ANADI

Análise de dados em Informática

Aulas Teóricas - Testes de Hipóteses Não Paramétricos

Ana Madureira, João Matos

Instituto Superior de Engenharia do Porto

Ano letivo 2023/2024

Introdução

- Os Testes não paramétricos requerem menos pressupostos relativamente à população;
  - Não exigem a normalidade nem se baseiam em parâmetros da distribuição.
- Baseiam-se nas estatísticas de ordem
- São especialmente úteis em variáveis ordinais
- Geralmente menos eficientes do que os testes paramétricos

Teste dos sinais para a mediana η de uma população

- Os dados devem ser no mínimo ordinais.
- Caso a distribuição da população seja simétrica também serve para localizar a média ($\mu = \eta$) .

Tem-se

$$H_0 : \ \eta = \eta_0 \text{ vs } H_1 \ : \quad \eta \neq \eta_0 \text{ ou } \eta > \eta_0 \text{ ou } \eta < \eta_0$$

- Dada uma amostra $X = (X_1, X_2, \ldots, X_n)$ a estatística teste é definida por

$$\boxed{T(X) = \text{ número de observações } X_i \text{ abaixo (ou acima) de } \eta_0}$$

que, sob H_0, segue uma distribuição binomial $Bi(n, p = 0.5)$.

- Sempre que na amostra se observarem valores iguais a η_0 não os consideramos o que implica a diminuição do tamanho da amostra.
- Para valores de $n \geq 30$ deve-se considerar a estatística

$$Z(X) = \frac{\widehat{P} - 0.5}{\frac{\sqrt{0.25}}{n}} \sim N(0,1)$$

onde $\widehat{P} = T(X)/n$.

- O teste do sinal é particularmente importante quando temos duas amostras emparelhadas $(X_i, Y_i),\ i = 1, 2, \ldots, n$ cujas escalas de medida é pelo menos ordinal. Neste caso faz-se um teste de sinal às diferenças $D_i = X_i - Y_i,\ i = 1, 2, \ldots, n$.
- Também se pode aplicar o teste do sinal para testar se as médias (ou medianas) das duas variáveis observadas são iguais desde que se suponha que as diferenças $D_i = X_i - Y_i,\ i = 1, 2, \ldots, n$ são i.i.d. e possuem distribuição simétrica, relativamente a $\mu = 0$.

Teste de Wilcoxon (à mediana de uma população)

- No teste do sinal os dados são transformados em sinais "+" caso estejam acima ou sinais "-" caso estejam abaixo de η_0
- No teste de Wilcoxson além de considerarmos os sinais também consideramos as diferenças entre os dados observados e η_0. Contudo temos que pressupor que a distribuição além de contínua é simétrica:
  - Inspeção do histograma e/ou calcular o coeficiente de assimetria de Pearson (função scipy.stats.skew() no Python)
    Regra prática:

$$\begin{aligned} |skewness| < 0.1 &\longleftarrow \text{Distribuição simétrica} \\ 0.1 < |skewness| < 1 &\longleftarrow \text{Distribuição moderadamente assimétrica} \\ |skewness| > 1 &\longleftarrow \text{Distribuição fortemente assimétrica} \end{aligned}$$

- Analogamente ao teste do sinal também é possível comparar medianas de duas amostras emparelhadas.

Dada uma amostra, $X = (x_1, x_2, \ldots, x_n)$, calculamos o valor observado da estatística teste da seguinte forma:

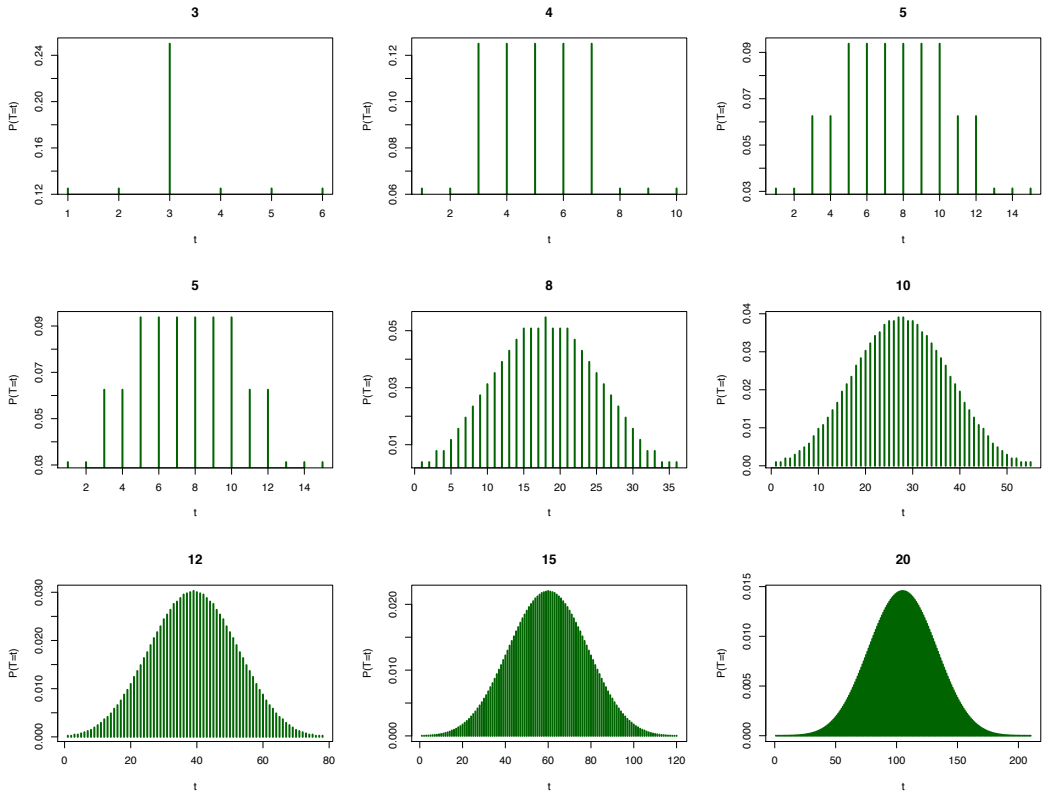
1. Calculam-se as diferenças $d_i = x_i - \eta_0$, $i = 1, 2, \ldots, n$ e ordenam-se de forma crescente os valores absolutos $|d_i|$ (supôr que não há empates)
2. A cada $|d_i|$, atribui-se um número de ordem n_i e um o sinal:

$$\begin{cases} + & \text{se } d_i > 0 \\ - & \text{se } d_i < 0 \end{cases}$$

3. Calculamos o valor da estatística que resulta da soma dos números de ordem, n_i, com argumento sinal "+", denotamos esta estatística por T_+. Calculamos o valor da estatística que resulta da soma dos números de ordem, n_i, com argumento sinal "−", denotamos esta estatística por T_-.
   - Se houver empates atribui-se o valor médio do número de ordem ocupado pelas observações.
   - Tem-se sempre $\sum n_i = \frac{n(n+1)}{2}$. Deste modo se H_0 for verdadeira as distribuições de T_+ e T_- são idênticas e simétricas em torno de $\frac{n(n+1)}{4}$ é pois indiferente usar T_+ ou T_-.
4. A estatística T_+ ou T_- segue uma distribuição com o mesmo nome do teste (Distribution of the Wilcoxon signed rank statistic). Para amostras superiores a 20 usa-se uma aproximação à distribuição normal.

$$\frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0,1).$$

Distribuição da estatística do teste de Wilcoxon para vários valores de n



Exemplo 1:

Pretende-se verificar se a média das notas nacionais a Matemática é superior a 63%. Escolheram-se aleatoriamente 10 alunos e registrou-se as respectivas notas: 66%,69%,40%,64%,67%,65%,82%,54%,70% e 74%. Assume-se que as notas obtidas são simétricas.

- Resolução com o Python:

```
from numpy import array
from scipy.stats import wilcoxon
notas = [66,69,40,64,67,65,82,54,70,74]
statistic, p_value = wilcoxon(array(notas) - 63, alternative='greater')
print(f'Test Statistic: {statistic}')
print(f'P-value: {p_value}')
Test Statistic: 38.0
P-value: 0.1611328125
```

- Decisão: Como o p-value é maior que α não se rejeita H_0.
- Note que tem-se de subtrair a hipótese nula aos dados

Exemplo 2 (Teste de Wilcoxon para duas amostras emparelhadas)

- Foi realizado um estudo sobre o efeito do álcool no tempo de reação. Solicitou-se a dez participantes que assistam a um vídeo e que pressionem um botão sempre que vissem um pequeno círculo vermelho. Passada uma semana os mesmos dez participantes repetiram a tarefa mas tomaram uma bebidas contendo 2 unidades de álcool. Pretende-se efectuar um teste de hipóteses para decidir se o álcool tem efeito no tempo de reação. Os dados obtidos estão na seguinte tabela.

Part.	1	2	3	4	5	6	7	8	9	10
Sem Álcool	30	20	19	20	19	25	23	39	18	24
Com Álcool	31	27	20	20	34	24	21	23	23	26

Resolução:

1. Seja x_i os tempos de reação sem álcool e y_i os tempos de reação com álcool, $i = 1, 2, \ldots, 10$. Consideramos as diferenças $d_i = x_i - y_i$, $i = 1, 2, \ldots, 10$.
2. Temos agora um teste de Wilcoxon a uma amostra $(d_1, d_2, \ldots, d_n)$ com,

$$H_0 : \ \eta = 0 \ vs \ H_1 \ \eta \neq 0$$

onde η é a mediana da v.a. $D = X - Y$.

Exemplo 2 (cont.)

No Python teriamos:

```
from scipy.stats import wilcoxon
antes=[30,20,19,20,19,25,23,39,18,24]
depois=[31,27,20,20,34,24,21,23,23,26]
statistic, p_value = wilcoxon(antes,depois)
print(f'Test statistic: {statistic}')
print(f'P-value: {p_value}')
Test statistic: 15.5
P-value: 0.40487306185858307
```

- Analisando o p-value (0.405) não temos evidências estatísticas para rejeitar H_0.

Teste de Mann-Whitney-U (Wilcoxon para duas amostras independentes)

- Este teste compara as medianas η_1 e η_2 de duas populações contínuas P_1 e P_2 com a mesma forma.
- Tem-se:

$$H_0 \ : \ \eta_1 = \eta_2 \text{ vs } \quad \begin{array}{l} H_1 \ : \ \eta_1 = \eta_2, \text{ ou} \\ H_1 \ : \ \eta_1 > \eta_2, \text{ ou} \\ H_1 \ : \ \eta_1 < \eta_2 \end{array}$$

- Supondo que se retira uma amostra aleatória de tamanho n_1 da população P_1 e uma amostra aleatória de tamanho n_2 da população P_2. Supondo $n_1 \geq n_2$, calculamos o valor observado da estatística teste na seguinte forma:
  1. Colocam-se em ordem crescente o conjunto de todas as $n = n_1 + n_2$ observações, e, atribui-se um número de ordem.
  2. O valor da estatística teste observada corresponde à soma, T, dos números de ordem da amostra com menor tamanho.
- Supondo H_0 verdadeira e n_1, n_2 grandes, usa-se a aproximação à distribuição normal

$$\frac{T - n_2(n+1)/2}{\sqrt{n_1 n_2 (n+1)/12}} \sim N(0,1)$$

- No Python usar a função scipy.stats.mannwhitneyu()

Teste de Friedman

- Supor que dispomos de $n \times k$ observações de uma v.a. quantitativa avaliada por n indivíduos sujeitos a k tratamentos de um determinado factor (variável qualitativa)

	Factor			
	Tratamento 1	Tratamento 2	...	Tratamento k
indivíduo 1	$x_{\mathbf{1},\mathbf{1}}$	$x_{\mathbf{1},\mathbf{2}}$	...	$x_{\mathbf{1},k}$
indivíduo 2	$x_{\mathbf{2},\mathbf{1}}$	$x_{\mathbf{2},\mathbf{2}}$	...	$x_{\mathbf{2},k}$
⋮	⋮	⋮		⋮
indivíduo n	$x_{n,\mathbf{1}}$	$x_{n,\mathbf{2}}$	...	$x_{n,k}$

- O teste de Friedman tem como hipóteses:

$$H_0 \ : \ \text{A distribuição dos } k \text{ tratamentos é a mesma}$$
$$H_1 \ : \ \text{A distribuição dos } k \text{ tratamentos é diferente}$$

Ou, em particular considerando η_i a mediana da população X_i, $i = 1, \ldots, k$, podemos testar

$$H_0 \ : \ \eta_1 = \eta_2 = \ldots = \eta_k \quad \text{vs} \quad H_1 \ : \ \exists i \neq j \ \ \eta_i \neq \eta_j$$

- Este teste pode ser visto como uma extensão do teste de Wilcoxon para duas amostras emparelhadas.

- Para se obter a estatística teste procede-se do seguinte modo:
  1. Em cada linha i ($i = 1, 2 \ldots, n$) atribui-se o número de ordem $R_{i,j}$, $j = 1, 2, \ldots k$. Em caso de empates procedemos da mesma forma do teste de Wilcoxon.
  2. Calcular $T_j = \sum_{i=1} R_{i,j}$ $j = 1, 2, \ldots, k$
  3. O valor da estatística observada é dada por

$$T = \frac{12}{n.k.(k+1)} \sum_{j=1}^{k} T_j^2 - 3n(k+1) \sim \chi^2_{k-1}$$

- Tem-se p-value $= P(T \geq T_{obs} \mid H_0)$
- Em Python usa-se a função scipy.stats.friedmanchisquare()

Teste de Kruskal-Wallis

- Este teste é a alternativa não paramétrica do teste One-Way ANOVA. Deve-se utilizar quando a hipótese da normalidade for rejeitada ou se o tamanho das amostras forem pequenas.
- Supor que dispomos de k amostras independentes:

$$(x_{1,1}, x_{1,2}, \ldots x_{1,n_1}), \ldots, (x_{k,1}, x_{k,2}, \ldots x_{k,n_k})$$

  retiradas aleatoriamente de k populações com distribuições continuas e mesma forma
- O objectivo do teste de Kruskal-Wallis é testar se uma dada variável qualitativa, designada de factor tem efeitos iguais sobre uma determinada variável quantitativa.
- Em particular, denotando por η_i a mediana da i-ésima população (i=1,2,. . . , k) temos as hipóteses

$$H_0 \ : \ \eta_1 = \eta_2 = \ldots = \eta_k \quad \text{vs} \quad H_1 \ : \ \exists i \neq j \ \ \eta_i \neq \eta_j$$

- O valor da estatística teste calcula-se da seguinte forma:
  1. atribuir números de ordem R_i, j $(i = 1, 2 \ldots k$ e $j = 1, 2 \ldots n_i)$ à amostra conjunta das $n = \sum_i^k n_i$ observações. Os empates são tratados como nos testes anteriores.
  2. calcular, para cada i, $T_i = \sum_{j=1}^{n_i} R_{i,j}$, $i = 1, 2, \ldots, k$
  3. A estatística teste é dada por

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1) \sim \chi^2_{(k-1)}$$

Exemplo:

A seguinte tabela contém os dados relativos ao número de vezes que três grupos de indivíduos, G_1, G_2 e G_3 vai ao cinema durante um mês. Será que os grupos apresentam a mesma distribuição?

G_1	20	4	7	2	17	3
G_2	12	21	9	0	14	1
G_3	8	22	10	5	6	20

Exemplo: (cont.)

Resolução: Tem-se:

$$H_0 \ : \ \text{Os grupos apresentam a mesma distribuição}$$
$$\text{vs}$$
$$H_1 \ : \ \text{Os grupos não apresentam a mesma distribuição}$$

Comandos no Python:

```
import scipy.stats as stats
G1=[20, 4, 7, 2, 17, 3]
G2=[12, 21, 9, 0, 14, 1]
G3=[8, 22, 10, 5, 6, 20]
res=stats.kruskal(G1,G2,G3)
print('valor de prova:',round(res.pvalue,4))
valor de prova: 0.6437
```

Decisão: O p-value leva-nos a não rejeitar H_0. Não existem diferenças significativas entre os grupos.

Testes de ajustamento

- São testes para averiguar se uma dada amostra pode ser considerada como sendo proveniente de uma certa distribuição teórica
- Têm especial interesse os testes de ajustamento à distribuição normal (Frequentemente, é um pressuposto para se usar um teste paramétrico)
- Investigar se duas amostras podem ser consideradas provenientes de uma distribuição comum
- Iremos começar com o teste de ajuste do χ^2 cujo procedimento é semelhante ao visto no teste a duas proporções efectuado pela função stats.chisquare()

Teste de ajuste do χ^2

- Baseia-se na comparação da distribuição empírica dos dados $X = (X_1, X_2, \ldots X_n)$ com a distribuição teórica à qual se suspeita que a amostra é proveniente.
- Tem-se:

$$H_0 \ : \ \text{A população possui a distribuição } F(x) \text{ vs } H_1 \ : \ \text{A população não possui a distribuição } F(x)$$

- Agrupa-se as observações em k classes (intervalos no caso da distribuição ser contínua)
- Calculam-se as frequências absolutas n_i, $i = 1, 2, \ldots, k$ das observações em cada classe ($n = \sum_{i=1}^{k} n_i$).
- Determina-se as frequências (teóricas) esperadas em cada classe $n.p_i$, onde p_i é, supondo H_0 verdadeira, a probabilidade de a variável aleatória pertencer à i-ésima classe
- A estatística teste é:

$$\mathrm{T(X)} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{(k-1)-r}$$

onde r representa o número de parâmetros da distribuição população estimados usando a amostra.

- A amostra deve ser grande ($n \geq 30$) e $n_i p_i \geq 5, \ \forall i \in \{1, 2, \ldots k\}$.

Teste de Kolmogorov-Smirnov (K-S)

- O teste K-S tem por base o ajuste entre o distribuição empírica $S_n(x)$ e a distribuição teórica $F_0(x)$
- Tem-se:

$$H_0 \ : \ \text{A função de distribuição da população } F(x) \text{ é igual a } F_0(x)$$
$$\text{vs}$$
$$H_0 \ : \ F(x) \neq F_0(x) \text{ Para algum valor de x}$$

- A estatística teste é

$$T = \sup|S_n(x) - F_0(x)|$$

que tem função distribuição conhecida.

Exemplo:

Geraram-se três amostras aleatórias A_1, A_2 e A_3 com 1000 elementos cada onde:

- A_1 foi retirada de uma população com distribuição $N(2,1)$
- A_2 foi retirada de uma população com distribuição $\chi^2_{(4)}$
- A_3 foi retirada de uma população com distribuição $T_{(30)}$

Use o teste K-S para verificar se alguma das populações tem distribuição normal com $\mu = 0$ e $\sigma = 1$.

```
import numpy as np
import scipy.stats as stats
np.random.seed(42)
A1 = np.random.normal(0,2,500)
A2 = np.random.chisquare(1,500)
A3 = np.random.standard_t(30,500)
print(np.mean(A1))
print(np.mean(A2))
print(np.mean(A3))
mean=0; std_dev=1,
r1=stats.kstest(A1,'norm',args=(mean,std_dev))
print('p-value do teste com A1:',round(r1.pvalue,4))
r2=stats.kstest(A2,'norm',args=(mean,std_dev))
print('p-value do teste com A2:',round(r2.pvalue,4))
r3=stats.kstest(A3,'norm',args=(mean,std_dev))
print('p-value do teste com A2:',round(r3.pvalue,4))
p-value do teste com A1: 0.0
p-value do teste com A2: 0.0
p-value do teste com A2: 0.3848
```

Conclusões?

Teste de Lilliefors

- O teste de K-S é efectuado para uma função de distribuição F(x) específica.
- Lilliefors fez uma correcção ao teste de K-S de modo a testar, independentemente dos valores da média e do desvio padrão se a amostra é proveniente de uma distribuição normal ou não.

Exemplo:

```
import numpy as np
from statsmodels.stats.diagnostic import lilliefors
np.random.seed(42)
A1 = np.random.normal(0,2,500)
res=lilliefors(A1)
print(res)
print('p-value (Lilliefors):',round(res[1],4))
```

Teste de Shapiro-Wilk

- Teste para verificar se a variável aleatória X, da qual foi retirada a amostra aleatória $(X_1, X_2, \ldots, X_n)$, segue uma distribuição normal (idêntico ao teste de Lilliefors)
- Tem-se:

  H_0 : X segue uma distribuição normal vs H_1 : X não segue uma distribuição normal

- Seja $(X_{(1)}, X_{(2)}, \ldots, X_{(2)})$ a amostra colocada por ordem crescente. Então a estatística teste é

$$T(X) = \frac{\sum_{i=1}^{n} a_i X_{(i)}}{\sum_{i=1}^{n} \left(X_{(i)} - \overline{X}\right)^2}$$

  onde,

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}$$

  sendo,
  - m^T o vector transposto composto pelos valores esperados das estatística de ordem de v.a. i.i.d. provenientes de uma distribuição normal reduzida
  - V a respectiva matriz das covariâncias.
- Sob H_0 verdadeiro T(X) tem uma distribuição conhecida

- Geralmente efectua-se o teste de Lilliefors para amostras grandes ($n \geq 30$) enquanto para amostras de dimensão mais reduzida é mais indicado efetuar o teste de Shapiro-Wilk.

Exemplo:

```
import numpy as np
from scipy.stats import shapiro
np.random.seed(42)
A2 = np.random.chisquare(1,500)
res2=shapiro(A2)
print('p-value (Shapiro):',res2[1])
p-value (Shapiro): 2.92488964446179e-29
```