Análisis de la varianza de dos factores

El problema anterior consideraba la comparación de k muestras para detectar diferencias entre las respectivas poblaciones. En el modelo de aleatorización N individuos son asignados al azar a uno de k tratamientos. Sin embargo, las diferencias entre tratamientos podrían verse "oscurecidas" por la variabilidad excesiva entre los individuos dentro de los grupos. Este problema puede resolverse dividiendo a los sujetos en subgrupos homogéneos o bloques y realizando las comparaciones entre los bloques. Se trata de una extensión del problema de muestras apareadas.

El diseño que consideraremos inicialmente es el *diseño en bloques completamente aleatorizado con una observación por celda*.

Se tienen N=nk sujetos divididos en n bloques y los sujetos, dentro de cada bloque, se asignan al azar a uno de los k tratamientos. Otro ejemplo de aplicación de este diseño es el problema de mediciones repetidas, en cuyo caso, el sujeto es el bloque y se realizan k mediciones sobre cada sujeto.

El modelo muestral puede definirse de dos formas:

1) $\vec{X}_i = (X_{i1},...,X_{ik})$ son n v.a. independientes con distribución $F_i(x_1-\theta_1,...,x_k-\theta_k)$ y $F_i(x_1,...,x_k) = F_i(x_{p_1},...,x_{p_k})$ para toda permutación ($p_1,...,p_k$), o sea que $X_1,...,X_k$ son canjeables.

2) $X_{ij} \sim F_i(x-\theta_j), F_i \in \Omega_o$ para $i=1,...,k$ y $j=1,...,n$. Es decir que F_i es la distribución de las observaciones del i-ésimo bloque y, dentro del bloque θ_j es la mediana del j-ésimo tratamiento.

Queremos testear

H_o: $\theta_1 = \theta_2 = ... = \theta_k$ vs H_1: existe al menos un par (i,j) tal que $\theta_i \neq \theta_j$

Test de Friedman (1973): Sea $R_{ij} = R(X_{ij})$ el rango de X_{ij} entre $X_{i1},...,X_{ik}$. Es decir que R_{ij} es el rango de X_{ij} dentro de su bloque. Entonces

$$R_{.j} = \sum_{i=1}^{n} R_{ij}$$

es la suma de los rangos correspondientes al tratamiento j. Esquemáticamente, colocando como columnas los tratamientos y los bloques (sujetos) como filas:

	1	2	....	k
1	R_{11}	R_{12}	....	R_{1k}
2	R_{21}	R_{22}	....	R_{2k}
...	....	....	....	....
n	R_{n1}	R_{n2}	....	R_{nk}
	$R_{.1}$	$R_{.2}$	....	$R_{.k}$

Bajo H_o, y suponiendo que no hay empates, conocemos la distribución de $R_{.1},...,R_{.k}$ pues

$$P(R_{ij} = s) = \frac{1}{k} \qquad 1 \le s \le k$$

$$P(R_{ij} = s, R_{il} = t) = \frac{1}{k(k-1)} \qquad s \neq t \quad j \neq l$$

Entonces,

$$E(R_{ij}) = \frac{k+1}{2} \qquad V(R_{ij}) = \frac{k^2-1}{12} \qquad \text{cov}(R_{ij}, R_{il}) = -\frac{k+1}{12} \quad \text{si } j \neq l$$

y por lo tanto

$$E(R_{.j}) = \frac{n(k+1)}{2} \qquad Var(R_{.j}) = \frac{n(k^2-1)}{12} \qquad \text{cov}(R_{.j}, R_{.l}) = -\frac{n(k+1)}{12} \quad \text{si } j \neq l$$

Friedman propuso el siguiente estadístico:

$$T_1 = \sum_{j=1}^{k} c_{jN}^2 \left(\frac{R_{.j} - E(R_{.j})}{\sqrt{Var(R_{.j})}} \right)^2$$

donde los c_{jN} se eligen de manera que el estadístico converja, bajo H_o a una distribución χ^2_{k-1}.

El estadístico toma la forma

$$T_1 = \sum_{j=1}^{k} \left(1 - \frac{1}{k}\right) \left(\frac{R_{.j} - n(k+1)/2}{\sqrt{\frac{n(k^2-1)}{12}}} \right)^2 = \left(\frac{12}{n\,k\,(k+1)} \sum_{j=1}^{k} R_{.j}^2 \right) - 3\,n\,(k+1)$$

que, bajo H_o, tiene distribución asintótica χ^2 con k-1 grados de libertad. Se rechazará H_o si

$$T_1 > \chi^2_{k-1,\alpha}$$

En caso en que haya empates, debe hacerse una modificación al estadístico del test de Friedman. Sean $A_1 = \sum_{i=1}^{n}\sum_{j=1}^{k} R_{ij}^2$ $\qquad C_1 = \frac{nk(k+1)^2}{4}$

Se define el estadístico modificado:

$$T_1 = \frac{(k-1)\left(\sum_{j=1}^{k} R_{.j}^2 - nC_1\right)}{A_1 - C_1} = \frac{(k-1)\sum_{j=1}^{k}\left(R_{.j} - \frac{n(k+1)}{2}\right)^2}{A_1 - C_1}$$

Si $A_1 = C_1$ se considera que se está en la región crítica y se rechaza H_o.

Otros estudios (Iman y Davenport, 1980) recomiendan utilizar, no T_1, sino el estadístico del test clásico de análisis de la varianza calculado sobre los rangos, que se puede expresar como función del estadístico T_1:

$$T_2 = \frac{(n-1)T_1}{n(k-1) - T_1}$$

Bajo H_o , T_2 tiene distribución F con (k-1) y (n-1)(k-1) grados de libertad. Se rechaza H_o si $T_2 > F_{(k-1),(k-1)(n-1),\alpha}$.

Ejemplo: 12 amas de casa son seleccionadas para participar en un experimento de siembra. A cada una de ellas se le pide que seleccione cuatro parcelas idénticas en su jardín y plante 4 tipos distintos de césped, uno en cada parcela. Después de cierto periodo, se les pide que ordenen los 4 tipos de césped por orden de preferencia, asignando el número 1 al césped menos preferido, 2 al siguiente, etc. La hipótesis nula implica que no hay diferencias entre las preferencias de los tipos de césped. Los resultados obtenidos son los siguientes:

Ama de casa	**Césped**			
	1	**2**	**3**	**4**
1	4	3	2	1
2	4	2	3	1
3	3	1.5	1.5	4
4	3	1	2	4
5	4	2	1	3
6	2	2	2	4
7	1	3	2	4
8	2	4	1	3
9	3.5	1	2	3.5
10	4	1	3	2
11	4	2	3	1
12	3.5	1	2	3.5
$R_{.j}$	38	23.5	24.5	34

Como hay empates, calculamos el valor del estadístico modificado

$$A_1 = 356.5 \qquad C_1 = 300 \qquad T_1 = 8.097$$

y obtenemos p-valor = $P\left(\chi_3^2 > 8.097\right) = 0.044$. Por lo tanto, a nivel 0.05, se rechaza la hipótesis nula.

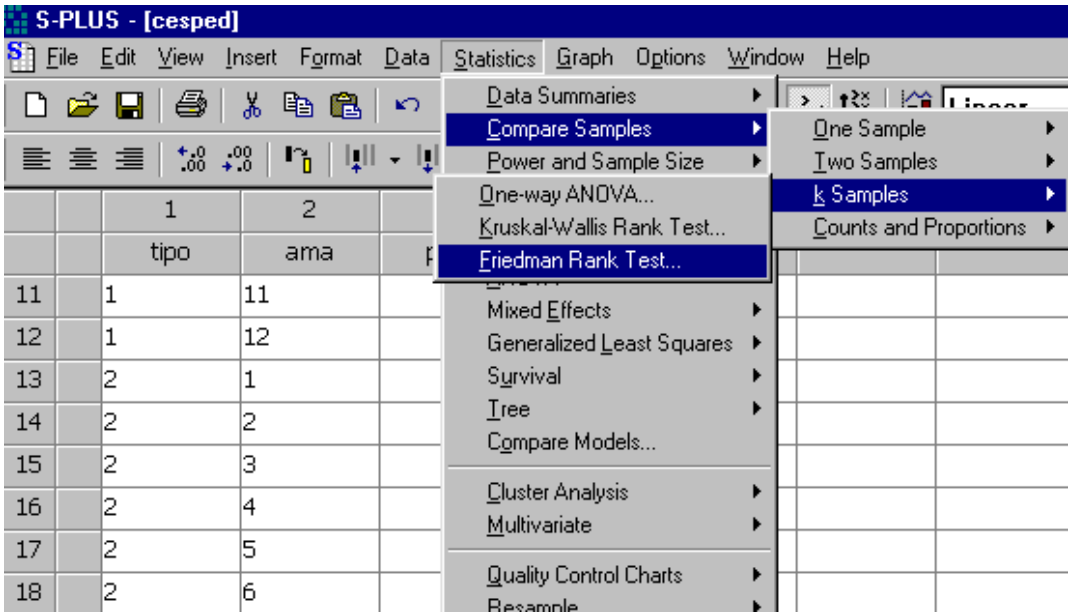
Calculemos el estadístico T_2.

$$T_2 = \frac{11(8.097)}{12(3) - 8.097} = 3.19$$

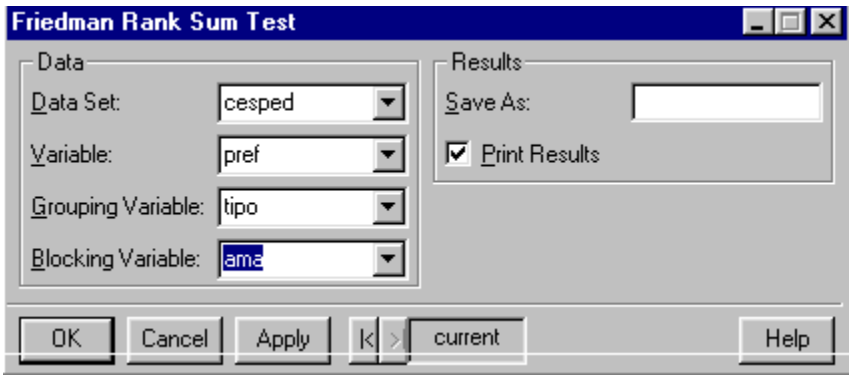
Como la región crítica de nivel 0.05 de la distribución F con 3 y 33 grados de libertad corresponde a valores del estadístico mayores que 2.90, se rechaza la hipótesis nula. El correspondiente p-valor es:

$$\text{p-valor} = P(F_{3,33} > 3.19) = 0.036$$

Procesamiento con S-PLUS: Los datos se ingresaron en un data set denominado "cesped" que contiene 3 variables: tipo (grupos o categorías), ama (bloque) y pref.



Se debe especificar cuál es la variable en estudio, qué variable define el agrupamiento y cuál el bloque.



S-PLUS no utiliza el estadístico T_2 sino T_1:

Friedman rank sum test

data: pref and tipo and ama from data set cesped
Friedman chi-square = 8.0973, df = 3, p-value = 0.044
alternative hypothesis: two.sided

Comparaciones múltiples: Cuando se rechaza H_o, es posible realizar tests de comparaciones múltiples en base a $R_{.1},...,R_{.k}$. En efecto, bajo H_o,

$$\frac{R_{.j} - R_{.i}}{\sqrt{Var(R_{.j} - R_{.i})}} \xrightarrow{d} N(0,1)$$

siendo $Var(R_{.j} - R_{.i}) = 2Var(R_{.j}) - 2\,\mathrm{cov}(R_{.j}, R_{.i}) = \frac{nk(k+1)}{6}$.

Entonces, diremos que θ_i es significativamente distinto de θ_j a nivel global α si

$$\left|R_{.j} - R_{.i}\right| \geq z_{\alpha'/2}\sqrt{\frac{nk(k+1)}{6}} \qquad \text{con } \alpha' = \frac{2\alpha}{k(k-1)}$$

Conover sugiere otro método. Diremos que θ_i es significativamente distinto de θ_j a nivel global α si

$$\left|R_{.j} - R_{.i}\right| \geq t_{(k-1)(n-1),\alpha/2}\left(\frac{2(nA_1 - \sum R_{.j}^2)}{(k-1)(n-1)}\right)^{1/2}$$

En este caso, el nivel α es el mismo que el utilizado en el test de Friedman. La ecuación anterior puede escribirse en función de T_1:

$$\left|R_{.j} - R_{.i}\right| \geq t_{(k-1)(n-1),\alpha/2}\left[\frac{(A_1 - C_1)2n}{(k-1)(n-1)}\left(1 - \frac{T_1}{n(k-1)}\right)\right]^{1/2}$$

Si no hay empates, $A_1 = nk(k+1)(2k+1)/6$ y $A_1 - C_1 = nk(k+1)(k-1)/12.$

Ejemplo: Dado que en el ejemplo anterior, se rechazó H_o a nivel 0.05, interesa detectar cuáles son los tipos de césped que difieren en cuanto a la preferencia. Aplicaremos el procedimiento sugerido por Conover. El valor crítico 0.025 de la distribución t con (11)(3) = 33 grados de libertad es 2.036, entonces

$$t_{(k-1)(n-1),\alpha/2}\left(\frac{2(nA_1 - \sum R_{.j}^2)}{(k-1)(n-1)}\right)^{1/2} = 11.49$$

El césped tipo 1 resulta ser significativamente diferente de los tipos 2 y 3 y no hay otra diferencia significativa.