Estadística Matemàtica

Dr. Francesc Carmona
Departament d'Estadística
Universitat de Barcelona

17 de juny de 2002

8 Estadística no paramètrica **61**
8.1 Introducció 61
8.2 Test dels signes 62
8.2.1 Test per a la mediana 62
8.2.2 Test dels signes per a dades aparellades 64
8.2.3 Test per a dades binaries 65
8.2.4 Test de McNemar 66
8.3 Test dels rangs amb signe de Wilcoxon 66
8.4 El test U de Mann-Whitney 70
8.5 Comparació de medianes 73
8.6 Test d'homogeneïtat de Kolmogorov-Smirnov 75
8.7 Test H de Kruskal-Wallis 77
8.8 Test de Friedman 79
8.9 Test de ratxes 81
8.9.1 Aleatorietat d'una mostra numèrica 83
8.9.2 Comparació de dues mostres 84
8.9.3 Distribució exacta de R 84
8.10 Coeficients de correlació 85
8.10.1 Coeficient τ de Kendall 85
8.10.2 Coeficient de correlació per rangs de Spearman 87
8.11 Solució de problemes amb R o S-PLUS 89
8.12 Taules 97

Capítol 8

Estadística no paramètrica

8.1 Introducció

En aquest capítol presentarem de forma breu alguns tests no paramètrics per a problemes d'una, dues o vàries mostres. L'objectiu d'aquests tests és disposar d'alternatives a les proves d'hipòtesis de comparació clàssiques quan no es coneix la forma de la distribució de les dades o la llei de les variables. En particular seran alternatives als tests sobre la mitjana o comparació de mitjanes quan no es verifica la suposició de normalitat de les dades. Ens referim als tests basats en poblacions normals com a *contrastos paramètrics* ja que es basen en comparar mitjanes o paràmetres de la llei normal. En contraposició, els que considerem aquí i que denominarem *contrastos no paramètrics* poden comparar medianes, quantils o fins i tot tota la distribució en bloc. Observem doncs que *no paramètrics* no significa que aquests tests no compararin algun paràmetre com la mediana, més aviat significa que no volem fer determinades suposicions sobre la funció de distribució de les variables. Des d'un punt de vista pràctic, que també és el que adopten molts programes informatics d'anàlisi estadística, distingirem entre:

- Problemes d'una mostra
- Problemes de dues mostres amb dades aparellades
- Problemes de dues mostres independents
- Problemes de k mostres independents.

Aquesta distinció ens permet classificar les tècniques que estudiem i comparar-les amb les corresponents proves paramètriques:

Problema	Test paramètric	Test no paramètric
Una mostra	Test t d'una mostra	Test dels signes Test dels rangs signats
Dades aparellades	Test t de dades aparellades	Test dels signes Test dels rangs signats
Dues mostres ind.	Test t per a dues mostres ind. (amb test F previ)	Test U de Mann-Whitney
k mostres ind.	ANOVA d'un factor	Test de Kruskal-Wallis

A més, alguns tests no paramètrics tenen d'altres utilitats com els tests d'aleatorietat, els tests de ratxes, etc

8.2 Test dels signes

8.2.1 Test per a la mediana

Sigui X una variable aleatòria amb distribució continua F_X desconeguda i $M = Q_{50}$ la seva mediana o quantil 50%, és a dir, el valor tal que

$$P(X \leq M) = 0.5$$

Suposem que volem contrastar les hipòtesis

$$\begin{aligned} H_0 &: M = m_0 \\ H_1 &: M \neq m_0 \end{aligned}$$

Donada una mostra $x_1, x_2, \ldots, x_n$, considerem el "signe" de cada valor mostral per comparació amb la mediana proposada per la hipòtesi H_0, és a dir,

$$\text{signe}(x_i) = \begin{cases} + & \text{si } x_i > m_0 \\ - & \text{si } x_i < m_0 \end{cases}$$

L'estadístic $B =$ "Nombre de signes positius" és:

$$B(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} I_{x_i > m_0}$$

on

$$I_{x_i > m_0} = \begin{cases} 1 & \text{si } x_i > m_0 \ (\text{signe}(x_i) = +) \\ 0 & \text{si } x_i < m_0 \ (\text{signe}(x_i) = -) \end{cases}$$

Si la hipòtesi nul·la és certa la distribució de l'estadístic B serà una binomial de paràmetres n i $1/2$ i és raonable esperar que B prengui valors pròxims a

$n/2$, mentre que quan sigui falsa és d'esperar que prengui valors a les cues de la distribució. Així doncs, una regió crítica per al test serà aquella en que el nombre de signes positius sigui massa alt o massa baix com per ser coherent amb la hipòtesi nul·la que implica que n'hi ha tants de positius com negatius. Podem agafar com regió crítica:

$$W = \{B(\mathbf{x}) \leq b_{\alpha/2}\} \cup \{B(\mathbf{x}) \geq b_{1-\alpha/2}\}$$

on $b_{\alpha/2}$ i $b_{1-\alpha/2}$ es determinen de forma que la probabilitat de les dues cues d'una distribució $B(n, \frac{1}{2})$ sigui igual al nivell de significació α (o una mica menor que α), és a dir,

$$\sum_{i=0}^{b_{\alpha/2}} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i} + \sum_{i=b_{1-\alpha/2}}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{n-i} \leq \alpha$$

Observacions

- Encara que amb probabilitat teòrica zero, perquè la variable considerada té funció de distribució continua, es pot donar el cas $x_i = m_0$ de signe indefinit. Si no és possible augmentar la precisió, s'aconsella eliminar aquest valor mostral i descomptar-ho conseqüentment de la mida de la mostra.

- Si la hipòtesi alternativa és $M < m_0$ o bé $M > m_0$ la regió crítica s'adapta a aquesta hipòtesi de la manera raonable, és a dir:

$$H_1 : M < m_0 \quad \Rightarrow \quad W = \{B(\mathbf{x}) \leq b_\alpha\}$$
$$H_1 : M > m_0 \quad \Rightarrow \quad W = \left\{B(\mathbf{x}) \geq b_{1-\alpha/2}\right\}$$

- Alguns llibres porten taules de la distribució binomial que podem fer servir per trobar els valors crítics. Per a $n \geq 20$ podem aproximar la binomial per una normal.

Exemple 8.2.1 *La següent taula recull una mostra de 40 notes en un examen. Contrasteu amb un nivell de significació 0.05 la hipòtesi que el valor mitjà (mediana) de les notes és 66.*

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52
67	95	70	43	70	73	57	64	60	83
73	40	78	70	64	86	76	62	95	66
+	+	−	−	+	0	+	−	+	−
+	−	+	+	+	−	+	+	−	−
+	+	+	−	+	+	−	−	−	+
+	−	+	+	−	+	+	−	+	0

Solució:
Si restem 66 de les notes observades i retenim només els signes de les diferències, s'obtenen 23 signes +, 15 signes − i 2 zeros. Descartats els zeros, $B = 23$ sobre un total de 38. Si fem un contrast bilateral amb l'aproximació normal, la regió d'acceptació és $\{-1.96 \leq z \leq 1.96\}$.
Donat que

$$z = \frac{(23 - 0.5) - 38 \cdot 0.5}{\sqrt{38 \cdot 0.5 \cdot 0.5}} = 1.14$$

acceptem la hipòtesi que la mediana és 66, al nivell 0.05.

8.2.2 Test dels signes per a dades aparellades

El test dels signes pot servir també en el cas de dades aparellades.
Considerem una mostra de dues variables X, Y

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

amb n observacions en dues situacions el més homogènies possible.
Suposem que les distribucions de les dues variables són semblants, excepte potser en un paràmetre de localització com la mediana. És a dir, les dues situacions considerades només poden traslladar la distribució i no modifiquen la forma.
Ara volem contrastar la hipòtesi que no hi ha diferència entre les dues situacions: les diferències observades entre els valors x_i i y_i són degudes a l'atzar, és a dir, les dues mostres $x_1, \ldots, x_n$ i $y_1, \ldots, y_n$ procedeixen de la mateixa població. Això es pot expressar estadísticament amb la hipòtesi H_0 d'igualtat de les distribucions de probabilitat, que amb les suposicions assumides és equivalent a la igualtat de medianes.
Si la hipòtesi H_0 és certa, i la distribució de la variable diferència $D = X - Y$ és simètrica en l'origen, es verificarà

$$P(X > Y) = P(X - Y > 0) = \frac{1}{2}$$

Així doncs, podem aplicar el test dels signes a la variable diferència $D = X - Y$. En general, però no necessàriament sempre, s'agafarà com valor de m_0 el 0.

Exemple 8.2.2 *Es vol comparar el número de peces defectuoses produïdes per dues màquines diferents. S'observa la producció en 10 dies, amb la mateixa producció diària per a les dues màquines encara que diferent cada dia,* *i els resultats són:*

Dia	1	2	3	4	5	6	7	8	9	10
Màquina 1	46	110	70	54	60	120	82	76	37	28
Màquina 2	42	87	75	50	48	108	80	67	40	25

Amb un nivell de significació $\alpha = 0.06$*, podem acceptar que la primera màquina produeix més peces defectuoses?*

Solució:
El fet que la producció total diària d'ambdues màquines sigui la mateixa permet considerar les dades aparellades. Que la producció diària sigui diferent cada dia aconsella utilitzar un test no paramètric.
Observem els signes de les diferències

Dia:	1	2	3	4	5	6	7	8	9	10
Signe:	+	+	−	+	+	+	+	+	−	+

De manera que $B = 8$ sobre $n = 10$. En aquest contrast, la regió crítica és unilateral i concretament és $W = \{8, 9, 10\}$, ja que

$$P(B \geq 8) = \sum_{i=8}^{10} \binom{10}{i} 0.5^{10} = 0.0547 < \alpha = 0.06$$

Donat que la freqüència observada és 8 i pertany a la regió crítica, rebutgem la igualtat i podem acceptar que la màquina 1 produeix més peces defectuoses.

8.2.3 Test per a dades binaries

En el cas d'una mostra de valors d'una variable dicotòmica com per exemple

$$a, a, b, b, b, a, a, b, a, \ldots, b$$

podem aplicar el test dels signes per contrastar l'equilibri de les probabilitats d'ambdós valors.

Exemple 8.2.3 *Davant d'un canvi en un servei públic es fa una enquesta a 300 persones, a les quals se'ls demana si el servei ha millorat o empitjorat, sense possibilitat de ser indiferent. Ha resultat que 197 persones han dit que el servei ha millorat i volem contrastar aquest fet amb un nivell de significació del 0.01.*

Solució:
Sota la hipòtesi nul·la d'equilibri, el número B de persones que afirmen que el servei ha millorat segueix una distribució binomial $B(300, 0.5)$ de forma que

$$z = \frac{(197 - 0.5) - 150}{\sqrt{300 \cdot 0.5 \cdot 0.5}} = 5.37$$

La regió crítica d'una cua és $W = \{z > 2.33\}$ per a un $\alpha = 0.01$, de manera que acceptem l'opinió que el servei ha millorat.

8.2.4 Test de McNemar

És una variant del test dels signes. Suposem que un conjunt d'individus es classifiquen en dues categories oposades, que podem indicar amb els signes $+$ i $-$. Després d'algun estímul, és possible que alguns individus canviïn de categoria, de forma que s'obté la taula de freqüències

		Després	
		$-$	$+$
Abans	$+$	a	b
	$-$	c	d

$$a + b + c + d = n$$

Només $a+d$ individus han canviat. Sota la hipòtesi nul·la que les proporcions no canvien, les probabilitats són

$$P(+ \to -) = P(- \to +) = 1/2$$

de manera que la freqüència esperada en aquest dos casos és $(a+d)/2$. Podem aplicar el test khi-quadrat

$$\chi^2 = \frac{(a - (a+d)/2)^2}{(a+d)/2} + \frac{(d - (a+d)/2)^2}{(a+d)/2} = \frac{(a-d)^2}{a+d} \qquad \text{amb 1 g.ll.}$$

Rebutjarem la hipòtesi d'equilibri si $\chi > \chi^2_\alpha$, on α és el nivell de significació. Si les freqüències són petites és convenient utilitzar la correcció de Yates

$$\chi = \frac{(|a-d| - 1)^2}{a+d}$$

8.3 Test dels rangs amb signe de Wilcoxon

Vist com extensió de l'anterior test dels signes, la idea d'aquest test és fer servir, a més del signe, la magnitud de les diferències.

El *rang* d'una observació és la posició que aquesta ocupa en la mostra ordenada. Per exemple, si considerem la mostra

$$x_1 = 3 \quad x_2 = 0 \quad x_3 = 5 \quad x_4 = 1.9$$

la mostra ordenada és

$$x_{(1)} = 0 \quad x_{(2)} = 1.9 \quad x_{(3)} = 3 \quad x_{(4)} = 5$$

de manera que els rangs valen:

$$r(0) = 1 \quad r(1.9) = 2 \quad r(3) = 3 \quad r(5) = 4$$

Una part important de l'estadística no paramètrica ha sorgit de la substitució dels valors quantitatius de les mostres pels seus rangs i l'obtenció d'estadístics de test anàlegs als utilitzats amb dades quantitatives com, per exemple, un test equivalent al test t de Student però basat en rangs o un coeficient de correlació amb la mateixa fórmula que el de Pearson però fent servir rangs.
Suposem que la hipòtesi nul·la és la mateixa que en el test de la mediana, és a dir:

$$\begin{aligned} H_0 &: M = m_0 \\ H_1 &: M \neq m_0 \end{aligned}$$

on M representa la mediana d'una variable o, sovint, de la diferència entre dues variables aparellades.
Wilcoxon proposà considerar els estadístics:

$$\begin{aligned} T^+ &= \text{Suma dels rangs de les observacions amb signe } + \\ T^- &= \text{Suma dels rangs de les observacions amb signe } - \end{aligned}$$

$$T^+ = \sum_{i=1}^{n} r(|x_i - m_0|) I_{x_i > m_0}$$

Si H_0 és certa, llavors esperem que $T^+ = T^-$.
L'estadístic T^+ es coneix amb el nom d'estadístic de Wilcoxon i està tabulat de forma que podem trobar uns valors $t_{\alpha/2}$ i $t_{1-\alpha/2}$ tals que

$$P[T^+ < t_{\alpha/2} | H_0] + P[T^+ > t_{1-\alpha/2} | H_0] \leq \alpha$$

i definir la regió crítica com

$$W = \{T^+ < t_{\alpha/2}\} \cup \{T^+ > t_{1-\alpha/2}\}$$

Observacions

- Per a valors grans de n es pot fer servir el fet que, sota H_0, l'estadístic T^+ és assimptòticament normal:

$$T^+ \sim AN(\mu_{T^+}, \sigma_{T^+}), \quad \mu_{T^+} = \frac{n(n+1)}{4}, \quad \sigma^2_{T^+} = \frac{n(n+1)(2n+1)}{24}$$

  i per tant per mostres grans podem basar-nos en l'estadístic

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

- Una alternativa a l'estadístic de test anterior és considerar l'estadístic

$$T = \text{mín}(T^+, T^-).$$

  Si H_0 és certa, llavors $T^+ = T^-$. Si no és certa, tindrem $T^+ > T^-$ o bé $T^+ < T^-$, de forma que el mínim serà un valor "petit". El test basat en aquest estadístic rebutjarà la hipòtesi nul·la si T és més petit que T_α on aquest valor crític s'obté d'una taula diferent de la taula de valors crítics per a T^+.

Exemple 8.3.1 *Donat que en l'exemple 8.2.1 les notes són numèriques podem utilitzar el test dels rangs amb signe per contrastar si $H_0 : M = 66$ en front de $H_1 : M \neq 66$ amb un nivell de significació del 0.05.*

Solució:
Per calcular l'estadístic T^+ hem d'assignar els rangs corresponents als valors positius de les diferències entre les observacions i el valor 66 proposat a la hipòtesi nul·la. Això es fa de forma relativament simple amb un full de càlcul com EXCEL. En la taula 8.1 podem observar l'ordenació que es fa en funció dels valors absoluts de les diferències. Observem també que cal repartir els rangs quan hi ha empats i que els dos zeros queden exclosos.
El resultat és que $T^+ = 465$ amb un $n = 38$, de manera que

$$z = \frac{465 - (38 \cdot 39)/4}{\sqrt{(38 \cdot 39 \cdot 77)/24}} = 1.37$$

que queda dins de la regió d'acceptació $\{-1.96 \leq z \leq 1.96\}$.

Exemple 8.3.2 *Donat que en l'exemple 8.2.2 les observacions són numèriques i aparellades, podem utilitzar els valors de les diferències amb el test dels rangs amb signe per contrastar si hi ha diferències entre les dues màquines.*

notes	notes-66
71	5
67	1
55	-11
64	-2
82	16
66	0
74	8
58	-8
79	13
61	-5
78	12
46	-20
84	18
93	27
72	6
54	-12
78	12
86	20
48	-18
52	-14
67	1
95	29
70	4
43	-23
70	4
73	7
57	-9
64	-2
60	-6
83	17
73	7
40	-26
78	12
70	4
64	-2
86	20
76	10
62	-4
95	29
66	0

dif	abs(dif)	rang	rang(+)	rang(-)
0	0			
0	0			
1	1	1,5	1,5	
1	1	1,5	1,5	
-2	2	4		4
-2	2	4		4
-2	2	4		4
4	4	7,5	7,5	
4	4	7,5	7,5	
4	4	7,5	7,5	
-4	4	7,5		7,5
5	5	10,5	10,5	
-5	5	10,5		10,5
6	6	12,5	12,5	
-6	6	12,5		12,5
7	7	14,5	14,5	
7	7	14,5	14,5	
8	8	16,5	16,5	
-8	8	16,5		16,5
-9	9	18		18
10	10	19	19	
-11	11	20		20
12	12	22,5	22,5	
-12	12	22,5		22,5
12	12	22,5	22,5	
12	12	22,5	22,5	
13	13	25	25	
-14	14	26		26
16	16	27	27	
17	17	28	28	
18	18	29,5	29,5	
-18	18	29,5		29,5
-20	20	32		32
20	20	32	32	
20	20	32	32	
-23	23	34		34
-26	26	35		35
27	27	36	36	
29	29	37,5	37,5	
29	29	37,5	37,5	
		suma=	465	276

Taula 8.1: Taula per a la suma dels rangs de l'exemple 8.3.1

Solució:
Primer calculem les diferències i les ordenem pel seu valor absolut (veure la taula 8.2). En aquest cas hi ha molts empats i s'han de repartir els rangs. L'estadístic T^+ és 46.5 amb un $n = 10$, un valor clarament superior a l'esperat si H_0 és certa.
Amb un nivell de significació de 0.05, la regió crítica que s'obté a la taula de Wilcoxon és $\{T^+ > 44\}$, de forma que rebutgem H_0 i admetem que la primera màquina fabrica més peces defectuoses.

maq1	maq2	dif
46	42	4
110	87	23
70	75	-5
54	50	4
60	48	12
120	108	12
82	80	2
76	67	9
37	40	-3
28	25	3

dif	abs(dif)	rang	rang(+)
2	2	1	1
-3	3	2,5	
3	3	2,5	2,5
4	4	4,5	4,5
4	4	4,5	4,5
-5	5	6	
9	9	7	7
12	12	8,5	8,5
12	12	8,5	8,5
23	23	10	10
		suma=	46,5

Taula 8.2: Taula per a la suma dels rangs de l'exemple 8.3.2

8.4 El test U de Mann-Whitney

Aquest test permet comparar dues poblacions amb mostres independents. Suposem que obtenim dues mostres independents

$$(x_1, ..., x_{n_1}), (y_1, ..., y_{n_2})$$

de dues poblacions X, Y amb funcions de distribució F_X, F_Y respectivament. Volem contrastar la hipòtesi $H_0 : F_X = F_Y$ front alguna de les alternatives

$$H_1 : F_X \neq F_Y \qquad H_1 : F_X < F_Y \qquad H_1 : F_X > F_Y$$

Si la hipòtesi nul·la és certa, llavors $P(X < Y) = \frac{1}{2}$. A més, com hi ha $n_1 \cdot n_2$ parelles possibles, el nombre de parelles d'observacions (x_i, y_j) que esperem que verifiquin $x_i < y_j$ estarà al voltant de $\frac{n_1 \cdot n_2}{2}$.
Un estadístic de test raonable per decidir si acceptem o rebutgem la hipòtesi nul·la és el nombre de parelles que verifiquen $x_i < y_j$ que definim com:

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I_{x_i < y_j}$$

Una desviació significativa de U respecte al valor esperat $\frac{n_1 \cdot n_2}{2}$ farà rebutjar la hipòtesi nul·la. Per decidir si U és significatiu consultarem la taula de Mann-Whitney-Wilcoxon que permet decidir si rebutgem H_0 en funció del nivell de significació escollit i de les mides mostrals n_1 i n_2.

Observacions

- Un procediment alternatiu per calcular U, i sovint més còmode, consisteix en formar la mostra conjunta, reunint les dues individuals, i assignar els rangs $1, 2, ..., n_1 + n_2$ a cadascun dels valors de la mostra ordenada. Es pot calcular U a partir de la relació:

$$U = W - \frac{n_2(n_2+1)}{2}$$

on W és la suma dels rangs de les observacions y_j

$$W = \sum_{j=1}^{n_2} r(y_j)$$

Aquest estadístic W per comparar dues poblacions va ser proposat per Wilcoxon però, per la relació anterior, és equivalent a l'estadístic U de Mann-Whitney.

- Si no hi ha ligadures o empats, la relació entre l'estadístic de Wilcoxon W (suma de rangs corresponents a les observacions Y) i l'estadístic U de Mann-Whitney (número de vegades que $x_i < y_j$ a la mostra conjunta ordenada) és

  $$W = \frac{n_2(n_2+1)}{2} + U$$

  Si W' és la suma dels rangs corresponents a les observacions X, llavors

  $$W + W' = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$$

  De forma que si U' és el número de vegades que $y_j < x_i$, resulta

  $$U + U' = n_1 n_2 \qquad W' = \frac{n_1(n_1+1)}{2} + U'$$

- Donats n_1, n_2 $(n_1, n_2 \leq 10)$ la taula de Mann-Whitney [1] proporciona l'enter a_p i la probabilitat p tal que

  $$P(U \leq a_p) = p$$

  Veiem com s'utilitzen aquests valors.

  En una prova unilateral amb hipòtesi alternativa $H_1 : F_X < F_Y$, el criteri de decisió serà rebutjar la hipòtesi nul·la si $\{U \leq a_p\}$, on p és l'aproximació per defecte del nivell de significació. Això és perquè, si és certa l'alternativa, llavors $F_X(s) \leq F_Y(s)$ $\forall s$ i $F_X(s) < F_Y(s)$ per a algun s (es diu que X és estocàsticament més gran que Y), el que

[1] Això depèn dels autors i cal mirar amb cura la definició de cada taula

implica que $P(X > Y) > 1/2$, de manera que és probable que U sigui petit.

L'estadístic U ha de verificar $U < n_1 n_2/2$, en cas contrari s'utilitza

$$U' = n_1 n_2 - U$$

i la regla és la mateixa, però fent servir U'.

Si l'alternativa és $H_1 : F_X > F_Y$, podem intercanviar X i Y.

En una prova bilateral es buscarà $a_{p/2}$ de forma que $P(U \leq a_{p/2}) = p/2$, on p és l'aproximació per defecte del nivell de significació, i la regió de rebuig és

$$\{U \leq a_{p/2}\} \cup \{U' \leq a_{p/2}\}$$

- Per a mostres "grans" es pot fer servir el fet que, sota H_0, l'estadístic U és assimptòticament normal:

$$U \sim AN(\mu_U, \sigma_U), \quad \mu_U = \frac{n_1 n_2}{2}, \quad \sigma_U^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

i per tant, per a $n_1 > 10$ o $n_2 > 10$ podem basar-nos en un estadístic de test

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

Exemple 8.4.1 *Per tal de comparar la resistència en kg/cm^2 d'un material subministrat per dos proveïdors es van mesurar dues mostres d'uns quants elements:*

Proveïdor A	*202, 229, 215, 220, 223, 233, 208, 228, 209*
Proveïdor B	*221, 207, 185, 203, 187, 190, 195, 204, 212*

Amb un nivell de significació del 0.05, indiqueu si hi ha diferències entre els materials subministrats pels dos proveïdors.

Solució:
És clar que es tracta de comparar dues mostres independents i que el fet que $n_1 = n_2 = 9$ no és important. La taula 8.3 indica el càlcul de les sumes de rangs per a cada mostra, de forma que

$$W = 56 \qquad W' = 115 \qquad W + W' = 171 = \frac{(9+9)(9+9+1)}{2}$$

d'on s'obté

$$U = W - \frac{9 \cdot 10}{2} = 56 - 45 = 11 < \frac{n_1 n_2}{2} = 40.5$$

La regió crítica per $n_1 = n_2 = 9$ i $\alpha = 0.05$ és $\{U \leq 22\}$, de manera que el valor observat $U = 11$ cau en ella i implica el rebuig de l'equivalència de les dades.

Proveïdor	x	rang	rang(A)	rang(B)
B	185	1		1
B	187	2		2
B	190	3		3
B	195	4		4
A	202	5	5	
B	203	6		6
B	204	7		7
B	207	8		8
A	208	9	9	
A	209	10	10	
B	212	11		11
A	215	12	12	
A	220	13	13	
B	221	14		14
A	223	15	15	
A	228	16	16	
A	229	17	17	
A	233	18	18	
		suma=	115	56

Taula 8.3: Taula per a la suma dels rangs de l'exemple 8.4.1

8.5 Comparació de medianes

Considerem una situació en la que es vol comparar dues poblacions continues amb distribucions d'igual forma i tractar de detectar desplaçaments entre ambdues distribucions.

Siguin $x_1, \ldots, x_{n_1}$ i $y_1, \ldots, y_{n_2}$ dues mostres aleatòries corresponents a cada població i independents entre sí. Si s'ordenen conjuntament les dues mostres en ordre creixent i es considera la mediana M de la mostra combinada, podem calcular l'estadístic

$$T = \sum_{i=1}^{n_1} I_{x_i < M}$$

que serveix per contrastar la hipòtesis $H_0 : M_X = M_Y$.

Si ambdues poblacions tenen la mateixa distribució, és d'esperar que T sigui pròxim a $n_1/2$. En canvi, si T resulta molt més gran que $n_1/2$, és raonable suposar que la mediana M_X de la primera població és inferior a la de la segona M_Y; mentre que si T és molt menor que $n_1/2$, això sembla indicar que M_X és superior a M_Y. Les regions crítiques són:

Alternativa	**Regió crítica**
$M_X < M_Y$	$\{T \geq k\}$
$M_X > M_Y$	$\{T \leq k'\}$
$M_X \neq M_Y$	$\{T \leq k_1\} \cup \{T \geq k_2\}$

Si la distribució d'ambdues poblacions és la mateixa, la distribució de T es pot trobar amb facilitat. Donat que les n_1+n_2 observacions són independents i igualment distribuïdes, les $\binom{n_1+n_2}{n_1}$ maneres d'assignar n_1 a la primera mostra (i les altres n_2 a la segona) són equiprobables. Si p és la part sencera de $(n_1+n_2)/2$, hi ha p de les n_1+n_2 observacions inferiors a M i serà $T=t$ en totes les assignacions en les que resultin t de la primera mostra d'entre les p primeres i n_1-t entre les n_1+n_2-p últimes. Així

$$P(T=t)=\frac{\binom{p}{t}\binom{n_1+n_2-p}{n_1-t}}{\binom{n_1+n_2}{n_1}}$$

on t pot variar entre $\max\{0,p-n_2\}$ i $\min\{n_1,p\}$. Es tracta doncs d'una distribució hipergeomètrica que pot aproximar-se, si n_1 i n_2 són grans, per una $N(n_1/2,\sqrt{n_1n_2/4(n_1+n_2)})$.

Exemple 8.5.1 *Amb les dades de l'exemple 8.4.1, calculeu l'estadístic T i compareu les medianes de les dues mostres.*

Solució:
La mediana conjunta és $M=208.5$, de manera que és evident que $T=2$. Els càlculs per a una distribució hipergeomètrica amb $n_1=n_2=9$ i $p=9$ ens proporcionen

$$P(T\leq 2)=P(T\geq 7)=0.02834$$

Unes cues que sumen una mica més del 0.05. El valor observat $T=2$ cau a la regió crítica i, en conseqüència, rebutgem la igualtat de medianes.

Un procediment alternatiu es pot aplicar amb el test khi-quadrat.
Considerem la mediana M de tots els valors mostrals conjuntament i dividim cada mostra original en dos grups: aquells que prenen valors inferiors o iguals a la mediana i els que prenen valors superiors.
S'obtenen així quatre classes d'efectius com es recull en la taula:

	Grup I	Grup II	
Observacions amb valors inferiors a la mediana	n_{11}	n_{12}	$n_{1\bullet}$
Observacions amb valors superiors a la mediana	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}=n_1$	$n_{\bullet 2}=n_2$	

Es calcula l'estadístic χ^2 ja que es tracta d'una comparació de proporcions o test d'homogeneïtat en una taula 2×2. La regió crítica es determina a partir de la distribució khi-quadrat o les alternatives estudiades.

Exemple 8.5.2 *Amb les dades de l'exemple 8.4.1, calculeu l'estadístic χ^2 i compareu les medianes de les dues mostres.*

Solució:
Com ja sabem, la mediana comuna de les dues mostres és $M = 208.5$. Llavors la taula pel test d'homogeneïtat resulta

	Pro. A	Pro. B	
valors inferiors	2	7	9
valors superiors	7	2	9
Total	9	9	

Així hem de calcular l'estadístic khi-quadrat amb la correcció de Yates

$$\chi^2 = \frac{(|2 \cdot 2 - 7 \cdot 7| - 18/2)^2}{9 \cdot 9 \cdot 9 \cdot 9} 18 = 3.556$$

Amb un grau de llibertat i per un nivell de significació del 0.05, la regió crítica comença en $\chi^2_{0.05} = 3.841$, de forma que podem acceptar la hipòtesi nul·la.

Amb mides mostrals grans, el test khi-quadrat és preferible si no hi ha constància que la forma d'ambdues distribucions sigui la mateixa, ja que el test T anterior té tendència a acceptar la homogeneïtat si $M_X = M_Y$ encara que la forma de les distribucions sigui diferent.
Per la mateixa raó es preferible el test de Kolmogorov-Smirnov que expliquem a la següent secció.

8.6 Test de Kolmogorov-Smirnov per a l'homogeneïtat

Quan disposem de dues mostres independents $x_1, x_2, \ldots, x_{n_1}$ i $y_1, y_2, \ldots, y_{n_2}$ de dues poblacions amb distribucions desconegudes F_X i F_Y respectivament i volem contrastar la seva coincidència, és a dir, la hipòtesi $H_0 : F_X = F_Y$ podem comparar les distribucions empíriques associades a cada mostra. Això és possible si coneixem els valors exactes de les observacions i, en aquest aspecte, és millor aquesta comparació que el test khi-quadrat d'homogeneïtat que utilitza freqüències i necessita moltes dades de cada població.

Les distribucions empíriques són:

$$F_{n_1}(z) = \frac{1}{n_1}\sum_{i=1}^{n_1} I_{x_i<z} \qquad G_{n_2}(z) = \frac{1}{n_2}\sum_{i=1}^{n_2} I_{y_i<z}$$

i l'estadístic de Kolmogorov-Smirnov

$$\Delta_{n_1,n_2} = \sup_{z\in\mathbb{R}} |F_{n_1}(z) - G_{n_2}(z)|$$

Si la hipòtesi H_0 és certa, les dues distribucions empíriques han d'estar molt pròximes i la mesura global de discrepància Δ_{n_1,n_2} serà petita. Pel contrari, quan $F_X \neq F_Y$ el valor de Δ_{n_1,n_2} serà més gran, de manera que la regió crítica que hem de considerar és de la forma

$$\{\Delta_{n_1,n_2} > a\}$$

El test es basa en el Teorema de Smirnov.

Teorema 8.1 *Si les distribucions continues de les dues poblacions coincideixen $F_X = F_Y$ i $n_1 \to \infty, n_2 \to \infty$, llavors per a cada λ*

$$P\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}}\Delta_{n_1,n_2} \leq \lambda\right) \to Q(\lambda) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2\lambda^2}$$

on $Q(\lambda)$ és la distribució asimptòtica de Kolmogorov-Smirnov.

Per aplicar aquest resultat, primer determinarem a la taula de Kolmogorov-Smirnov el valor tal que $Q(\lambda_\alpha) = 1-\alpha$, on α és el nivell de significació. Llavors, tot suposant que n_1 i n_2 són grans, acceptarem H_0 si

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\Delta_{n_1,n_2} \leq \lambda_\alpha$$

i rebutjarem H_0 en cas contrari.
Per a valors petits de n_1 i n_2 existeixen unes taules calculades per Massey que proporcionen els valors crítics a_α, de forma que rebutgem H_0 si $\Delta_{n_1,n_2} > a_\alpha$. Alguns llibres porten exclusivament la taula en el cas $n_1 = n_2$.

Exemple 8.6.1 *Amb les dades de l'exemple 8.4.1, calculeu l'estadístic de Kolmogorov-Smirnov i compareu les distribucions de les dues mostres.*

Solució:
A la taula 8.4 veiem els càlculs fets amb un full EXCEL. Observem els increments de les freqüències a raó de 1/9.
L'estadístic és $\Delta = 0.667$ i cau justament a la frontera de la regió crítica $\Delta_{9,9} > 0.667$ de la taula de Massey pel nivell de significació 0.05 i $n_1 = n_2 = 9$.

Proveïdor	x	F	G	abs(dif)
B	185	0,000	0,111	0,111
B	187	0,000	0,222	0,222
B	190	0,000	0,333	0,333
B	195	0,000	0,444	0,444
A	202	0,111	0,444	0,333
B	203	0,111	0,556	0,444
B	204	0,111	0,667	0,556
B	207	0,111	0,778	0,667
A	208	0,222	0,778	0,556
A	209	0,333	0,778	0,444
B	212	0,333	0,889	0,556
A	215	0,444	0,889	0,444
A	220	0,556	0,889	0,333
B	221	0,556	1,000	0,444
A	223	0,667	1,000	0,333
A	228	0,778	1,000	0,222
A	229	0,889	1,000	0,111
A	233	1,000	1,000	0,000
			max=	0,667

Taula 8.4: Taula per calcular Δ amb les dades de l'exemple 8.4.1

8.7 Test H de Kruskal-Wallis

El test U és un test no paramètric per decidir si dues mostres independents provenen o no de la mateixa població. El test H de Kruskal-Wallis és una generalització per a k mostres agafades en k poblacions. Així doncs, és una versió no paramètrica d'un ANOVA d'un factor.
Considerem k mostres de mides $n_1, n_2, \ldots, n_k$ recollides en les k poblacions i tals que $n_1 + n_2 + \cdots + n_k = n$. Suposem que ordenem totes les observacions de forma conjunta i calculem les sumes dels rangs per a les k mostres $R_1, R_2, \ldots, R_k$, respectivament. Si definim l'estadístic

$$H = \left(\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(n+1)$$

es demostra que, si existeix homogeneïtat entre les distribucions dels k grups, la seva distribució en el mostratge està molt pròxima a una khi-quadrat amb $k-1$ graus de llibertat quan les mides mostrals n_i són grans. Així, exigirem sempre que $n_1, n_2, \ldots, n_k$ siguin tots ells superiors a 5. Per a valors petits és necessari consultar unes taules especials.

Observacions

- L'estadístic de Kruskal-Wallis es pot posar en la forma

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i (R_{\bullet i} - R_{\bullet\bullet})^2$$

on $R_{\bullet i} = R_i/n_i$, $R_{\bullet\bullet} = (n+1)/2$. D'aquesta forma, el test basat en H si sembla molt al test F en un disseny d'un factor i rèpliques.

- Si existeixen observacions repetides, l'estadístic H es corregeix amb un factor de manera que el nou estadístic és

$$H' = \frac{H}{1 - \frac{\sum_{j=1}^{r}(t_j^3 - t_j)}{n^3 - n}}$$

on t_j és el número d'observacions en la mostra combinada repetides per a un rang donat i r el número de repeticions. Aquesta correcció no té molt d'efecte sobre el valor de H, fins i tot en presència de moltes observacions repetides.

Exemple 8.7.1 *Es vol comparar el pes en grams d'un producte envasat per tres fabricants amb mostres de mida 6 en els tres casos.*

Fabr. A	*251*	*250*	*249*	*255*	*258*	*258*
Fabr. B	*247*	*246*	*250*	*241*	*240*	*242*
Fabr. C	*228*	*236*	*240*	*225*	*236*	*230*

Estudieu si hi ha diferències entre els tres fabricants amb el test de Kruskal-Wallis.

Solució:

Fabr.	pes	rang	rang(A)	rang(B)	rang(C)
C	225	1			1
C	228	2			2
C	230	3			3
C	236	4,5			4,5
C	236	4,5			4,5
B	240	6,5		6,5	
C	240	6,5			6,5
B	241	8		8	
B	242	9		9	
B	246	10		10	
B	247	11		11	
A	249	12	12		
A	250	13,5	13,5		
B	250	13,5		13,5	
A	251	15	15		
A	255	16	16		
A	258	17,5	17,5		
A	258	17,5	17,5		
		suma=	91,5	58	21,5

Taula 8.5: Taula per calcular H amb les dades de l'exemple 8.7.1

L'estadístic per al contrast és

$$H = \frac{12}{18 \cdot 19}\left(\frac{91.5^2}{6} + \frac{58^2}{6} + \frac{21.5^2}{6}\right) - 3 \cdot 19 = 14.34$$

El valor crític per a un nivell de significació del 0.05 que trobem a la taula de la khi-quadrat amb 2 graus de llibertat és 5.99 i, com queda superat per l'estadístic, rebutgem la igualtat entre els fabricants.
Encara que hi ha lligadures, en aquest cas no cal calcular $H' > H$.

8.8 Test de Friedman

Aquest test està pensat per comprovar si existeixen diferències significatives entre k tractaments o condicions experimentals aplicats a n individus.

	Tractament					
Individu	1	2	$\dots$	j	$\dots$	k
1	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1k}
2	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{nk}

Els individus s'han d'escollir a l'atzar i de forma independent, de forma que les files són independents entre sí. Però, com els individus són els mateixos, les columnes són dependents.
El test de Friedman serveix per provar si hi ha diferències entre els k tractaments (efecte columna), amb la presència dels efectes individuals (efecte fila). És una versió no paramètrica del disseny de dos factors sense interacció. És clar que la hipòtesi nul·la és la igualtat de resposta o d'efecte dels diferents tractaments, mentre que l'alternativa és que hi ha, com a mínim, dos tractaments amb resposta diferent.
Per calcular l'estadístic no paramètric, per cada fila per separat, s'assignen els rangs que corresponen als valors observats. Una vegada convertida la taula original en rangs, es calculen les sumes de rangs R_j per a cada columna o tractament $j = 1, \dots, k$. L'estadístic és

$$S = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

La distribució aproximada de S per a valors grans de n és una khi-quadrat amb $k-1$ graus de llibertat. Per a valors molt petits de n ($n < 10$) cal consultar unes taules especials. La regió crítica és de la forma $\{S \geq c\}$.

Quan hi ha lligadures en una fila, s'han de promitjar els rangs dels valors repetits i calcular l'estadístic de Friedman modificat per un factor de correcció

$$S' = \frac{12 \sum_{j=1}^{k} R_j^2 - 3n^2 k(k+1)^2}{nk(k+1) - \frac{1}{k-1} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{g_i} t_{ij}^3 - k \right\}}$$

on g_i és el número de grups d'observacions lligades en la fila i i t_{ij} el número d'observacions lligades en el grup j de la fila i. Quan no hi ha lligadures es considera per conveni que $g_i = k$, $t_{ij} = 1$, i llavors, el terme de correcció pel individu i és $\sum_{j=1}^{g_i} t_{ij}^3 - k = k - k = 0$. Si això passa en totes les files, llavors $S' = S$.

Exemple 8.8.1 *S'ha consultat a un grup de 12 persones per tal que opinin sobre cinc marques de xampú. En concret les seves classificacions es troben a la següent taula.*

	Xampú				
Ind.	*A*	*B*	*C*	*D*	*E*
1	*5*	*3*	*2*	*4*	*1*
2	*4*	*3*	*5*	*2*	*1*
3	*3*	*5*	*4*	*2*	*1*
4	*4*	*5*	*1*	*2*	*3*
5	*3*	*4*	*5*	*1*	*2*
6	*5*	*3*	*4*	*2*	*1*
7	*2*	*5*	*4*	*3*	*1*
8	*3*	*5*	*4*	*1*	*2*
9	*3*	*4*	*5*	*2*	*1*
10	*4*	*5*	*3*	*1*	*2*
11	*5*	*3*	*2*	*4*	*1*
12	*5*	*4*	*3*	*2*	*1*

Esbrineu si hi ha diferències significatives entre els xampus.

Solució:
En aquest cas les dades de la taula coincideixen directament amb els rangs i, a més, no hi ha lligadures. Les sumes dels rangs per columnes són

$$R_A = 46 \quad R_B = 49 \quad R_C = 42 \quad R_D = 26 \quad R_E = 17$$

De forma que l'estadístic és

$$S = \frac{12}{12 \cdot 5 \cdot 6}(46^2 + 49^2 + 42^2 + 26^2 + 17^2) - 3 \cdot 12 \cdot 6 = 25.53$$

Un valor clarament superior al de la taula de la khi-quadrat $\chi^2_{4,0.05} = 9.488$. Així hem de rebutjar la igualtat entre els xampus.

Observacions

- Es pot veure que

$$S = \frac{12n}{k(k+1)} \sum_{j=1}^{k} (R_{\bullet j} - R_{\bullet\bullet})^2$$

  on $R_{\bullet j} = R_j/n$ i $R_{\bullet\bullet} = (k+1)/2$. Això posar de manifest la relació d'aquest test amb el test F per detectar l'efecte columna en el disseny de dos factors sense interacció.

8.9 Test de ratxes

En aquesta secció ens plantegem si una mostra observada és realment una mostra aleatòria simple, és a dir, si les observacions han estat plenament independents. Qualsevol condició de mostratge sense les degudes garanties d'aleatorietat pot afectar a la independència. Per exemple, l'aprenentatge d'un aparell de mesura pot modificar les observacions amb el temps. Aquesta qüestió és molt important perquè molts mètodes de contrast que hem estudiat parteixen sempre de la consideració de mostres aleatòries simples.
Així doncs, la nostra intenció és contrastar les hipòtesis

$$H_0 : \text{la mostra observada és aleatòria}$$
$$H_1 : \text{la mostra observada no és aleatòria}$$

Pensem en una variable aleatòria que només pren el valors a i b i que observem les mostres

$$aaaaaaaabbbbbb \qquad \text{ó} \qquad abababababab$$

És evident que presenten un aspecte poc aleatori. En canvi, la sospita no apareix si, per exemple, les mostres observades són

$$abbaaabbaabbba \qquad \text{ó} \qquad baaabbabbbbaaa$$

Una ratxa és una sèrie de símbols idèntics (o relacionats) continguda entre dos símbols diferents o un de sol si estem al principi o al final de la seqüència. Si, per exemple, separem les ratxes amb una barra vertical tenim

$$\begin{array}{ll} aaaaaaaa|bbbbbb & a|b|a|b|a|b|a|b|a|b|a|b|a|b \\ a|bb|aaa|bb|aa|bbb|a & b|aaa|bb|a|bbbb|aaa \end{array}$$

En el primer cas hi ha dues ratxes, molt poques, i en el segon masses, en realitat el número màxim. En canvi, els altres dos casos són més naturals, més aleatoris. Sembla clar doncs que existeix una relació entre el número de ratxes i l'aleatorietat. La seqüència es considera no aleatòria quan hi ha masses o massa poques ratxes.
Es pot estudiar la distribució en el mostratge del número R de ratxes quan es formen totes les seqüències possibles amb n_1 símbols a i n_2 símbols b. Es demostra que la mitjana i la variància d'aquesta distribució són

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 \qquad \sigma_R^2 = \frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

Quan n_1 i n_2 són grans (superiors a 8) la distribució asimptòtica és normal i per contrastar l'aleatorietat podem utilitzar l'estadístic

$$z = \frac{R - \mu_R}{\sigma_R}$$

amb una regió crítica de dues cues.

Exemple 8.9.1 *Considerem els valors observats en la producció d'una màquina que indiquem per c quan la peça fabricada és correcte i d quan és defectuosa.*

c	c	c	c	d	d	c	c	c	c	c	c	c	c	c
c	d	d	d	c	c	c	c	c	c	c	c	c	c	c
c	c	d	c	c	c	c	c	c	d	d	c	c	c	c
c	c	d	c	c	c	d	d	c	c	c	c	c	c	c
c	c	c	c	c	c	c	c	c	c	d	d	d	c	c

Apliqueu el test de les ratxes per comprovar l'aleatorietat d'aquesta mostra.

Solució:
El recompte de les ratxes per files ens diu que $R = 15$ i $n_1 = 61, n_2 = 14$. Així doncs, els paràmetres de la distribució de R són

$$\mu_R = \frac{2 \cdot 61 \cdot 14}{61 + 14} + 1 = 23.7733$$
$$\sigma_R^2 = \frac{2 \cdot 61 \cdot 14 \cdot (2 \cdot 61 \cdot 14 - 61 - 14)}{(61 + 14)^2(61 + 14 - 1)} = 6.7007$$

i l'estadístic $z = (15 - \mu_R)/\sigma_R = -3.39$. Aquest valor cau a la regió crítica de dues cues de la distribució normal per al nivell del 0.05, de manera que hem de posar en dubte l'aleatorietat de la mostra. Les peces defectuoses apareixen de forma agrupada.

8.9.1 Aleatorietat d'una mostra numèrica

Per tal de determinar si una mostra de dades numèriques és o no aleatòria, considerem la seqüència de dades tal com van ser observades, és a dir, en el mateix ordre. Calculem la mediana i assignem el símbol $-$ o $+$ a cada observació en funció que aquest valor estigui per sota o per sobre de la mediana. Si un valor coincideix amb la mediana el suprimirem, de forma que quasi sempre resultarà $n_1 = n_2 = m$. L'aleatorietat de la mostra es contrasta amb l'estadístic R amb paràmetres

$$\mu_R = m + 1 \qquad\qquad \sigma_R^2 = \frac{m(m-1)}{2m-1}$$

Exemple 8.9.2 *Comproveu l'aleatorietat de les dades de l'exemple 8.2.1.*

Solució:
La mediana de la mostra és 70, de forma que s'obtenen els símbols de la taula.

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52
67	95	70	43	70	73	57	64	60	83
73	40	78	70	64	86	76	62	95	66
$+$	$-$	$-$	$-$	$+$	$-$	$+$	$-$	$+$	$-$
$+$	$-$	$+$	$+$	$+$	$-$	$+$	$+$	$-$	$-$
$-$	$+$		$-$		$+$	$-$	$-$	$-$	$+$
$+$	$-$	$+$		$-$	$+$	$+$	$-$	$+$	$-$

El recompte de les ratxes per files ens diu que $R = 26$ i, malauradament en aquest cas, els empats ens porten a $n_1 = 19 \neq n_2 = 18$. Així doncs, els paràmetres de la distribució de R són

$$\mu_R = \frac{2 \cdot 19 \cdot 18}{19 + 18} + 1 = 19.4865$$
$$\sigma_R^2 = \frac{2 \cdot 19 \cdot 18 \cdot (2 \cdot 19 \cdot 18 - 19 - 18)}{(19+18)^2(19+18-1)} = 8.9795$$

i l'estadístic $z = (26 - \mu_R)/\sigma_R = 2.17$. Aquest valor cau a la regió crítica de dues cues de la distribució normal per al nivell del 0.05, de manera que hem de posar en dubte l'aleatorietat de la mostra.

Una altra aplicació del test de ratxes consisteix en substituir cada valor mostral, excepte el primer, per $-$ ó $+$ segons que aquest sigui inferior o superior a l'anterior. Així detectarem si hi ha hagut períodes d'observació amb tendències de creixement o decreixement, no atribuïbles a l'atzar.

Exemple 8.9.3 *Comproveu l'aleatorietat dels períodes de creixement i decreixement de les dades de l'exemple 8.2.1.*

Solució:
La següent taula recull els símbols $-$ i $+$ que expressen si un valor és inferior o superior a l'anterior.

71	67	55	64	82	66	74	58	79	61
78	46	84	93	72	54	78	86	48	52
67	95	70	43	70	73	57	64	60	83
73	40	78	70	64	86	76	62	95	66
	$-$	$-$	$+$	$+$	$-$	$+$	$-$	$+$	$-$
$+$	$-$	$+$	$+$	$-$	$-$	$+$	$+$	$-$	$+$
$+$	$+$	$-$	$-$	$+$	$+$	$-$	$+$	$-$	$+$
$-$	$-$	$+$	$-$	$-$	$+$	$-$	$-$	$+$	$-$

El recompte de les ratxes per files ens proporciona $R = 27$ per $n_1 = 19$ i $n_2 = 20$. També, com en l'exemple anterior, és evident que el número de ratxes és excessiu.

8.9.2 Comparació de dues mostres

El test de les ratxes es pot aplicar també a la comparació de dues mostres independents, encara que proporciona un resultat més pobre que el test U de Mann-Whitney. La idea és ordenar les dues mostres de forma conjunta, assignar un símbol a la primera mostra i un altre a la segona i, finalment, comptar les ratxes que en resulten.
Si les dues mostres procedeixen de la mateixa distribució poblacional, els seus valors haurien d'estar barrejats. En canvi, si existeix una diferència de posició entre les distribucions poblacionals, els signes estaran més agrupats i el número de ratxes serà més reduït. Així doncs, en aquest cas la regió crítica és de la forma $\{R \leq k\}$.

8.9.3 Distribució exacta de R

Suposem que a la mostra hi ha n_1 símbols a i n_2 símbols b. Llavors la distribució de R, tot suposant que la mostra és aleatòria, es pot trobar explícitament:

$$P(R = 2r) = 2\frac{\binom{n_1-1}{r-1}\binom{n_2-1}{r-1}}{\binom{n_1+n_2}{n_1}}$$

$$P(R = 2r+1) = \frac{\binom{n_1-1}{r-1}\binom{n_2-1}{r} + \binom{n_1-1}{r}\binom{n_2-1}{r-1}}{\binom{n_1+n_2}{n_1}}$$

on $r \leq \min\{n_1, n_2\}$.
En efecte, $n_1 + n_2$ símbols es poden ordenar de $(n_1 + n_2)!$ formes, igualment probables. Per formar seqüències amb r ratxes de a i r ratxes de b, les a es poden ordenar de $n_1!$ formes, i dividir-se després en r grups, amb l'elecció de $r - 1$ dels $n_1 - 1$ forats existents entre ells. Anàlogament, les b poden ser ordenades de $n_2!$ maneres i dividides en r grups de $\binom{n_2-1}{r-1}$ formes. Després s'han d'intercalar els grups formats, tot començant amb el primer grup de a, o bé amb el primer grup de b. En total, la probabilitat que hi hagi r ratxes de cada tipus és

$$\frac{n_1!\binom{n_1-1}{r-1}n_2!\binom{n_2-1}{r-1}2}{(n_1+n_2)!} = 2\frac{\binom{n_1-1}{r-1}\binom{n_2-1}{r-1}}{\binom{n_1+n_2}{n_1}}$$

De manera similar es calcula la probabilitat que hi hagi r ratxes de a i $r + 1$ ratxes de b i la probabilitat que hi hagi $r + 1$ ratxes de a i r ratxes de b. Això proporciona els dos sumands de la segona expressió. Naturalment, el número de ratxes d'un tipus i de l'altre es diferencien, com a molt, en 1.

8.10 Coeficients de correlació

En aquesta secció proposem dos coeficients no paramètrics que permeten mesurar la dependència estocàstica de dues mostres aparellades en poblacions contínues. També estem interessats en els contrastos d'independència que es puguin formular amb aquests coeficients.

8.10.1 Coeficient τ de Kendall

Considerem una mostra aleatòria simple $(x_1, y_1), \ldots, (x_n, y_n)$ d'una distribució bidimensional. Sabem que la freqüència relativa de les parelles tals que $(x_i - x_j)(y_i - y_j) > 0$ serà un estimador del paràmetre

$$\pi_+ = P\{(X - X')(Y - Y') > 0\}$$

on (X, Y) i (X', Y') són independents i tenen la mateixa distribució conjunta poblacional.
La continuïtat de les distribucions implica que $P\{(X-X')(Y-Y') = 0\} = 0$, de manera que

$$\pi_- = P\{(X - X')(Y - Y') < 0\} = 1 - \pi_+$$

Llavors,

$$\tau = \pi_+ - \pi_- = 2\pi_+ - 1$$

és l'anomenat coeficient d'associació de Kendall i mesura, en certa forma, la dependència de les variables. De fet, si X i Y són independents,

$$\begin{aligned}\pi_+ &= P(X < X')P(Y < Y') + P(X > X')P(Y > Y')\\ &= P(X > X')P(Y < Y') + P(X < X')P(Y > Y') = \pi_-\end{aligned}$$

de forma que $\tau = 0$. El recíproc no és cert, pot ser $\tau = 0$ sense que necessàriament les dues variables X i Y siguin independents.
Si P i N representen el número de parelles tals que $(x_i - x_j)(y_i - y_j) > 0$ i $(x_i - x_j)(y_i - y_j) < 0$ respectivament, entre les $\binom{n}{2}$ possibles, l'estimador natural de τ és

$$T = \frac{P}{\binom{n}{2}} - \frac{N}{\binom{n}{2}} = \frac{2}{n(n-1)}(P - N)$$

Però a més, sabem que $P + N = \binom{n}{2}$ i llavors

$$T = \frac{4P}{n(n-1)} - 1$$

Per a una mostra concreta, P es calcula fàcilment amb la mostra ordenada per la primera component: és el número de parelles amb $i < j$ per a les que $y_i < y_j$.
L'estadístic T pren valors entre -1 i 1 i un valor llunyà del zero ens indica que $\tau \neq 0$ i, per tant, que les variables X i Y no són independents.
La distribució exacta de T es pot calcular per a valors moderats de n. Per a $n \leq 10$ hi ha una taula de valors crítics tals que $P(|T| > k_\alpha) \leq \alpha$. Per a $n > 10$ pot considerar-se la distribució aproximada

$$T \sim N\left(0, \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$$

Exemple 8.10.1 *La longitud i l'amplada d'una mostra de 11 fulles d'una certa planta és*

Fulla	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Long.	*6.60*	*7.11*	*9.80*	*6.62*	*7.10*	*6.83*	*6.54*	*7.14*	*7.13*	*12.52*	*10.41*
Ampl.	*4.24*	*5.41*	*5.26*	*5.53*	*3.25*	*4.22*	*3.98*	*3.29*	*3.43*	*5.57*	*6.01*

Trobeu el coeficient de correlació de Kendall i estudieu si és significatiu.

Solució:
Si ordenem les parelles per la primera component, l'ordenació de les y_i és

3.98 4.24 5.53 4.22 3.25 5.41 3.43 3.29 5.26 6.01 5.57

D'aquesta forma el recompte de les parelles tals que $(x_i - x_j)(y_i - y_j) > 0$ és

$$P = 7 + 5 + 2 + 4 + 6 + 2 + 3 + 3 + 2 + 0 = 34$$

Així tenim

$$T = \frac{4 \cdot 34}{11 \cdot 10} - 1 = 0.2364$$

Es pot comprovar que aquest valor no és suficient per rebutjar la hipòtesi d'independència.

8.10.2 Coeficient de correlació per rangs de Spearman

El coeficient de correlació de Spearman és el coeficient de correlació ordinari, anomenat correlació de Pearson, si substituïm les parelles de valors $(x_1, y_1), \ldots, (x_n, y_n)$ pels seus rangs. El resultat és

$$r_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

on $d_i = r(x_i) - r(y_i)$, $i = 1, \ldots, n$ són les diferències dels rangs dels valors en les seves respectives mostres.
Aquest coeficient s'utilitza quan les variables són mesurades en una escala ordinal i l'ordre en la mostra és la informació més rellevant.

Càlcul del coeficient

Com hem dit, el coeficient de correlació r_S de Spearman és el coeficient de correlació ordinari en substituir els valors mostrals $(x_1, y_1), \ldots, (x_n, y_n)$ per $(a_1, b_1), \ldots, (a_n, b_n)$, on $a_i = r(x_i)$ és el rang de la dada x_i quan hem ordenat les x i $b_i = r(y_i)$ el rang que ocupa y_i en l'ordenació creixent de les y.
Efectivament,

$$r_S = \frac{\sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^n (a_i - \bar{a})^2 \sum_{i=1}^n (b_i - \bar{b})^2}}$$

on $\bar{a} = \bar{b} = (1/n) \sum_{i=1}^n i = (n+1)/2$ i

$$\sum_{i=1}^n (a_i - \bar{a})^2 = \sum_{i=1}^n (b_i - \bar{b})^2 = \sum_{i=1}^n i^2 - n\frac{(n+1)^2}{4} = \frac{n(n^2-1)}{12}$$

de manera que

$$r_S = \frac{n(n^2-1)}{12} \sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b})$$

D'altra banda

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(a_i - b_i)^2 = \sum_{i=1}^{n}(a_i - \bar{a} + \bar{b} - b_i)^2$$
$$= \frac{n(n^2-1)}{6} - 2\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})$$
$$= \frac{n(n^2-1)}{6} - 2\frac{n(n^2-1)}{12} r_S = \frac{n(n^2-1)}{6}(1 - r_S)$$

d'on resulta l'expressió amb la que hem definit el coeficient r_S.
Sota la hipòtesi d'independència es pot calcular la distribució en el mostratge de r_S i, per tant, es pot determinar els punts crítics del contrast per a diferents nivells de significació i $n \leq 10$. Per a $n > 10$, pot provar-se que r_S es aproximadament $N(0, 1/\sqrt{n-1})$.

Exemple 8.10.2 *Calculeu el coeficient de correlació per rangs de Spearman per a les dades de l'exemple 8.10.1 i estudieu si és significatiu.*

Solució:
Si juguem amb l'ordenació dels valors i els seus rangs tenim

rang x	long	ampl	rang y	d_i	d_i^2
1	6,54	3,98	4	-3	9
2	6,60	4,24	6	-4	16
3	6,62	5,53	9	-6	36
4	6,83	4,22	5	-1	1
5	7,10	3,25	1	4	16
6	7,11	5,41	8	-2	4
7	7,13	3,43	3	4	16
8	7,14	3,29	2	6	36
9	9,80	5,26	7	2	4
10	10,41	6,01	11	-1	1
11	12,52	5,57	10	1	1
				suma=	140

Taula 8.6: Taula per a la suma de les diferències al quadrat dels rangs de l'exemple 8.10.2

El resultat és $\sum_{i=1}^{n} d_i^2 = 140$ i per tant

$$r_S = 1 - \frac{6 \cdot 140}{11(11^2-1)} = 0.3636$$

Aquest valor no és suficient per rebutjar la independència entre les variables.

El paràmetre poblacional

Si la distribució conjunta de (X, Y) és $F(x, y)$ i les distribucions marginals són $F_X(x)$ i $F_Y(y)$, llavors ρ_S és el coeficient de correlació ordinari entre les variables $V_1 = F_X(X)$ i $V_2 = F_Y(Y)$ amb distribució uniforme totes dues. Amb conseqüència es pot provar que

$$\begin{aligned}\rho_S &= 12 \iint_{\mathbb{R}^2} (F(x,y) - F_X(x)F_Y(y))\, dF_X(x)dF_Y(y) \\ &= 12 \iint_{\mathbb{R}^2} F(x,y)\, dF_X(x)dF_Y(y) - 3\end{aligned}$$

La versió probabilística de la correlació τ de Kendall és

$$\tau = 4 \int_{\mathbb{R}^2} (F(x,y) - F_X(x)F_Y(y))\, dF(x,y)$$

i es verifica la relació

$$-1 \leq 3\tau - 2\rho_S \leq 1$$

Observem que $\rho_S = \tau = 0$ si $F(x, y) = F_X(x)F_Y(y)$, és a dir, si X, Y són estocàsticament independents.
Per contrastar la hipòtesi nul·la $H_0 : \rho_S = 0$, on ρ_S és la correlació poblacional, podem calcular l'estadístic

$$t = \sqrt{n-2}\frac{r_S}{\sqrt{1 - r_S^2}}$$

que té una distribució aproximada (si $n \geq 10$) t de Student amb $n - 2$ graus de llibertat.

8.11 Solució de problemes amb R o S-PLUS

Exemple 8.2.1

Primer introduïm les notes i calculem quants valors són superiors a 66, és a dir, el número de signes positius.

```
> notes<-c(71,67,55,64,82,66,74,58,79,61,78,46,84,93,72,54,78,86,48,52,67,
  95,70,43,70,73,57,64,60,83,73,40,78,70,64,86,76,62,95,66)
> notes[notes>66]
 [1] 71 67 82 74 79 78 84 93 72 78 86 67 95 70 70 73 83 73 78 70 86 76 95
> B<-length(notes[notes>66])
> B
[1] 23
```

I finalment, calculem l'aproximació normal de l'estadístic i el seu p-valor.

```
> n<-length(notes)-length(notes[notes==66])
> n
[1] 38
> z<-(B-0.5-n*0.5)/sqrt(n*0.5*0.5)
> round(z,2)
[1] 1.14
> 2*(1-pnorm(z))
[1] 0.2561450
> 2*(1-pnorm(B-0.5,mean=n*0.5,sd=sqrt(n*0.5*0.5)))
[1] 0.2561450
```

El p-valor superior al nivell de significació fa que acceptem la hipòtesi nul·la $H_0 : M = 66$.

Exemple 8.2.2

Introduïm les dades en dos vectors de la mateixa longitud. Calculem la diferència i el número de valors positius. A partir d'aquest número, calculem la probabilitat de la cua dreta d'una binomial, ja que la hipòtesi alternativa així ho demana.

```
> maq1<-c(46,110,70,54,60,120,82,76,37,28)
> maq2<-c(42,87,75,50,48,108,80,67,40,25)
> dif<-maq1-maq2
> dif
 [1]  4 23 -5  4 12 12  2  9 -3  3
> B<-length(dif[dif>0]);B
[1] 8
> dbinom(8,10,0.5)+dbinom(9,10,0.5)+dbinom(10,10,0.5)
[1] 0.0546875
> 1-pbinom(B-1,10,0.5)
[1] 0.0546875
```

El p-valor és inferior al nivell de significació 0.06 i, per tant, rebutgem la hipòtesi nul·la i acceptem que la màquina 1 produeix més peces defectuoses.

Exemple 8.2.3

En aquest cas els càlculs són molt senzills.

```
> z<-(197-0.5-300*0.5)/sqrt(300*0.5*0.5);z
[1] 5.369358
> 1-pnorm(z)
[1] 3.950882e-08
```

I el p-valor és clarament inferior a 0.01, de manera que acceptem l'alternativa.

Exemple 8.3.1

En aquest exemple farem servir la funció `wilcox.test` amb el vector `notes` de l'exemple 8.2.1.

```
> wilcox.test(notes,mu=66,alternative="two.sided",exact=F)

        Wilcoxon signed rank test with continuity correction

data:  notes
V = 465, p-value = 0.1726
alternative hypothesis: true mu is not equal to 66
> z<-(465-0.5-38*39/4)/sqrt(38*39*77/24);z
[1] 1.363214
> round(2*(1-pnorm(z)),2)
[1] 0.17
```

Aquesta funció calcula l'estadístic $T^+ = 465$ i el seu p-valor amb correcció per continuïtat. En aquest problema hi ha, a més de dos zeros, un grapat d'empats o lligadures, de manera que la funció `wilcox.test` no pot calcular el p-valor exacte i per això li hem indicat amb `exact=F`. L'estadístic z que hem calculat de forma directa i el seu p-valor, sense tenir en compte les lligadures, són força semblants als que l'algorisme calcula. Si no indiquem res sobre aquest paràmetre, sortiran dos missatges d'advertència sobre aquest fet.

```
> wilcox.test(notes,mu=66,alternative="two.sided")

        Wilcoxon signed rank test with continuity correction

data:  notes
V = 465, p-value = 0.1726
alternative hypothesis: true mu is not equal to 66

Warning messages:
1: Cannot compute exact p-value with ties in:
   wilcox.test.default(x, mu = 66, alternative = "two.sided")
2: Cannot compute exact p-value with zeroes in:
   wilcox.test.default(x, mu = 66, alternative = "two.sided")
```

Més informació sobre la funció amb

```
> help(wilcox.test)
```

Exemple 8.3.2

En aquest exemple també s'utilitza la funció `wilcox.test` amb els dos vectors de dades de l'exemple 8.2.2.

```
> wilcox.test(maq1,maq2,mu=0,paired=T,alternative="greater",exact=F)

        Wilcoxon signed rank test with continuity correction

data:  maq1 and maq2
V = 46.5, p-value = 0.02942
alternative hypothesis: true mu is greater than 0
```

Observem que en aquest cas hem fet servir l'opció `paired=T` per indicar que les dades són aparellades. També hem identificat l'alternativa correcte amb `alternative="greater"`. A més, com en l'exemple anterior, les lligadures no permeten calcular el p-valor exacte. El p-valor aproximat 0.029 ens indica el rebuig de la hipòtesi nul·la.

Exemple 8.4.1

Ara farem servir la funció `wilcox.test` amb els dos vectors de dades però tindrem en compte que són independents, que és l'opció per defecte.

```
> pro.A<-c(202,229,215,220,223,233,208,228,209)
> pro.B<-c(221,207,185,203,187,190,195,204,212)
> wilcox.test(pro.A,pro.B,alternative="two.sided")

        Wilcoxon rank sum test

data:  pro.A and pro.B
W = 70, p-value = 0.007775
alternative hypothesis: true mu is not equal to 0
```

No ens hem de deixar confondre per la notació, l'estadístic calculat és el que nosaltres hem anomenat $U' = 70 > n_1 n_2/2 = 40.5$. En tot cas, el p-valor és molt explícit i implica el rebuig de la hipòtesi nul·la d'equivalència.

Exemple 8.5.1

Observem el càlcul de la mediana conjunta amb `median(c(pro.A,pro.B))`.

```
> T<-length(pro.A[pro.A<median(c(pro.A,pro.B))])
> T
[1] 2
> phyper(T,9,9,9)
[1] 0.02834225
> 1-phyper(6,9,9,9)
[1] 0.02834225
```

La distribució hipergeomètrica ens permet trobar els límits de la regió crítica.

Exemple 8.5.2

En primer lloc hem d'introduir les freqüències de la taula.

```
> numero<-cbind(expand.grid(M=c("inferior a M","superior a M"),
  grup=c("A","B")))
> fr<-c(2,7,7,2)
> attach(numero)
> taula<-table(M,grup)*fr
> taula
              grup
M               A B
  inferior a M 2 7
  superior a M 7 2
```

I amb aquesta taula calculem el test d'homogeneïtat.

```
> chisq.test(taula)

        Pearson's Chi-squared test with Yates' continuity correction

data:  taula
X-squared = 3.5556, df = 1, p-value = 0.05935

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(taula)
```

El resultat és l'acceptació de la igualtat de medianes. La discrepància amb l'exemple anterior és possible per la manca d'observacions.

Exemple 8.6.1

Les dades han estat introduïdes en els vectors `pro.A` i `pro.B` i el test es calcula amb la funció `ks.test`.

```
> ks.test(pro.A,pro.B,alternative="two.sided")

        Two-sample Kolmogorov-Smirnov test

data:  pro.A and pro.B
D = 0.6667, p-value = 0.03357
alternative hypothesis: two.sided
```

En aquest cas, el p-valor ens indica el rebuig de la hipòtesi nul·la.

Exemple 8.7.1

Per fer el test de Kruskal-Wallis fem servir la funció `kruskal.test` que ens proporciona l'estadístic H o si cal, com en aquest cas, l'estadístic H'.

```
> pes<-c(251,250,249,255,258,258,247,246,250,241,240,242,
  228,236,240,225,236,230)
> fabr<-c(rep(1,6),rep(2,6),rep(3,6))
> kruskal.test(pes,fabr)

        Kruskal-Wallis rank sum test

data:  pes and fabr
Kruskal-Wallis chi-squared = 14.3957, df = 2, p-value = 0.0007482
```

El p-valor es prou significatiu del rebuig de la hipòtesi nul·la.

Exemple 8.8.1

El test de Friedman s'aplica amb la funció `friedman.test`.

```
> nota<-c(5,3,2,4,1,4,3,5,2,1,3,5,4,2,1,
          4,5,1,2,3,3,4,5,1,2,5,3,4,2,1,
          2,5,4,3,1,3,5,4,1,2,3,4,5,2,1,
          4,5,3,1,2,5,3,2,4,1,5,4,3,2,1)
> individu<-c(rep(1,5),rep(2,5),rep(3,5),
              rep(4,5),rep(5,5),rep(6,5),
              rep(7,5),rep(8,5),rep(9,5),
              rep(10,5),rep(11,5),rep(12,5))
> xampu<-c(rep(seq(1,5,1),12))
> friedman.test(nota,xampu,individu)

        Friedman rank sum test

data:  nota, xampu and individu Friedman chi-squared = 25.5333, df
= 4, p-value = 3.929e-05
```

El p-valor ens indica clarament la significació de les diferències.

Exemple 8.9.1

```
> peces<-c(0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,
           0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,
           0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,
           0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,
           0,0,0,0,0,0,0,0,0,0,1,1,1,0,0)
> c<-1
> for (i in 1:(length(peces)-1)) if (peces[i]!=peces[i+1]) c<-c+1
> c
```

```
[1] 15
> n2<-sum(peces);n2
[1] 14
> n1<-length(peces)-n2;n1
[1] 61
> mu<-(2*n1*n2)/(n1+n2)+1;mu
[1] 23.77333
> sigma2<-(2*n1*n2*(2*n1*n2-n1-n2))/((n1+n2)^2*(n1+n2-1));sigma2
[1] 6.700694
> z<-(c-mu)/sqrt(sigma2);z
[1] -3.389259
> 2*pnorm(z)
[1] 0.0007008184
```

També podem construir un test.

```
> ratxes.test<-function(s){
+ c<-1
+ for (i in 1:(length(s)-1)) if (s[i]!=s[i+1]) c<-c+1
+ n2<-sum(s)
+ n1<-length(s)-n2
+ mu<-(2*n1*n2)/(n1+n2)+1
+ sigma2<-(2*n1*n2*(2*n1*n2-n1-n2))/((n1+n2)^2*(n1+n2-1))
+ z<-(c-mu)/sqrt(sigma2);z
+ }
> ratxes.test(peces)
[1] -3.389259
```

Exemple 8.9.2

Calculem la mediana del vector `notes` i, per estalviar-nos dificultats, eliminem del vector de dades els valors que coincideixen amb la mediana.

```
> median(notes)
[1] 70
> notes2<-c(71,67,55,64,82,66,74,58,79,61,78,46,84,93,72,54,78,86,48,52,67,
  95,43,73,57,64,60,83,73,40,78,64,86,76,62,95,66)
> signes<-ifelse(notes2<median(notes),0,1);signes
 [1] 1 0 0 0 1 0 1 0 1 0 1 0 1 1 1 0 1 1 0 0 0 1 0 1 0 0 0 1 1 0 1 0 1 1 0 1 0
> z<-ratxes.test(signes);z
[1] 2.173642
> 2*(1-pnorm(z))
[1] 0.029732
```

Exemple 8.9.3

Tenim les dades en el vector `notes`

```
> notes
 [1] 71 67 55 64 82 66 74 58 79 61 78 46 84 93 72 54 78 86 48 52 67 95 70 43 70
[26] 73 57 64 60 83 73 40 78 70 64 86 76 62 95 66
> notes2<-notes[2:length(notes)]
> cre<-ifelse(notes[1:(length(notes)-1)]<=notes2,1,0);cre
 [1] 0 0 1 1 0 1 0 1 0 1 0 1 1 0 0 1 1 0 1 1 1 0 0 1 1 0 1 0 1 0 0 1 0 0 1 0 0 1
[39] 0
> z<-ratxes.test(cre);z
[1] 2.115198
> 2*(1-pnorm(z))
[1] 0.03441305
```

Exemple 8.10.1

Pel càlcul dels coeficients de correlació s'utilitza la funció `cor.test` amb el paràmetre

```
method = c("pearson", "kendall", "spearman")"
```

segons si volem el coeficient clàssic de Pearson, la τ de Kendall o el coeficient per rangs de Spearman.

```
> long<-c(6.60,7.11,9.80,6.62,7.10,6.83,6.54,7.14,7.13,12.52,10.41)
> ampl<-c(4.24,5.41,5.26,5.53,3.25,4.22,3.98,3.29,3.43,5.57,6.01)
> cor.test(long,ampl,method="kendall")

        Kendall's rank correlation tau

data:  long and ampl
T = 34, p-value = 0.3587
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.2363636
```

Observem que el número de parelles *positives* és 34, estadístic que nosaltres hem anomenat P.
El procediment calcula el p-valor exacte per a $n < 50$. En aquest cas el p-valor ens mostra que l'estadístic no és significatiu.

Exemple 8.10.2

Amb les dades de l'exemple 8.10.1 tenim:

```
> cor.test(long,ampl,method="spearman")

        Spearman's rank correlation rho

data:  long and ampl
```

```
S = 140, p-value = 0.2732
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3636364
```

8.12 Taules

- **Taula de probabilitats binomials:** D. Peña. Estadística. Modelos y métodos. Vol. I. Alianza Universidad. Tabla 2. pàg. 333
- **Taula de Wilcoxon:** C.M. Cuadras. Problemas de probabilidades y Estadística Vol II. Tablas Estadísticas. Tabla XIV.
- **Taula de l'estadístic U de Mann-Whitney:** C.M. Cuadras. Problemas de probabilidades y Estadística Vol II. Tablas Estadísticas. Tabla XII.
- **Taula de Massey per al test d'homogeneïtat de Kolmogorov-Smirnov:** R. Vélez Ibarrola y A. García Pérez. Cálculo de Probabilidades y Estadística Matemática. Principios de Inferencia Estadística. UNED. Tablas 9 y 10.
- **Taula de l'estadistic τ de Kendall:** R. Vélez Ibarrola y A. García Pérez. Cálculo de Probabilidades y Estadística Matemática. Principios de Inferencia Estadística. UNED. Tabla 13.
- **Taula de l'estadistic r_S de Spearman:** R. Vélez Ibarrola y A. García Pérez. Cálculo de Probabilidades y Estadística Matemática. Principios de Inferencia Estadística. UNED. Tabla 14.