Testes de Permutação para Comparação de Distribuições Multivariadas

André F. B. Menezes

Universidade Estadual de Maringá

Departamento de Estatística

3 de Outubro de 2017

Organização

Teste de Permutação

Testes de Hipóteses Clássicos

- Desenvolvido por Jerzy Neyman e Egon Pearson.
- Encontrar uma estatística de teste cuja distribuição não dependa do parâmetro de interesse.
- Muitas vezes tem-se uma distribuição assintótica.
- Várias suposições são impostas sobre os dados.

Teste de Permutação

Os dados devem ser provenientes de ao menos duas populações.

Conceito

1. A estatística de teste é calculada para os dados observados;
2. As observações são permutadas sobre todas as possíveis disposições dos dados;
3. A estatística de teste selecionada é calculada;
4. A proporção de permutações com valores da estatística de teste iguais ou mais extremos do que a estatística de teste observada fornece a probabilidade exata de observar o valor da estatística de teste calculada.

Teste de Permutação

Vantagens sobre os Teste de Hipóteses Clássicos

- Podem ser utilizados em amostras não aleatórias.
- São completamente dependente dos dados.
  - Implicitamente, esta entendido que a inferência estatística esta limitada ao experimento ou pesquisa realizada.
- Não fazem suposições a respeito de distribuições.
- Fornecem valores exatos de probabilidade.

Teste de Permutação

Testes de Permutação Exatos

- Enumera todas as combinações igualmente prováveis dos dados observados.
- Suponha que:

$$X_1, \ldots, X_n \sim F_X \quad \text{e} \quad Y_1, \ldots, Y_m \sim F_Y;$$
$$\mathcal{H}_0 : F_X = F_Y \quad \text{vs.} \quad \mathcal{H}_1 : F_X \neq F_Y$$

- Seja Z o conjunto ordenado $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ indexado pelos índices $\nu = \{1, \ldots, n, n+1, \ldots, n+m\} = \{1, \ldots, n+m\}$.
- Então existem $\binom{n+m}{n}$ diferentes formas de particionar o conjunto Z em dois subconjuntos de tamanhos n e m.

Teste de Permutação

Testes de Permutação Aproximados por Reamostragem

Conforme Rizzo (2008) para o obter um teste de permutação aproximado por reamostragem procedemos:

1. Calcule a estatística observada $\widehat{\theta}$;
2. Para cada replica, indexada $j = 1, \ldots, M$:
   (1) Gere uma permutação aleatória;
   (2) Calcule a estatística para a permutação obtida;
3. Se valores grandes de $\widehat{\theta}$ são favoráveis a hipótese alternativa, calcule o valor-*p* empírico por:

$$\widehat{p} = \frac{1 + \sum_{j=1}^{M} I\left(\widehat{\theta}^{(j)} \geq \widehat{\theta}\right)}{M + 1}.$$

4. Rejeita-se $\mathcal{H}_0$ em nível de significância α se $\widehat{p} \leq \alpha$.

Testes Multivariados para Igualdade de Distribuições

Teste T^2 de Hotelling

- Utilizado para avaliar se existe diferença entre dois grupos quando $p \geq 2$ variáveis são mensuradas.
- Exige que as duas amostras sejam independentes e provenientes de uma distribuição Normal multivariada com matriz de covariância iguais (RENCHER, 2002).
- A estatística de teste é definida por:

$$T^2 = n\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)^\top \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pl}\right]^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)$$

em que

$$\mathbf{S}_{pl} = \frac{1}{n_1 + n_2 - 2}\left[\left(n_1 - 1\right)\mathbf{S}_1 + \left(n_2 - 1\right)\boldsymbol{S}_2\right],$$

Testes Multivariados para Igualdade de Distribuições

Teste baseado no k-NN

- Proposto por Schilling (1986) e Henze (1988) é um teste não paramétrico para comparar se duas ou mais distribuições multivariadas são iguais.
- Para formalizar sua teoria vamos supor que

$$\mathbf{X} = \{X_1, \ldots, X_{n_1}\} \in \mathbb{R}^p, \qquad \mathbf{Y} = \{Y_1, \ldots, Y_{n_2}\} \in \mathbb{R}^p$$

são amostras independentes e aleatórias, $p \geq 2$.

Testes Multivariados para Igualdade de Distribuições

Teste baseado no k-NN

- A matriz de dados agrupados $\mathbf{Z}$, uma matriz $n \times p$ com observações nas linhas

$$\mathbf{Z}_{n\times p} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_1,1} & x_{n_1,2} & \cdots & x_{n_1,p} \\ y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ y_{n_2,1} & y_{n_1,2} & \cdots & y_{n_1,p} \end{bmatrix}$$

em que $n = n_1 + n_2$.

Testes Multivariados para Igualdade de Distribuições

Teste baseado no k-NN

- Seja $NN_k(Z_i)$ o k-ésimo vizinho mais próximo de Z_i.

$$I_i(k) = \begin{cases} 1, & \text{se } NN_k(Z_i) \text{ pertence a mesma amostra de } Z_i. \\ 0, & \text{caso contrário.} \end{cases}$$

A estatística de teste é definida como

$$T_{k,n} = \frac{1}{n\,k} \sum_{i=1}^{n} \sum_{r=1}^{k} I_i(r),$$

- ou seja, a proporção de todas as k comparações do vizinhos mais próximo em que uma observação e seu vizinho são membros da mesma amostra.

Aplicações

Aplicação I

- Dados retirados de Manly (2004).
- Consistem de cinco variáveis relacionadas a medidas do corpo dos 49 pardais moribundos e mais uma variável indicando se o pardal morreu ou não.
- Objetivo é verificar se existe diferença nas medidas corporais entre os pardais vivos e mortos.

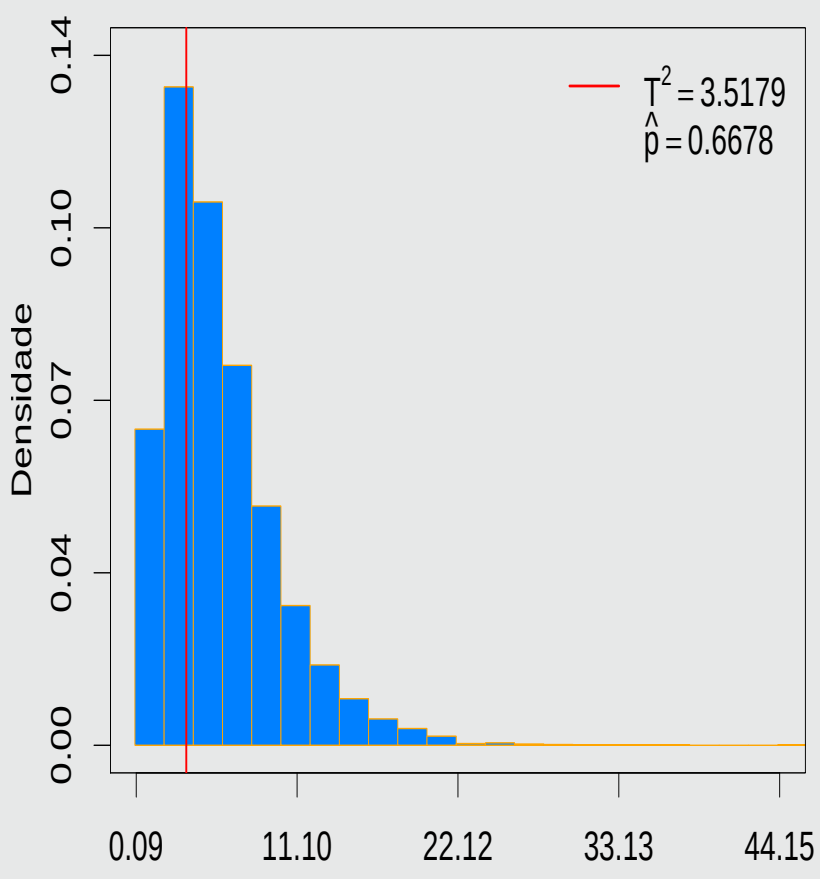
Aplicações

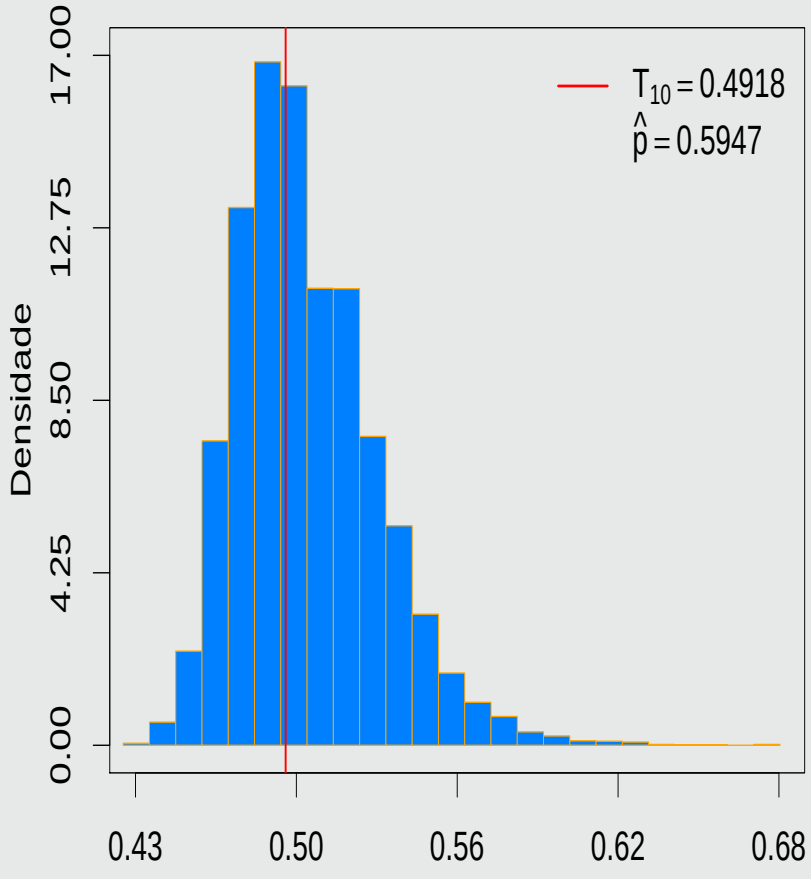
Aplicação I

ct	ea	cbc	cdu	cqe	sobrev
156	245	31.60	18.50	20.50	S
154	240	30.40	17.90	19.60	S
153	240	31.00	18.40	20.60	S
153	236	30.90	17.70	20.20	S
155	243	31.50	18.60	20.30	S
⋮	⋮	⋮	⋮	⋮	⋮
155	235	30.70	17.70	19.60	N
162	247	31.90	19.10	20.40	N
153	237	30.60	18.60	20.40	N
162	245	32.50	18.50	21.10	N
164	248	32.30	18.80	20.90	N

Aplicações

Aplicação I



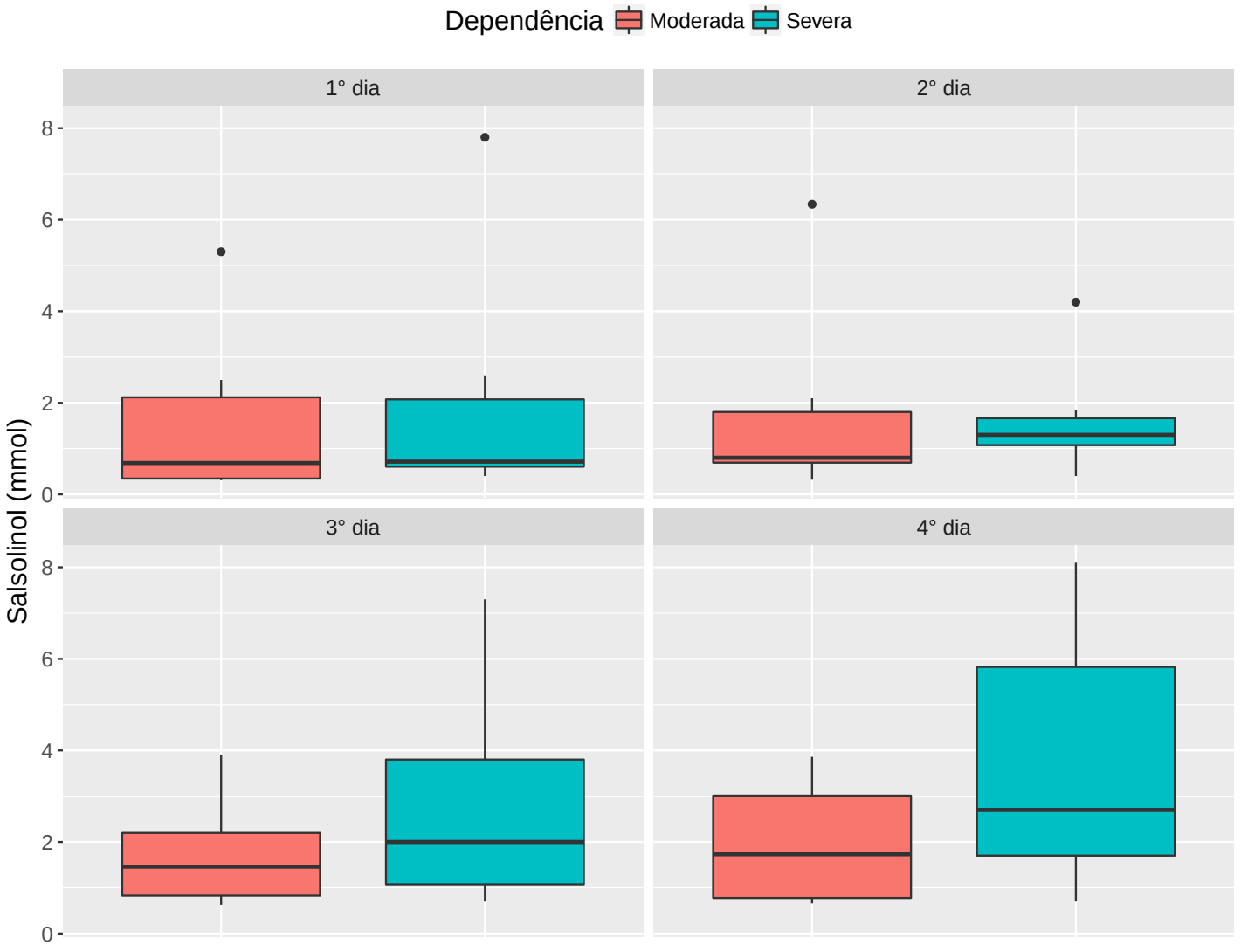


Aplicações

Aplicação II

- Dados retirados de Hand (1987).
- Referem-se ao comportamento de um agente bioquímico denominado salsolinol em 14 pacientes com dependência química.
- O salsolinol foi medido nos 14 paciente a partir de uma análise bioquímica da urina durante quatro dias.
- Dos 14 indivíduos observados nos quatro dias, 8 foram considerados “severamente” dependentes e 6 foram categorizados como sendo “moderadamente” dependentes.
- Objetivo é verificar se existe diferença estatisticamente significativa entre os o grupo de paciente com dependência severa e moderada.

Aplicações


Dependência
Moderada
Severa
1° dia
2° dia
3° dia
4° dia
Salsolinol (mmol)
0
2
4
6
8

Aplicações

Aplicação II

- Dados são medidas repetidas do mesmo sujeito ao longo de 4 dias.
- Abordagem I: tratar os dados em formato multivariado.
- Abordagem II: considerar formato univariado.

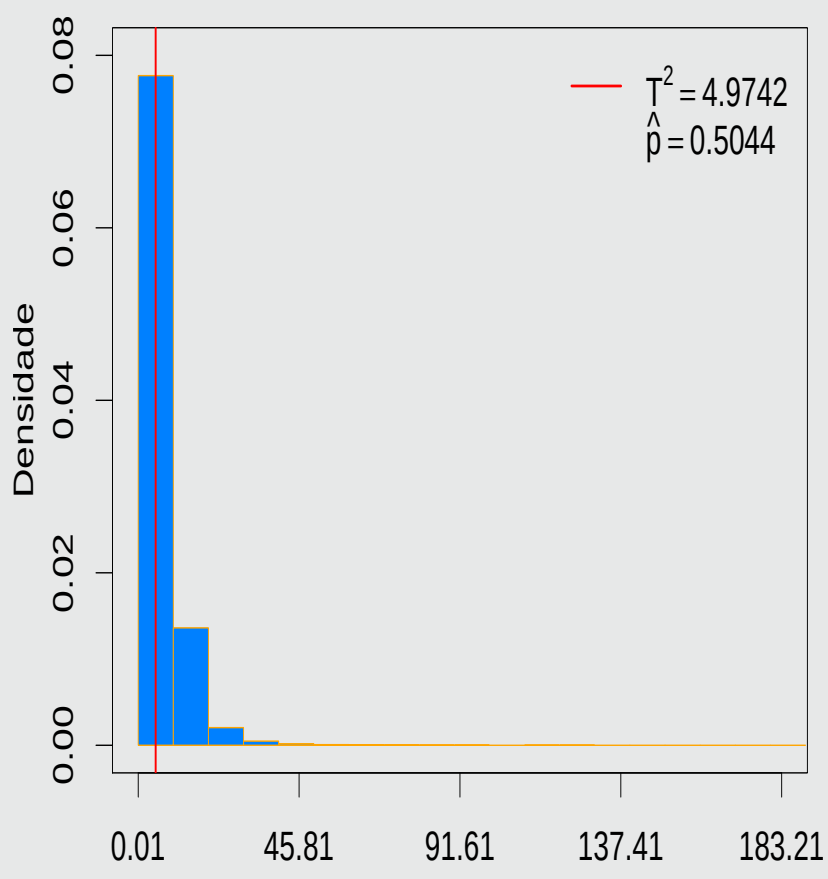
Aplicação II

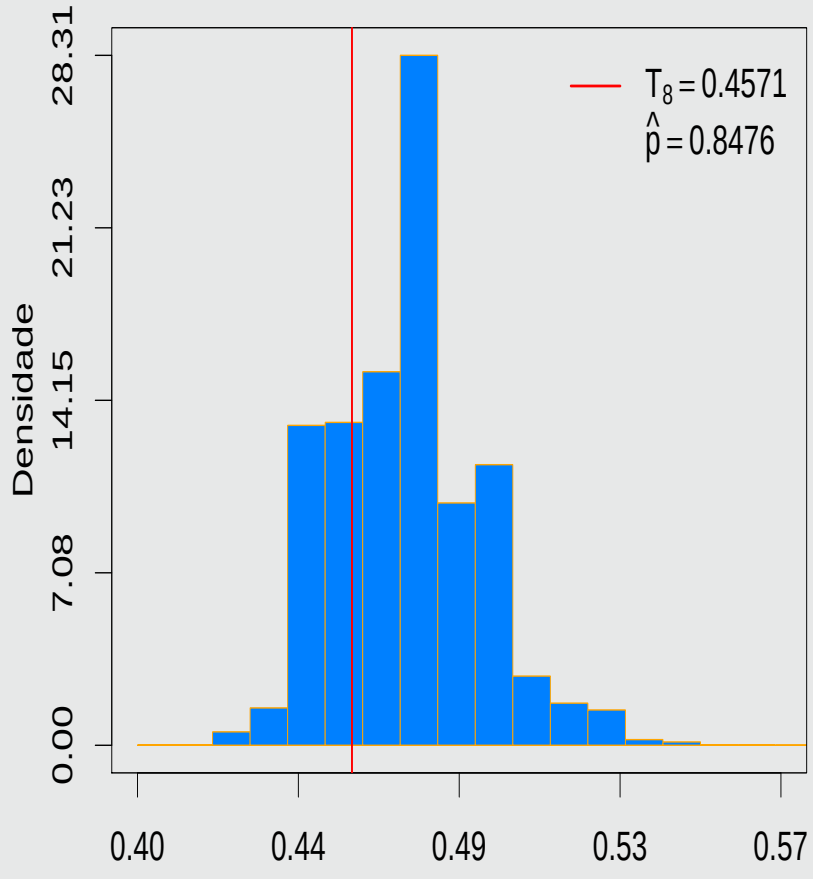
Abordagem I

grupo	D1	D2	D3	D4
2	0.64	0.70	1.00	1.40
1	0.33	0.70	2.33	3.20
2	0.73	1.85	3.60	2.60
2	0.70	4.20	7.30	5.40
2	0.40	1.60	1.40	7.10
2	2.60	1.30	0.70	0.70
2	7.80	1.20	2.60	1.80
1	5.30	0.90	1.80	0.70
1	2.50	2.10	1.12	1.01
2	1.90	1.30	4.40	2.80
1	0.98	0.32	3.91	0.66
1	0.39	0.69	0.73	2.45
1	0.31	6.34	0.63	3.86
2	0.50	0.40	1.10	8.10

Aplicação II

Abordagem I





Aplicação II

Abordagem II

ind	grupo	dia	salsolinol
1	2	1	0.64
2	1	1	0.33
3	2	1	0.73
4	2	1	0.70
5	2	1	0.40
6	2	1	2.60
7	2	1	7.80
⋮	⋮	⋮	⋮
8	1	4	0.70
9	1	4	1.01
10	2	4	2.80
11	1	4	0.66
12	1	4	2.45
13	1	4	3.86
14	2	4	8.10

Aplicação II

Abordagem II

Seja Y_{ijk} o valor de salsolinol do k-ésimo paciente do i-ésimo grupo no j-ésimo dia, então o modelo proposto é escrito na forma

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\,\beta)_{ij} + \epsilon_{ijk}$$

com $i = 1, 2$, $j = 1, 2, 3, 4$ e $k = 1, 2, \ldots, 14$. Sendo:

- μ: média global;
- α_i: efeito do i-ésimo grupo;
- β_j: efeito do j-ésimo dia;
- ϵ_{ijk}: o erro aleatório.

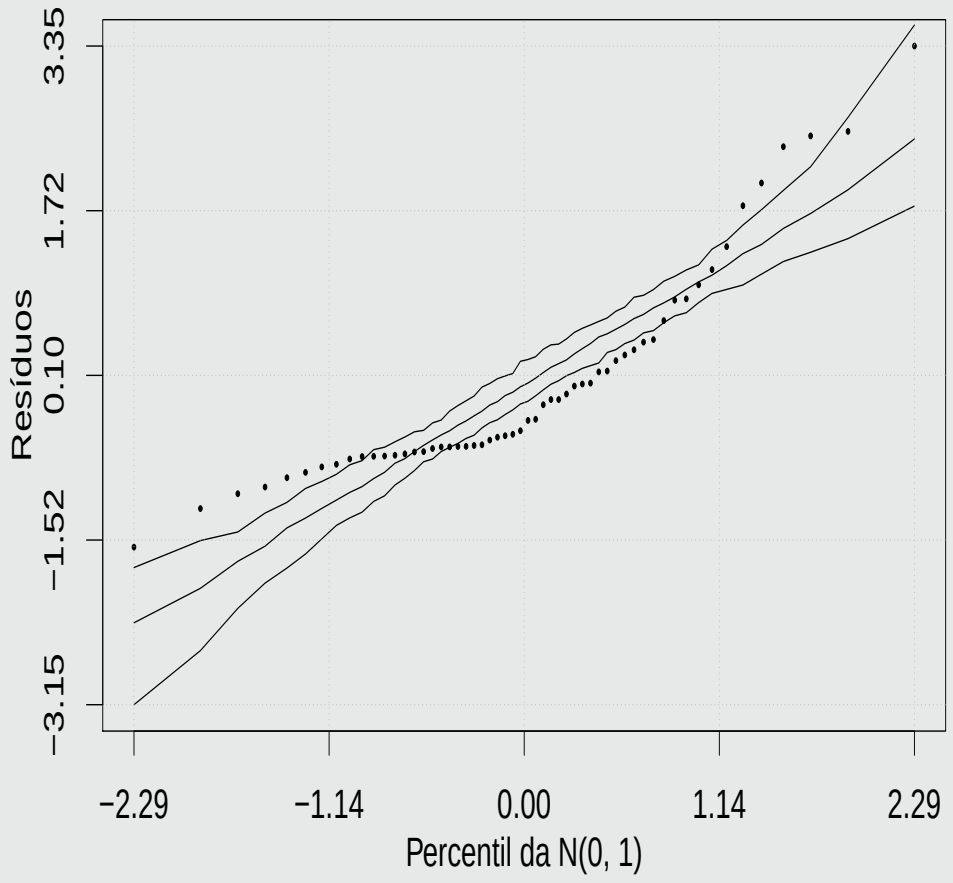
Aplicação II

Abordagem II – Análise de Variância

Fontes de variação	GL	SQ	QM	F_0	Pr(F_0 >F)	Pr(Exata)
dia	3	14.8877	4.9626	1.1562	0.3362	0.3397
grupo	1	6.5649	6.5649	1.5295	0.2222	0.2256
dia*grupo	3	8.0293	2.6764	0.6236	0.6032	0.6117
Resíduos	48	206.0214	4.2921			

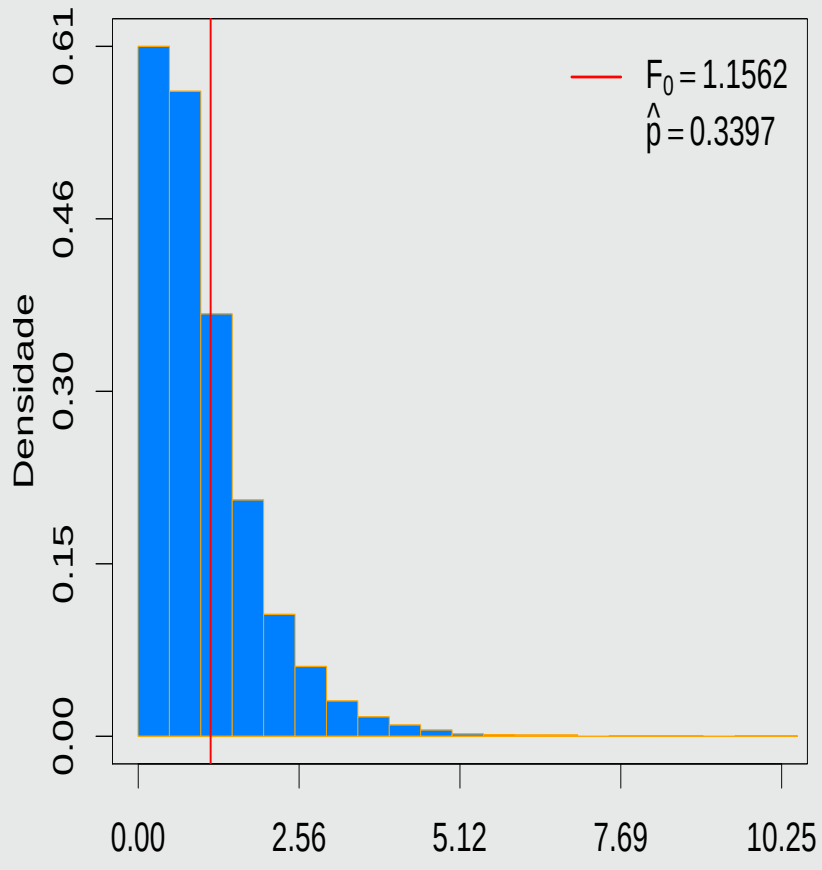
Aplicação II

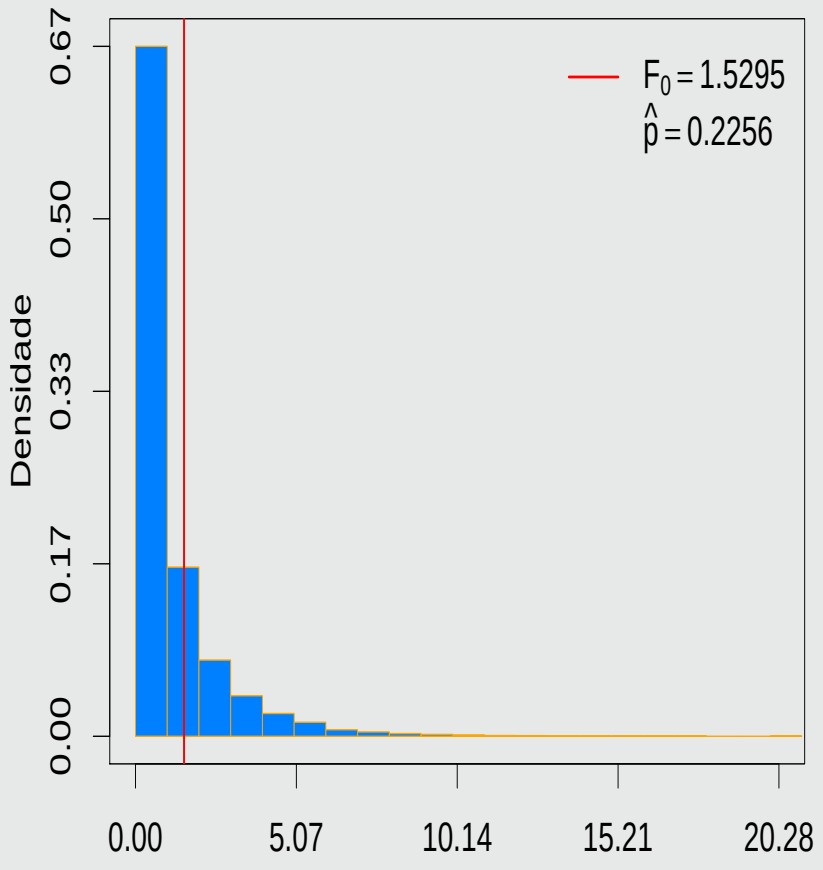
Abordagem II – Normalidade dos Resíduos



Aplicação II

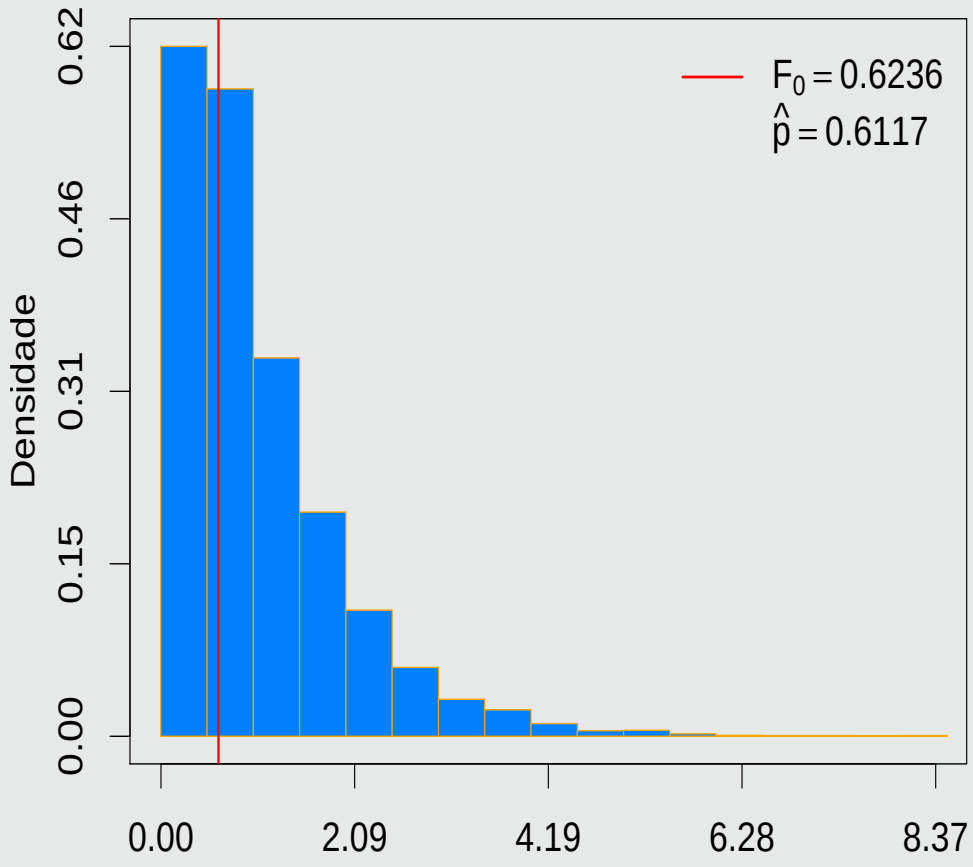
Abordagem II





Aplicação II

Abordagem II



Referências

[1] Berry, K. J., Mielke Jr, P. W., Johnston, J. E., 2016. **Permutation Statistical Methods An Integrated Approach.** Springer.

[2] Hand, D., Taylor, C. C., 1987. **Multivariate Analysis of Variance and Repeated Measures A practical approach for behavioural scientists.** Chapman & Hall/CRC.

[3] Henze, N., 1988. **A multivariate two-sample test based on the number of nearest neighbor type coincidences.** Ann. Statist. 16 (2), 772–783.

Referências

[4] Manly, B. F. J., 2004. **Multivariate Statistical Methods**, Third Edition. Chapman & Hall/CRC.

[5] Rizzo, M. L., 2008. **Statistical Computing with R.** Chapman & Hall/CRC.

[6] Schilling, M. F., 1986. **Multivariate two-sample tests based on nearest neighbors.** Journal of the American Statistical Association 81 (395), 799–806.

[6] Rencher, A. C., 2002. **Methods of Multivariate Analysis**, 2nd Edition. John Wiley & Sons, Inc.