CC0291 - Estatística Não Paramétrica

Correlação de Spearman - 18/04/2023.

Prof. Maurício

1. Vamos enunciar o exercício 16 ,página 525, do Mood: Uma medida de associação de duas variáveis aleatórias X e Y é a correlação de postos ou correlação de Spearman. Os valores de X são substituídos por seus postos e os correspondentes valores de Y também o são. Por exemplo, considere amostra de tamanho 5.

X	20,4	19,7	21,8	20,1	20,7
Y	9,2	8,9	11,4	9,4	10,3

Esta tabela é trocada por:

RX=R	3	1	5	2	4
RY=S	2	1	5	3	4

Sejam $r(X_i) = R_i$ posto de X_i e $r(Y_i) = S_i$ posto de Y_i. Agora use estes pares de postos e calcule o coeficiente de correlação simples entre RX e RY.

Seja r o coeficiente de correlação de Spearman ele é calculado como:

$$r = \frac{\sum_{i=1}^{n} \left[(R_i - \bar{R})(S_i - \bar{S}\right]}{\sqrt{\sum_{i=1}^{n} \left[(R_i - \bar{R})^2\right] \sum_{i=1}^{n} \left[(S_i - \bar{S})^2\right]}}$$

com

$$\bar{R} = \bar{S} = \frac{\sum_{i=1}^{n} R_i}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n+1}{2}$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} S_i^2 = \frac{n(n+1)(2n+1)}{6}.$$

Além disso:

$$\sum_{i=1}^{n} \left[(R_i - \bar{R})^2\right] = \sum_{i=1}^{n} R_i^2 - n\bar{R}^2 = \frac{n(n+1)(2n+1)}{6} - n\frac{(n+1)^2}{4}$$

$$= n(n+1)\left[\frac{2n+1}{6} - \frac{n+1}{4}\right] = n(n+1) \times \frac{n-1}{12}$$

$$= \frac{(n-1)n(n+1)}{12} = \frac{n(n^2-1)}{12} = \frac{n^3 - n}{12}.$$

Note que:

$$\sum_{i=1}^{n}\left[(S_i-\bar{S})^2\right]=\sum_{i=1}^{n}\left[(R_i-\bar{R})^2\right].$$

$$\sum_{i=1}^{n}\left[(R_i-\bar{R})(S_i-\bar{S})\right]=\sum_{i=1}^{n}R_i\,S_i-n\bar{R}\bar{S}$$

$$=\sum_{i=1}^{n}R_i\,S_i-n\;\frac{n+1}{2}\times\frac{n+1}{2}=\sum_{i=1}^{n}R_i\,S_i-\frac{n(n+1)^2}{4}$$

$$\sum_{i=1}^{n}\left[(R_i-\bar{R})(S_i-\bar{S})\right]=\frac{4\sum_{i=1}^{n}R_i\,S_i-n(n+1)^2}{4}.$$

Mostre que o coeficiente de correlação de Spearman pode ser dado por:

$$r=\frac{12\sum_{i=1}^{n}R_i\,S_i-3n(n+1)^2}{n(n^2-1)}.$$

Prova:

Considere

$$NUM=\sum_{i=1}^{n}\left[(R_i-\bar{R})(S_i-\bar{S})\right]=\frac{4\sum_{i=1}^{n}R_i\,S_i-n(n+1)^2}{4}$$

$$DEN=\sqrt{\sum_{i=1}^{n}\left[(R_i-\bar{R})^2\right]\;\sum_{i=1}^{n}\left[(S_i-\bar{S})^2\right]}=\frac{n(n^2-1)}{12}$$

$$r=\frac{NUM}{DEN}=\frac{4\sum_{i=1}^{n}R_i\,S_i-n(n+1)^2}{4}\times\frac{12}{n(n^2-1)}.$$

$$r=\frac{12\sum_{i=1}^{n}R_i\,S_i-3n(n+1)^2}{n(n^2-1)}.$$

Uma forma mais operacional pode ser dada como:

$$r=1-\quad\frac{6\;\sum_{i=1}^{n}D_i^2}{n^3-n},$$

em que

$$D_i = R_i - S_i, i = 1, 2, \ldots, n.$$

Prova:

Note que:

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (R_i - S_i)^2 = \sum_{i=1}^{n} R_i^2 + \sum_{i=1}^{n} S_i^2 - 2\sum_{i=1}^{n} R_i S_i$$

$$\sum_{i=1}^{n} D_i^2 = 2\sum_{i=1}^{n} R_i^2 - 2\sum_{i=1}^{n} R_i S_i$$

$$\sum_{i=1}^{n} D_i^2 = \frac{n(n+1)(2n+1)}{3} - 2\sum_{i=1}^{n} R_i S_i$$

$$2\sum_{i=1}^{n} R_i S_i = \frac{n(n+1)(2n+1)}{3} - \sum_{i=1}^{n} D_i^2$$

$$12\sum_{i=1}^{n} R_i S_i = 2n(n+1)(2n+1) - 6\sum_{i=1}^{n} D_i^2$$

$$r = \frac{2n(n+1)(2n+1) - 6\sum_{i=1}^{n} D_i^2 - 3n(n+1)^2}{n(n^2-1)}$$

Mas

$$2n(n+1)(2n+1) - 3n(n+1)^2 = n(n+1)(4n+2-3n-3) = n(n+1)(n-1) = n(n^2-1).$$

Assim,

$$r = \frac{n(n^2-1) - 6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}.$$

Exemplo 1: Resolva a questão do Mood usando o **R**:

Calcule a correlação de Spearman:

```
>
> X=c(20.4,19.7,21.8,20.1,20.7)
>
>
> Y=c(9.2,8.9,11.4,9.4,10.3)
```

```
>
>
> n=length(X);n
[1] 5
>
> i=1:n;i
[1] 1 2 3 4 5
> RX=rank(X);RX
[1] 3 1 5 2 4
>
> RY=rank(Y);RY
[1] 2 1 5 3 4
>
> mRX=mean(RX);mRX
[1] 3
> mRY=mean(RY);mRY
[1] 3
>
> Num=sum( (RX-mRX)*(RY-mRY));Num
[1] 9
>
> Den2=sum( (RX-mRX)^2)*sum( (RY-mRY)^2);Den2
[1] 100
> Den=sqrt(Den2);Den
[1] 10
>
> S=Num/Den;S
[1] 0.9
>
>
> ####Outra maneira
>
>
>
> D=RX-RY;D
[1]  1  0  0 -1  0
>
> n=5
> D2=D^2
> sum(D2)
[1] 2
> aux=n^3-n;aux
[1] 120
>
> 1- (6*sum(D^2))/aux
[1] 0.9
>
>
>
>
> tab=cbind(i,X,Y,RX,RY,D,D2);tab
i    X    Y RX RY  D D2
```

```
[1,] 1 20.4  9.2  3  2  1  1
[2,] 2 19.7  8.9  1  1  0  0
[3,] 3 21.8 11.4  5  5  0  0
[4,] 4 20.1  9.4  2  3 -1  1
[5,] 5 20.7 10.3  4  4  0  0
>
>
> ######Direto no R
>
>
>
>
>
>
> cor(X,Y)########## Correlação de Pearson
[1] 0.9621163
>
>
> cor(X,Y,method="spearman") #####Tira toda a emoção!!!!!!!
[1] 0.9
>
>
>
>
```

Empates

No caso de empates entre os valores de X e de Y procedemos do modo usual. Para os empates só de X ou só de Y, consideramos a média dos postos que seriam àqueles valores, caso não houvesse o empate.

Neste caso usamos

$$r_1 = \frac{12\sum_{i=1}^{n} R_i\, S_i - 3n(n+1)^2}{\sqrt{[n(n^2-1)-t'][n(n^2-1)-u']}},$$

em que:

g = número de grupos de empates entre os X;

h = número de grupos de empates entre os Y;

$$t' = \sum_{i=1}^{g} t_i(t_i^2 - 1),$$

t_i número de observações no i-ésimo grupo de X.

$$u' = \sum_{j=1}^{h} u_j(u_j^2 - 1),$$

u_j número de observações no j-ésimo grupo de Y.

Exemplo 2

Vamos supor a seguinte estrutura de postos:

i	1	2	3	4	5	6	7	8	9	10	11	12	13
R(X)	7	12	8	2	4	5	10	13	9	1	3	6	11
S (Y)	8	9,5	9,5	2	7	5	13	12	3	1	4	11	6

Percebemos que não ha empates na variável X e um grupo de 2 observações empatada em Y. Logo

$$t' = 0 \quad u' = 2^2 - 2 = 6.$$

Como $n = 13$ temos

$$n(n^2 - 1) = 13 \times 168 = 1184$$

$$n(n^2 - 1) - u' = 2184 - 6 = 2178.$$

Veja o cálculo na saída do **R**:

```
>
> n=13
>
> a=n*(n^2-1)
> a;
[1] 2184
> a-6
[1] 2178
>
>
> Den2=a*(a-6);Den2
[1] 4756752
>
> Den=sqrt(Den2);Den
[1] 2180.998
>
> R=c(7,12,8,2,4,5,10,13,9,1,3,6,11)
> sum(R)
[1] 91
> S=c(8,9.5,9.5,2,7,5,13,12,3,1,4,11,6)
> sum(S)
[1] 91
>
>
> sum(R*S)
[1] 761
>
>
> Num=12*sum(R*S)-3*n*(n+1)^2;Num
[1] 1488
```

```
>
> r=Num/Den;r
[1] 0.6822565
>
> cor(R,S)
[1] 0.6822565
>
```

Teste de Hipóteses: Seja (X, Y) um população contínua bivariada com correlação de Spearman ρ_S.

Queremos testar se X e Y são independentes ($\rho_S = 0$)

Seja α o nível de significância desejado.

$$H_0 : \rho_S = 0.$$

Contra uma das hipóteses alternativas:

$$H_1 : \rho_S > 0,$$

isto é, X e Y são positivamente correlacionados.

Neste caso rejeitamos H_0 se

$$r \geq r(\alpha, n)$$

em que

$$P_0(r \geq r(\alpha, n)) = \alpha$$

são encontrados na tabela 13 dada em sala de aula.

Ou

$$H_1 : \rho_S < 0,$$

isto é, X e Y são negativamente correlacionados.

Neste caso rejeitamos H_0 se

$$r \leq -r(\alpha, n)$$

Ou

$$H_1 : \rho_S \neq 0,$$

isto é, X e Y são dependentes.

Neste caso rejeitamos H_0 se

$$|r| \geq r(\frac{\alpha}{2}, n)$$

Lembrando ainda que a distribuição de R é simétrica em torno da origem e :

Se H_0 é verdade e não há empates temos que

$$E(r) = 0 \quad e \quad Var(r) = \frac{1}{n-1},$$

Aproximação Normal:

Para valores grandes de n temos:

$$Z = \frac{r - E_0(r)}{\sqrt{V_0(r)}} = \frac{r}{1/\sqrt{(n-1)}} = r\sqrt{n-1}$$

é aproximadamente normal padrão.

Assim, para testarmos

$$H_0 : \rho_S = 0 \quad versus \quad H_0 : \rho_S \neq 0,$$

rejeitamos H_0 se

$$|r| \geq z_{tab},$$

com

$$P(Z \geq z_{tab}) = \frac{\alpha}{2}.$$

Distribuição Nula de r:

Para a obtenção da distribuição nula de r vamos utilizar a fórmula

$$r = 1 - \frac{6\sum_{i=1}^{n}(R_i - S_i)^2}{n(n^2-1)} = 1 - \frac{6A}{n(n^2-1)} = 1 - B.$$

Vamos fazer para o caso $n = 4$.

Inicialmente vamos fixar os postos de X em:

$$R_1 = 1 \ ; \ R_2 = 2 \ , \ R_3 = 3 \ ; R_4 = 4.$$

Vamos supor que os postos de Y são:

$$S_1 = 1 \ ; \ S_2 = 2 \ , \ S_3 = 3 \ ; S_4 = 4.$$

Note que

$$n(n^2-1) = 4 \times 15 = 60$$

$$B = \frac{6A}{n(n^2-1)} = \frac{6A}{60} = \frac{A}{10}.$$

$$A = \sum_{i=1}^{4}(R_i - S_i)^2 = (1-1)^2 + (2-2)^2 + (3-3)^2 + (4-4)^2 = 0$$

Logo

$$B = \frac{A}{10} = \frac{0}{10} = 0,$$

finalmente

$$r = 1 - B = 1 - 0 = 1.$$

Vamos supor agora que os postos de Y são:

$$S_1 = 4 \ ; \ \ S_2 = 3 \ , \ \ S_3 = 2 \ ; S_4 = 1.$$

$$A = \sum_{i=1}^{4}(R_i - S_i)^2 = (1-4)^2 + (2-3)^2 + (3-2)^2 + (4-1)^2 = 20$$

Logo

$$B = \frac{A}{10} = \frac{20}{10} = 2,$$

finalmente

$$r = 1 - B = 1 - 2 = -1.$$

Temos $4! = 24$ maneiras de atribuir os postos a $Y <$isto gera a seguinte tabela:

i	Y_1	Y_2	Y_3	Y_4	A	B	r
1	1	2	3	4	0	0,00	1,00
2	1	2	4	3	2	0,20	0,80
3	1	3	2	4	2	0,20	0,80
4	1	3	4	2	6	0,60	0,40
5	1	4	2	3	6	0,60	0,40
6	1	4	3	2	8	0,80	0,20
7	2	1	3	4	2	0,20	0,80
8	2	1	4	3	4	0,40	0,60
9	2	3	1	4	6	0,60	0,40
10	2	3	4	1	12	1,20	-0,20
11	2	4	1	3	10	1,00	0,00
12	2	4	3	1	14	1,40	-0,40
13	3	1	2	4	6	0,60	0,40
14	3	1	4	2	10	1,00	0,00
15	3	2	1	4	8	0,80	0,20
16	3	2	4	1	14	1,40	-0,40
17	3	4	1	2	16	1,60	-0,60
18	3	4	2	1	18	1,80	-0,80
19	4	1	2	3	12	1,20	-0,20
20	4	1	3	2	14	1,40	-0,40
21	4	2	1	3	14	1,40	-0,40
22	4	2	3	1	18	1,80	-0,80
23	4	3	1	2	18	1,80	-0,80
24	4	3	2	1	20	2,00	-1,00

Vamos resumir a tabela anterior:

Note que a distribuição nula de r é dada por:

$$P_0(r = r_0) = \frac{f_i}{n!} = \frac{f_i}{24}.$$

r_0	f_i	$P(r = r_0)$	$P(r \leq r_0)$	$P(r \geq r_0)$
-1	1	0,042	0,042	1
-0,8	3	0,125	0,167	0,958
-0,6	1	0,042	0,208	0,833
-0,4	4	0,167	0,375	0,792
-0,2	2	0,083	0,458	0,625
0	2	0,083	0,542	0,542
0,2	2	0,083	0,625	0,458
0,4	4	0,167	0,792	0,375
0,6	1	0,042	0,833	0,208
0,8	3	0,125	0,958	0,167
1	1	0,042	1,000	0,042

Veja a saída do **R**:

```
>
> r_0=seq(-1,1,0.2);r_0
[1] -1.0 -0.8 -0.6 -0.4 -0.2  0.0  0.2  0.4  0.6  0.8  1.0
```

```
>
> f=c(1,3,1,4,2,2,2,4,1,3,1);sum(f)
[1] 24
>
>
> pr=f/factorial(4);sum(pr)
[1] 1
>
> Pr=cumsum(pr)
> aux=c(0,Pr[-11]);aux
[1] 0.00000000 0.04166667 0.16666667 0.20833333 0.37500000 0.45833333
[7] 0.54166667 0.62500000 0.79166667 0.83333333 0.95833333
>
> Sr=1-aux;Sr
[1] 1.00000000 0.95833333 0.83333333 0.79166667 0.62500000 0.54166667
[7] 0.45833333 0.37500000 0.20833333 0.16666667 0.04166667
> tab=cbind(r_0,f,pr,Pr,Sr);tab
r_0 f         pr         Pr         Sr
[1,] -1.0 1 0.04166667 0.04166667 1.00000000
[2,] -0.8 3 0.12500000 0.16666667 0.95833333
[3,] -0.6 1 0.04166667 0.20833333 0.83333333
[4,] -0.4 4 0.16666667 0.37500000 0.79166667
[5,] -0.2 2 0.08333333 0.45833333 0.62500000
[6,]  0.0 2 0.08333333 0.54166667 0.54166667
[7,]  0.2 2 0.08333333 0.62500000 0.45833333
[8,]  0.4 4 0.16666667 0.79166667 0.37500000
[9,]  0.6 1 0.04166667 0.83333333 0.20833333
[10,]  0.8 3 0.12500000 0.95833333 0.16666667
[11,]  1.0 1 0.04166667 1.00000000 0.04166667
>
```

```
>
> Er=sum(r_0*pr);Er;round(Er,10)
[1] 4.857226e-17
[1] 0
> n=4
> Vr=sum(r_0^2*pr);Vr;1/(n-1)
[1] 0.3333333
[1] 0.3333333
>
> require(MASS)
> fractions(Vr)
[1] 1/3
>
```

Exemplo 3: Ferreira (1970), estudando a densidade básica média do Eucalyptus grandis Hill

ex maiden, considerou:

X = Densidade média ao nível do D.A.P.(Diâmetro à altura do peito), amostras Pressler(Volume determinado através de paquímetros e do diâmetro da sonda de Pressler de 0,5 cm)

Y = Densidade média ao nível do D.A.P.(seções tranversais do tronco).

Os dados seguintes constituem parte dos seus resultados e são referidos em g/cm^3

i	X	Y
1	0,602	0,619
2	0,636	0,620
3	0,604	0,621
4	0,548	0,538
5	0,590	0,616
6	0,592	0,601
7	0,625	0,664
8	0,641	0,652
9	0,606	0,579
10	0,502	0,501
11	0,588	0,590
12	0,594	0,622
13	0,626	0,606

Responda ao que se pede:

a. Estime a correlação de Pearson (ρ) , de Spearman (ρ_S) e de Kendall (τ)?

b. Teste se X e Y são positivamente correlacionados usando um teste paramétrico. Use um nível de significância de 5%. Explique com detalhes a saída do **R**.

c. Teste se X e Y são positivamente correlacionados usando um teste não paramétrico. Use um nível de significância de 5%. Calcule o nível descritivo usando a tabela 13. Calcule o nível descritivo aproximado.

Explique com detalhes a saída do **R**.

```
 X=c(602,636,604,548,590,592,625,641,606,502,588,594,626)/1000;X
[1] 0.602 0.636 0.604 0.548 0.590 0.592 0.625 0.641 0.606 0.502 0.588 0.594
[13] 0.626
> n=length(X);n
[1] 13
>
> RX=rank(X);RX
[1]  7 12  8  2  4  5 10 13  9  1  3  6 11
>
> Y=c(619,620,621,538,616,601,664,652,579,501,590,622,606)/1000
>
> RY=rank(Y);RY
[1]  8  9 10  2  7  5 13 12  3  1  4 11  6
>
>
```

```
> D=RX-RY;D
[1] -1  3 -2  0 -3  0 -3  1  6  0 -1 -5  5
> aux=sum(D^2);aux
[1] 120
>
> Den=n*(n^2-1);Den
[1] 2184
>
> r=1- 6*aux/Den;r
[1] 0.6703297
>
> cor(X,Y,method="spearman")
[1] 0.6703297
> cor(X,Y,method="kendall")
[1] 0.5384615
>
> cor(X,Y,method="pearson")
[1] 0.8891707
>
> cor.test(X,Y, alternative="greater")

Pearson's product-moment correlation

data:  X and Y
t = 6.4449, df = 11, p-value = 2.388e-05
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
0.7152263 1.0000000
sample estimates:
cor
0.8891707

>
> cor.test(X,Y, method="spearman",alternative="greater")

Spearman's rank correlation rho

data:  X and Y
S = 120, p-value = 0.007459
alternative hypothesis: true rho is greater than 0
sample estimates:
rho
0.6703297

>
> cor.test(X,Y, method="kendall",alternative="greater")

Kendall's rank correlation tau

data:  X and Y
T = 60, p-value = 0.005059
alternative hypothesis: true tau is greater than 0
```

```
sample estimates:
tau
0.5384615

>
```