INF 222 – Computação Experimental
Testes não-paramétricos

Nome: Danilo Freitas Vieira
Matrícula: 108201

Nome: Yuri Cardoso Bragine
Matrícula: 108199

1. No exercício 1 foram dadas as idades das melhores atrizes e dos melhores atores premiados com o Oscar. Uma das perguntas foi: “Existe uma tendência das melhores atrizes serem mais jovens que os melhores atores?” Na época, isto foi respondido com base em medidas de centro, como média e mediana. Agora pode ser respondido com testes de hipótese, mais especificamente com testes não paramétricos.

a) Selecione uma amostra (ao acaso) de tamanho apropriado das idades dos melhores atores e verifique a hipótese nula que a mediana da população dos atores é igual à 33.5, a mediana da idade das atrizes, contra a hipótese alternativa que é maior. Use o teste dos sinais e o valor-p. Comente o resultado.

Partindo dos dados

Idade dos Melhores Atores:

44 41 62 52 41 34 34 52 41 37 38 34 32 40 43

56 41 39 49 57 41 38 42 52 51 35 30 39 41 44

49 35 47 31 47 37 57 42 45 42 44 62 43 42 48

49 56 38 60 30 40 42 36 76 39 53 45 36 62 43

51 32 42 54 52 37 38 32 45 60 46 40 36 47 29 43

Idade das Melhores Atrizes

22 37 28 63 32 26 31 27 27 28 30 26 29 24 38

25 29 41 30 35 35 33 29 38 54 24 25 46 41 28

40 39 29 27 31 38 29 25 35 60 43 35 34 34 17

37 42 41 36 32 41 33 31 74 33 50 38 61 21 41

26 80 42 19 33 35 45 49 39 34 26 25 33 35 35 28

Sendo h0 a hipótese de que a mediana da idade da população dos atores é igual à 33.5 e h1 sendo a hipótese de que é maior faremos o teste dos sinais. Selecionando a amostra [56, 41, 39, 49, 57, 41, 38, 42, 52, 51, 35, 30, 39, 41, 44, 49, 35, 47, 31, 47, 37, 57, 42, 45,

42, 44, 62, 43, 42, 48] temos 28 idades maiores que 33.5. Ou seja, encontraremos valor-p = P(S >= 28) usando a distribuição normal, com n = 30, μ = 30/2 = 15 e σ = √30/2 = 2.73861.

Valor-p = P(Z >= (27.5 – 15)/2.73861) = 1 – Φ(4.54) ≈ 0.0

Portanto se rejeita H0 dado que o valor p é pequeno

Fazendo a confirmação com Python:

```
from statsmodels.stats.descriptivestats import sign_test

amostra = [56, 41, 39, 49, 57, 41, 38, 42, 52, 51, 35, 30, 39, 41, 44,
49, 35, 47, 31, 47, 37, 57, 42, 45, 42, 44, 62, 43, 42, 48]

valorp = sign_test(amostra, mu0=33.5)

print(valorp) # valorp = 8.67992639541626e-07, ou seja, praticamente 0
```

b) Faça o mesmo, com os mesmos dados, mas usando o valor-p do teste dos postos sinalizados de Wilcoxon. Comente eventuais diferenças em relação ao resultado anterior.

Usando o Python para a realização do teste:

```
from scipy.stats import wilcoxon

amostra = [56, 41, 39, 49, 57, 41, 38, 42, 52, 51, 35, 30, 39, 41, 44,
49, 35, 47, 31, 47, 37, 57, 42, 45, 42, 44, 62, 43, 42, 48]

amostra = [i - 33.5 for i in amostra] # Normaliza o vetor

resp = wilcoxon(amostra)

print(resp.pvalue)   #valor-p   =   4.6566128730773926e-08,   ou   seja,
praticamente 0
```

Notamos que esse teste é um pouco mais preciso, já que além de levar em conta apenas a informação se o dado está acima ou abaixo da mediana, também leva em conta as posições relativas de cada um. Vemos uma pequena diferença no valor-p, que tem uma precisão maior nesse teste.

c) Selecione duas amostras ao acaso, uma das idades das atrizes e outra das idades dos atores, e use o teste da soma dos postos para responder à pergunta do exercício 1 citada acima. Escreva as hipóteses nulas e alternativas que está verificando, apresente e comente o resultado.

Fx: Atrizes, Fy: Atores

H0: A distribuição das idades das atrizes é igual a dos atores (Fx = Fy)

H1: A distribuição das idades das atrizes é menor que dos atores (Fx < Fy)

Utilizando o teste em Python:

```
from scipy.stats import mannwhitneyu

amostraHomens=[44,41,62,52,41,34,42,52,51,35,30,39,35,47,31,47,37,57,39
,53,45,36,62,43,60,46,40,36]

amostraMulheres=[28,63,32,26,31,29,38,54,24,25,60,43,35,34,34,41,36,32,
41,33,39,34,26,25,33]

resp = mannwhitneyu(x=amostraMulheres, y=amostraHomens,
alternative="less")

print(resp.pvalue) # Valor-p = 0.0003634027697788937
```

Como o valor-p é muito baixo, temos evidências suficientes para rejeitar H0 em favor de H1, ou seja, temos evidências de que as idades das melhores atrizes tendem a ser menores que as idades dos melhores atores. O que responde à pergunta do enunciado.

2. A planilha Ex-08-dados contém dados da UNICEF sobre mortalidade infantil: número de mortes por 1000 crianças de 1 a 4 anos. Nesta questão, considere os dados de 2018.

a) Selecione ao acaso 10 países da América do Sul e verifique a hipótese que a mediana do índice de mortalidade infantil apresentado é maior que 2 em 1000. Informe qual o teste feito, como foi feito e comente o resultado.

H0: A mediana da mortalidade infantil dos países da América do Sul é igual a 2
H1: A mediana da mortalidade infantil dos países da América do Sul é maior que 2
Utilizando o Teste dos Sinais faremos com os dados a seguir:

País	2018
Argentina	1,08
Bolívia	5,20
Brasil	1,64
Chile	1,05
Colômbia	1,99
Equador	1,96
Guiana	5,17
Paraguai	2,99
Uruguai	1,02
Venezuela	3,23

A amostra é pequena, apenas 10 países, portanto é necessário seguir a distribuição binomial em que n=10 e p = 0.5 e 4 países possuem taxa de mortalidade maior que 2. Portanto, na fórmula de valor-p = P(S >= 4) = P(S <= 6) = 0.828

Como o valor-p é grande, não rejeitamos h0, logo, não podemos afirmar que a mediana da amostra é diferente de 2.

b) Faça o mesmo considerando países da Europa.
H0: A mediana da mortalidade infantil dos países da Europa é igual a 2

H1: A mediana da mortalidade infantil dos países da Europa é maior que 2
Utilizando o Teste dos Sinais faremos com os dados a seguir:

País	2018
Bélgica	0,78
Croácia	0,70
Dinamarca	0,54
Espanha	0,55
Estônia	0,53
Grécia	0,51
Holanda	0,58
Itália	0,48
Letônia	0,61
Lituânia	0,74

A amostra é pequena, apenas 10 países, portanto é necessário seguir a distribuição binomial em que n=10 e p = 0.5 e 4 países possuem taxa de mortalidade maior que 2. Portanto, na fórmula de valor-p = P(S >= 0) = P(S <= 10) = 1

Como o valor-p é grande, não rejeitamos h0, logo, não podemos afirmar que a mediana da amostra é diferente de 2.

c) Com as amostras de tamanho 10 selecionadas, faça um teste para verificar a hipótese que a mortalidade é menor na Europa que na América do Sul. Informe os dados, o teste e o resultado.

Podemos utilizar o teste da soma dos postos para a comparação das duas amostras:

Fx = Europa

Fy = América do Sul

H0: A mediana da mortalidade infantil dos países da Europa é igual à da América do Sul (Fx = Fy)

H1: A mediana da mortalidade infantil dos países da Europa é menor que na América do Sul (Fx < Fy)

```
from scipy.stats import mannwhitneyu

amostraEuropa=[0.78,0.70,0.54,0.55,0.53,0.51,0.58,0.48,0.61,0.74,0.52,0
.45]
amostraAmerica = [1.08, 5.20,1.64,1.05,1.99,1.96,5.17,2.99,1.02,3.23]

resp=mannwhitneyu(x=amostraEuropa,y=amostraAmerica,alternative="less")

print(resp.pvalue) #P-valor = 4.3669632941789134e-05, ou seja, quase 0
```

Como o valor-p é muito baixo, temos dados suficientes para rejeitar H0 em favor de H1, ou seja, aceitamos a hipótese de que a mediana da mortalidade infantil dos países da Europa é menor que na América do Sul

d) Faça o mesmo que a letra (c), mas para África e América do Sul.

País	2018
África do Sul	7,15
Botsuana	9,75
Angola	27,04
Argélia	3,45
Camarões	27,40
Congo	13,34
Egito	3,12
Costa do Marfim	23,39
Etiópia	15,81
Gana	13,57

Podemos utilizar o teste da soma dos postos para a comparação das duas amostras:

Fx = América do Sul

Fy = África

H0: A mediana da mortalidade infantil dos países da América do Sul é igual à da África (Fx = Fy)

H1: A mediana da mortalidade infantil dos países da América do Sul é menor que na África (Fx < Fy)

```
from scipy.stats import mannwhitneyu

amostraAmerica = [1.08, 5.20,1.64,1.05,1.99,1.96,5.17,2.99,1.02,3.23]
amostraAfrica=[7.15,9.75,27.04,3.45,27.40,13.34,3.12,23.39,15.81,13.57]

resp=mannwhitneyu(x=amostraAmerica,y=amostraAfrica,alternative="less")
print(resp.pvalue) #P-valor = 0.00038426945658138323
```

Como o valor-p é muito baixo, temos dados suficientes para rejeitar H0 em favor de H1, ou seja, aceitamos a hipótese de que a mediana da mortalidade infantil dos países da América do Sul é menor que na África.

3. Verifique se houve redução no índice de mortalidade de 2010 para 2018, das seguintes formas:

a) Teste pareado: selecione ao acaso alguns países e colete seus dados nos anos de 2010 e 2018.

País	2018	2010
Alemanha	0,59	0,70
Filipinas	5,92	7,10
Finlândia	0,44	0,55
México	2,07	2,81
Marrocos	2,95	4,50

A partir disso, podemos testar se a média das diferenças entre as taxas de mortalidade infantil nos dois anos para cada país é nula ou não para verificarmos se houve ou não redução.

Dados = {-0.11, -1.18, -0.11, -0.74, -1.55}

H0: A média das diferenças é nula
H1: A média das diferenças é menor que 0

Variância = 0.0121 + 1.3924 + 0.0121 + 0.5476 + 2.4025 = 4.3667 / 5 = 0.87334
Desvio Padrão(s) = 0.9345
Média amostral(x) = -0.738

Descobrindo o valor-p:

t0 = $(x - u0)/(s/\sqrt{n})$

t0 = (-0.738 - 0)/(0.9345 / $\sqrt{5}$**)**
t0 = -1.7658
Valor-P = P{T <= -1.7658} = P{T >= 1.7658} com 4 graus de liberdade
Para 4 graus de liberdade, 1.7658 é um valor que indica um valor-p entre 0.1 e 0.05, logo, é um valor-p em que já podemos rejeitar H0 em favor de H1, com uma confiança de 90% a 95%. Portanto, podemos dizer que a média das diferenças é menor que 0, ou seja, de 2010 para 2018 houve uma certa redução nas taxas de mortalidade infantil.

b) Teste não-pareado: selecione ao acaso vários dados de 2010 e outros de 2018. Certifique-se que a amostra seja aleatória e independente (sem escolher os países por conveniência).

País	2018	2010
Chile	1,05	1,29
Alemanha		0,70
China	2,10	3,26
Lituânia	0,74	1,17
Brasil	1,64	2,02
Tuvalu		5,44
Canadá	0,65	0,77
Catar	0,96	1,32
Grécia		0,48
Argentina	1,08	1,56

A partir disso, temos:
Dados 2010 = {1.29, 0.70, 3.26, 1.17, 2.02, 5.44, 0.77, 1.32, 0.48, 1.56}
Dados 2018 = {1.05, 2.10, 0.74,1.64, 0.65, 0.96, 1.08}

Média 2010 (x1) = 1.801
Média 2018 (x2) = 1.1743

n1 = 10
n2 = 7
$s1^2$ (Variância de 2010) = (1.6641 + 0.49 + 10.6276 + 1.3689 + 4.0804 + 29.5936 + 0.5929 + 1.7424 + 0.2304 + 2.4336)/10 = 52.8239/10 = 5.28239
$S2^2$ (Variância de 2018) = (1.1025 + 4.41 + 0.5476 + 2.6896 + 0.4225 + 0.9216 + 1.1664)/7 = 11.2602/7 = 1,6086

H0: A média das diferenças é nula
H1: A média das diferenças é menor que 0

t0 = $(x1 - x2 - u0)/(\sqrt{((s1^2/n1) + (s2^2/n2))})$
t0 = (1.801 - 1.1743)/$(\sqrt{((5.28239/10) + (1,6086/7))})$ = 0.7198

Valor-P = P{T < 0.7198}

Para qualquer que seja o número de graus de liberdade, esse valor não está na tabela, logo, o valor-P é maior que 0.1, ou seja, não podemos rejeitar H0. Assim, não há evidências suficientes para constatarmos que as taxas de mortalidade infantil diminuíram com