Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Industriales

Estudio de simulación para una mejor comprensión de MANOVA frente a ANOVA

TRABAJO FIN DE GRADO

Grado en Ingeniería en Tecnologías Industriales

Tutor: José Manuel Mira McWilliams

Roberto Delgado Ferrezuelo

Madrid, febrero de 2022

AGRADECIMIENTOS

Agradecer a mi tutor Pepe, que me ha facilitado las cosas desde el primer momento, ayudándome no solo en el ámbito académico sino en el personal. Sin él no habría sido capaz de afrontar toda la presión de combinar el curso académico con el proyecto.

A mis compañeros y amigos que me han acompañado durante este tiempo y en especial a Santi, que desde el primer día ha estado conmigo, dando fuerza y ánimos y para mí es lo más valioso que me llevo de esta etapa.

A mi familia por inculcarme los valores que me hacen ser quien soy y darlo todo por verme feliz.

RESUMEN

Introducción y objetivos

Las técnicas del análisis de la varianza están presentes en multitud de campos de investigación como pueden ser la psicología, botánica, marketing, medicina… Estas técnicas permiten a los investigadores analizar la influencia de uno o varios factores sobre una serie de variables respuesta.

En este proyecto se analizarán dos herramientas para el análisis de la varianza con procedimientos muy similares, al ser una la extensión de la otra. De lo que se trata es de adquirir un conocimiento más profundo de ambas herramientas y analizar su comportamiento bajo determinadas condiciones impuestas para ver que técnica es la que mejor resultado proporciona en cada caso.

Diseños experimentales

Los diseños experimentales son considerados como una rama de la estadística aplicada en la que, mediante la manipulación de unas variables, que reciben el nombre de variables independientes, se mide el efecto que tienen sobre las variables de estudio o variable dependiente.

El estadístico Ronald Aylmer Fisher (1890-1962) es considerado como el pionero de los diseños experimentales. Una de sus mayores contribuciones ha sido la definición de la varianza y su relación con la media poblacional, Fisher propuso que la varianza es una buena medida de la variabilidad interna dentro de una población. Esta definición es el fundamento de las herramientas con las que se trabajará en este proyecto.

En 1919 Fisher comienza a trabajar como estadístico en Rothamsted, una de las instituciones basadas en la investigación agrícola más antiguas del mundo, donde permaneció durante 14 años (desde los 29 a los 43). Durante su estancia en la estación Fisher fue encargado de analizar conjuntos de datos de experimentos en el campo de la agricultura recopilados durante un largo periodo de tiempo. Mientras analizaba estos datos es cuando se dio cuenta de la cantidad de errores que contenían y de la dificultad de darle uso a muchos de los datos, principalmente por el mal diseño de los experimentos. En este punto Fisher comienza a pensar en cómo debería llevarse a cabo el proceso de recolección de datos.

En base a sus experimentos en el campo de la agricultura Fisher publicó su libro “The Design of Experiments” en 1935. Previo a la publicación de su libro, la forma de analizar la influencia de un tratamiento sobre un conjunto muestral (en su caso los experimentos solían ser con plantas), consistía en someter a un grupo al tratamiento a analizar y al otro no. A medida que crecían la cantidad de tratamientos, el número de experimentos crecía de forma exponencial. Fisher sugirió diseñar experimentos en los que se someta a distintos tratamientos de forma simultánea a un mismo grupo de datos. Para ello desarrolló las herramientas estadísticas necesarias para poder analizar la contribución de cada uno de los tratamientos en el resultado.

A raíz de su publicación es cuando aparecen los diseños factoriales, experimentos en los que se emplean dos o más factores (variables independientes), cada uno de ellos formado por distintos niveles.

En este contexto es donde las herramientas del análisis univariante de la varianza (ANOVA) y análisis multivariante de la varianza (MANOVA) permiten a los investigadores llevar a cabo sus estudios de una forma mucho más eficiente.

Metodología

En este proyecto se aplican técnicas de simulación estocástica, como es la de Monte Carlo, para adquirir una mejor comprensión de las herramientas de ANOVA y MANOVA. El estudio se desarrolla en un contexto estocástico, en el sentido de que los modelos de relación entrada-salida no son deterministas, ya que tienen un término de error aleatorio que recoge de forma agregada la influencia de todas las variables sobre las respuestas y que no se incluyen de forma explícita en el modelo.

La ventaja de la simulación sobre datos reales es que se conocen los valores de los parámetros (α_i, β_j, $\alpha\beta_{ij}, \sigma^2$) ya que se han impuesto sus valores para generar las muestras de datos, de forma que Monte Carlo permite complementar los resultados obtenidos en las expresiones analíticas.

El grueso de este estudio es la técnica multivariante MANOVA, ya que el número de variables que intervienen en este análisis es superior al del ANOVA y su comportamiento, por tanto, es bastante más complejo. Se ha visto en las simulaciones como sus resultados están ligados a multitud de factores y como a medida que se varían estos se producen unas diferencias notables en los resultados entregados por el análisis.

Para ello, se han separado los experimentos en dos, en el primero se ha llevado a cabo un análisis más sencillo para poder analizar el comportamiento de MANOVA en un entorno más controlado. Con los resultados obtenidos en el primer experimento se ha tratado de extender las conclusiones a un caso más complejo.

Los parámetros con los que se han trabajado han sido cuatro, los tamaños del efecto de las variables dependientes, las correlaciones, las varianzas y los tamaños muestrales. Se han fijado unos valores de estos en los experimentos computacionales que difieren ligeramente con los que se simulan los modelos estocásticos.

Conclusiones

En el primero de los experimentos se ha empleado una sola variable independiente y en el otro dos, de los resultados se han obtenido principalmente las siguientes conclusiones:

- La potencia del MANOVA dependerá de una combinación del valor de la correlación entre las variables dependientes con su tamaño del efecto (α_i, β_j, $\alpha\beta_{ij}$), además de la varianza residual y del tamaño muestral.
- Cuanto más grande sea la cantidad de datos con la que se trabaja, mayor será la evidencia de que existen diferencias entre los niveles formados en las variables independientes.

Palabras clave

MANOVA, ANOVA, p-valor, estadístico F, error tipo I, error tipo II, factor, variable dependiente, residuos, correlación, distancia de Mahalanobis, varianza, tamaño del efecto.

Códigos UNESCO

1209.09, 1209.05, 1203.26, 1203.23.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. SOFTWARE Y MÉTODO EMPLEADO PARA EL ESTUDIO DE SIMULACIÓN

1.1. RStudio

RStudio es una interfaz del lenguaje R destinada a cálculos estadísticos, es un software gratuito y de código abierto.

A nivel corporativo se han consolidado como una organización de beneficio público, que cumple con los requisitos del B-Lab, la cual es una organización sin ánimo de lucro que certifica a aquellas compañías que cumplen con unos estándares de transparencia, responsabilidad y sostenibilidad con objetivo de aportar valor a la sociedad, su CEO es Joseph J. Allaire.

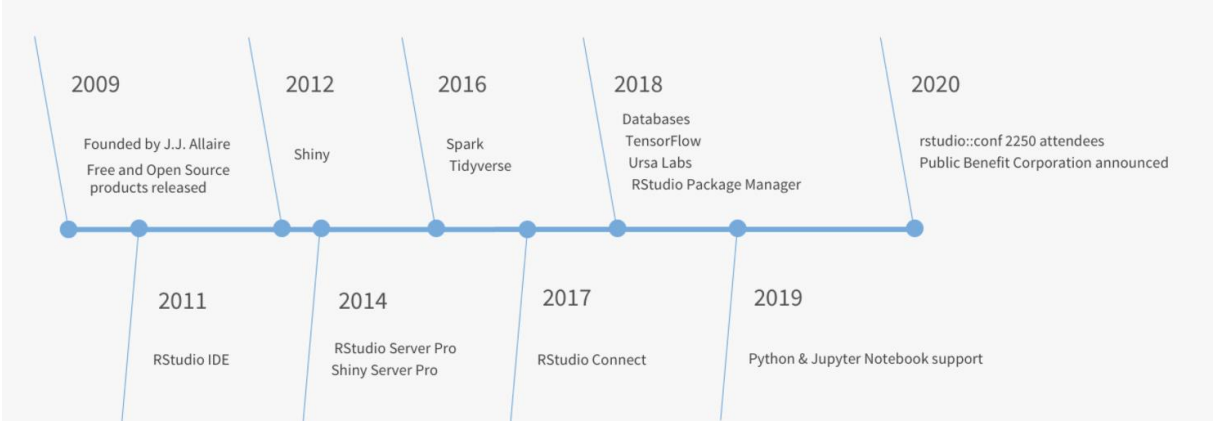


Figura 1.1. Línea de tiempo de RStudio. Adaptado de *RStudio´s Timeline*, RStudio (https://www.rstudio.com/about/)

En la Figura 1.1 se muestra la línea de tiempo de la compañía desarrolladora del software, desde que se fundó por J.J. Allaire hasta que recibió en 2020 el certificado B que le acredita como organización de beneficio público.

Su uso está muy extendido, siendo utilizado por millones de usuarios y cada vez está cobrando mayor importancia en el mundo empresarial, siendo una herramienta adoptada por más de 1200 compañías actualmente.

RStudio ha sido la herramienta empleada en este estudio de simulación para hacer todos los cálculos y gran parte de las figuras explicativas.

1.2. Métodos de Monte Carlo

Existen distintas formas de clasificar las matemáticas, entre otras muchas, se pueden clasificar como experimentales y teóricas, mientras que las teóricas deducen resultados de postulados, las experimentales infieren conclusiones de observaciones. Los métodos de Monte Carlo se corresponden con la rama de las matemáticas experimentales, han tenido un gran uso en campos como la física nuclear.

Los problemas resueltos mediante estos métodos son de dos tipos, probabilistas y deterministas. En los probabilistas se escogen números aleatorios que reflejen el comportamiento del sistema real sujeto a estudio y se infieren conclusiones a partir del comportamiento observado de los

números. En los deterministas se parte de una ecuación que representa la estructura que subyace un problema real, la cual se obtiene por medio de las matemáticas teóricas, que se resuelve, de igual forma, mediante generación de números aleatorios con estos métodos, al no poder encontrarse una solución con los medios teóricos, para ello se deriva el problema original a otro de características similares y cuya solución numérica es idéntica a la del original. Hay que tener en cuenta que las conclusiones que se puedan extraer del Monte Carlo no son exactas al provenir de una serie de números aleatorios, no obstante, existen distintas formas de reducir esta incertidumbre.

Ya desde la segunda mitad del siglo XIX se llevaron a cabo algunas variantes tempranas de estos métodos. En uno de los experimentos, un grupo de personas dedujo el valor de π tirando agujas al azar en un tablero con líneas rectas equidistantes entre sí, viendo el número de intersecciones de las agujas con las rectas.

Según Hammersley y Handscomb (1964), el uso de los métodos de Monte Carlo como herramienta de investigación comienza con el desarrollo de la bomba atómica en la segunda guerra mundial, donde se utilizó para resolver problemas probabilísticos relacionados con la difusión de neutrones en material fisible. Siendo más tarde aplicado a problemas deterministas por Stanislaw Ulam, John von Neumann y Enrico Fermi, es entonces, en 1948, cuando Ulam y Fermi junto con Nicholas Metropolis estiman los autovalores de la ecuación de Schrödinger aplicando estos métodos.

En este proyecto, para llevar a cabo las simulaciones se ha empleado el método de Monte Carlo en la generación de números aleatorios de acuerdo a una distribución normal de media y desviación típica dadas. Existen diferentes formas de llevar a cabo la generación de números, en este caso, se ha trabajado con la función *rnorm()* de RStudio. Esta función trabaja empleando el método de la transformada inversa, que hace uso de la generación de números aleatorios para le creación de variables con una función de distribución determinada. Es uno de los pilares del método de Monte Carlo.

1.2.1. Método de la transformada inversa

Supóngase que se desea generar de forma aleatoria una variable que tiene una función de densidad de la forma $f(x)$ con x en el rango $-\infty < x < \infty$, se sabe entonces que su función de distribución en un punto x_0 se calcula como la integral hasta ese punto de la función de densidad, $F(x_0) = \int_{-\infty}^{x_0} f(x)dx$, cuyo valor está comprendido en el rango $[0,1]$.

Si se genera un valor y uniformemente distribuido en el intervalo (0,1), es posible encontrar un punto x tal que

$$y = F(x),$$

cuyo valor se puede expresar a partir de la inversa de la función de distribución

$$x = F^{-1}(y).$$

En la Figura 1.2 se muestra el procedimiento descrito para el caso de una función de distribución continua.

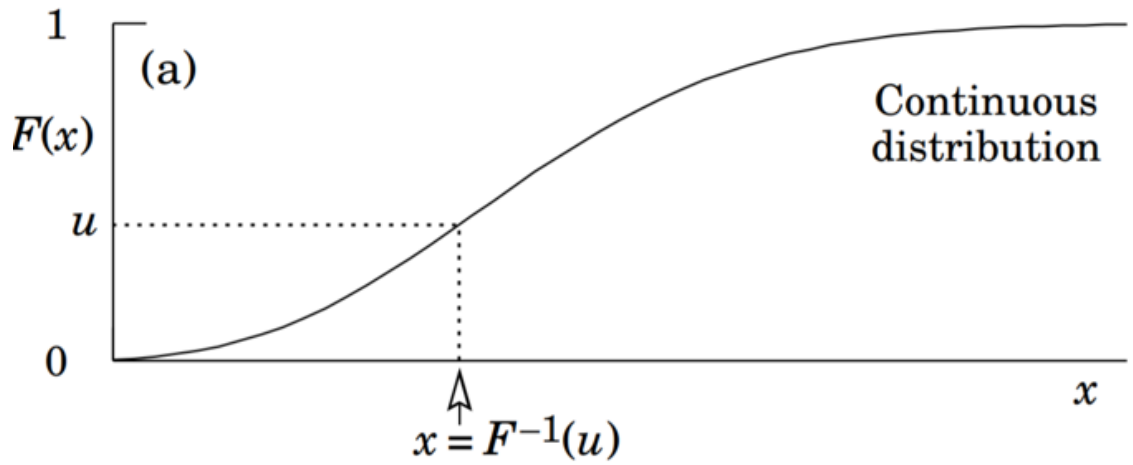


Figura 1.2. Generación de números aleatorios por el método de la transformada inversa. P.A. Zyla et al., 2020, particle data group (https://pdg.lbl.gov/2020/reviews/rpp2020-rev-monte-carlo-techniques.pdf)

Este es el método que se utiliza por defecto en RStudio para generar datos con una distribución normal. Al no existir una expresión analítica directa para la inversa de la función de distribución normal se hace una aproximación utilizando polinomios de grado moderado. En la Figura 1.3 se muestran las curvas empleadas para la generación del número aleatorio empleadas por RStudio.

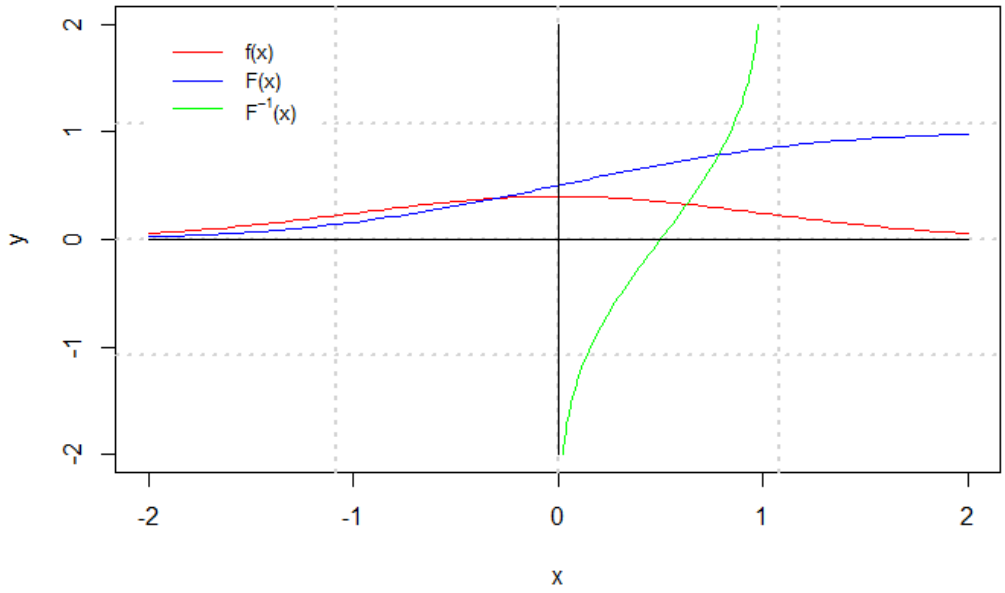


Figura 1.3. Método de la transformada inversa usando *rnorm()* en RStudio

2. ANOVA

Dadas las similitudes entre las herramientas de ANOVA y MANOVA, se comenzará dando una descripción detallada de lo que es un ANOVA y de cómo se modelan los distintos parámetros y variables aleatorias que intervienen en este, para luego extender las explicaciones al caso de MANOVA sin necesidad de repetir todas aquellas características que tienen en común ambos análisis.

2.1. Introducción

ANOVA es una herramienta estadística empleada para analizar la igualdad de medias entre grupos de datos, haciendo uso de la varianza. En este método se hace una partición de la varianza del conjunto de datos en dos, la varianza debida a las diferencias que existen entre los grupos y la varianza interna propia de cada grupo o varianza residual. En base a las proporciones de cada tipo de varianza se aceptará o no la hipótesis de igualdad de medias. Además, el análisis entrega un nivel de significación que indica el nivel de seguridad con el que se puede se puede aceptar o rechazar la hipótesis planteada.

Este concepto fue introducido por Fisher y su uso se generalizó a raíz de su publicación "Statistical Methods for Research Workers" en 1925, la cual se convirtió en la guía de referencia para multitud de investigadores de todo el mundo. Su publicación pasó por 13 ediciones y una última que terminó de ser escrita por uno de sus compañeros y publicada en 1970, tras su fallecimiento.

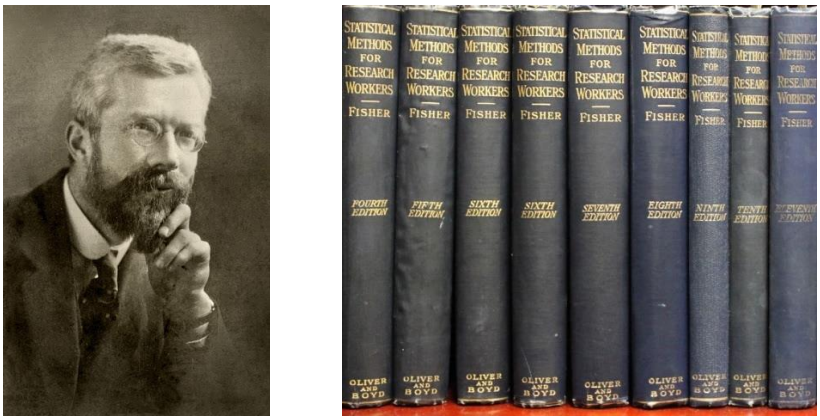


Figura 2.1. Ronald Fisher y las distintas ediciones de su publicación "Statistical Methods for Research Workers". Adaptado de *R.A. FISHER*, por Indian Statistical Institute, Google Arts & Culture (https://artsandculture.google.com/asset/r-a-fisher/iAFkKK47btRx5Q)

En el análisis se emplean una única variable cuantitativa, la variable dependiente y una o varias cualitativas, las variables independientes o factores, que estarán formadas por diferentes niveles, se trata de analizar si existen diferencias entre los distintos niveles de las variables independientes en base a los resultados recogidos en la dependiente.

Dependiendo del número de variables independientes que se usen el ANOVA se formulará de una forma diferente, si únicamente se emplea una variable, recibe el nombre de ANOVA de un factor, en caso de usar dos, de dos factores. Se comentará el modelo que subyace estos dos, siendo el resto de análisis multi factores (dos o más variables independientes) una extensión del caso de dos factores.

En el caso particular de que en el ANOVA de un factor la variable independiente esté formada únicamente por dos niveles, el ANOVA es equiparable a otro test estadístico conocido como el test de la t de Student.

2.2. ANOVA de un factor

En cuanto a los distintos tipos de ANOVAs, como se comentaba, el más sencillo es el ANOVA de un factor, compuesto por una sola variable independiente con diferentes niveles.

2.2.1. Modelo

En un ANOVA de un factor, la variable dependiente se puede descomponer en dos partes, una totalmente predecible y el término aleatorio, que representa la parte no predecible

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \to N(0, \sigma^2).$$

El primer componente es la parte predecible que representa la media de cada grupo y el segundo es el término de error, con una distribución normal de media 0 y desviación típica σ.

Al no disponerse de estos parámetros, se utilizan los estimadores calculados por el método de máxima verosimilitud. En el caso de la media se utilizará la media muestral

$$\mu_i \to \bar{y}_i = \sum_{j=1}^{n_i} \frac{y_{ij}}{n_i},$$

el subíndice *i* hace referencia al grupo y el *j* a la observación, n_i es el número de observaciones en el grupo *i* de la variable independiente, el término de error se estimará como la diferencia entre el valor real y la estimación del valor previsto y recibirá el nombre de residuo

$$\varepsilon_{ij} \to e_{ij} = y_{ij} - \bar{y}_i,$$

$$\sum_{j=1}^{n_i} e_{ij} = 0, \quad \forall i.$$

La varianza de los residuos se estima a partir de la varianza residual que se calcula como

$$\sigma^2 \to \hat{s}_R{}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} e^2{}_{ij}}{n-k},$$

siendo n el número total de observaciones y k el número de niveles formados en la variable independiente. En la Figura 2.2 se muestra de forma gráfica cada una de las componentes mencionadas para un conjunto de datos generados de forma aleatoria, asignando una media diferente a tres niveles, formados cada uno por cinco observaciones. El término $\bar{y}_1 - \bar{y}$ representa el efecto de uno de los niveles del factor sobre la muestra, se puede representar como α_i y debe cumplir que

$$\sum_{i=1}^{n_i} \widehat{\alpha_i} = 0$$

ya que la variación se produce con respecto al valor de la media global y, por tanto, la suma del conjunto de los efectos tiene que dar un resultado nulo.

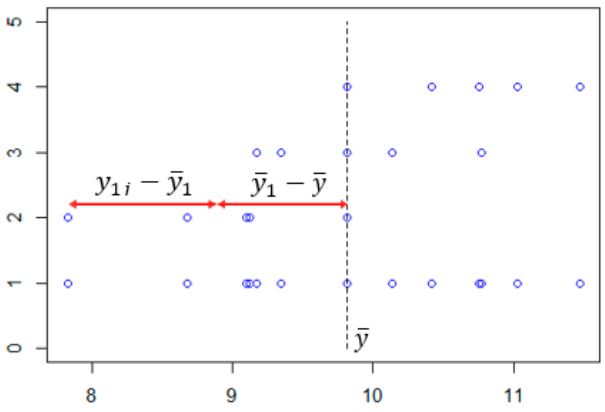


Figura 2.2. Descomposición del modelo ANOVA

Las hipótesis a contrastar en el análisis son dos:

$$H_0: \mu_1 = \mu_2 \cdots = \mu_k$$

$$H_1: alguna\ media\ es\ distinta\ de\ otra$$

Como se ha explicado, ANOVA contrasta estas hipótesis haciendo una descomposición de la varianza en dos, la varianza interna y la varianza entre grupos. Para ello, se parte del modelo ya definido

$$y_{ij} = \bar{y}_i + y_{ij} - \bar{y}_i,$$

restando la media global en ambos lados de la igualdad

$$y_{ij} - \bar{y} = \bar{y}_i - \bar{y} + y_{ij} - \bar{y}_i,$$

elevando todo al cuadrado, haciendo el sumatorio para $i,\ j$ y recordando que $\sum_{j=1}^{n_i} e_{ij} = 0, \quad \forall i$, y que, por tanto, $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_i - \bar{y})(y_{ij} - \bar{y}_i) = 0$ ya que el primer miembro del producto es constante para todos los valores de *j*, se llega a

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2.$$

A estos términos se les conoce como:

- $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$, suma total de cuadrados: Representa la variabilidad total. Tiene un total de $n - 1$ grados de libertad. Se expresará mediante la abreviatura VT.
- $\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$, suma de cuadrados entre grupos. Es la parte de variabilidad que queda explicada por las diferencias entre grupos asociada al efecto del factor. El número de grados de libertad es igual al número de niveles menos uno $(k - 1)$. Se expresará mediante VE.
- $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$, suma de cuadrados dentro del grupo. Es la variabilidad que queda sin explicar, debida al término de error. Tiene $n - k$ grados de libertad. Se expresará como VNE.

Sin embargo, para el análisis lo que se emplea son las varianzas, resultado de dividir las sumas de cuadrados entre sus grados de libertad

$$\hat{s}_T^2 = \frac{VT}{n-1}, \quad \hat{s}_E^2 = \frac{VE}{k-1}, \quad \hat{s}_R^2 = \frac{VNE}{n-k}.$$

Con estos valores se calcula el estadístico F_0

$$F_0 = \frac{\hat{s}_E^2}{\hat{s}_R^2} \sim F_{k-1,n-k},$$

válido bajo el cumplimiento de la hipótesis nula. Este estadístico sigue una distribución de Fisher-Snedecor, nombrada así por sus dos desarrolladores. Su función de densidad viene dada por

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{x^{\frac{m-2}{2}}}{\left(1+\frac{mx}{n}\right)^{\frac{m+n}{2}}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Γ representa la función gamma, *m* y *n* son los grados de libertad del numerador y denominador respectivamente. En la Figura 2.3 se muestran algunos ejemplos de esta distribución para distintas combinaciones de los grados de libertad.

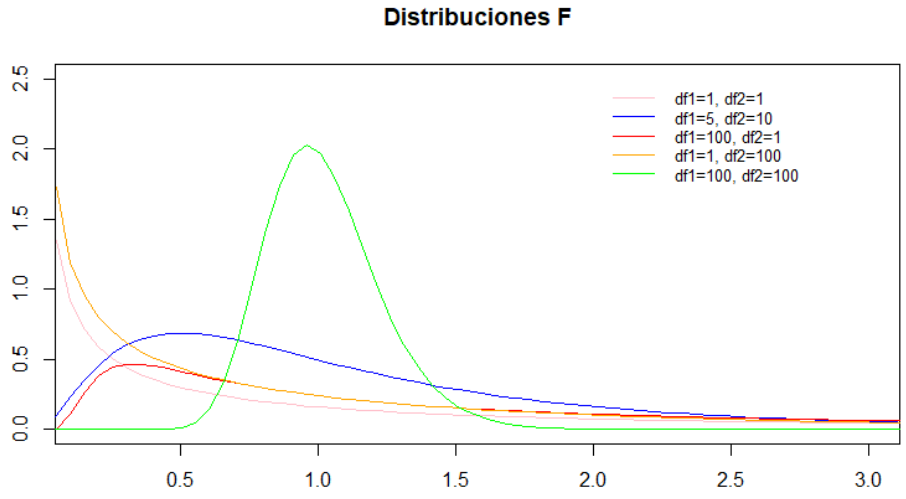


Figura 2.3. Distribución de Fisher-Snedecor para distintos grados de libertad

El valor de F_0 obtenido se compara con un valor F_α, que representa el valor extremo para el que la probabilidad de que F sea superior a este sea α, también conocido como nivel de significación, su valor típico suele ser de 0,05. En la Figura 2.4 se indica de forma gráfica,

$$Pr(F \geq F_\alpha) = \alpha.$$

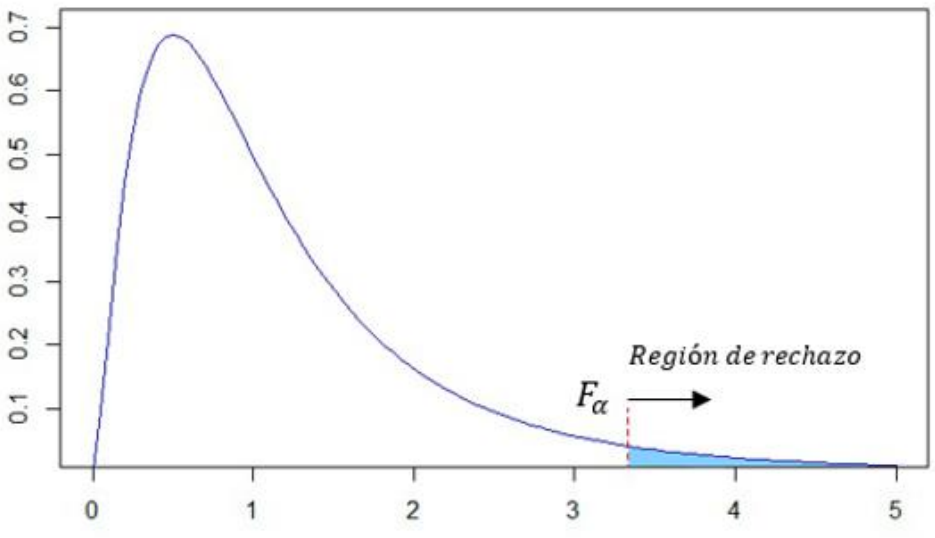


Figura 2.4. Valor de $F_{5,10,\alpha}$ para un $\alpha = 0{,}05$

En la práctica, el resultado que devuelve el software suele ser el p-valor, que representa la probabilidad de que el estadístico sea igual o mayor que un cierto valor, tomando la hipótesis nula como cierta,

$$p - valor = P(valor\ obtenido\ igual\ o\ más\ extremo\ que\ el\ observado |\ H_0).$$

Es una medida de la significación estadística del dato observado, para poder descartar aquellos que son debidos al azar de los que no. Si el p-valor es superior al del nivel de significación α se tomará la hipótesis nula como cierta, si no, se rechazará.

Al nivel de significación α, también se le conoce como probabilidad de error tipo I, es la probabilidad de rechazar la hipótesis nula sabiendo que esta se cumple. Por otro lado, existe la probabilidad de error tipo II, o error tipo β, que representa la probabilidad de aceptar la hipótesis nula cuando esta no se cumple. Al complementario de este último se le conoce como potencia del ANOVA y es uno de los principales parámetros que se ha tenido en cuenta a la hora de realizar este proyecto. En la Tabla 2.1 se recogen los distintos casos posibles.

	H0 es cierta	**H1 es cierta**
Se acepta H0	$1 - \alpha$	Error tipo β
Se acepta H1	Error tipo α	$1 - \beta$

Tabla 2.1. Errores en función de la hipótesis aceptada

Con toda esta información se construiría la tabla de ANOVA que es lo que devuelve el software una vez introducido el código. La forma de la tabla para un ANOVA de un factor con *k* niveles en la variable independiente sería la mostrada en la Tabla 2.2.

	Grados de libertad	Suma de cuadrados	Varianzas	Valor de F	p-valor
Tratamiento	$k-1$	$\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2$	$\frac{\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2}{k-1}$	$\frac{(n-k)\sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}$	$P(F \geq F_0)$
Residual	$n-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$	$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2}{n-k}$		

Tabla 2.2. Tabla ANOVA de un factor

2.2.2. Hipótesis

Es importante que para que los resultados que se obtengan del análisis se puedan considerar como válidos se cumpla con una serie de supuestos.

1.1.1.1. Normalidad

En primer lugar, la variable dependiente tiene que seguir una distribución normal en cada uno de los niveles formados en las variables independientes. En aquellos casos en que la falta de normalidad no sea excesivamente grande, se puede decir que, a mayores tamaños muestrales, menor será el impacto que tenga la falta de normalidad en los resultados.

Una forma rápida de comprobarlo es mediante el uso de los gráficos quantile-quantile (Q-Q). Aunque esta no es una prueba formal, da buenos resultados y permite rechazar algunos experimentos a falta del cumplimiento de la hipótesis.

Para elaborar el gráfico se ordenan los residuos de menor a mayor y se calcula las probabilidades asociadas a los cuantiles como

$$\frac{i-0.5}{n},$$

donde *i* es la posición que ocupa el residuo en su ordenación, siendo 1 para el más pequeño, *n* es el total de residuos. A continuación, haciendo uso de la inversa de la función de densidad normal estándar se obtiene el valor del cuantil que corresponde a esa probabilidad y este se multiplica por la estimación de la desviación típica del conjunto de datos, este valor obtenido sería el valor teórico y es con el que se compara el del residuo original. Es decir,

$$eteórico_i = \Phi^{-1}\left(\frac{i-0.5}{n}\right)\hat{s}_R, \quad i = 1,2\ldots,n.$$

Se representa en una gráfica, en el eje de abscisas los valores teóricos y en el de ordenadas el experimental, si la hipótesis se cumple, la nube de puntos debe representar una línea recta.

Esto se puede elaborar de forma práctica haciendo uso de la herramienta RStudio. Programando un código sencillo se puede crear un conjunto de datos aleatorio, que siga una distribución normal, por ejemplo, con una varianza de 10 y una media que tendrá que ser 0, ya que se está hablando de residuos. Mediante el comando *rnorm()*, se crea un conjunto de 100 datos de acuerdo a esta características, que representará el valor de los residuos.

Orden	Residuo	Probabilidad	Cuantil	Cuantil teórico
i	e_{ij}	(i-0,5)/n	ϕ^{-1}(Prob)	Φ^{-1}(Prob)* $\hat{s}_R$
1	-23,091	0,005	-2,576	-23,871
2	-19,666	0,015	-2,170	-20,111
3	-16,867	0,025	-1,960	-18,164
4	-15,488	0,035	-1,812	-16,792
5	-12,654	0,045	-1,695	-15,712
6	-12,65	0,055	-1,598	-14,811
7	-12,207	0,065	-1,514	-14,032
8	-11,381	0,075	-1,440	-13,341
9	-11,231	0,085	-1,372	-12,717
10	-10,718	0,095	-1,311	-12,146
11	-10,678	0,105	-1,254	-11,617
12	-10,264	0,115	-1,200	-11,124
13	-10,26	0,125	-1,150	-10,661
14	-10,186	0,135	-1,103	-10,222
15	-7,289	0,145	-1,058	-9,806

Tabla 2.3. Obtención de los cuantiles teóricos para la elaboración del gráfico Q-Q

En la Tabla 2.3 se muestra el proceso seguido para el cálculo de los primeros 15 cuantiles a partir de los datos generados. Si se representan en una misma gráfica los datos enmarcados en azul junto con la recta de ajuste por mínimos cuadrados se llega al resultado mostrado en la Figura 2.5.

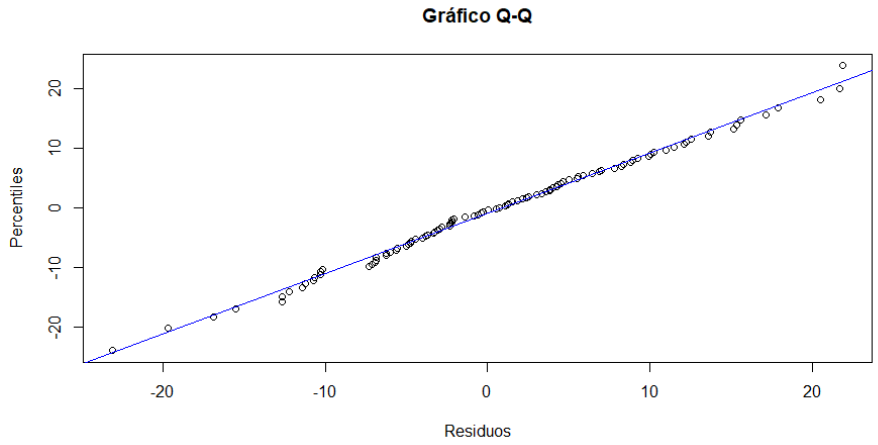


Figura 2.5. Gráfico Q-Q para datos que sigan una distribución normal

Se observa que la nube de puntos se ajusta bastante bien a la recta, resultado que era el esperado ya que los residuos se han generado utilizando una función normal. Si en cambio, se generan los residuos mediante una función de distribución gamma, el resultado sería el mostrado en la Figura 2.6.

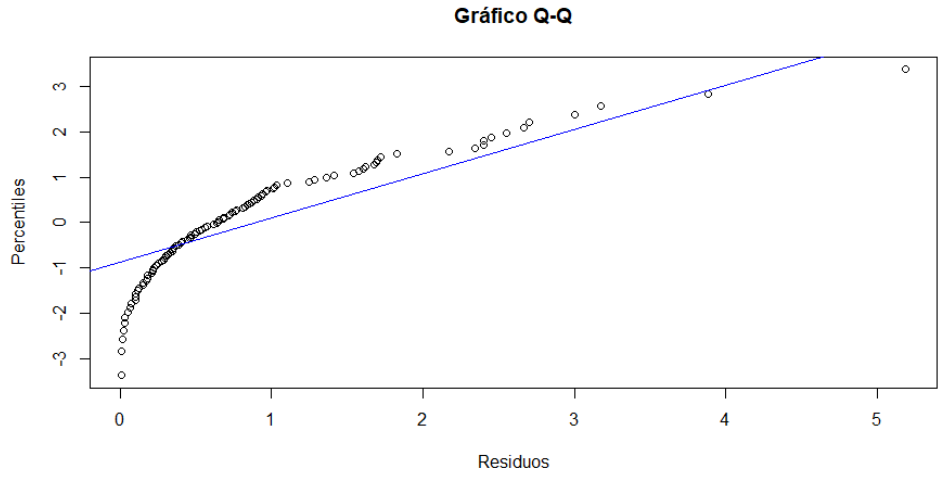


Figura 2.6. Gráfico Q-Q para datos que no siguen una distribución normal

En este caso se observa que la nube de puntos se ajusta bastante mal a una línea recta y, por tanto, se podría rechazar la hipótesis de normalidad.

Conviene examinar los residuos extremos, ya que pueden provenir de observaciones atípicas y estas pueden introducir sesgo en el análisis, siendo aconsejable eliminarlas.

Como pruebas formales, se puede llevar a cabo, por ejemplo, la de Kolmogorov-Smirnov, en estas simulaciones se ha usado la de Shapiro-Wilks, mediante el comando *shapiro.test()* de RStudio.

1.1.1.2. Homocedasticidad

Este supuesto considera que la varianza tiene que ser constante a lo largo de todos los nivles. De nuevo, en los casos en los que la falta de homocedasticidad no sea excesivamente grande, a mayor cantidad de datos por nivel, menor será la influencia de la falta de cumplimiento en esta hipótesis. Es útil, además, disponer de un diseño equilibrado, un número de observaciones por grupo suficientemente grande y un cociente de la desviación estándar entre los diferentes nivles que en algunos libros se recomienda que sea inferior a 3.

Para ver si se cumple la hipótesis se puede volver a utilizar una prueba no formal que permita descartar aquellos modelos que no la cumplan de forma rápida. En este caso, se representará, en una gráfica, en el eje de abscisas la media de los valores correspondientes a cada tratamiento y en el de ordenadas el valor de los residuos de cada grupo. Se aceptará la hipótesis en caso de que la distribución de los residuos en todos los niveles se encuentre comprendida en una franja horizontal. Si la dispersión varía mucho en función de la media de los niveles, se rechazará y se considerará que los datos son heterocedásticos.

Volviendo a hacer uso de RStudio, se crean dos conjuntos de datos, simulando cada uno un conjunto muestral que ha sido sometido a un tratamiento. En el primer conjunto de datos se establece la misma varianza para todos los niveles (en este caso se han establecido nueve niveles) y en el segundo se incrementa a medida que aumenta la media de cada nivel. Los resultados son los mostrados en la Figura 2.7.

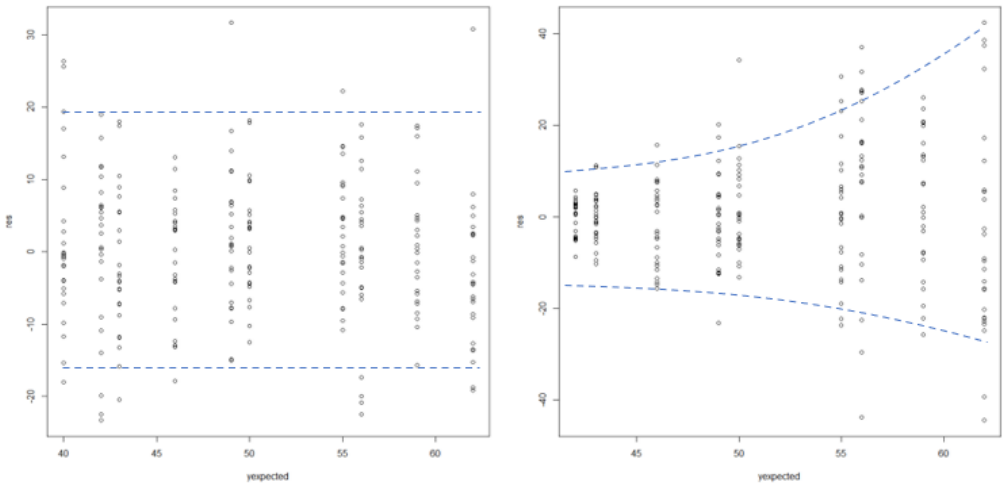


Figura 2.7. Residuos frente a valores previstos

En la gráfica izquierda se aprecia como los residuos están comprendidos en una franja horizontal, al haber sido generados utilizando la misma varianza en todos los niveles, mientras que en la de la derecha la dispersión aumenta conforme la media crece.

En caso de que no se cumpla la hipótesis, se puede recurrir a las transformaciones de Box-Cox, esto consiste en hacer un cambio en la variable dependiente de la siguiente forma

$$z(p) = \begin{cases} \dfrac{y^p - 1}{p} & si\ p \neq 0 \\ log(y) & si\ p = 0 \end{cases}$$

Para que la transformación sea válida, las medidas que se estén tomando de la variable dependiente deben ser valores positivos. En la Figura 2.8 se representan las curvas de transformación más típicas, que son aquellas comprendidas entre valores de p de -3 a 3.

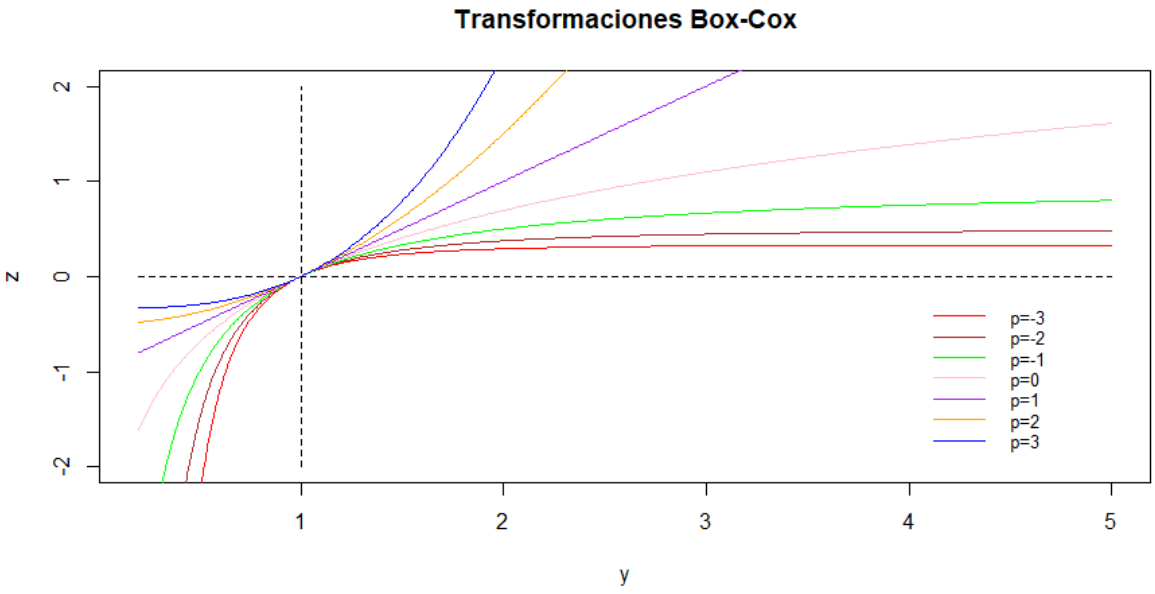


Figura 2.8. Transformaciones de Box-Cox

Si a los datos anteriores que seguían una distribución heterocedástica se les aplica esta transformación con $p = -1$, el resultado sería el mostrado en la Figura 2.9.

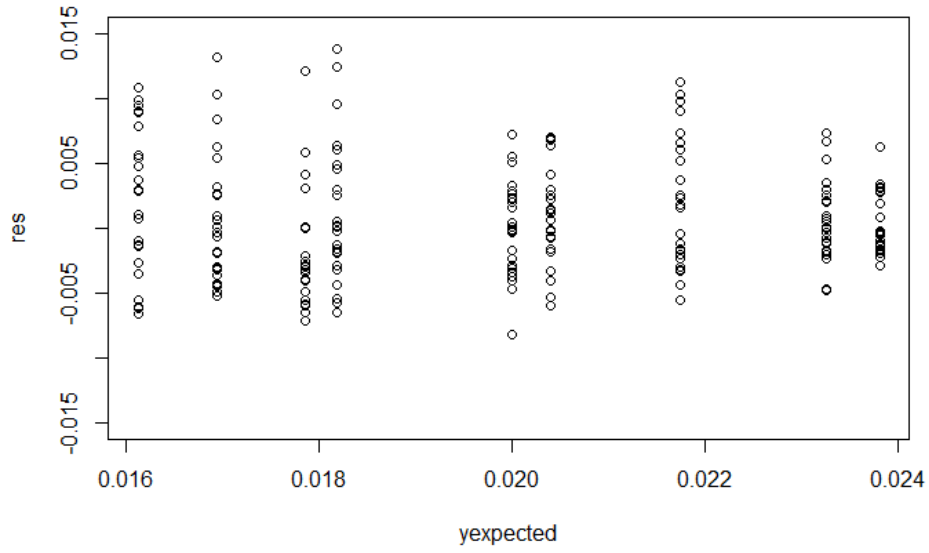


Figura 2.9. Datos heterocedásticos transformados

Como se observa, la dispersión de la nube de puntos en los distintos niveles se reduce de manera significativa.

Existen diferentes pruebas formales para obtener resultados más precisos, entre algunas de ellas se encuentran la de Bartlett, Brown-Forsyth o la de Fligner-Killeen. Destacar la prueba de Levene que es la que se ha usado en este trabajo para rechazar aquellas simulaciones con falta de homocedasticidad, se puede llevar a cabo en RStudio mediante el comando *leveneTest()* del paquete *heplots*.

1.1.1.3. Independencia

Es considerada como la hipótesis más importante, hay que evitar que las observaciones dependan las unas de las otras. No existe ningún test que permita afirmar el cumplimiento de esta hipótesis, por tanto, dependerá principalmente de que la recogida de datos en la investigación se haga de forma adecuada. Comprobadas todas estas hipótesis se puede pasar a extraer las conclusiones del análisis.

1.1.2. Comparaciones post-hoc

Una vez llevado a cabo el análisis, si se ha rechazado la hipótesis nula, habría que llevar a cabo una serie de comparaciones dos a dos para encontrar aquellas medias que se diferencian entre sí. Esto se puede hacer mediante la prueba de la t de Student, que como se comentó previamente es equivalente a realizar un ANOVA de un factor con dos niveles en el factor.

Supóngase el caso que tras realizar un ANOVA se desea comparar si las medias de los grupos 1 y 2 son distintas entre sí, para ello se dispone de las puntuaciones de la variable dependiente en ambos grupos,

$$y_{1j} \rightarrow N(\mu_1, \sigma^2), \quad j = 1,2,\ldots,n_1,$$

$$y_{2j} \rightarrow N(\mu_2, \sigma^2), \quad j = 1,2,\ldots,n_2.$$

Tomando la media muestral de cada uno de los grupos, se sabe que su distribución toma la siguiente forma

$$\bar{y}_i \rightarrow N(\mu_i, \frac{\sigma^2}{n_i}),$$

por tanto, la diferencia entre las medias de ambas observaciones sigue una distribución

$$\bar{y}_1 - \bar{y}_2 \rightarrow N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}),$$

que si se le resta la media y se divide entre la desviación típica se puede expresar como una distribución normal estándar

$$\frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow N(0,1).$$

Si se sustituye la desviación típica por su estimador, que ya se ha visto como es su cálculo, el término anterior sigue una distribución conocida como la distribución de la t de Student

$$\frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{\hat{s}_R\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow t_{n_1+n_2-2}.$$

Si la diferencia entre las medias es nula, se obtendría el siguiente valor del estadístico t_0

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\hat{s}_R\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

cuyo valor debe estar comprendido para un determinado nivel de significación en el intervalo $[-t_{\frac{\alpha}{2}},\ t_{\frac{\alpha}{2}}]$.

A diferencia de la distribución F, este es un contraste bilateral. Si el valor absoluto del estadístico es superior al valor $t_{\frac{\alpha}{2}}$, se rechaza la hipótesis nula, asumiendo que la diferencia entre ambas medias es estadísticamente significativa.

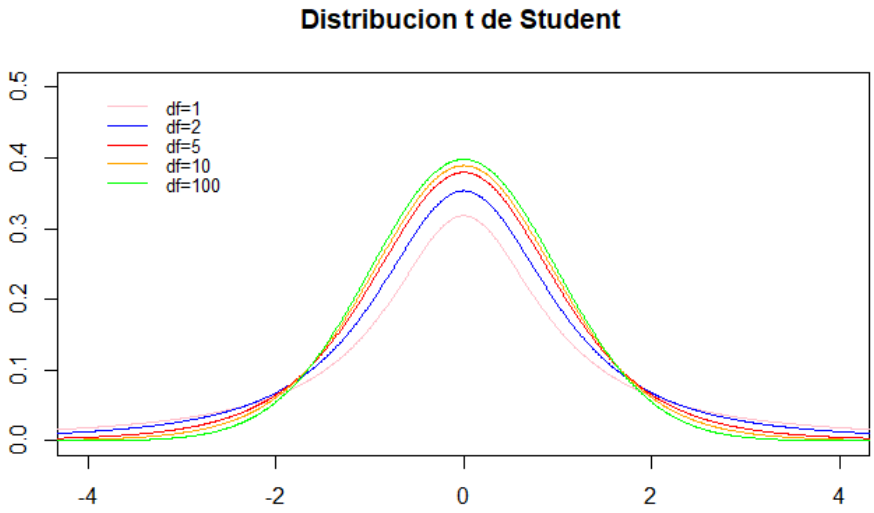


Figura 2.10. Distribución t de Student en función de los grados de libertad

En la Figura 2.10 se han recogido algunos ejemplos de esta distribución variando los grados de libertad. Como se observa, la función de densidad es simétrica con respecto al 0 en las abscisas y de ahí que se emplee el valor absoluto del estadístico para el contraste de igualdad de medias. El nivel de significación para el cálculo del estadístico *t* a partir del que se rechazará la hipótesis nula, se parte en dos, ya que el valor obtenido puede ser extremo en sentidos negativos o positivos.

Al estar trabajando con múltiples comparaciones aparece el fénomeno conocido como la probabilidad de error conjunto, que consiste en una inflación del probabilidad de error tipo I, que crece de la forma

$$1 - (1 - \alpha)^n,$$

siendo *n* el número de comparaciones llevadas a cabo. Para controlar este fénomeno se emplean diferentes ajustes, como el de la diferencia significativa propuesto por Tukey (Tukey HSD), o el de Holm–Bonferroni. En este caso se empleará la corrección de Bonferroni que consiste en dividir el nivel de significación original entre el número total de comparaciones. Es un método muy conservador, no obstante, de esta forma se consigue que la tasa de error conjunto sea inferior a un cierto umbral,

$$\alpha_{corregido} = \frac{\alpha_{original}}{n} \rightarrow 1 - (1 - \alpha_{corregido})^n \leq \alpha_{original}.$$

En la Figura 2.11 se da una idea gráfica del significado de los dos tipos de errores para el caso de un contraste bilateral.

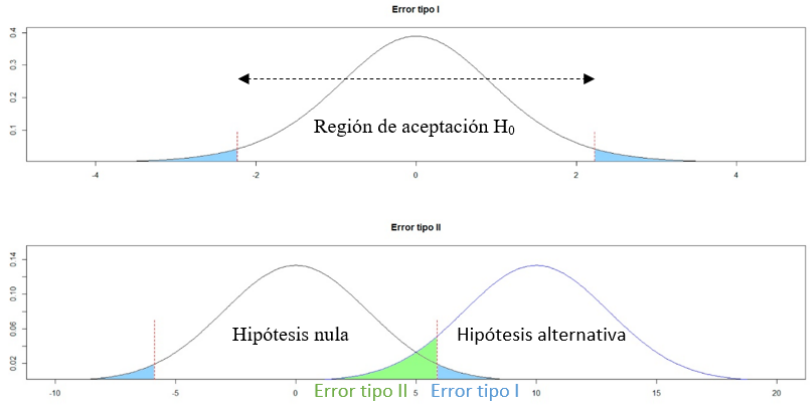


Figura 2.11. Probabilidad de error tipo I y probabilidad de error tipo II para contraste bilateral

Aunque este proyecto es un estudio de simulación, con la finalidad de adquirir una mejor compresión del funcionamiento de las herramientas que intervienen se dará un ejemplo utilizando datos reales al final de cada una de las explicaciones.

1.1.3. Ejemplo

Supóngase que se esta haciendo un estudio para ver la influencia que tiene el número de cilindros en el consumo de un vehículo. Para ello se dispone de tres grupos de vehículos, de 4, 6 y 8 cilindros y se tiene de cada uno de ellos el consumo expresado en millas por galón (equivalente a unos 235 litros cada 100 kilómetros). Hay datos de un total de 32 vehículos. Se muestran los datos referidos a los primeros 6 vehículos.

```
##                    mpg cyl
## Mazda RX4         21.0   6
## Mazda RX4 Wag     21.0   6
## Datsun 710        22.8   4
## Hornet 4 Drive    21.4   6
## Hornet Sportabout 18.7   8
## Valiant           18.1   6
```

En la Figura 2.12 se representa el boxplot de los datos para ver como están distribuidos en los distintos grupos y examinar la presencia de osbervaciones atípicas.

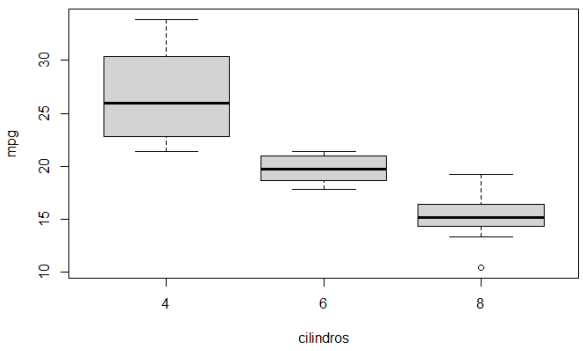


Figura 2.12. Boxplot millas por galón en función del número de cilindros

No se tienen valores atípicos, se aprecia algo de diferencia en la amplitud de las cajas, que sugiere algo de heterocedasticidad y que por tanto habrá que corregir con alguna transformación.

Se comprueban las hipótesis haciendo uso de los gráficos mostrados en la Figura 2.13.

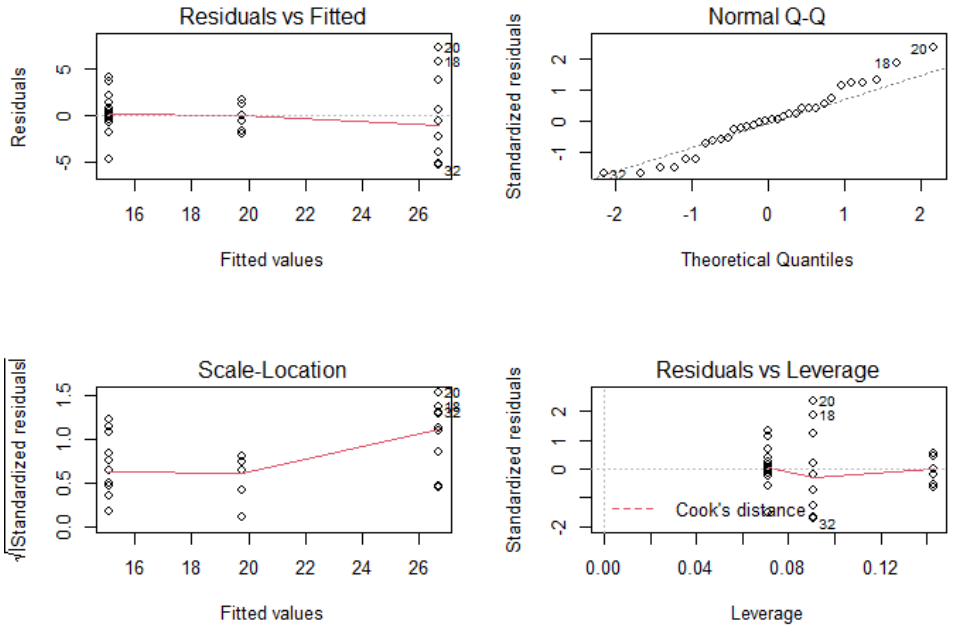


Figura 2.13. Gráficos para comprobar hipótesis de normalidad y homocedasticidad ANOVA de un factor

Se observa falta de normalidad y un cierto grado de heterocedasticidad, no obstante, se comprueba formalmente haciendo uso de las pruebas comentadas.

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2  5.5071 0.00939 **
##       29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  Shapiro-Wilk normality test
##
## data:  mpg
## W = 0.94756, p-value = 0.1229
```

Según los resultados, hay alta probabilidad de que exista heterocedasticidad, pero se puede considerar que los datos siguen aproximadamente una distribución normal. Por tanto se efectúa las transformaciones de Box-Cox, ecnontrandose el mejor resultado para $p = 0$.

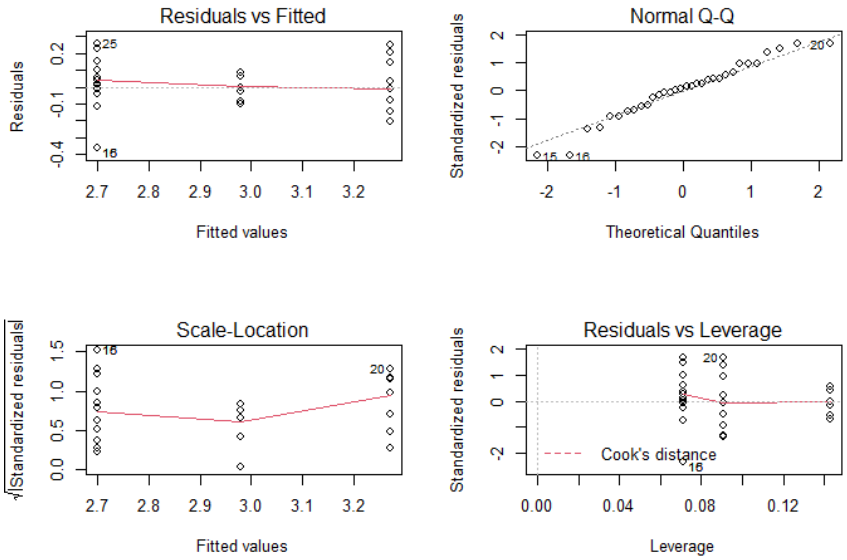


Figura 2.14. Gráficos para comprobar hipótesis de normalidad y homocedasticidad ANOVA de un factor tras la transformación de los datos

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  1.4428 0.2527
##       29
```

```
##  Shapiro-Wilk normality test
##
## data:  log(mpg)
## W = 0.97668, p-value = 0.699
```

Tanto en los gráficos de la Figura 2.14 como en los resultados de los contrastes formales se aprecia una mejora.

A continuación se efectuará el cálculo de la tabla ANOVA para ver si existen diferencias entre los grupos.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cilindros    2 2.0081  1.0040   39.31 5.52e-09 ***
## Residuals   29 0.7407  0.0255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Empleando un nivel de significación de 0,05 se puede rechazar la hipótesis nula, por tanto, faltaría comprobar entre que grupos existen estas diferencias, para lo que se hará una serie de contrastes pareados utilizando la prueba de la t de Student con el ajuste del nivel de significación por el método de Bonferroni.

```
##  Pairwise comparisons using t tests with pooled SD
##
## data:  mpg and cilindros
##
##   4       6
```

```
## 6 0.00036 -
## 8 2.6e-09 0.01246
##
## P value adjustment method: bonferroni
```

Se detectan diferencias significativas entre todos los grupos. Se representará la forma de la distribución que sigue el conjunto de datos y la de cada uno de los grupos usando sus estimadores máximos verosímiles para estudiar cual es la relación entre el número de cilindros y el consumo del vehículo.

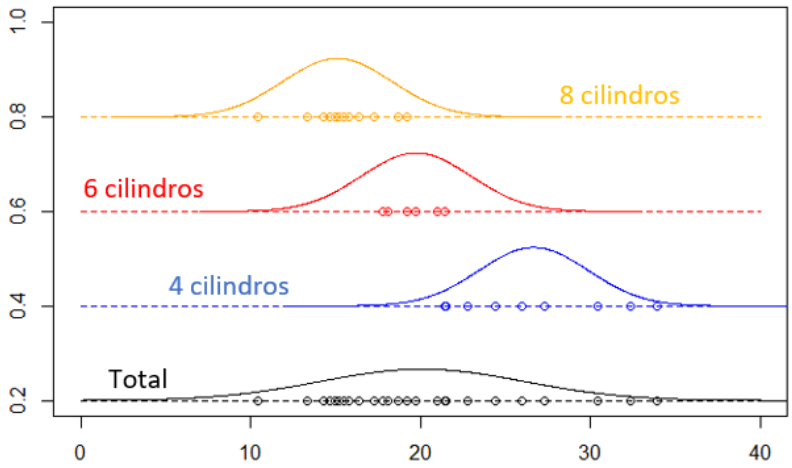


Figura 2.15. Distribuciones del conjunto de observaciones y de cada grupo por separado

Atendiendo a las formas de las distribuciones mostradas en la Figura 2.15 y a los resultados del ANOVA, se puede concluir que existe una cierta probabilidad de que a mayor número de cilindros el consumo del vehículo aumenta.

Fuente de los datos

Henderson and Velleman (1981), Building multiple regression models interactively. Biometrics, 37, 391–411.

1.2. ANOVA de dos factores

Como se ha comentado en la introducción a los diseños de experimentos, una de las mayores aportaciones de Fisher fue la de trabajar con más de un factor simultaneamente, el ANOVA de dos factores es una herramienta muy útil para este tipo de experimentos. De lo que se tratará a continuación es de extender la idea del ANOVA de un factor a un caso un poco más complejo, pasando de tener un único factor a tener dos.

Al igual que en el anterior caso, previo a realizar el análisis conviene comprobar que se cumplen todas las hipótesis y hacer alguna transformación en caso de ser necesario, todo esto mediante las técnicas ya comentadas.

1.2.1. Modelo

En el ANOVA de dos factores, la variable dependiente se vuelve a descomponer en la parte predicible y la aleatoria, sin embargo, ahora se expresa en función de los efectos de los factores y no de las medias de los niveles

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \to N(0, \sigma^2).$$

En esta nueva expresión μ vuelve a ser la media global del conjutno de observaciones, α_i es el efecto del nivel *i* del factor 1 y β_j el efecto del nivel j del factor 2, ε_{ijk} sigue siendo el término aleatorio que sigue una distribución normal de media 0 y desviación típica σ. Al estar trabajando con dos factores aparece un nuevo término, la intreacción entre los niveles de ambos, $\alpha\beta_{ij}$.

La interacción representa el efecto combinado de los dos factores sobre la variable dependiente. Si un conjunto de datos presenta interacción, la influencia que tiene el nivel de un factor en la variable dependiente dependerá del nivel del otro factor.

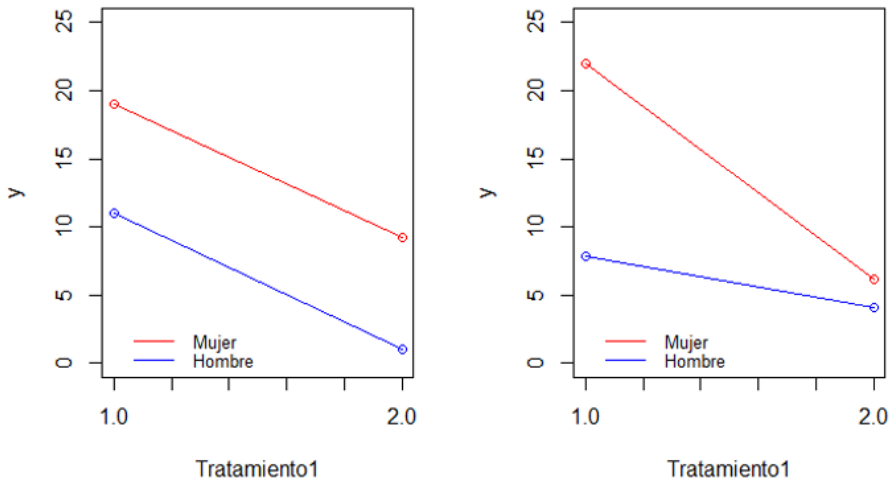


Figura 2.16. Comparación de tratamientos con y sin interacción

Para ilustrar este efecto se crearán dos conjuntos de datos haciendo uso de RStudio, ambos conjuntos con dos factores con dos niveles cada uno, a uno de de estos conjuntos se le introducirá una cierta interacción y al otro no. Supóngase que la variable dependiente está midiendo como varía un indicador del nivel de salud de las personas, resultado de un tratamiento médico. Se estudia en conjunto el efecto del tratamiento y el género del paciente, los resultados para cada conjunto de datos generado serían los mostrados en la Figura 2.16. En azul se muestran los resultados de hombres y en rojo los de mujeres, el grupo Tratamiento1 1, son aquellos pacientes sometidos al tratamiento médico, mientras que los del 2 son los que no.

Se considerará que en este ejemplo el efecto de la varianza residual es suficientemente pequeño como para que el efecto de la interacción en la gráfica derecha sea significativo, mientras que en la izquierda no, enntonces se pude deducir que la gráfica de la izquierda corresponde al conjunto sin interacción, como se observa , las líneas son prácticamente paralelas, lo que indica que independientemente del género, el efecto del tratamiento aumenta la punutación del indicador de forma similar en ambos grupos. En el segundo caso, el efecto del tratamiento parece aumentar de forma mucho más significativa la puntuación en el grupo de las mujeres que en el de los hombres, se podría considerar en este caso que ambos factores tienen una cierta relación de dependencia. Si la varianza residual es muy grande, el planteamiento definido para la interacción podría dejar de ser válido y no ser considerable en ninguno de los dos casos.

El caso descrito es el de una interacción ordinal, en la que ambos grupos varían en la misma dirección cuando se someten al tratamiento, si las direcciones fuesen opuestas se hablaría de

interacción desordinal, es importante tener en cuenta qué tipo de interacción se da a la hora de extraer conclusiones del ANOVA.

De nuevo habrá que estimar cada uno de los parámetros que intervienen en el análisis. Volviendo a utilizar los estimadores calculados por el método de máxima verosimiltud se llega a los siguientes resultados.

$$\mu \to \bar{y}_{\bullet\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\frac{y_{ijk}}{n},$$

$$\alpha_i \to \hat{\alpha}_i = \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet}, \quad \bar{y}_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{k=1}^{m}\frac{y_{ijk}}{mJ},$$

$$\beta_j \to \hat{\beta}_j = \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet}, \quad \bar{y}_{\bullet j\bullet} = \sum_{i=1}^{I}\sum_{k=1}^{m}\frac{y_{ijk}}{mI},$$

$$\alpha\beta_{ij} \to \widehat{\alpha\beta}_{ij} = \bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet}, \quad \bar{y}_{ij\bullet} = \sum_{k=1}^{m}\frac{y_{ijk}}{n},$$

$$u_{ijk} \to e_{ijk} = \bar{y}_{ijk} - (\bar{y}_{\bullet\bullet\bullet} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{\alpha\beta}_{ij}) = \bar{y}_{ijk} - \bar{y}_{ij\bullet},$$

$$\sigma^2 \to \hat{s}^2{}_R = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m} e_{ijk}^2}{IJ(m-1)}.$$

En las anteriores expresiones se ha considerado que el número de observaciones por cada combinación de niveles de los dos factores es constatnte y de valor *m*. La media global se estima a partir de la media poblacional, los efectos de los factores se calculan como la desviación de las medias de los niveles de cada factor con respecto a la estimación de la media global, la interacción se interpreta como la desviación de cada uno de los grupos formados por la combinación de los niveles de ambos factores con respecto a sus efectos individuales, los residuos serán la diferencia entre el valor real y el estimado a partir de la parte predecible del modelo y por último, la varianza se vuelve a expresar en función de la varianza residual. De los estimadores, por la forma en la que están definidos se sabe que

$$\sum_{i=1}^{I}\hat{\alpha}_i = 0, \quad \sum_{j=1}^{J}\hat{\beta}_j = 0,$$

$$\sum_{i=1}^{I}\widehat{\alpha\beta}_{ij} = 0, \ \forall j, \quad \sum_{j=1}^{J}\widehat{\alpha\beta}_{ij} = 0, \qquad \forall i.$$

Además, la estimación del error, esto es, los resiudos , cumplen que

$$\sum_{k=1}^{m} e_{ijk} = 0, \ \forall i, j.$$

En este nuevo caso existen tres hipótesis a contrastar y para ello se necesitará descomponer la varianza de la misma forma que en el ANOVA de un factor, para ello se retoma la expresión de la variable dependiente, pero empleando los estimadores calculados.

$$y_{ijk} = \bar{y}_{\bullet\bullet\bullet} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{\alpha\beta}_{ij} + e_{ijk},$$

$$y_{ijk} - \bar{y}_{\bullet\bullet\bullet} = \hat{\alpha}_i + \hat{\beta}_j + \widehat{\alpha\beta}_{ij} + e_{ijk}.$$

Elevando todo al cuadrado, haciendo el sumatorio para todo *i*, *j*, *k* y teniendo en cuenta que $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\hat{\alpha}_i\hat{\beta}_j = 0$, $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\hat{\alpha}_i\widehat{\alpha\beta}_{ij} = 0$, $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\hat{\beta}_j\widehat{\alpha\beta}_{ij} = 0$, $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\hat{\alpha}_i e_{ijk} = 0$, $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\hat{\beta}_j e_{ijk} = 0$, $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\widehat{\alpha\beta}_{ij} e_{ijk} = 0$.

Ya que en todos los casos uno de los miembros del producto permanece constante para los distintos valores de uno de los índices y la suma del otro vale 0, dadas las condiciones impuestas por la definición del modelo. Se llega entonces a la siguiente expresión,

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\hat{\alpha}_i^{\,2} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\hat{\beta}_j^{\,2} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}\widehat{\alpha\beta}_{ij}^{\,2} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2},$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 = mJ\sum_{i=1}^{I}\hat{\alpha}_i^{\,2} + mI\sum_{j=1}^{J}\hat{\beta}_j^{\,2} + m\sum_{i=1}^{I}\sum_{j=1}^{J}\widehat{\alpha\beta}_{ij}^{\,2} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2}.$$

Las variabilidades se pueden separar en:

- $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}(y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2$, variabilidad total. Tiene $n-1$ grados de libertad y se abrevia como VT.
- $mJ\sum_{i=1}^{I}\hat{\alpha}_i^{\,2}$, variabilidad explicada debida al efector del factor 1. Tiene un total de $I-1$ grados de libertad y se representa por VE(Factor 1).
- $mI\sum_{j=1}^{J}\hat{\beta}_j^{\,2}$, variabilidad explicada debida al efector del factor 2. Tiene un total de $J-1$ grados de libertad y se representa por VE(Factor 2).
- $m\sum_{i=1}^{I}\sum_{j=1}^{J}\widehat{\alpha\beta}_{ij}^{\,2}$, variabilidad explicada debida a la interacción entre los niveles de los factores. Tiene $(I-1)(J-1)$ grados de libertad y se representa por VE(Factor 1×Factor 2).
- $\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2}$, variabilidad no explicada. Tiene $IJ(m-1)$ grados de libertad y se expresa como VNE.

Las tres hipótesis a contrastar son la influencia de cada uno de los factores y si la itneracción entre ambos es significativa, dependiendo de si esta última hipótesis es cierta o no habrá que llevar a cabo el análisis considerando el efecto conjunto de los dos factores o separado.

La tabla ANOVA se construye de la misma forma que en el anterior caso, diviendo las sumas de cuadrados entre sus grados de libertad. En la Tabla 2.4 se muestra su forma para el caso en el que se considere la interacción.

	Grados de libertad	Suma de cuadrados	Varianzas	Valor de F	p-valor
Factor 1	*I-1*	$mJ\sum_{i=1}^{I}\widehat{\alpha}_i^{\,2}$	$\frac{mJ\sum_{i=1}^{I}\widehat{\alpha}_i^{\,2}}{I-1}$	$\frac{IJ(m-1)mJ\sum_{i=1}^{I}\widehat{\alpha}_i^{\,2}}{(I-1)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2}}$	$P(F_1 \geq F_{01})$
Factor 2	*J-1*	$mI\sum_{j=1}^{J}\widehat{\beta}_j^{\,2}$	$\frac{mI\sum_{j=1}^{J}\widehat{\beta}_j^{\,2}}{J-1}$	$\frac{IJ(m-1)mI\sum_{j=1}^{J}\widehat{\beta}_j^{\,2}}{(J-1)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2}}$	$P(F_2 \geq F_{02})$
Factor 1: Factor 2	*(I-1)(J-1)*	$m\sum_{i=1}^{I}\sum_{j=1}^{J}\widehat{\alpha\beta}_{ij}^{\,2}$	$\frac{m\sum_{i=1}^{I}\sum_{j=1}^{J}\widehat{\alpha\beta}_{ij}^{\,2}}{(I-1)(J-1)}$	$\frac{IJ(m-1)m\sum_{i=1}^{I}\sum_{j=1}^{J}\widehat{\alpha\beta}_{ij}^{\,2}}{(I-1)(J-1)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2}}$	$P(F_{12} \geq F_{012})$
Residual	*IJ(m-1)*	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2}$	$\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}e_{ijk}^{\,2}}{IJ(m-1)}$		

Tabla 2.4. Tabla ANOVA de dos factores

1.2.2. Comparaciones post-hoc

Una vez realizado el ANOVA, si se han encontrado diferencias en los grupos habrá que volver a realizar una serie de comparaciones para ver entre qué grupos se da la diferencia. En este caso, se va a emplear otra técnica distinta a la que se ha comentado para el ANOVA de un factor. Se construirá los intervalos de confianza de las medias para un cierto nivel de significación y se analizará si existe solapamiento entre los intervalos de las distintas medias, en caso de existir se concluirá que la diferencia entre ambos grupos no es significativa.

A la hora de realizar las comparaciones hay que distinguir entre aquellos casos en los que exista interacción y los que no.

1.2.2.1. Sin interacción

Si no hay interacción se construirán los intervalos de confianza de las medias de los grupos formados en cada uno de los factores de forma independiente y se estudiará las diferencias en los grupos de los factores que hayan dado un p-valor inferior al nivel de significación en el ANOVA.

El cálculo del intervalo de confianza para las medias de los grupos formados en el factor 1 sería de la forma similar a como se efectuaban las comparaciones en la prueba de la t de Student. Se aproxima la distribución de la media muestral en uno de los grupos del factor a una normal estándar al dividir la diferencia entre la media muestral y la teórica entre la desviación típica.

$$y_{ijk} \to N(\mu, \sigma^2), \quad y_{i\bullet\bullet} \to N(\mu + \alpha_i, \frac{\sigma^2}{mJ}),$$

$$\frac{y_{i\bullet\bullet} - (\mu + \alpha_i)}{\frac{\sigma}{\sqrt{mJ}}} \to N(0,1).$$

Se sustituye en esa nueva expresión la desviación típica por la raíz de la varianza residual, resultando la forma de una distribución *t* de Student con los mismos grados de libertad que los residuos,

$$\frac{y_{i\bullet\bullet} - (\mu + \alpha_i)}{\frac{\hat{s}_R}{\sqrt{mJ}}} \to t_{IJ(m-1)}.$$

Se calcula un intervalo de confianza $1-\alpha$ para la media teórica a partir de la muestral,

$$Pr(-t_{\alpha/2} \leq \frac{y_{i\bullet\bullet} - (\mu + \alpha_i)}{\frac{\hat{s}_R}{\sqrt{mJ}}} \leq t_{\alpha/2}) = 1-\alpha,$$

$$\mu + \alpha_i \in y_{i\bullet\bullet} \pm t_{\alpha/2} \frac{\hat{s}_R}{\sqrt{mJ}}.$$

Todo esto se puede hacer en RStudio mediante la prueba de Tukey, que calcula los intervalos de confianza para la diferencia de medias a partir de los intervalos de ambas medias, utilizando una tasa de error conjunto dada, evitando así el problema visto de la inflación de la probabilidad error de tipo I.

En el caso del factor 2 el cálculo sería análogo al visto para el factor 1, resultando el intervalo

$$\mu + \beta_j \in y_{\bullet j\bullet} \pm t_{\alpha/2} \frac{\hat{s}_R}{\sqrt{mI}}.$$

1.2.2.2. Con interacción

Si se da el caso de que exista interacción hay que estudiar los dos factores juntos y por tanto habrá que construir el intervalo de confianza de la media de todas las combinaciones de los niveles de los factores. El razonamiento que se sigue es el mismo, se parte de la media de uno de los grupos y se estandariza.

$$y_{ijk} \to N(\mu, \sigma^2), \quad \bar{y}_{ij\bullet} \to N(\mu + \alpha_i + \beta_j + \alpha\beta_{ij}, \frac{\sigma^2}{m}),$$

$$\frac{y_{i\bullet\bullet} - (\mu + \alpha_i + \beta_j + \alpha\beta_{ij})}{\frac{\sigma}{\sqrt{m}}} \to N(0,1).$$

Se sustituye la desviación típica por su estimación,

$$\frac{\bar{y}_{ij\bullet} - (\mu + \alpha_i + \beta_j + \alpha\beta_{ij})}{\frac{\hat{s}_R}{\sqrt{m}}} \to t_{IJ(m-1)}.$$

Por último, se calcula el intervalo de confianza $1-\alpha$,

$$\mu + \alpha_i + \beta_j + \alpha\beta_{ij} \in \bar{y}_{ij\bullet} \pm t_{\alpha/2} \frac{\hat{s}_R}{\sqrt{m}}.$$

Esto se llevará a cabo en RStudio utilizando la función *interIC()* desarrollada por la Unidad Docente de Estadística de la Escuela Técnica Superior de Ingenieros Industriales en la Universidad Politécnica de Madrid.

Para concluir con este apartado de ANOVA se llevará a cabo otro experimento distinto haciendo uso de dos factores.

1.2.3. Ejemplo

El departamento de marketing de una empresa está interesado en saber cómo evoluciona el número de consultas que reciben en base a sus espacios publicitarios en distintas secciones de periódicos y al día de la semana, para de esta forma poder definir una mejor campaña de publicidad. Para ello se han recogido los datos de las consultas recibidas cada día de la semana a raíz de sus publicaciones a lo largo de un mes, las secciones en las que publicitaron fueron la de noticias, deportes y negocios.

De igual manera que antes, se comprobará si el modelo cumple con las hipótesis antes de interpretar los resultados del ANOVA.

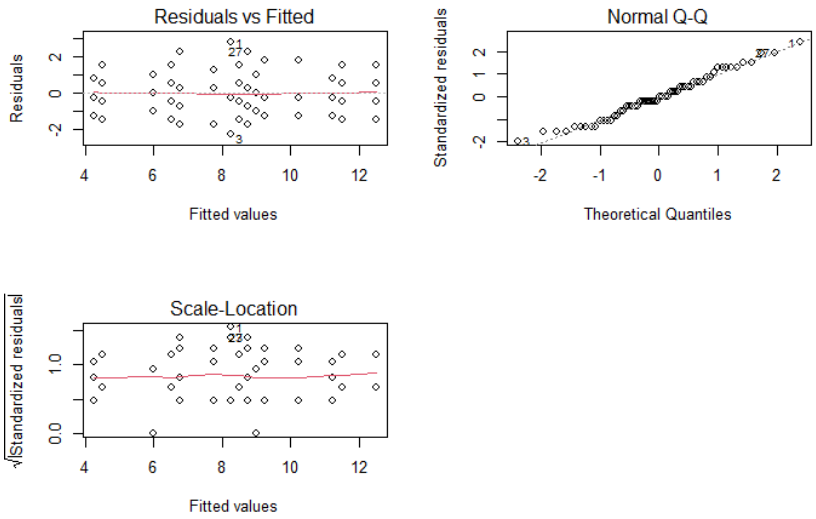


Figura 2.17. Gráficos para comprobar hipótesis de normalidad y homocedasticidad ANOVA de dos factores

```
Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 14  0.3768  0.975
##       45
```

```
##  Shapiro-Wilk normality test
##
## data:  dat$Inquiries
## W = 0.97647, p-value = 0.2977
```

Tanto de los contrastes formales como de las gráficas se puede concluir que no es necesario transformar los datos. Por tanto, se pasa a construir la tabla ANOVA.

```
## Analysis of Variance Table
##
## Response: Inquiries
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Section      2  53.733  26.867 15.3038 8.503e-06 ***
## Day          4 146.833  36.708 20.9098 8.518e-10 ***
## Section:Day  8 135.767  16.971  9.6669 1.125e-07 ***
```

```
## Residuals   45  79.000   1.756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tanto los efectos de los factores por separado como su interacción son influyentes. Quedaría comprobar entre qué grupos se dan estas diferencias, para ello se hará uso de la herramienta *interIC()*.

```
##     Section       Day media  2.5% 97.5%
## 1  Business    Friday  9.00  7.67 10.33
## 2  Business    Monday 11.50 10.17 12.83
## 3  Business  Thrusday  7.75  6.42  9.08
## 4  Business   Tuesday  8.75  7.42 10.08
## 5  Business Wednesday  8.50  7.17  9.83
## 6      News    Friday 12.50 11.17 13.83
## 7      News    Monday  8.25  6.92  9.58
## 8      News  Thrusday  4.25  2.92  5.58
## 9      News   Tuesday 10.25  8.92 11.58
## 10     News Wednesday  9.25  7.92 10.58
## 11   Sports    Friday 11.25  9.92 12.58
## 12   Sports    Monday  4.50  3.17  5.83
## 13   Sports  Thrusday  6.00  4.67  7.33
## 14   Sports   Tuesday  6.50  5.17  7.83
## 15   Sports Wednesday  6.75  5.42  8.08
```

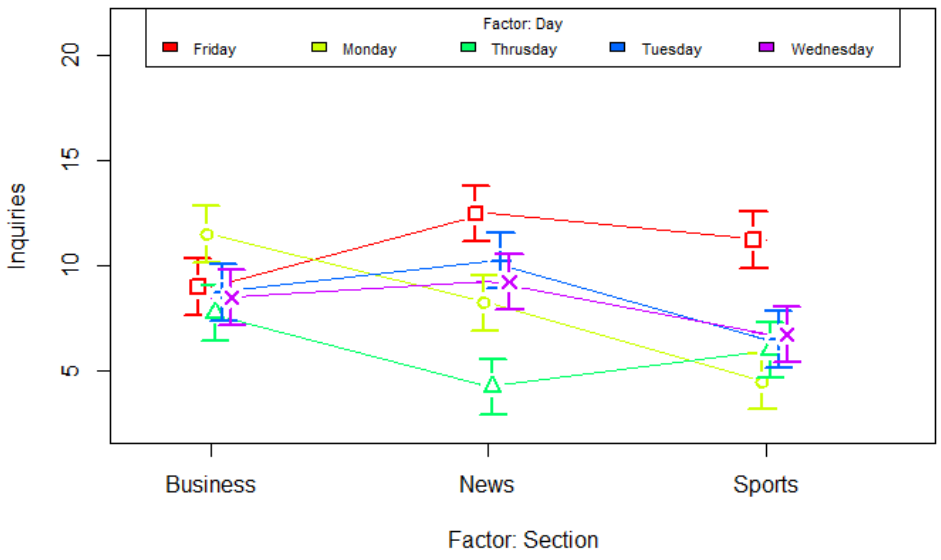


Figura 2.18. Gráfico de intervalos de confianza para interacción entre los factores

De la Figura 2.18 y de los resultados numéricos se desprende que las mayores diferencias se dan en la sección de deportes, en concreto entre los viernes y lunes y en la sección de noticias, en los jueves y viernes, parece que en la sección de negocios las consultas recibidas no varían significativamente con el día de la semana.

Por tanto, el departamento de marketing debería tener en cuenta que los lunes es un mal día para publicitarse en la sección de deportes y quizás le convenga más hacerlo en la de negocios ese día y esperar al viernes para hacer publicidad en la sección de deportes. Algo similar ocurre en la de noticias, mientras que los jueves parece no ser el día indicado, los viernes, martes y miércoles parecen ser una mejor opción.

Fuente de los datos

https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/twan/frames/frame.html

2. MANOVA

MANOVA surge como una extensión multivariante de ANOVA, permitiendo realizar estudios en los que intervengan más de una variable dependiente.

2.1. Introducción

Este procedimiento fue inicialmente desarrollado por Samuel Stanley Wilks (1906-1964) en 1932. Se basó en el principio de la razón de verosimilitud generalizada, apoyándose sobre las bases asentadas por Fisher en el principio de la estimación de máxima verosimilitud. Inicialmente fueron Jerzy Neyman y Egon Pearson los que en su publicación "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference" en 1928 dieron el primer paso al empleo de este principio en el contraste de hipótesis.

Figura 2.1. Samuel Stanley Wilks. Adaptado de *Samuel Stanley Wilks,* por School of Mathematics and Statistics University of St Andrews, MacTutor (https://mathshistory.st-andrews.ac.uk/Biographies/Wilks/pictdisplay/)

Para entender mejor este concepto, supóngase que se tienen *n* variables aleatorias idénticamente distribuidas.

$$X = [X_1, X_2, \dots, X_n\ ]$$

Un vector con una serie de muestras observadas de las variables aleatorias.

$$x = [x_1, x_2, \dots, x_n\ ]$$

Cada valor x_i está representando una magnitud vectorial. Y un vector con los parámetros desconocidos de la función de densidad conjunta.

$$\theta = [\theta_1, \theta_2, \dots, \theta_n\ ]$$

Lo que dice el principio de máxima verosimilitud es que la mejor estimación de los parámetros de la función de densidad conjunta será aquella que haga máxima la probabilidad de observar los valores obtenidos en la muestra definida por el vector *x*.

$$\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n \rightarrow\ \ f(x_1, x_2, \dots, x_n; \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n\ ) = \max f(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_n\ )$$

Los estimadores de los parámetros se suelen representar con un circunflejo sobre su símbolo. El principio de la razón de verosimilitud se basa en comparar la probabilidad anterior para dos muestras que cumplan con distintas hipótesis cada una. Si el cociente entre ambas es cercano a la unidad se puede considerar que se cumple la hipótesis nula y no la alternativa, mientras que a medida que se acerca a 0 mayores será las diferencias.

En MANOVA las hipótesis a contrastar son la de igualdad de medias de las variables dependientes.

$$H_0: \begin{bmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{bmatrix} = \begin{bmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{bmatrix} = \cdots = \begin{bmatrix} \mu_{1k} \\ \mu_{2k} \\ \vdots \\ \mu_{pk} \end{bmatrix}$$

$$H_1: algún\ vector\ de\ medias\ es\ distinto$$

Para ello se usa el test basado en el estadístico Λ, Lambda de Wilks, que se formula a partir de la definición del principio de la razón de verosimilitud. En este test se utiliza el estadístico Λ para determinar qué hipótesis es la que se aceptará, su valor se puede expresar como

$$\Lambda = \frac{|W|}{|T|},$$

donde W y T son una extensión a un caso multivariante de lo que sería en el ANOVA las variabilidades internas de los grupos y la variabilidad total. En la siguiente sección se dará un mayor detalle de cómo se obtienen estas variables y de cómo el valor del estadístico se transforma en uno que siga una distribución de Fisher.

Al igual que en ANOVA, dependiendo del número de variables independientes que se utilicen MANOVA se formulará de una forma, se explicará el procedimiento del MANOVA de un factor y del de dos factores, siendo el resto de los análisis multivariantes en los que intervenga más de un factor una extensión de lo explicado para el de dos factores.

2.2. MANOVA de un factor

La forma de operar del MANOVA es similar a la de un ANOVA, se calcula un valor de F, que en este caso será un valor aproximado y si el valor es extremo se rechaza la hipótesis nula. A pesar de las similitudes con el ANOVA no hay que entender MANOVA como una serie de análisis univariantes para cada variable dependiente.

En primer lugar, MANOVA proporciona un mayor control sobre la tasa de error conjunto, ya que, si se llevan a cabo varios ANOVAs seguidos para analizar la influencia de una serie de factores en cada una de las variables dependientes, seguido de sus respectivas comparaciones post-hoc se dispara la probabilidad de error tipo I. Además, en MANOVA se tiene en cuenta las relaciones de dependencia que puedan existir entre las variables dependientes, de forma que el resultado del análisis puede ser significativo, mientras que el de alguno de los univariantes no serlo.

2.2.1. T^2 de Hotelling

Se comenzará primero definiendo el caso en el que solamente se tienen dos niveles formados en la variable independiente, en este caso se puede utilizar una particularización de MANOVA, la prueba de la T^2 de Hotelling, es una extensión de la prueba de la t de Student vista para el caso de ANOVA.

Para entender la prueba es necesario aclarar previamente unos conceptos referidos a las distancias entre observaciones en el espacio.

2.2.1.1. Distancia estadística

Existen distintas formas de medir las distancias entre dos puntos de una misma variable aleatoria, supóngase $X \rightarrow N(0,2)$ y se quiere medir la distancia entre las observaciones $x_1 = -2$ y $x_2 = 2$, una forma habitual de hacerlo es empleando la distancia euclídea entre ambos puntos.

$$d_{12} = \sqrt{(x_1 - x_2)^2}$$

Sin embargo, si atendemos a la Figura 2.2, parece que la distancia para las dos observaciones mostradas no es la misma para las de una variable aleatoria con distribución N(0,2) que una con N(0,1).

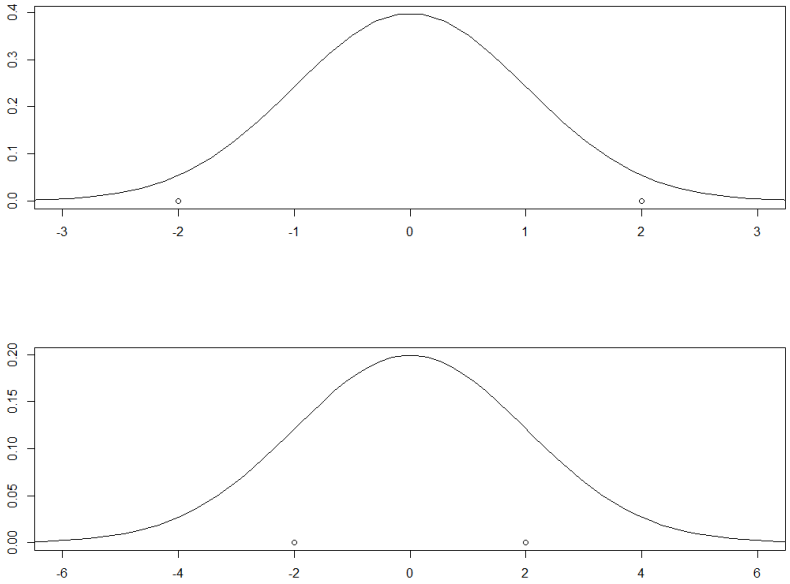


Figura 2.2. Distancia entre dos puntos bajo diferente distribución

La distribución superior sigue una distribución N(0,1) y la inferior N(0,2), la probabilidad de obtener un valor comprendido entre los dos puntos de la N(0,1) es de 0,954, mientras que en la segunda es de 0,683, por tanto, parece lógico emplear una corrección de la distancia euclídea atendiendo a la varianza de la distribución, es esta corrección la que recibe el nombre de distancia estadística

$$SD_{ij}^2 = \left(\frac{x_i - x_j}{\sigma}\right)^2.$$

Esta expresión se puede extender para el caso en el que se tengan p variables

$$SD_{ik}^2 = \sum_{j=1}^{p} \left(\frac{x_{ij} - x_{kj}}{\sigma_j}\right)^2.$$

Siendo *i* y *k* las observaciones, *j* la variable dependiente en la que se mide la distancia y σ_j la desviación típica de la variable aleatoria.

2.2.1.2. Distancia de Mahalanobis

La distancia de Mahalanobis se puede definir como la distancia entre dos observaciones tomando en consideración la correlación existente entre las variables y sus varianzas individuales. Si se tienen dos variables correlacionadas, la distancia entre las observaciones *i* y *k* viene dada por

$$MD_{ik}^2 = \frac{1}{1-\rho^2}\left[\frac{(x_{i1} - x_{k1})^2}{\sigma_1^2} + \frac{(x_{i2} - x_{k2})^2}{\sigma_2^2} - 2\frac{\rho(x_{i1} - x_{k1})(x_{i2} - x_{k2})}{\sigma_1^2\sigma_2^2}\right],$$

donde ρ es la correlación, se observa que para el caso en el que $\rho = 0$ la expresión anterior se reduce a la suma de distancias estadísticas de un caso univariante. Generalizando la fórmula anterior al caso de *p* variables se obtiene la siguiente expresión,

$$MD_{ik} = (x_i - x_k)^T \Sigma^{-1} (x_i - x_k),$$

siendo x_i y x_k los vectores de coordenadas de las observaciones *i* y *k*, de dimensión $p \times 1$ y Σ^{-1} la inversa de la matriz de covarianzas de dimensión $p \times p$.

A partir de esta última definición se construye el estadístico T^2 para la prueba de Hotelling,

$$T^2 = \left(\frac{n_1 \times n_2}{n_1 + n_2}\right) MD^2,$$

siendo n_1y n_2 el número de observaciones de cada uno de los niveles con los que se está haciendo el análisis y MD^2 la distancia de Mahalanobis entre los centroides de ambos niveles. La expresión anterior se puede transformar en un valor de F exacto,

$$F_{p,(n_1+n_2-p-1)} = \frac{(n_1 + n_2 - p - 1)}{2(n_1 + n_2 - p)} T^2.$$

De la misma forma que en ANOVA, se comprobaría si el valor de F es tan extremo como para considerarse significativa la diferencia entre ambos niveles.

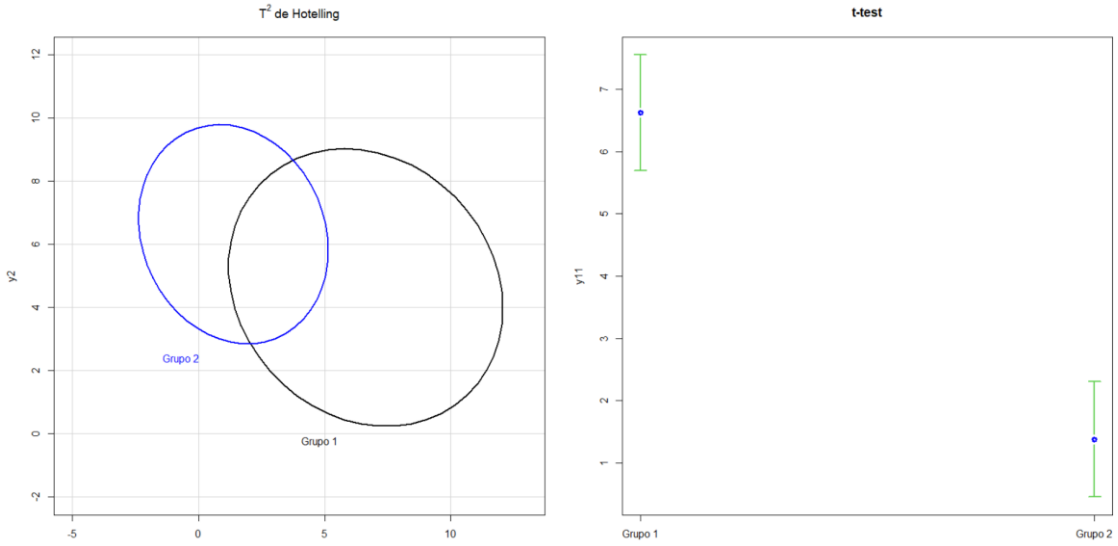


Figura 2.3. Similitudes entre la prueba de la t de Student y la T^2 de Hotelling

En la Figura 2.3 se representa como la prueba de la T^2 de Hotelling es una extensión de la prueba de la t de Student, pasando de trabajar con intervalos de confianza de las medias a usar la distancia de Mahalanobis entre los puntos en un espacio bidimensional.

Para los casos en los que el número de niveles formados en la variable independiente sea superior a dos, el procedimiento a seguir será igual que el caso de ANOVA, en primer lugar, se analiza si existen diferencias de manera global y luego se estudiará individualmente cada una de las variables dependientes, ya por último, se analizará las diferencias entre los niveles formados por el factor en las variables dependientes cuyo ANOVA individual haya dado un p-valor inferior al nivel de significación.

2.2.2. Modelo

Para formular el modelo general del MANOVA de un factor se partirá de la misma formulación de la variable dependiente que se hizo en ANOVA

$$y_{uij} = \mu_{ui} + \varepsilon_{uij}, \quad \varepsilon_{uij} \to N(0, \sigma_u^2),$$

$$u = 1,2 \dots, p, \quad i = 1,2, \dots I, \quad j = 1,2, \dots, n_i.$$

Donde el subíndice u está indicando la variable dependiente, mientras que *i* y *j* se siguen refiriendo al grupo y a la observación respectivamente. Esto expresado en función de sus estimadores quedaría como

$$y_{uij} = \bar{y}_{ui} + y_{uij} - \bar{y}_{ui}.$$

La variabilidad total se puede descomponer en las partes explicada y no explicada, pero a diferencia de ANOVA, estos términos se expresarán de forma matricial. La variabilidad interna o no explicada se expresa a partir de la matriz de suma de cuadrados y productos cruzados dentro del grupo, sus términos se calculan de la siguiente forma,

$$SS_{wu} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{uij} - \bar{y}_{ui})^2,$$

$$SCP_{wuv} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{uij} - \bar{y}_{ui})(y_{vij} - \bar{y}_{vi}), \quad u \neq v.$$

SS_{wu} es la suma de cuadrados de la variable *u* y SCP_{wuv} es la suma de productos cruzados, el subíndice *w* hace referencia a que se trata de la variabilidad interna del grupo, es decir, variabilidad no explicada, n_i es el número de observaciones en el grupo *i*, habiendo un total de *k* niveles. Con estos valores se construye la matriz de suma de cuadrados y productos cruzados dentro del grupo ($SSCP_w$). La matriz tiene dimensión $p \times p$, esta explicación se reducirá al caso de $p = 2$,

$$SSCP_w = \begin{pmatrix} SS_{w1} & SCP_w \\ SCP_w & SS_{w2} \end{pmatrix}.$$

La matriz es simétrica, ya que como se puede ver en la expresión anterior $SCP_{wuv} = SCP_{wvu}$. Se la denominará también matriz de error (E). Tiene $n - k$ grados de libertad, con *n* el número total de observaciones.

La variabilidad explicada se expresa mediante la matriz de suma de cuadrados y productos cruzados entre grupos, sus términos son,

$$SS_{bu} = \sum_{i=1}^{k} n_i (\bar{y}_{ui} - \bar{y}_u)^2,$$

$$SCP_{buv} = \sum_{i=1}^{k} n_i (\bar{y}_{ui} - \bar{y}_u)(\bar{y}_{vi} - \bar{y}_v), \quad u \neq v,$$

en este caso el subíndice *b* se refiere a la variabilidad entre los grupos. De la misma forma se construye la matriz de suma de cuadrados y productos cruzados para el caso $p = 2$,

$$SSCP_b = \begin{pmatrix} SS_{b1} & SCP_b \\ SCP_b & SS_{b2} \end{pmatrix}.$$

Se la conoce también como matriz de hipótesis (H). Tiene $k - 1$ grados de libertad.

El cálculo de F a partir de estas matrices no es directo, existiendo diferentes test para ello, todos basados en el principio enunciado para la prueba de Wilks pero con distintas parametrizaciones, aquí se comentarán los cuatro más conocidos. Para ello se ha empleado como referencia el artículo de Hintze (2007).

2.2.3. Lamba de Wilks

El estadístico Λ se ha definido según lo visto en la introducción a MANOVA como

$$\Lambda = \frac{|E|}{|E + H|},$$

que también puede expresarse en función de los autovalores de la matriz $H(E + H)^{-1}$,

$$\Lambda = \prod_{j=1}^{p} (1 - \theta_j),$$

donde θ_j es el autovalor *j* de la matriz $H(E + H)^{-1}$ y *p* el número de variables dependientes. La aproximación del valor de F bajo cumplimiento de la hipótesis nula se hace de la siguiente forma,

$$F_{ph,ft-g} = \frac{(ft - g)(1 - \Lambda^{\frac{1}{t}})}{ph\Lambda^{\frac{1}{t}}},$$

$f = e - \frac{1}{2}(p - h + 1)$,

$g = \frac{ph-2}{2}$,

$$t = \begin{cases} \sqrt{\frac{p^2h^2-4}{p^2+h^2-5}} & si\ p^2 + h^2 - 5 > 0 \\ 1 & si\ p^2 + h^2 - 5 \leq 0 \end{cases}$$

siendo *e* y *h* los grados de libertad de las matrices E y H respectivamente.

2.2.4. La traza de Hotelling

$$T_g{}^2 = e \sum_{j=1}^{s} \phi_j,$$

$s = min(p, h)$,

donde ϕ_j es cada uno de los autovalores de la matriz HE^{-1}. Su aproximación al valor de F se expresa como,

$$F_{a,b} = \frac{T_g^{\,2}}{ce},$$

$a = ph,$

$b = 4 + \frac{(a+2)}{(B-1)},$

$c = \frac{a(b-2)}{b(e-p-1)},$

$B = \frac{(e+h-p-1)(e-1)}{(e-p-3)(e-p)}.$

2.2.5. La traza de Pillai

$$V = \sum_{j=1}^{s} \theta_j = tr(H(E+H)^{-1}),$$

$s = min(p, h).$

La aproximación al valor de F se hace de la siguiente forma,

$$F_{s(2m+s+1,\ s(2n+s+1)} = \frac{(2n+s+1)V}{(2m+s+1)(s-V)},$$

$m = \frac{(|p-h|-1)}{2},$

$n = \frac{(e-p-1)}{2}.$

2.2.6. La mayor raíz de Roy

$$F_{2w_1+2,2w_2+2} = \frac{2w_2+2}{2w_1+2}\phi_{max},$$

$w_1 = \frac{(|p-h|-1)}{2},$

$w_2 = \frac{(e-p-1)}{2},$

siendo ϕ_{max} el mayor de los *s* primeros autovalores, con $s = min(p, h)$.

Con todos estos valores se puede construir la tabla de MANOVA, que para el MANOVA de un factor sería la mostrada en la Tabla 2.1.

	Grados de libertad	Pillai	F estimado	Grados de libertad numerador	Grados de libertad denominador	p-valor
Tratamiento	$k-1$	$\sum_{j=1}^{s} \theta_j$	$\frac{(2n+s+1)V}{(2m+s+1)(s-V)}$	$\frac{(\lvert p-h\rvert-1)}{2}$	$\frac{(e-p-1)}{2}$	$P(F \geq F_0)$
Residual	$n-k$					

Tabla 2.1. Tabla MANOVA de un factor

Se ha expresado en función del estadístico utilizado en la traza de Pillai ya que es la prueba que por defecto calcula RStudio. Igual que ANOVA, con el valor obtenido de F se calcula el p-valor y atendiendo al nivel de significación que se esté usando se decidirá si se acepta o se rechaza la hipótesis nula.

2.2.7. Tamaño del efecto

En este punto conviene explicar qué es el tamaño del efecto, ya que ha sido una de las medidas en la que se han basado los posteriores experimentos.

El tamaño del efecto (α_i, β_j, $\alpha\beta_{ij}$) de una variable independiente es una medida del grado en el que esta afecta a las variables dependientes. Existen diferentes medidas para computar su cálculo, la más básica podría ser la diferencia entre las medias de ambos grupos, sin embargo, esta medida oculta múltiples factores como ya se ha visto al comentar los inconvenientes de usar la distancia euclídea, por ello conviene emplear otras como la ya comentada distancia de Mahalanobis o la eta-squared (η^2). Esta última para el caso univariante se calcula como

$$\eta^2 = 1 - \Lambda = \frac{SS_b}{SS_t},$$

en el multivariante,

$$\eta^2 = 1 - \Lambda = \frac{|SSCP_b|}{|SS_t|}.$$

Da una medida de la proporción de varianza que es debida a las diferencias entre los grupos, de forma que cuanto mayor sea su valor, incluso pequeñas diferencias entre los grupos se interpretarán como significativas, al ser debidas al efecto del factor y no a la parte aleatoria.

2.2.8. Hipótesis

Al igual que en ANOVA, para que el análisis sea válido hay que verificar el cumplimiento de una serie de hipótesis.

2.2.8.1. Normalidad

Existen pocas pruebas para comprobar la hipótesis de normalidad en el caso multivariante, la prueba gráfica es muy similar a la explicada para el caso univariante. Se comienza calculando la distancia de Mahalanobis de cada observación al centroide de la muestra y se ordenan de menor a mayor las distancias. A continuación, se calculan de la misma forma que en el gráfico Q-Q, las probabilidades asociadas a los percentiles y con estas se calculan a partir de una distribución chi-cuadrado (χ^2) de *p* grados de libertad, con *p* el número de variables, el valor de los percentiles asociados. En la Tabla 2.2 se muestra un ejemplo de estos cálculos para un caso en el que se generen las variables utilizando una función normal multivariante.

Orden	Distancia de Mahalobis	Probabilidad	Cuantil
i	MD^2	(i-0,5)/n	χ_2^2
1	0,661	0,02	0,04
2	0,765	0,06	0,124
3	0,79	0,1	0,211
4	1,078	0,14	0,302
5	1,097	0,18	0,397
6	1,111	0,22	0,497
7	1,21	0,26	0,602
8	1,323	0,30	0,713
9	1,34	0,34	0,831
10	1,597	0,38	0,956
11	1,901	0,42	1,089
12	1,995	0,46	1,232
13	2,021	0,5	1,386
14	2,11	0,54	1,553
15	2,378	0,58	1,735

Tabla 2.2. Cálculos para construir el gráfico para verificar la hipótesis de normalidad multivariante

Si se representa en una gráfica en el eje de abscisas los valores de MD^2 y en ordenadas los de χ^2, la nube de puntos resultante debe seguir de forma aproximada una línea recta.

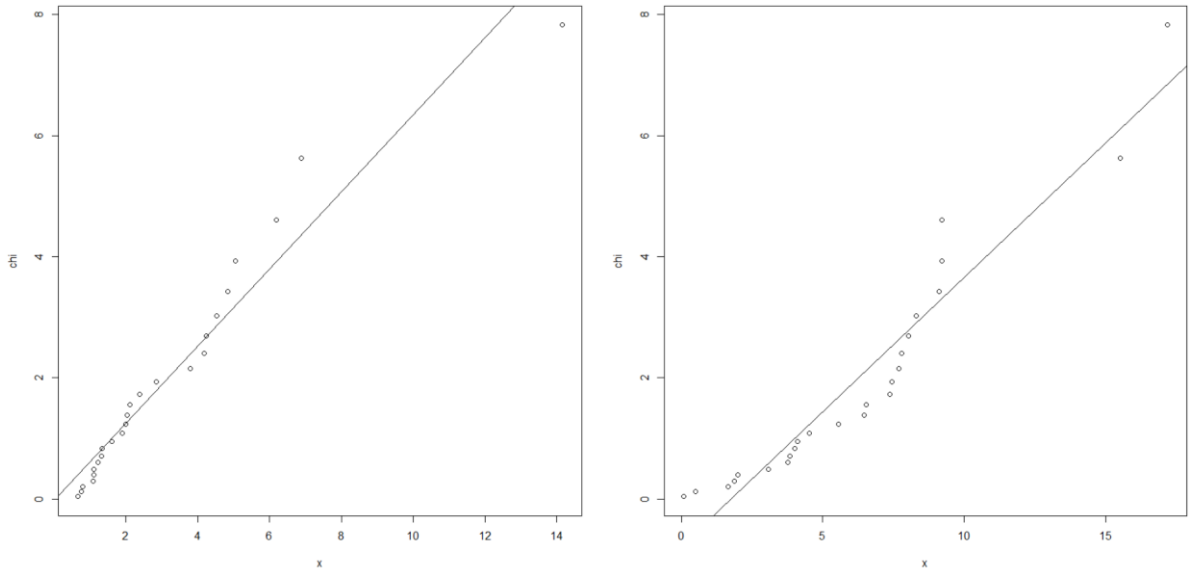


Figura 2.4. Nube de puntos de una distribución que cumple con la hipótesis de normalidad y otra que no

Se han generado dos conjuntos de datos, uno utilizando una función normal multivariante y en el otro usando distribuciones gamma, los resultados son los mostrados en la Figura 2.4, la gráfica izquierda son los datos con distribución normal, que se ajustan mucho mejor a la recta que los de la distribución gamma.

Si existe falta de normalidad se pueden aplicar las transformaciones de la misma forma que se vio para el análisis univariante.

2.2.8.2. Homocedasticidad

Esta hipótesis es mucho más restrictiva que en el caso del análisis univariante ya que exige que todos los elementos de la matriz de covarianzas sean iguales a lo largo de todos los niveles. La falta de cumplimiento de esta hipótesis penaliza severamente la precisión de cálculo de la probabilidad de error tipo I, en especial en casos en los que el número de observaciones por grupo no sea constante.

La mayoría de las pruebas formales que existen son sensibles a la falta de normalidad en el conjunto, por tanto, previo a las pruebas hay que comprobar la normalidad y aplicar las transformaciones necesarias en caso de que no se cumpla.

La prueba más empleada es la de la M de Box, en RStudio se lleva a cabo con el comando *boxM()* del paquete *heplots*.

2.2.8.3. Independencia

Es muy importante que los resultados de una observación no presenten ninguna relación de dependencia con otra. Depende de las condiciones en las que se haya llevado a cabo la recogida de datos por parte del investigador.

Es la única hipótesis en la que tamaños muestrales grandes pueden dar peores resultados, ya que un mayor número de observaciones, referidas a distintas muestras, dependientes las unas de las otras resultan en un mayor incumplimiento de la hipótesis.

Se han enunciado las tres hipótesis más importantes, no obstante, existen otras como la de linealidad y multicolinealidad o la de presencia de valores atípicos.

2.2.9. Comparaciones post-hoc

Al igual que en ANOVA, el resultado de MANOVA indica si existen diferencias entre los niveles formados en la variable independiente, pero no entre cuáles, además al estar trabajando con más de una variable dependiente, es interesante saber para cuál de las variables se dan las diferencias, para ello se llevan a cabo las comparaciones post-hoc.

Sigue existiendo el problema de la inflación de la probabilidad de error de tipo I, además ahora al estar trabajando con múltiples variables dependientes, previo al análisis de las diferencias entre niveles se llevan a cabo una serie de ANOVAs individuales para ver qué variables son influyentes, lo que aumenta aún más la probabilidad de error tipo I.

Para corregir este efecto, se seguirán empleando las correcciones vistas para el caso univariante. Los procedimientos más habituales son los de Tukey HSD, que ya se ha comentado anteriormente, el método de Scheffé o la diferencia menos significativa de Fisher (LSD) entre otros.

Como alternativa a las comparaciones post-hoc existen las comparaciones planeadas, en este tipo de comparaciones el investigador define cuáles son los grupos que quiere analizar en vez de analizar las diferencias entre todos. Al reducir el número de comparaciones a las necesarias, la probabilidad de error tipo I está mucho más controlada, pero para que estas comparaciones sean útiles, el investigador debe tener ciertas evidencias de que puedan existir diferencias entre esos grupos. Una vez definidos los grupos a analizar se lleva a cabo el contraste haciendo una

combinación lineal de las medias de esos grupos, para posteriormente obtener un valor de F con el que llevar a cabo el test estadístico pertinente.

Aquí se seguirán empleando las comparaciones post-hoc en línea con lo visto para los análisis univariantes. Hay que tener en cuenta también, como se ha comentado anteriormente, que es posible que el MANOVA indique diferencias significativas entre los grupos y los ANOVAs individuales no, ya que en MANOVA se tienen en cuenta las relaciones existentes entre las variables dependientes.

A continuación, para facilitar la compresión de lo enunciado, se verá un caso práctico de un MANOVA de un factor.

2.2.10. Ejemplo

En un estudio preliminar para el diseño de cascos para jugadores de fútbol americano se desea saber si los requerimientos en cuanto a dimensión de los jugadores varían atendiendo a la categoría de juego. Para ello se recopilan medidas de distintas dimensiones de la cabeza, el ancho, la circunferencia, distancia entre nuca y frente medida a la altura de los ojos, distancia de los ojos y oídos a lo alto de la cabeza y medidas de la mandíbula. Los datos han sido recopilados para tres grupos distintos: jugadores de instituto, jugadores de universidad y no jugadores.

En primer lugar, se comenzará verificando que se cumplen las hipótesis del modelo.

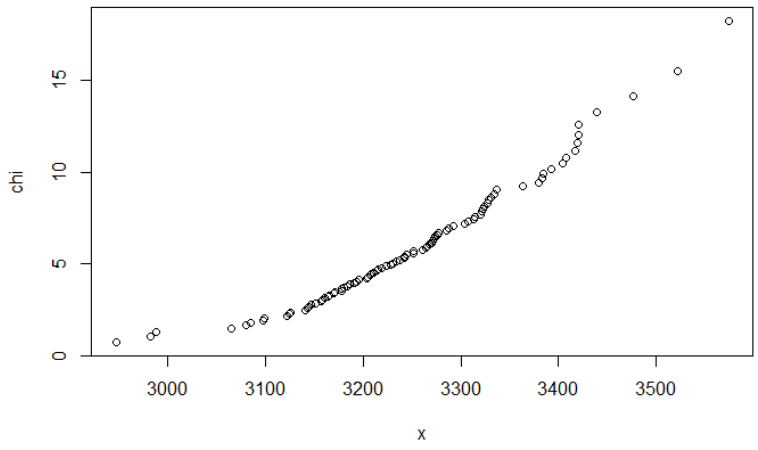


Figura 2.5. Gráfico para comprobar normalidad

```
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  y
## Chi-Sq (approx.) = 57.472, df = 42, p-value = 0.05622
```

De la gráfica se comprueba que la nube de puntos se distribuye a lo largo de una línea recta y por tanto se cumple la normalidad, del resultado de la prueba de la M de Box, empleando un nivel de significación $\alpha = 0{,}05$, se verifica la hipótesis de homocedasticidad, no es necesario aplicar transformación.

Se procede entonces a calcular la tabla de MANOVA usando la prueba de Pillai.

```
##               Df  Pillai approx F num Df den Df    Pr(>F)
## FootHead[, 1]  2 0.76116   8.4994     12    166 1.876e-12 ***
## Residuals     87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado indica que existen diferencias entre las dimensiones de las cabezas de los distintos grupos. A continuación, se llevarán a cabo unos ANOVAs individuales para cada variable dependiente, utilizando la corrección del p-valor por el método de Bonferroni.

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## tipo         2  143.4   71.68   58.16 <2e-16 ***
## Residuals   87  107.2    1.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El resultado indicado para este primer ANOVA se refiere a la distancia de los ojos a la parte superior de la cabeza.

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## tipo         2  27.72  13.861   22.43 1.4e-08 ***
## Residuals   87  53.77   0.618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La otra variable para la que se encuentran diferencias entre los grupos es la de la distancia de los oídos a la parte superior de la cabeza.

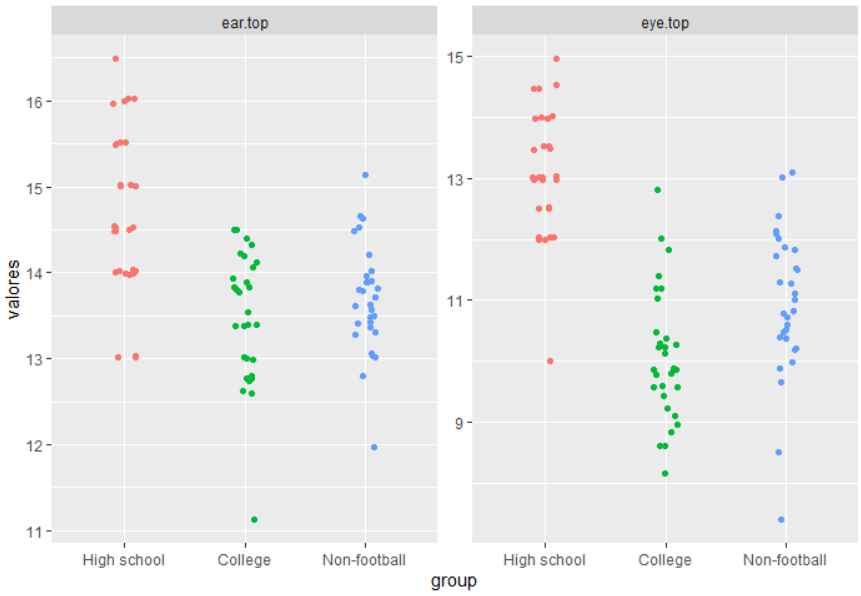


Figura 2.6. Observaciones por grupos para las dos variables influyentes

En la Figura 2.6 se muestra la nube de puntos de las observaciones en estas dos variables, agrupadas por tipo de jugador. Para saber los grupos en los que se dan las diferencias en cada una de las variables se llevará a cabo la prueba de Tukey HSD, manteniendo una tasa de error conjunto del 5%.

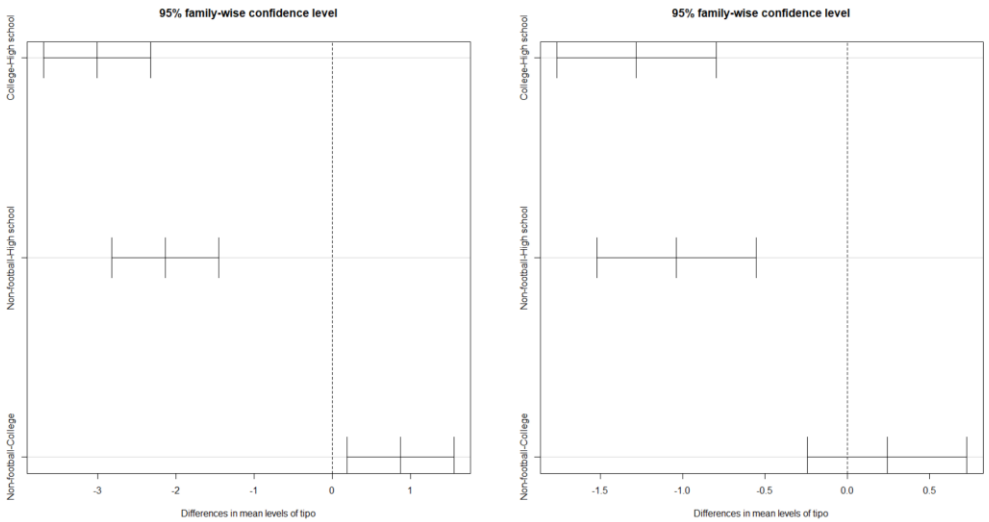


Figura 2.7. Tukey HSD para las variables con diferencias significativas

En la Figura 2.7 se muestran los diagramas con los intervalos de confianza de la diferencia entre las medias para ambas variables, a la izquierda la de la distancia de los ojos a la parte superior de la cabeza y a la derecha la distancia tomada desde las orejas. En la medida de tomada desde los ojos todos los grupos parecen discernir entre sí, mientras que en la tomada desde las orejas no es significativa la diferencia entre los que no juegan al fútbol y los que juegan en universidad.

Fuente de los datos

Rencher, A. C. (1995), Methods of Multivariate Analysis, New York: Wiley, Table 8.3.

2.3. MANOVA de dos factores

Por último, se dará una exposición del MANOVA de dos factores a partir de todo lo enunciado para el de un factor. De la misma forma que en los ANOVAs, estos dos análisis se diferencian en la cantidad de variables independientes que intervienen, pasando a ser de dos en el MANOVA de dos factores.

2.3.1. Modelo

La expresión de las variables dependientes era de la forma

$$y_{uijk} = \mu_u + \alpha_{ui} + \beta_{uj} + \alpha\beta_{uij} + \varepsilon_{uijk}, \quad \varepsilon_{uijk} \to N(0, \sigma_u^2).$$

$$u = 1,2,\dots,p, \;\; i = 1,2\dots,I, \;\; j = 1,2,\dots,J, \;\; k = 1,2,\dots,m.$$

Las estimaciones para cada uno de los parámetros se sigue haciendo de la misma forma que en el caso univariante,

$$y_{uijk} = \bar{y}_{u\bullet\bullet\bullet} + \hat{\alpha}_{ui} + \hat{\beta}_{uj} + \widehat{\alpha\beta}_{uij} + e_{uijk}, \quad e_{uijk} \to N(0, s_{Ru}^2).$$

La variabilidad total se descompone de nuevo en la parte explicada por cada uno de los factores , la explicada por la interacción entre los factores y la parte no explicada. Al estar trabajando con múltiples variables dependientes se expresará en función de las matrices de sumas de cuadrados y productos cruzados, la expresión de cada una de ellas se dará para el caso de $p = 2$.

En primer lugar, la matriz asociada al efecto principal del factor 1 estará compuesta por

$$SS_{F1u} = mJ\sum_{i=1}^{I}(\bar{y}_{ui\bullet\bullet} - \bar{y}_{u\bullet\bullet\bullet})^2,$$

$$SCP_{F1uv} = mJ\sum_{i=1}^{I}(\bar{y}_{ui\bullet\bullet} - \bar{y}_{u\bullet\bullet\bullet})(\bar{y}_{vi\bullet\bullet} - \bar{y}_{v\bullet\bullet\bullet}), \quad u \neq v\,.$$

Se ha supuesto, igual que se hizo en ANOVA, que el número de observaciones por combinaciones de niveles de factores permanece constante y de valor *m*, *I* es el número de niveles del factor 1 y *J* el núero de niveles de formados en el factor 2. El término SS_{bF1u} es la suma de cuadrados de las diferencias entre grupos debidas al efecto del factor 1 en la variable dependiente *u*. Para el caso de $p = 2$ se construye la matriz de suma de cuadrados y productos cruzados igual que en al anterior caso

$$SSCP_{F1} = \begin{pmatrix} SS_{F11} & SCP_{F112} \\ SCP_{F121} & SS_{F12} \end{pmatrix},$$

que sigue siendo simétrica y representa la parte de la variabilidad que viene determinada por las diferencias entre grupos debida al efecto del factor 1, tiene asociados $I - 1$ grados de libertad. La debida al efecto del factor 2 se expresa de forma similar,

$$SS_{F2u} = mI\sum_{j=1}^{J}(\bar{y}_{u\bullet j\bullet} - \bar{y}_{u\bullet\bullet\bullet})^2,$$

$$SCP_{F2uv} = mI\sum_{j=1}^{J}(\bar{y}_{u\bullet j\bullet} - \bar{y}_{u\bullet\bullet\bullet})(\bar{y}_{v\bullet j\bullet} - \bar{y}_{v\bullet\bullet\bullet}), \quad u \neq v,$$

$$SSCP_{F2} = \begin{pmatrix} SS_{F21} & SCP_{F212} \\ SCP_{F221} & SS_{F22} \end{pmatrix}.$$

Tiene $J - 1$ grados de libertad. Para el efecto de la interacción, el razonamiento a seguir es análogo al de los efectos principales de los factores, teniendo en cuenta la expresión de la variabilidad explicada por la interacción descrita en el apartado de ANOVA,

$$SS_{F1F2u} = m\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{uij\bullet} - \bar{y}_{ui\bullet\bullet} - \bar{y}_{ui\bullet\bullet} + \bar{y}_{u\bullet\bullet\bullet})^2\,,$$

$$SCP_{F1F2uv} = m\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{uij\bullet} - \bar{y}_{ui\bullet\bullet} - \bar{y}_{ui\bullet\bullet} + \bar{y}_{u\bullet\bullet\bullet})(\bar{y}_{vij\bullet} - \bar{y}_{vi\bullet\bullet} - \bar{y}_{vi\bullet\bullet} + \bar{y}_{v\bullet\bullet\bullet}), \quad u \neq v,$$

$$SSCP_{F1F2} = \begin{pmatrix} SS_{F1F21} & SCP_{F1F212} \\ SCP_{F1F221} & SS_{F1F22} \end{pmatrix}.$$

Esta matriz tiene $(I - 1)(J - 1)$ grados de libertad. Por último, la matriz asociada a la variabilidad no explicada se construye como

$$SS_{wu} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}(\bar{y}_{uijk} - \bar{y}_{uij\bullet})^2,$$

$$SCP_{wuv} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}(\bar{y}_{uijk} - \bar{y}_{uij\bullet})(\bar{y}_{vijk} - \bar{y}_{vij\bullet}), \quad u \neq v,$$

$$SSCP_w = \begin{pmatrix} SS_{w1} & SCP_{w12} \\ SCP_{w21} & SS_{w2} \end{pmatrix}.$$

Esta última matriz consta de $IJ(m-1)$ grados de libertad. Con todas las matrices definidas y recordando las estimaciones vistas para el valor de F a partir de las diferentes pruebas disponibles se puede construir la tabla del MANOVA de dos factores.

	Grados de libertad	Pillai	F estimado	Grados de libertad numerador	Grados de libertad denominador	p-valor
Factor 1	$I-1$	$\sum_{j=1}^{s}\theta_j$	$\frac{(2n+s+1)V}{(2m+s+1)(s-V)}$	$\frac{(\lvert p-h\rvert-1)}{2}$	$\frac{(e-p-1)}{2}$	$P(F_1 \geq F_{02})$
Factor 2	$J-1$		ídem			$P(F_2 \geq F_{02})$
Factor 1:Factor 2	$(I-1)(J-1)$		ídem			$P(F_{12} \geq F_{012})$
Residual	$IJ(m-1)$					

Tabla 2.3. Tabla MANOVA de dos factores

En la Tabla 2.3 se muestra su forma. El cálculo del valor de F es idéntico en todos los efectos, lo único que cambia es la expresión de las matrices de hipótesis que se utilizarán para su cálculo, siendo las tres descritas anteriormente, en todos los análisis se compararán con la misma matriz de error.

De la misma manera que en el MANOVA de un factor, en este tipo de análisis hay que comprobar el cumplimiento de las hipótesis ya mencionadas.

2.3.2. Comparaciones post-hoc

Ya se han comentado anteriormente los inconvenientes de realizar múltiples comparaciones post-hoc y la alternativa de realizar contrastes planeados, no obstante, esta explicación se centrará en las comparaciones post-hoc al no disponer siempre de evidencias de existencia de diferencias entre niveles.

Para las comparaciones se deberá tener en cuenta, en primer lugar, en qué variables dependientes se dan las diferencias, para a continuación analizar las diferencias entre los niveles formados en esa variable.

De la misma manera que en el caso univariante, dependiendo de si existe o no interacción se estudiarán las diferencias atendiendo a los niveles formados por la combinación de ambos factores o a los efectos de los factores por separado. Las herramientas que se utilizan para este tipo de comparaciones siguen siendo las ya explicadas, primero se lleva a cabo el MANOVA, si el resultado es significativo, se sigue de los ANOVAs individuales para cada variable, en aquellas en las que se encuentren diferencias significativas, si la interacción también lo es, se representa el intervalo de confianza de las medias para los grupos formados por ambos factores

y si la interacción no es significativa se estudian las diferencias entre los niveles formados en cada factor de manera independiente.

A continuación, se realizará un ejemplo para concluir con la explicación del MANOVA de dos factores.

2.3.3. Ejemplo

Para este ejemplo se tomará como referencia el estudio realizado por Hartman (2016) y Heinrichs et al. (2015), en el que se quería analizar el rendimiento en medidas neurocognitivas de personas que padecieran esquizofrenia y trastorno esquizoafectivo, utlizando una batería de preguntas específicamente diseñada para personas con psicosis. El objetivo de este estudio era ver si existían diferencias entre los grupos y qué variables permitían distinguir entre los grupos de esquizofrenia y trastorno esquizoafectivo.

Para ello se sometió a estudio a tres grupos de personas, un grupo de control, que no padeciese ningún trastorno, otro de personas con esquizofrenia y otro de trastorno esquizoafectivo. Además, se separaron los sujetos en función de su sexo, conformando estas dos variables los factores. Se recopilaron ocho medidas numéricas: su velocidad, su atención, su memoria, su aprendizaje verbal, aprendizaje visual, capacidad de razonamiento, dominio de la cognición social y su edad, formando todas estas las variables dependientes.

El primer paso es analizar los datos para verificar el cumplimiento de las hipótesis.

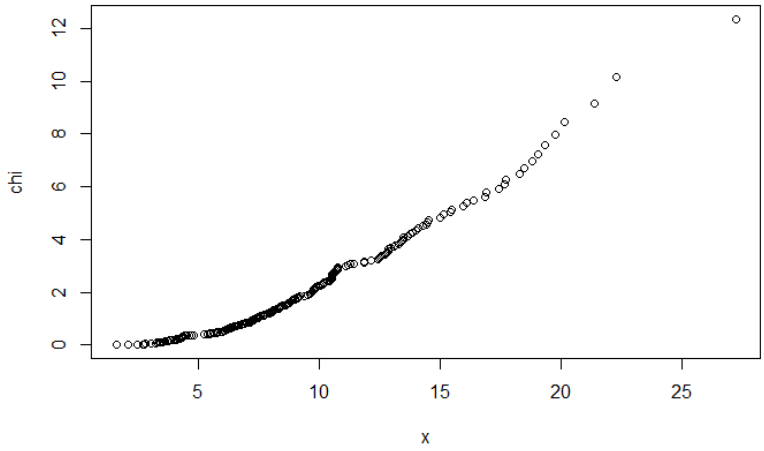


Figura 2.8. Gráfico para comprobar normalidad multivariante

```
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 209.61, df = 180, p-value = 0.06465
```

El gráfico de normalidad parece aproximarse a una línea recta y la prueba de la M de Box, empleando un nivel de significación de 0,05 no resulta significativa, con lo que se puede considerar que los datos son normales y homocedásticos. Se procede entonces a realizar el MANOVA.

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## Sex         1 0.11097   3.5729      8    229 0.0006315 ***
## Dx          2 0.35337   6.1699     16    460 1.806e-12 ***
## Sex:Dx      2 0.09511   1.4355     16    460 0.1205454
## Residuals 236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La interacción entre los dos factores no parece ser significativa, mientras que los efectos principales de cada uno de ellos sí. Como se comentaba, el interés del estudio está en ver, además de si existen diferencias, en qué variables se dan, para lo que se llevarán a cabo los ANOVAs individuales empleando el ajuste del p-valor por el método de Bonferroni.

Factor	Variable dependiente	Suma de cuadrados	Grados de libertad	Varianza	F	p-valor ajustado
Sexo	Velocidad	2	1	2	0,017	1
	Atención	30	1	30	0,186	1
	Memoria	45	1	45	0,333	1
	Verbal	477	1	477	5,648	0,146
	Visual	173	1	173	1,556	1
	Razonamiento	124	1	124	1,504	1
	Cognición social	836	1	836	5,979	0,122
	Edad	729	1	729	6,252	0,105
Trastorno						
	Velocidad	8379	2	4189,5	37,023	8,01E-14
	Atención	5549	2	2774,5	17,234	8,19E-07
	Memoria	3837	2	1918,5	14,174	1,22E-05
	Verbal	4493	2	2246,5	26,578	3,05E-10
	Visual	3605	2	1802,5	16,2	2,03E-06
	Razonamiento	4264	2	2132	25,834	5,62E-10
	Cognición social	4972	2	2486	17,781	5,08E-07
	Edad	289	2	144,5	1,237	1

Tabla 2.4. Resultados ANOVA para el efecto de los factores

En la Tabla 2.4 se recoge un resumen de los resultados de los análisis. Como se observa, el efecto del sexo no es influyente en ninguna variable dependiente, aunque en el MANOVA sí se considera como influyente, esta discrepancia puede deberse en parte a la corrección del p-valor y a que MANOVA tiene en cuenta las relaciones de dependencia que se pueden dar entre las variables dependientes. Respecto al efecto del trastorno, se considera influyente en todas las variables dependientes excepto en la edad.

En la Figura 2.9 se muestran las nubes de puntos en cada uno de los grupos formados dependiendo del tipo de trastorno. El objetivo del estudio era analizar si existen diferencias, que ya se ha comprobado mediante el MANOVA y estudiar en qué variables respuesta se encuentran las diferencias entre los grupos de trastorno esquizofrénico y esquizoafectivo, para lo que se podría llevar a cabo las comparaciones pareadas mediante la aplicación de la prueba de Tukey HSD.

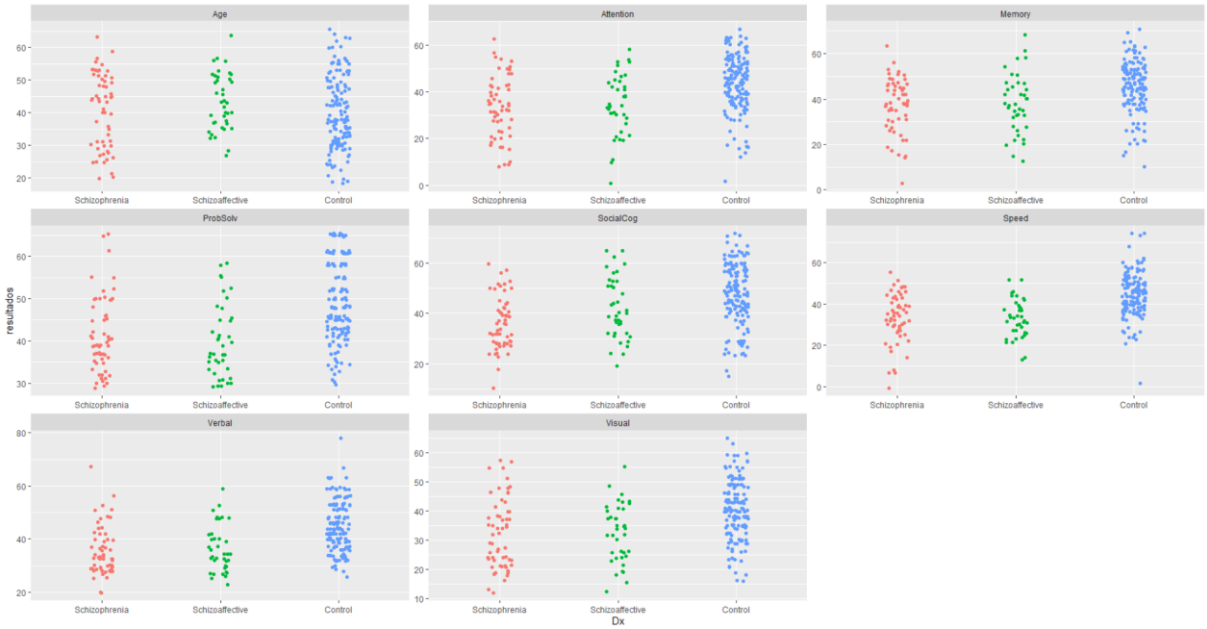


Figura 2.9. Puntuaciones de las variables dependientes en función del tipo de trastorno

De todas las variables en las que se han econtrado diferencias entre los grupos, la que tiene un intervalo de confianza más alejado del 0 para la diferencia entre los grupos de interés es la asociada al dominio de la cognición social, representado en la Figura 2.10, aunque sigue conteniendo al 0. Por tanto, para esas variables observadas no se podría concluir, en primera instancia, la existencia de diferencias entre ambos grupos.

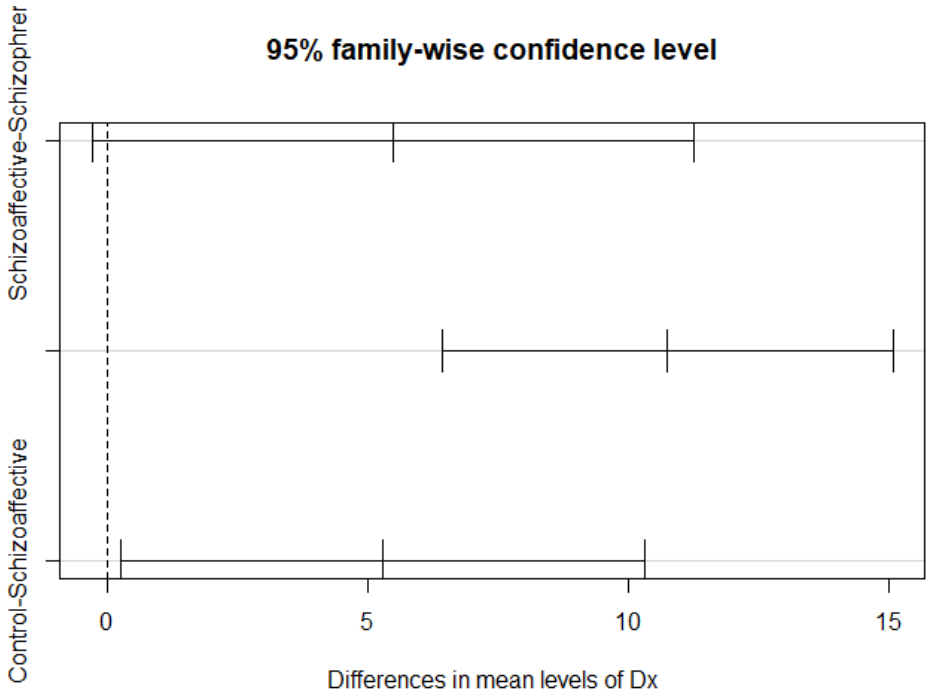


Figura 2.10. Prueba de Tukey para analizar diferencias entre grupos en la variable cognición social

Fuente de los datos

Hartman, L. I. (2016). Schizophrenia and Schizoaffective Disorder: One Condition or Two? Unpublished PhD dissertation, York University.

Heinrichs, R.W., Pinnock, F., Muharib, E., Hartman, L.I., Goldberg, J.O., & McDermid Vaz, S. (2015). Neurocognitive normality in schizophrenia revisited. Schizophrenia Research: Cognition, 2 (4), 227-232. doi: 10.1016/j.scog.2015.09.001

3. EXPERIMENTOS COMPUTACIONALES

Tanto MANOVA como ANOVA son empleados para analizar la significación estadística de las diferencias entre los distintos niveles que se encuentran en un conjunto de datos, usando una serie de variables, que reciben el nombre de variables dependientes cuyo valor dependerá del tipo de tratamiento al que se someta la muestra experimental, o de lo que se conoce como variables independientes.

La diferencia entre ambos es que MANOVA hace uso de múltiples variables métricas dependientes para encontrar las diferencias mientras que ANOVA emplea una única, o lo que en la jerga de la estadística se explica como que ANOVA es un análisis univariante, mientras que MANOVA es multivariante.

De lo que trata el experimento es de observar las diferencias en los resultados de ambos análisis variando una serie de parámetros, estos son:

- El tamaño del efecto, medido como la diferencia de medias en las variables dependientes.
- La varianza de los residuos. Se ha fijado un valor para cada una de las varianzas de los residuos de las variables dependientes.
- La correlación entre las variables dependientes, para lo que se han correlacionado sus términos de error y se ha variado la correlación para unas combinaciones de tamaños del efecto fijas.
- El tamaño muestral. Se han ido variando los tamaños muestrales, viendo los resultados con distintos valores de la correlación para cada uno de los tamaños.

3.1. MANOVA de un factor

En primer lugar, se comienza con un experimento básico, consistente en un MANOVA empleando dos variables dependientes y una sola variable independiente, que tendrá dos niveles, el grupo 1 y el 2, esta variable simboliza un tratamiento ficticio al que se está sometiendo el conjunto de muestras. Ya se ha definido anteriormente la forma de las variables dependientes en este tipo de análisis,

$$y_{uij} = \mu_{ui} + \varepsilon_{uij}, \quad \varepsilon_{uij} \to N(0, \sigma_u^2),$$

$$\mu_{ui} = \mu_u + \alpha_{ui},$$

Todas las simulaciones se ejecutarán un total de unas 200 veces seguidas, dependiendo del caso y de la complejidad del análisis, los cálculos que se obtengan de cada una serán un promedio del total de replicaciones, garantizando que el comportamiento del modelo es representativo de un número amplio de casos y no es debido al azar, reduciendo así la incertidumbre asociada al método de Monte Carlo, problema ya comentado en el primer capítulo.

Se comenzará con una primera simulación en la que se fijará el tamaño muestral en 100 observaciones y se establecerá un tamaño del efecto similar en ambas variables dependientes, en este caso será un tamaño del efecto pequeño en comparación con la varianza.

3.1.1. Las dos variables dependientes con efecto del factor pequeño

El único parámetro que se variará será la correlación entre los términos de error de las variables dependientes, que irá desde -0.9 hasta 0.9 y se contrastarán los diferentes valores tanto del p-valor como del estadístico F para cada uno de los análisis.

Se ha garantizado que los datos cumplen con las hipótesis de ambos análisis. La independencia y normalidad se aseguran generando datos aleatorios mediante una distribución normal multivariante, usando el comando *mvrnorm()* del paquete *MASS* de RStudio. A esta función se le ha introducido entre otros de los parámetros de entrada, la matriz de covarianzas, estableciendo la misma para los distintos niveles y por tanto asegurando también la hipótesis de homocedasticidad, que además se ha comprobado utilizando la prueba de la M de Box, filtrando únicamente aquellas replicaciones que cumplan con esta hipótesis.

Con todo ello, se han llevado a cabo 200 replicaciones utilizando, para el caso de MANOVA, las distintas pruebas estadísticas comentadas, aunque se espera que el resultado sea el mismo en todas ellas, ya que el modelo cumple con todas las hipótesis más restrictivas. El conjunto de datos se ha generado con las siguientes características,

$$y_1\colon\ \mu_{11} = 2\ \ \mu_{12} = 4\ \ \sigma_1^2 = 100,$$

$$y_2\colon\ \mu_{21} = 4\ \ \mu_{22} = 6\ \ \sigma_2^2 = 100.$$

		r= −0.9	**r= −0.7**	**r= −0.5**	**r= −0.3**	**r= −0.1**	**r= 0.1**	**r= 0.3**	**r= 0.5**	**r= 0.7**	**r= 0.9**
ANOVAy1	F	2.114	1.815	1.777	2.057	1.888	2.058	1.956	1.916	2.041	1.978
	p	0.379	0.376	0.384	0.371	0.384	0.357	0.375	0.388	0.34	0.364
ANOVAy2	F	2.153	2.123	2.287	1.946	2.032	2.135	1.973	2.245	2.172	1.973
	p	0.337	0.365	0.347	0.393	0.388	0.358	0.365	0.349	0.335	0.367

Tabla 3.1. Resultados de los ANOVAs individuales efecto pequeño-pequeño

		r= −0.9	**r= −0.7**	**r= −0.5**	**r= −0.3**	**r= −0.1**	**r= 0.1**	**r= 0.3**	**r= 0.5**	**r= 0.7**	**r= 0.9**
Hotelling	F	11.751	4.333	3.024	2.435	2.085	2.015	1.733	1.715	1.701	1.461
	p	0.004	0.093	0.191	0.266	0.312	0.308	0.346	0.373	0.346	0.391
Pillai	F	11.751	4.333	3.024	2.435	2.085	2.015	1.733	1.715	1.701	1.461
	p	0.004	0.093	0.191	0.266	0.312	0.308	0.346	0.373	0.346	0.391
Roy	F	11.751	4.333	3.024	2.435	2.085	2.015	1.733	1.715	1.701	1.461
	p	0.004	0.093	0.191	0.266	0.312	0.308	0.346	0.373	0.346	0.391
Wilks	F	11.751	4.333	3.024	2.435	2.085	2.015	1.733	1.715	1.701	1.461
	p	0.004	0.093	0.191	0.266	0.312	0.308	0.346	0.373	0.346	0.391

Tabla 3.2. Resultados para el MANOVA efecto pequeño-pequeño

De la Tabla 3.2 se desprende como el valor de F y del p-valor (representado por la letra *p* en la tabla) es el mismo para todos los test utilizados y, por tanto, de ahora en adelante se trabajará usando únicamente el de Pillai. Se observa, además, como el valor de F en el MANOVA va

creciendo a medida que la correlación se acerca a -1, alejándose del que entrega el ANOVA. El valor de F es similar para ambos análisis cuando las correlaciones son cercanas a 0, a pesar de esto, el p-valor para estas correlaciones ya comienza a ser notablemente distinto. Esto se debe a que las distribuciones empleadas en cada uno de los análisis tienen distintos grados de libertad, mientras que la distribución F para la prueba de Pillai, que ya fue definida en el apartado de MANOVA, es de la forma

$$F_{s(2m+s+1),s(2n+s+1)},$$

$$s = \min(p, h), \qquad m = \frac{(|p-h|-1)}{2}, \qquad n = \frac{e-p-1}{2},$$

la del ANOVA se puede expresar como

$$F_{h,e.}$$

Donde h y e se recuerda que son los grados de libertad de la matriz hipótesis y error respectivamente. Para este caso, la distribución F en el MANOVA tiene 2 grados de libertad en el numerador y 97 en el denominador, mientras que en el ANOVA tiene 1 en el numerador y 98 en el denominador.

Como se ha dicho, el conjunto de variables se ha generado de forma que el tamaño del efecto sea pequeño, en este caso si se toma como medida de referencia η^2, cuya expresión era de la forma $\frac{|SSCP_b|}{|SS_t|}$, el valor para cada una de las variables es cercano a 0.01, es decir, las diferencias medidas explican el 1% de la variabilidad total del modelo, siendo el resto variabilidad residual. Por tanto, se esperaría que ambos análisis no encuentren diferencias significativas entre ambos niveles, sin embargo, en el caso de MANOVA no es así. Para correlaciones negativas la potencia de MANOVA aumenta a valores considerablemente altos, o lo que es lo mismo, la probabilidad de error tipo II disminuye.

Este suceso en MANOVA está ligado a la distancia de Mahalanobis (MD) entre los centroides de las dos variables dependientes. Para este caso particular de MANOVA en el que se dispone de dos variables dependientes y dos niveles en la variable independiente, se ha visto que se puede expresar el valor de F en función del de la T^2 de Hotelling y por tanto de MD,

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - p)2}\left(\frac{n_1 \times n_2}{n_1 + n_2}\right) MD^2,$$

siendo n_1 y n_2 en el número de observaciones en los grupos 1 y 2 respectivamente, que para este caso es de 50. MD se puede expresar en función del tamaño del efecto, medido en este caso como la diferencia de medias entre ambos grupos $(\Delta\bar{y}_u)$ y de la matriz de covarianzas (S) como sigue,

$$MD^2 = [\Delta\bar{y}_1\ \Delta\bar{y}_2]\mathrm{S}^{-1}\begin{bmatrix}\Delta\bar{y}_1\\ \Delta\bar{y}_2\end{bmatrix}$$

Si ${s_{11}}^{-1}, {s_{22}}^{-1}\ y\ {s_{12}}^{-1}$ representan los términos de la inversa de la matriz de covarianzas anterior y se considera que $\Delta\bar{y}_1 \approx \Delta\bar{y}_2$, ya que ambas variables se han generado de forma que las diferencias entre las medias de cada grupo sean de 2, se llega a la expresión

$$MD^2 = {\Delta\bar{y}_1}^2({s_{11}}^{-1} + 2{s_{12}}^{-1} + {s_{22}}^{-1}),$$

Donde *s* es la varianza muestral, teniendo en cuenta que $s_{11} \approx s_{22}$, los elementos de la matriz inversa se pueden expresar como

$$s_{11}{}^{-1} = {s_{11}^2}\Big/{|S|} = s_{22}{}^{-1},$$

$$s_{12}{}^{-1} = {-s_{12}^2}\Big/{|S|},$$

donde el determinante de la matriz equivale a

$$|S| = s_{11}{}^4 - s_{12}{}^4.$$

Sabiendo además que $s_{12}^2 = r s_{11}^2$, con r el coeficiente de correlación entre y_1 e y_2, entonces MD^2 se puede reescribir como

$$MD^2 = \frac{2\Delta \bar{y}_1{}^2}{s_{11}^2(1+r)}.$$

Como se observa en el resultado, al acercarse *r* a valores cercanos a -1 el valor del denominador crece y con ello la distancia de Mahalanobis aumenta, ocurriendo lo contrario para valores positivos de *r*. Gráficamente se puede observar un menor solapamiento entre los dos niveles cuando la correlación se hace negativa.

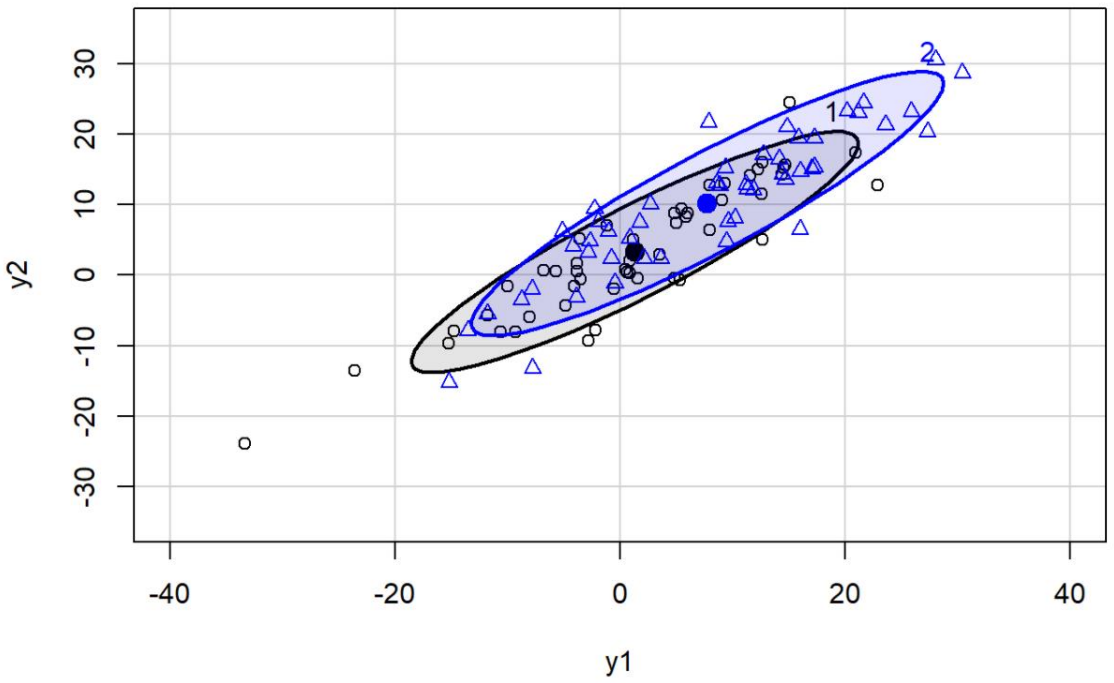


Figura 3.1. Elipse del 80% de confianza para $r = 0.9$

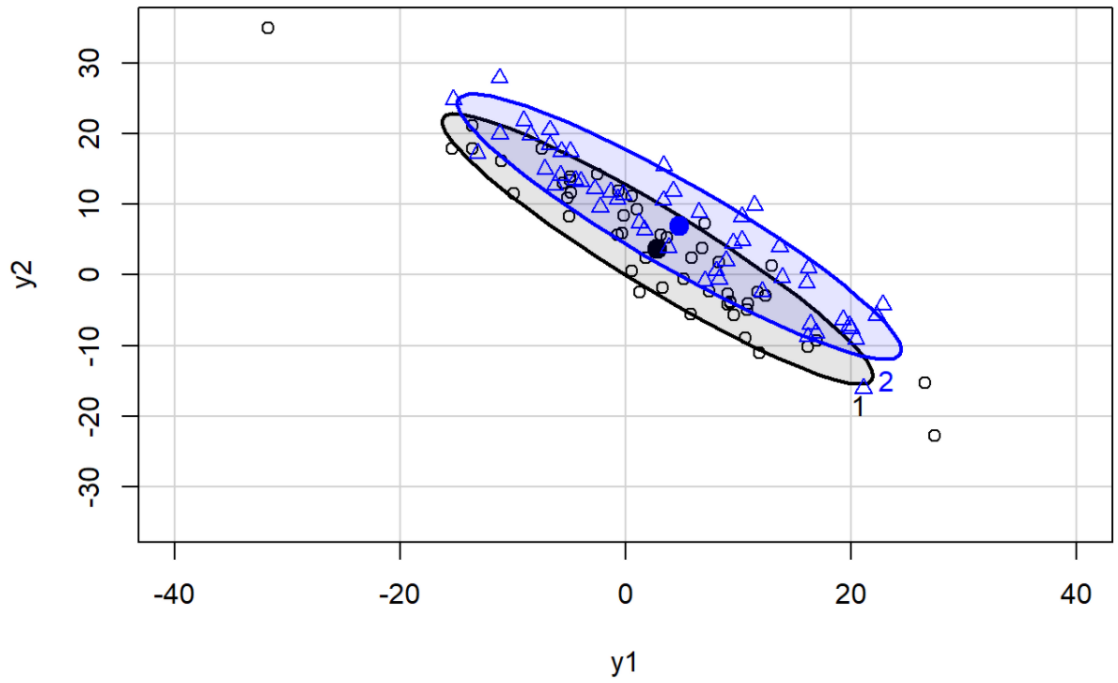


Figura 3.2. Elipse del 80% de confianza para $r = -0.9$

Las Figuras 3.1 y 3.2 reflejan la esencia de lo que se está tratando de analizar en este estudio de simulación, se indica como son las elipses de las nubes de puntos para una confianza del 80%, se ha elegido este valor porque el área de las elipses es más pequeña que para valores grandes de la confianza, siendo así más fácil de observarse las diferencias, ya que al estar usando tamaños del efecto pequeños es difícil apreciar la separación entre los dos niveles. En azul se representan los datos referentes al grupo 2 y en negro los del grupo 1. En la Figura 3.1, que corresponde a una correlación con un valor positivo alto se observa como el solapamiento de los datos es mucho mayor que cuando las correlaciones son negativas, sugiriendo por tanto una menor distancia. Esto que se acaba de enunciar es de vital importancia para comprender los resultados de los análisis.

Todo esto se ve reflejado en la diferencia en el valor del estadístico F de ambos análisis, que como ya se ha visto en la expresión anterior es directamente proporcional al valor de MD^2.

Por tanto, a valores negativos de la correlación, en un caso en el que se disponga de dos variables con un tamaño del efecto pequeño, la distancia entre los conjuntos de datos aumenta, lo que repercute sobre el valor de F y por tanto del p-valor, incrementando las diferencias entre el ANOVA y el MANOVA a medida que el valor de *r* se hace más pequeño.

Se puede medir la repercusión que esto tiene sobre las probabilidades de error de tipo I y II en el MANOVA. Con lo visto, la potencia de MANOVA, o la capacidad de aceptar la hipótesis alternativa cuando es cierta, debería aumentar para correlaciones negativas, se medirá como el porcentaje de veces del total de simulaciones en las que el p-valor del MANOVA es inferior al nivel de significación.

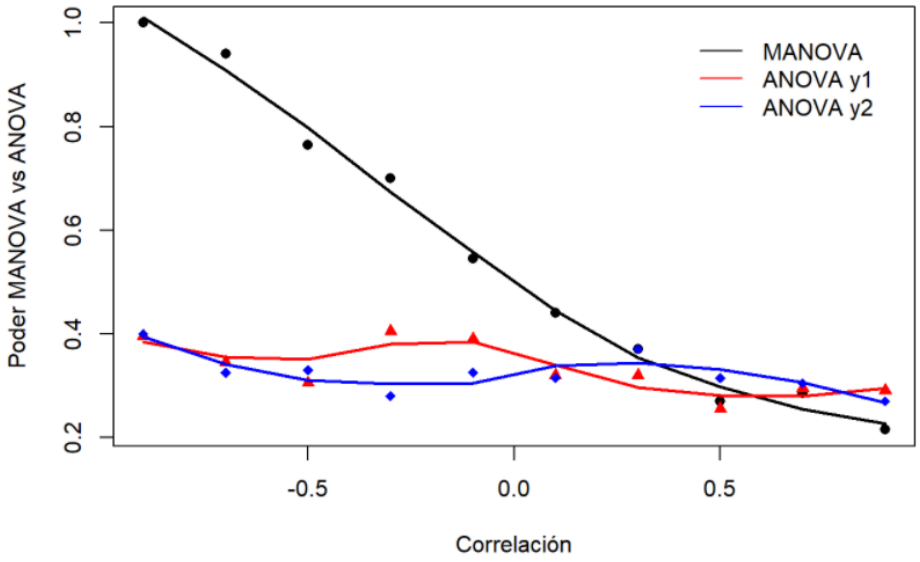


Figura 3.3. Comparación de la potencia de los análisis efecto pequeño-pequeño

En la Figura 3.3 se ha representado cómo evoluciona con la correlación la potencia (en determinados textos se expresa también como "poder") de los ANOVAs individuales y la del MANOVA utilizando un nivel de significación $\alpha = 0.05$. Se observa que los ANOVAs fluctúan debido únicamente a la aleatoriedad en la generación de datos, dando resultados similares al MANOVA para correlaciones grandes, en cambio, MANOVA tiene una tendencia decreciente a medida que aumenta la correlación ya que al aumentar el solapamiento entre los datos se dejan de percibir las diferencias y se considera más veces la hipótesis nula como la cierta.

En este modelo el valor de F en el ANOVA se puede expresar directamente a partir del de T^2, sabiendo que la distancia de Mahalanobis para el caso univariante tiene la forma

$$MD^2 = \frac{\Delta\bar{y}_u{}^2}{s_u^2},$$

donde u es la variable dependiente para la que se esté realizando el análisis y s_u su varianza, entonces F quedaría como

$$F = \left(\frac{n_1 \times n_2}{n_1 + n_2}\right)\frac{\Delta\bar{y}_u{}^2}{s_u^2}.$$

Como se observa, no depende de la relación existente entre las variables dependientes, de ahí que la potencia de los ANOVAs no presente ningún tipo de tendencia con variaciones en la correlación.

3.1.2. Una variable dependiente con efecto del factor grande y la otra pequeño

Se analizarán a continuación las diferencias que se producen cuando una de las variables dependientes tiene un tamaño del efecto grande y la otra pequeño. Se empleará el mismo modelo, con los mismos parámetros definidos anteriormente, cambiando únicamente el valor

de las diferencias entre las medias de los grupos formados en ambas variables dependientes, de forma que $\Delta\mu_1 = 0$ y $\Delta\mu_2 = 6$, se probará para varias correlaciones.

		r= −0.9	**r= −0.7**	**r= −0.5**	**r= −0.3**	**r= −0.1**	**r= 0.1**	**r= 0.3**	**r= 0.5**	**r= 0.7**	**r= 0.9**
ANOVA y1	F	0.922	1.009	1.053	0.836	0.999	1.11	1.34	1.231	1.038	1.113
	p	0.513	0.499	0.496	0.509	0.5	0.491	0.461	0.43	0.502	0.505
ANOVA y2	F	9.591	9.626	10.5	10.094	10.277	10.551	9.847	10.373	10.585	9.49
	p	0.04	0.051	0.029	0.033	0.029	0.043	0.031	0.036	0.034	0.041
MANOVA	F	24.533	9.49	7.083	5.921	5.7	5.877	6.243	7.069	10.607	25.555
	p	0	0.009	0.022	0.049	0.05	0.052	0.044	0.029	0.006	0

Tabla 3.3. Resultados ANOVA y MANOVA efecto grande-pequeño

Ahora los ANOVAs dan resultados diferentes, mientras que el perteneciente a la variable con efecto grande percibe mayores diferencias entre los niveles, el de la de efecto pequeño, que por simplificación se ha establecido como nulo se mantiene similar al caso anterior. Era el resultado esperado, ya que, al aumentar el tamaño del efecto en la variable y_2, aumenta la variabilidad del modelo que queda explicada por las diferencias entre los grupos, dando lugar a valores de F más grandes y, por tanto, p-valores bajos. En el MANOVA, sin embargo, el valor de F sigue una especie de parábola, para valores de $|r|$ grandes F crece y cuando *r* se acerca a 0, alcanza valores mínimos, comportamiento que va en contra de la intuición, ya que lo esperado es una cancelación de ruido en correlaciones negativas, pero no en positivas. Esto se puede observar de nuevo analíticamente en la variación de la distancia de Mahalanobis. En este caso, al tener tamaños del efecto diferentes, el valor de MD^2 quedaría de la siguiente forma

$$MD^2 = \Delta\bar{y}_1{}^2 s_{11}{}^{-1} + 2\Delta\bar{y}_1\Delta\bar{y}_2 s_{12}{}^{-1} + \Delta\bar{y}_2{}^2 s_{22}{}^{-1}.$$

Si se sigue considerando que $s_{11}^2 \approx s_{22}^2$ y se sustituyen los términos de la matriz inversa por su correspondiente valor, se llega a

$$MD^2 = \frac{\Delta\bar{y}_1{}^2 + \Delta\bar{y}_2{}^2 - 2r\Delta\bar{y}_1\Delta\bar{y}_2}{s_{11}^2(1-r^2)}.$$

Si además se tiene en cuenta que este caso se ha configurado de forma que $\Delta\bar{y}_1 \approx 0$,

$$\mathrm{MD}^2 = \frac{\Delta\bar{\mathrm{y}}_2{}^2}{s_{11}^2(1-\mathrm{r}^2)}.$$

Por tanto, el valor de MD^2 se verá aumentado a medida que $|r|$ se vaya acercando a 1. Si además se recuerda que para este caso específico F se puede relacionar con T^2, que depende directamente de MD^2, el valor de este se incrementará para valores de $|r|$ cercanos a 1, como se ha comprobado en la simulación. Se puede volver a observar gráficamente el efecto que tiene esto sobre la potencia del MANOVA.

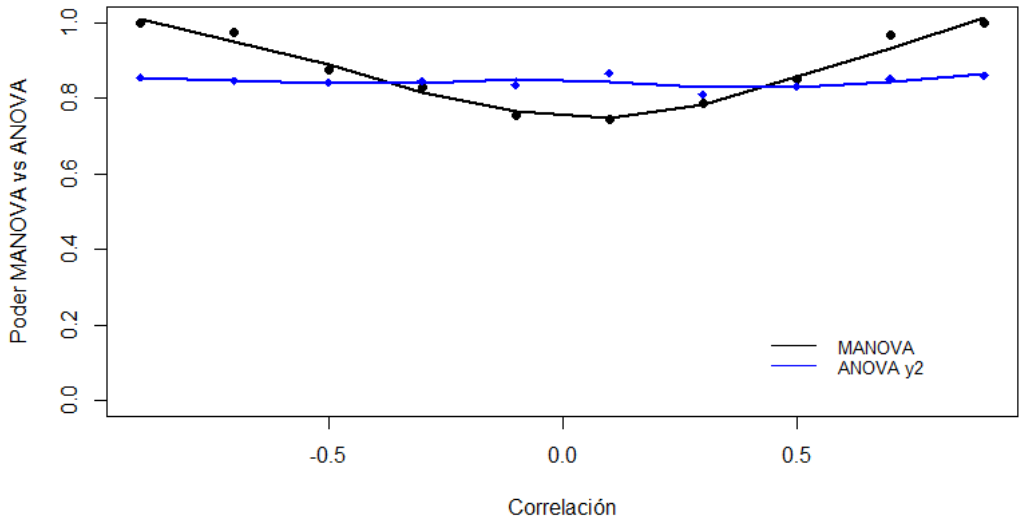


Figura 3.4. Comparación de la potencia de los análisis efecto pequeño-grande

En la Figura 3.4 se ve como varía la potencia en función de la correlación, se aprecia como al llegar a correlaciones cercanas a 0 el MANOVA empieza a aceptar más veces la hipótesis nula debido al valor más bajo de F, mientras que en los extremos la potencia aumenta. La potencia del ANOVA de y_2 se mantiene cercana al 0.8 y la de y_1 no se ha representado ya que no se han establecido diferencias entre los niveles y por tanto la hipótesis nula es la que se cumple para esa variable.

3.1.3. Las dos variables dependientes con efecto del factor grande

Para el caso en el que las dos variables dependientes tengan un tamaño del efecto grande, los resultados esperados son similares a lo que se ha visto para las dos variables de efecto pequeño, salvo que, en este caso, en los ANOVAs, el valor del estadístico F debería aumentar dando lugar a p-valores más bajos. Se ha ejecutado la simulación asignando un tamaño del efecto a cada una de las variables igual al de la de efecto grande de la anterior simulación, diferenciándose ambas en el valor de su media global. El valor de σ que se está utilizando es el mismo en todas las simulaciones, $\sigma^2 = 100$ para las dos variables.

$$y1:\ \mu_1 = 5,\ \alpha_{11} = -3,\ \alpha_{12} = 3,$$

$$y2:\ \mu_2 = 7,\ \alpha_{21} = -3,\ \alpha_{22} = 3.$$

En la Tabla 3.4 se comprueba que el resultado es el esperado, siendo el MANOVA más capaz de detectar las diferencias entre las variables cuando el valor de la correlación es cercano a -1, mientras que a medida que crece hacia valores positivos ocurre lo contrario. En este caso al estar trabajando con tamaños del efecto grandes se aprecia mejor como cambia el valor del estadístico F que en el caso en el que se empleaban tamaños pequeños.

		r= −0.9	r= −0.7	r= −0.5	r= −0.3	r= −0.1	r= 0.1	r= 0.3	r= 0.5	r= 0.7	r= 0.9
ANOVA y1	F	10.64	10.69	9.279	10.6	9.369	9.923	9.949	10.93	10.89	10.37
	p	0.042	0.026	0.046	0.028	0.03	0.036	0.039	0.027	0.032	0.031
ANOVA y2	F	9.722	10.88	10.99	9.755	10.60	10.83	9.373	10.45	10.01	10.28
	p	0.038	0.033	0.023	0.043	0.025	0.039	0.038	0.033	0.033	0.031
MANOVA	F	90.82	33.3	19.63	14.09	11.01	9.608	7.67	7.52	6.543	5.84
	p	0	0	0	0.001	0.003	0.011	0.021	0.024	0.039	0.048

Tabla 3.4. Resultados ANOVA y MANOVA efecto grande-grande

Para esta simulación se observa que el p-valor está por debajo de un nivel de significación $\alpha = 0.05$, pero a medida que aumenta la correlación se va acercando a este. Se puede observar gráficamente en la Figura 3.5 la reducción de la potencia de MANOVA con valores positivos en la correlación, siendo superado por la de los ANOVAs a partir de un cierto umbral.

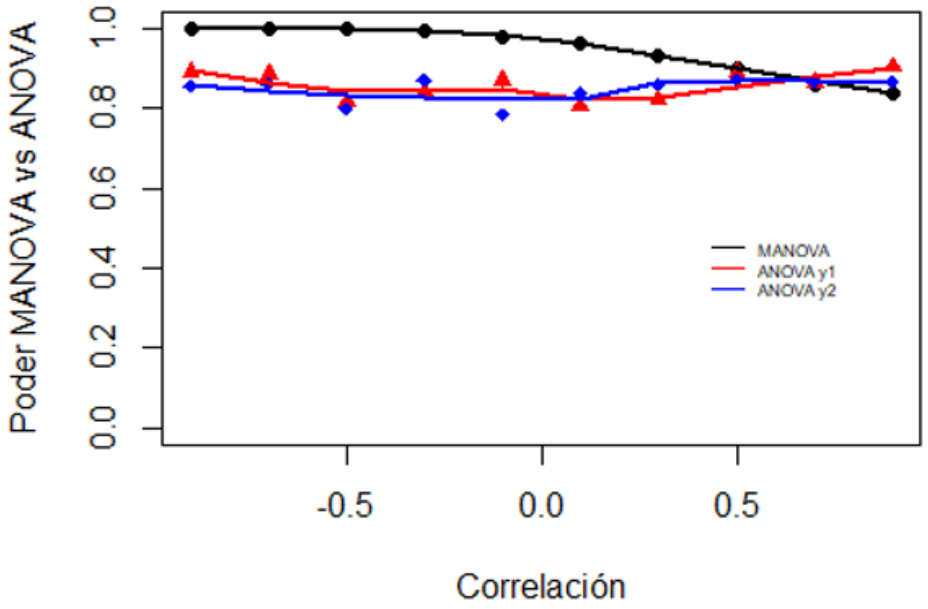


Figura 3.5. Comparación de la potencia de los análisis efecto grande-grande

Con este último experimento se habrían analizado las diferencias en los resultados proporcionados por ambos análisis dependiendo tanto del tamaño del efecto como de la correlación. Pero como se ha comentado, también se quiere ver qué papel desempeña un cambio en el tamaño muestral de los grupos, que en estas simulaciones se ha mantenido fijo en 50 por grupo, por tanto, en esta última simulación del MANOVA de un factor se analizará el efecto en el cambio del tamaño muestral.

3.1.4. Tamaño muestral

Para observar el comportamiento de los análisis se volverá a emplear el modelo definido en el primer apartado

$$y_1\colon\ \mu_{11} = 2,\ \ \mu_{12} = 4,\ \ \sigma_1^2 = 100,$$

$$y_2\colon\ \mu_{21} = 4,\ \ \mu_{22} = 6,\ \ \sigma_2^2 = 100,$$

es decir, un tamaño del efecto pequeño para ver mejor como aumenta o disminuye p-valor con el número de muestras, ya que para efectos grandes el valor observado será prácticamente 0 en todas las ocasiones.

		r= −0.9	**r= −0.7**	**r= −0.5**	**r= −0.3**	**r= −0.1**	**r= 0.1**	**r= 0.3**	**r= 0.5**	**r= 0.7**	**r= 0.9**
ANOVA y1	F	10.229	10.211	11.219	11.572	10.853	11.269	10.649	11.5	11.088	11.051
	p	0.031	0.044	0.032	0.031	0.029	0.029	0.024	0.023	0.021	0.024
ANOVA y2	F	11.696	11.816	11.292	11.012	10.619	11.153	10.585	11.748	11.144	11.187
	p	0.019	0.019	0.028	0.028	0.019	0.027	0.017	0.03	0.022	0.026
MANOVA	F	102.34	34.159	21.424	15.634	11.779	10.304	8.345	8.033	6.969	6.283
	p	0	0	0	0	0.001	0.005	0.011	0.018	0.024	0.038

Tabla 3.5. Resultados para un tamaño muestral, nmues=1000

		r= −0.9	**r= −0.7**	**r= −0.5**	**r= −0.3**	**r= −0.1**	**r= 0.1**	**r= 0.3**	**r= 0.5**	**r= 0.7**	**r= 0.9**
ANOVA y1	F	1.817	1.825	1.971	2.173	2.01	2.086	2.501	2.213	1.972	1.93
	p	0.393	0.391	0.378	0.395	0.386	0.361	0.314	0.362	0.365	0.386
ANOVA y2	F	2.379	1.863	1.828	2.135	1.824	2.187	2.286	2.255	1.792	2.079
	p	0.309	0.386	0.379	0.373	0.384	0.343	0.358	0.342	0.386	0.349
MANOVA	F	12.132	4.161	2.828	2.565	2.025	2.044	2.117	1.809	1.577	1.653
	p	0.002	0.12	0.211	0.256	0.304	0.288	0.287	0.348	0.381	0.361

Tabla 3.6. Resultados para un tamaño muestral, nmues=100

En las Tablas 3.5 y 3.6 se muestran los resultados proporcionados por RStudio para un tamaño muestral de 1000 y 100 muestras respectivamente. Claramente se aprecia una disminución en el p-valor cuando el número de muestras es más alto y, por tanto, un aumento del estadístico F.

Recordando la fórmula del valor de F en MANOVA para el caso en el que se tenga dos variables con tamaño de efecto pequeño se tiene que

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - p)2} T^2.$$

En este caso se está trabajando con $n_1 = n_2$ y además se puede considerar que el valor de p es despreciable frente al número de muestras, aproximación más válida cuanto mayor sea el tamaño muestral, por tanto, expresando T en función de MD

$$F = \frac{\Delta\bar{y}_1{}^2 n_1}{2s_{11}^2(1 + r)},$$

para ANOVA quedaría de forma similar

$$MD^2 = \frac{\Delta\bar{y}_1{}^2}{s_{11}^2},$$

$$F = T^2 = \frac{n_1 \times n_2}{n_1 + n_2} MD^2 = \frac{n_1 \Delta\bar{y}_1{}^2}{2s_{11}^2}.$$

Al aumentar el tamaño muestral ocurren dos efectos contrapuestos, en primer lugar, se produce un aumento en el tamaño del efecto, ya que si se tienen dos grupos diferenciados entre sí, a mayores tamaños muestrales más representativa será la muestra sujeta a estudio y por tanto mayores serán las evidencias de existencia de diferencias entre sí, sin embargo, también se está incrementando la correlación entre las variables, ya que al trabajar con un mayor número de datos, las variables dependientes pasan a estar más relacionadas entre sí. Esto puede tener un efecto perjudicial sobre la potencia cuando se dan correlaciones positivas altas, como se ha visto en dos de las tres simulaciones que se han llevado a cabo. De la combinación de estos dos efectos resulta una ganancia neta de potencia según los resultados observados en las tablas, que se ve claramente como el valor del estadístico F aumenta en la simulación para un tamaño muestral de 1000.

Hay que tener en cuenta que además de variar el valor de F se están cambiando los grados de libertad de la distribución y por tanto el p-valor debería ser distinto para iguales valores del estadístico F en ambos análisis ya que la forma de la distribución cambia, consecuencia del aumento en el número de muestras.

En este caso, en el que número de grados de libertad del denominador de la distribución F es muy alto para ambos tamaños muestrales y el del numerador es bajo, apenas se aprecian diferencias en su función de densidad y por tanto se puede considerar que el p-valor será prácticamente idéntico ante iguales valores de F en ambos modelos. En la Figura 3.6 se ha representado la función de densidad para los dos tamaños muestrales con los que se está trabajando para el caso de MANOVA, en ANOVA sería prácticamente igual ya que las diferencias entre ambas distribuciones no son grandes[1] .

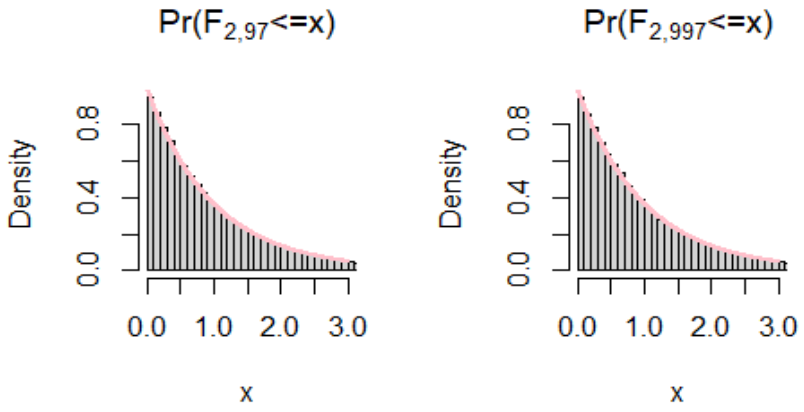


Figura 3.6. Funciones de densidad de F en MANOVA para nmues=100 y nmues=1000

[1] En MANOVA, se tienen 2 grados de libertad en el numerador para ambos tamaños muestrales, mientras que en el denominador para 100 muestras hay 97 y para 1000 muestras, 997. En ANOVA es 1 grado de libertad en el numerador en ambos casos, 98 en el denominador para 100 muestras y 998 para 1000.

En las gráficas superiores se aprecian las similitudes entre las funciones de densidad de ambas distribuciones[2], por tanto, se puede considerar que un aumento en el número de observaciones con las que se trabaja da lugar a una ganancia neta de potencia tanto en MANOVA como en ANOVA.

En la Figura 3.7 se ha representado el efecto del número de muestras sobre la potencia para cada uno de los análisis con los datos especificados para este apartado. Se observa como la potencia de estos crece considerablemente con el tamaño muestral. De la gráfica se puede deducir que la capacidad de discriminar entre los grupos del MANOVA es mucho superior a la de los ANOVAs si se trabaja con correlaciones negativas, tamaños muestrales pequeños y tamaños del efecto de las variables pequeño.

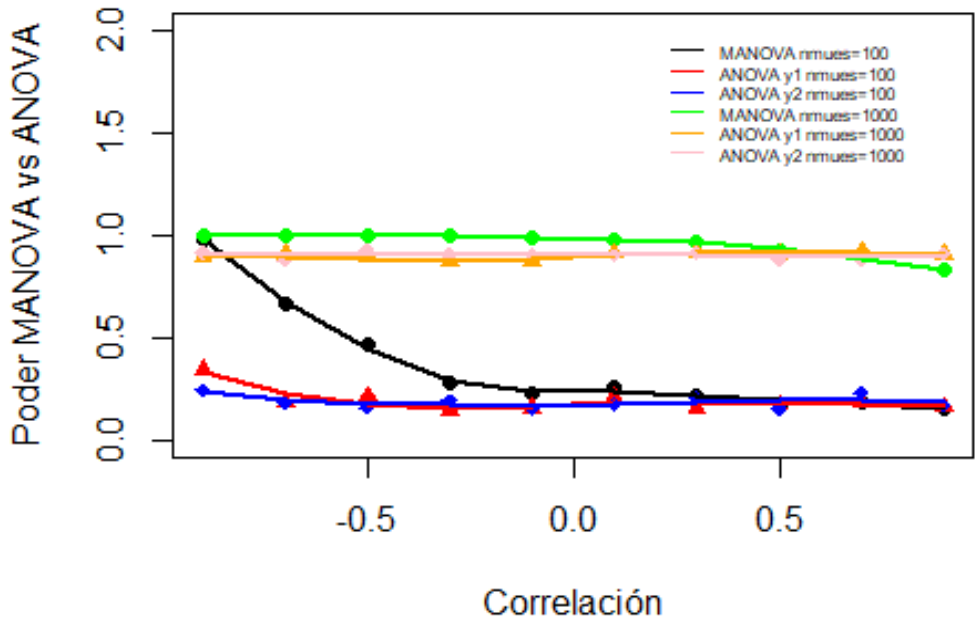


Figura 3.7. Potencia de los análisis en función del tamaño muestral

El siguiente paso en el estudio de simulación sería estudiar el comportamiento de ambos análisis en casos algo más complejos y representativos de muchos procesos reales. Se trasladarán las conclusiones extraídas de esta sección a una comparación entre una serie de ANOVAs individuales y un MANOVA de dos factores.

3.2. MANOVA de dos factores

En este segundo experimento se pasará a trabajar con dos variables independientes en vez de una. Estas variables independientes estarán formadas cada una por tres niveles distintos y el

[2] Notar que, aunque el valor de la función de densidad es superior a 1 en un intervalo pequeño de x cercano a 0, esto se debe al proceso de generación de los datos con RStudio, la función de densidad se genera de tal forma que $\int_{\mathbb{R}^+} f(x)dx = 1$ y, por tanto, puede ser que en un intervalo pequeño de x $f(x) > 1$.

número de variables dependientes[3] pasará a ser de tres. Como se ha comentado, lo que se busca es extender las conclusiones de un experimento sencillo a uno más complejo y observar si las reglas definidas para el MANOVA de un factor son propias también de casos más completos.

Como ya se ha mencionado anteriormente, las variables dependientes, en el caso en el que se trabaje con dos factores, se pueden modelar como

$$y_{uijk} = \mu_u + \alpha_{ui} + \beta_{uj} + \alpha\beta_{uij} + \varepsilon_{uijk}, \;\; \varepsilon_{uijk} \to N(0, \sigma_u^2),$$

donde μ_u sigue siendo la media global de la variable *u*, α_{ui} la desviación en la variable *u* con respecto a la media global debido al efecto del grupo *i* del primer factor y β_j la desviación asociada al otro factor. Para este experimento se trabajará con interacción entre las variables independientes, efecto representado por $\alpha\beta_{uij}$, es decir, la interacción entre los niveles *i* y *j* de los dos factores en la variable *u*.

Se volverá a correlacionar las tres variables a partir de los términos de error y junto con el tamaño del efecto se analizarán las diferencias entre los ANOVAs y el MANOVA para cada uno de los casos. El número de replicaciones que se hará de cada una de las simulaciones será de unas 300 mínimo, dependiendo del caso, ya que al aumentar la cantidad de datos que se están generando, la variabilidad que se introduce es mayor, de esta forma se busca un balance entre coste computacional y rigor estadístico. Al igual que antes, los valores que se representen y los cálculos realizados se harán sobre el promedio de todas las replicaciones.

Ahora al trabajar con tres variables las correlaciones se fijarán de forma que una de las tres variables tenga una correlación fija con las dos variables cuyo valor de la correlación se modificará. El tamaño del efecto de las dos variables de las que se irá cambiando el valor de la correlación se escogerá dependiendo del caso en el que se esté y en la tercera se fijará otro.

3.2.1. Cambio en la correlación de las variables con efecto grande de los factores

En este nuevo experimento se comenzará analizando las diferencias que se dan entre ambos análisis cuando se modifica la correlación entre las variables con un tamaño del efecto grande en comparación con la varianza residual. Para ello, los datos se han generado de acuerdo a las siguientes características,

$$\mu_1 = 4, \;\; \mu_2 = 5, \;\; \mu_3 = 8,$$

$$\alpha_1 = [2 \quad 1 \quad -3], \;\; \alpha_2 = [2 \quad 1 \quad -3], \;\; \alpha_3 = [0 \quad 0 \quad 0],$$

$$\beta_1 = [-3 \quad 1 \quad 2], \;\; \beta_2 = [-3 \quad 1 \quad 2], \;\; \beta_3 = [0 \quad 0 \quad 0],$$

$$\alpha\beta_1 = \begin{bmatrix} 3 & 1 & -4 \\ 1 & -3 & 2 \\ -4 & 2 & 2 \end{bmatrix}, \;\; \alpha\beta_2 = \begin{bmatrix} 3 & 1 & -4 \\ 1 & -3 & 2 \\ -4 & 2 & 2 \end{bmatrix}, \;\; \alpha\beta_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\sigma_1^2 = 100, \;\; \sigma_2^2 = 100, \;\; \sigma_3^2 = 100,$$

[3] Se ha intentado también con 4, pero en la mayoría de las simulaciones la matriz (E+H) era singular y al ser necesario trabajar con su inversa en MANOVA, se obtenía para muchas de las correlaciones errores en el código.

donde el subíndice hace referencia a la variable dependiente a la que pertenece ese parámetro. Como se observa, los parámetros se generan cumpliendo que

$$\sum_{i}^{I} \alpha_{ui} = 0, \forall u, \qquad \sum_{j}^{J} \beta_{uj} = 0, \forall u, \qquad \sum_{i}^{I} \alpha\beta_{uij} = 0, \forall j, u, \qquad \sum_{j}^{J} \alpha\beta_{uij} = 0, \ \ \forall i, u.$$

Se supondrá que el tamaño del efecto de los factores sobre las variables y_1 e y_2 es el mismo en todos los niveles, que representarán las variables con efecto grande, mientras que y_3 se ha escogido de forma que su tamaño del efecto sea nulo y, por tanto, represente a una variable de efecto pequeño, es equivalente a suponer que las únicas diferencias entre los distintos niveles es debido a la parte aleatoria.

Se realizará una primera simulación del nuevo experimento con estos parámetros, se comenzará trabajando con un conjunto muestral de 180 observaciones, 20 por cada grupo sometido a los pares de tratamientos correspondientes. Se fijará la correlación de la variable y_3 con y_1 e y_2 en 0.3. La correlación[4] de las otras dos variables, las de tamaño del efecto grande, se variará de -0.7 a 0.7, con todo lo definido, se procede a ejecutar la primera simulación.

		$r=-0.7$	$r=-0.5$	$r=-0.3$	$r=-0.1$	$r=0.1$	$r=0.3$	$r=0.5$	$r=0.7$
ANOVA y1	F	5.28	5.151	5.245	5.109	5.023	5.414	5.418	5.463
	p	0.062	0.068	0.064	0.075	0.071	0.07	0.051	0.058
ANOVA y2	F	5.054	5.428	5.293	4.956	5.084	5.435	5.497	5.388
	p	0.075	0.056	0.063	0.068	0.066	0.057	0.049	0.056
ANOVA y3	F	1.004	1.007	1.05	1.001	1.057	1.075	0.996	1.006
	p	0.5	0.502	0.474	0.501	0.486	0.487	0.504	0.501
MANOVA	F	17.48	8.569	5.881	4.43	3.748	3.491	3.116	2.847
	p	0	0	0.002	0.009	0.019	0.035	0.052	0.073

Tabla 3.7. Resultados ANOVA y MANOVA cambiando la correlación en las variables con tamaño del efecto grande para el efecto del factor 1

En la Tabla 3.7 se han representado los resultados obtenidos para el efecto del factor 1, el comportamiento que se observa vuelve a ser igual al del anterior experimento, en los ANOVAs de las dos variables con tamaño del efecto grande se dan p-valores más bajos que en el de la variable de efecto pequeño. En MANOVA se tienen valores del p-valor decrecientes a medida que disminuye la correlación, ocurriendo lo contrario para F, que aumenta su valor a medida que la correlación se acerca a -1.

Este efecto se puede volver a expresar a partir de la distancia de Mahalanobis, pero en este caso hay que tener en cuenta que se está trabajando con tres niveles diferentes y por tanto se tienen tres distancias distintas, la distancia entre los niveles 1 y 2, 1 y 3 y entre 2 y 3. Además, al estar

[4] Se ha tratado de emplear valores más extremos, al igual que en el anterior experimento, sin embargo, a la hora de generar los términos de error correlacionados con la distribución normal multivariante, se requería que la matriz de covarianzas fuese definida positiva, lo que se ha visto impedido para valores tan extremos.

trabajando con dos variables independientes hay que separar las distancias en función del factor que se vaya considerar, o considerando ambos en caso de que se quiera analizar el efecto de la interacción.

Para este estudio se analizarán las diferencias dadas en el factor 1 entre los niveles 1 y 2. El razonamiento matemático que hay que aplicar es el mismo que en el experimento anterior, teniendo en cuenta las modificaciones debidas a los aspectos comentados. La distancia de Mahalanobis para esta prueba se definiría como

$$MD^2 = [\Delta\bar{y}_1 \quad \Delta\bar{y}_2 \quad \Delta\bar{y}_3] \begin{bmatrix} {s_{11}}^{-1} & {s_{12}}^{-1} & {s_{13}}^{-1} \\ {s_{21}}^{-1} & {s_{22}}^{-1} & {s_{23}}^{-1} \\ {s_{31}}^{-1} & {s_{32}}^{-1} & {s_{33}}^{-1} \end{bmatrix} \begin{bmatrix} \Delta\bar{y}_1 \\ \Delta\bar{y}_2 \\ \Delta\bar{y}_3 \end{bmatrix}.$$

Hay que tener en cuenta que el tamaño del efecto para y_3 se ha establecido como nulo y el de las variables y_1 e y_2 es el mismo, se puede considerar que $\Delta\bar{y}_1 \approx \Delta\bar{y}_2 = \Delta\bar{y}$. Además, se está trabajando con la misma varianza en las tres variables, por tanto, $s_{11}^2 \approx s_{22}^2 \approx s_{33}^2 = s^2$. Por último, las correlaciones de la variable con tamaño del efecto pequeño con las variables con efecto grande se han fijado de forma que sean idénticas, $r_{y1y3} \approx r_{y2y3} = r_{gp}$.

Si se opera en la anterior expresión se llega a

$$MD^2 = \Delta\bar{y}^2({s_{11}}^{-1} + {s_{12}}^{-1} + {s_{21}}^{-1} + {s_{22}}^{-1}),$$

sustituyendo los términos de la matriz inversa por su correspondiente valor

$${s_{11}}^{-1} = s^4(1 - {r^2}_{gp})/|\mathrm{S}|,$$

$${s_{12}}^{-1} = s^4({r^2}_{gp} - r_{y1y2})/|\mathrm{S}|,$$

$${s_{22}}^{-1} = s^4(1 - {r^2}_{gp})/|\mathrm{S}|,$$

al estar trabajando con una matriz simétrica regular, su inversa sigue siendo simétrica, de forma que ${s_{12}}^{-1} = {s_{21}}^{-1}$. El determinante de la matriz de covarianzas es

$$|\mathrm{S}| = s^6(1 + 2r_{y1y2}{r^2}_{gp} - 2{r^2}_{gp} - {r^2}_{y1y2}),$$

reagrupando términos del determinante de la matriz

$$|\mathrm{S}| = s^6(1 - r_{y1y2})(1 + r_{y1y2} - 2{r^2}_{gp}).$$

Remplazando todo lo anterior en la expresión de MD^2 se obtiene

$$MD^2 = \frac{2s^4(1 - r_{y1y2})\Delta\bar{y}^2}{s^6(1 - r_{y1y2})(1 + r_{y1y2} - 2{r^2}_{gp})},$$

$$MD^2 = \frac{2\Delta\bar{y}^2}{s^2(1 + r_{y1y2} - 2{r^2}_{gp})}.$$

Al ser MD^2 una expresión positiva y su numerador también, el denominador será de igual manera positivo, por tanto, se observa que un aumento en sentido positivo de la correlación entre las variables con efecto grande da lugar a menores diferencias entre los grupos, el mismo comportamiento que se tenía para el MANOVA de un factor. En este nuevo experimento deja de ser aplicable la T^2 de Hotelling ya que se está trabajando con dos factores con más de dos

niveles cada uno. No obstante, para el cálculo de F se sigue empleando la variabilidad explicada por cada uno de los factores y la residual, de tal manera que, a mayor distancia entre los niveles, mayor será la variabilidad entre ellos, lo que se verá reflejado en las matrices de hipótesis y de error, dando lugar a distintos autovalores en función de la separación entre grupos.

Esto se puede volver a observar de forma gráfica, extendiendo el caso a un modelo con tres dimensiones, pasando de trabajar con elipses a utilizar elipsoides. Para la representación gráfica se han incrementado las distancias entre los niveles debidas al efecto del factor 1, de forma que sea más fácil apreciar los cambios en la nube de puntos para cada una de las correlaciones.

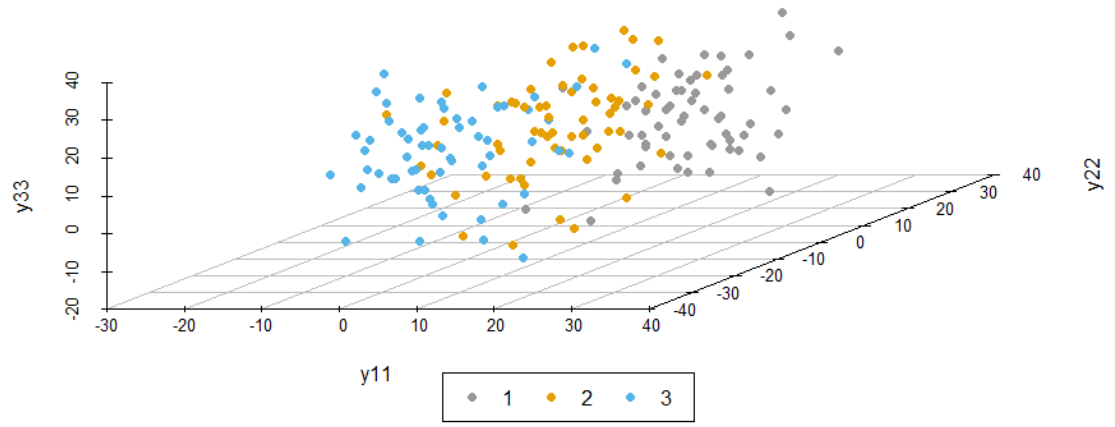


Figura 3.8. Scatterplot para el efecto del factor 1 $r_{y1y2} = -0.7$

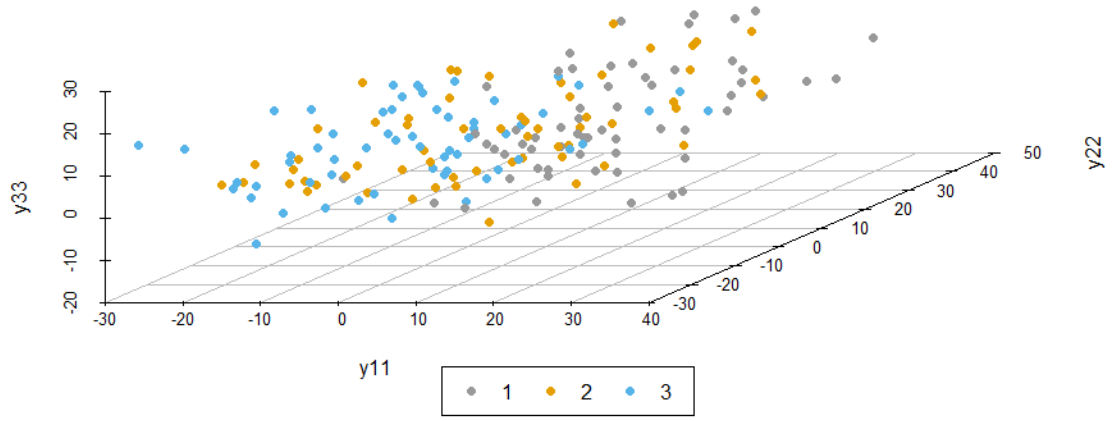


Figura 3.9. Scatterplot para el efecto del factor 1 $r_{y1y2} = 0.7$

En las Figuras 3.8 y 3.9 se han representado los casos más extremos examinados. Se puede apreciar una mayor separación entre los tres niveles cuando la correlación entre las variables de

efecto grande es negativa, lo que tiene sentido según los cálculos que se han llevado a cabo. Si se proyecta la figura sobre los ejes y_1 e y_2 se tendría de nuevo las elipses con las que se explicó este comportamiento en el MANOVA de un factor.

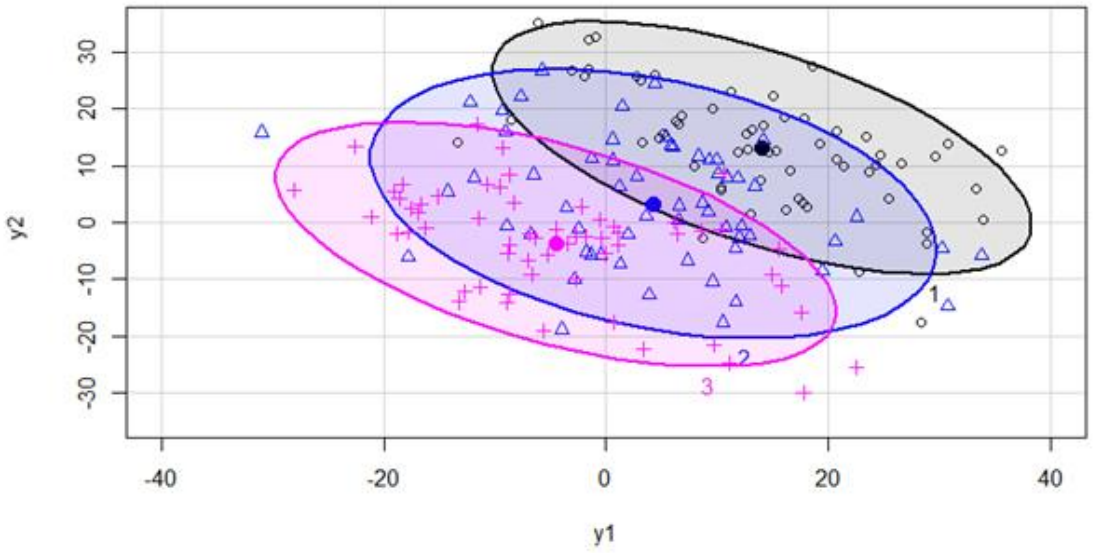


Figura 3.10. Elipses del 80% de confianza para el efecto del factor 1, $r_{y1y2} = -0.7$

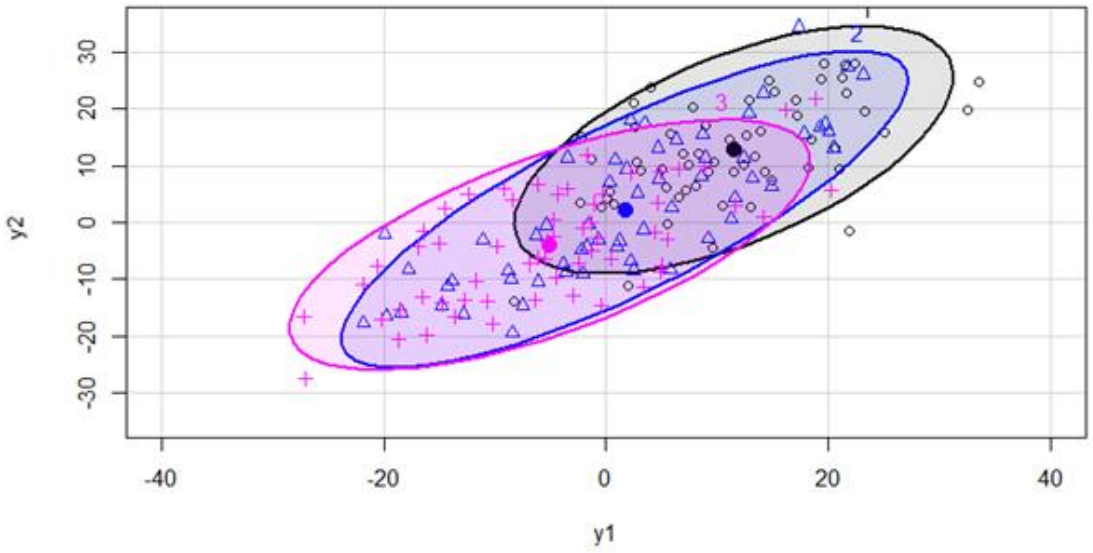


Figura 3.11. Elipses del 80% de confianza para el efecto del factor 1, $r_{y1y2} = 0.7$

Las elipses muestran el mismo comportamiento que el antes visto, extendido a un caso en el que se está trabajando con tres niveles y por tanto tres elipses diferentes, mientras que en la figura superior se observa una separación entre los niveles, en la inferior hay un claro solapamiento entre los tres.

Se representará, al igual que en el MANOVA de un factor, cómo varía la potencia para el efecto del factor 1 en MANOVA, comparada con la de los ANOVAs individuales.

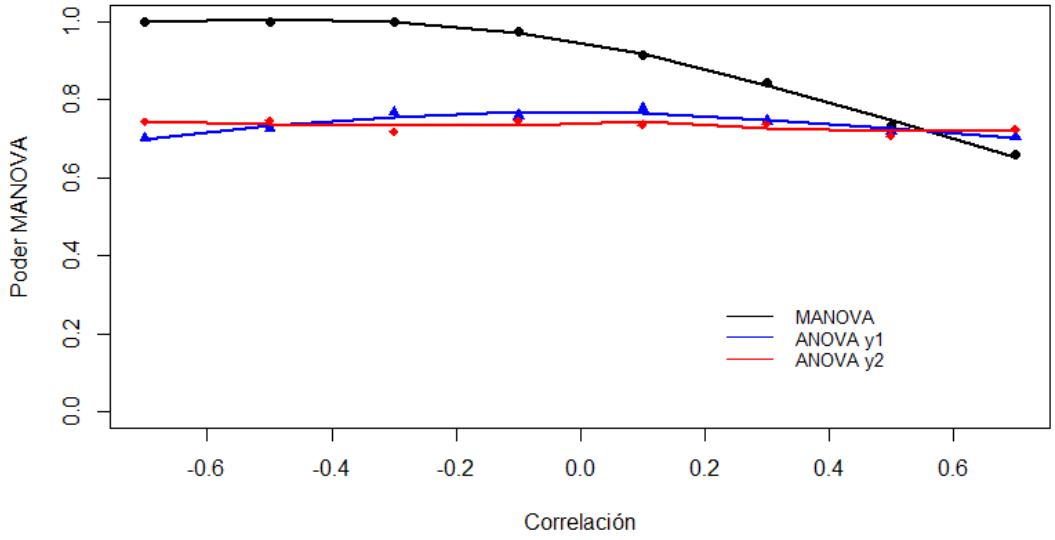


Figura 3.12. Potencia MANOVA y ANOVA factor 1, en función de r_{y1y2} para tamaño del efecto grande

En la Figura 3.12 se muestra gráficamente la disminución de la potencia de MANOVA, que para valores de *r* cercanos a 1 queda por debajo de la de los ANOVAs, lo esperado tras ver los valores de F y del p-valor en la tabla.

Las conclusiones obtenidas en esta primera simulación se han centrado en el efecto del factor 1, por tanto, sería interesante comprobar que ocurre algo parecido con la interacción o con el factor 2.

		$r=-0.7$	$r=-0.5$	$r=-0.3$	$r=-0.1$	$r=0.1$	$r=0.3$	$r=0.5$	$r=0.7$
ANOVA y1	F	4.324	3.987	4.155	4.371	4.052	3.994	4.087	4.334
	p	0.037	0.038	0.039	0.035	0.045	0.048	0.043	0.03
ANOVA y2	F	4.122	4.429	4.197	4.152	4.217	4.298	4.17	4.411
	p	0.037	0.033	0.038	0.042	0.041	0.039	0.038	0.036
ANOVA y3	F	1.062	1.027	1.058	1.023	1.016	1.026	1.044	0.987
	p	0.481	0.494	0.477	0.5	0.5	0.502	0.489	0.493
MANOVA	F	10.663	6.022	4.408	3.541	3.032	2.678	2.47	2.325
	p	0	0	0	0.003	0.014	0.02	0.031	0.046

Tabla 3.8. Resultados ANOVA y MANOVA cambiando la correlación en las variables con tamaño del efecto grande para el efecto de la interacción

En la Tabla 3.8 se han recogido los resultados para el mismo modelo analizando el efecto de la interacción. De nuevo se vuelve a apreciar como los ANOVAs de las variables con un tamaño de efecto grande dan lugar a p-valores bajos en comparación con el de la variable de efecto pequeño. En el resultado del MANOVA se observa que el p-valor crece a medida que la correlación aumenta a valores positivos, por tanto, se puede considerar que lo estudiado para el factor 1 se extiende al efecto de la interacción.

3.2.2. Cambio en la correlación de las variables con efecto pequeño de los factores

A continuación, se pasa a discutir el caso en el que las dos variables en las que se cambie la correlación sean las de efecto pequeño, se supondrá para este apartado que el tamaño del efecto de y_2 se mantiene constante, representado una variable de efecto grande y se pone a 0 el de y_1, y_3 se mantendrá igual que en el caso anterior. Se realizará la simulación y se analizarán las diferencias.

		$r=-0.7$	$r=-0.5$	$r=-0.3$	$r=-0.1$	$r=0.1$	$r=0.3$	$r=0.5$	$r=0.7$
ANOVA y1	F	1.046	1.007	0.98	1.054	1.014	1.037	0.963	0.942
	p	0.494	0.5	0.501	0.493	0.502	0.493	0.507	0.518
ANOVA y2	F	5.224	5.313	5.115	5.16	5.186	5.557	5.502	5.096
	p	0.057	0.067	0.053	0.062	0.067	0.053	0.051	0.06
ANOVA y3	F	1.055	0.936	0.967	0.944	0.995	0.986	0.966	0.94
	p	0.484	0.514	0.509	0.514	0.498	0.5	0.499	0.524
MANOVA	F	4.375	3.023	2.755	2.654	2.623	2.663	2.573	2.464
	p	0.01	0.062	0.069	0.083	0.091	0.085	0.091	0.107

Tabla 3.9. Resultados ANOVA y MANOVA cambiando la correlación en las variables de efecto pequeño para el efecto del factor 1

En la Tabla 3.9 se vuelve a ver la tendencia creciente del p-valor de MANOVA cuando la correlación entre las variables de efecto pequeño aumenta. La expresión analítica de MD^2 en este caso sería

$$MD^2 = \Delta\bar{y}^2 {s_{22}}^{-1},$$

$${s_{22}}^{-1} = s^4(1 - r_{y1y3}^2)/|\mathrm{S}|.$$

La expresión |S| es la misma que en el anterior caso, pero teniendo en cuenta que ahora r_{gp} será la correlación entre las variables y_1 e y_2, que será igual que la de y_3 con y_2, por tanto,

$$MD^2 = \frac{\Delta\bar{y}^2(1 + r_{y1y3})}{s^2(1 + r_{y1y3} - 2{r_{gp}}^2)},$$

$\Delta\bar{y}$ representa el tamaño del efecto de la variable y_2. Si se considera otro conjunto de tres variables con iguales tamaños del efecto que en el caso definido, misma correlación entre

variables de efecto grande y pequeño y varianzas, pero distinta correlación en las variables de efecto pequeño y se comparan sus distancias se obtiene la expresión

$$MD^2 - MD^{2*} = \frac{\Delta\bar{y}^2(1 + r_{y1y3})}{s^2(1 + r_{y1y3} - 2{r_{gp}}^2)} - \frac{\Delta\bar{y}^2(1 + {r_{y1y3}}^*)}{s^2(1 + {r_{y1y3}}^* - 2{r_{gp}}^2)},$$

$$MD^2 - MD^{2*} = \frac{-2{r_{gp}}^2({r_{y1y3}} - {r_{y1y3}}^*)}{s^2(1 + r_{y1y3} - 2{r_{gp}}^2)(1 + {r_{y1y3}}^* - 2{r_{gp}}^2)}.$$

Sabiendo que el denominador tiene que ser positivo al corresponder cada uno de los términos del producto a una expresión de MD^2 con numerador positivo, MD^2 será mayor que MD^{2*} si $r_{y1y3} < {r_{y1y3}}^*$, llegándose a la misma conclusión que en el MANOVA de un factor, a mayores correlaciones positivas entre las variables de efecto pequeño menor es la distancia entre los grupos.

En la Figura 3.13 se ha representado la variación de la potencia. La tendencia vuelve a ser decreciente, pero en este caso al ser dos de las variables de efecto pequeño se nota que la potencia de MANOVA para correlaciones positivas queda superado por el del ANOVA de la variable con tamaño del efecto grande en mayor medida que cuando se tenían dos variables con efecto grande.

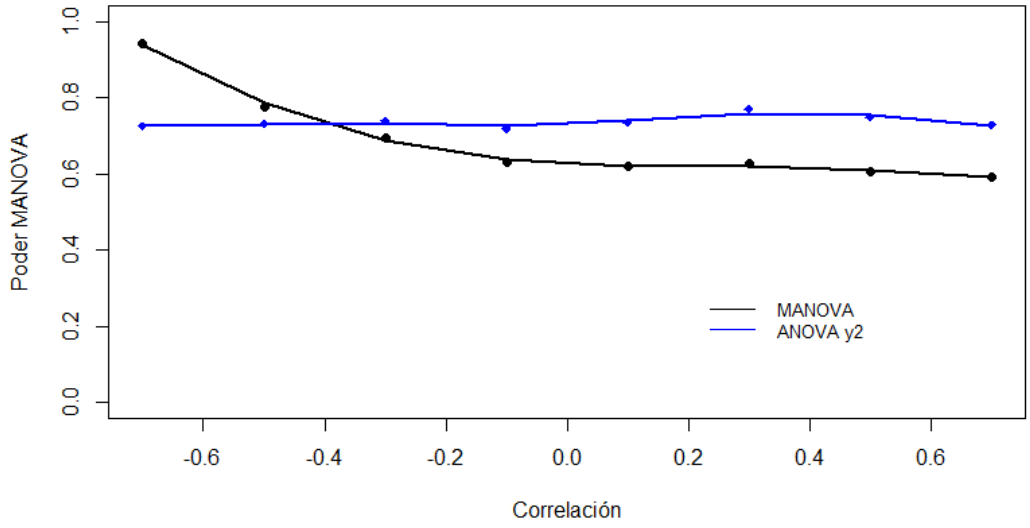


Figura 3.13. Potencia MANOVA y ANOVA factor 1, en función de r_{y1y3} para tamaño del efecto pequeño

3.2.3. Cambio en la correlación de variables con efecto pequeño y grande de los factores

El último caso que se analizará será el cambio de la correlación de una variable con tamaño del efecto grande, respecto a una con un tamaño del efecto pequeño.

Si se retoma la expresión de MD del segundo apartado de este segundo experimento, se tenía que

$$MD^2 = \frac{\Delta\bar{y}^2(1 + r_{y1y3})}{s^2(1 + r_{y1y3} - 2{r_{gp}}^2)}.$$

Estando en el denominador el cuadrado de la correlación entre las variables de efecto pequeño y la de efecto grande, que en ese caso se ha establecido la misma para las dos de efecto pequeño con respecto a la de efecto grande. Si se aumentase el valor absoluto de r_{gp}, el solapamiento entre los datos debería disminuir. Se representará el cambio en el valor de F y del p-valor cuando se mantiene fija la correlación entre las variables de efecto pequeño, en un valor de 0.3 y se modifica la correlación de las dos de efecto pequeño con la de efecto grande, en el intervalo comprendido de -0.7 a 0.7.

		$r = -0.7$	$r = -0.5$	$r = -0.3$	$r = -0.1$	$r = 0.1$	$r = 0.3$	$r = 0.5$	$r = 0.7$
ANOVA y1	F	1.02	1.046	0.944	0.991	1.067	0.995	0.974	1.014
	p	0.492	0.488	0.516	0.501	0.474	0.5	0.514	0.499
ANOVA y2	F	5.383	5.41	4.92	5.114	5.361	5.316	5.401	5.231
	p	0.059	0.062	0.071	0.07	0.059	0.069	0.058	0.06
ANOVA y3	F	1.056	1.021	0.952	1.028	1.107	0.961	0.992	0.978
	p	0.498	0.506	0.519	0.503	0.479	0.509	0.499	0.51
MANOVA	F	6.225	3.197	2.426	2.334	2.463	2.54	3.169	6.199
	p	0.001	0.049	0.115	0.131	0.118	0.109	0.054	0.001

Tabla 3.10. Resultados ANOVA y MANOVA cambiando r_{gp} para el efecto del factor 1

Al igual que en el experimento anterior, el p-valor crece a medida que la correlación se acerca a 0 y aumenta a medida que se aleja en ambos sentidos.

Si se tienen en cuenta las dos correlaciones que están interviniendo en el análisis, debería cumplirse que para valores cercanos a 1 en la correlación entre las variables de efecto pequeño disminuya la potencia de MANOVA, que sumado a valores cercanos a 0 de la correlación entre las variables de efecto pequeño y la de efecto grande da lugar a potencias todavía más pequeñas. En la Figura 3.14 se muestra la comparativa de la potencia de MANOVA en función de ambas correlaciones.

En azul, rojo y negro se ha representado la evolución de la potencia de MANOVA, correspondiendo cada curva de color a un valor de r_{gp}, para distintos valores de r_{y1y3} en el eje de abscisas.

Lo mismo debería ocurrir cuando se modifica la correlación entre las variables de tamaño del efecto grande y la de efecto pequeño en el primer apartado. Si se retoma su expresión de MD,

$$MD^2 = \frac{2\Delta y^2}{s^2(1 + r_{y1y2} - 2{r^2}_{gp})},$$

el denominador también contiene el cuadrado de la correlación de las variables de efecto grande con la de efecto pequeño. En la Figura 3.15 se ha representado la variación de la potencia, tanto

con la correlación de las variables de efecto pequeño y grande, como de las dos de efecto grande, usando los mismos parámetros que en el primer apartado. Cada curva de color corresponde a un valor [5] de r_{gp} y a lo largo del eje x se encuentran los distintos valores de r_{y1y2}.

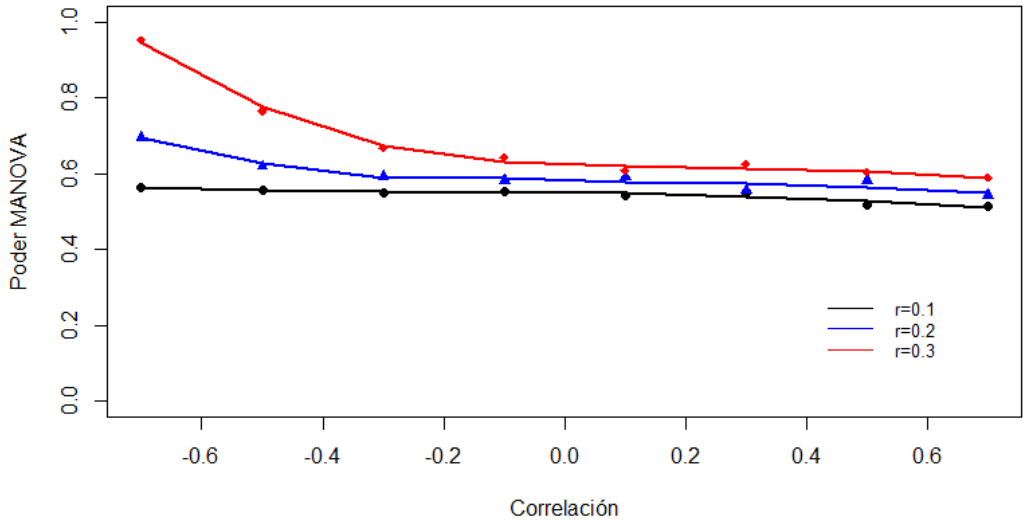


Figura 3.14. Potencia de MANOVA en función de r_{gp} y r_{y1y3}

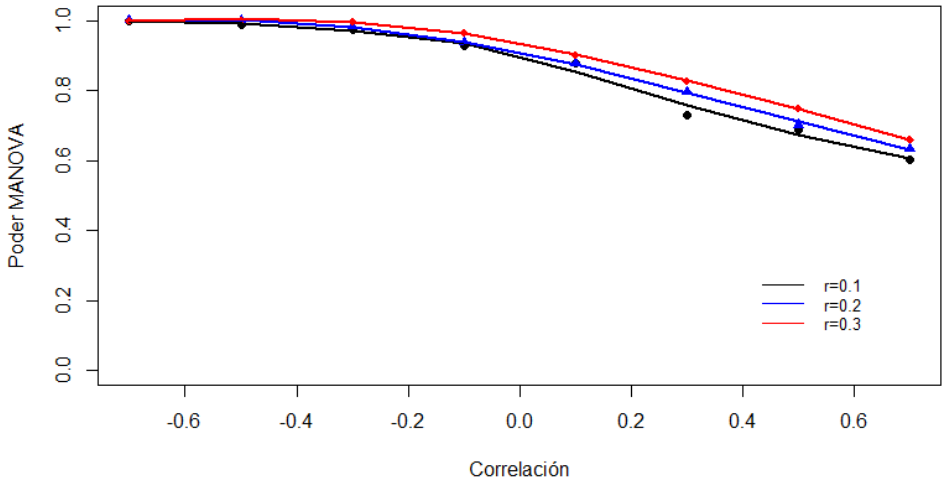


Figura 3.15. Potencia de MANOVA para el factor 1 en función de r_{gp} y r_{y1y2}

El ANOVA, en cambio, como es de esperar, no se ve influenciado por un cambio en la correlación, ni de las variables con efecto grande ni de las de efecto pequeño, en la Figura 3.16 se representa la evolución de la potencia del ANOVA para el efecto del factor 2 de una de las variables de efecto grande de la primera simulación de este segundo experimento.

[5] Se ha probado solamente con esos valores de r_{gp}, ya que para valores absolutos más grandes se vuelve a dar el problema de que la matriz de correlaciones deja de ser definida positiva.

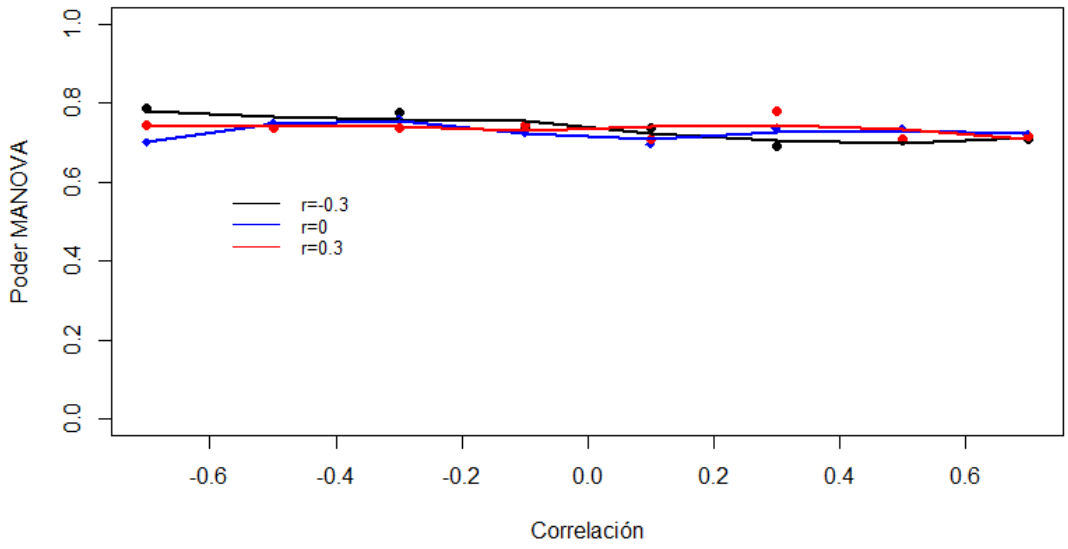


Figura 3.16. Potencia ANOVA de y_1 para el factor 2 en función de r_{gp}

Las tres curvas siguen una evolución similar sin verse influenciadas por la correlación, la expresión de la distancia de Mahalanobis usando solo una variable se recuerda que es

$$MD^2 = \frac{\Delta \bar{y}_u{}^2}{s_u^2},$$

donde MD^2 está representando la distancia entre dos de los niveles para la variable dependiente u, como se observa, no interviene el valor ni de r_{gp} ni de ninguna otra correlación, idéntico al experimento del MANOVA de un factor. En este caso la distancia entre los dos niveles queda determinada por la distancia estadística, al no existir segunda variable dependiente con la que correlacionar esta. Por tanto, a valores grandes de $|r_{gp}|$ junto con valores de r_{y1y2} cercanos a -1, en el caso en el que y_1 e y_2 sean variables con tamaño del efecto grande, la potencia de MANOVA será mucho mayor en comparación con la de los ANOVAs individuales.

3.2.4. Tamaño muestral

En el MANOVA de un factor se observó como un aumento en el tamaño muestral suponía p-valores más pequeños, tanto en los ANOVAs como el MANOVA. Se analizará para el nuevo modelo si el efecto que se produce es similar.

Para ello, se duplicará el número de muestras por grupo, que antes era de 20 y ahora pasará a ser de 40. En la Tabla 3.11 se muestran los datos para el efecto del factor 1 en el caso en el que se esté cambiando la correlación en las variables con tamaño del efecto pequeño. Las tres variables se han generado usando los mismos parámetros que en el segundo apartado del MANOVA de dos factores, salvo el tamaño muestral.

		$r=-0.7$	$r=-0.5$	$r=-0.3$	$r=-0.1$	$r=0.1$	$r=0.3$	$r=0.5$	$r=0.7$
ANOVA y1	F	0.948	0.946	0.926	0.935	1.011	0.993	0.936	0.963
	p	0.512	0.52	0.527	0.525	0.5	0.503	0.525	0.511
ANOVA y2	F	9.877	9.65	9.38	9.395	9.361	9.38	9.351	9.105
	p	0.006	0.005	0.008	0.008	0.008	0.008	0.008	0.009
ANOVA y3	F	0.951	0.955	1.099	1.05	1.062	0.995	1.03	0.981
	p	0.509	0.513	0.481	0.498	0.493	0.5	0.494	0.51
MANOVA	F	7.626	5.275	4.607	4.381	4.212	4.146	4.037	3.878
	p	0	0.003	0.009	0.013	0.016	0.016	0.018	0.018

Tabla 3.11. Resultados ANOVA y MANOVA cambiando la correlación en las variables de efecto pequeño para el efecto del factor 1 con un total de 360 observaciones

Si se compara con la del segundo apartado, se observa que los valores de F del MANOVA y del ANOVA de la variable y_3 prácticamente se han duplicado con respecto a los del otro. Los ANOVA de y_1 e y_3 no varían al haberse anulado su tamaño del efecto como simplificación. Por tanto, de la misma manera que en el MANOVA de un factor, se puede decir que un aumento en el número de muestras da lugar p-valores más bajos.

Se representa en las Figuras 3.17 y 3.18 la comparativa de la potencia de MANOVA y del ANOVA de y_1. Se ha hecho para tres tamaños muestrales diferentes, usando las variables con las que se realizó la simulación del primer apartado del MANOVA de dos factores y para el efecto del factor 1. En ambos casos la potencia aumenta, debido a la disminución del p-valor, la única diferencia es que en ANOVA no existe tendencia decreciente con la correlación entre las variables de efecto pequeño.

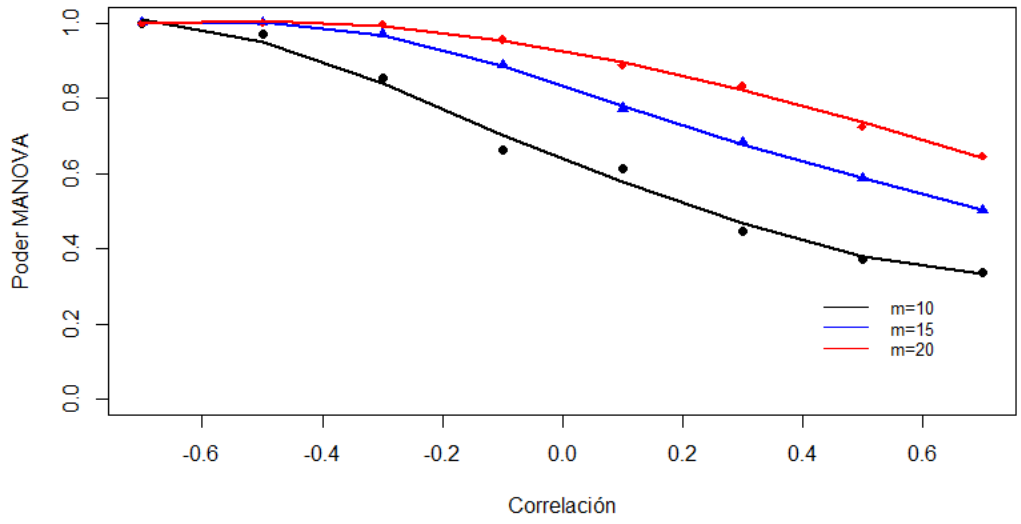


Figura 3.17. Potencia de MANOVA en función del número de observaciones m

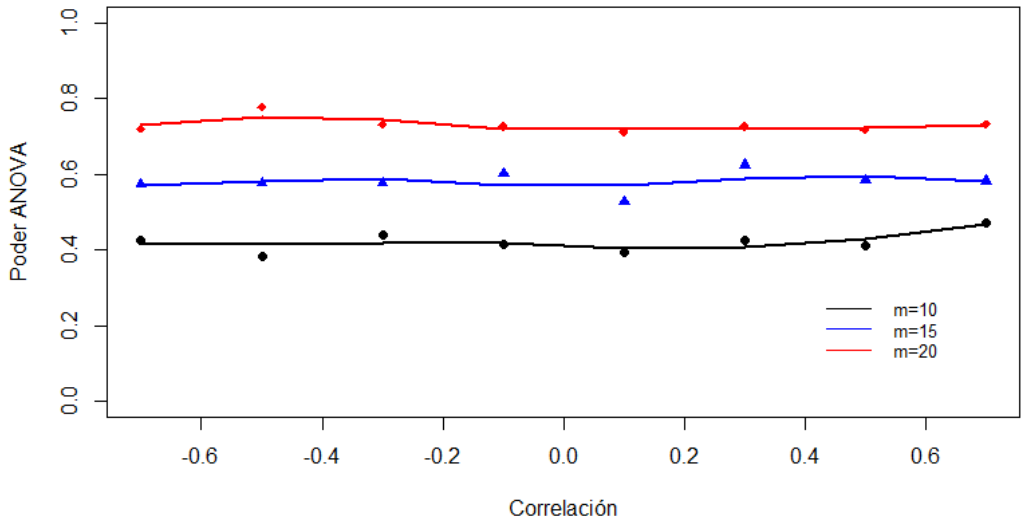


Figura 3.18. Potencia del ANOVA de y_1 en función del número de observaciones m

Con los resultados de este último apartado se puede concluir que lo visto en el primer experimento se puede ampliar a un caso con una mayor complejidad.

4. CONCLUSIONES Y LÍNEAS FUTURAS

4.1. Conclusiones

En este estudio se han llevado a cabo una serie de simulaciones que nos permiten tener un mayor conocimiento de cómo funcionan las herramientas de MANOVA y ANOVA. Se han generado un conjunto de variables de forma aleatoria, estableciendo unos parámetros como fijos y se ha estudiado el impacto que tiene en los resultados el cambiar uno o varios parámetros de forma simultánea.

Se ha visto que la potencia del MANOVA depende de una combinación conjunta del tamaño del efecto, de la correlación, de la varianza y del tamaño muestral. Para ello se han estudiado tres situaciones distintas, en una se han correlacionado dos variables con tamaños del efecto grandes, en otra pequeños y en otra se ha considerado que una de las variables tenía un tamaño del efecto despreciable frente a la otra, fijando en cada una de ellas un tamaño muestral y una misma varianza para todas las variables dependientes.

Se extraen las siguientes conclusiones de cada uno de los casos estudiados:

- En aquellos experimentos en los que existan variables dependientes con tamaños del efecto considerablemente grandes, correlaciones negativas darán lugar a mayores potencias. Esta situación se puede dar cuando se toman medidas cuyas diferencias entre los distintos grupos vayan en sentidos opuestos. Por ejemplo, Pabalan, Davey y Packe en uno de sus estudios sobre el efecto de la captividad y maltrato de las abejas en sus capacidades reproductivas, tomaron como variables dependientes el tiempo de exposición al tratamiento y su índice de desarrollo ovárico, siendo este último menor a medida que estaban expuestas más tiempo al tratamiento.
- Puede ser beneficioso, en un estudio, recopilar datos no solo de las variables en las que se espera gran repercusión de la variable independiente, sino también en aquellas que, estando relacionadas, las diferencias que se esperan percibir sean mucho menores (tamaño del efecto pequeño), a mayores correlaciones, tanto positivas como negativas, de esta variable con las de tamaño del efecto grande se tendrá una mayor potencia del MANOVA. En el ejemplo del MANOVA de dos factores se comentó un estudio en el que se medía el rendimiento de unas personas que padecían una serie de trastornos, una de las variables dependientes medidas fue la edad, en la que no se encontraban grandes diferencias entre los grupos, una mayor correlación de esta con la cognición social, por ejemplo, resulta en un aumento de la potencia del MANOVA.
- Algo similar al caso en el que se tengan variables en las que se esperan grandes diferencias entre los grupos ocurre para aquellas en las que se esperan pequeñas diferencias. Si a medida que se están recopilando los datos, el investigador descubre la presencia de dos variables con tamaños del efecto pequeño será beneficioso que estén negativamente correlacionadas entre sí.
- Aumentos en la cantidad de datos recopilados acerca de un fenómeno, por un lado, aumentan el tamaño del efecto, al disponer de una muestra más representativa del grupo sometido a estudio, pero por otro, aumenta la correlación existente entre las variables medidas, que en casos en los que disponga de variables con tamaños del efecto parecidos se ha visto que se traduce en una pérdida de potencia. De las simulaciones se ha

comprobado que el resultado final es una ganancia neta de potencia independientemente del aumento en la correlación.

4.2. Líneas futuras

Los experimentos se han reducido a un número limitado de casos, no obstante, se pueden ampliar a otros más extensos, se describirá algunos de estos:

- Las simulaciones se han hecho con un número máximo de tres variables dependientes, se podría ampliar y estudiar el efecto que tiene el añadir otras variables dependientes con distintos tamaños del efecto correlacionadas con las ya estudiadas.
- Se ha trabajado con unos determinados tamaños del efecto, fijando en 0 el de las variables de efecto pequeño, excepto en la primera simulación del MANOVA de un factor, algo que es bastante poco probable en un experimento real. Se podrían estudiar otros valores en las variables de efectos grandes y pequeños, viendo hasta qué punto se puede considerar que son de cada uno de los tipos.
- El número máximo de niveles formados en la variable independiente ha sido de 3, pudiendo ampliarse y ver cómo aumenta o disminuye el solapamiento entre los datos cuando se trabaja con mayor número de niveles.
- Desarrollar expresiones analíticas a partir del test de la intersección de la unión y contrastarlas con nuevas simulaciones de Monte Carlo.

5. PLANIFICACIÓN TEMPORAL Y PRESUPUESTO

5.1. Planificación temporal

El proyecto se ha separado en cinco puntos principales, en la Figura 5.1 se han representado los cinco puntos y los paquetes de trabajo asociados.

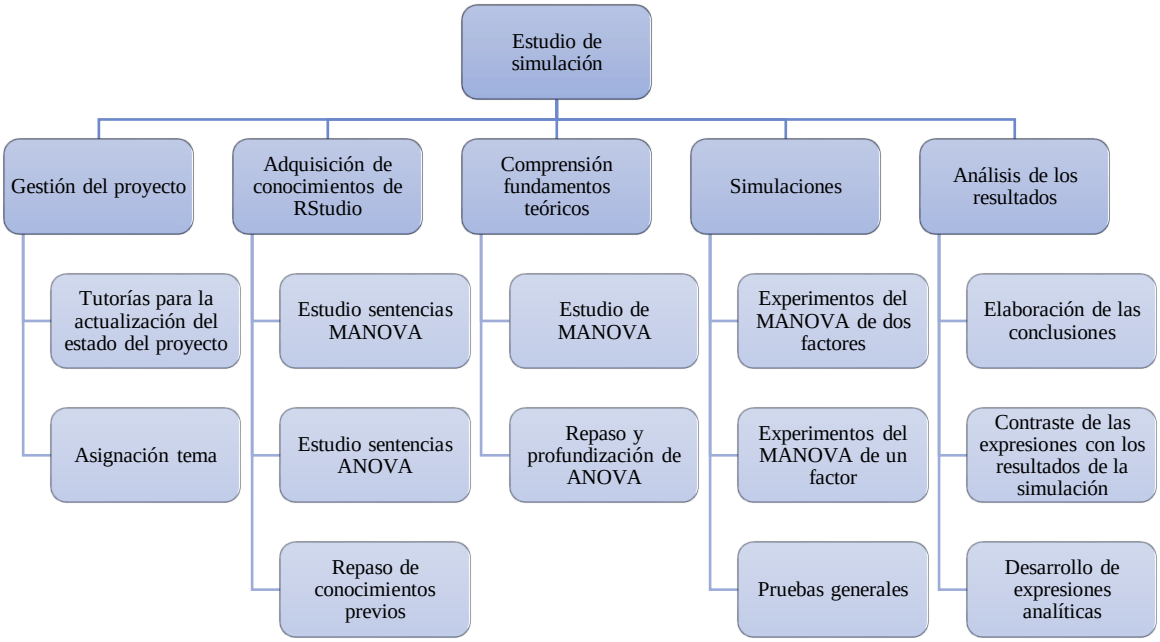


Figura 5.1. Estructura de descomposición del proyecto (EDP)

En cuanto a los tiempos dedicados a cada una de una de las tareas de la EDP, se han representado en la Figura 5.2, detallando el conjunto de actividades que se han llevado a cabo en cada paquete de trabajo. En rojo se muestra el camino crítico.

La duración ha sido de 124 días, comenzando desde el 20 de julio, día en el que se recibe la asignación del trabajo. No se ha considerado el tiempo que ha llevado la elaboración de la memoria, por ello, se establece como fecha de fin el 10 de enero, día en el que se dan por concluidos los resultados y conclusiones extraídas de las simulaciones. El tiempo medio dedicado al día ha sido aproximadamente de 3 horas, dependiendo de la carga de trabajo que se haya tenido del máster, dedicando más tiempo en el periodo previo al inicio del curso y en aquellas semanas libres de exámenes. Se consideran sábados y domingos como festivos.

5.2. Presupuesto

Los costes se recogerán en tres categorías generales, los costes directos, los de inversión en equipos y los indirectos.

5.2.1. Costes directos

Los costes directos imputables en este proyecto son los de mano de obra. El número de personas que han intervenido son dos, el alumno y el tutor del trabajo.

El sueldo de un alumno se podría aproximar al de un asistente de investigación. Atendiendo a ofertas de empleo en esta posición de la Universidad Politécnica de Madrid, se observan salarios de unos 10.000 € anuales para un contrato de media jornada, por tanto, se estimará en unos 9 €/h, dando un coste total del de

$$C_{alumno} = 9\ €/h \times 124\ días \times 3\ horas/día = 3.348\ €.$$

El salario del tutor se estima en base a consultas en diversas fuentes en unos 25 €/h y el tiempo de dedicación, teniendo en cuenta tutorías y revisiones del trabajo por cuenta propia en unas 20 horas.

$$C_{tutor} = 25\ €/h \times 20\ h = 500\ €.$$

Unos costes directos totales de 3.848 €.

5.2.2. Costes indirectos

El principal coste indirecto que se repercutirá es el de suministros. En este caso, al estar trabajando únicamente con un ordenador portátil, se estimarán los costes del consumo eléctrico.

De media se puede considerar que un ordenador portátil consume aproximadamente 1 kWh al día. El precio medio de la luz durante los meses de agosto a diciembre de 2021, atendiendo a los datos publicados en la web de statista (https://es.statista.com/estadisticas/993787/precio-medio-final-de-la-electricidad-en-espana/) es de unos 0,19 €/kWh, aplicando un factor corrector al consumo del ordenador en base al tiempo diario que se utiliza se llega a

$$C_{suministros} = 0{,}19\ €/kWh \times 1\ kWh/día \times {}^{3}/_{8} \times 124\ días = 9\ €.$$

5.2.3. Coste de equipos

El único equipo empleado para el estudio ha sido un portátil Lenovo IdeaPad 5i. El precio de adquisición fue de 750 €, el 3 de mayo de 2021. La parte que se repercute de su coste es la correspondiente a la amortización o pérdida de valor del activo.

Atendiendo a las normas contables, el plazo máximo de amortización para un equipo electrónico es de 10 años, se utilizará la amortización lineal, por tanto, el coste incurrido en los meses correspondientes es de

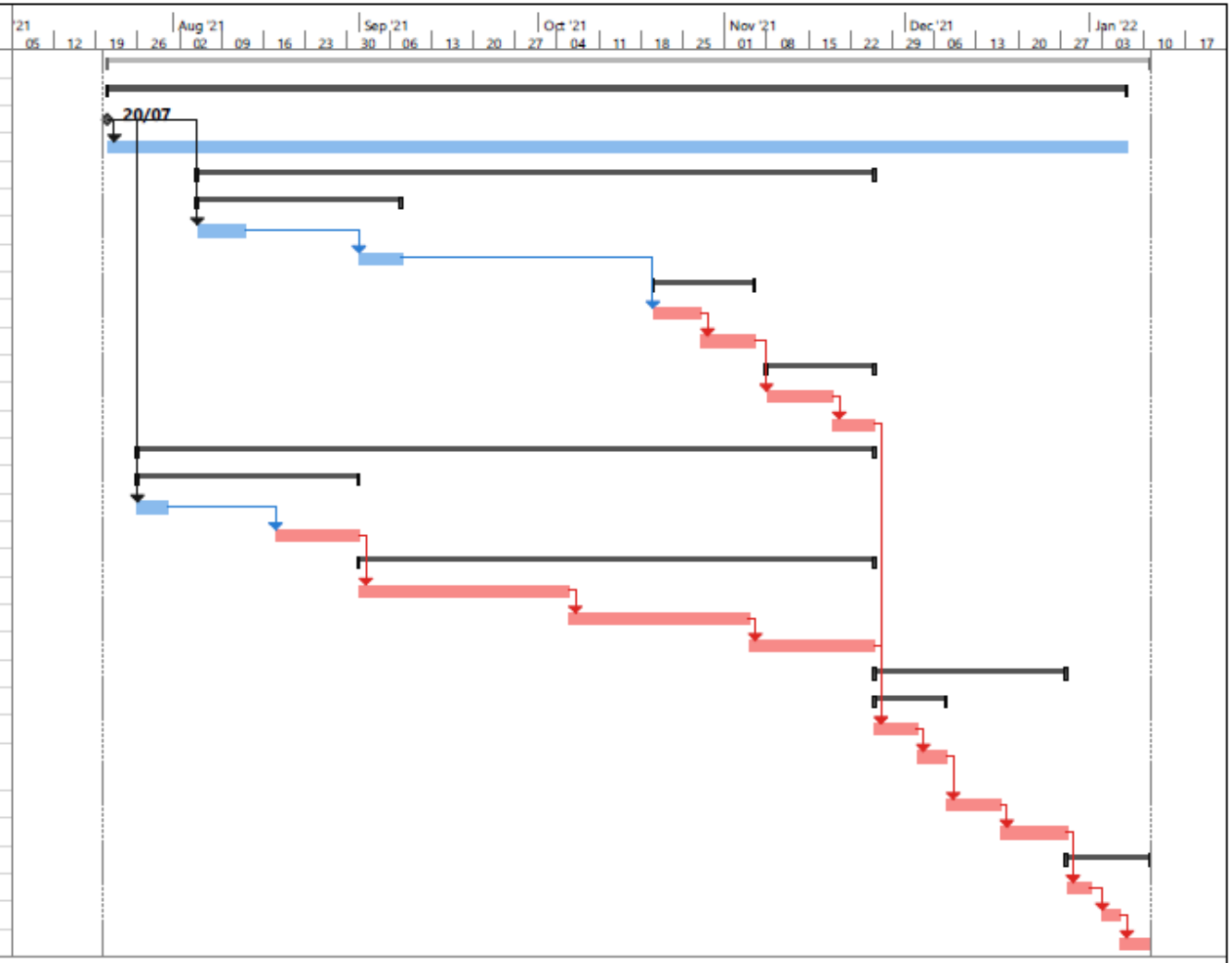
$$C_{equipos} = {}^{750\ €}/_{10\ años} \times {}^{5}/_{12}\ años = 31\ €.$$

En la Tabla 5.1 se recoge el resumen de los costes

Concepto	Coste (€)
Salarios	3.848
Suministros	9
Equipos	31
Total	**3.888**

Tabla 5.1. Presupuesto final del proyecto

ID	Task Name	Duration	Start	Finish
0	**Estudio de simulación**	**124 days**	**Tue 20/07/21**	**Mon 10/01/22**
1	**Gestión del proyecto**	**122 days**	**Tue 20/07/21**	**Thu 06/01/22**
2	Asignación tema	0 days	Tue 20/07/21	Tue 20/07/21
3	Tutorias	122 days	Wed 21/07/21	Thu 06/01/22
4	**Conocimientos RStudio**	**81 days**	**Thu 05/08/21**	**Thu 25/11/21**
5	**Repaso de conocimientos previos**	**24 days**	**Thu 05/08/21**	**Tue 07/09/21**
6	Repaso general	6 days	Thu 05/08/21	Thu 12/08/21
7	Instalación y pruebas con R Markdowr	5 days	Wed 01/09/21	Tue 07/09/21
8	**Sentencias ANOVA**	**13 days**	**Wed 20/10/21**	**Fri 05/11/21**
9	ANOVA de un factor	6 days	Wed 20/10/21	Wed 27/10/21
10	ANOVA de dos factores	7 days	Thu 28/10/21	Fri 05/11/21
11	**Sentencias MANOVA**	**14 days**	**Mon 08/11/21**	**Thu 25/11/21**
12	MANOVA de un factor	9 days	Mon 08/11/21	Thu 18/11/21
13	MANOVA de dos factores	5 days	Fri 19/11/21	Thu 25/11/21
14	**Fundamentos teóricos**	**89 days**	**Mon 26/07/21**	**Thu 25/11/21**
15	**ANOVA**	**27 days**	**Mon 26/07/21**	**Tue 31/08/21**
16	Repaso	5 days	Mon 26/07/21	Fri 30/07/21
17	Profundización en aspectos de ANOVA	10 days	Wed 18/08/21	Tue 31/08/21
18	**Estudio de MANOVA**	**62 days**	**Wed 01/09/21**	**Thu 25/11/21**
19	Lectura libro Sharma	25 days	Wed 01/09/21	Tue 05/10/21
20	Lectura libro Multivariate Data Analys	22 days	Wed 06/10/21	Thu 04/11/21
21	Lectura libro Mardia	15 days	Fri 05/11/21	Thu 25/11/21
22	**Simulaciones**	**22 days**	**Fri 26/11/21**	**Mon 27/12/21**
23	**Primeras pruebas de simulación**	8 days	Fri 26/11/21	Tue 07/12/21
24	Creación de tablas	5 days	Fri 26/11/21	Thu 02/12/21
25	Generación de números aleatorios y bucles	3 days	Fri 03/12/21	Tue 07/12/21
26	MANOVA de un factor	7 days	Wed 08/12/21	Thu 16/12/21
27	MANOVA de dos factores	7 days	Fri 17/12/21	Mon 27/12/21
28	**Análisis resultados**	**10 days**	**Tue 28/12/21**	**Mon 10/01/22**
29	Expresiones analíticas	4 days	Tue 28/12/21	Fri 31/12/21
30	Contraste con simulaciones	3 days	Mon 03/01/22	Wed 05/01/22
31	Conclusiones	3 days	Thu 06/01/22	Mon 10/01/22



Project: Estudio de simulación
Date: Sun 06/02/22

Task
Split
Milestone
Summary
Project Summary
Inactive Task
Inactive Milestone
Inactive Summary
Manual Task
Duration-only
Manual Summary Rollup
Manual Summary
Start-only
Finish-only
External Tasks
External Milestone
Deadline
Critical
Critical Split
Progress
Manual Progress

Figura 5.2. Diagrama de Gantt

REFERENCIAS

Amat, J. (enero de 2016). *ANOVA análisis de varianza para comparar múltiples medias*. Ciencia de Datos. https://www.cienciadedatos.net/documentos/19_anova.

Cole, D., Maxwell, S., Arvey, R. y Salas, E. (1994). *How the Power of MANOVA Can Both Increase and Decrease as a Function of the Intercorrelations Among the Dependent Variables*. Psychological Bulletin 115: 465–74.

Everitt, B. y Howell, D. (2005). Encyclopedia of Statistics in Behavioral Science. (Vol. 2). John Wiley & Sons, Ltd.

Hammersley, J. y Handscomb, D. (1964). *Monte Carlo Methods*. Chapman & Hall.

Hair, J., Black, W., Babin, B., Anderson, R. (2019). *Multivariate Data Analysis*. (8ª ed.). Cengage.

Hintze, J. (2007). *User's Guide IV Multivariate Analysis, Clustering, Meta-Analysis, Forecasting / Time Series, Operations Research, Mass Appraisal*. NCSS Statistical Software. https://www.ncss.com/wp-content/uploads/2012/09/NCSSUG4.pdf.

Indian Statistical Institute. *R.A.Fisher*. Google Arts & Culture https://artsandculture.google.com/asset/r-a-fisher/i AFkKK47btRx5Q.

Laboratorio de Estadística. *Diseño de Experimentos y Regresión*. http://www.etsii.upm.es/ingor/estadistica/Grado/dLibro.pdf. Universidad Politécnica de Madrid. Escuela Técnica Superior de Ingenieros Industriales.

Mardia, K., Kent, J. y Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Murphy, T. (2022). *Statistical Design and Analysis of Experiments with R*. Emory University. https://tjmurphy.github.io/jabstb/.

Particle Data Group. (1 de junio de 2020). *Monte Carlo Techniques*. https://pdg.lbl.gov/2020/reviews/rpp2020-rev-monte-carlo-techniques.pdf.

Rosenfeld, B. *History of Statistics 8. Analysis of Variance and the Design of Experiments. R. A. Fisher (1890-1962)*. Similarweb. https://higherlogicdownload.s3.amazonaws.com/AMS__TAT/1484431b-3202-461e-b7e6-ebce10ca8bcd/UploadedImages/Classroom_Activities/HS_8__FISHER_and_Design_of_experiments.pdf.

RStudio. *About RStudio*. https://www.rstudio.com/about/.

School of Mathematics and Statistics University of St Andrews. *Samuel Stanley Wilks*. MacTutor. https://mathshistory.st-andrews.ac.uk/Biographies/Wilks/pictdisplay/.

Sharma, S. (1996). *Applied Multivariate Techniques*. University of South Carolina.

ANEXO

Sentencias de R

Simulación del MANOVA de un factor

```
#Se cargan las librerías necesarias
library(MASS)
library(heplots)

library(tidyverse)

library(flextable)

#Se establece el número de veces que se repetirá la simulación
repeticiones=100
#Se declaran las variables para crear las tablas
Fanovay1=array(rep(0,repeticiones),repeticiones)
Fanovay2=array(rep(0,repeticiones),repeticiones)
FmanovaPillai=array(rep(0,repeticiones),repeticiones)
FmanovaWilks=array(rep(0,repeticiones),repeticiones)
FmanovaHotelling=array(rep(0,repeticiones),repeticiones)
FmanovaRoy=array(rep(0,repeticiones),repeticiones)
pvaloranovay1=array(rep(0,repeticiones),repeticiones)
pvaloranovay2=array(rep(0,repeticiones),repeticiones)
pvalormanovaPillai=array(rep(0,repeticiones),repeticiones)
pvalormanovaWilks=array(rep(0,repeticiones),repeticiones)
pvalormanovaRoy=array(rep(0,repeticiones),repeticiones)
pvalormanovaHotelling=array(rep(0,repeticiones),repeticiones)
Fanovay1tabla=array(rep(0,repeticiones),10)
Fanovay2tabla=array(rep(0,repeticiones),10)
FmanovatablaPillai=array(rep(0,repeticiones),10)
FmanovatablaWilks=array(rep(0,repeticiones),10)
FmanovatablaHotelling=array(rep(0,repeticiones),10)
FmanovatablaRoy=array(rep(0,repeticiones),10)
pvaloranovay1tabla=array(rep(0,repeticiones),10)
pvaloranovay2tabla=array(rep(0,repeticiones),10)
pvalormanovatablaPillai=array(rep(0,repeticiones),10)
pvalormanovatablaWilks=array(rep(0,repeticiones),10)
pvalormanovatablaHotelling=array(rep(0,repeticiones),10)
pvalormanovatablaRoy=array(rep(0,repeticiones),10)
k=1
#Se usará un bucle for para variar la correlación
for (corrcoef in seq(-0.9, 0.9, 0.2)) {
i=1
j=0
repeat{
  #generación de las variables dependientes
mu11=2;mu12=4;mu21=4;mu22=6;
  nmues=100;
  errormu=c(0,0)
  sigma1=10;sigma2=10;

sigma=array(c(sigma1^2,sigma1*sigma2*corrcoef,sigma1*sigma2*corrcoef,sigma2^
```

```
2),c(2,2))#matriz de covarianzas del término aleatorio
  error=mvrnorm(n=nmues,errormu,sigma,tol=1e-6,empirical=FALSE,EISPACK =
FALSE)
  y1=array(rep(0,nmues),nmues)
  y2=array(rep(0,nmues),nmues)
  y1[1:(nmues/2)]=mu11+error[1:(nmues/2),1]
  y2[1:(nmues/2)]=mu21+error[1:(nmues/2),2]
  y1[((nmues/2)+1):nmues]=mu12+error[((nmues/2)+1):nmues,1]
  y2[((nmues/2)+1):nmues]=mu22+error[((nmues/2)+1):nmues,2]
  Tratamiento=c(rep(1,nmues/2),rep(2,nmues/2))
  Tratamiento=factor(Tratamiento,labels=c(1,2))
  Tratamiento=factor(Tratamiento,labels=c(1,2))

  Output=cbind(y1,y2)
  #verificamos cumplimiento hipótesis
  if((length(which(Tratamiento==2))>2)&(length(which(Tratamiento==1))>2)){

if((boxM(Output,Tratamiento)$p.value>0.05)&&(leveneTest(y1~Tratamiento)$`Pr(>
F)`[[1]]>0.05)
       &&(leveneTest(y2~Tratamiento)$`Pr(>F)`[[1]]>0.05)){
      j=j+1
      m1=aov(y1~Tratamiento)
      m2=aov(y2~Tratamiento)
      m3=manova(Output~Tratamiento)
      Fanovay1[j]=summary(m1)[[1]][["F value"]][[1]]
      Fanovay2[j]=summary(m2)[[1]][["F value"]][[1]]
      FmanovaPillai[j]=summary(m3)$stat[5]
      FmanovaWilks[j]=summary(m3,test="Wilks")$stat[5]
      FmanovaHotelling[j]=summary(m3,test="Hotelling")$stat[5]
      FmanovaRoy[j]=summary(m3,test="Roy")$stat[5]
      pvaloranovay1[j]=summary(m1)[[1]][["Pr(>F)"]][[1]]
      pvaloranovay2[j]=summary(m2)[[1]][["Pr(>F)"]][[1]]
      pvalormanovaPillai[j]=summary(m3)$stat[11]
      pvalormanovaWilks[j]=summary(m3,test="Wilks")$stat[11]
      pvalormanovaHotelling[j]=summary(m3,test="Hotelling")$stat[11]
      pvalormanovaRoy[j]=summary(m3,test="Roy")$stat[11]
      #Hay que tener en cuenta que no todas las replicaciones cumplen el
if
    }
  }
  if(i==repeticiones) break
  i=i+1
}
#Fin del total de replicaciones

Fanovay1tabla[k]=mean(Fanovay1[1:j])
Fanovay2tabla[k]=mean(Fanovay2[1:j])
FmanovatablaPillai[k]=mean(FmanovaPillai[1:j])
FmanovatablaWilks[k]=mean(FmanovaWilks[1:j])
FmanovatablaHotelling[k]=mean(FmanovaHotelling[1:j])
FmanovatablaRoy[k]=mean(FmanovaRoy[1:j])
pvaloranovay1tabla[k]=mean(pvaloranovay1[1:j])
pvaloranovay2tabla[k]=mean(pvaloranovay2[1:j])
pvalormanovatablaPillai[k]=mean(pvalormanovaPillai[1:j])
```

```
pvalormanovatablaWilks[k]=mean(pvalormanovaWilks[1:j])
pvalormanovatablaHotelling[k]=mean(pvalormanovaHotelling[1:j])
pvalormanovatablaRoy[k]=mean(pvalormanovaRoy[1:j])
k=k+1
}
#Se hace la media de todos los valores obtenidos
```

Simulación del MANOVA de dos factores

```
#Se cargan las librerías
library(heplots)

library(car)
library(mvnfast)

#Se define el número de veces que se ejecutará la simulación
repeticiones=500
#Declaración de las variables
Fanovay1T1=array(rep(0,repeticiones),repeticiones)
Fanovay1T2=array(rep(0,repeticiones),repeticiones)
Fanovay1T1T2=array(rep(0,repeticiones),repeticiones)
Fanovay2T1=array(rep(0,repeticiones),repeticiones)
Fanovay2T2=array(rep(0,repeticiones),repeticiones)
Fanovay2T1T2=array(rep(0,repeticiones),repeticiones)
Fanovay3T1=array(rep(0,repeticiones),repeticiones)
Fanovay3T2=array(rep(0,repeticiones),repeticiones)
Fanovay3T1T2=array(rep(0,repeticiones),repeticiones)
Fanovay4T1=array(rep(0,repeticiones),repeticiones)
Fanovay4T2=array(rep(0,repeticiones),repeticiones)
Fanovay4T1T2=array(rep(0,repeticiones),repeticiones)
FmanovaPillaiT1=array(rep(0,repeticiones),repeticiones)
FmanovaPillaiT2=array(rep(0,repeticiones),repeticiones)
FmanovaPillaiT1T2=array(rep(0,repeticiones),repeticiones)
pvaloranovay1T1=array(rep(0,repeticiones),repeticiones)
pvaloranovay1T2=array(rep(0,repeticiones),repeticiones)
pvaloranovay1T1T2=array(rep(0,repeticiones),repeticiones)
pvaloranovay2T1=array(rep(0,repeticiones),repeticiones)
pvaloranovay2T2=array(rep(0,repeticiones),repeticiones)
pvaloranovay2T1T2=array(rep(0,repeticiones),repeticiones)
pvaloranovay3T1=array(rep(0,repeticiones),repeticiones)
pvaloranovay3T2=array(rep(0,repeticiones),repeticiones)
pvaloranovay3T1T2=array(rep(0,repeticiones),repeticiones)
pvalormanovaPillaiT1=array(rep(0,repeticiones),repeticiones)
pvalormanovaPillaiT2=array(rep(0,repeticiones),repeticiones)
pvalormanovaPillaiT1T2=array(rep(0,repeticiones),repeticiones)
Fanovay1T1tabla=array(rep(0,repeticiones),8)
Fanovay1T2tabla=array(rep(0,repeticiones),8)
Fanovay1T1T2tabla=array(rep(0,repeticiones),8)
Fanovay2T1tabla=array(rep(0,repeticiones),8)
Fanovay2T2tabla=array(rep(0,repeticiones),8)
Fanovay2T1T2tabla=array(rep(0,repeticiones),8)
Fanovay3T1tabla=array(rep(0,repeticiones),8)
Fanovay3T2tabla=array(rep(0,repeticiones),8)
Fanovay3T1T2tabla=array(rep(0,repeticiones),8)
FmanovaPillaiT1tabla=array(rep(0,repeticiones),8)
```

```
FmanovaPillaiT2tabla=array(rep(0,repeticiones),8)
FmanovaPillaiT1T2tabla=array(rep(0,repeticiones),8)
pvaloranovay1T1tabla=array(rep(0,repeticiones),8)
pvaloranovay1T2tabla=array(rep(0,repeticiones),8)
pvaloranovay1T1T2tabla=array(rep(0,repeticiones),8)
pvaloranovay2T1tabla=array(rep(0,repeticiones),8)
pvaloranovay2T2tabla=array(rep(0,repeticiones),8)
pvaloranovay2T1T2tabla=array(rep(0,repeticiones),8)
pvaloranovay3T1tabla=array(rep(0,repeticiones),8)
pvaloranovay3T2tabla=array(rep(0,repeticiones),8)
pvaloranovay3T1T2tabla=array(rep(0,repeticiones),8)
pvalormanovaPillaiT1tabla=array(rep(0,repeticiones),8)
pvalormanovaPillaiT2tabla=array(rep(0,repeticiones),8)
pvalormanovaPillaiT1T2tabla=array(rep(0,repeticiones),8)

#Configuración parámetros variables dependientes
I=3;J=3;m=20;
mu1=4;mu2=5;mu3=8;
alfa1=array(c(0,0,0),dim=c(3,1))
alfa2=array(c(2,1,-3),dim=c(3,1))
alfa3=array(c(0,0,0),dim=c(3,1))
beta1=array(c(0,0,0),dim=c(3,1))
beta2=array(c(-3,1,2),dim=c(3,1))
beta3=array(c(0,0,0),dim=c(3,1))
alfabeta1=array(c(0,0,0,0,0,0,0,0,0),dim=c(3,3))
alfabeta2=array(c(3,1,-4,1,-3,2,-4,2,2),dim=c(3,3))
alfabeta3=array(c(0,0,0,0,0,0,0,0,0),dim=c(3,3))
y11=array(runif(I*J*m,min=0,max=1))
y22=array(runif(I*J*m,min=0,max=1))
y33=array(runif(I*J*m,min=0,max=1))
sigma1=10
sigma2=10
sigma3=10
corrcoef12=0.3
corrcoef13=0.3
corrcoef23=0.3
t=1
#Se crea un bule for para ir variando la correlación de las variables
for(corrcoef13 in seq(-0.7,0.7,0.2)){
  s=0
  r=1
  repeat{
    #Matriz de covarianzas del error
dt2<-
rmvn(m*I*J,mu=c(0,0,0),matrix(c(sigma1^2,sigma1*sigma2*corrcoef12,sigma1*sigm
a2*corrcoef13,

sigma1*sigma2*corrcoef12,sigma2^2,sigma1*sigma2*corrcoef23,

sigma1*sigma2*corrcoef13,sigma1*sigma2*corrcoef23,sigma3^2),3,3))
Tratamiento1=array(runif(I*J*m,min=0,max=1),dim=c(I*J*m))#Factor 1
Tratamiento2=array(runif(I*J*m,min=0,max=1),dim=c(I*J*m))#Factor 2
contador=0
for (i in 1:I)
```

```
{
  for (j in 1:J)
  {
    for (k in 1:m)
    {
      contador=contador+1
      y11[contador]=mu1+alfa1[i]+beta1[j]+alfabeta1[i,j]+dt2[contador,1]
      y22[contador]=mu2+alfa2[i]+beta2[j]+alfabeta2[i,j]+dt2[contador,2]
      y33[contador]=mu3+alfa3[i]+beta3[j]+alfabeta3[i,j]+dt2[contador,3]
      Tratamiento1[contador]=i
      Tratamiento2[contador]=j
    }
  }
}
Tratamiento1=factor(Tratamiento1,labels=c(1,2,3))
Tratamiento2=factor(Tratamiento2,labels=c(1,2,3))
Output=cbind(y11,y22,y33)
mod=lm(Output~Tratamiento1*Tratamiento2)

  s=s+1
  m1=aov(y11~Tratamiento1*Tratamiento2)
  m2=aov(y22~Tratamiento1*Tratamiento2)
  m3=aov(y33~Tratamiento1*Tratamiento2)
  m5=manova(Output~Tratamiento1*Tratamiento2)
  Fanovay1T1[s]=summary(m1)[[1]][["F value"]][[1]]
  Fanovay1T2[s]=summary(m1)[[1]][["F value"]][[2]]
  Fanovay1T1T2[s]=summary(m1)[[1]][["F value"]][[3]]
  Fanovay2T1[s]=summary(m2)[[1]][["F value"]][[1]]
  Fanovay2T2[s]=summary(m2)[[1]][["F value"]][[2]]
  Fanovay2T1T2[s]=summary(m2)[[1]][["F value"]][[3]]
  Fanovay3T1[s]=summary(m3)[[1]][["F value"]][[1]]
  Fanovay3T2[s]=summary(m3)[[1]][["F value"]][[2]]
  Fanovay3T1T2[s]=summary(m3)[[1]][["F value"]][[3]]
  FmanovaPillaiT1[s]=summary(m5,tol=0)$stat[9]
  FmanovaPillaiT2[s]=summary(m5,tol=0)$stat[10]
  FmanovaPillaiT1T2[s]=summary(m5,tol=0)$stat[11]
  pvaloranovay1T1[s]=summary(m1)[[1]][["Pr(>F)"]][[1]]
  pvaloranovay1T2[s]=summary(m1)[[1]][["Pr(>F)"]][[2]]
  pvaloranovay1T1T2[s]=summary(m1)[[1]][["Pr(>F)"]][[3]]
  pvaloranovay2T1[s]=summary(m2)[[1]][["Pr(>F)"]][[1]]
  pvaloranovay2T2[s]=summary(m2)[[1]][["Pr(>F)"]][[2]]
  pvaloranovay2T1T2[s]=summary(m2)[[1]][["Pr(>F)"]][[3]]
  pvaloranovay3T1[s]=summary(m3)[[1]][["Pr(>F)"]][[1]]
  pvaloranovay3T2[s]=summary(m3)[[1]][["Pr(>F)"]][[2]]
  pvaloranovay3T1T2[s]=summary(m3)[[1]][["Pr(>F)"]][[3]]
  pvalormanovaPillaiT1[s]=summary(m5,tol=0)$stat[21]
  pvalormanovaPillaiT2[s]=summary(m5,tol=0)$stat[22]
  pvalormanovaPillaiT1T2[s]=summary(m5,tol=0)$stat[23]

if(r==repeticiones) break
r=r+1
  }#Fin de las replicaciones
  Fanovay1T1tabla[t]=mean(Fanovay1T1[1:s])
  Fanovay1T2tabla[t]=mean(Fanovay1T2[1:s])
```

```
  Fanovay1T1T2tabla[t]=mean(Fanovay1T1T2[1:s])
  Fanovay2T1tabla[t]=mean(Fanovay2T1[1:s])
  Fanovay2T2tabla[t]=mean(Fanovay2T2[1:s])
  Fanovay2T1T2tabla[t]=mean(Fanovay2T1T2[1:s])
  Fanovay3T1tabla[t]=mean(Fanovay3T1[1:s])
  Fanovay3T2tabla[t]=mean(Fanovay3T2[1:s])
  Fanovay3T1T2tabla[t]=mean(Fanovay3T1T2[1:s])
  FmanovaPillaiT1tabla[t]=mean(FmanovaPillaiT1[1:s])
  FmanovaPillaiT2tabla[t]=mean(FmanovaPillaiT2[1:s])
  FmanovaPillaiT1T2tabla[t]=mean(FmanovaPillaiT1T2[1:s])
  pvaloranovay1T1tabla[t]=mean(pvaloranovay1T1[1:s])
  pvaloranovay1T2tabla[t]=mean(pvaloranovay1T2[1:s])
  pvaloranovay1T1T2tabla[t]=mean(pvaloranovay1T1T2[1:s])
  pvaloranovay2T1tabla[t]=mean(pvaloranovay2T1[1:s])
  pvaloranovay2T2tabla[t]=mean(pvaloranovay2T2[1:s])
  pvaloranovay2T1T2tabla[t]=mean(pvaloranovay2T1T2[1:s])
  pvaloranovay3T1tabla[t]=mean(pvaloranovay3T1[1:s])
  pvaloranovay3T2tabla[t]=mean(pvaloranovay3T2[1:s])
  pvaloranovay3T1T2tabla[t]=mean(pvaloranovay3T1T2[1:s])
  pvalormanovaPillaiT1tabla[t]=mean(pvalormanovaPillaiT1[1:s])
  pvalormanovaPillaiT2tabla[t]=mean(pvalormanovaPillaiT2[1:s])
  pvalormanovaPillaiT1T2tabla[t]=mean(pvalormanovaPillaiT1T2[1:s])
  t=t+1
}#Se calcula la media de todos los valores
```