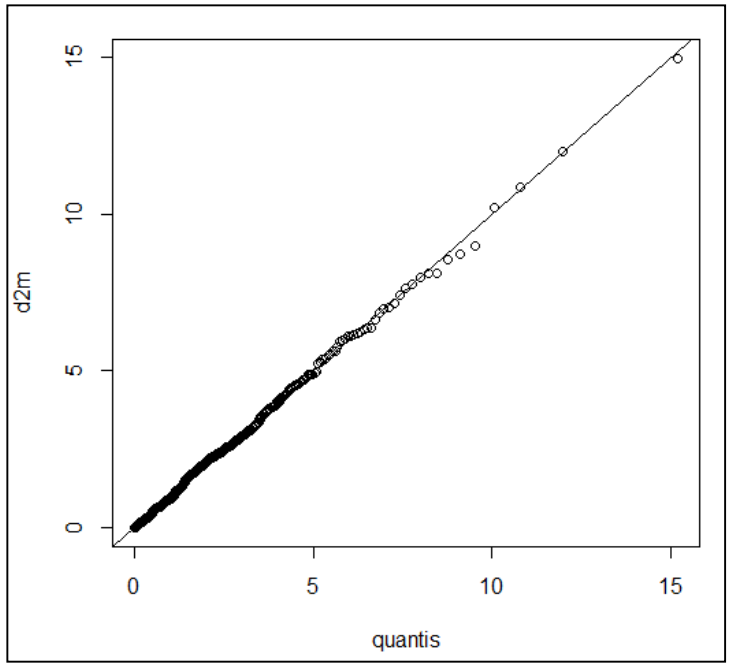
$$\Rightarrow Y_{i\ p\times 1} \sim N_p\left(\mu_{p\times 1}; \Sigma_{p\times p}\right)$$

$$d_M^2 = \left(Y_i - \bar{Y}\right)' S^{-1} \left(Y_i - \bar{Y}\right) \sim \chi_p^2$$






```
library(MASS)
mu<-c(0,0)
sigma<-matrix(c(2,1,1,2),ncol=2)
n<-500
y<-mvrnorm(n,mu,sigma)
mi<-colMeans(y)
s<-cov(y)
par(mfrow=c(1,2))
bivbox(y, method ="O") #solução robusta
# Copy Everitt's bivbox function
d2m<-mahalanobis(y,mi,s)
plot(seq(1,500),d2m)
abline(h=qchisq(0.95,df=2))
quantis <- qchisq(ppoints(length(y)),df=2)
qqplot(quantis, d2m)
abline(0,1)
```



Motivação

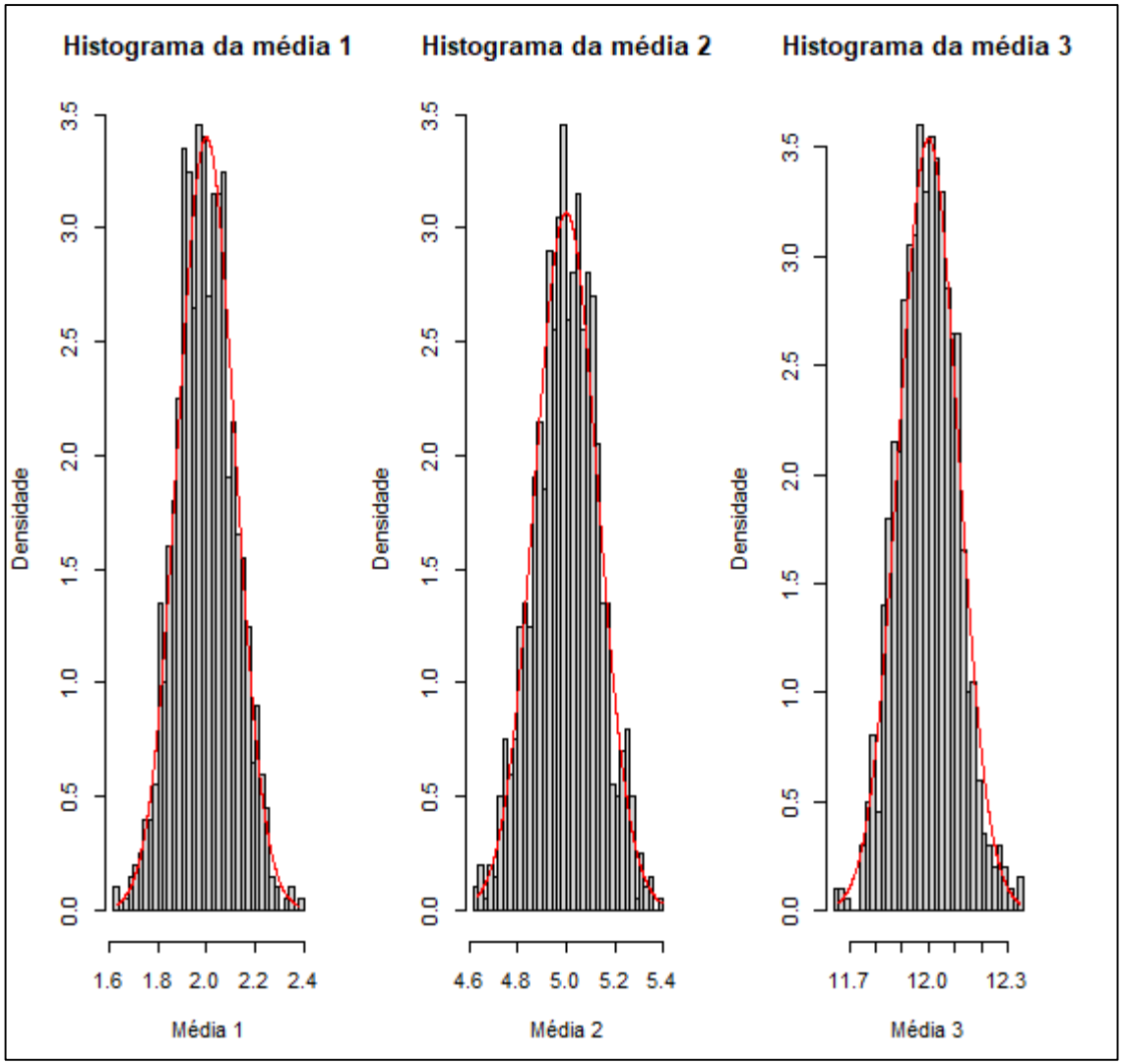
Normal Trivariada:

```
> media_pop #vetor centróide
[1]  2  5 12
> cov_pop #matriz de covariâncias Sigma populacional
     [,1] [,2] [,3]
[1,] 5.17 1.86 2.22
[2,] 1.86 4.57 2.27
[3,] 2.22 2.27 6.79
```

1. Gerar L=1000 amostras de tamanho n=500 da N3
2. De cada amostra estimar o vetor centróide
3. Mostrar que a distribuição da media amostral de cada variável é Normal univariada

$$\Rightarrow Y_{i\,3\times1} \sim N_3\left(\mu_{3\times1};\Sigma_{3\times3}\right)$$

$$\bar{Y}_j \sim N_1\left(\mu_j;\sigma_j^2/n\right)$$



Motivação

Normal Trivariada:

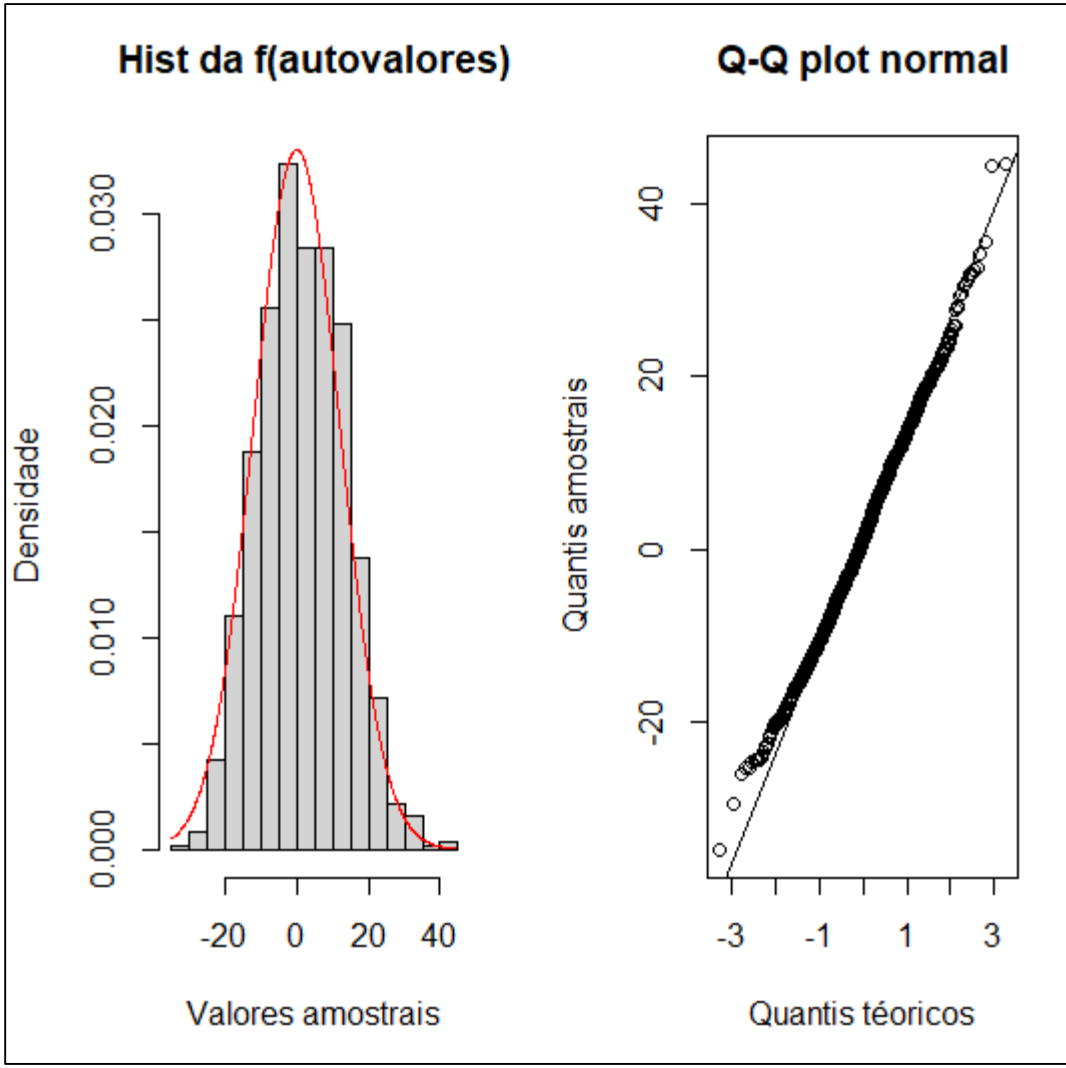
```
> media_pop #vetor centróide
[1]  2  5 12
> cov_pop #matriz de covariâncias Sigma populacional
     [,1] [,2] [,3]
[1,] 5.17 1.86 2.22
[2,] 1.86 4.57 2.27
[3,] 2.22 2.27 6.79

> eigen(cov_pop)
eigen() decomposition
$values
[1] 8.95074 8.49655 4.24271

$vectors
           [,1]       [,2]       [,3]
[1,] 0.6163374 -0.4493308  0.6467071
[2,] 0.4489783  0.8751869  0.1801841
[3,] 0.6469519 -0.1793033 -0.7411502
```

$$\Rightarrow Y_{i\,3\times1} \sim N_3\left(\mu_{3\times1}; \Sigma_{3\times3}\right)$$

$$\sqrt{n}\left(\hat{\lambda} - \lambda\right) \sim N_3\left(0; 2\Lambda^2\right)$$



Simulação Monte Carlo

International Journal of Mathematics and Computational Science
Vol. 4, No. 1, 2018, pp. 18-33
http://www.aiscience.org/journal/ijmcs
ISSN: 2381-7011 (Print); ISSN: 2381-702X (Online)

A Practical Guide for Creating Monte Carlo Simulation Studies Using R

Mohamed Reda Abonazel*

Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University, Cairo, Egypt

Abstract

This paper considers making Monte Carlo simulation studies using R language. Monte Carlo simulation techniques are very commonly used in many statistical and econometric studies by many researchers. So, we propose a new algorithm that provides researchers with basics and advanced skills about how to create their R-codes and then achieve their simulation studies. Our algorithm is a general and suitable for creating any simulation study in statistical and econometric models. Moreover, we provide some empirical examples in econometrics as applications on this algorithm.

Keywords

Autocorrelation Problem, Econometric Modeling, Graphical Presentation Methods, Ridge Estimation, Seemingly Unrelated Regressions Model

Simulação Monte Carlo

Existem problemas que são difíceis de resolver por métodos simples de MC.

Principais dificuldades:

- Obter amostras independentes: uso de métodos MCMC (Monte Carlo de Cadeias de Markov) em que, ao invés de amostrar pontos independentes, amostras são obtidas de Cadeias de Markov, cuja distribuição limite é a distribuição que se deseja amostrar.

- Garantir estimativas MC precisas: uso de métodos QMC (Quase-Monte Carlo) que aceleram a convergência para soluções “ótimas”

Métodos de Reamostragem

Efron and Tibshirani, 1998

⇒ **Reamostrar os dados** (amostra efetivamente observada) por MC pode ser útil para, por exemplo:

- Estimar a variância de um estimador
- Obter Intervalos de Confiança para um parâmetro de interesse

Exemplos:

- Qual é a distribuição dos autovalores de uma matriz? Qual é o erro padrão dos autovalores?
- Qual é a distribuição do “escore de risco poligênico”?
- Qual é a distribuição do “escore de risco cardiovascular” de Framingham?

Métodos de Reamostragem

Efron and Tibshirani, 1998

O **método de reamostragem** não requer assumir uma **DISTRIBUIÇÃO AMOSTRAL** para uma estatística de interesse.
Exemplo: a estatística t, sob premissas clássicas, tem distribuição t com n-1 graus de liberdade.
O **método de reamostragem** permite construir uma **DISTRIBUIÇÃO EMPÍRICA** da estatística de interesse, com base em centenas ou milhares de amostras extraídas **com reposição** dos dados.

Assim, sob esse procedimento, não há violação de pressuposições! MAS é preciso garantir que os dados são uma amostra "aleatória" e "representativa" da população sob estudo e "L é grande".

Principais Métodos de Reamostragem (dos dados observados):

Bootstrap: extração das amostras COM reposição

Jackknife: extração das amostras SEM reposição

Métodos de Reamostragem

Reamostragem Jackknife (precurssor do Bottstrap): obter estimativas do erro padrão de estatísticas

$y = (y_1, y_2, ..., y_n)$: amostra aleatória efetivamente observada (observações *iid*)

$y_{-i} = (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$: i-ésima amostra Jack (sem a observação i) i=1,2,…,n

1. Obter n amostras Jack
2. Para cada amostra calcular o valor do estimador de interesse $\hat{\theta}_j, j = 1, ..., n$
3. Calcular a média das estimativas para as replicatas Jack $\bar{\theta}$
4. A estimativa Jackknife para o Erro Quadrático Médio do estimador é:

$$\frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_j - \bar{\theta}\right)^2; \quad \bar{\theta} = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_j$$

Métodos de Reamostragem

Reamostragem Jackknife (precurssor do Bottstrap):

```
> x <- c(10,27,31,40,46,50,52,104,146)

> mean(x)   # 56.22222
> var(x)    # 1799.194
> var(x)/length(x) #erro padrão ao quadrado: variância da média amostral
[1] 199.9105

> #Estimando a média
> jack <- numeric(length(x))
> for (i in 1:length(x)){ jack[i] <- mean(x[-i]) }
> jack
[1] 62.000 59.875 59.375 58.250 57.500 57.000 56.750 50.250 45.000
> mean(jack) #comparar com mean(x)
[1] 56.22222
> ((length(x)-1)/length(x))*sum((jack-mean(jack))^ 2)
[1] 199.9105 #comparar com a variância da media amostral
```

Métodos de Reamostragem

Reamostragem Bottstrap:

$y = (y_1, y_2, ..., y_n)$: amostra aleatória efetivamente observada (*iid*)

$y_b = (y_{(1)}, y_{(2)}, ..., y_{(n)})$: i-ésima amostra Bottstrap (i=1,2,…,B) extraída aleatoriamente e com reposição de y

1. Obter B amostras Bottstrap
2. Para cada amostra calcular o valor do estimador de interesse $\hat{\theta}_b, b = 1, ..., B$
3. Calcular a média das estimativas para as replicatas Bootstrap, $\bar{\theta}$
4. A estimativa Bottstrap para o Erro Quadrático Médio do estimador é (no caso de estimar a média):

$$\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2; \quad \bar{\theta} = \frac{1}{n}\sum_{b=1}^{B}\hat{\theta}_b$$

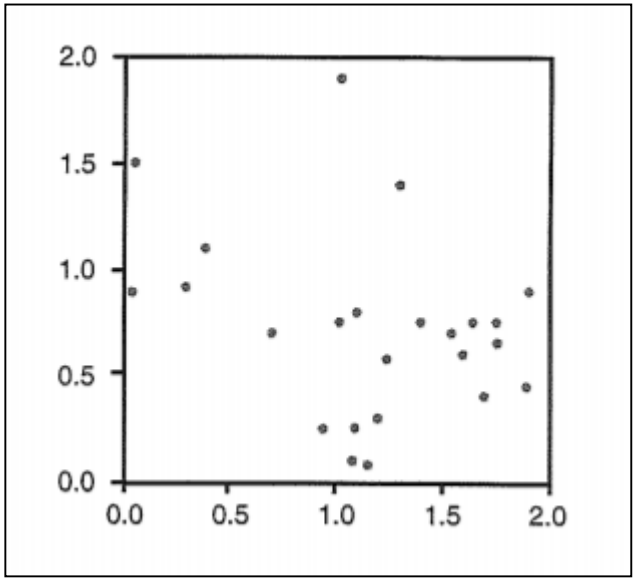
Reamostragem Bootstrap

```
> x <- c(10,27,31,40,46,50,52,104,146)
> mean(x)  # 56.22222
> var(x)/length(x) #erro padrão ao quadrado: variância da média
[1] 1799.194/9 = 199.9105

> #Estimando a verdadeira média da população em que x amostrou
> fc <- function(d){mean(d)}
> boot <-numeric(1000)
> for (i in 1:1000) boot[i] <- fc(sample(x,replace=T))
> head(boot)
[1] 67.55556 38.66667 50.44444 51.22222 84.00000 56.33333
> mean(boot)
[1] 57.33933
> var(boot) #erro quadrático médio do estimador bootstrap
[1] 171.0338
> quantile(boot,0.025) #Limite de confiança de Efron
   2.5% 35.65278
> quantile(boot,0.975)
  97.5% 83.78333
> bias <- mean(boot) - mean(x)
> bias   #1.117111
> mean(boot) - bias - 1.96*sqrt(var(boot))  #IC Bootstrap corrigido
[1] 30.58937
> mean(boot) - bias + 1.96*sqrt(var(boot))
[1] 81.85508
```

Bootstrap – Distribuições Aleatórias

Posições de 24 plantas em um quadrante (Manly, 2007)



Questão: As plantas estão posicionadas aleatoriamente ou há um padrão causal?

Uso de Estatísticas do Vizinho-mais-próximo

g1: distância da planta ao primeiro vizinho-mais próximo
g2: distância da planta ao segundo vizinho-mais próximo
Assim por diante, até
g10: distância da planta ao décimo vizinho-mais próximo

Amostra bootstrap: amostrar 24 pontos com reposição dentre os pontos observados

Médias observadas e Médias de **amostras bootstrap** (B=999) para os vizinhos mais próximos (de 1 a 10). Indicação da significância (bootstrap) do valor observado.

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
Observed	0.217	0.293	0.353	0.419	0.500	0.559	0.606	0.646	0.698	0.739
Mean	0.227	0.354	0.455	0.544	0.625	0.708	0.780	0.846	0.909	0.968
Significance	0.729	0.080	0.009	0.003	0.009	0.005	0.008	0.004	0.004	0.002

Teste χ^2 de Aderência: g1 e g2 não rejeitar H0 $\Rightarrow$ g1 e g2 mostram padrão aleatório!

Bootstrap - Bioequivalência

Dados da resposta de 8 pacientes aos tratamentos Placebo, Medicamento Antigo e Medicamento Novo

Placebo	Med_Antigo	Med_Novo
9243	17649	16449
9671	12013	14614
11792	19979	17274
13357	21816	23798
9055	13850	12560
6290	9806	10157
12412	17208	16570
18806	29044	26325

Questão: O Medicamento Novo é equivalente ao Medicamento Antigo?

$$ANVISA: \quad \theta = \frac{\left|\mu_{NOVO} - \mu_{ANTIGO}\right|}{\mu_{ANTIGO} - \mu_{PLACEBO}} \leq 0,20$$

Critério para declarer que o Medicamento Novo é equivalência ao Medicamento Antigo.

Como testar a Bioequivalência?

Bootstrap - Bioequivalência

Resposta de 8 pacientes aos tratamentos Placebo, Medicamento Antigo e Novo

Placebo	Med_Antigo	Med_Novo	Antigo-Placebo	Novo-Antigo
9243	17649	16449	8406	-1200
9671	12013	14614	2342	2601
11792	19979	17274	8187	-2705
13357	21816	23798	8459	1982
9055	13850	12560	4795	-1290
6290	9806	10157	3516	351
12412	17208	16570	4796	-638
18806	29044	26325	10238	-2719

↑ ↑

$$ANVISA: \quad \theta = \frac{|\mu_{NOVO} - \mu_{ANTIGO}|}{\mu_{ANTIGO} - \mu_{PLACEBO}} \leq 0,20 \qquad \hat{\theta} = 0,0713$$

Conclusão? Há evidência para declarar bioequivalência?

Testar a bioequivalência via amostras Bootstrap dos dados observados **(Como?)**
Construir IC Booststrap, construer um Teste de Aleatorização.

Amostra booststrap: como os dados são pareados, de cada linha, amostrar com reposição, dentre os 3 valores, as respostas ao Placebo e Medicamentos Antigo e Novo

Bootstrap – Componentes Principais

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis*

(Manly, 2007)

	Yellow				Pink				Brown	
Colony	UB	MB	FB	OB	UB	MB	FB	OB	UB	B
1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
3	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
4	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
5	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
6	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
7	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
8	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
9	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
10	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
11	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
12	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
13	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
14	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
15	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
16	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
17	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

Matriz de Dados:

$$Y_{17\times10}$$

Os dados são composicionais: cada linha soma 100%

$\Rightarrow$obrigatoriamente a matriz de correlação tem valores negativos!

- Transformar os dados: Aitchison (1986) ou
- Eliminar uma das colunas

Bootstrap – Componentes Principais

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis*

(Manly, 2007)

	Yellow				Pink				Brown	
Colony	UB	MB	FB	OB	UB	MB	FB	OB	UB	B
1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
3	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
4	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
5	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
6	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
7	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
8	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
9	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
10	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
11	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
12	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
13	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
14	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
15	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
16	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
17	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

Matriz de Dados (eliminando a última coluna):

$$Y_{17\times 9}$$

Como avaliar propriedades dos CPs? Por ex., quantos autovalores são maiores que a média deles (critério de seleção do número de CP)?

Proponha uma solução Bootstrap! **Como construir as amostras Bootstrap?**

Bootstrap – Componentes Principais

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis*

(Manly, 2007)

	Yellow				Pink				Brown	
Colony	UB	MB	FB	OB	UB	MB	FB	OB	UB	B
1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
3	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
4	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
5	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
6	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
7	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
8	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
9	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
10	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
11	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
12	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
13	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
14	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
15	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
16	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
17	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

Matriz de Dados:

$$Y_{17\times 9}$$

Amostras Bootstrap:

Como temos dados multivariados, amostrar com reposição n=17 vetores (p=9) da amostra. Proceder assim B vezes.

Bootstrap – Correlação Canônica

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis*

(Manly, 2007)

	Yellow				Pink				Brown	
Colony	UB	MB	FB	OB	UB	MB	FB	OB	UB	B
1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
3	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
4	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
5	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
6	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
7	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
8	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
9	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
10	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
11	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
12	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
13	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
14	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
15	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
16	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
17	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

Matriz de Dados com dois conjuntos de variáveis (Yellow e Pink):

$$Y_{17\times 8}$$

p+q=4+4=8

Obtenha um Intervalo de Confiança Bootstrap para o coeficiente de correlação Canônico!

Como construir as amostras Bootstrap?

Bootstrap – Correlação Canônica

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis*

(Manly, 2007)

	Yellow				Pink				Brown	
Colony	UB	MB	FB	OB	UB	MB	FB	OB	UB	B
1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
3	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
4	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
5	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
6	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
7	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
8	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
9	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
10	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
11	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
12	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
13	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
14	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
15	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
16	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
17	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

Matriz de Dados com dois conjuntos de variáveis (Yellow e Pink):

$$Y_{17\times 8}$$

Amostras Bootstrap:

Fixar um dos conjuntos de variáveis, digamos Pink, e amostrar com reposição 17 vetores da amostra de Yellow (p=4) pareando com as observações de Pink. Prodecer assim B vezes.

Bootstrap – Análise Discriminante

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis* in Colonies from Six Habitat Types in Southern England

(Manly, 2007)

Habitat	Colony	Yellow				Pink				Brown	
		UB	MB	FB	OB	UB	MB	FB	OB	UB	B
Downland beech	1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
	2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
Oakwood	1	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
	2	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
	3	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
Mixed deciduous woods	1	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
	2	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
	3	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
	4	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
	5	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
Hedgerows	1	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
	2	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
	3	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
	4	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
Downside long coarse grass	1	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
	2	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
Downside short turf	1	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

Matriz de Dados Agrupados (G=5) eliminando a última coluna:

$$Y_{17\times 9}$$

Eliminar esse grupo

17=2+3+5+4+2+1

A função discriminante é obtida da decomposição spectral de $W^{-1}B$, sendo W e B as matrizes de SQPC dentro e entre grupos.

$$l_k' Y_{17\times 9} = l_{k1} Y_1 + \ldots + l_{k9} Y_9$$

↑ Pesos=autovetores

Bootstrap – Análise Discriminante

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis* in Colonies from Six Habitat Types in Southern England

(Manly, 2007)

Habitat	Colony	Yellow				Pink				Brown	
		UB	MB	FB	OB	UB	MB	FB	OB	UB	B
Downland beech	1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
	2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
Oakwood	1	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
	2	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
	3	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
Mixed deciduous woods	1	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
	2	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
	3	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
	4	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
	5	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
Hedgerows	1	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
	2	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
	3	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
	4	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
Downside long coarse grass	1	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
	2	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
Downside short turf	1	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

Matriz de Dados Agrupados (G=5) eliminando a última coluna: $Y_{17\times 9}$

Como avaliar o poder de discriminação das funções discriminantes?

Existem até k=min(n,p,G-1) funções.

Verificar, por um teste de aderência, para cada k, se a formação dos grupos é aleatória. O primeiro k que rejeitar H0 tem poder de discriminação significante:

$$\left(n-1-\frac{p+G}{2}\right)\ln\left(1+\lambda_k\right)\sim\chi^2_{p+G-2k}$$

Como construir um teste de aleatorização neste caso?

Bootstrap – Análise Discriminante

Percentages of Shells with Different Colors and Banding for Samples of *C. nemoralis* in Colonies from Six Habitat Types in Southern England

(Manly, 2007)

Habitat	Colony	Yellow				Pink				Brown	
		UB	MB	FB	OB	UB	MB	FB	OB	UB	B
Downland beech	1	9.6	15.4	0.0	0.0	48.7	25.0	0.0	0.0	0.6	0.6
	2	10.6	16.0	0.0	0.0	26.3	4.5	0.0	0.0	36.5	5.8
Oakwood	1	1.2	4.7	1.2	1.2	5.8	25.6	4.7	1.2	34.9	19.8
	2	0.0	13.0	0.5	0.0	27.6	43.2	3.8	8.6	0.0	3.2
	3	1.5	1.0	0.7	0.5	23.2	28.4	10.7	23.8	10.2	0.0
Mixed deciduous woods	1	3.0	1.5	4.5	0.0	50.0	6.1	16.7	6.1	12.1	0.0
	2	0.4	3.1	14.8	12.6	6.7	6.7	24.2	17.9	10.8	2.7
	3	0.0	0.0	11.4	5.7	17.1	14.3	25.7	14.3	11.4	0.0
	4	1.0	4.0	18.2	14.1	17.2	3.0	8.1	10.1	23.2	1.0
	5	13.5	0.0	5.8	7.7	13.5	0.0	23.1	21.2	15.4	0.0
Hedgerows	1	9.5	4.8	47.6	14.3	0.0	0.0	9.5	9.5	0.0	4.8
	2	6.1	9.0	16.5	9.3	2.5	9.7	21.1	21.9	3.2	0.7
	3	16.1	6.5	16.1	19.4	12.9	3.2	16.1	9.7	0.0	0.0
	4	1.3	1.3	13.2	2.6	0.0	1.3	55.3	25.0	0.0	0.0
Downside long coarse grass	1	0.0	15.5	43.3	5.4	0.0	7.5	19.2	2.6	1.9	4.7
	2	0.0	7.1	13.5	21.9	12.3	3.9	21.3	20.0	0.0	0.0
Downside short turf	1	28.7	12.8	0.0	3.7	3.0	7.9	0.0	4.3	29.9	9.8

Note: UB = unbanded; MB = mid-banded; FB = fully banded; OB = other banding types; B = banded.

$$\left(n-1-\frac{p+G}{2}\right)\ln\left(1+\lambda_k\right) \sim \chi^2_{p+G-2k}$$

Como construir um teste de aleatorização neste caso?
Não usar a distribuição Qui-quadrado.

Amostrar com reposição, dentre os 17 vetores de respostas, 2 vetores para compor G1, 3 vetores para G2 e assim por diante até 1 para G=6. Para cada amostra formada calcular a estatística. Construir a distribuição de referência a avaliar nela a significância do resultado observado.