Trabajo Fin de Grado

INFERENCIA NO PARAMÉTRICA PARA DATOS MULTIVARIANTES

Facultad de Matemáticas

GRADO EN MATEMÁTICAS

Inmaculada Rodríguez Montero
Sevilla, Junio de 2021

Índice general

Resumen

El análisis de varianza multivariante (en adelante, MANOVA) es un procedimiento estadístico para comparar vectores de medias multivariantes, es decir, comparar los valores esperados de diversas variables entre dos o más grupos, subpoblaciones o condiciones de experimentación. Cuando se usa el MANOVA, deben cumplirse algunas hipótesis bastante restrictivas como la normalidad multivariante. Esta hipótesis restringe la aplicación de métodos paramétricos multivariantes, por lo que a veces se cambian por métodos univariantes debido a la falta de métodos no paramétricos. Estos métodos univariantes no suelen funcionar bien, pues hay que analizar cada variable individualmente, sin tener en cuenta las posibles correlaciones entre las variables.

El objetivo de este trabajo es llevar a cabo una descripción teórica de técnicas no paramétricas multivariantes de comparación de poblaciones, así como la implementación en R de dos funciones que llevan a cabo dichas técnicas. Además, con objeto de ilustrar la aplicabilidad de estas técnicas, se incluyen referencias sobre artículos científicos recientes en los que se ha hecho uso de las mismas.

Abstract

The multivariate analysis of variance (MANOVA, henceforth) is an statistical procedure to compare multivariate mean vectors, that is, to compare the expected values of some variables between two or more groups, subpopulations or experimental conditions. When MANOVA is used, some quite restrictive hypotheses like multivariate normality must be fulfilled. This hypothesis restricts the application of multivariate parametric methods, so sometimes they are switched to univariate methods due to lack of non-parametric methods. These univariate methods don't usually work well, since each variable must be analysed individually, without considering possible correlations between variables.

The aim of this work is to carry out a theoretical description of nonparametric multivariate techniques to compare populations, as well as the implementation in R of two functions that carry out these techniques. Moreover, in order to illustrate the applicability of these techniques, some references about current scientific articles where they have been used are included.

Capítulo 1

INTRODUCCIÓN

El análisis de varianza multivariante (en adelante, MANOVA) es un procedimiento estadístico para comparar vectores de medias multivariantes, es decir, comparar los valores esperados de diversas variables entre dos o más grupos, subpoblaciones o condiciones de experimentación. Cuando se usa el MANOVA, deben cumplirse algunas hipótesis bastante restrictivas como la normalidad multivariante. Esta hipótesis restringe la aplicación de métodos paramétricos multivariantes, por lo que a veces se cambian por métodos univariantes debido a la falta de métodos no paramétricos. En estudios relacionados con la medicina, estos métodos (univariantes) tampoco suelen funcionar bien, pues hay que analizar cada variable individualmente, sin tener en cuenta las posibles correlaciones entre las variables.

Por ello, se hace necesario profundizar en el estudio de métodos no paramétricos con idénticos objetivos al MANOVA, como hicieron Puri y Sen [21], Thompson [25] o Munzel y Brunner ([16] y [17]). Estos métodos tienen menos limitaciones y una aplicabilidad más amplia debido a que se pueden usar en estudios donde las variables puedan no ser solamente continuas, sino también binarias, ordinales o una mezcla de ambas. Las técnicas no parámetricas tienen varias ventajas: una mayor consistencia, la utilidad cuando los datos contienen valores atípicos o sesgados y una invariancia al aplicar transformaciones a los datos. Para aplicar estas técnicas el único requisito es que las distribuciones sean no degeneradas.

El objetivo de este Trabajo de Fin de Grado es desarrollar el fundamento teórico de los tests no paramétricos que sustituyen y se fundamentan en cuatro tests paramétricos multivariantes que funcionan bien bajo hipótesis de normalidad: el test Lambda de Wilks, el test tipo ANOVA, el test de Lawley-Hotelling y el test de Bartlett-Nanda-

Pillai. Además, se presenta su implementación en R.

Previamente a este desarrollo, veremos una descripción de la alternativa paramétrica bajo hipótesis de normalidad. Esta alternativa es, como se ha recogido anteriormente, el Análisis de la Varianza Multivariante (MANOVA).

1.1. Análisis de la varianza multivariante (MANOVA)

En este apartado se va a exponer un resumen del análisis de la varianza multivariante, más conocido como MANOVA como acrónimo de su denominación en inglés (Multivariate ANalysis Of Variance) y se presentarán algunos ejemplos en los que se puede aplicar dicho análisis. Por si se desea tener más información sobre este tema, este capítulo está basado en Peña [18] , Jhonson y Wichern [11] y Cuadras [7].

Consideremos la siguiente situación experimental: disponemos de muestras aleatorias de a poblaciones normales multivariantes, es decir, tenemos a matrices de datos independientes (todas las filas son independientes) $\mathbf{X}_1, \ldots, \mathbf{X}_a$ de dimensiones $n_1 \times p$, $\ldots$, $n_a \times p$ que provienen de distribuciones $N_p(\mu_1, \Sigma), \ldots, N_p(\mu_a, \Sigma)$, siendo p el número de variables o la dimensión del vector que describe las poblaciones y $n_1, \ldots, n_a$ el tamaño de muestras de cada población. En este caso vamos a suponer que la matriz de varianzas-covarianzas Σ sea igual para todas las poblaciones, es decir, se supone la homocedasticidad de las distribuciones.

El MANOVA se usa para analizar si los vectores de medias de las poblaciones son iguales.

El modelo MANOVA de un factor o una vía es el siguiente:

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij},\ j = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, a \tag{1.1}$$

donde ε_{ij} son variables normales multivariantes independientes $N_p(0, \Sigma)$ y representan los errores aleatorios o perturbaciones aleatorias. El vector μ es la media general y τ_i representa el i-ésimo efecto del "tratamiento" o grupo sobre el comportamiento medio de las variables incluidas en el vector aleatorio bajo estudio. Por ello, dado que representa el efecto diferencial entre grupos y μ el efecto global o común, se puede suponer: $\sum_{i=1}^{a} n_i \tau i = 0$

La matriz de covarianzas Σ es una matriz de parámetros desconocidos, por lo que obtenemos su estimación centrada o insesgada,

$$\mathbf{S} = \frac{1}{n-a}\sum_{i=1}^{a}(n_i - 1)\mathbf{S}_i$$

y el vector de medias generales

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{a} n_i\overline{x_i}$$

siendo $\mathbf{S}_i$ la matriz de covarianzas muestrales de la población $i - ésima, i = 1, \ldots, a$, $\overline{x_i}$ el vector de medias muestrales de la población $i - ésima$ y $n = \sum_{i=1}^{a} n_i$.

De acuerdo con el modelo anterior, cada componente de cada vector x_{ij} satisface el modelo univariante

$$x_{ij}^{(k)} = \mu^{(k)} + \tau_i^{(k)} + \varepsilon_{ij}^{(k)},\ k = 1, \ldots, p$$

Un vector de observaciones (1.1) se puede descomponer como sigue:

$$x_{ij} = \overline{x} + (\overline{x_i} - \overline{x}) + (x_{ij} - \overline{x_i})$$

Esta descomposición conduce a la conocida como descomposición de la variabilidad multivariante o descomposición MANOVA de las matrices sumas de cuadrados y productos cruzados.

Primero, vamos a descomponer el siguiente producto:

$$(x_{ij} - \overline{x})(x_{ij} - \overline{x})' = [(x_{ij} - \overline{x_i}) + (\overline{x_i} - \overline{x})]\,[(x_{ij} - \overline{x_i}) + (\overline{x_i} - \overline{x})]' =$$

$$= (\mathrm{x}_{ij} - \overline{x_i})(x_{ij} - \overline{x_i})' + (x_{ij} - \overline{x_i})(\overline{x_i} - \overline{x})' + (\overline{x_i} - \overline{x})(x_{ij} - \overline{x_i})' + (\overline{x_i} - \overline{x})(\overline{x_i} - \overline{x})'$$

La suma en j de los dos términos intermedios son cero, porque $\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i}) = 0$. Por lo tanto, sumando los productos cruzados sobre i y j obtenemos lo siguiente:

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \overline{x})(x_{ij} - \overline{x})' = \sum_{i=1}^{a} n_i(\overline{x_i} - \overline{x})(\overline{x_i} - \overline{x})' + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})(x_{ij} - \overline{x_i})'$$

Así se pueden introducir las siguientes matrices:

Variación entre grupos: mide las desvaciones de la media de cada grupo respecto al vector de medias generales

$$\mathbf{B} = \sum_{i=1}^{a} n_i(\overline{x_i} - \overline{x})(\overline{x_i} - \overline{x})'$$

Variación dentro de los grupos: mide las desviaciones de los datos de cada grupo respecto a su media

$$\mathbf{W} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})(x_{ij} - \overline{x_i})'$$

Variación total: mide las desviaciones de todos los datos respecto al vector de medias (es la suma total de cuadrados y productos cruzados)

$$\mathbf{T} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})(x_{ij} - \overline{x})' = \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})'$$

Se tiene que $\mathbf{W} = (n-a)\mathbf{S}$ y la relación

$$\mathbf{T} = \mathbf{B} + \mathbf{W} \tag{1.2}$$

que se traduce en:
Variabilidad total ($\mathbf{T}$) = Variabilidad explicada($\mathbf{B}$) + Variabilidad residual($\mathbf{W}$)

Dicha descomposición es la habitual del análisis de la varianza y se conoce como **descomposición MANOVA**.

El objetivo fundamental será construir un test para decidir si se puede aceptar la igualdad de medias de las a poblaciones. El contraste de hipótesis es

$$H_0 : \mu_1 = \cdots = \mu_a$$

$$H_1 : \text{no todas las } \mu_i \text{ son iguales}$$

Este contraste se llevará a cabo analizando los tamaños de las matrices $\mathbf{W}$ y $\mathbf{B}$, aunque a veces también se realiza comparando los tamaños de $\mathbf{T}$ y $\mathbf{W}$.

Dado que será necesario determinar las distribuciones de los estadísticos de contraste bajo las condiciones impuestas por la hipótesis nula, necesitamos conocer la distribución de Wilks, por lo que se introduce a continuación.
Sean dos matrices $\mathbf{A}$ y $\mathbf{B}$, de orden $p \times p$ estocásticamente independientes que siguen distribuciones Wishart $W_p(\Sigma, m)$ y $W_p(\Sigma, n)$ respectivamente, entonces la distribución del cociente de los determinantes

$$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|}$$

se define como la distribución lambda de Wilks, denotada como $\Lambda(p, m, n)$. Esta distribución, como se indica en sus parámetros, no depende de la matriz Σ, solamente de la dimensión de las matrices y los grados de libertad asociados.

Volviendo a lo anterior, si la hipótesis nula es cierta, se verifica que

$$\mathbf{B} \sim W_p(\Sigma, a-1), \mathbf{W} \sim W_p(\Sigma, n-a), \mathbf{T} \sim W_p(\Sigma, n-1)$$

y también que **B,W** son estocásticamente independientes. Si H_0 es cierta,

$$\Lambda = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{B}|} \sim \Lambda(p, n-a, a-1)$$

Por lo tanto, en el contraste de hipótesis anterior, se deberá rechazar H_0 si el estadístico Λ alcanza un valor pequeño, dado que ello se alcanza cuando la dispersión entre los grupos, representada en la matriz $\mathbf{B}$, es significativamente grande. Este es el estadístico resultante del test de razón de verosimilitudes bajo las condiciones de normalidad y homocedasticidad.

Debido a su relevancia en este trabajo, se definen a continuación el estadístico de Lawley-Hotelling o T^2 de Hotelling generalizado y el estadístico traza de Pillai.
Sean:

$$\mathbf{A} \sim W_p(\Sigma, m), \mathbf{B} \sim W_p(\Sigma, n)$$

independientes y no singulares con $m, n > p$. Se definen los siguientes estadísticos:
Estadístico de Lawley-Hotelling o T^2 de Hotelling generalizado:

$$T_g^2 = m\, traza(\mathbf{A}\mathbf{B}^{-1})$$

Estadístico traza de Pillai:

$$V = traza\left[\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\right]$$

Su particularización en el MANOVA de un factor será:

$$T_g^2 = (n-a) traza(\mathbf{B}\mathbf{W}^{-1})$$

$$V = traza\left[\mathbf{B}(\mathbf{W}+\mathbf{B})^{-1}\right]$$

Ambos estadísticos se pueden aplicar en el test de hipótesis de igualdad de vectores medias y resultan prácticamente equivalentes al estadístico Lambda de Wilks.

1.2. Normalidad y homocedasticidad

Resuelta la parte teórica bajo las citadas condiciones distribucionales, también resulta de interés recoger diversas consideraciones sobre la parte práctica de MANOVA y las situaciones prácticas donde puede aplicarse.

En primer lugar, para poder llevar a cabo un análisis de este tipo debemos asegurarnos de que los datos cumplen las hipótesis de normalidad multivariante y homocedasticidad.

Para analizar la normalidad multivariante, podrá aplicarse el Test de Mardia. Dicho test, tanto en el caso univariante como para el caso multivariante, está basado en la asimetría y en la kurtosis.

En el caso univariante, se sabe que la distribución normal es simétrica en torno a su media. Por lo tanto, la forma más común de desviarse de la normalidad es la falta de simetría, por lo que parece lógico construir un método de contraste en base a cierta medida de simetría. Independientemente de lo visto anteriormente, sea $X_1, \ldots, X_n$ una muestra aleatoria simple. El coeficiente de simetría muestral se define como:

$$A = \frac{(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^3}{S^3} = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{X_i - \overline{X}}{S}\Big)^3$$

siendo $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$ la media muestral y $S^2 = (1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2$ la varianza muestral.

Si la distribución de los datos muestrales es normal, entonces el coeficiente de asimetría tiene distribución asintótica normal de media cero y varianza $6/n$, por lo que se puede emplear estadístico de contraste el siguiente:

$$\sqrt{\frac{n}{6}}A \sim N(0,1)$$

Se rechazará la normalidad cuando el estadístico anterior supere, en términos absolutos, a los cuantiles $\alpha/2$ y $(1-\alpha/2)$ de la distribución $N(0,1)$, siendo α el nivel de significación deseado.

Siguiendo con el caso univariante, el coeficiente de kurtosis se define como

$$K = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{X_i - \overline{X}}{S}\Big)^4 - 3$$

Si la distribución de los datos muestrales es normal, entonces la kurtosis tiene distribución asintótica normal de media cero y varianza $24/n$, por lo que se puede emplear

como estadístico de contraste el siguiente:

$$\sqrt{\frac{n}{24}}K \sim N(0,1)$$

A continuación vamos a desarrollar los tests anteriores (basados en la asimetría y la kurtosis) para el caso multivariante.

En primer lugar se estandarizan los datos multivariantes, pues esta transformación permite suprimir la media y la matriz de covarianzas. Al estandarizar los datos originales $X_1, \ldots, X_n$ se transforman en:

$$Z_i = \mathbf{S}^{-1/2}(X_i - \overline{X})$$

siendo $\overline{X}$ la media muestral y $\mathbf{S}$ la matriz de covarianzas muestral.

Las distancias de Mahalanobis de las observaciones del vector de medias se expresan:

$$r_{ii} = (X_i - \overline{X})S^{-1}(X_i - \overline{X}) = Z_i'Z_i$$

Por otro lado, se consideran los valores:

$$r_{ij} = (X_i - \overline{X})S^{-1}(X_j - \overline{X}) = Z_i'Z_j = \parallel Z_i \parallel \cdot \parallel Z_j \parallel \cdot cos(\theta_{ij})$$

siendo θ_{ij} el ángulo que forman las observaciones estandarizadas Z_i y Z_j. Las raíces cuadradas de r_{ii} y r_{jj} son las distancias de dos puntos a y b, respectivamente, al vector de medias.De este modo, r_{ij} refleja si dos puntos están en el mismo lado respecto a la media o si se encuentran en lados opuestos.

La medida de asimetría multivariante introducida por Mardia es:

$$A_m = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}^3$$

Siempre se cumple que:

$$\sum_{i=1}^{n} r_{ij} = 0 = \sum_{j=1}^{n} r_{ij}$$

esto es, se cancelan los valores positivos con los negativos. Así, si cada punto de la muestra tuviera otro punto simétrico en la propia muestra, el coeficiente de asimetría valdría cero. Si hubiera asimetría, este coeficiente sería mayor que cero. Como test de normalidad, y dado que la distribución normal es simétrica, rechazaremos la normalidad, si el coeficiente A_m es demasiado grande, comparado con ciertos valores correspondientes a muestras normales.

Por otro lado, el coeficiente de kurtosis propuesto por Mardia es el siguiente:

$$K_m = \frac{1}{n}\sum_{i=1}^{n} r_{ii}^2$$

Al ser r_{ii} el cuadrado de la distancia del dato i-ésimo al vector de medias, los valores r_{ii} que aparecen en la definición anterior son las potencias cuartas de estas distancias.

De este modo, para el coeficiente de kurtosis no importa la dirección o sentido de la desviación respecto a la media, sino el tamaño de dicha desviación.

En relación con el contraste de normalidad, el test basado en la kurtosis rechazará la normalidad si el valor absoluto del coeficiente de kurtosis K_m es demasiado grande.

Finalmente, Mardia [13] propuso una prueba de normalidad multivariante, como se ha puntualizado anteriormente, basado en la asimetría y en la kurtosis. Bajo la hipótesis nula de normalidad multivariante, el coeficiente de asimetría verifica

$$(n/6)A_m \sim \chi^2_{p(p+1)(p+2)/6}$$

De una forma similar para la kurtosis,

$$K_m \sim N(p(p+2), 8p(p+2)/n)$$

O, expresado de forma equivalente,

$$(K_m - p(p+2))\sqrt{\frac{n}{8p(p+2)}} \approx N(0,1)$$

Para la hipótesis de homocedasticidad, un test que podría aplicarse es el Test M de Box [2].

Siguiendo en la línea de lo visto hasta ahora, supongamos que tenemos a poblaciones independientes y se quiere comprobar si las matrices de covarianzas son iguales. La hipótesis nula del test sería

$$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_a$$

Sean $\mathbf{S}_1, \ldots, \mathbf{S}_a$ las matrices de covarianzas muestrales de cada población.Sea

$$\mathbf{S} = \frac{1}{n-a}\sum_{i=1}^{a}(n_i - 1)\mathbf{S}_i$$

Se definen también:

$$M = (n-a)log \mid \mathbf{S} \mid - \sum_{i=1}^{a} (n_i - 1) log \mid \mathbf{S}_i \mid$$

$$f_1 = \frac{p(p+1)(a-1)}{2}$$

$$\rho = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(a-1)} \sum_{i=1}^{a} \Big(\frac{1}{n_i - 1} - \frac{1}{n-a} \Big)$$

$$\tau = -\frac{(p-1)(p+2)}{6(a-1)} \sum_{i=1}^{a} \Big(\frac{1}{(n_i - 1)^2} - \frac{1}{(n-a)^2} \Big)$$

$$f_2 = \frac{f_1 + 2}{\mid (\tau - (1-\rho)^2)}$$

$$\gamma = (\rho - \frac{f_1}{f_2})/f_1$$

Finalmente, el test M de Box se basa en el estadístico $F = M\gamma$ que, bajo las condiciones de normalidad y la hipótesis nula de homocedasticidad se distribuye según una ley F de Snedecor:

$$F \sim F(f_1, f_2)$$

1.3. Aplicaciones

Existe un amplio abanico de situaciones prácticas donde se puede aplicar MANOVA. Por ejemplo, en el ámbito sanitario, al aplicar un tratamiento a un grupo (suficientemente grande) de personas, se puede concretar si dicho tratamiento es efectivo o no; analizar la efectividad de un tratamiento fungicida para plantaciones en agricultura; determinar si el nivel socioeconómico de la población influye en el nivel de lectura y la cultura de la misma. Estos casos serían MANOVA de un factor, pero también se puede extender a situaciones con más de un factor, como en el caso del ANOVA, cubriendo todo el amplio ámbito del diseño estadístico de experimentos cuando el objetivo queda descrito por diversas variables.

En general puede decirse que MANOVA se usa en experimentos en agricultura, medicina, biología, sociología y otras ciencias, en las que las variables pueden ser tanto

continuas como ordinales o incluso binarias.

A continuación se recogen diversos trabajos de reciente publicación en diversos ámbitos científicos, en los que se ha de aplicar la comparación de vectores medias, pero no es posible aplicar el MANOVA anteriormente descrito bien por la presencia de variables no cuantitativas, bien por la falta de normalidad o bien por la falta de homocedasticidad. En dichos trabajos se aplican las técnicas multivariantes no paramétricas recogidas en el capítulo 2.

El primer trabajo pertenece al campo de la medicina, realizado por Senker y otros [23]. Este trata sobre la relación entre fumar y una operación llamada "fusión espinal mínimamente invasiva", y su finalidad es determinar si la probabilidad de complicación de esta cirugía es mayor en personas fumadoras. Para ello, se evalúa si un historial de, al menos un paquete al año antes de la operación podría ser usado para predecir efectos adversos en pacientes que se fueran a someter a una cirugía de este tipo. En este estudio, se evalúa la afectividad clínica de MIS (cirugía mínimamente invasiva) en una población de 187 pacientes y los ratios de complicación en el perioperatorio en cirugías mínimamente invasivas de fusión espinal en pacientes fumadores y no fumadores.

Los resultados del estudio son los siguientes: los fumadores eran significativamente más jóvenes que los fumadores. No se encontró infección en el lugar de la operación o problemas en la cicatrización de las heridas en el caso de los fumadores. No se registraron diferencias entre los grupos de fumadores y no fumadores, en vistas al ratio de complicación en el perioperatorio o posoperatorio, pérdida de sangre, o el tiempo de estancia en el hospital. Sí se encontró una influencia significativa de fumar en el ratio general de complicación en el perioperatorio. La conclusión, por lo tanto, es que las técnicas de fusión MIS parecen ser una herramienta adecuada para tratar enfermedades degenerativas espinales en fumadores.

El segundo estudio se sitúa en el campo de la sociología. Buckley y otros [3] exploran los prototipos de la definición de "ingeniero/a inteligente".

Los varones están sobrerrepresentados en las ingenierías, pero la magnitud de esta varianza difiere en los países y los campos de la ingeniería. Este estudio examinó las percepciones sobre "un/a ingeniero/a inteligente" según estudiantes de ingeniería de Irlanda y Suecia, algunos que no se habían graduado aún y estudiantes que ya

habían graduado. Estos dos países fueron seleccionados basados en sus niveles de representación femenina en ingenierías. La hipótesis realizada fue que habría diferencias significativas de las percepciones entre los países. Una muestra aleatoria de estos estudiantes respondieron a dos encuestas: la primera preguntaba sobre las características de un/a ingeniero/a inteligente, y la segunda pedía un ranking de la importancia de cada característica.

Los resultados indicaron que un/a ingeniero/a inteligente estaría descrito por siete factores: resolver problemas prácticos, meticulosidad, dirección, conocimiento, razonamiento, atributos negativos y curiosidad. Esto resultó cuando los datos se analizaron los dos países conjuntamente; sin embargo, cuando los datos de cada país se analizaron independientemente, solo se pudieron interpretar los siguientes atributos: resolver problemas prácticos, meticulosidad, dirección, conocimiento y atributos negativos. Se observó una interacción del país y el género para cada uno de esos cinco factores. Los resultados fueron similares respecto a los factores que denotan inteligencia en Irlanda y Suecia, pero se percibieron diferencias en términos de cuán importantes son dichos factores.

El tercer estudio se sitúa en el campo de la ecología. Diesburg y otros [8] tratan de determinar las consecuencias de la invasión de un insecto en las redes tróficas de las zonas ribereñas en los Apalaches Centrales. En primer lugar, una red trófica es la interconexión natural de las cadenas alimenticias y generalmente es una representación gráfica de quién se come a quién en una comunidad ecológica. Por otro lado, las zonas ribereñas son la interfase entre la tierra y un río o un arroyo.

El invasor terrestre "adélgido lanudo de la tuya" (Adelges tsugae) diezma la cicuta oriental (Tsuga canadensis); éste domina las zonas ribereñas de los arroyos de los bosques apalaches. Sin embargo, las consecuencias ecológicas para los ecosistemas enlazados acuáticos y terrestres son inciertas. Se trataron de medir los vínculos tróficos en riberas con arroyos en 21 sitios de Ohio, Virginia del Este, y Virginia representando la cronología de la disminución de T. canadensis. Se midieron recíprocamente flujos de recursos basales (perifiton, detritos terrestres), la tasa de flujo de insectos acuáticos emergentes y la composición de este flujo, la densidad de arañas tejedoras de orbes ribereñas, y la posición trófica estimada de la araña y dependencia de energía de origen acuático. Los resultados fueron los siguientes: se encontró una mayor biomasa de perifiton en los arroyos de zonas no invadidas que en los sitios invadidos y la composición del

flujo de detritos de la tierra al agua cambió con el declive de T.canadensis. La composición de insectos acuáticos se explica en parte por este mismo declive. Las densidades de araña tejedora de orbes fueron más altas en los sitios con una severa disminución de T.canadensis, pero no estaban vinculados a tasas de flujo de emergencia de insectos. En general, las consecuencias ecológicas de este invasor fueron más claras en los niveles tróficos inferiores, con impactos más sutiles en las arañas ribereñas.

Capítulo 2

TÉCNICAS NO PARAMÉTRICAS MULTIVARIANTES DE COMPARACIÓN DE POBLACIONES

En este capítulo se va a hacer una descripción teórica de las técnicas y los test estadísticos que han sido mencionados en la introducción.
Para cada test se hará un breve desarrollo teórico, que abordará el contraste de hipótesis, la determinación del estadístico de contraste y la distribución de dicho estadístico.

2.1. Modelo e hipótesis previas

Se considera la misma notación que en el caso paramétrico: p para el número de variables, a para el número de tratamientos, niveles del factor o poblaciones a tener en cuenta, y n o n_i para el tamaño de la muestra de cada tratamiento. La principal diferencia entre la inferencia paramétrica y no paramétrica es que en la inferencia paramétrica se supone conocida la distribución de los vectores aleatorios, salvo parámetros, mientras que en la no paramétrica no se realiza hipótesis de distribución alguna. Por lo tanto, la inferencia paramétrica podemos aplicarla en los distintos parámetros de las distribuciones, pero la inferencia no paramétrica hay que aplicarla directamente sobre las distribuciones.

Consideremos un p-vector aleatorio X definido en a poblaciones, con función de distribución en la población i-ésima F_i . Sean $X_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(p)})' \sim F_i, i = 1, \ldots, a, j = 1, \ldots, n_i$ vectores independientes con distribuciones multivariantes F_i. Los vectores X_{ij} tendrán componentes dependientes $X_{ij}^{(k)}$ con distribuciones marginales

$$F_i^{(k)}(x) = \frac{1}{2}\left(P(X_{ij}^{(k)} \leq x) + P(X_{ij}^{(k)} < x)\right), k = 1, \ldots, p$$

Esta versión de la función de distribución se llama versión normalizada, y su uso tiene ventajas cuando se usan tests no paramétricos porque sirve para todo tipo de variables, ya sean cuantitativas, ordinales o incluso binarias.

$X_{ij}^{(k)}$ es la observación de la variable k en el sujeto j del tratamiento i.

Sea $R_{ij}^{(k)}$ el rango de $X_{ij}^{(k)}$ a través de las $N = \sum_{i=1}^{a} n_i$ variables aleatorias $X_{11}^{(k)}$, …, $X_{an_a}^{(k)}$.

Sea el vector columna $R_{ij} = (R_{ij}^{(1)}, \ldots, R_{ij}^{(p)})'$, que contiene todos los rangos de una observación multivariante, mientras que la matriz $R = (R_{11}, \ldots, R_{1n_1}, R_{21}, \ldots, R_{an_a})$ contiene los rangos de todas las observaciones para cada variable.

Bathke et al.[1] basó el desarrollo de estos tests en rangos, por lo que la hipótesis nula será la igualdad de las funciones de distribución de las distintas poblaciones.

El contraste de hipótesis que se va a considerar en todos los test es

$$H_0 : F_1 = \cdots = F_a$$

$$H_1 : F_i \neq F_j \text{ para algún } i, j \in \{1, \ldots, a\}$$

Se considera que un modelo es balanceado cuando $n_1 = \ldots = n_a = n$, por lo que el número total de observaciones es $N = n \cdot a$. Para este tipo de modelos, se definen las siguientes matrices, que serán necesarias para obtener el estadístico :

$$\mathbf{H} = \frac{1}{a-1}\sum_{i=1}^{a} n(\overline{R_{i\cdot}} - \overline{R_{\cdot\cdot}})(\overline{R_{i\cdot}} - \overline{R_{\cdot\cdot}})'$$

$$\mathbf{G} = \frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n}(R_{ij} - \overline{R_{i\cdot}})(R_{ij} - \overline{R_{i\cdot}})'$$

Donde $\overline{R_{i\cdot}}$ denota la media sobre el índice que corresponde al punto. Estas matrices representan las matrices suma de cuadrados y los productos cruzados definidas sobre

los rangos, definidas de forma idéntica a las matrices $\mathbf{B}$ y $\mathbf{W}$, respectivamente, sobre las observaciones originales.

Para el modelo no balanceado, es decir, cuando los tamaños de las muestras $n_1,\ldots,n_a$ son distintos, vamos a definir las siguientes matrices :

$$\mathbf{H}_1 = \frac{1}{a-1}\sum_{i=1}^{a} n_i(\overline{R_{i.}} - \overline{R_{..}})(\overline{R_{i.}} - \overline{R_{..}})',$$

$$\mathbf{H}_2 = \frac{1}{a-1}\sum_{i=1}^{a}(\overline{R_{i.}} - \tilde{R}_{..})(\overline{R_{i.}} - \tilde{R}_{..})',$$

$$\mathbf{G}_1 = \frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(R_{ij} - \overline{R_{i.}})(R_{ij} - \overline{R_{i.}})',$$

$$\mathbf{G}_2 = \frac{1}{a-1}\sum_{i=1}^{a}\left(1 - \frac{n_i}{N}\right)\frac{1}{n_i-1}\sum_{j=1}^{n_i}(R_{ij} - \overline{R_{i.}})(R_{ij} - \overline{R_{i.}})',$$

$$\mathbf{G}_3 = \frac{1}{a}\sum_{i=1}^{a}\frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}(R_{ij} - \overline{R_{i.}})(R_{ij} - \overline{R_{i.}})'$$

Se usa la siguiente notación para indicar dos formas distintas de calcular la media general. Los elementos del vector $\overline{R}_{..}$ son

$$\overline{R}_{..}^{(k)} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^{(k)}$$

mientras que los elementos de $\tilde{R}_{..}$ son

$$\tilde{R}_{..}^{(k)} = \frac{1}{a}\sum_{i=1}^{a}\overline{R}_{i.}^{(k)} = \frac{1}{a}\sum_{i=1}^{a}\frac{1}{n_i}\sum_{j=1}^{n_i} R_{ij}^{(k)}$$

Si el diseño es balanceado, se pueden usar ambas expresiones indistintamente pues serían idénticas.

En las dos matrices $\mathbf{H}_1$ y $\mathbf{H}_2$ se calculan los cuadrados medios debido al tratamiento, pero la matriz $\mathbf{H}_1$ usa una media ponderada, mientras que $\mathbf{H}_2$ no. Por otro lado, las matrices $\mathbf{G}_1$,$\mathbf{G}_2$ y $\mathbf{G}_3$ calculan la suma de cuadrados debida al error, pero con diferentes ponderaciones.

En el caso del modelo no balanceado, los test están basados en uno de los pares $(\mathbf{H}_1, \mathbf{G}_1), (\mathbf{H}_1, \mathbf{G}_2)$ ó $(\mathbf{H}_2, \mathbf{G}_3)$, que corresponden al par $(\mathbf{H}, \mathbf{G})$ cuando estamos en el modelo balanceado.

En las siguientes secciones vamos a centrarnos en el desarrollo teórico de los tests para aproximaciones finitas basados en teoría asintótica cuando a y n_i son moderadamente pequeños, basándonos en Bathke et al. [1] y Liu et al. [12]. Esto es debido a que la mayoría de experimentos comprenden un número pequeño y moderado tanto de poblaciones como de tamaños muestrales. En ambos artículos se incluye el desarrollo teórico cuando $a \to \infty$ y cuando algún $n_i \to \infty$.

A continuación vamos a describir los estadísticos y su distribución.

2.2. Estadístico tipo ANOVA

Para el modelo balanceado, este estadístico Brunner [6] y Srivastava y Fujikoshi [24] lo definen como

$$\mathrm{T}_A = \frac{tr(\mathbf{H})}{tr(\mathbf{G})}$$

y sigue aproximadamente, bajo hipótesis nula, una distribución F con grados estimados de libertad:

$$\hat{f} = (a-1)\frac{tr(\mathbf{G})^2}{tr(\mathbf{G}^2)} \text{ y } \hat{f}_2 = \frac{a^2}{(a-1)\sum_{i=1}^{a}\frac{1}{n_i-1}}\hat{f}$$

Para el modelo no balanceado, Munzel y Brunner ([16] y [17]) definen el estadístico como

$$\mathrm{T}_A = \frac{tr(\mathbf{H}_2)}{tr(\mathbf{G}_3)}$$

y sigue aproximadamente, bajo hipótesis nula, una distribución F con grados estimados de libertad:

$$\hat{f}_1 = (a-1)\frac{tr(\mathbf{G}_3)^2}{tr(\mathbf{G}_3^2)} \text{ y } \hat{f}_2 = \frac{a^2}{(a-1)\sum_{i=1}^{a}\frac{1}{n_i-1}}\hat{f}_1$$

Obsérvese que la traza de la matriz $\mathbf{H}$ (o bien $\mathbf{H}_2$) representa la suma, para cada una de las p variables, de la varianza muestral medida 'entre' las poblaciones o grupos. Asimismo, la traza de la matriz $\mathbf{G}$ (o bien $\mathbf{G}_3$) representa la suma, para cada una de las p variables, de la varianza muestral medida "dentro"de cada población o grupo. De ahí que, por su similitud al análisis de la varianza univariante, se denominen estadísticos tipo ANOVA.

2.3. Aproximación de McKeon para el test de Lawley-Hotelling

Para el modelo balanceado el estadístico lo define McKeon [14] como

$$\mathrm{U} = \frac{(a-1)}{(N-a)}\mathrm{tr}(\mathbf{H}\mathbf{G}^{-1})$$

Recuérdese que el estadístico de Lawley-Hotelling bajo condiciones de normalidad se define como función de la traza de $\mathbf{B}\mathbf{W}^{-1}$, es decir, el producto de la matriz variabilidad "entre" grupos por la inversa de la matriz variabilidad "dentro"de los grupos. En este caso, se define de forma similar para las matrices que miden la variabilidad en los rangos. La distribución de U se aproxima, bajo hipótesis nula, por $g \cdot \mathrm{F}_{K,D}$, una distribución F "ensanchada" siendo K y D los grados de libertad, donde

$$K = p(a-1)$$

$$D = 4 + \frac{K+2}{B-1}$$

$$B = \frac{(N-p-2)(N-a-1)}{(N-a-p)(N-a-p-3)}$$

y

$$g = \frac{p(a-1)(D-2)}{(N-a-p-1)D}$$

Para el modelo no balanceado el estadístico se define en [5] como

$$\mathrm{U}_n = \frac{(a-1)}{(N-a)}\mathrm{tr}(\mathbf{H}_1\mathbf{G}_1^{-1})$$

La distribución de U_n se aproxima, bajo hipótesis nula, por $g \cdot \mathrm{F}_{K,D}$, una distribución F "ensanchada" siendo K y D los grados de libertad, donde

$$K = p(a-1)$$

$$D = 4 + \frac{K+2}{B-1}$$

$$B = \frac{(N-p-2)(N-a-1)}{(N-a-p)(N-a-p-3)}$$

y

$$g = \frac{p(a-1)(D-2)}{(N-a-p-1)D}$$

2.4. Aproximación de Muller para el test de Bartlett-Nanda-Pillai

De forma similar al estadístico traza de Pillai recogido en el Capítulo 1, definido para las matrices de dispersión de los datos originales, se puede considerar un estadístico definido sobre las matrices de dispersión de los rangos. Sea

$$\mathrm{V}= \mathrm{tr}\left\{(a-1)\mathbf{H}\left[(a-1)\mathbf{H}+(N-a)\mathbf{G}\right]^{-1}\right\}$$

Para el modelo balanceado, Muller [15] define el estadístico como

$$\mathrm{BNP} = \frac{(V/\gamma)/\nu_1}{(1-V/\gamma)/\nu_2}$$

y sigue una aproximadamente, bajo hipótesis nula, distribución F con ν_1 y ν_2 grados de libertad, siendo

$$\gamma = \min(a-1, p)$$
$$\nu_1 = \frac{p(a-1)}{\nu(N-1)}\left[\frac{\nu(N-a+\gamma-p)(N+2)(N-1)}{(N-a)(N-p)}-2\right]$$
$$\nu_2 = \frac{N-a+\gamma-p}{N}\left[\frac{\gamma(N-a+\gamma-p)(N-1)}{(N-a)(N-p)}-2\right]$$

Sea

$$\mathrm{V}_1 = \mathrm{tr}\left\{(a-1)\mathbf{H}_1\left[(a-1)\mathbf{H}_1+(N-a)\mathbf{G}_1\right]^{-1}\right\}$$

Para el modelo no balanceado, el estadístico se define en [5] como

$$\mathrm{BNP} = \frac{(V_1/\gamma)/\nu_1}{(1-V_1/\gamma)/\nu_2}$$

y sigue, bajo hipótesis nula, una distribución F con ν_1 y ν_2 grados de libertad, siendo

$$\gamma = \min(a-1, p)$$
$$\nu_1 = \frac{p(a-1)}{\nu(N-1)}\left[\frac{\nu(N-a+\gamma-p)(N+2)(N-1)}{(N-a)(N-p)}-2\right]$$
$$\nu_2 = \frac{N-a+\gamma-p}{N}\left[\frac{\gamma(N-a+\gamma-p)(N-1)}{(N-a)(N-p)}-2\right]$$

2.5. Estadístico Lambda de Wilks

De forma similar a los anteriores, se puede considerar la versión no paramétrica del estadístico Lambda de Wilks, aplicada a las matrices de rangos. Sea

$$\lambda = \det\left(\frac{(N-a)\mathbf{G}}{(N-a)\mathbf{G} + (a-1)\mathbf{H}}\right)$$

Liu [12] define el estadístico como

$$\mathrm{F} = \frac{(1-\lambda^{1/t})}{(\lambda^{1/t})}\frac{df_2}{df_1}$$

y sigue aproximadamente, bajo hipótesis nula, una distribución F de df_1 y df_2 grados de libertad, siendo

$$df_1 = p(a-1)$$

$$df_2 = rt - (p(a-1)-2)/2$$

y

$$r = (N-a) - (p-(a-1)+1)/2$$

Si $p(a-1) = 2$, entonces $t = 1$, y en otro caso

$$t = \sqrt{\frac{p^2(a-1)^2 - 4}{p^2 + (a-1)^2 - 5}}$$

2.6. Tests basados en permutaciones y algoritmos de subconjuntos

En las secciones anteriores se han mostrado cuatro estadísticos cuya distribución es una aproximación de una distribución F. Como se ha visto anteriormente, estos tests se usan para comprobar la igualdad de las funciones de distribución de las distintas poblaciones. Cuando el tamaño de las muestras es pequeño o moderado, Liu et al. [12] proponen los tests multivariantes de permutaciones como alternativa para analizar dicho contraste de hipótesis, así como Bathke et al. [5] proponen el algoritmo de subconjuntos.

2.6.1. Tests basados en permutaciones

Para los tests multivariantes de permutaciones, se usa como base uno de los cuatro estadísticos. El algoritmo que se sigue es el siguiente:

1. En primer lugar, se escoge uno de ellos y se calcula el estadístico que corresponda, que se denotará por t_1.

2. Sea $\mathbf{X}$ la $(N \times p)$-matriz que contiene los N vectores de datos originales independientes, de forma que las n_1 primeras filas corresponden a la primera población, las n_2 filas siguientes a la segunda población, y así sucesivamente.

3. Los tests de permutaciones se basan en determinar las permutaciones en el orden de las filas, y por tanto, en la asignación de los casos a las poblaciones y obtener el estadístico para cada una de las permutaciones generadas. Así, se genera una permutación de los datos, es decir, de las filas de la matriz $\mathbf{X}$, obteniendo una matriz $\mathbf{X^*}$. En esta matriz, se considera que las primeras n_1 filas corresponden al primer tratamiento, las siguientes n_2 filas corresponden al tratamiento 2, y así sucesivamente.

4. Para esta matriz se vuelve a calcular el estadístico, que denotaremos por t_1^*.

5. Repitiendo el proceso anterior con todas las permutaciones posibles de los datos, se obtiene una distribución de valores del estadístico t_1. Dado que esa distribución de valores se ha obtenido mezclando los datos en las poblaciones, se puede suponer que dicha distribución se aproxima a la distribución de t_1 bajo la hipótesis de igualdad de poblaciones. Así, el valor original t_1 se compara con todos los valores que se obtengan en cada permutación, para establecer el cuantil de t_1 en la distribución de permutación.

6. Pueden llevarse a cabo $N!$ permutaciones o bien un número predeterminado de permutaciones elegidas o generadas aleatoriamente. Este último se usa, en general, como alternativa al test de permutación cuando N es demasiado grande, ya que esto puede convertirse en un problema al tener que calcular computacionalmente todos los valores de t_1.

2.6.2. Tests basados en algoritmos de subconjuntos

El algoritmo de subconjuntos se usa para analizar qué variables muestran diferencias significativas de las demás y también qué factores contribuyen a resultados significativos. El algoritmo que se implementa es el siguiente: supongamos que tenemos 4 tratamientos.

La hipótesis de igualdad de los tratamientos 1 y 3,

$$H_0^{(1,3)} : F_1 = F_3$$

se puede rechazar sólo si todas las hipótesis en las que se consideren el tratamiento 1 o el 3 se rechazan también, incluyendo la anterior. Es decir, además de $H_0^{(1,3)}$, deben ser rechazadas las siguientes hipótesis:

$$H_0^{(1,2,3)} : F_1 = F_2 = F_3$$

$$H_0^{(1,3,4)} : F_1 = F_3 = F_4$$

$$H_0^{(1,2,3,4)} : F_1 = F_2 = F_3 = F_4$$

y

$$H_0^{(1,3)(2,4)} : (F_1 = F_3) \wedge (F_2 = F_4)$$

Por lo tanto, el algoritmo empieza con el test global multivariante ($H_0 : F_1 = F_2 = F_3 = F_4$) a un nivel α fijado. El α que se fija es el mismo para todos los tests. Además, hay que elegir un test de los cuatro presentados anteriormente como base, el cual se realiza sobre los datos originales.

Si se rechaza la hipótesis del test global, se pasa a la siguiente etapa. Si, por el contrario, se acepta la igualdad de distribuciones, no se realiza ningún test más. En el caso de que se pase a la siguiente etapa, se llevarían a cabo los cuatro tests cuya hipótesis nula contempla la igualdad de distribuciones de tres poblaciones, éstas son:

$$H_0^{(1,2,3)} : F_1 = F_2 = F_3$$

$$H_0^{(1,3,4)} : F_1 = F_3 = F_4$$

$$H_0^{(1,2,4)} : F_1 = F_2 = F_4$$

y

$$H_0^{(2,3,4)} : F_2 = F_3 = F_4$$

a un nivel $\frac{3}{4}\alpha$. Si alguna de ellas se rechazara, se seguiría sucesivamente con los tests a un nivel $\frac{2}{4}\alpha$. El número máximo de tests que se llevarían a cabo es $2^{min(a,p)} - 1$

2.7. ¿Qué test elegir?

Como hemos expuesto hasta ahora en esta memoria, para el análisis de datos que no cumplan los requisitos para aplicar metodos paramétricos, se dispone de diversos métodos alternativos. En la práctica, ninguno de ellos muestra mejores resultados que los demás. Dependiendo del caso práctico en el que nos encontremos, será más recomendable centrarnos en uno u otro. En general, todos suelen coincidir, pero en algunos casos difieren ligeramente. Basándose en diferentes estudios de simulación, Bathke [5] sugiere lo siguiente:

1. El estadístico Lambda de Wilks se usa en todas las situaciones donde sea posible.

2. Para datos de dimensiones demasiado grandes, el único que se puede usar es el ANOVA-type. Además, se usa cuando el estadístico Lambda de Wilks no se puede.

3. Para $N < 10$ se lleva a cabo el test de permutación. Para $10 \leq N < 30$, el test aleatorio se lleva a cabo con 10000 permutaciones. Para $N \geq 30$, se usa la aproximación de F. Esto es debido al coste computacional que conllevan los cálculos cuando N es demasiado grande.

Como se indicó al principio de este capítulo, sólo se ha llevado a cabo el desarrollo teórico de los tests que se implementarán en el siguiente capítulo. Es decir, aquellos que abarcan el caso en el que tanto los tamaños muestrales como el número de poblaciones son finitos. Sin embargo, tanto en Bathke et al.[1] como en Liu et al.[12] se desarrollan una gran cantidad de tests para los demás casos. Dichos casos son, tanto para el caso balanceado como para el caso no balanceado:

- $a \to \infty$, n y p fijados.

- $n \to \infty$, o $\min n_i \to \infty$, a y p fijados.

En el caso de que a, n y p son finitos, en Bathke et al.[1] se encuentran algunos estadísticos como una expansión de la distribución χ^2 para los tests Lawley-Hotelling y Bartlett-Nanda-Pillai y la expansión Fujikoshi, también para los tests Lawley-Hotelling y Bartlett-Nanda-Pillai.

Capítulo 3

IMPLEMENTACIÓN EN R

3.1. Descripción del paquete npmv

El paquete de R **npmv** (Burchett y Ellis [5]) lleva a cabo un análisis de un conjunto de datos multivariantes usando técnicas no paramétricas. Realiza una comparación de las distribuciones multivariantes para una sola variable explicativa.

Para el análisis de la varianza multivariante paramétrico clásico (MANOVA) es necesario que se cumplan algunas hipótesis como la de normalidad multivariante, pero son muy restrictivas, además de no proporcionar mucha información útil sobre las variables respuesta o los niveles del factor. Las técnicas del paquete **npmv** contemplan algunas mejoras con respecto a lo anterior: las variables respuesta pueden ser binarias, ordinales o cuantitativas, no hace falta la hipótesis de normalidad multivariante y se identifican los subconjuntos de variables respuesta o niveles del factor que son significativos.

Este paquete contiene las funciones *nonpartest* y *ssnonpartest*.

3.2. La función *nonpartest*

Esta función lleva a cabo los tests no paramétricos (aproximaciones de F), sus tests de permutación y calcula los efectos relativos de cada variable respuesta.

La orden es :

`nonpartest(formula,data,permtest = TRUE , permreps = 1000, plots=TRUE, tests= c(1,1,1,1), releffects=TRUE)`,

siendo los argumentos de entrada:

1. `formula`: es un objeto de clase 'formula' con una sola variable explicativa (grupo) y varias variables respuesta.

2. `data`: es un objeto 'data.frame' que contiene todas las variables.

3. `permtest`: si es TRUE, devuelve los p-valores de los test de permutación.

4. `permreps`: número de repeticiones de los tests de permutación.

5. `plots`: si es TRUE, devuelve los boxplots de cada variable respuesta frente a la variable explicativa.

6. `tests`: especifica los tests que se van a calcular. Si es 1 se calcula el test correspondiente, y si es 0 no se calcula. El orden es: tipo ANOVA, tipo Lawley-Hotelling, tipo Bartlett-Nanda-Pillai y Lambda de Wilks.

7. `releffects`: si es TRUE, devuelve los efectos relativos de las variables respuesta. El efecto relativo del tratamiento "k" se define como *"la probabilidad de que un sujeto escogido aleatoriamente del tratamiento "k" muestre un valor respuesta mayor que un sujeto escogido aleatoriamente de cualquier tratamiento, incluido el "k"*([5]).

Lo primero que devuelve la función es un diagrama de cajas de cada variable respuesta frente a la variable explicativa, donde se representa la mediana y los valores atípicos de cada una, para tener una "idea" visual de los datos. Lo siguiente que devuelve son dos cuadros de datos: el primero contiene los valores de los estadísticos y los grados de libertad de cada uno, los p-valores de cada estadístico correspondientes a los tests no paramétricos y los tests de permutación. El segundo contiene los efectos relativos que corresponde a cada variable respuesta (este segundo se devuelve si `releffects`=TRUE en la orden).
Si en el conjunto de datos falta alguno, se produce un "warning".

3.3. La función *ssnonpartest*

La función se basa en un algoritmo de subconjuntos para ver qué variables y subconjuntos de variables causan resultados significativos y si hay algún nivel del factor (tratamiento) que destaque sobre los demás. El algoritmo que se implementa es el que se muestra en la Sección 2.6.2.

La orden es:
`ssnonpartest(formula,data,alpha=0.5,test=c(1,0,0,0),`
`factors.and.variables=TRUE)`,
siendo los argumentos de entrada:

1. `formula`: es un objeto de clase 'formula' con una sola variable explicativa (grupo) y varias variables respuesta.

2. `data`: es un objeto 'data.frame' que contiene todas las variables.

3. `alpha`: es el nivel global de significación con el que se realizan los tests de hipótesis. Es 0.05 por defecto.

4. `test`: especifica el test que se va a calcular, el 1 corresponde al test que se va a realizar. El orden es: tipo ANOVA, tipo Lawley-Hotelling, tipo Bartlett-Nanda-Pillai y Lambda de Wilks.

5. `factors.and.variables`: si es TRUE el algoritmo se realiza para variables y factores. Es FALSE por defecto. Si $p \leq a$,el algoritmo usa subconjuntos de variables y, en otro caso, subconjuntos de los factores.

Devuelve el resultado del algoritmo, mostrando los subconjuntos de variables y de factores que muestran resultados significativos.

3.4. Ilustración

3.4.1. Datos "sberry"

A continuación se implementan las dos funciones anteriores en un conjunto de datos llamado *sberry* incluido en el paquete **npmv**, basado en un trabajo de Horst y otros [10]

Estos datos fueron recogidos en un estudio de una plantación de fresas para evaluar los efectos de tres fungicidas (tratamientos). La plantación se dividió en 16 parcelas, y en cada 4 de ellas se usó aleatoriamente un fungicida, siendo las 4 parcelas restantes el grupo de control. Las 6 variables respuesta que se consideran son:

- X_1: treatment - fungicida (tratamiento) aplicado

- X_2: replication

- X_3: weight - peso total de las fresas recogidas
- X_4: bot - porcentaje de Botrytis (un tipo de hongo)
- X_5: fungi - porcentaje causado por otros tipos de hongos
- X_6: rating - marcador de síntomas causados por Phomosis (otro tipo de hongo). Esta es la única variable ordinal que abarca de 0-3, asignando 0 cuando no hay síntomas y 3 cuando hay más del 40 % del follaje dañado.

El objetivo es comprobar si hay diferencias entre los tratamientos y, en ese caso, entre qué variables respuestas y qué tratamientos.
A continuación se muestra la tabla de datos:

	treatment	replication	weight	bot	fungi	rating
1	3	1	6.90	4.0956	17.2355	1.0
2	3	2	8.30	5.1348	5.6482	1.0
3	3	3	8.40	6.0698	8.8012	1.5
4	3	4	7.95	2.7174	9.5109	1.5
5	6	1	8.60	1.1945	17.0649	1.0
6	6	2	8.50	0.5533	12.8631	1.0
7	6	3	8.20	0.7353	6.7647	0.5
8	6	4	9.50	0.9929	1.8440	1.0
9	8	1	6.20	4.2857	4.6428	1.0
10	8	2	9.00	1.5640	3.0303	3.0
11	8	3	6.80	0.8757	5.6042	0.0
12	8	4	8.50	2.4249	8.6606	2.0
13	9	1	7.50	15.5975	13.0817	1.0
14	9	2	6.70	10.2819	14.4279	1.0
15	9	3	8.70	13.2895	10.9211	2.5
16	9	4	7.40	18.3824	16.0295	3.0

En primer lugar, vamos a aplicar el test de Mardia para comprobar si los datos siguen una distribución normal multivariante.

```
mardiaTest(sberry)
```

Los p-valores obtenidos son **0.427133 y 0.208007**. correspondientes a los tests del coeficiente de simetría y coeficiente de kurtosis, respectivamente. Por lo tanto, no podemos rechazar la hipótesis de normalidad multivariante. Si se aceptara la hipótesis de homocedasticidad, el conjunto de datos cumpliría las dos hipótesis necesarias para aplicar MANOVA (normalidad multivariante y homocedasticidad). Para analizar la homocedasticidad se aplica el Test M de Box, visto en la Sección 1.2.

```
MBox(sberry[,2:6],sberry[,1])
```

El p-valor que obtenemos es 0, por lo tanto se rechaza la hipótesis de homocedasticidad. Esto nos conduce a aplicar las técnicas no paramétricas multivariantes vistas anteriormente. Además, este conjunto de datos contiene variables no cuantitativas (rating), por lo que igualmente no se podría aplicar MANOVA.

En primer lugar, se aplica la función `nonpartest`:

```
nonpartest(weight|bot|fungi|rating ~ treatment,sberry,permreps=1000)
```

Esta función se implementa para analizar cómo afecta el tratamiento a las variables respuesta, eliminando la variable *replication*, que se usará más tarde. Como vimos anteriormente, lo primero que devuelve la función son los diagramas:

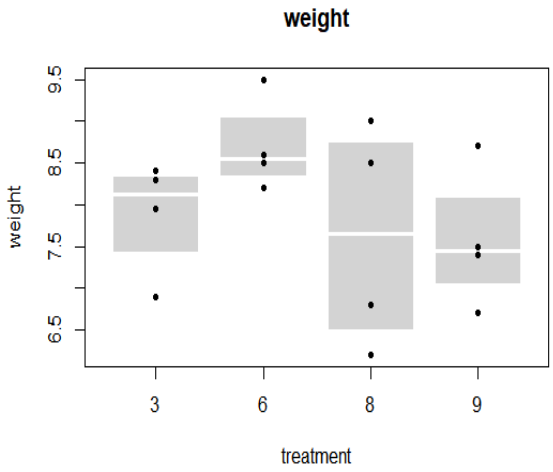

Figura 3.1: Peso frente a tratamiento

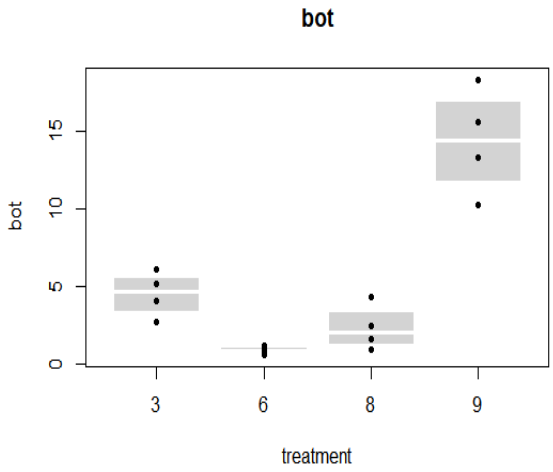

Figura 3.2: Bot frente a tratamiento

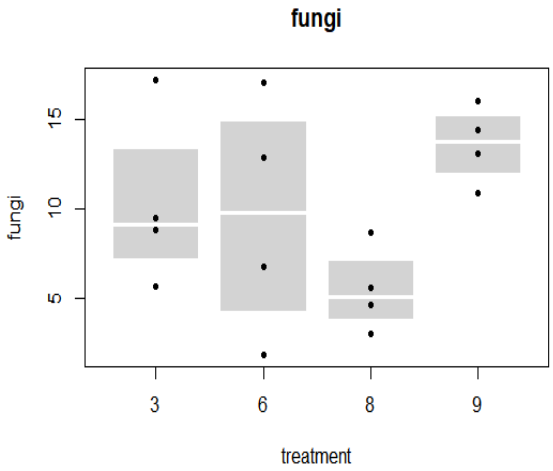

Figura 3.3: Fungi frente a tratamiento

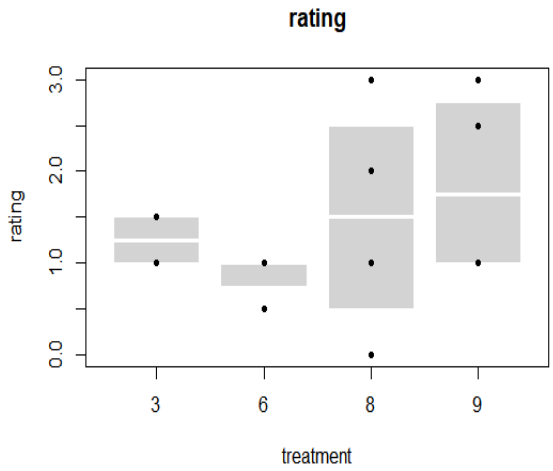

Figura 3.4: Rating frente a tratamiento

Como podemos ver en la Figura 3.2, el tratamiento influye en el porcentaje de Botrytis, pues las parcelas en las que se aplicó el tratamiento 6, por ejemplo, tienen porcentajes mucho más bajos que las parcelas en las que se aplicó el tratamiento 9.

A continuación se muestra el valor de los estadísticos de los tests no paramétricos y sus grados de libertad correspondientes, así como los p-valores de cada estadístico correspondientes a los tests no paramétricos y los tests de permutación.

```
$results
                Test Statistic     df1      df2  p-value  PT p-value
ANOVA type Test          2.984   6.836  27.3426    0.019       0.009
LHT Test                 5.769   12.00    12.00    0.002       0.006
BNP Test                 2.501  15.967  41.1641    0.009       0.007
Wilks Lambda             4.166   12.00  24.1033    0.001       0.002
```

En esta tabla hay algunas abreviaciones, siendo `PT p-value` el p-valor del test de permutación, `LHT Test` la aproximación de McKeon para el test de Lawley Hotelling, y `BNP Test` la aproximación de Muller para el test de Bartlett-Nanda-Pillai.

Observando los p-valores, vemos que todos son menores que 0.05 (nivel de significación), por lo que se llega a la conclusión de que el tratamiento es bastante significativo. Es decir, el tratamiento que se aplique influye en el valor de las variables respuesta.

Por último, se devuelven los efectos relativos, que expresan las tendencias observadas en las variables respuesta en términos de probabilidad.

```
$releffects
   weight     bot   fungi  rating
3  0.4375  0.5938  0.5625  0.5313
6  0.7266  0.1563  0.4843  0.3047
8  0.4453  0.3750  0.2188  0.5313
9  0.3906  0.8750  0.7344  0.6328
```

La interpretación de los efectos relativos es la siguiente: la probabilidad de que una planta escogida aleatoriamente del grupo 3 tenga un porcentaje de Botrytis mayor que cualquier otra planta de toda la plantación (incluyendo las del tratamiento 3) es de 0.5938. Se observa que las probabilidades correspondientes al factor 9 (grupo de control) son más altas en las variables bot, fungi y rating respecto a los demás factores, es decir, donde se ha aplicado algún tratamiento. Además, la probabilidad de la variable weigth es menor en dicho grupo respecto a los demás. Esto significa que cuando no se aplica ningún tratamiento fungicida, las probabilidades de que la plantación sufra daños debido a algún tipo de hongo es mayor que cuando se aplican tratamientos y, además, el peso total de la plantación será menor.

A continuación aplicamos la función `ssnonpartest`:
`ssnonpartest(weight|bot|fungi|rating ~ treatment,data=sberry,test=c(0,0,0,1), alpha=0.05, factors.and.variables=TRUE)`.

Al poner el último 1 en el vector, el test que se usará en cada subconjunto es el Lambda de Wilks. Además, al activar factors.and.variables, el algoritmo se llevará a cabo tanto para las variables respuesta como para los tratamientos.

La función devuelve lo siguiente:

```
The Wilks' Lambda type statistic will be used in the following test
The Global Hypothesis is significant, subset algorithm will continue
Performing the Subset Algorithm based on Factor levels
The Hypothesis of equality between factor levels 3 6 8 9 is rejected
The Hypothesis of equality between factor levels 6 8 9 is rejected
The Hypothesis of equality between factor levels 3 8 9 is rejected
The Hypothesis of equality between factor levels 3 6 9 is rejected
The Hypothesis of equality between factor levels 3 6 8 is rejected
The Hypothesis of equality between factor levels 3 6 is rejected
All appropriate subsets using factor levels have been checked using a closed
multiple testing procedure, which controls the maximum overall type I error rate
at alpha= 0.05
```

```
Performing the Subset Algorithm based on Response Variables
The Hypothesis of equality using response variables weight bot fungi rating is
rejected
The Hypothesis of equality using response variables bot fungi rating is
rejected
The Hypothesis of equality using response variables weight bot rating is rejected
The Hypothesis of equality using response variables weight bot fungi is rejected
The Hypothesis of equality using response variables bot rating is rejected
The Hypothesis of equality using response variables bot fungi is rejected
The Hypothesis of equality using response variables weight bot is rejected
The Hypothesis of equality using response variables bot is rejected
All appropriate subsets using response variables have been checked using a
multiple testing procedure, which controls the maximum overall type I error rate
at alpha= 0.05
```

En la primera línea se indica que se usará el test Lambda de Wilks. La segunda línea señala que hay resultados significativos usando todas las variables y todos los tratamientos, es decir, hay diferencias entre las variables respuestas y los distintos tratamientos. Si no se hubieran apreciado diferencias, no se seguiría con el algoritmo. A continuación se muestran cuáles son las variables y los tratamientos significativos.

En primer lugar están los tests basados en los tratamientos. Sólo se muestran los subconjuntos significativos, y el algoritmo llega hasta los subconjuntos con dos tratamientos, pues no tiene sentido hacer un test para un sólo tratamiento. En este caso, se encuentran resultados significativos en todos los subconjuntos de 3 tratamientos, y también en el subconjunto formado por el tratamiento 3 y 6. Es decir, se encuentran diferencias entre estos dos tratamientos, por lo que los valores de las variables respuesta diferirán dependiendo si el tratamiento es el 3 o el 6.

En segundo lugar se encuentran los tests basados en las variables respuesta. En este caso sí tiene sentido llegar hasta los subconjuntos formados por una sola variable, pues una variable puede ser significativa por sí sola. En este ejemplo, la variable Botrytis resulta significativa, y también los subconjuntos formados por esa variable y cualquier otra. Esto significa que la variable Botrytis toma valores muy distintos en función del tratamiento que se haya aplicado.

3.4.2. Datos "enzimas"

Este conjunto de datos consiste en 218 paciences con enfermedades hepáticas, basado en un trabajo de Plomteux [19]. Se consideran cuatro enfermedades:

1. Hepatitis vírica aguda (AVH) (n1 = 57 pacientes)
2. Hepatitis crónica persistente (PCH) (n2 = 44)
3. Hepatitis crónica agresiva (ACH) (n3 = 40)
4. Cirrosis posnecrótica (PNC) (n4 = 77)

El AVH se diagnostica con signos biológicos y clínicos. PCH, ACH y PNC se diagnostican con laparoscopia y biopsia. Los casos correspondientes a cada grupo son del 1 al 57, del 58 al 101, del 102 al 141 y del 142 al 218, respectivamente.
El objetivo de este estudio es obtener un diagnosis diferencial de las cuatro enfermedades hepáticas consideradas mediante un perfil enzimático. Para formar este perfil enzimático, se escogen tres variables respuesta (enzimas):

- X_1 : Aspartato aminotransferasa (ASP)
- X_2 : Alanina aminotransferasa (ALA)
- X_3 : Glutamato deshidrogenasa (GLU)

Todas están expresadas en unidades internacionales por litro. Las variables observadas han sido transformadas mediante logaritmos.

Para la obtención del diagnosis diferencial, se plantea, previamente, la siguiente cuestión: ¿el perfil enzimático definido por las tres variables tiene un comportamiento diferente entre los grupos de pacientes?

Los primeros datos son los siguientes:

	ASP	ALA	GLU	DIS
1	5.46	6.37	2.30	DIS1-AVH
2	3.56	5.55	3.00	DIS1-AVH
3	4.08	5.50	3.00	DIS1-AVH
4	4.52	4.79	1.79	DIS1-AVH
5	4.60	5.86	2.56	DIS1-AVH
6	4.47	5.94	1.95	DIS1-AVH
7	5.31	6.06	2.71	DIS1-AVH
8	5.34	6.29	2.64	DIS1-AVH
9	4.65	6.14	1.39	DIS1-AVH
10	4.55	6.09	2.56	DIS1-AVH

Para resolver la anterior cuestion, se aplicará previamente el test de Mardia para comprobar si los datos cumplen la hipótesis de normalidad multivariante. Para ello, aplicamos una función para el test de Mardia previamente cargada en R, eliminando la columna correspondiente al tipo de enfermedad:

```
mardiaTest(datos1)
```

Obtenemos los p-valores **5.78466e-06** y **0.2891351** correspondientes a los tests del coeficiente de simetría y al coeficiente de kurtosis, respectivamente. Por lo tanto, no podemos aceptar la normalidad multivariante de los datos. Al no poder aplicar técnicas paramétricas, para resolver la cuestión anterior sobre si el perfil enzimático definido por las tres variables tiene un comportamiento diferente entre los grupos de pacientes debemos aplicar de nuevo las funciones del apartado anterior.

En primer lugar, se aplica la función `nonpartest`:

```
nonpartest(ASP|ALA|GLU ~ DIS,datos,permreps=1000)
```

Con esta función obtenemos, en primer lugar, los diagramas:

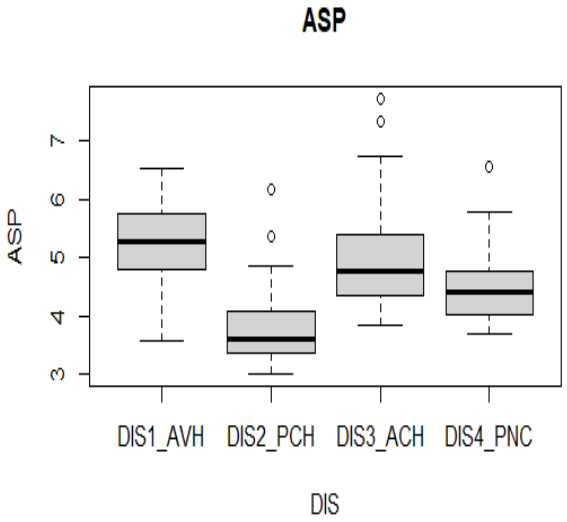

Figura 3.5: ASP frente a enfermedad

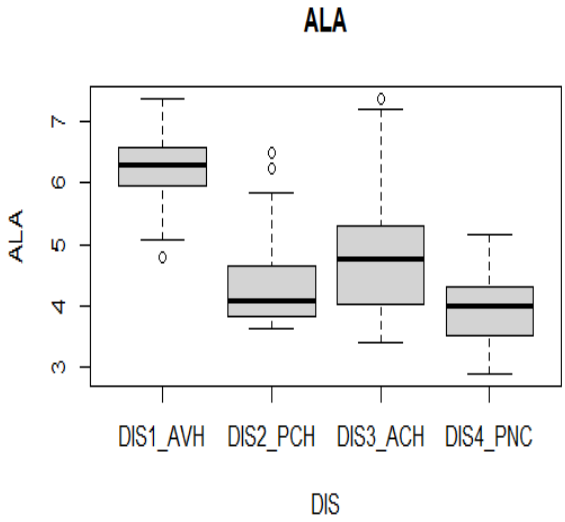

Figura 3.6: ALA frente a enfermedad

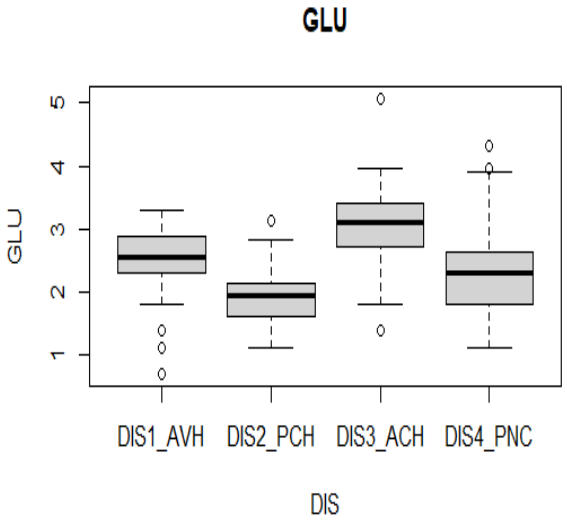

Figura 3.7: GLU frente a enfermedad

En estos gráficos se muestran los datos de cada variable respuesta frente al tipo de enfermedad, señalando las medianas y los valores atípicos. Se puede sospechar que el perfil enzimático será distinto en las distintas enfermedades pues, si nos fijamos por

ejemplo en la enzima "ALA", se puede apreciar que toma valores muy altos en el caso de la enfermedad "AVH", mientras que en las demás toma valores más bajos.

En segundo lugar se muestra el valor de los estadísticos correspondientes a los tests no paramétricos y sus grados de libertad correspondientes, así como los p-valores de cada estadístico correspondientes a los tests no paramétricos y los tests de permutación.

```
$results
                Test Statistic    df1       df2  p-value  PT p-value
ANOVA type Test          51.239  5.210  347.7052        0           0
LHT Test                 66.667  9.000  329.9743        0           0
BNP Test                 46.180  9.182  651.9535        0           0
Wilks Lambda             60.109  9.000  516.1029        0           0
```

Como se puede observar, todos los p-valores de los tests realizados son 0, por lo tanto se rechaza en todos los casos la hipótesis nula, es decir, la igualdad del perfil enzimático definido en las cuatro enfermedades.

Por último, devuelve los efectos relativos:

```
$releffects
             ASP      ALA      GLU
DIS1-AVH  0.72541  0.84239  0.57657
DIS2-PCH  0.20043  0.39778  0.26559
DIS3-ACH  0.60808  0.52678  0.74438
DIS4-PNC  0.44817  0.29105  0.45032
```

Esta tabla de datos se muestra las probabilidades de cada enzima en frente a cada enfermedad. Por ejemplo, la probabilidad de que una persona escogida aleatoriamente del grupo "DIS1-AVH" tenga un porcentaje de "ALA" mayor que cualquier otra persona de todos los grupos (incluyendo las del grupo "DIS1-AVH") es de 0.84239. Podemos apreciar que las probabilidades correspondientes al grupo "DIS2-PCH" son mucho más bajas que en los demás grupos, sobre todo respecto al grupo "DIS1-AVH". Esto también indica que el perfil enzimático será distinto en las diferentes enfermedades.

A continuación aplicamos la función `ssnonpartest`:
`ssnonpartest(ASP|ALA|GLU ~ DIS,datos,test=c(1,0,0,0),alpha=.05, factors.and.variables=TRUE).`

Esta función devuelve lo siguiente:
Al haber seleccionado `factors.and.variables`, en primer lugar lleva a cabo el algoritmo de subconjuntos descrito en la Sección 2.6.2 para analizar los conjuntos de factores (grupos de pacientes) que son significativos, es decir, que muestran diferencias entre ellos :

```
The ANOVA type statistic will be used in the following test
The Global Hypothesis is significant, subset algorithm will continue
Performing the Subset Algorithm based on Factor levels
The Hypothesis of equality between factor levels DIS1-AVH DIS2-PCH DIS3-ACH
DIS4-PNC is rejected
The Hypothesis of equality between factor levels DIS2-PCH DIS3-ACH DIS4-PNC
is rejected
The Hypothesis of equality between factor levels DIS1-AVH DIS3-ACH DIS4-PNC
is rejected
The Hypothesis of equality between factor levels DIS1-AVH DIS2-PCH DIS4-PNC
is rejected
The Hypothesis of equality between factor levels DIS1-AVH DIS2-PCH DIS3-ACH
is rejected
The Hypothesis of equality between factor levels DIS3-ACH DIS4-PNC is rejected
The Hypothesis of equality between factor levels DIS2-PCH DIS4-PNC is rejected
The Hypothesis of equality between factor levels DIS2-PCH DIS3-ACH is rejected
The Hypothesis of equality between factor levels DIS1-AVH DIS4-PNC is rejected
The Hypothesis of equality between factor levels DIS1-AVH DIS3-ACH is rejected
The Hypothesis of equality between factor levels DIS1-AVH DIS2-PCH is rejected
All appropriate subsets using factor levels have been checked using a closed
multiple testing procedure, which controls the maximum overall type I error rate
at alpha= 0.05
```

La interpretación de estos resultados es la siguiente: en la primera línea señala el

tipo de test que hemos escogido, en este caso el tipo ANOVA. La segunda línea indica que la hipótesis global es significativa, lo que señala que hay resultados significativos usando las tres variables respuesta y los cuatro grupos de pacientes. Al obtener este resultado, se continua con el algoritmo para determinar qué variables son significativas y entre qué grupos se encuentran diferencias.

En primer lugar se lleva a cabo el algoritmo de subconjuntos para los factores. En el primer paso se rechaza la igualdad de los cuatro factores. A continuación se llevan a cabo los cuatro tests posibles de igualdad entre tres grupos de pacientes, los cuales también se rechazan. Siguiendo con el algoritmo, se realizan los correspondientes tests de igualdad entre dos grupos de pacientes, los cuales se rechazan de nuevo. Por lo tanto, el resultado de este algoritmo señala que se encuentran diferencias significativas en el comportamiento del perfil enzimático en los cuatro grupos de pacientes.

En segundo lugar, lleva a cabo el mismo algoritmo, en este caso para analizar qué variables respuesta son significativas:

```
Performing the Subset Algorithm based on Response Variables
The Hypothesis of equality using response variables ASP ALA GLU is rejected
The Hypothesis of equality using response variables ALA GLU is rejected
The Hypothesis of equality using response variables ASP GLU is rejected
The Hypothesis of equality using response variables ASP ALA is rejected
The Hypothesis of equality using response variables GLU is rejected
The Hypothesis of equality using response variables ALA is rejected
The Hypothesis of equality using response variables ASP is rejected
All appropriate subsets using response variables have been checked using a
multiple testing procedure, which controls the maximum overall type I error rate
at alpha= 0.05
```

La interpretación de estos resultados es la siguiente: en primer lugar, se rechaza la igualdad entre las tres variables respuesta, lo que indica que las variables no toman los mismos valores en los distintos grupos de pacientes. Siguiendo con el algoritmo, se llevan a cabo los tres tests de igualdad entre dos variables respuesta, los cuales se rechazan. En este caso, se puede hacer el test con una sola variable porque, si este se rechaza, indica que la variable es significativa. En este caso, las tres variables respuesta son significativas al rechazarse los tres tests con una sola variable. Esto significa que las

tres variables toman valores muy distintos dependiendo el grupo de pacientes en el que nos encontremos, es decir, dependiendo de la enfermedad que padezcan. Esto refuerza el resultado del algoritmo anterior, que ya mostraba que el comportamiento del perfil enzimático definido por las tres variables respuesta sería distinto dependiendo del grupo de pacientes en el que se encontrara.

Bibliografía

[1] Bathke,A.C.,Harrar, S.W.,Madden, L.V.(2008). "How to compare small multivariate samples using nonparametric tests".*Computational Statistics and Data Analysis*. **52**,4951-4965.

[2] Box, G.E.P. (1949). "A General Distribution Theory for a Class of Likelihood Criteria". *Biometrika*. **36** (3–4): 317–346. doi:10.1093/biomet/36.3-4.317.

[3] Buckley, J., Hyland, T., Gumaelius, L., Seery, N., Pears, A. (2021). "Exploring the Prototypical Definitions of Intelligent Engineers Held by Irish and Swedish Higher Education Engineering Students."*Psychological Reports*. https://doi.org/10.1177/00332941211000667

[4] Burchett,W.,Ellis, A.(2017)."Nonparametric Comparison of Multivariante Samples".

[5] Burchett,W.,Ellis, A.,Harrar, S.W.,(2017)."Nonparametric Inference for Multivariate Data: The R Package npmv" *Journal of Statistical Software*, **76**(4),1-18.

[6] Brunner, E., Dette, H., Munk, A., (1997). "Box-type approximations in nonparametric factorial designs". *J. Amer. Statist. Assoc.* **92**, 1494–1502.

[7] Cuadras,C.M.(2014). *Nuevos métodos de Análisis Multivariante*. CMC Editions, Barcelona.

[8] Diesburg, K.M., Sullivan, S.M.P. Manning, D.W.P. (2021) "Consequences of a terrestrial insect invader on stream-riparian food webs of the central Appalachians, USA."*Biol Invasions*, **23**, 1263–1284. https://doi.org/10.1007/s10530-020-02435-x

[9] Harrar, S.,Bathke A.(2008). "Nonparametric methods for unbalanced multivariate data and many factor levels". *Journal of Multivariate Analysis* **99**, 1635-1664.

[10] Horst LE, Locke J, Krause CR, McMahon RW, Madden LV, Hoitink HAJ (2005). "Suppression of Botrytis Blight of Begonia by Trichoderma Hamatum 382 in Peat and Compost- Amended Potting Mixes." *Plant Disease*, **89**(11), 1195–1200. doi:10.1094/pd-89-1195.

[11] Johnson,R.A.,Wichern D.W.,(2007) *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Nueva Jersey.

[12] Liu,C.,Bathke,A.C.,Harrar S.W.,(2011). "A nonparametric version of Wilks' lambda-Asymptotic results and small sample approximations"*Statistics and Probability Letters*, **81**, 1502-1506.

[13] Mardia, K.V. (1970). "Measures of multivariate skewness and kurtosis with applications". *Biometrika*, **57**, 519–530.

[14] McKeon, J.J., (1974). "F approximations to the distribution of Hotelling's T_0^2. *Biometrika*,**61**(2), 381–383.

[15] Muller, K.E., (1998). "A new F approximation for the Pillai-Bartlett trace under H_0. *J. Comp. Graph. Statist.* **7**(1), 131–137.

[16] Munzel, U., Brunner, E., (2000a). "Nonparametric methods in multivariate factorial designs". *J. Statist. Plann. Inference* **88** (1), 117–132.

[17] Munzel, U., Brunner, E., (2000b). "Nonparametric tests in the unbalanced multivariate one-way design". *Biom. J.* **42** (7), 837–854; (2001). *Biom. J.* **43** (6), 791–792 (Erratum).

[18] Peña,D.(2002).*Análisis de Datos Multivariantes*. McGraw-Hill, Madrid.

[19] Plomteux, G. (1980) "Multivariate analysis of an enzyme profile for the differential diagnosis of viral hepatitis". *Clinical Chemistry*, **26**, 1897–1899.

[20] Porras,J.C.,(2016)."Comparación de pruebas de normalidad multivariada" *Anales Científicos* **77**(2), 141-146.

[21] Puri, M.L., Sen, P.K., (1966). "On a class of multivariate multisample rank-order tests". *Sankhya*, **28**, 353–376.

[22] Sánchez,C.A. (2010). "Contraste de la normalidad multivariante"

[23] Senker, W., Stefanits, H., Gmeiner, M., Trutschnig, W., Radl, C., Gruber, A. (2021). "The influence of smoking in minimally invasive spinal fusion surgery." *Open medicine*, **16(1)**, 198–206. https://doi.org/10.1515/med-2021-0223

[24] Srivastava, M.S., Fujikoshi, Y., (2006). "Multivariate analysis of variance with fewer observations than the dimension". *J. Multivariate. Anal.* **97**, 1927–1940.

[25] Thompson, G.L., (1990). "Asymptotic distribution of rank statistics under dependencies with multivariate application". *J. Multivariate. Anal.* **33**, 183–211