Testes não paramétricos

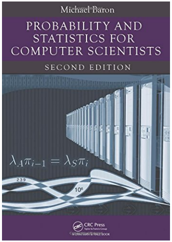
André Gustavo dos Santos

Departamento de Informática
Universidade Federal de Viçosa

INF222 - 2022/2

– Fonte do material

Conteúdo e figuras do tópico 3 são do livro complementar da disciplina (seção 10.2):



Baron, Michael.
Probability and statistics for computer scientists.
Chapman and Hall/CRC, 2013.

Tópicos da aula

Exemplos em Python

- Além da teoria, os slides contêm exemplos de como fazer os testes em Python
- Serão usados os seguintes pacotes

```
# Pacotes para os dados
import numpy as np
import pandas as pd

# Pacotes estatísticos
import scipy.stats as stats
import statsmodels.api as sm

# Pacote para gráficos
import pylab
```

Exemplos em Python – dados

- Os dados podem ser inicializados diretamente

```
var1 = np.array([1,2,3,4,5])
var2 = np.array([6,7,8,9,10])
```

- Podem ser lidos de planilha

```
dados = pd.read_excel('aula1.xlsx')
idator = dados.loc[:,'HOMENS'].values.astype(float)
idatriz = dados.loc[:,'MULHERES'].values.astype(float)
```

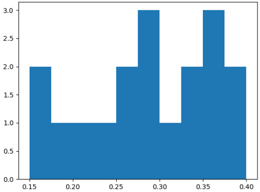
- Podem ser gerados aleatoriamente

```
x1 = np.random.uniform(size = 20)
x2 = np.random.normal(size = 100)
x3 = np.random.normal(loc = 5, scale = 2, size = 200)
```

Teste de normalidade

- Vários dos métodos para testes de hipótese vistos até então partem da premissa de normalidade da distribuição
  - No exemplo da análise da influência de gravação de CD na duração da bateria dos laptops assumimos uma distribuição normal dos dados.
  - Em outro exemplo, analisamos o tempo médio entre as teclas de alguém que digitou nome e senha para entrar em um sistema. Consideramos uma distribuição normal, mas um histograma da amostra não confirma (nem nega) isso:

```
import matplotlib.pyplot as plt
x = np.array([.24,.22,.26,.34,.35,.32,.33,.29,.19,
              .36,.30,.15,.17,.28,.38,.40,.37,.27])
plt.hist(x)
plt.show()
```



- Poderíamos fazer os testes de hipóteses sem essa suposição?
- E como verificar se os dados seguem aproximadamente uma distribuição normal?

Teste de normalidade

- O teste de Shapiro-Wilk para normalidade tem por objetivo fornecer uma estatística de teste para avaliar se uma amostra tem distribuição normal.
- O teste pode ser utilizado para amostras de qualquer tamanho.
- A estatística de teste para normalidade é definida como:

$$W = \frac{(\sum a_i x_i)^2}{\sum (x_i - \bar{x})^2}$$

  em que x_i é a variável aleatória observada e a_i são coeficientes tabelados para cada α, tendo como referência a distribuição normal.
- Se o valor calculado W for menor que o tabelado, rejeita-se a hipótese de normalidade ao nível de significância α.

- Em Python

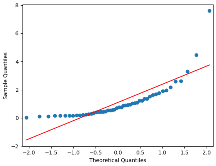
```
import scipy.stats as stats
stats.shapiro(x3)
```

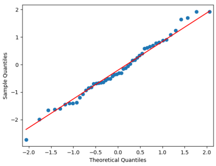
Gráfico q-q

- Uma alternativa mais subjetiva é a análise gráfica dos dados
- O gráfico q-q (quantil-quantil) apresenta os quantis amostrais contra os quantis teóricos da distribuição colocados como referência em uma linha reta.
- Costuma ser melhor verificar se uma reta se ajusta aos pontos do que se uma curva de densidade se ajusta a um histograma ou uma curva de probabilidade acumulada se ajusta à acumulada empírica.

- Em Python

```
import statsmodels.api as sm
import pylab
xe = np.random.exponential(size = 50)
sm.qqplot(xe,line='s')
pylab.show()
xn = np.random.normal(size = 50)
sm.qqplot(xn,line='s')
pylab.show()
```





Procedimentos não-paramétricos

- **Métodos paramétricos:** baseados em amostras aleatórias provenientes de uma família paramétrica particular de distribuições. Possuem requisitos sobre a natureza dos dados (em geral, pressupõem a distribuição populacional normal).
- **Métodos não paramétricos:** métodos livres de hipóteses sobre a distribuição dos dados (não exigem normalidade). São chamados de testes livres de distribuição.

Procedimentos não-paramétricos

- Em procedimentos não paramétricos, os dados são tratados de forma qualitativa: são ordenados e arranjados segundo algum critério.
- Como não utilizam toda a informação fornecida pela amostra, será menos eficiente que o procedimento paramétrico correspondente quando a população em questão for normal.
- Geralmente, se os métodos paramétricos e não paramétricos forem aplicáveis a um problema particular, devemos usar o paramétrico, por ser mais eficiente.
- Mas, esta perda de eficiência frequentemente não é grande, especialmente quando consideramos amostras grandes.
- Os métodos não paramétricos podem ser aplicados a dados categóricos.

Teste dos sinais

- O teste dos sinais é um teste não-paramétrico usado para testar hipóteses a respeito da mediana de uma distribuição qualquer:

$$H_0 : M = m$$

$$H_1 : M \neq m$$

- Estamos testando se exatamente metade da população está abaixo de m e metade acima de m (em distribuição contínua, $\mathbf{P}\{X_i = m\} = 0$)
- Para o teste dos sinais, contamos quantas observações estão acima de m

$$S_{obs} = S(X_1, \ldots, X_n) = \text{ número de } \{i : X_i > m\}$$

- Se H_0 é verdadeira, m é a mediana e cada X_i tem chances iguais de estar acima ou abaixo de m; então S tem distribuição Binomial com parâmetros n e $p = 0.5$
- Para n grande e X_i com distribuição contínua, S é aproximadamente normal: $S \approx N(n/2, \sqrt{n}/2)$

Teste dos sinais

- O teste unilateral à direita será sustentado por valores maiores que S_{obs}, então o valor-p dever ser $P = \mathbf{P}\{S \geq S_{obs}\}$. No teste à esquerda, $P = \mathbf{P}\{S \leq S_{obs}\}$. No teste bilateral, o menor dos dois: $P = 2 \times \min(\mathbf{P}\{S \leq S_{obs}\}, \mathbf{P}\{S \geq S_{obs}\})$.

- Quando a distribuição for simétrica, sua mediana é igual à média, então o teste dos sinais pode ser usado para testar hipóteses sobre a média
- Em especial, o teste pode ser usado para a média de uma distribuição normal

- Ao aplicar o teste de sinais, estatísticos costumavam marcar os dados acima do valor testado ($X_i > m$) com '+' e os valores abaixo com '-'. Daí o nome do teste...

Exemplo

Exemplo: detecção de intruso (de novo)

Em um exemplo de outra aula, alguém digitou o nome e a senha corretas para entrar no sistema com os seguintes tempos entre as teclas:

.24, .22, .26, .34, .35, .32, .33, .29, .19, .36, .30, .15, .17, .28, .38, .40, .37, .27 segundos

Como primeiro passo para detectar se é um intruso, havíamos calculado o intervalo de confiança de 99% para a média do tempo entre as teclas, assumindo distribuição normal.

Vamos fazer o teste sem essa suposição.

Solução:

- Hipótese nula $H_0 : M = 0.2$ versus hipótese alternativa $H_1 : M \neq 0.2$
- Dos 18 valores da amostra, $S_{obs} = 15$ são maiores que 0.2
- Usando distribuição Binomial com $n = 18$, $p = 0.5$, encontramos

$$\text{valor-}p = 2 \times \min(\mathbf{P}\{S \leq 15\}, \mathbf{P}\{S \geq 15\}) = 2 \times \min(0.9993, 0.0038) = 0.0076$$

- O teste dos sinais rejeitará H_0 para qualquer $\alpha > 0.0076$, uma evidência de acesso não autorizado

Dica para consultar na tabela: $\mathbf{P}\{S \geq 15\} = \mathbf{P}\{S \leq 3\}$ *pois a distribuição binomial é simétrica*

Exemplo

Exemplo: acima da velocidade

Ao passar de carro na reta da UFV, um estudante notou que vários veículos pareciam estar acima do limite de velocidade permitido, de 50km/h. Como experimento, no dia seguinte ele passou pela reta exatamente a 50km/h e contou os carros. Foram 56 acima de sua velocidade e 44 abaixo. Isto confirma sua suspeita que a maioria está acima (ou seja, mediana > 50)?

Solução:

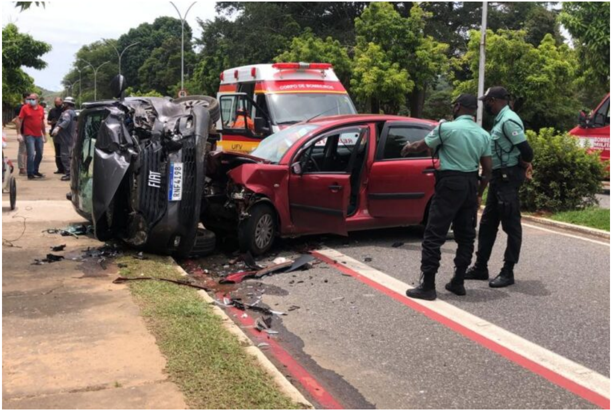
- Hipótese nula $H_0 : M = 50$ versus hipótese alternativa $H_1 : M > 50$
- A rejeição de H_0 indica que mais da metade dos carros que passam pela reta excede a velocidade permitida
- A estatística do teste dos sinais é $S = 56$ e o tamanho da amostra é $n = 100$
- A distribuição nula de S é aproximadamente normal com $\mu = n/2 = 50$ e $\sigma = \sqrt{n}/2 = 5$
- Calculando o valor-p (ajuste de -0.5 para correção de continuidade)

$$\text{valor-}p = \mathbf{P}\{S \geq 56\} = \mathbf{P}\{Z \geq \frac{55.5 - 50}{5}\} = 1 - \Phi(1.1) = 1 - 0.8643 = 0.1357$$

- Não há evidência significativa que a mediana da velocidade dos carros esteja acima do limite

Note que foram usados dados ordinais, não foi necessário saber a velocidade exata para o teste dos sinais

No início do ano...



Teste dos postos sinalizados de Wilcoxon

- O teste dos postos sinalizados de Wilcoxon é um teste não-paramétrico também usado para testar hipóteses a respeito da mediana, mas para uma distribuição aproximadamente simétrica:

$$H_0 : M = m$$

$$H_1 : M \neq m$$

- Um posto é um número atribuído a um item amostral individual de acordo com sua posição na lista ordenada: o primeiro item da lista tem posto 1, o segundo item tem posto 2, etc.
- Assim, o teste incorpora e usa mais informação, tendendo a resultar em conclusões que refletem melhor a verdadeira natureza dos dados.

Teste dos postos sinalizados de Wilcoxon

Exemplo de posto

Considere a amostra $3, 7, 5, 6, 5, 4$

- A observação de menor valor, 3, tem posto 1
- A de segundo menor valor, 4, tem posto 2
- As duas seguintes têm valor 5; elas têm posto 3.5, a média dos postos 3 e 4
- A seguinte na ordem é 6, que tem posto 5
- A de maior valor é 7, que recebe posto 6

Os postos dos itens da amostra são, então:

- $R_1 = 1, R_2 = 6, R_3 = 3.5, R_4 = 5, R_5 = 3.5, R_6 = 2$

Teste dos postos sinalizados de Wilcoxon

Teste dos postos sinalizados de Wilcoxon[1] para a hipótese nula $H_0 : M = m$:

1. Calcule as distâncias entre as observações e o valor testado: $d_i = |X_i - m|$
2. Ordene as distâncias e calcule os postos R_i (das distâncias d_i, não de X_i)
3. Considere somente postos das observações $X_i > m$. Sua soma é a estatística W:

$$W = \sum_{i:X_i>m} R_i$$

4. Valores altos de W sugerem rejeição de H_0 em favor de $H_1 : M > m$, enquanto valores baixos sugerem rejeição de H_0 em favor de $H_1 : M < m$; ambos sugerem a alternativa bilateral $H_1 : M \neq m$

- Os valores críticos são encontrados em tabelas (geralmente para $n \leq 30$)
- Wilcoxon propôs usar postos sinalizados com sinal '+' para posto R_i se $X_i > m$ e sinal '-' para posto R_i se $X_i < m$. A estatística W é a soma dos postos de sinal +[2]

[1] proposto pelo químico e estatístico irlandês-americano Frank Wilcoxon

[2] algumas pessoas preferem somar os postos de sinal '-'; são equivalentes, qualquer um pode ser usado

Exemplo

Exemplo - oferta e demanda

Você foi contratado para gerenciar um laboratório de computadores com impressoras. Uma de suas responsabilidades é não deixar acabar o papel das impressoras. Nos primeiros seis dias de seu trabalho, foram consumidas as seguintes quantidades de resmas:

$$7,\ 5.5,\ 9.5,\ 6,\ 3.5,\ 9$$

Ter oferta suficiente para atender a demanda requer ajuda de estatística. Os dados indicam, com nível de significância de 5%, que a mediana de consumo diário é maior que 5 resmas?
(vamos assumir que a quantidade em cada dia é independente e os dados representam uma boa amostra aleatória)

Solução:

- Vamos testar $H_0 : M = 5$ versus $H_1 : M > 5$
- Consultando numa tabela $n = 6$ e $\alpha = 0.05$, rejeitaremos H_0 se a estatística $W \geq 19$
- Para calcular W, ordenamos as diferenças $d_i = |X_i - 5|$ e somamos os postos "positivos"

i	X_i	$X_i - 5$	d_i	R_i	sinal
1	7	2	2	4	+
2	5.5	0.5	0.5	1	+
3	9.5	4.5	4.5	6	+
4	6	1	1	2	+
5	3.5	-1.5	1.5	3	-
6	9	4	4	5	+

- Soma dos postos de sinal '+': $W = 4 + 1 + 6 + 2 + 5 = 18$, menor que o valor para rejeição
- Não há evidência com significância de 5% que a mediana do consumo diário é maior que 5

Teste dos postos sinalizados de Wilcoxon

- Usando postos em vez de sinais, o teste leva em consideração as magnitudes das diferenças e não apenas o número de sinais do desvio do valor a ser testado
- Para $n \geq 15$ podemos usar uma aproximação normal: $W \approx N(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}})$, com a devida correção para variável contínua
- Somando-se os postos com sinal positivo[3] (W) e temos a estatística do teste z:

$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

- O teste estatístico consiste em avaliar a magnitude ordenada das diferenças, em relação aos valores de uma tabela, para um dado nível de significância α.

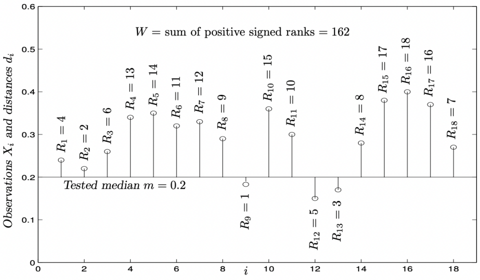
[3] ou os postos com sinal negativo

Exemplo

detecção de intruso (última vez...)

Aplicando o teste dos postos sinalizados de Wilcoxon no teste $H_0 : M = 0.2$ versus $H_1 : M \neq 0.2$:

- calculamos as diferenças $d_1 = |0.24 - 0.2| = 0.04, \ldots, d_{18} = |0.27 - 0.2| = 0.07$
- ordenamos as diferenças e atribuímos postos sinalizados (veja figura)
  (no caso, apenas as observações 9, 12 e 13 estão abaixo do valor testado $m = 0.2$)
- calculando a soma dos postos de sinal '+', obtemos $\sum_{i: X_i > 0.2} R_i = 162$
- encontramos o valor-p: $p = 2 \times \min(\mathbf{P}\{W \leq 162\}, \mathbf{P}\{W \geq 162\}) < 2 \times 0.001 = 0.002$
  (consultado na tabela com $n = 18$)
- o teste mostra forte evidência que se trata de intruso



Exemplo

detecção de intruso (cont...)

Como $n \geq 15$, também podemos fazer o cálculo por aproximação da normal:

- $\mathbf{E}(W|H_0) = \frac{(18)(19)}{4} = 85.5$
- $\text{Std}(W|H_0) = \sqrt{\frac{(18)(19)(37)}{24}} = 23.0$
- fazendo a correção de continuidade (W era discreta) de 162 para 161.5 temos $P = 2 \times \mathbf{P}\{Z \geq \frac{161.5-85.5}{23.0}\} = 2(1 - \Phi(3.30)) = 2 \times 0.0005 = 0.001$
- forte evidência que se trata de intruso

- Note que, comparado com o resultado do teste de sinais, o teste dos postos sinalizados de Wilcoxon obteve um resultado mais forte
- Geralmente é assim, pois usa mais informação dos dados
- Ele é mais sensível à violação de H_0; isto é, se H_0 não é verdadeira, o teste de Wilcoxon tem mais chance de detectar isso

Teste dos postos sinalizados de Wilcoxon

- Em Python

```
import scipy.stats as stats

# Para comparar duas variáveis
stats.wilcoxon(var1,var2)

# Para comparar uma variável com mediana de referência zero:
stats.wilcoxon(var1,None)
```

- Exemplo (detecção de intruso):

```
# dados da amostra
tempo = [.24,.22,.26,.34,.35,.32,.33,.29,.19,
         .36,.30,.15,.17,.28,.38,.40,.37,.27]

# shift 0.2 em cada um pois o teste verifica mediana 0
diftempo = [x - 0.2 for x in tempo]

stats.wilcoxon(diftempo)
```

Teste da soma dos postos de Mann-Whitney-Wilcoxon

- E no caso de duas populações? Como comparar suas medianas?
- O teste da soma dos postos é um teste não paramétrico que usa postos de dados amostrais de duas populações independentes para testar a hipótese nula de que as duas amostras independentes provêm de populações com distribuições iguais:

$$H_0 : F_X = F_Y$$

$$H_1 : F_X \neq F_Y$$

- Para duas amostras independentes $X_1, \ldots, X_{n_1}$ e $Y_1, \ldots, Y_{n_2}$ de duas populações
  1. Combinamos todos X_i e Y_j em uma amostra só e atribuímos postos (de 1 a $n_1 + n_2$)
  2. A estatística de teste U é a soma de todos os postos de X_i
  3. Valores pequenos de U sugerem que Y é estocasticamente maior que X, pois tem maiores postos (nesse caso, $F_Y(t) < F_X(t)$, X "acumula" mais rápido)
- Alternativamente, se $n_1, n_2 > 10$, U é aproximadamente normal, com estatística

$$z = \frac{U - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}}$$

- Wilcoxon propôs o teste para amostras de tamanhos iguais; posteriormente Henry Mann and Donald Whitney o estenderam para tamanhos diferentes

UFV

Exemplo

Promoção em *e-commerce*

Os responsáveis por um portal de *e-commerce* suspeitam que terão mais clientes se oferecerem algum incentivo nas compras. Para verificar esta hipótese, escolheram 12 dias ao acaso. Em 6 desses dias, escolhidos aleatoriamente, ofereceram 5% de desconto em anúncios que levavam ao portal; e em 6 dias não ofereceram desconto.

Nos dias de anúncio com desconto, receberam 1200, 1700, 2600, 1500, 2400 e 2100 acessos (arredondados para centena). Já nos dias sem incentivo, 1400, 900, 1300, 1800, 700 e 1000 acessos. Os dados sustentam a hipótese?

Solução:

- X sem incentivo, Y com incentivo; o teste é $H_0 : F_X = F_Y$ versus $H_1 : F_X > F_Y$
- Combinando as amostras (com as de X sublinhadas):

$$\underline{700}, \underline{900}, \underline{1000}, 1200, \underline{1300}, \underline{1400}, 500, 1700, \underline{1800}, 2100, 2400, 2600$$

- A soma dos postos de X é $U = 1 + 2 + 3 + 5 + 6 + 9 = 26$
- Teste unilateral à esquerda com $n = 6$ e $m = 6$: na tabela, valor-$p \in (0.01, 0.025]$
- Há certa evidência que o incentivo aumenta os acessos, mas não é tão forte; há evidência da afirmação para um nível de significância > 0.025

Exemplo

Pings

Tempos de *ping* para dois locais diferentes são mostrados abaixo:

- Local I: 0.0156, 0.0210, 0.0215, 0.0280, 0.0308, 0.0327, 0.0335, 0.0350, 0.0355, 0.0396, 0.0419, 0.0437, 0.0480, 0.0483, 0.0543 segundos
- Local II: 0.0039, 0.0045, 0.0109, 0.0167, 0.0198, 0.0298, 0.0387, 0.0467, 0.0661, 0.0674, 0.0712, 0.0787 segundos

Há evidência que a mediana do tempo de ping depende do local?

Solução:

- Sejam X e Y tempo de ping nos locais; vamos testar $H_0 : F_X = F_Y$ versus $H_1 : F_X \neq F_Y$
- Combinando as amostras, a soma dos postos dos pings X é $U = 213$
- Amostras de tamanhos $n_1 = 15$ e $n_2 = 12$, podemos usar aproximação normal
- $\mathbf{E}(U|H_0) = \frac{n_1(n_1+n_2+1)}{2} = 210$; $\text{Std}(U|H_0) = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}} = \sqrt{420}$
- valor-p: $P = 2 \times \min(\mathbf{P}\{U \leq 213.5\}, \mathbf{P}\{U \geq 212.5\}) =$[a]
  $2 \times \min(\mathbf{P}\{Z \leq \frac{213.5-210}{\sqrt{420}}\}, \mathbf{P}\{Z \geq \frac{212.5-210}{\sqrt{420}}\}) = 2(1 - \Phi(0.12)) = 2(0.4522) = 0.9044$
- Não há evidência que os pings dos dois locais tenham distribuição diferente

[a] com correção de continuidade, de U discreta para distribuição contínua

Exemplo

Competição por contrato

Duas fabricantes de computador, A e B, disputam um contrato com uma grande empresa, ambas alegando que seus computadores são mais rápidos. Como a empresa pode escolher entre elas?

A empresa fez um teste: iniciou simultaneamente um mesmo programa em 7 computadores de cada empresa e anotou a ordem em que eles terminaram a execução (sem anotar o tempo real).

Os dois primeiros foram da empresa A, seguidos de 3 da empresa B, depois 5 da A e então 4 da B. Com esses dados, pode-se dizer que os fabricados por A são estocasticamente mais rápidos?

Solução:

- Não há dados numéricos, então faremos teste não paramétrico
- Testamos a hipótese nula $H_0 : F_B = F_B$ versus $H_1 : X_A$ estocasticamente menor que X_B, sendo X_A e X_B os tempos de execução e F_A e F_B suas distribuições acumuladas.
- Dos resultados, A A B B B A A A A A B B B B, vemos que A tem postos 1, 2, 6, 7, 8, 9, 10
- A soma dos postos é $U = 43$
- Teste unilateral à esquerda, com $n_1 = n_2 = 7$, e $U = 43$, o valor-p está entre 0.1 e 0.2
- Não há evidência significativa que computadores de A são (estocasticamente) mais rápidos

Teste da soma dos postos de Mann-Whitney-Wilcoxon

- Em Python

```
import scipy.stats as stats

# Para comparar duas variáveis
stats.ranksums(var1,var2)
```

- Exemplo (pings):

```
# dados das amostras
local1 = [0.0156, 0.0210, 0.0215, 0.0280, 0.0308, 0.0327, 0.0335, 0.0350, 0.0355,
          0.0396, 0.0419, 0.0437, 0.0480, 0.0483, 0.0543]
local2 = [0.0039, 0.0045, 0.0109, 0.0167, 0.0198, 0.0298, 0.0387, 0.0467, 0.0661,
          0.0674, 0.0712, 0.0787]
stats.ranksums(local1, local2)
```

Teste de Kruskal-Wallis

- O teste de Kruskal-Wallis (também chamado teste H) é um teste não paramétrico que usa postos de amostras aleatórias simples de três ou mais populações independentes para testar a hipótese nula de que as populações têm a mesma distribuição
- Quando for possível assumir que as distribuições têm mesma forma e escala, pode ser usado para testar a hipótese nula de mesma mediana

$$H_0 : M_1 = M_2 = \ldots = M_k$$

$$H_1 : \text{ao menos uma é diferente}$$

- O teste calcula uma estatística de teste H, cuja distribuição pode ser aproximada pela distribuição χ^2, desde que cada amostra tenha, no mínimo, 5 observações.

Teste de Kruskal-Wallis

- A estatística de teste depende da variância das somas dos postos R_i:

$$H = \frac{12}{N(N+1)}\left(\frac{R_1^2}{n_1} + \ldots + \frac{R_k^2}{n_k}\right) - 3(N+1)$$

- Se os postos são distribuídos igualmente entre os grupos amostrais, então H deve ser um número relativamente pequeno.
- Se as amostras são muito diferentes, então os postos serão excessivamente baixos em alguns grupos e altos em outros, com o efeito final de tornar H grande.
- Consequentemente, apenas valores grandes de H levarão à rejeição da hipótese nula de que as amostras provêm de populações com medianas iguais.
- Entretanto, o teste não indica qual(is) população(ões) difere(m) das demais
- No caso de Kruskal-Wallis indicar a hipótese alternativa, outros testes podem ser usados para agrupar as populações, como Tukey e Scott-Knott

Teste de Kruskal-Wallis

- Em Python

```
import scipy.stats as stats

# Para comparar várias variáveis
stats.kruskal(var1,var2,var3)
```