

Teste de Kuskal-Wallis

Linguagem R

2023

Apresentação do teste

Os dados (X) consistem de $N = \sum_{j=1}^k n_j$ observações, sendo que existem n_j observações do j -ésimo grupo, $j = 1, \dots, k$, $k \geq 2$.

Suposições

1. As N variáveis aleatórias $\{X_{1j}, X_{2j}, \dots, X_{n_j, j}\}$, $j = 1, \dots, k$ são mutuamente independentes.
2. Para cada $j \in \{1, \dots, k\}$, as n_j variáveis aleatórias $\{X_{1j}, X_{2j}, \dots, X_{n_j, j}\}$ formam uma amostra aleatória de uma distribuição contínua com função distribuição F_j .
3. As funções distribuição F_1, \dots, F_k estão relacionadas da forma

$$F_j(t) = F(t - \tau_j), \quad t \in \mathbb{R}, \quad j = 1, \dots, k,$$

em que F é uma função distribuição de uma variável aleatória contínua com mediana (desconhecida) θ e τ_j representa o efeito (desconhecido) do j -ésimo grupo.

Estas suposições correspondem ao modelo de um fator (grupo, no caso) dado por

$$X_{ij} = \theta + \tau_j + \epsilon_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, k,$$

em que θ é a mediana geral, τ_j é o efeito do j -ésimo grupo (ou tratamento) e ϵ_{ij} , $i = 1, \dots, n_j$ e $j = 1, \dots, k$, formam uma amostra aleatória de uma distribuição contínua com mediana igual a 0. A mediana de X_{ij} é $\theta + \tau_j$.

Hipóteses

$$H_0 : \tau_1 = \dots = \tau_k \quad \text{contra} \quad H_1 : \tau_j \neq \tau_\ell \text{ para pelo menos um par } (j, \ell), j \neq \ell. \quad (1)$$

Procedimento Todas as N observações dos k grupos são combinadas e dispostas em ordem crescente. O posto de X_{ij} na amostra combinada ordenada é denotado por r_{ij} .

Se H_0 for verdadeira, a soma total dos postos, que é igual a $N(N+1)/2$, é dividida proporcionalmente entre as k amostras de acordo com os tamanhos amostrais n_1, \dots, n_k . Sendo assim, a soma esperada dos postos para a j -ésima amostra seria

$$\frac{n_j}{N} \frac{N(N+1)}{2} = \frac{n_j(N+1)}{2}, \quad j = 1, \dots, k.$$

A soma dos postos das observações da j -ésima amostra e sua média são

$$R_j = \sum_{i=1}^{n_j} r_{ij} \quad \text{e} \quad \bar{R}_j = \frac{R_j}{n_j}, \quad j = 1, \dots, k.$$

Portanto, se H_0 for verdadeira, \bar{R}_j tende a ser “próxima” de $E_0(\bar{R}_j) = (N + 1)/2$. Quanto maiores forem as diferenças, $j = 1, \dots, k$, mais forte a indicação de rejeição de H_0 .

Estatística de teste

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left(\bar{R}_j - \frac{N+1}{2} \right)^2, \quad (2)$$

que é uma soma ponderada dos quadrados das diferenças entre \bar{R}_j e $E_0(\bar{R}_j)$, $j = 1, \dots, k$.

Distribuição exata (sem empates) Na amostra combinada os postos são $1, \dots, N$. Sob H_0 , todas as $N! / \prod_{j=1}^k n_j!$ atribuições de n_1 postos às observações do grupo 1, n_2 postos às observações do grupo 2, \dots , n_k postos às observações do grupo k são igualmente prováveis. Para cada sequência de postos calculamos H . A distribuição exata de H é dada pela tabela de frequências relativas dos valores de H . Tabelas da distribuição exata de H foram preparadas para alguns valores de k e n_1, \dots, n_k .

O valor- p do teste é dado por $P_0(H \geq H_{\text{obs}})$. O teste é livre de distribuição.

Distribuição assintótica Quando H_0 é verdadeira e $\min(n_1, \dots, n_k) \rightarrow \infty$, a distribuição de H tende a uma distribuição χ_{k-1}^2 .

O valor- p aproximado é dado por $P(Q \geq H_{\text{obs}})$, com $Q \sim \chi_{k-1}^2$. O teste é assintoticamente livre de distribuição.

Empates Se ocorrerem empates entre as observações X_{ij} , para cada grupo de observações de mesmo valor tomamos a média dos postos na situação sem empates. A estatística H é calculada com a expressão (2).

Para utilizar a distribuição exata (como uma aproximação) ou a distribuição assintótica, substituímos H por

$$H^* = \frac{H}{1 - \sum_{\ell=1}^g (f_\ell^3 - f_\ell) / (N^3 - N)},$$

em que g denota o número de observações diferentes e f_ℓ é a frequência de cada uma das diferentes observações. A obtenção da distribuição exata *condicional* de H em (2) é mais trabalhosa. Caso não ocorram empates, temos $g = N$ e $f_\ell = 1$, $\ell = 1, \dots, N$, de modo que $H^* = H$.

Comparações múltiplas Quando a hipótese H_0 em (1) é rejeitada, quaisquer dois grupos j e ℓ podem ser comparados, com $1 \leq j < \ell \leq k$, totalizando $k(k-1)/2$ comparações. A estatística de teste é

$$Z_{j\ell} = \frac{|\bar{R}_j - \bar{R}_\ell|}{\left\{ \frac{N(N+1)}{12} \left(\frac{1}{n_j} + \frac{1}{n_\ell} \right) \right\}^{1/2}}.$$

Se $Z_{j\ell, \text{obs}} \geq z_{\alpha/\{k(k-1)\}}$, a diferença entre os grupos j e ℓ é significativa, sendo que $P(Z \geq z_{\alpha/\{k(k-1)\}}) = \alpha/\{k(k-1)\}$, $Z \sim N(0, 1)$ e α é o nível de significância global.

Se, por exemplo, o grupo 1 ($j = 1$) é o grupo de referência, um teste semelhante pode ser proposto para as $k-1$ comparações possíveis utilizando as diferenças $\bar{R}_\ell - \bar{R}_1$, $\ell = 2, \dots, k$.

Estes dois procedimentos, desenvolvidos por Dunn (1964) e baseados na desigualdade de Bonferroni, podem ser conservadores.

Exemplo

```
# Separador decimal: ","
options(OutDec = ",")

## Dados
# Ex 8., p. 200 em Hollander Wolfe (1999, 2nd Edition)
# Número de sítios receptores por célula de leucócito (em milhares)
# Grupos: normal, hairy-cell anemia, chronic lymphatic leukemia,
# chronic myelocytic leukemia e acute leukemia.

dados <- list(normal = c(3.5, 3.5, 3.5, 4, 4, 4, 4.3, 4.5, 4.5, 4.9,
  5.2, 6, 6.75,8), hca = c(5.71, 6.11, 8.06, 8.08, 11.4),
  cll = c(2.93, 3.33, 3.58, 3.88, 4.28, 5.12),
  cml = c(6.32, 6.86, 11.4, 14),
  al = c(3.23, 3.88, 7.64, 7.89, 8.28, 16.2, 18.25, 29.9))

names(dados)

## [1] "normal" "hca" "c11" "cml" "al"
cat("\n Número de grupos =", length(dados))

##
## Número de grupos = 5
cat("\n Tamanhos amostrais:", as.numeric(lapply(dados, length)))

##
## Tamanhos amostrais: 14 5 6 4 8

# Estatísticas descritivas
lapply(dados, summary)

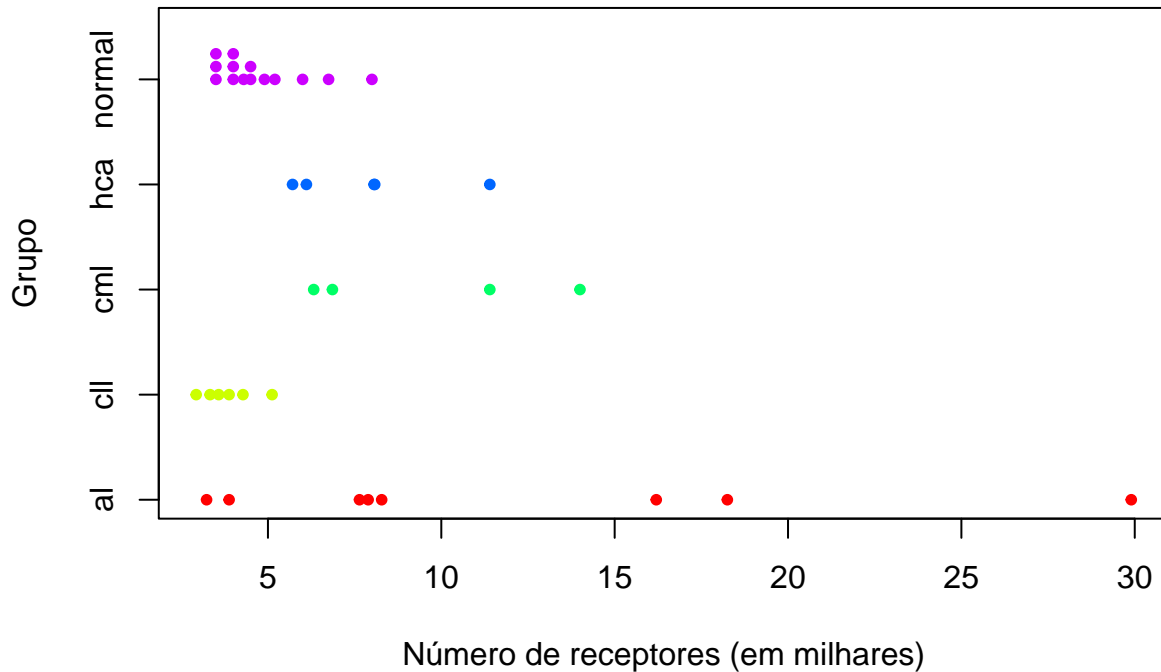
## $normal
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  3,500  4,000   4,400   4,761  5,125   8,000
##
## $hca
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  5,710  6,110   8,060   7,872  8,080  11,400
##
## $c11
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  2,930  3,393   3,730   3,853  4,180   5,120
##
## $cml
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  6,320  6,725   9,130   9,645 12,050  14,000
##
## $al
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  3,230  6,700   8,085  11,909 16,712  29,900

# Gráfico de pontos
vdados <- unlist(dados) # dados em um vetor
cores <- rainbow(length(dados))
```

```

grupo <- factor(rep(names(dados), times = lapply(dados, length)))
stripchart(vdados ~ grupo, method = "stack", pch = 20,
  xlab = "Número de receptores (em milhares)",
  col = cores, ylab = "Grupo")

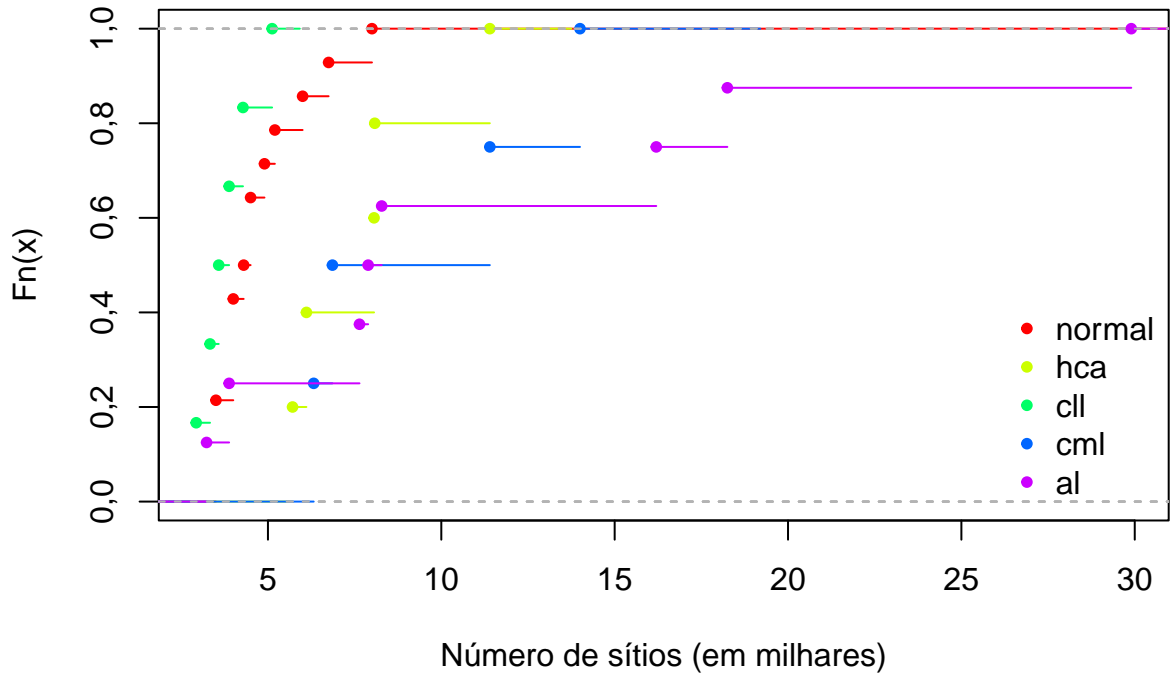
```



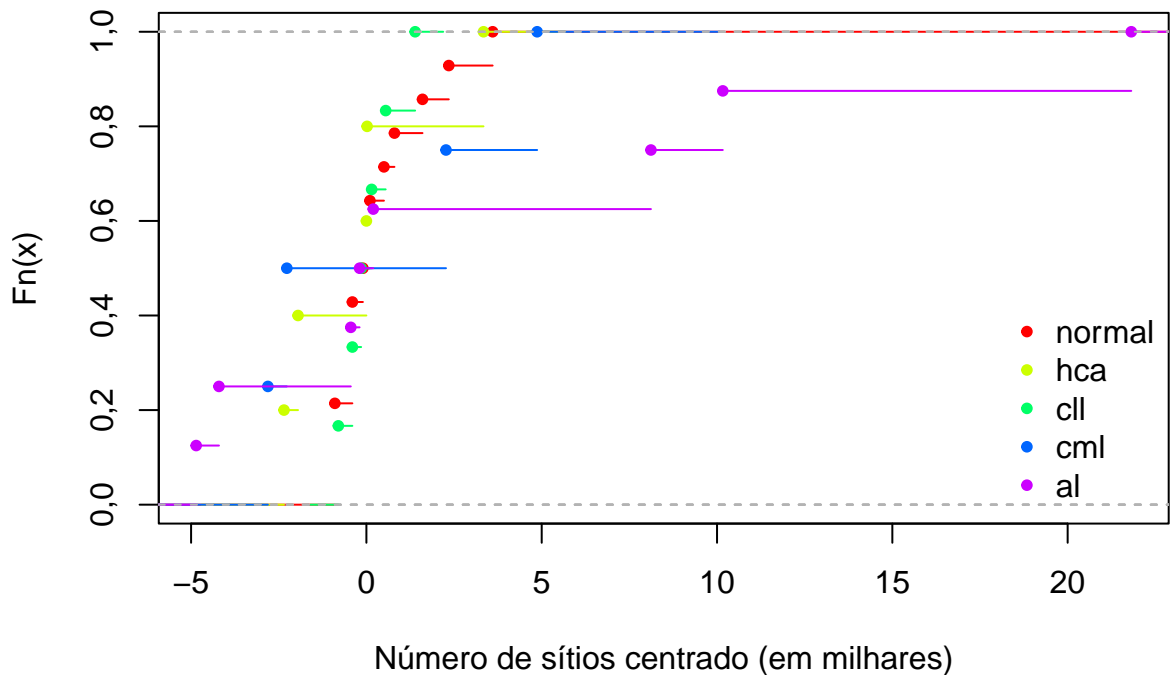
```

# Distribuições empíricas
plot(ecdf(dados[[1]]), main = "", col = cores[1], xlim = range(dados),
  xlab = "Número de sítios (em milhares)", pch = 20)
for (j in 2:length(dados)) {
  lines(ecdf(dados[[j]]), main = "", col = cores[j], pch = 20)
}
legend("bottomright", names(dados), pch = 20, col = cores, bty = "n")

```



```
# Distribuições empíricas dos dados centrados (pela mediana)
estimat <- unlist(lapply(dados, median)) # Medianas amostrais
plot(ecdf(dados[[1]] - estimat[[1]]), main = "", col = cores[1],
     pch = 20, xlim = range(unlist(dados) - rep(estimat, times =
       lapply(dados, length))), xlab = "Número de sítios centrado (em milhares)")
for (j in 2:length(dados)) {
  lines(ecdf(dados[[j]] - estimat[[j]]), main = "", col = cores[j],
       pch = 20)
}
legend("bottomright", names(dados), pch = 20, col = cores, bty = "n")
```



Nota 1. O gráfico acima justifica a suposição de que as distribuições diferem apenas pela mediana?

```
# Teste de Kruskal-Wallis
kruskal.test(dados)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: dados
## Kruskal-Wallis chi-squared = 16,668, df = 4, p-value =
## 0,002242
```

```
## Usando o pacote coin
library(coin)
# Distribuição assintótica
kruskal_test(vdados ~ grupo, distribution = "asymptotic")
```

```
##
## Asymptotic Kruskal-Wallis Test
##
## data: vdados by grupo (al, cll, cml, hca, normal)
## chi-squared = 16,668, df = 4, p-value = 0,002242
```

Adotando um nível de significância de 5%, os resultados indicam que há diferenças significativas entre as distribuições ($H = 16,7$ com 4 g.l., $p = 0,0022$).

Com o pacote *coin*, no caso de empates, a aproximação assintótica é usada por *default*. O valor- p baseado na distribuição exata condicional pode ser calculado para $k = 2$ grupos (`distribution = "exact"`).

```
# Comparações múltiplas com o teste de Wilcoxon para duas amostras
pairwise.wilcox.test(vdados, grupo, p.adjust = "bonferroni",
  alternative = "two.sided")
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
## compute exact p-value with ties
```

```
##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data: vdados and grupo
##
##      al    cll    cml    hca
## cll  0,387 -      -      -
```

```
## cml    1,000 0,095 -      -  
## hca    1,000 0,043 1,000 -  
## normal 0,512 1,000 0,089 0,061  
##  
## P value adjustment method: bonferroni
```

Se $p_{j\ell}$ denota o valor- p do teste da soma dos postos de Wilcoxon quando são comparados os grupos j e ℓ , verifique que na matriz acima os resultados são dados por $\min(10p_{j\ell}, 1)$. Por que 10?