Testes de Permutação para Comparação de Distribuições Multivariadas

André Felipe Berdusco Menezes

Departamento de Estatística, Universidade Estadual de Maringá

Maringá, PR, Brazil

1 Testes de Permutação

O uso de métodos de permutação para inferências exatas teve orgiem com Fisher em 1925 (Fisher, 1925). Os métodos estatísticos de permutação foram inicialmente desenvolvidos para validar as suposições de normalidade e homogeneidade dos métodos da estatística clássica. Desde então, a praticidade de tais métodos aumentou de forma constante com o poder da computação.

Para definir de forma geral os métodos de permutação vamos discutir dois tipos de modelos para realizar inferência estatística. Conforme Berry et al. (2016) os dois modelos de inferência estatística são: o modelo de população e o modelo de permutação. Os modelos de população, formalmente propostos por Jerzy Neyman e Egon Pearson (Neyman and Pearson, 1928a,b), assumem que a amostra aleatória é retirada de uma ou mais populações específicas. Sob esta suposição, o nível de significância estatística que resulta da aplicação de um teste estatístico aos resultados de um experimento ou amostra corresponde a frequência com que a hipótese nula seria rejeitada em amostragens aleatorizadas repetidas da(s) mesma(s) população(ões) especificada(s). Como a amostragens repetidas da(s) população(ões) especificada(s) é geralmente impraticável, presume-se que a distribuição amostral das estatísticas do teste geradas sob amostragem aleatória repetida está em conformidade com uma distribuição teórica aproximada.

Por outro lado, em um teste estatístico de permutação a única suposição é que a variabilidade experimental causou o resultado observado. Assim, a estatística de teste é calculada para os dados observados, então as observações são permutadas sobre todas as disposições possíveis dos dados e a estatística de teste selecionada é calculada para cada combinação igualmente prováveis dos dados observados. 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.

Os testes de permutação diferem dos tradicionais testes paramétricos de várias formas. Primeiro, os testes de permutação são dependente dos dados (data-dependent), no sentido de que toda informação necessária para analise está contida no conjunto de dados observados. Implicitamente, esta entendido que a inferência estatística esta limitada ao experimento ou pesquisa realizada. Segundo, os testes de permutação são apropriados para amostras não aleatórias. Terceiro, os testes de permutação são livres de distribuição (free-distribution) na medida de que não dependem de pressupostos associados aos testes paramétricos tradicionais, como a normalidade e a homogeneidade da variância. Quarto, os testes de permutação fornecem valores de probabilidade exatos, ao invés de valores de probabilidade aproximados com base em uma distribuição teórica.

De acordo com Berry et al. (2016) os três tipos mais comuns de testes de permutação são: exato, aproximação por momentos (moment-aproximation) e aproximação por reamostragem (resampling-aproximation). Neste relatório discutirei apenas os teste de permutação exatos e aproximação por reamostragem.

1.1 Testes de Permutação Exatos

Um teste de permutação exato enumera exaustivamente todos as combinações (arranjos) igualmente prováveis dos dados observados. Então, para cada arranjo a estatística desejada é calculada. Os dados observados fornecem o valor observado da estatística do teste. A probabilidade de obter o valor observado da estatística de teste, ou mais um extremo, é a proporção das estatísticas calculadas para cada arranjo com valores iguais ou maiores do que o valor observado da estatística.

A título de ilustração considere o exemplo discutido em Rizzo (2008). Suponha que duas amostras aleatórias independentes $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ são observadas das distribuições F_X e F_Y, respectivamente. O interesse é testar $\mathcal{H}_0 : F_X = F_Y$ versus $\mathcal{H}_1 : F_X \neq F_Y$. 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.

Na prática, ao menos que as amostras sejam pequenas, calcular a estatística de teste para todas as $\binom{n+m}{n}$ permutações é computacionalmente excessivo. Uma aproximação reamostrando os dados sem reposição um número grande de vezes sera discutida a seguir.

1.2 Testes de Permutação Aproximados por Reamostragem

Os testes de permutação aproximados por reamostragem examinam subconjuntos aleatórios de Monte Carlo de todos as permutações possíveis e igualmente prováveis do conjunto de dados. Com um número suficiente de reamostragens é possível calcular a probabilidade de observar determinado valor da estatística. Berry et al. (2016) recomendam utilizar $M = 1.000.000$ de reamostras.

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 $j = 1, \ldots, M$:

    (a) Gere uma permutação aleatória;

(b) 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:

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

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

2 Testes Multivariados para Igualdade de Distribuições

Nesta seção será discutida duas abordagens para comparar distribuições multivariadas. Analogamente ao caso univariado o objetivo consiste em comparar se duas amostras são provenientes do mesma distribuição de probabilidade. O primeiro teste exige em uma série de suposição dos dados observado, no entanto se elas forem atendidas têm-se um teste assintoticamente mais poderoso. O segundo teste é menos restrito e não faz nenhum suposição sobre a distribuição dos dados.

2.1 Teste T^2 de Hotelling

A forma mais usual de comparar duas amostras quando $p \geq 2$ variáveis são observadas em cada unidade amostral é por meio do teste T^2 de Hotelling. Como aponta Berry et al. (2016) o teste T^2 de Hotelling é bastante utilizada para avaliar se existe diferença entre dois grupos quando $p \geq 2$ variáveis são mensuradas. Entretanto, este teste 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

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

em que $\bar{\mathbf{x}}_1$ e $\bar{\mathbf{x}}_2$ denotam os vetores de médias do grupo 1 e 2, respectivamente, n_1 e n_2 são os tamanhos amostrais do grupo 1 e 2, respectivamente e

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

sendo $\mathbf{S}_1$ e $\mathbf{S}_2$ as matrizes de covariância amostrais do grupo 1 e 2.

O valor observado da estatística T^2 de Hotelling, T_0^2, é convencionalmente transformada em uma estatística F observada por

$$F_0 = \frac{n_1 + n_2 - p - 1}{(n_1 - n_2 - 2)\,p}\,T_0^2 \tag{2}$$

na qual tem distribuição assintótica F de Snedecor com $\nu_1 = p$ e $\nu_2 = n_1 + n_2 - p - 1$.

Note que para a teoria clássica paramétrica funcionar é necessário que as amostras sejam independentes e provenientes de distribuições Normais multivariadas com matriz de covariância iguais. Por outro lado, o método de permutação não requer nenhuma dessa suposições, ou seja, podemos calcular estatística (2) para a amostra observada e comparar com a distribuição de permutação dela.

2.2 Teste baseado no k-NN

Proposto por Schilling (1986) e Henze (1988) é um teste não paramétrico baseado na proporção de todas as k comparações do vizinho mais próximo em que observações e seus vizinhos pertencem à mesma amostra.

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, $p \geq 2$. 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 $n = n_1 + n_2$.

Em geral, se as distribuições amostradas forem iguais, então a amostra agrupada $\mathbf{Z}$ tem, em média, menos vizinhos mais próximas coincidentes do que sob a hipótese alternativa. Usando esse ponto de vista os autores propuseram a estatística deste teste.

Vamos denotar o k-ésimo vizinho mais próximo de Z_i por $NN_k(Z_i)$. Além disso, define-se a seguinte função indicadora

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

A estatística considerada por Schilling (1986) para testar $\mathcal{H}_0 : F(x) = G(y)$ versus $\mathcal{H}_0 :$

$F(x) \neq G(y)$ é definida como

$$T_{k,n} = \frac{1}{n\,k} \sum_{i=1}^{n} \sum_{r=1}^{k} I_i(r), \tag{3}$$

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

Como mencionado anteriormente, sob a hipótese de distribuições iguais, a amostra agrupada tem em média menos vizinhos mais próximos coincidentes do que sob a hipótese alternativa, logo o teste rejeita a hipótese nula para grandes valores de $T_{k,n}$. Henze (1988) derivou expressões para a média e variância de (3) e mostrou a estatística tem distribuição assintoticamente Normal.

O inconveniente deste teste é determinar uma medida de distância entre as observações, bem como o número de vizinhos. Recentemente Lai et al. (2011) utilizaram a aproximação para a normal do teste multivariado baseado no k-NN para determinar o número de vizinhos mais próximos a ser utilizado no método de classificação k-NN.

2.3 Implementação

A implementação computacional dos testes foi realizada no software R (R Core Team, 2016). Para o teste T^2 de Hotelling o valor-p foi calculado utilizando a distribuição assintótica da estatística e a distribuição de permutação. A versão assintótica do teste baseado no k-NN não foi implementada. Na sequência temos os códigos em R das implementações dos testes.

Código 1: Implementação teste T^2 de Hotelling.

```
T2.Hotelling <-function(X1, X2, M)
{
   p  <- ncol(X1)
   n1 <- nrow(X1)
   n2 <- nrow(X2)
   n  <- n1 + n2
   T2 <- function(X1, X2, perm = FALSE)
   {
    if(perm == TRUE)
   {
    Z  <- rbind(X1, X2)
    Z  <- apply(Z, 2, sample, replace = FALSE, size = n)
    X1 <- Z[1:n1, ]
    X2 <- Z[(n1+1):n, ]
   }
   M1   <- apply(X1, 2, mean)
   S1   <- var(X1)
   M2   <- apply(X2, 2, mean)
   S2   <- var(X2)
   S    <- ((n1 - 1) * S1 + (n2 - 1) * S2) / (n - 2)
   stat <- ((n1 * n2) / n) * t(M1 - M2) %*% solve(S) %*% (M1 - M2)
   return(as.numeric(stat))
   }
   T2.orig <- T2(X1, X2)
   F0.orig <- (n - p - 1) / (p * (n -2)) * T2.orig
   T2.perm <- replicate(M, T2(X1, X2, perm = TRUE))
   F0.perm <- (n - p - 1) / (p * (n -2)) * T2.perm
```

```
    p.asym  <- 1 - pf(F_0, p, n - p - 1)
    p.perm  <- (1 + sum(T2.perm >= T2.orig)) / (M + 1)
    output  <- list(statistic = c(T2 = T2.orig, F0 = F0.orig), p.value = c(asymptotic = p.asym,
         permutation = p.perm),
    T2.perm = T2.perm, F0.perm = F0.perm)
    return(output)
}
```

Código 2: Implementação teste baseado no k-NN.

```
NN.test <- function(X, Y, k, M)
{
    Z    <- rbind(X, Y)
    n1   <- nrow(X)
    n    <- nrow(Z)
    Tn.k <- function(Z, k)
    {
     NN   <- FNN::get.knn(Z, k = k)$nn.index
     TN.k <- (sum(NN[1:n1, ] < n1 + 0.5) + sum(NN[(n1 + 1):n, ] > n1 + 0.5)) / (n * k)
     return(TN.k)
    }
    TNk.orig <- Tn.k(Z, k)
    TNk.perm <- replicate(n = M, expr = Tn.k(Z = apply(Z, 2, sample, replace = FALSE), k = k))
    p.value  <- (1 + sum(TNk.perm >= TNk.orig)) / (M + 1)
    output   <- list(statistic = TNk.orig, p.value = p.value, TNk.perm = TNk.perm)
    return(output)
}
```

3 Aplicações

Para ilustrar o método de permutação nos testes discutidos na seção anterior vamos considerar dois conjuntos de dados discutidos nos livros de Manly (2004) e Hand and Taylor (1987).

Aplicação I

Após uma tempestade em 1 de Fevereiro de 1898, um número de pardais moribundos foram levados a um laboratório em Rhode Island nos Estados Unidos. Posteriormente mais da metade dos pardais morreram. Hermon Bumpus, um pesquisador da Brown University, viu isso como uma oportunidade para encontrar subsídios a teoria da seleção natural de Charles Darwin. Em síntese, os dados consistem de cinco variáveis relacionadas a medidas do corpo dos 49 pardais moribundos e mais uma variável indicando se o pardal morreu ou não. Eles estão disponíveis na Tabela 1.1 do livro de Manly (2004). Aqui, o objetivo é verificar se existe diferença nas medidas corporais entre os pardais vivos e mortos.

Nesse sentido os testes T^2 de Hotteling e o teste baseado no k-NN foram utilizados. Na Figura 1 apresenta-se a distribuição de permutação aproximada para ambas as estatísticas. Percebe-se que o valor observado, isto é, valor calculado na amostras de ambas as estatísticas (linha vertical vermelha) não demostra ser muito improvável, ou seja, este valor não ocorreu por mero acaso. Logo, não rejeitamos a hipótese nula de que as medidas corporais dos pardais vivos diferem dos pardais mortos.

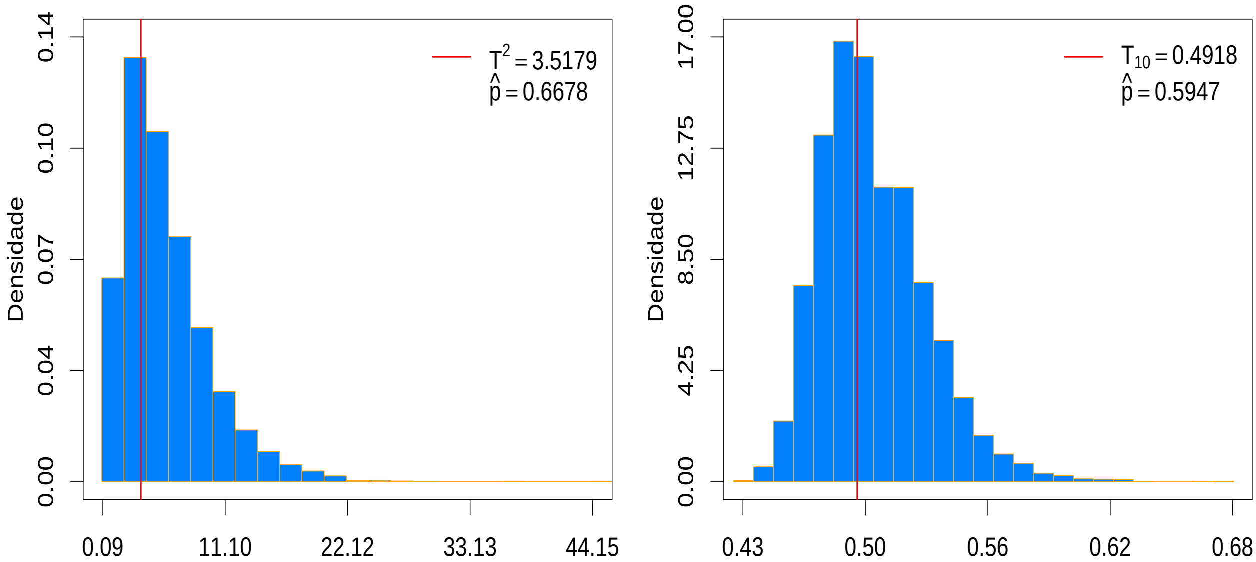

Figura 1: Distribuição de permutação aproximada das estatísticas T^2 de Hotelling (esquerda) e $T_{6,n}$ (direita) considerando $M = 10000$.

A partir da distribuição de permutação, foi possível estimar a probabilidade da estatística observada de ambos os teste ocorrer por mero acaso, isto é, o velho conhecido valor-p . Para a estatística T^2 de Hotelling obtemos $\widehat{p} = 0.6636$, enquanto que para a estatística baseada k-NN, obtemos $\widehat{p} = 0.5947$. Ressalta-se que para a estatística k-NN, foi considerado os 10 vizinhos mais próximos.

Aplicação II

Na segunda aplicação, foi considerado o conjunto de dados apresentado no capítulo *Study B* do livro de Hand and Taylor (1987). Os dados 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. Uma representação gráfica dos dados é exibida na Figura 2.

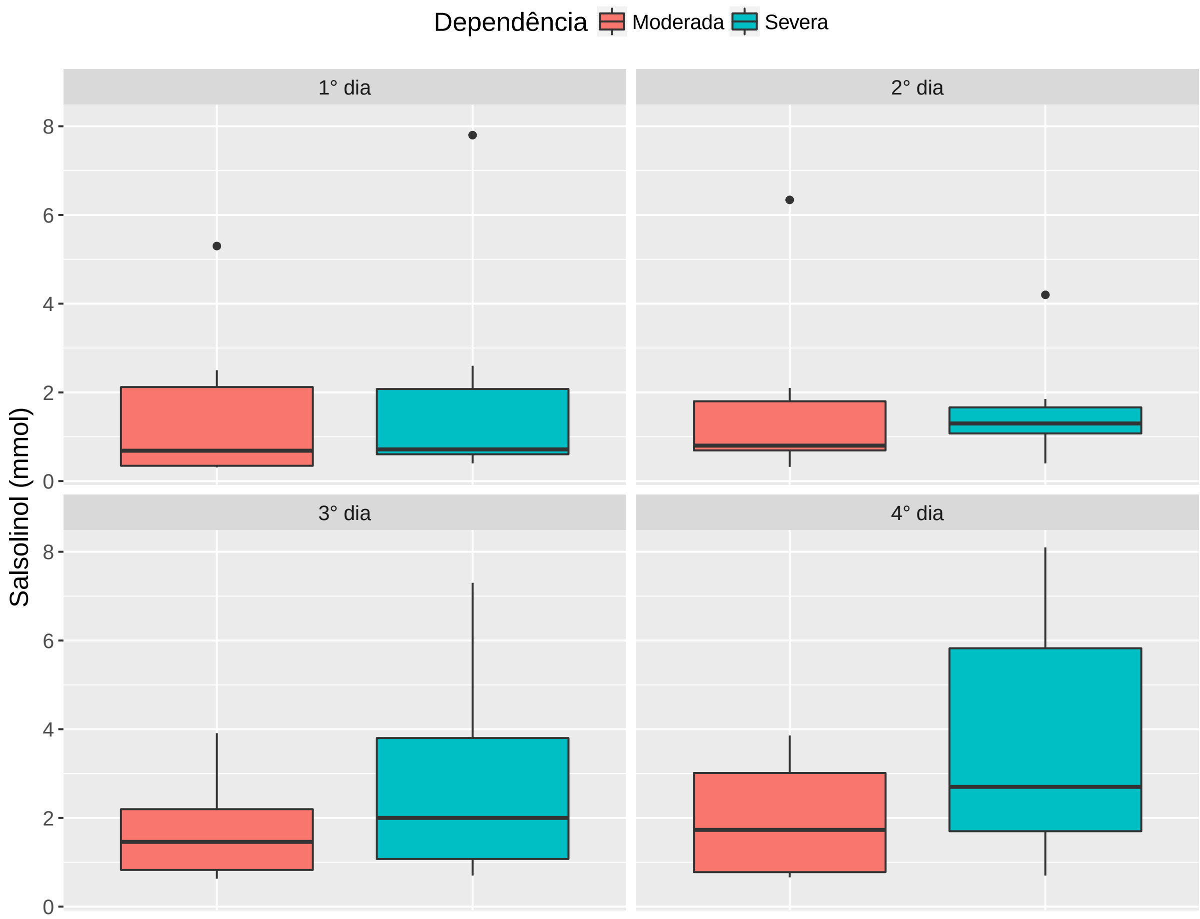

Figura 2: Comportamento do salsolinol durante os 4 dias e conforme o grupo.

O objetivo nesta análise é verificar se existe diferença estatisticamente significativa entre os o grupo de paciente com dependência severa e moderada. Os dados observados podem ser classificados com medidas repetidas ao longo do tempo. Uma abordagem para analisar medidas repetidas de dados é considerar uma estrutura multivariada, em que as variáveis são as mensurações da variável resposta ao longo do tempo. Neste formato, podemos utilizar os testes T^2 de Hotelling e o teste baseado no k-NN para verificar se existe diferença entre os grupos.

Apresenta-se na Figura 3 a distribuição de permutação das estatísticas T^2 e $T_{8,n}$. Percebe-se que o valor observado a partir dos dados não parece ter ocorrido por mero acaso, logo há evidências de que não existe diferença entre os grupos com moderada e severa dependência.

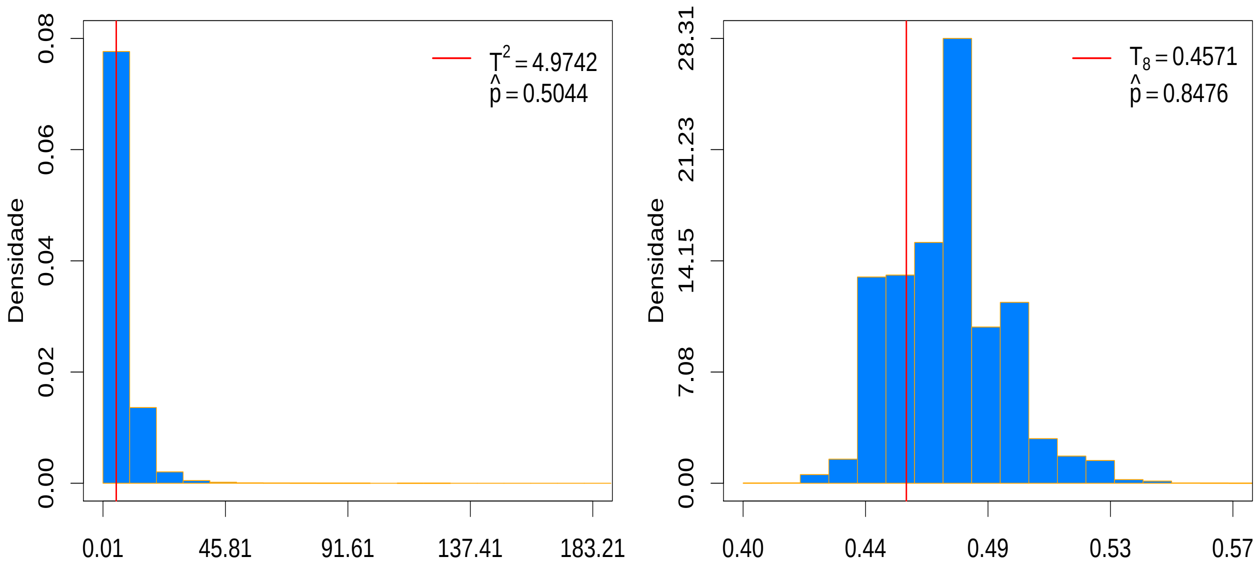

Figura 3: Distribuição de permutação aproximada das estatísticas T^2 de Hotelling (esquerda) e $T_{6,n}$ (direita) considerando $M = 10000$.

Tendo em vista a distribuição de permutação, obtemos estimar a probabilidade da estatística observada ocorrer por mero acaso, isto é, o valor-p. Para a estatística T^2 de Hotelling obtemos $\widehat{p} = 0.5044$, ao passo que para a estatística baseada no k-NN têm-se $\widehat{p} = 0.8476$. Conclui-se que não existe diferença em relação ao salsolinol, entre os pacientes com moderada e severa dependência de álcool.

Referências

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

Fisher, R., 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

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

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

Lai, Y., Wu, B., Zhao, H., 2011. A permutation test approach to the choice of size k for the nearest neighbors classifier. Journal of Applied Statistics 38 (10), 2289–2302.

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

Neyman, J., Pearson, E. S., 1928a. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. Biometrika 20A (1/2), 175–240.

Neyman, J., Pearson, E. S., 1928b. On the use and interpretation of certain test criteria for purposes of statistical inference: Part ii. Biometrika 20A (1/2), 263–294.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

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

Rizzo, M. L., 2008. Statistical Computing with R. Chapman & Hall/CRC.

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