MÈTODES NO PARAMÈTRICS I DE REMOSTREIG

Grau en Estadística UB-UPC. Curs 2014-2015. Prova parcial.

Cada pregunta compta 1 punt: puntuació màxima 10.

Respon als mateixos fulls de l'enunciat. Si et falta espai pots utilitzar el que hi ha al final desprès dels llistats, i com a darrer recurs fulls addicionals. En tot moment pots considerar un nivell de significació de 0,05 o de confiança de 0,95. Quan realitzis una prova d'hipòtesis has d'expressar clarament les hipòtesis nul·la i alternativa, la conclusió final i el procés que hi ha conduït.

lloc	agost	novembre
1	8.1	11.2
2	10.0	16.3
3	16.5	15.3
4	13.6	15.6
5	9.5	10.5
6	8.3	15.5
7	18.3	12.7
8	13.3	11.1
9	7.9	19.9
10	8.1	20.4
11	8.9	14.2
12	12.6	12.7
13	13.4	36.8

Problema 1. En un estudi sobre els efectes de la contaminació en els boscos, es van escollir 13 llocs a l'atzar d'una zona molt contaminada, i per cada lloc es va mesurar el nivell d'alumini (en micrograms per gram de fusta) d'un pollancre. Per cada lloc la mesura es va fer el mes d'agost i el mes de novembre.

Respon les següents qüestions, utilitzant els llistats del final de l'enunciat de l'examen quan ho creguis convenient.

1) Indica el nom d'una prova d'hipòtesis basada en rangs que sigui adequada per a intentar demostrar que la contaminació (expressada com la mediana del nivell d'alumini) ha crescut de l'agost al novembre. Indica les condicions de validesa de la prova que has triat.

Es tracta clarament d'una situació de dades aparellades, dos mesos diferents però pel mateix lloc. Per tant la prova de rangs candidata més clara seria la prova de Wilcoxon dels rangs amb signe.

Aquesta prova requereix treballar amb les diferències dins cada lloc. La variable diferència ha de tenir distribució contínua i **simètrica**, com a conseqüència de la suposició de continuïtat i d'igualtat distribucional, excepte en la localització, entre les dades d'agost i novembre.

2) Realitza la prova anterior sobre les dades d'aquest problema. Què demostra el resultat? Per a major simplicitat, no tinguis en compte la possible presència d'empats.

Si δ indica la mediana de les diferències **agost – novembre**, tenim una prova **unilateral** (es vol demostrar que el nivell d'alumini ha pujat d'agost a novembre) $H_0 : \delta = 0$ vs $H_1 : \delta < 0$. Cal basar-se en els càlculs senyalats en verd (precedits d'uns càlculs que correspondrien a dues mostres independents, aquests **no** tenen res a veure amb aquest apartat). Intuïtivament, si fos certa l'alternativa haurien de predominar les diferències negatives. La suma de rangs positius de les diferències, R^+, val 16, i la de rangs negatius 75. Per tant tenim que

 NOM I COGNOMS: **FIRMA:**

L'estadístic de Wilcoxon és $T = \min\{R^+, R^-\} = 16$, més extrem (menor) que el valor que trobaríem a les taules per la prova **unilateral** sota un nivell 0.05 per 13 dades (= 21). Com que a més predominen els rangs negatius ($R^+ < R^-$) podem rebutjar H_0, la conclusió és que hi ha hagut un increment en mediana del nivell d'alumini.

3) Indica justificadament el valor de l'estimació puntual i l'interval de confiança **bilateral** per al canvi experimentat en la mediana de les diferències de concentració d'alumini.

Continuant amb els llistats ressaltats, la mediana de totes les diferències (-3.1) té sentit com a estimació puntual de l'efecte estudiat. Però l'estimació puntual associada al test de Wilcoxon per dades aparellades és la mediana de les semisumes de diferències, -4.1. Restant aquest valor a les diferències uniformitzaríem al màxim la mostra, en el sentit de maximitzar el p-valor o equilibrar els rangs de diferències negatives i positives.

L'interval de confiança bilateral associat al test anterior correspon a $[D_{(\lambda)}, D_{(\nu)}]$ on $D_{()}$ és el **vector de diferències ordenat de menor a major** (és a dir: -23.4, -12.3, -12.0, -7.2, -6.3, -5.3, -3.1, -2.0, -1.0, -0.1, 1.2, 2.2, 5.6) i les posicions ν i λ es determinen com:

$$\upsilon^* = \frac{n+1}{2} + \frac{1}{2} z_\alpha \sqrt{n} = \frac{13+1}{2} + \frac{1}{2} 1.96\sqrt{13} = 10.53$$

$$\upsilon = \left\{ \begin{array}{ll} \upsilon^* \ \text{si} \ \upsilon^* & \text{és enter} \\ \left[\upsilon^* + 1\right] & \text{en cas contrari} \end{array} \right\} = 11 \quad \lambda = n - \upsilon + 1 = 13 - 11 + 1 = 3$$

$$z_\alpha \ \text{valor t.q.} \ \Pr\left(|Z| \le z_\alpha\right) = 1 - \alpha, \ Z \sim N(0,1)$$

Per tant l'interval de confiança seria $[D_{(3)}, D_{(11)}]=[-12, 1.2]$.

4) Suposa que, en les mateixes condicions d'abans, el nivell d'alumini s'ha mesurat en més de 2 mesos (per exemple: agost, octubre i desembre). Indica el nom d'una prova basada en rangs adequada per a intentar demostrar que la mediana del nivell d'alumini ha variat segons els mesos.

Tractant-se d'un disseny en blocs (cada lloc un bloc, mesurat en 3 instants diferents) una prova de rangs adient seria la de **Friedman**.

5) Realitza un prova de permutacions per intentar demostrar que el nivell d'alumini **mitjà** ***ha variat*** d'agost a novembre.

Els llistats ressaltats en groc ens permeten fer un test de permutacions exacte, enumerant totes les permutacions, que no són gaires: 2^{13} = 8192. Si ara δ designa la mitjana poblacional de les diferències ***agost – novembre***, tindrem un test bilateral $H_0 : \delta = 0$ vs $H_1 : \delta \neq 0$. Segons els llistats, la mitjana mostral de les diferències sobre la mostra original val -4.9. La mitjana sobre les diferències de dades permutades ha estat superior o igual a aquest valor 8070 vegades, ha estat inferior o igual 123 vegades o bé ha estat tant o més extrema (en negatiu o en positiu) 246 vegades:

```
> sum(abs(m.perm) >= abs(m.d))
[1] 246
```

Aquest és el valor que cal utilitzar per calcular el p-valor ja que el test plantejat és **bilateral**. El p-valor exacte és 246 / 8192 = 0.03 (arrodonint a 2 decimals) i per tant rebutgen la hipòtesi nul·la, la conclusió final és que hi ha hagut variació del nivell mitjà d'alumini.

(Fixeu-vos que aquí no tindria sentit fer (246 + 1) / (8192 + 1), no estem estimant el p-valor sinó que l'estem calculant de forma exacta, enumerant totes les possibles permutacions.)

Problema 2. Encara que no correspon a la situació real, suposa ara que les dades del problema anterior corresponen a 26 llocs diferents, 13 llocs seleccionats a l'atzar a l'agost i 13 llocs al novembre.

1) Realitza una prova de permutacions per demostrar que la contaminació mitjana ha augmentat d'agost a novembre.

Ara estaríem en una situació a la qual compararíem dues mostres independents de mides 13 i 13. Caldria per tant permutar lliurement les 26 dades, la qual cosa donaria un total de $26! > 4 \times 10^{26}$ permutacions possibles. Com sabem, en realitat només caldria enumerar les permutacions amb repetició, $26!/(13!13!)$, però continua sent un nombre molt gran, més de 10 milions. És quasi impossible enumerar-les totes de forma exhaustiva, per tant caldria realitzar una prova de permutacions de Monte Carlo.

Si δ designa la diferència de mitjanes poblacionals, les hipòtesis serien H_0: $\delta = 0$ vs $H1$: $\delta < 0$. Per un contrast com aquest, un estadístic de test adient és la diferència de mitjanes mostrals –també ho seria la suma dels valors d'un grup, en tractar-se d'un cas balancejat, però no disposem d'aquesta informació als llistats. Segons els llistats situats entre els ressaltats en verd i en groc, tenim que de 9999 diferències de mitjanes mostrals calculades sobre permutacions generades aleatòriament, 74 han estat menors que la diferència de mitjanes mostrals obtinguda sobre les dades reals, -4.9. Per tant l'estimació del p-valor és $(74 + 1) / (9999 + 1) = 0.0075 < 0.05$, de manera que podrem rebutjar H_0, hem obtingut evidència a favor d'un augment de la contaminació d'agost a novembre, expressada en el sentit que indiquen les hipòtesis.

2) Realitza una prova basada en rangs per demostrar que la contaminació mediana ha augmentat d'agost a novembre.

Atès que ara estaríem considerant una situació de dues mostres independents, tindria sentit aplicar la prova de Mann-Whitney-Wilcoxon.

Les sumes dels rangs de les 13 dades d'agost i de les 13 de novembre, dins el vector conjunt de les 26 dades, serien respectivament $R_{1\cdot} = 129$ i $R_{2\cdot} = 222$, de manera que

$$U = \min\left\{R_{1\cdot} - \frac{n(n+1)}{2}, R_{2\cdot} - \frac{m(m+1)}{2}\right\} =$$
$$\min\left\{129 - \frac{13(13+1)}{2}, 222 - \frac{13(13+1)}{2}\right\} =$$
$$\min\{38, 131\} = 38.$$

Per mides mostrals $n = m = 13$ trobem a les taules un valor crític unilateral $u_{0,05}(13,13) = 51$. Atès que el valor observat de l'estadístic és més extrem (menor) que el valor crític de les taules, $U = 38 < 51$, i que la mitjana de rangs de les dades "agost" és menor que la mitjana dels rangs de les dades "novembre", podem rebutjar H_0, hem obtingut evidència a favor d'un augment de la contaminació d'agost a novembre, expressada en el sentit que indiquen les hipòtesis. (No s'ha demanat, però aquest increment de la contaminació, "estimació de l'efecte", es podria dir que seria de –3.2, la mediana de les 169 diferències agost – novembre.)

Problema 3. Tornem a la interpretació inicial de les dades feta al **Problema 1**: 13 llocs del bosc, amb dades de cada lloc, a l'agost i al novembre. Atès que cada parella de valors de contaminació per alumini es refereix al mateix lloc del bosc, seria lògic sospitar que hi pot haver un cert grau de dependència entre les variables X = 'agost' i Y = 'novembre'.

1) Ignorant el empats, calcula el coeficient tau de Kendall entre X i Y i determina si és significatiu. El resultat del test anterior, demostra que els nivells d'alumini d'agost i novembre són estocàsticament independents?

El nombre de concordances entre diferències dels valors de X i de Y és $n_C = 33$ i el de discordances $n_D = 43$. El total de parelles de valors de X o de Y a formar és de $n(n-1)/2 = 78$, per tant el coeficient de correlació de Kendall mostral és $\hat{\tau} = (33 - 43) / 78 = -0.128$. (Com que hi ha empats, és possible que algú hagi dividit per 33 + 43 = 76; també ho donem per bo, però en realitat tots aquests càlculs no són del tot correctes i caldria calcular l'anomenat tau-B o tau-C, estimacions alternatives millors en presència d'empats. Si algú ho fa, perfecte, però a efectes d'aquest examen ja està bé l'anterior.)

Per estudiar-ne la significació es plantejaria el contrast H_0 : "X i Y estocàsticament independents" vs $H_1 : \tau(X,Y) \neq 0$, "el coeficient de correlació de Kendall poblacional no és nul". Rebutjarem H_0 si $|\hat{\tau}| > \tau_\alpha(n)$ on $\tau_{0.05}(13) = 0.436$ correspon al valor crític a les taules de la prova bilateral pel nivell 0.05 i mida mostral 13.

Com que $|-0.128| = 0.128 < 0.436$, **NO** podem rebutjar H_0.

Però recordem que **no rebutjar una hipòtesi nul·la NO la demostra**, això no serveix com a demostració d'independència entre X i Y. Deduir que hem demostrat que aquestes variables són independents seria un error metodològic que hem d'evitar, l'únic que podem afirmar és que no hem pogut demostrar dependència (en el sentit de la tau de Kendall), que no és el mateix; potser amb una mostra més gran hauríem aconseguit rebutjar H_0.

2) Justificant el resultat, indica el valor del coeficient de correlació de Spearman.

Del resultat de `cor.test(rank(x), rank(y))`, se'n dedueix que el valor demanat és -0.138 (arrodonint a 3 decimals).

Aquesta pregunta solament es considerarà 100% ben contestada si l'elecció d'aquest valor es justifica, és a dir si es diu clarament que el coeficient de correlació de Spearman correspon al coeficient de correlació de Pearson ***però entre els rangs.***

També seria correcte el càlcul directe del coeficient a partir de les dades, que hauria de donar el mateix resultat.

3) Realitza una prova de permutacions per a determinar si el coeficient de correlació lineal de Pearson entre *X* i *Y* és diferent de zero.

H_0 : "X i Y estocàsticament independents" vs H_1 : $\rho(X,Y) \neq 0$, "el coeficient de correlació de Pearson poblacional no és nul". Si H_0 fos certa, la mostra observada seria una més de les possibles permutacions a les quals deixem igual un dels vectors d'observacions (indistintament, X o Y) i permutem l'altre. Aquí el nombre de permutacions possibles de les observacions Y (o X) és molt gran, 13! = 6227020800, és casi impossible fer una prova de permutacions exacta, enumerant totes les permutacions possibles. Utilitzarem l'aproximació de Monte Carlo realitzada en part al llistat final on s'han generat 9999 permutacions aleatòries (permutant els valors d'Y, `sample(y, replace = FALSE)`*) i sobre cadascuna d'elles s'ha calculat el coeficient de correlació de Pearson mostral. La hipòtesi alternativa és bilateral al voltant de zero, per tant comptarem com a "extrems" tant els valors negatius com positius però "grans" en valor absolut en comparació amb el valor realment observat a la mostra original, r = 0.01669772. De les 9999 mostres permutades, 9645 proporcionen un valor de correlació mostral tant o més extrem (en negatiu o positiu) que 0.01669772. Per tant, l'estimador de Dwass del p-valor, (9645 + 1) / (9999 + 1) = 0.9646* ***NO*** *permet rebutjar H_0. Novament, això NO demostra que hi hagi independència, senzillament no tenim prou evidències que hi hagi correlació.*

```
> pollancres = read.table("pollancres.txt", header = TRUE)
>
> agost = pollancres$agost
> novembre = pollancres$novembre
>
> # Totes les dades en un únic vector:
> alumini = c(agost, novembre)
>
> N = length(alumini)
> N
[1] 26
> n1 = length(agost)
> n1
[1] 13
> n2 = length(novembre)
> n2
[1] 13
>
> # rang de cada observació dins el total de N = 26 valors:
> rangs <- rank(alumini)
> rangs
 [1]  2.5  7.0 22.0 16.0  6.0  4.0 23.0 14.0  1.0  2.5  5.0 11.0 15.0 10.0 21.0
[16] 18.0 20.0  8.0 19.0 12.5  9.0 24.0 25.0 17.0 12.5 26.0
> # rangs de les observacions "agost"
> rangs[1:n1]
 [1]  2.5  7.0 22.0 16.0  6.0  4.0 23.0 14.0  1.0  2.5  5.0 11.0 15.0
> # rangs de les observacions "novembre"
> rangs[(n1+1):N]
 [1] 10.0 21.0 18.0 20.0  8.0 19.0 12.5  9.0 24.0 25.0 17.0 12.5 26.0
>
> # Sumes de rangs dins cada grup:
> # Agost:
> sum(rangs[1:n1])
[1] 129
> # Novembre:
> sum(rangs[(n1+1):N])
[1] 222
>
> # Totes les diferències possibles (13 * 13 = 169 valors)
> # entre "agost" i "novembre":
> dd = outer(agost, novembre, "-")
> dd
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]  [,9] [,10] [,11] [,12] [,13]
 [1,] -3.1 -8.2 -7.2 -7.5 -2.4 -7.4 -4.6 -3.0 -11.8 -12.3  -6.1  -4.6 -28.7
 [2,] -1.2 -6.3 -5.3 -5.6 -0.5 -5.5 -2.7 -1.1  -9.9 -10.4  -4.2  -2.7 -26.8
 [3,]  5.3  0.2  1.2  0.9  6.0  1.0  3.8  5.4  -3.4  -3.9   2.3   3.8 -20.3
 [4,]  2.4 -2.7 -1.7 -2.0  3.1 -1.9  0.9  2.5  -6.3  -6.8  -0.6   0.9 -23.2
 [5,] -1.7 -6.8 -5.8 -6.1 -1.0 -6.0 -3.2 -1.6 -10.4 -10.9  -4.7  -3.2 -27.3
 [6,] -2.9 -8.0 -7.0 -7.3 -2.2 -7.2 -4.4 -2.8 -11.6 -12.1  -5.9  -4.4 -28.5
 [7,]  7.1  2.0  3.0  2.7  7.8  2.8  5.6  7.2  -1.6  -2.1   4.1   5.6 -18.5
 [8,]  2.1 -3.0 -2.0 -2.3  2.8 -2.2  0.6  2.2  -6.6  -7.1  -0.9   0.6 -23.5
 [9,] -3.3 -8.4 -7.4 -7.7 -2.6 -7.6 -4.8 -3.2 -12.0 -12.5  -6.3  -4.8 -28.9
[10,] -3.1 -8.2 -7.2 -7.5 -2.4 -7.4 -4.6 -3.0 -11.8 -12.3  -6.1  -4.6 -28.7
[11,] -2.3 -7.4 -6.4 -6.7 -1.6 -6.6 -3.8 -2.2 -11.0 -11.5  -5.3  -3.8 -27.9
[12,]  1.4 -3.7 -2.7 -3.0  2.1 -2.9 -0.1  1.5  -7.3  -7.8  -1.6  -0.1 -24.2
[13,]  2.2 -2.9 -1.9 -2.2  2.9 -2.1  0.7  2.3  -6.5  -7.0  -0.8   0.7 -23.4
>
> # mediana de les 169 diferències:
> median(dd)
[1] -3.2
>
> # Les 169 diferències ordenades:
> sort(dd)
  [1] -28.9 -28.7 -28.7 -28.5 -27.9 -27.3 -26.8 -24.2 -23.5 -23.4 -23.2 -20.3
 [13] -18.5 -12.5 -12.3 -12.3 -12.1 -12.0 -11.8 -11.8 -11.6 -11.5 -11.0 -10.9
 [25] -10.4 -10.4  -9.9  -8.4  -8.2  -8.2  -8.0  -7.8  -7.7  -7.6  -7.5  -7.5
 [37]  -7.4  -7.4  -7.4  -7.4  -7.3  -7.3  -7.2  -7.2  -7.2  -7.1  -7.0  -7.0
 [49]  -6.8  -6.8  -6.7  -6.6  -6.6  -6.5  -6.4  -6.3  -6.3  -6.3  -6.1  -6.1
 [61]  -6.1  -6.0  -5.9  -5.8  -5.6  -5.5  -5.3  -5.3  -4.8  -4.8  -4.7  -4.6
 [73]  -4.6  -4.6  -4.6  -4.4  -4.4  -4.2  -3.9  -3.8  -3.8  -3.7  -3.4  -3.3
 [85]  -3.2  -3.2  -3.2  -3.1  -3.1  -3.0  -3.0  -3.0  -3.0  -2.9  -2.9  -2.9
 [97]  -2.8  -2.7  -2.7  -2.7  -2.7  -2.6  -2.4  -2.4  -2.3  -2.3  -2.2  -2.2
[109]  -2.2  -2.2  -2.1  -2.1  -2.0  -2.0  -1.9  -1.9  -1.7  -1.7  -1.6  -1.6
[121]  -1.6  -1.6  -1.2  -1.1  -1.0  -0.9  -0.8  -0.6  -0.5  -0.1  -0.1   0.2
```

```
[133]    0.6    0.6    0.7    0.7    0.9    0.9    0.9    1.0    1.2    1.4    1.5    2.0
[145]    2.1    2.1    2.2    2.2    2.3    2.3    2.4    2.5    2.7    2.8    2.8    2.9
[157]    3.0    3.1    3.8    3.8    4.1    5.3    5.4    5.6    5.6    6.0    7.1    7.2
[169]    7.8
>
>
> # Diferència dins cada lloc (13 valors possibles):
> d = agost - novembre
> d
 [1]   -3.1  -6.3   1.2  -2.0  -1.0  -7.2   5.6   2.2 -12.0 -12.3  -5.3  -0.1
[13] -23.4
> n = length(d)
> n
[1] 13
> # Valors absoluts de les diferències:
> abs.d = abs(d)
> abs.d
 [1]   3.1  6.3  1.2  2.0  1.0  7.2  5.6  2.2 12.0 12.3  5.3  0.1 23.4
> #
> # Rangs dels valors absoluts de les diferències:
> rabs.d = rank(abs.d)
> rabs.d
 [1]   6  9  3  4  2 10  8  5 11 12  7  1 13
> #
> # Suma de rangs de diferencies positives:
> r.plus = sum(rabs.d[d > 0])
> r.plus
[1] 16
> # Suma de rangs de diferencies negatives:
> r.minus = sum(rabs.d[d < 0])
> r.minus
[1] 75
> #
> #
> # Mediana de totes les diferències:
> median(d)
[1] -3.1
> # Mediana de totes les semisumes entre diferències:
> sSums = outer(d, d, "+") / 2
> median(sSums[lower.tri(sSums, diag = TRUE)])
[1] -4.1


> # Diferència de les mitjanes de les dades "agost" i "novembre":
> dMeans = mean(agost) - mean(novembre)
> dMeans
[1] -4.9
>
> # 9999 permutacions aleatòries de les 26 dades.
> # Per cada permutació, càlcul de la diferència de la mitjana de les n1
> # primeres i les n2 últimes:
> nperm = 9999
> dMeansPerm = replicate(nperm,
+   {
+     alumini.perm = sample(alumini)
+     mean(alumini.perm[1:n1]) - mean(alumini.perm[(n1+1):N])
+   }
+ )
>
>
> sum(dMeansPerm <= dMeans)
[1] 74
> sum(dMeansPerm >= dMeans)
[1] 9926
> sum(abs(dMeansPerm) >= abs(dMeans))
[1] 153


> # Permutacions sobre el vector de 13 diferències.
> # Enumeració de TOTES les permutacions possibles, maneres segons les quals podem
> # permutar DINS cada parella de valors (agost, novembre). En altres paraules,
> # maneres possibles segons les quals podem donar un signe - o + a les diferències:
> sgn = c(-1, +1)
> signsTab = expand.grid(as.data.frame(matrix(rep(sgn, n), ncol = n)))
> signsTab = apply(signsTab, 1, "*", abs.d)
> # Cada columna de 'signsTab' conté les diferències sobre una permutació possible.
> # Per exemple les 10 primeres:
```

```
> signsTab[,1:10]
     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
v1   -3.1   3.1  -3.1   3.1  -3.1   3.1  -3.1   3.1  -3.1   3.1
v2   -6.3  -6.3   6.3   6.3  -6.3  -6.3   6.3   6.3  -6.3  -6.3
v3   -1.2  -1.2  -1.2  -1.2   1.2   1.2   1.2   1.2  -1.2  -1.2
v4   -2.0  -2.0  -2.0  -2.0  -2.0  -2.0  -2.0  -2.0   2.0   2.0
v5   -1.0  -1.0  -1.0  -1.0  -1.0  -1.0  -1.0  -1.0  -1.0  -1.0
v6   -7.2  -7.2  -7.2  -7.2  -7.2  -7.2  -7.2  -7.2  -7.2  -7.2
v7   -5.6  -5.6  -5.6  -5.6  -5.6  -5.6  -5.6  -5.6  -5.6  -5.6
v8   -2.2  -2.2  -2.2  -2.2  -2.2  -2.2  -2.2  -2.2  -2.2  -2.2
v9  -12.0 -12.0 -12.0 -12.0 -12.0 -12.0 -12.0 -12.0 -12.0 -12.0
v10 -12.3 -12.3 -12.3 -12.3 -12.3 -12.3 -12.3 -12.3 -12.3 -12.3
v11  -5.3  -5.3  -5.3  -5.3  -5.3  -5.3  -5.3  -5.3  -5.3  -5.3
v12  -0.1  -0.1  -0.1  -0.1  -0.1  -0.1  -0.1  -0.1  -0.1  -0.1
v13 -23.4 -23.4 -23.4 -23.4 -23.4 -23.4 -23.4 -23.4 -23.4 -23.4
> # Nombre de permutacions possibles:
> nperm = ncol(signsTab)
> nperm
[1] 8192
> #
> # Estimació de la mitjana de les diferències sobre cada possible permutació:
> m.perm = apply(signsTab, 2, mean)
> #
> # La mitjana de les diferències a la mostra original és:
> m.d = mean(d)
> m.d
[1] -4.9
> #
> sum(m.perm >= m.d)
[1] 8070
> sum(abs(m.perm) >= abs(m.d))
[1] 246
> sum(m.perm <= m.d)
[1] 123
```

```
> x = pollancres$agost
> y = pollancres$novembre
>
> # Taula amb totes les possibles diferències entre x[i] i x[j]:
> difs.x = outer(x, x, "-")
> # Descartem les diferències de la diagonal (i == j) i de la meitat triangular superior:
> difs.x = difs.x[ltri <- lower.tri(difs.x)]
> # Totes les possibles diferències entre y[i] i y[j]:
> difs.y = outer(y, y, "-")[ltri]
> #
> # Nombre de concordances:
> concor = sum(sign(difs.x)*sign(difs.y) > 0)
> concor
[1] 33
> # Nombre de discordances:
> discor = sum(sign(difs.x)*sign(difs.y) < 0)
> discor
[1] 43
> cor.test(rank(x), rank(y))

        Pearson's product-moment correlation

data:  rank(x) and rank(y)
t = -0.4612, df = 11, p-value = 0.6536
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6401437  0.4471824
sample estimates:
      cor
-0.137741

> # Coeficient de correlació lineal de Pearson sobre la mostra original:
> r = cor(x,y)
> # Una permutació aleatòria del vector 'y':
> y.perm = sample(y, replace = FALSE)
> # 9999 permutacions aleatòries i càlcul de la correlació:
> nperm = 9999
> set.seed(5719)
> r.perms = replicate(nperm, cor(x, sample(y, replace = FALSE)))
> sum(r.perms >= r)
[1] 4291
```

```
> sum(abs(r.perms) >= abs(r))
[1] 9645
> sum(r.perms <= r)
[1] 5708
```