FACULTAD DE MATEMÁTICAS

DOBLE GRADO EN MATEMÁTICAS Y ESTADÍSTICA

DISEÑO DE EXPERIMENTOS CON SOFTWARE ESTADÍSTICO.

Trabajo fin de grado presentado por:
Víctor Salas Aranda

Universidad Ð Sevilla

Junio de 2018

Supervisado por:
JOAQUÍN A. GARCÍA DE LAS HERAS
JOSÉ LUIS PINO MEJÍAS

A mi familia y amigos que me inspiraron a estudiar matemáticas.

Índice general

Resumen

Hoy en día, y sobre todo debido al auge de las tecnologías, el hecho de que la estadística se encuentre detrás de casi cualquier proceso es indudable. Detrás de cada invento revolucionario hay un proceso previo de experimentación, que es tan antigua como el ser humano. Es por ello que a través de este proyecto trataremos de dar nociones básicas del maravilloso y amplio mundo de los diseños de experimentos, que incluye desde el diseño del estudio de la mejora en la producción de patatas realizado por Fisher en 1935, hasta el diseño del estudio del aumento en la potencia de un acelerador de partículas.

Gracias a los diseños de experimentos podremos no solo conocer qué factores influyen más o menos en la respuesta de un determinado invento, sino que podremos disminuir la variabilidad que se produce en dicha respuesta debida al azar, y poder así optimizar tiempo y recursos para ensayos posteriores.

Durante esta memoria nos encargaremos de indagar más a fondo, en los principales diseños de experimentos. A modo de ejemplo realizaremos un estudio sociodemográfico, a través del sotware informático *R*, con datos reales de cómo influyen ciertos factores en el rendimiento escolar de los alumnos del territorio andaluz. Así pues, desarrollaremos diferentes enfoques del estudio para poder conocer las diferencias entre cada uno de los diseños empleados y obtener, además, un posible guion para que el lector pueda realizar, si así lo desea, su propio diseño de experimentos.

No obstante, el objetivo de este proyecto no es realizar las demostraciones necesarias para el análisis de cada uno de los modelos estudiados, por lo que no se verán explicados cada uno de los desarrollos teóricos. Como base a este trabajo, nos basaremos principalmente en asignaturas estudiadas previamente en el grado, en el *Tutorial Agricolae* [10], así como en el libro: *Statistical Analysis of Designed Experiments: Theory and Applications* [15].

Abstract

Nowadays, there is a universal, indubitable truth: behind almost every single process there is statistics involved, and even more if we consider the current rise of technologies. Behind each revolutionary invention there is a prior process of experimentation, which is as old as humanity itself. That is the reason why, thanks to this present project, we will try to expound basic notions concerning the wonderful and wide world of the designs of experiments, which covers from the design of the study of the improvement in the production of potatoes made by Fisher in 1935, to the design of the study of the increase in the power of a particle accelerator.

Thanks to the designs of experiments, we cannot only know what the factors which influence —to a greater or lesser degree— the response of a particular invention are, but also we can reduce the variability that occurs in such response due to chance, and thus we can optimize time and resources for subsequent trials.

Throughout this memory we will explore the main designs of experiments. A sociodemographic study will be carried out to exhibit it as an example, using the computer software *R* and real data on how certain factors influence the academic performance of students in the Andalusian territory. Hence, we will develop different study approaches so as to discern the differences between each design used. In addition, we will also obtain a possible script so that the reader can perform, if he wants to, his own design of experiments.

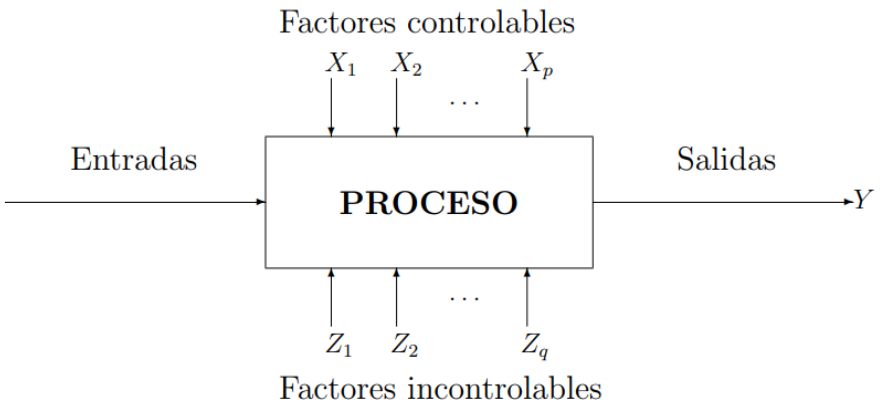
However, the objective of this project is not to carry out the necessary demonstrations for the analysis of every single model studied. Consequently, each of the theoretical developments will not be explained. As a basis for this work, we will mainly rely on subjects previously studied in the degree, in the *Tutorial Agricolae* [10], as well as in the book: *Statistical Analysis of Designed Experiments: Theory and Applications* [15].

Capítulo 1

Introducción al Diseño de Experimentos.

Comenzaremos con un breve recorrido histórico obtenido de [16]: A pesar de que la técnica del Diseño de Experimentos no surgió hasta el siglo XX, la experimentación es tan vieja como la existencia del hombre. Durante muchos años se utilizó la estrategia de un-factor-por-vez (OFAT), conocida como el método científico y atribuida a Francis Bacon en el siglo XVII, aunque basada en la cultura helénica. La estrategia de "un factor-por-vez" consiste en ir modificando cada vez un solo factor (variable) y distinguir los efectos que tiene sobre la respuesta cada factor por separado. Esta estrategia, a pesar de ser la más utilizada en las empresas y seguir el "método científico", es ineficiente para hallar las mejores condiciones del proceso. Dicha metodología tuvo su apogeo con Thomas Edison, aplicó la estrategia OFAT para inventar la bombilla de luz. Estas estrategias quedaron obsoletas, cuando Sir Ronald Fisher en los años 1920 descubrió un método mucho más eficiente para experimentar basado en los diseños factoriales. El diseño de experimentos fue aplicado por primera vez por Fisher en Inglaterra en la agricultura, y sus experiencias le llevaron a publicar en 1935, su libro "Design of Experiments". Sus estudios estaban centrados en mejorar la producción de patatas, trabajando para la Estación Agrícola Experimental de Rothamstede en Londres.

Un experimento diseñado se define como una prueba o serie de pruebas en las que se introducen cambios deliberados en las variables de entrada de un proceso o sistema, de manera que sea posible observar o identificar las causas de los cambios en la respuesta de salida. De una forma más esquematizada, podemos ilustrarlo de la siguiente manera:



A modo de ejemplo, podemos citar los siguientes casos, obtenidos de [2], de dónde podemos utilizar un diseño de experimentos:

– En el rendimiento de un determinado tipo de máquina (unidades producidas por día): se desea estudiar la influencia del trabajador que la maneja y la marca de la máquina.

– Se quiere estudiar la influencia de un tipo de pila eléctrica y de la marca, en la duración de las pilas.

–Una compañía telefónica está interesada en conocer la influencia de varios factores en la variable duración de una llamada telefónica. Los factores que se consideran son los siguientes: hora a la que se produce la llamada; día de la semana en que se realiza la llamada; zona de las ciudades desde la que se hace la llamada; sexo del que realiza la llamada; tipo de teléfono (público o privado) desde el que se realiza la llamada.

– Una compañía de software está interesada en estudiar la variable porcentaje en que se comprime un fichero, al utilizar un programa de comprensión teniendo en cuenta el tipo de programa utilizado y el tipo de fichero que se comprime.

–Se quiere estudiar el rendimiento de los alumnos en una asignatura y, para ello, se desean controlar diferentes factores: profesor de la asignatura; sexo del alumno, etc.

Así pues, podemos visualizar el proceso como una combinación de métodos y otros recursos que transforman una entrada (a menudo un material) en una salida, que tiene una o más respuestas observables.

Además, hay que tener en cuenta que algunas de las variables del proceso son controlables, las variables $X_1, X_2, \ldots, X_p$, mientras que otras, $Z_1, Z_2, \ldots, Z_q$, son incontrolables.

Entre los objetivos del experimento pueden incluirse, entre otros:

- Determinar las principales causas de variación en la respuesta.
- Encontrar las condiciones experimentales con las que se consigue un valor extremo en la variable de interés o respuesta.
- Comparar las respuestas en diferentes niveles de observación de variables controladas.
- Obtener un modelo estadístico-matemático que permita hacer predicciones de respuestas futuras.

La metodología del diseño de experimentos se basa en la experimentación. Es sabido que, si se repite un experimento en condiciones indistinguibles, los resultados presentan una cierta variabilidad. Si la experimentación se realiza en un laboratorio donde la mayoría de las causas de variabilidad están muy controladas, el error experimental será pequeño y habrá poca variación en los resultados del experimento. Pero si se experimenta en procesos industriales o administrativos la variabilidad será mayor en la mayoría de los casos.

Por tanto, se producen distintos resultados o salidas aun manteniendo las mismas fuentes de variación controladas, bien por la posible influencia de variables no controladas, bien por la naturaleza aleatoria de la respuesta.

El objetivo del diseño de experimentos es estudiar si cuando se utiliza un determinado tratamiento se produce una mejora en el proceso o no. Para ello se debe experimentar aplicando el tratamiento y no aplicándolo. Si la variabilidad experimental es grande, solo se detectará la influencia del uso del tratamiento cuando éste produzca grandes cambios en relación con el error de observación.

En resumen, estamos ante experimentos sujetos a **errores experimentales** y/o incertidumbre. En éstos, para la extracción de conclusiones es necesario la utilización de técnicas, tanto en el análisis de datos como en el propio diseño de la experiencia, a través del conjunto de técnicas englobadas en lo que se conoce como el diseño estadístico de experimentos.

1.1. Diseño estadístico de experimentos.

El diseño estadístico de experimentos es el proceso de planificación de un experimento para obtener datos que puedan ser analizados mediante métodos estadísticos, con el fin de obtener conclusiones válidas y objetivas. La metodología estadística es un enfoque objetivo para analizar un problema que involucre datos sujetos a errores experimentales. Por tanto, en cualquier problema experimental pueden distinguirse los siguientes aspectos:

– El diseño del experimento.

– El análisis estadístico de los datos.

Ambas cuestiones están estrechamente relacionadas, ya que el método de análisis depende directamente del diseño empleado. Para la creación del diseño, se requieren ciertos parámetros:

- **Factor:** Variable cuyo efecto experimental debe ser medido.
- **Nivel o tratamiento:** Estados o modalidades de dicho factor.
- **Unidad experimental:** Elemento del experimento sobre el que se aplica un tratamiento.
- **Bloque:** Grupo de unidades experimentales homogéneas frente a un determinado factor.
- **Error experimental:** Variación de la respuesta entre unidades experimentales tratadas de forma semejante.

1.2. Principios básicos del diseño estadístico de experimentos.

El diseño de experimentos se basa en los siguientes principios:

- **Replicación:** que consiste en la repetición un número determinado de veces del experimento bajo las mismas condiciones de las fuentes de variación controladas.
- **Aleatorización:** se entiende por Aleatorización al hecho de que tanto la asignación de material experimental como el orden en que se realizan las pruebas o ensayos individuales, se determina aleatoriamente.
- **El análisis por bloques:** es una técnica utilizada para incrementar la precisión del experimento, que consiste en dividir las unidades experimentales en subconjuntos homogéneos (bloques).

1.3. Directrices para el diseño estadístico de experimentos.

Es necesario tener con antelación una idea clara de qué es lo que se pretende, para poder determinar cómo deben ser obtenidos los datos y cómo deben ser analizados los datos. Así pues, se recomienda seguir las siguientes etapas:

1. **Reconocimiento y estado del problema.-** Un planteamiento claro del problema contribuye de forma sustancial a un mejor conocimiento del fenómeno. Es necesario plantear adecuadamente los objetivos, y tras una revisión bibliográfica de experiencias semejantes, determinar todas las partes, instrumentos y personal que va a intervenir.
2. **Elección de factores y niveles.-** Elegir los factores de interés y los niveles específicos bajo los cuales se realizará el experimento. Los diagramas causa-efecto ayudan a la identificación de factores que deben ser considerados.
3. **Selección de la variable respuesta.-** Se debe tener la certeza de que la elegida proporcione información útil sobre el proceso y sobre el objetivo principal. También ha de tenerse en cuenta la capacidad de medición, dado que un error de medida grande provocará variabilidad que se recogerá en el error experimental, y en consecuencia se necesitarán grandes diferencias entre los niveles de los factores para que éstos puedan ser considerados como significativos. En ocasiones la respuesta puede estar medida por varias variables.
4. **Elección del diseño experimental.-** Además del diseño, con la inclusión o no de bloques y de la interacción entre los factores, se ha de determinar: tamaño muestral, orden de realización del experimento y asignación del material experimental.
5. **Realización del experimento.-** Es vital vigilar el proceso cuidadosamente para asegurar que todo se haga conforme a lo planeado. En esta fase, los errores en el procedimiento pueden anular la validez experimental.
6. **Análisis de datos.-** La técnica estadística que se aplica en este tipo de estudios es el Análisis de la Varianza, es decir, descomposición de la variabilidad total en una serie de componentes asociados a cada una de las fuentes de variación consideradas en el estudio: factores controlados (principal o de bloqueo, con la posibilidad de interacción entre ellos) y el error experimental.
7. **Conclusiones y recomendaciones.-** Una vez analizados los datos, el experimentador debe extraer conclusiones prácticas de los resultados y recomendar un curso de acción.

La experimentación es un método esencial para la búsqueda del conocimiento sobre un proceso determinado, sobre el que se formulan hipótesis, se realizan experiencias para investigar dichas hipótesis y, sobre la base de los resultados, se formulan nuevas hipótesis y así sucesivamente. Es decir, tanto la experimentación como la generación de conocimiento son procesos iterativos, por lo que a veces es conveniente no invertir todos los recursos en la primera experiencia.

1.4. Conceptos generales sobre Modelos Lineales.

Las técnicas que se desarrollan en esta memoria forman parte de una teoría más amplia: El Modelo Lineal General. Por este motivo, realizaremos a continuación un breve estudio sobre el mismo. El esquema general del modelo lineal es el siguiente:

$$y = f(x_1, x_2, \ldots, x_p; \beta_1, \beta_2, \ldots, \beta_p) + \varepsilon \tag{1.1}$$

donde:

$\boldsymbol{y}$ es la variable que se desea estudiar.

$\boldsymbol{f(x_1, x_2, \ldots, x_p; \beta_1, \beta_2, \ldots, \beta_p)}$ es una función lineal de las variables $x_1, \ldots, x_p$, que representan los valores conocidos (denominadas predictoras o explicativas), y de los parámetros $\beta_1, \ldots, \beta_p$, que son desconocidos.

$\boldsymbol{\varepsilon}$ es una variable aleatoria no observable que representa la desviación o perturbación del modelo respecto de la realidad.

El modelo se dice que es lineal porque cada observación $\boldsymbol{y_i}$ es expresada como una combinación lineal de las variables $\boldsymbol{x_i}$. En el caso de tener más de una observación, $\boldsymbol{y_1, \ldots, y_n}$, entonces el modelo sería:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon_i \qquad con \quad 1 = 1, 2, ..., n. \tag{1.2}$$

donde $\boldsymbol{x_{ij}}$ representa la j-ésima coordenada para la i-ésima observación.

Además, podemos expresarlo matricialmente de manera que:

$$Y = X\beta + \varepsilon \tag{1.3}$$

donde $Y = (y_1, \ldots, y_n)^t$ es un vector compuesto por n observaciones, X es una matriz $n \times p$ de constantes conocidas, β es un vector de parámetros desconocidos de dimensión p y ε es un vector de errores aleatorios.

La matriz $X = (x_{ij})$ es usualmente denominada la matriz del diseño. Es una matriz con n filas y p columnas. En los modelos asociados al diseño de experimentos, los elementos de esta matriz tomarán sólo los valores 0 y 1.

Los elementos del vector β, el vector paramétrico, serán usualmente constantes desconocidas. Nuestro principal objetivo se centrará en estimar las componentes de β, estimar funciones de las componentes de β, y realizar contrastes sobre ellas.

1.4.1. Hipótesis distribucionales.

En el modelo lineal la i-ésima observación consta de dos componentes:

- $\sum_{j=1}^{p} x_{ij}\beta_j$.
- ε_i: error en la i-ésima observación.

Supondremos que los errores satisfacen las siguientes hipótesis:

1. Tienen media 0, $E(\varepsilon_i) = 0$, $\forall i$, y por tanto $E(Y) = X\beta$.
2. Están incorrelados, $E(\varepsilon_i \varepsilon_j) = 0$, $\forall i \neq j$
3. Sigue una distribución Normal, $\varepsilon \sim N_n(0, \sigma^2 I_n)$, lo que implica que $Y \sim N_n(X\beta, \sigma^2 I_n)$ (Hipótesis de Normalidad).
4. Tiene igual varianza , $E(\varepsilon_i) = \sigma^2$, $\forall i$ (hipótesis de homocedasticidad).
5. De (2) y (4) podemos obtener que $V(Y) = V(\varepsilon) = \sigma_2 I_n$.

1.4.2. Método de mínimos cuadrados.

Para estimar los parámetros desconocidos, β_j, podemos utilizar el método de mínimos cuadrados, que consiste en estimar $\beta_1, \ldots, \beta_p$ mediante $\widehat{\beta_1}, \ldots, \widehat{\beta_p}$ de modo que minimicen la suma de cuadrados de los residuos, donde los residuos se definen como sigue:

- Sea $\widehat{\beta}$ un vector de estimadores (puede no ser único).
- Sea $\widehat{y_i}$ el correspondiente estimador de $E(y_i)$, obtenido sustituyendo β por $\widehat{\beta}$, i.e., $\widehat{y_i} = \sum_j x_{ij}\widehat{\beta_j}$.

Entonces, se define el i-ésimo residuo como el valor real observado menos el valor estimado en el modelo, es decir: $e_i = y_i - \widehat{y_i}$. El método de mínimos cuadrados estima β mediante $\widehat{\beta}$ tal que minimice $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$, cuya solución constituye lo que se denomina el **Sistema de Ecuaciones Normales (SEN):**

$$X^t Y = X^t X \widehat{\beta} \tag{1.4}$$

La matriz $X^t X$ es una matriz $p \times p$ simétrica semidefinida (o definida) positiva con el mismo rango que X. Supondremos que $n \geq p$, y por tanto el rango de X es el número de columnas de X que son linealmente independientes.

1.4.2.1. Caso de rango total.

Este es el caso usual en los modelos de regresión múltiple y la excepción en los modelos de diseños de experimentos. No obstante lo examinaremos brevemente como introducción. Si las p columnas de X son linealmente independientes, entonces $rg(X) = rg(X^t X) = p$, por tanto existe $(X^t X)^{-1}$ y el SEN (1.4) tiene solución única dada por:

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

El estimador obtenido es insesgado y tiene por varianza $V(\hat{\beta}) = \sigma^2 (X^t X)^{-1}$

1.4.2.2. Caso singular.

Si X no es de rango total, es decir, sea r su rango tal que $(r < p)$, entonces X^tX es singular y por tanto no existe su inversa. Por lo que el sistema de ecuaciones normales (1.4) es compatible y la solución general del mismo viene dada por:

$$\tilde{\beta} = (X^tX)^-X^tY + (I - H)Z \qquad con \quad \begin{array}{l} H = (X^tX)^-(X^tX). \\ Z \in \mathbb{R}^p, arbitraria \end{array}$$

donde hemos usado la definición de **inversa generalizada de una matriz**: Sea A una matriz $n \times m$, se dice que la matriz A^- de dimensión $m \times n$ es una inversa generalizada de A si: $AA^-A = A$

1.4.3. Descomposición de la variabilidad total.

La variabilidad observada en los datos es debida a la naturaleza propia de las variables o medidas que analizamos, pero también es imputable a los niveles o tratamientos en el caso que afecten de manera desigual a la variable respuesta.

El estudio de la variabilidad permite considerar herramientas (estadísticos) que separan la variabilidad debida al azar de la variabilidad imputable a los tratamientos o niveles. Estos estadísticos se definen a partir de las variables que configuran las n observaciones.

Además, el interés reside en que para comparar los efectos de los distintos niveles de un factor se emplea la técnica estadística denominada análisis de la varianza, abreviadamente ANOVA, que está basada en la descomposición de la variabilidad total de los datos en distintas componentes.

Una medida de la variabilidad total de los datos es la **suma de cuadrados total**. Dado que:

$$Y^tY = \hat{Y}^t\hat{Y} + (Y - \hat{Y})^t(Y - \hat{Y})$$

entonces, si denotamos:

$$\begin{aligned} Y^tY &= SC_T \quad (suma\ de\ cuadrados\ total) \\ Y^t\hat{Y} &= SC_{mod} \quad (suma\ de\ cuadrados\ debido\ al\ modelo) \\ (Y - \hat{Y})^t(Y - \hat{Y}) &= SC_\varepsilon \quad (suma\ de\ cuadrados\ debido\ al\ error) \end{aligned}$$

se tiene lo que se denomina **descomposición de la variabilidad total**:

$$SC_T = SC_{mod} + SC_\varepsilon$$

Además, si $r = rg(X)$, podemos definir los cuadrados medios como:

- **El cuadrado medio debido al modelo:** $CM_{mod} = \frac{SC_{mod}}{r}$
- **El cuadrado medio debido al error:** $CM_{\varepsilon} = \frac{SC_{\varepsilon}}{n-r}$

1.4.4. Método de máxima verosimilitud.

Otro método para estimar los parámetros desconocidos β_j con $j = 1, \ldots, p$, es a través de los estimadores de máxima verosimilitud.

Bajo la hipótesis de normalidad, es decir, sea $y \sim Nn(\mu, \sigma^2 Id_{nxn})$ y dado que el vector de medias cumple que: $\mu = X\beta$, la función de verosimilitud viene dada por:

$$\mathbb{L}(\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}(Y-\mu)^t(Y-\mu)\right) \tag{1.5}$$

que tomando logaritmos de (1.5) y sustituyendo $\mu = X\beta$ nos queda:

$$\ln \mathbb{L}(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y-X\beta)^t(Y-X\beta) \tag{1.6}$$

Entonces, los EMV serán el resultado del siguiente sistema de ecuaciones:

$$\begin{aligned} X^tX\beta &= X^tY \\ n\sigma^2 &= (Y-X\beta)^t(Y-X\beta) \end{aligned}$$

obteniéndose:

$$\begin{aligned} &\widehat{\beta}_{MV} = \tilde{\beta}, \\ &\widehat{\sigma}^2_{MV} = \frac{1}{n}(Y-X\tilde{\beta})^t(Y-X\tilde{\beta}) = \frac{SC_{\varepsilon}}{n} \end{aligned}$$

Por tanto, el EMV de β coincide con el obtenido mediante el método de mínimos cuadrados, coincidiendo también el estimador de σ^2 salvo constante multiplicativa.

1.5. Diagnosis del modelo.

Como veremos en los siguientes capítulos, será de gran interés realizar comparaciones de los distintos niveles de un factor, utilizando la técnica estadística denominada análisis de la varianza o ANOVA. Para esto, hace falta que se verifiquen una serie de características del modelo tales como la normalidad de los términos de error o la independencia en los datos investigados.

Sin embargo, cuando se selecciona un modelo para un conjunto de datos, frecuentemente no se puede estar seguro a priori de que ese modelo sea adecuado. Por lo tanto, es importante examinar la adecuación del modelo a los datos antes de realizar un análisis de los mismos basado en dicho modelo.

Como hemos comentado anteriormente, las hipótesis que las muestras deben cumplir son las siguientes:

– **Hipótesis de Normalidad**.

– **Hipótesis de Independecia**.

– **Hipótesis de Homocedasticidad**.

En esta sección se presentan métodos para comprobar estas suposiciones, así como algunas soluciones que a menudo resultan útiles cuando éstas no se cumplen. A modo de ejemplo, lo haremos para el caso de un sólo factor, aunque es análogo al resto de casos. Las herramientas principales para el diagnóstico están basadas en los residuos:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.}$$

Podemos por tanto, considerar que en el estudio de un experimento se debe seguir un proceso secuencial formado por los pasos siguientes:

1. Plantear un modelo que explique los datos.
2. Examinar la adecuación del modelo planteado. Si el modelo no es adecuado, tomar las medidas correctoras, tales como empleo de transformaciones de los datos, o modificar el modelo.
3. Una vez comprobado las hipótesis necesarias se realiza el análisis estadístico de los datos.

La comprobación de la idoneidad o adecuación del modelo la realizaremos mediante varias gráficas de estos residuos.

1.5.1. Hipótesis de Normalidad.

Un método gráfico para contrastar esta hipótesis es representar los pares:

$$\left(e_{(r)}, \Phi^{-1}\left(\frac{r-0.5}{n}\right)\right), \qquad con \quad r = 1, \ldots, n$$

donde $e_{(r)}$ denota el r-ésimo residuo ordenado y Φ es la función de distribución de la $N(0,1)$.

Si la hipótesis de normalidad fuera cierta, los puntos se encontrarían en torno a una recta del tipo $y = \sigma x$. Teóricamente estarían sobre la recta pero, debido a la varibilidad muestral, esto no ocurrirá así exactamente sino que fluctuarán en torno a la misma, sobre todo para muestras pequeñas. Al visualizar esta gráfica, hay que poner más énfasis en los valores centrales que en los extremos, pues éstos presentan una mayor variabilidad. Un ejemplo de esta representación sería:

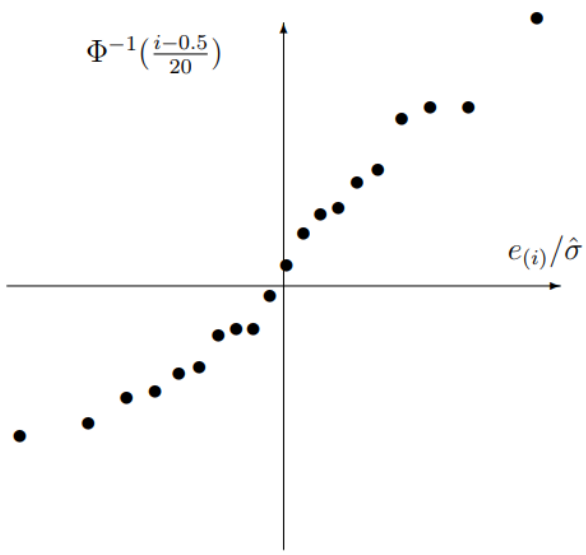


Figura 1.1: Gráfico de Normalidad.

Otra forma de contrastar la hipótesis de Normalidad es a través del siguiente test:

– **Test de Cramér-Von Mises**: Este test se basa en la relación de una función de distribución acumulada F^* supuesta, comparada con una función de distribución empírica F_n. Sean $y_1, y_2, \ldots, y_n$ los valores observados en orden creciente, entonces el estadístico que utiliza es el siguiente:

$$T = n \int_{-\infty}^{\infty} [F_n(y) - F_n^*(y)]^2 \cdot dF^*(y) = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(y_i) \right]^2$$

Si este valor es mayor que el valor tabulado, se puede rechazar la hipótesis de que los datos provienen de la distribución F.

La falta de normalidad en los errores tiene poca influencia en el contraste (que veremos más adelante) F del análisis de la varianza, y en la comparación entre medias; ya que, por el Teorema Central del Límite, su distribución será aproximadamente normal. Por tanto, los resultados de estos contrastes son sustancialmente válidos, aunque los datos sean no normales, y, en este sentido, podemos afirmar que el análisis de la varianza es una técnica robusta frente a desviaciones de la normalidad. Por ejemplo, si los errores vienen de una distribución uniforme, y hay cinco grupos con cinco observaciones por grupo, el efecto de la no normalidad se traduce en que el contrate F calculado con $\alpha = 0.05$, en la hipótesis de normalidad, tiene ahora un nivel de $\alpha = 0.053$.

Sin embargo, la falta de normalidad afecta mucho a la estimación de la varianza, y si los datos son marcadamente no normales, tendremos que desconfiar de intervalos de confianza para el error calculados a partir de la distribución χ^2. Por ejemplo, con datos procedentes de una distribución uniforme, estos intervalos son conservadores, y el 95 % cubre aproximadamente el 99.5 %. Sin embargo, si el error tiene colas más largas que la normal, la confianza del intervalo al 95 % puede caer hasta un 60 %.

1.5.2. Hipótesis de Independencia.

Cuando el experimento se ha realizado secuencialmente, los residuos deben dibujarse en función del tiempo, para detectar posibles tendencias en los datos u otros hechos inesperados.

Si la hipótesis de independencia es cierta, los puntos representados no deben mostrar patrón alguno. En caso contrario, por ejemplo, alternancia de rachas de residuos positivos y negativos, indicarían que esta hipótesis no se está cumpliendo. Una posible ilustración de esto sería la gráfica de la derecha:

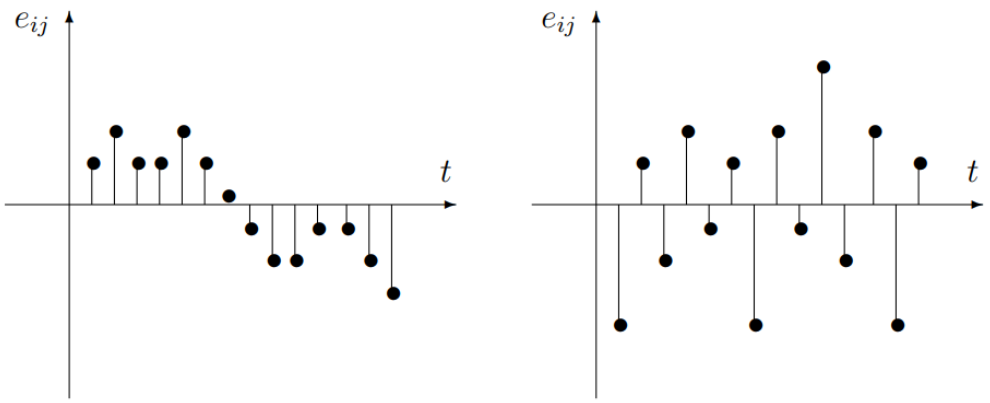


Figura 1.2: Gráfico de Independencia.

El efecto de dependencia entre las observaciones puede ser muy grave, ya que las fórmulas para las varianzas de las distribuciones muestrales de las medias $\bar{y}_{i\cdot}$ no son válidas en este caso, por lo que todos los cálculos sobre la precisión de los estimadores serán erróneos. Este problema es difícil de corregir. El modo más eficaz para prevenir la dependencia es realizar un procedimiento apropiado de aleatorización.

1.5.3. Hipótesis de Homocedasticidad.

Una manera sencilla de comprobar la hipótesis de que la variabilidad de las observaciones es la misma en todos los grupos, es representar los residuos frente a las medias de cada grupo. Un ejemplo de dicha representación sería:

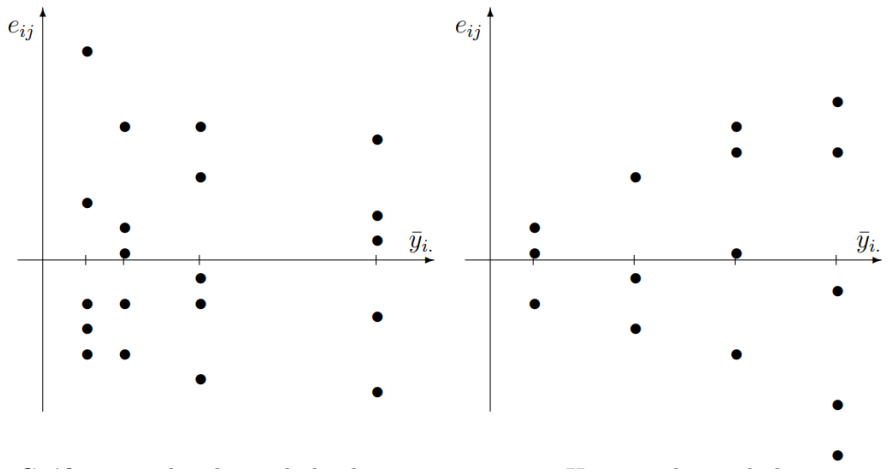


Figura 1.3: Gráfico de Homocedasticidad.

Si el aspecto de la nube de puntos se distribuye de forma aleatoria como en la primera gráfica, esto es señal de que se cumple la igualdad de varianzas. Aunque si por el contrario, observamos alguna tendencia al igual que en la segunda gráfica, ésta puede ser indicio de heterocedasticidad.

Sin embargo, la manera más segura de comprobar la homocedasticidad es a través de test de hipótesis que tienen como objetivo la comparación de las k varianzas:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$
$$H_1 : \sigma_i^2 \neq \sigma_j^2, \quad \textit{para algunos } i \neq j$$

Aunque existen varios test de este tipo, nosotros vamos a utilizar el siguiente:

– **Test de Levene**: Es un procedimiento relativamente insensible a fallos de normalidad. Para contrastar la hipótesis de homocedasticidad, este test trabaja no con las observaciones, sino con las desviaciones en valor absoluto de éstas a la media del grupo al que pertenecen. El estadístico que utiliza es el siguiente:

$$W = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i (\bar{z}_{i.} - \bar{z}_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i.})^2}, \qquad con \quad z_{ij} = |y_{ij} - \bar{y}_{i.}|$$

que bajo H_0 sigue una distribución F-Snedecor con $k-1$ y $N-k$ grados de libertad, luego la región crítica se verifica si el valor de dicho estadístico es mayor que el correspondiente valor teórico de la distribución F con $k-1$ y $N-k$ grados de libertad al nivel de significación α, es decir:

$$Rechazar\ H_0\ si\ W \geq \mathcal{F}_{k-1,N-k,1-\alpha}$$

Capítulo 2

Estudio del rendimiento escolar de los alumnos.

En este capítulo vamos a tomar el primer contacto con datos reales. Indagaremos y analizaremos a fondo cada una de las variables que nos servirán, a modo de ejemplo, para desarrollar algunos de los diseños de experimentos que veremos en el siguiente capítulo. Además, para su realización nos apoyaremos en el programa estadístico ***R***, con el interfaz gráfico ***RStudio***. Las librerías empleadas se muestran en el Apéndice de la memoria.

Para ilustrar la aplicación de las técnicas de los Diseños de Experimentos, lo ideal es realizar la experimentación tras realizar el diseño. No obstante, debido al tiempo disponible para esta memoria, es díficil encontrar un problema adecuado que cumpla estas condiciones. Por ello, se ha optado por realizar la aplicación empleando un conjunto de datos ya disponibles, en concreto los datos obtenidos en la *Encuesta Social de 2010 de Educación y Hogares en Andalucía.*

2.1. Presentación del estudio.

Se trata de una encuesta que se encuadra dentro del objetivo estadístico específico de suministrar información sobre la educación en Andalucía. El objetivo principal es encontrar los factores que influyen en el rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz. Para ello se entrevistaron a padres e hijos residentes en viviendas familiares con menores nacidos en 1994 y/o 1998. La información difundida recoge las principales características sociodemográficas y del entorno familiar, escolar y social del colectivo considerado.

Este estudio sociodemográfico es de dominio público, y su extracción se puede obtener fácilmente a través de la página oficial del Instituto Nacional de Estadística (INE):

https://www.juntadeandalucia.es/institutodeestadisticaycartografia/encsocial/2010/index.htm

Cuenta con un fichero de microdatos a modo de guía en formato PDF, donde aparece explicado cada una de las variables sociodemográficas y la metodología usada en el estudio.

Por otro lado, cuenta también con los ficheros que contienen las bases de datos con las cohortes de los menores nacidos en 1998 y en 1994 en formato *SPSS*. Basaremos nuestra memoria en este último, en la cohorte de menores nacidos en 1994, estudio que contiene 2.584 encuestas completas (cuestionarios hogar + padres + hijos).

2.2. Importación de los datos y obtención de las variables.

Comenzaremos leyendo la base de datos y creando el *Dataframe* a partir de ésta, con el que podamos trabajar:

```
data<-read.spss("CHIJOS_NACIDOS_1994.sav")
datos<-as.data.frame(data)
dim(datos)
```

```
## [1] 2584  429
```

Vemos que la base de datos consta de 429 variables con 2584 observaciones cada una. Dado que la cantidad de variables distintas es cuantiosa y, por tanto, sería improductivo trabajar con cada una de éstas, nos vemos obligados a trabajar sólo con algunas de ellas.

A continuación, pasaremos a extraer las variables seleccionadas que nos servirán de apoyo durante toda la memoria.

2.2.1. Nota media de los alumnos.

La variable *Nota media de los alumnos* es la más importante del estudio puesto que, a partir de ésta, mediremos el rendimiento escolar de los alumnos y alumnas nacidos en 1994 que asisten a centros educativos del territorio andaluz. Crearemos esta variable a través de la media aritmética de las notas obtenidas en 3º de ESO de las asignaturas de:

– Ciencias Naturales.

– Ciencias Sociales, Geografía e Historia.

– Inglés.

– Lengua Castellana y Literatura.

– Matemáticas.

Cabe añadir que cada una de las notas obtenidas para cada asignatura están evaluadas por enteros del 1 al 10, pero también existen otros valores en las observaciones que nos dificultan el estudio. Se trata de aquellos alumnos que o no se presentaron a alguno de los exámenes, o bien tienen valores perdidos. A modo de ejemplo, veamos como vienen determinados las notas de *Ciencias Naturales* de los alumnos de 3º de ESO:

```
Nota1<- datos$CN08
summary(Nota1)
```

```
## -1            1              10             2              3
##           300            158             69            103            114
## 4             5              6              7              8
##            41            558            370            263            180
## 9             No Presentado
##           137            291
```

Al ser la variable *Nota media* una variable cuantitativa, lo mejor será excluir estos casos problemáticos del estudio. Puesto que las notas de aquellos alumnos que no se presentaron vienen expresados por un *"NA"*, y las notas de los alumnos con valores perdidos por un *"-1"*, bastará con transformar los valores *"-1"* en valores *"NA"*; y eliminar, más adelante, todas las observaciones que tengan éstos valores.

Pasemos a la lectura y obtención de la variable Nota media de los alumnos:

```
Nota1 <-as.numeric(as.character(Nota1))
Nota2 <- as.numeric(as.character(datos$CSGH08))
Nota3 <- as.numeric(as.character(datos$Ingles08))
Nota4 <- as.numeric(as.character(datos$LCL08))
Nota5 <- as.numeric(as.character(datos$Mates08))
#A)Para Ciencias Naturales.
Nota1[Nota1 == "-1"] <- "NA"
Nota1<-as.numeric(as.character(Nota1))
#B)Para Ciencias Sociales, Geografía e Historia.
Nota2[Nota2 == "-1"] <- "NA"
Nota2<-as.numeric(as.character(Nota2))
#C)Para Inglés.
Nota3[Nota3 == "-1"] <- "NA"
Nota3<-as.numeric(as.character(Nota3))
#D)Para Lengua Castellana y Literatura.
Nota4[Nota4 == "-1"] <- "NA"
Nota4<-as.numeric(as.character(Nota4))
#E)Para Matemáticas.
Nota5[Nota5 == "-1"] <- "NA"
Nota5<-as.numeric(as.character(Nota5))
```

Por lo que, una vez que hemos depurado los valores perdidos, pasaremos a la creación de la variable continua *Nota media* de los alumnos de 3º de ESO:

```
NotaMedia<-(Nota1+Nota2+Nota3+Nota4+Nota5)/5
```

Análisis descriptivo.

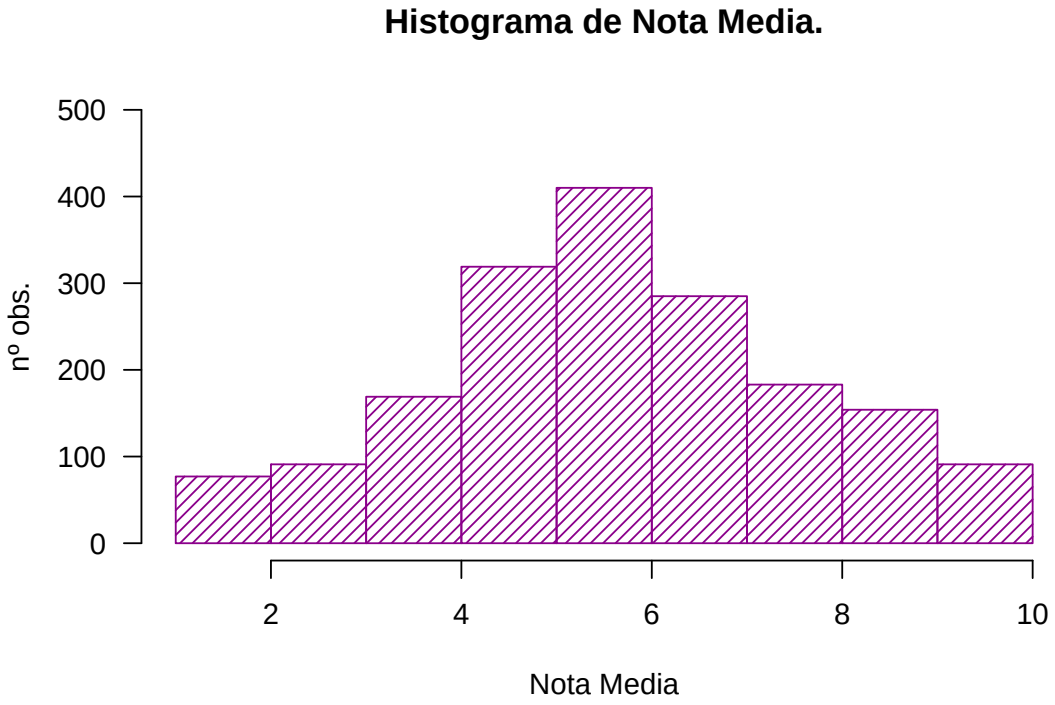
Procedemos a realizar un análisis descriptivo de nuestra variable de interés: Nota media de los alumnos de Andalucía nacidos en 1994 obtenidas en 3º de ESO. Este análisis nos permitirá controlar la presencia de posibles errores, conocer el número de valores perdidos y, además, nos proporcionará una idea de la forma que tienen los datos a través de medidas tales como los cuartiles, la media y la mediana:

```
summary(NotaMedia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.600   5.600   5.726   7.000  10.000     805
```

Al ser una variable cuantitativa, podemos realizar el correspondiente histograma que refleja la frecuencia absoluta, es decir, el número de observaciones, para cada intervalo. Si lo representamos, nos queda:

```
Histograma<- hist(NotaMedia,col=colors()[84],frequency=1,
                  las=1,density=20, xlab="Nota Media", ylab="nº obs.",
                  ylim = c(0, 500), main= "Histograma de Nota Media.")
```

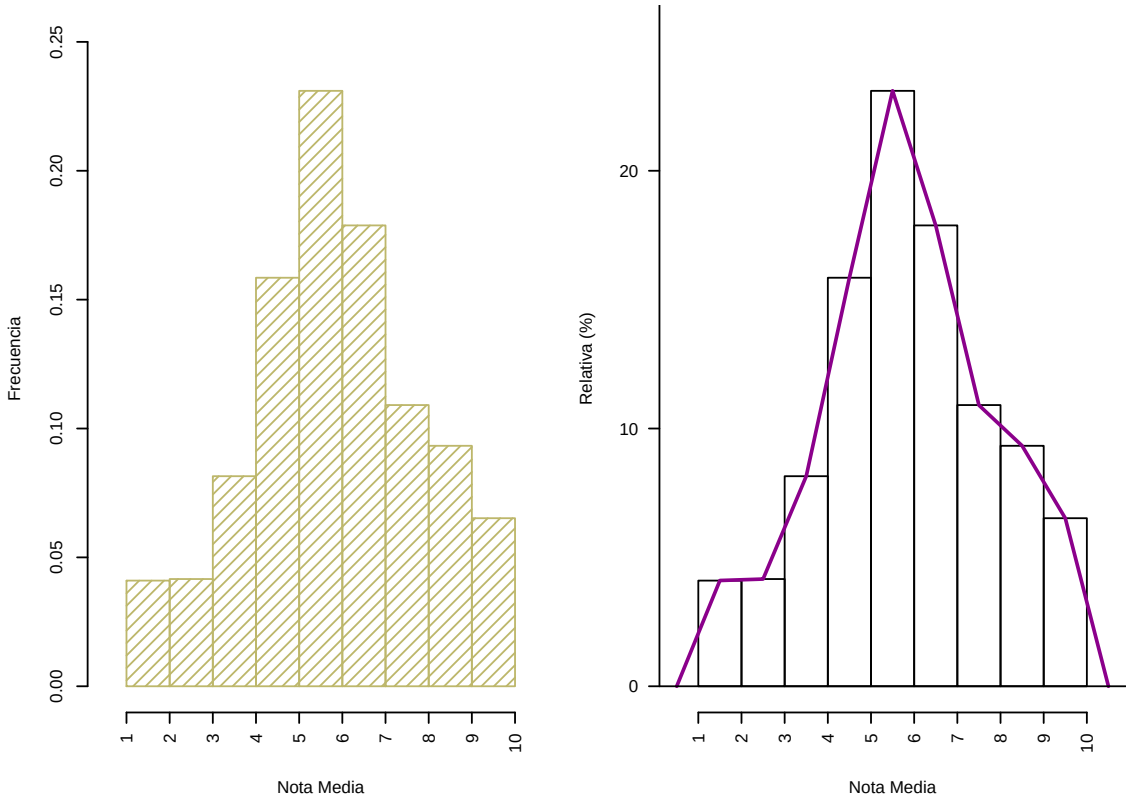


Vemos que la mayor parte de las observaciones, pertenecen al intervalo de nota media comprendido entre el 5 y el 6, mientras que disminuye (por lo general) al alejarse de éste.

También podemos realizar una representación de la frecuencia relativa tal que:

```
par(mfrow=c(1,2),mar=c(4,4,0,1),cex=0.6)
breaks2= c( 1,  2,  3,  4,  5,  6,  7,  8,  9, 10)
h1<- graph.freq(NotaMedia,col=colors()[83],frequency=2,
                las=3,density=20,xlab="Nota Media", ylab="Frecuencia",
                breaks =breaks2)
x<-h1$breaks
h2<- plot(h1, frequency =2, axes= FALSE,xlab="Nota Media",
          ylab="Relativa (%)")
polygon.freq(h2, col=colors()[84], lwd=2, frequency =2)
axis(1,x,cex=0.6,las=2)
y<-seq(0,0.4,0.1)
```

```
axis(2, y,y*100,cex=0.6,las=1)
```



2.2.2. Nota de Matemáticas de los alumnos.

Será de gran interés realizar también el análisis de una variable que no sea absolutamente continua, como por ejemplo la variable discreta: Nota de Matemáticas de los alumnos. Así podremos analizar el rendimiento de los alumnos en matemáticas y, además, saber cómo actuar en el caso de que nos encontremos frente a un estudio donde falle la hipótesis de Normalidad.

Al igual que en el apartado anterior, las notas de matemáticas están evaluadas por enteros del 1 al 10, existiendo también observaciones que nos dificultarán el estudio, como los valores perdidos *"No Consta"* o los *"No Presentado"*. Veamos un breve resumen:

```
NotaMates <- datos$Mates08
summary(NotaMates)
```

```
## -1              1             10              2              3
##             299            175             60            118            148
## 4               5              6              7              8
##              41            608            362            221            162
## 9              No Presentado
##             106            284
```

Como vimos anteriormente, debemos excluir los casos problemáticos citados del estudio. Actuaremos de forma análoga:

```
NotaMates<-as.numeric(as.character(NotaMates))
NotaMates[NotaMates == "-1"] <- "NA"
NotaMates<-as.numeric(as.character(NotaMates))
```

Análisis descriptivo.

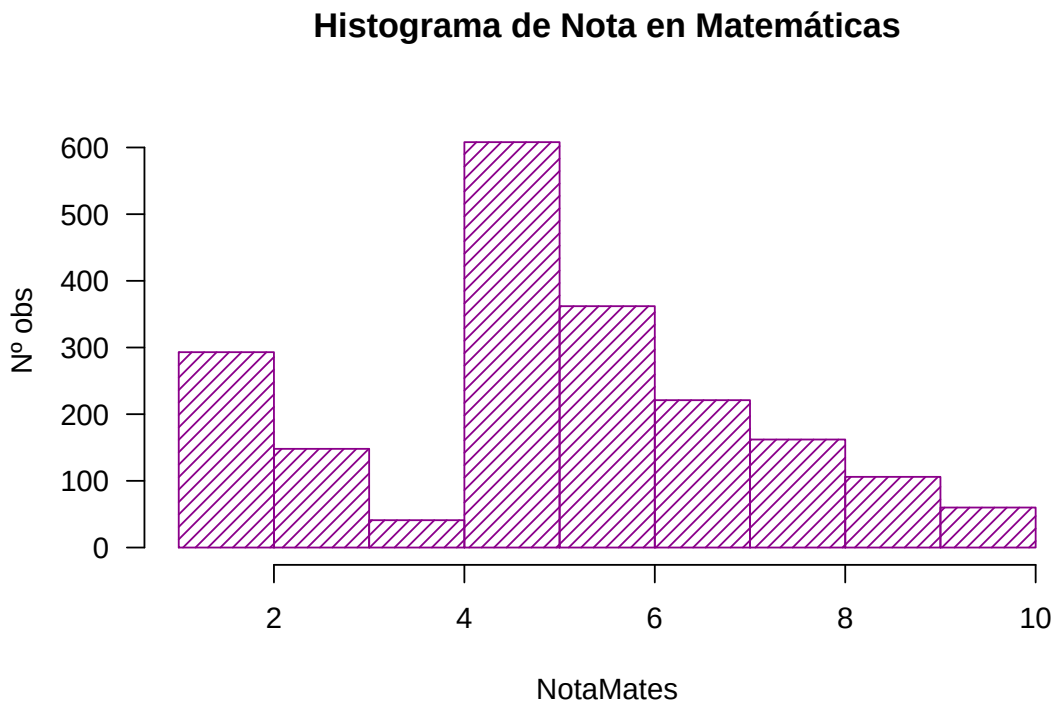
Procedemos a realizar un análisis descriptivo de nuestra variable cuantitativa de interés: Nota de Matemáticas de los alumnos de Andalucía nacidos en 1994 obtenidas en 3º de ESO. Tenemos que:

```
summary(NotaMates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   5.000   5.000   5.311   7.000  10.000     583
```

A continuación, realizamos el correspondiente histograma:

```
Histograma<- hist(NotaMates,col=colors()[84],frequency=1,las=1,density=20,
                  ylim=c(0,650), breaks = breaks2, ylab="Nº obs",
                  main= "Histograma de Nota en Matemáticas")
```



Observamos que la mayoría de los alumnos han obtenido una calificación con un valor de 5 en matemáticas y, por el contrario, la menos obtenida ha sido con un valor de 4.

2.2.3. Nivel de estudios máximo de los padres/tutores.

La segunda variable que usaremos durante toda la memoria es la variable cualitativa ordinal nivel de estudios máximo de los padres/tutores. Por ejemplo, si en un hogar la madre del menor tiene estudios superiores y el padre ha finalizado hasta el bachillerato, este hogar contabiliza como un hogar con estudios superiores.

Esta variable sociodemográfica nos permitirá conocer en qué tipo de hogar se ha criado un alumno o alumna, ya sea por ejemplo en uno donde ambos padres no tengan ningún tipo de estudio o, por el contrario, en el que al menos uno de ellos haya obtenido un título de licenciatura.

Los distintos valores que puede tomar esta variable son:

```
estudios <- datos$STUDIOSC
levels(estudios)
```

```
## [1] "No consta"
## [2] "Sin educación formal o inferior primaria"
## [3] "Educación primaria"
## [4] "Enseñanza secundaria de 1º etapa"
## [5] "Educación secundaria de 2ª etapa"
## [6] "Educación post secundaria pero no terciaria"
## [7] "Estudios universitarios"
```

```
levels(estudios)<-c("NC","SE","EPr","ES1","ES2","EPS","EU")
summary(estudios)
```

```
##  NC  SE EPr ES1 ES2 EPS  EU
##   2  96 306 969 559 182 470
```

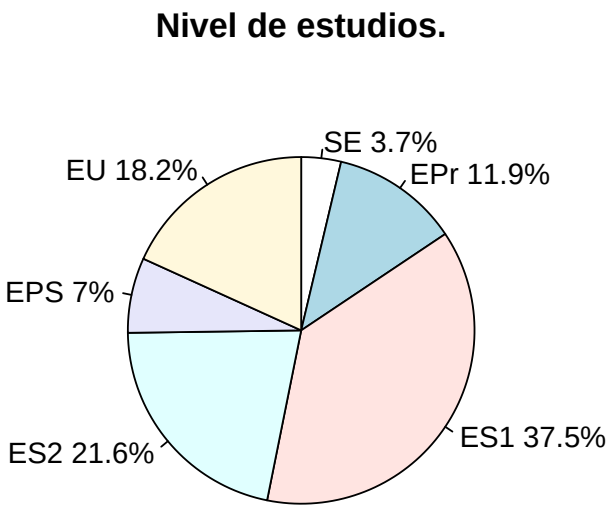
Al igual que antes, nos encontramos con valores perdidos "No Consta". Por lo que, actuaremos de forma análoga a los apartados anteriores:

```
Estudios<-as.character(estudios)
Estudios[Estudios == "NC"] <- "NA"
```

Análisis descriptivo.

Para el caso de una variable cualitativa, podemos representarla a través del correspondiente diagrama de sectores:

```
par(mfrow=c(1,1))
Cuadro3<-table(estudios)
c3porc<-(round(((Cuadro3/margin.table(Cuadro3))*100),1))
etiquetas<-c("SE","EPr","ES1","ES2","EPS","EU")
etiquetas<-paste(etiquetas, c3porc)
etiquetas<-paste(etiquetas,"%",sep="")
pie(c3porc,labels = etiquetas, clockwise = TRUE,
    main = "Nivel de estudios.")
```



Observamos que la mayor parte de alumnos, en concreto el 37.5 %, tienen padres o tutores con nivel máximo de estudios *"Enseñanza secundaria de 1º etapa"*, mientras que sólo el 3.7 % de ellos, tienen padres *"Sin educación formal o inferior primaria"*.

Por otra parte, será de interés para los estudios que realizaremos en capítulos posteriores, conocer el valor de la media de la variable *Nota media*, en función de cada uno de los niveles de la variable *Nivel de estudios máximo de los padres/tutores.* Así pues, pasaremos a realizar una tabla compuesta con cada una de las medias mencionadas:

```
DataEstudios<-cbind(NotaMedia,Estudios)
DataEstudios<- as.data.frame(DataEstudios)
DataEstudios$NotaMedia <-as.numeric(as.character(NotaMedia))
DataEstudios<-na.omit(DataEstudios)
levels(DataEstudios$Estudios)
```

```
## [1] "EPr" "EPS" "ES1" "ES2" "EU"  "NA"  "SE"
```

```
A<-mean(DataEstudios[which(DataEstudios$Estudios==
                           levels(DataEstudios$Estudios)[7]),]$NotaMedia)
B<-mean(DataEstudios[which(DataEstudios$Estudios==
                           levels(DataEstudios$Estudios)[1]),]$NotaMedia)
C<-mean(DataEstudios[which(DataEstudios$Estudios==
                           levels(DataEstudios$Estudios)[3]),]$NotaMedia)
D<-mean(DataEstudios[which(DataEstudios$Estudios==
                           levels(DataEstudios$Estudios)[4]),]$NotaMedia)
E<-mean(DataEstudios[which(DataEstudios$Estudios==
                           levels(DataEstudios$Estudios)[2]),]$NotaMedia)
F<-mean(DataEstudios[which(DataEstudios$Estudios==
                           levels(DataEstudios$Estudios)[5]),]$NotaMedia)
```

```
Medias<-c(A,B,C,D,E,F)
Tratamientos<-c("SE","EPr", "ES1","ES2","EPS","EU")
TablaMedias = data.frame(rbind(Tratamientos,Medias))
colnames(TablaMedias)<-NULL
TablaMedias
```

```
##
## Tratamientos                SE               EPr               ES1
## Medias       4.24285714285714 5.05487804878049 5.26576980568012
##
## Tratamientos               ES2               EPS                EU
## Medias       5.92281553398058 6.06620689655172 6.73179190751445
```

2.2.4. Lugar de residencia del alumno.

La siguiente variable que extraeremos será la variable cualitativa nominal lugar de residencia del alumno. Puesto que este es un estudio de Educación y hogares en Andalucía, los posibles valores que pueden tomar son las 8 provincias andaluzas.

```
Provincia <- datos$CPRO
levels(Provincia)
```

```
## [1] "Almería" "Cádiz"   "Córdoba" "Granada" "Huelva"  "Jaén"    "Málaga"
## [8] "Sevilla"
```

Esta variable es de gran interés puesto que, nos permitirá saber en qué provincia reside el alumno en cuestión. Sin embargo, por falta de observaciones en alguno de los cruces necesarios en análisis posteriores, es necesario agrupar algunas de estas provincias. Por lo que, los distintos valores que puede tomar esta variable son:

- Almería.
- Cádiz.
- Córdoba y Jaén.
- Granada
- Sevilla y Huelva.
- Málaga.

```
levels(Provincia)<-c("Alm","Cad","CoryJae","Gra","SevyHue","CoryJae",
                     "Mal", "SevyHue")
summary(Provincia)
```

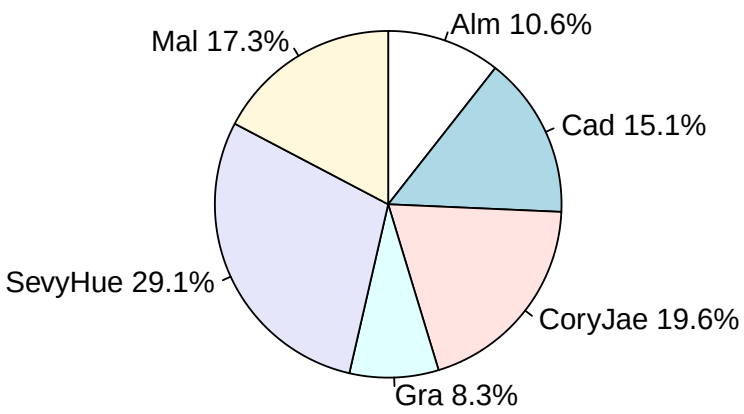
```
##     Alm     Cad CoryJae     Gra SevyHue     Mal
##     273     390     507     215     751     448
```

```
Provincia<-as.character(Provincia)
```

Análisis descriptivo.

Análogamente, al ser una variable cualitativa, la representaremos a través del diagrama de sectores:

```
par(mfrow=c(1,1))
Cuadro3<-table(Provincia)
c3porc<-(round(((Cuadro3/margin.table(Cuadro3))*100),1))
etiquetas<-c("Alm","Cad","CoryJae","Gra","SevyHue","Mal")
etiquetas<-paste(etiquetas, c3porc)
etiquetas<-paste(etiquetas,"%",sep="")
pie(c3porc,labels = etiquetas, clockwise = TRUE,
    main = "Lugar de residencia del alumno.")
```



Por secciones, vemos que la mayor parte de los alumnos del estudio residen en Sevilla junto con Huelva (con un 29.1 %), y los que aparecen en menor porcentaje que el resto son aquellos que residen en Granada (con un 8.3 %).

Al igual que en el apartado anterior, pasaremos a obtener una tabla compuesta con cada una de las medias de la variable *Nota media*, en función de cada uno de los niveles de la variable *Lugar de residencia del alumno.*

```
DataProvincia<-cbind(NotaMedia,Provincia)
DataProvincia<- as.data.frame(DataProvincia)
DataProvincia$NotaMedia <-as.numeric(as.character(NotaMedia))
DataProvincia<-na.omit(DataProvincia)
levels(DataProvincia$Provincia)
```

```
## [1] "Alm"     "Cad"     "CoryJae" "Gra"     "Mal"     "SevyHue"
```

```
A<-mean(DataProvincia[which(DataProvincia$Provincia==
                            levels(DataProvincia$Provincia)[1]),]$NotaMedia)
B<-mean(DataProvincia[which(DataProvincia$Provincia==
                            levels(DataProvincia$Provincia)[2]),]$NotaMedia)
C<-mean(DataProvincia[which(DataProvincia$Provincia==
                            levels(DataProvincia$Provincia)[3]),]$NotaMedia)
D<-mean(DataProvincia[which(DataProvincia$Provincia==
                            levels(DataProvincia$Provincia)[4]),]$NotaMedia)
E<-mean(DataProvincia[which(DataProvincia$Provincia==
                            levels(DataProvincia$Provincia)[5]),]$NotaMedia)
F<-mean(DataProvincia[which(DataProvincia$Provincia==
                            levels(DataProvincia$Provincia)[6]),]$NotaMedia)
Medias<-c(A,B,C,D,E,F)
Tratamientos<-c("Alm","Cad","CoryJae","Gra","SevyHue","Mal")
TablaMedias = data.frame(rbind(Tratamientos,Medias))
colnames(TablaMedias)<-NULL
TablaMedias
```

```
##
## Tratamientos              Alm              Cad          CoryJae
## Medias       5.73296089385475 5.71927272727273 5.95297157622739
##
## Tratamientos              Gra          SevyHue              Mal
## Medias       6.13818181818182 5.51170568561873 5.53459915611814
```

2.2.5. Sexo del alumno.

La última variable a tener en cuenta será el sexo de los alumnos. Esta es también una variable cualitativa nominal que toma los valores: "Hombre" o "Mujer". Veámoslo:

```
sexo <- datos$SEXO_EGO
levels(sexo)
```

```
## [1] "Hombre" "Mujer"
```

```
levels(sexo)<-c("H","M")
summary(sexo)
```

```
##    H    M
## 1355 1229
```

```
Sexo<-as.character(sexo)
```

Análisis descriptivo.

Análogamente, al ser una variable cualitativa, la representaremos a través del diagrama de sectores:

```
par(mfrow=c(1,1))
Cuadro3<-table(sexo)
c3porc<-(round(((Cuadro3/margin.table(Cuadro3))*100),1))
```

```
etiquetas<-c("Hombre","Mujer")
etiquetas<-paste(etiquetas, c3porc)
etiquetas<-paste(etiquetas,"%",sep="")
pie(c3porc,labels = etiquetas, clockwise = TRUE,
    main = "Sexo del alumno.")
```

Observamos que la mayor parte de los alumnos del estudio son "hombres" o "chicos" (con un 52.4 %) y, por tanto, la menor parte son "mujeres" o "chicas" (con un 47.6 %).

Por último, pasaremos a obtener una tabla compuesta con cada una de las medias de la variable *Nota media*, en función de cada uno de los niveles de la variable *Sexo del alumno*.

```
DataSexo<-cbind(NotaMedia,Sexo)
DataSexo<- as.data.frame(DataSexo)
DataSexo$NotaMedia <-as.numeric(as.character(NotaMedia))
DataSexo<-na.omit(DataSexo)
levels(DataSexo$Sexo)
```

```
## [1] "H" "M"
```

```
A<-mean(DataSexo[which(DataSexo$Sexo==
                       levels(DataSexo$Sexo)[1]),]$NotaMedia)
B<-mean(DataSexo[which(DataSexo$Sexo==
                       levels(DataSexo$Sexo)[2]),]$NotaMedia)


Medias<-c(A,B)
Tratamientos<-c("Hombre","Mujer")
TablaMedias = data.frame(rbind(Tratamientos,Medias))
```

```
colnames(TablaMedias)<-NULL
TablaMedias
```

```
##
## Tratamientos           Hombre            Mujer
## Medias       5.48552486187845 5.97551487414188
```

2.3. Proceso de depuración de los datos y creación de la muestra.

Como ya aclaramos al principio, debido al factor tiempo, es inviable para esta memoria realizar primero el diseño y esperar a obtener los resultados para su posterior análisis. Es necesario aclarar que nuestro objetivo no es obtener resultados sobre un estudio concreto, sino más bien establecer un guion con ejemplos reales que permita al lector poder realizar su propio estudio basado en el diseño de experimentos. Es por eso que es necesario simular la obtención de los datos a través de muestras de nuestra población: *alumnos y alumnas de la Encuesta Social de 2010 de Educación y Hogares en Andalucía*, para poder así explicar posteriormente la metodología y el análisis necesario. A continuación, pasaremos a crear las distintas muestras que usaremos a lo largo de la memoria.

NOTA: Debido a la traba que supondría para el lector la gran cantidad de códigos necesarios para obtener las distintas muestras de nuestro estudio, la obtención detallada de la muestra se dejará como lectura opcional en el Apéndice del final de la memoria.*

2.3.1. Muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres.

Nuestro objetivo será crear una muestra ponderada de tamaño 400 donde aparezcan observaciones de los alumnos con su correspondiente *Nota Media* y el correspondiente *Nivel de estudios máximo de los padres o tutores*. Lo primero que haremos es crear el *Dataframe* con los datos de la Nota media del alumno y Nivel de estudios máximo de los padres para cada una de las observaciones posibles:

```
Datos1factor<-cbind(Estudios,NotaMedia)
Datos1factor<- as.data.frame(Datos1factor)
Datos1factor$NotaMedia <-as.numeric(as.character(NotaMedia))
```

Por otra parte, para la aleatoriedad de la muestra añadiremos a nuestro *Dataframe* una tercera columna formada por valores aleatorios de una Uniforme con parámetros 0 y 1 (i.e U(0,1)), y lo ordenaremos según éstos.

Además, como expusimos al principio del capítulo, lo más óptimo es descartar aquellas observaciones que tengan valores perdidos. Por lo que también procederemos a la eliminación de estos.

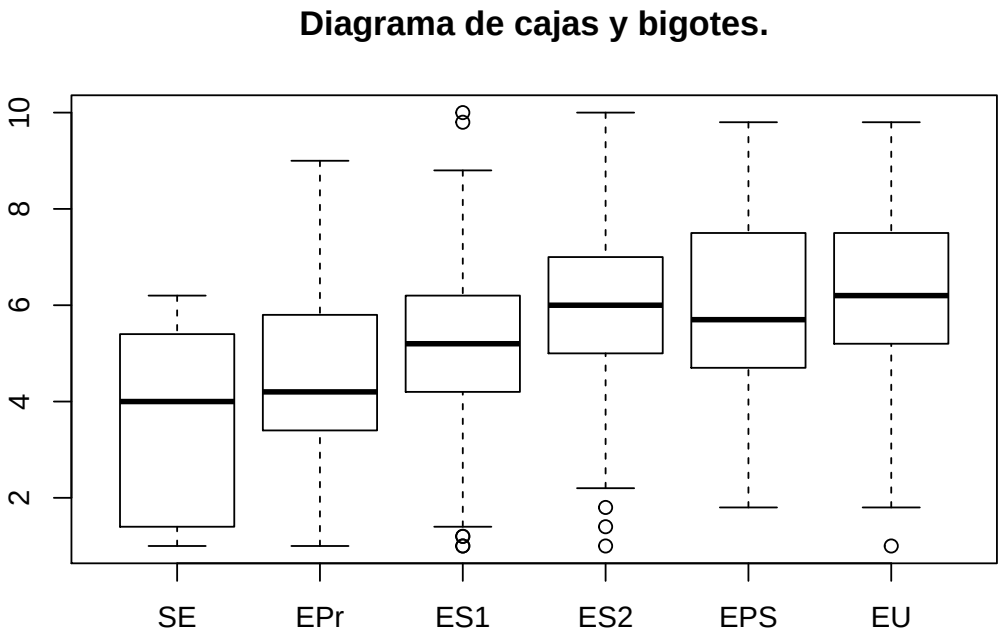
Una vez realizado estos pasos, nos queda la siguiente muestra:

```
summary(Muestra1factor)
```

```
##  Estudios    NotaMedia
##  SE :  9   Min.   : 1.00
##  EPr: 37   1st Qu.: 4.40
##  ES1:150   Median : 5.60
##  ES2: 93   Mean   : 5.54
##  EPS: 32   3rd Qu.: 6.80
##  EU : 79   Max.   :10.00
```

Podemos representar estos resultados a través del diagrama de cajas y bigotes, que permite realizar el estudio de los valores atípicos de la *Nota Media* del alumno para cada uno de los tratamientos del *Nivel de estudios máximo de los padres o tutores*, de modo que:

```
par(mfrow=c(1,1))
boxplot(Muestra1factor$NotaMedia~Muestra1factor$Estudios,
        main="Diagrama de cajas y bigotes.")
```



2.3.2. Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

Para este caso obtendremos una muestra balanceada más pequeña, concretamente de tamaño 30, dónde aparezcan observaciones de los alumnos con su correspondiente

Nota Media, *Nivel de estudios máximo de los padres o tutores* y *Lugar de residencia*. Comenzaremos con la creación del *Dataframe*:

```
DatosBIB<-cbind(NotaMedia,Estudios,Provincia)
DatosBIB<- as.data.frame(DatosBIB)
DatosBIB$NotaMedia <-as.numeric(as.character(NotaMedia))
```

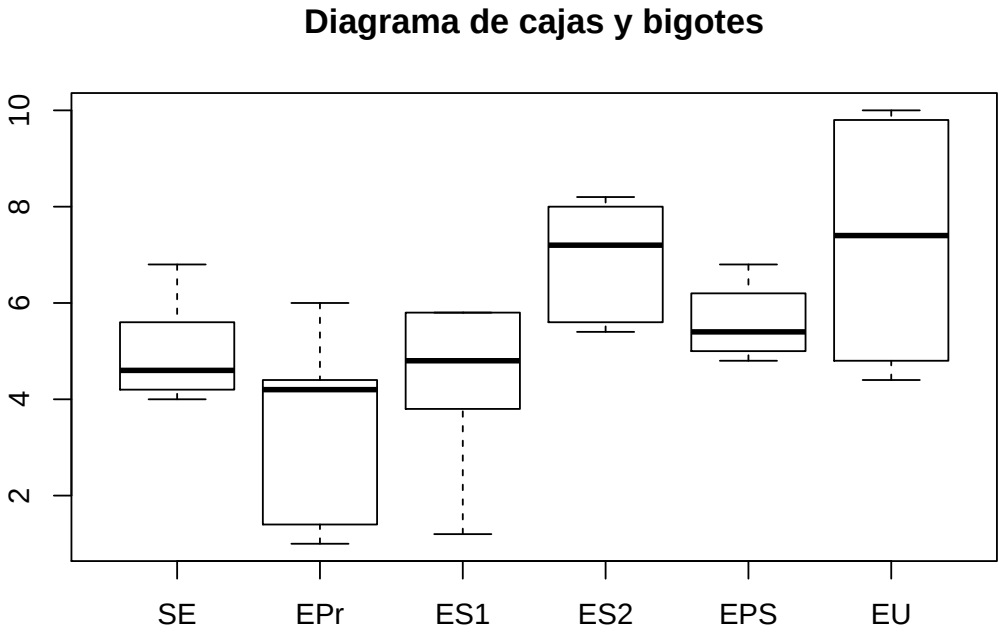
Análogamente al apartado anterior, pasemos a obtener dicha muestra:

```
summary(MuestraBIB)
```

```
##    NotaMedia      Estudios   Provincia
##  Min.   : 1.00   SE :5    Alm    :5
##  1st Qu.: 4.40   EPr:5    Cad    :5
##  Median : 5.40   ES1:5    CoryJae:5
##  Mean   : 5.42   ES2:5    Gra    :5
##  3rd Qu.: 6.65   EPS:5    Mal    :5
##  Max.   :10.00   EU :5    SevyHue:5
```

Al igual que en al apartado anterior, representaremos estos resultados a través del diagrama de cajas y bigotes para así visualizar los valores observados atípicos de la *Nota Media* del alumno para cada uno de los tratamientos del *Nivel de estudios máximo de los padres o tutores*.

```
par(mfrow=c(1,1))
boxplot(MuestraBIB$NotaMedia~MuestraBIB$Estudios,
        main="Diagrama de cajas y bigotes")
```



2.3.3. Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Sexo del alumno.

A continuación, obtendremos una muestra balanceada de tamaño 400 donde aparezcan observaciones de los alumnos con sus correspondiente *Nota Media*, *Nivel de estudios máximo de los padres o tutores* y *Sexo del alumno.*

```
Datos2factores<-cbind(NotaMedia,Estudios,Sexo)
Datos2factores<- as.data.frame(Datos2factores)
Datos2factores$NotaMedia <-as.numeric(as.character(NotaMedia))
```

Cabe añadir, que debido al bajo número de observaciones con padres cuyo nivel máximo de estudio sea el valor *"Sin educación formal o inferior primaria"* y, por tanto, a la imposibilidad de realizar algunos cruces, nos vemos obligados a anexionarlo para esta muestra junto al nivel *"Educación primaria"*, creando así el nuevo nivel *"Estudios Básicos"*:

```
summary(Datos2factores)
```

```
##    NotaMedia      Estudios  Sexo
##  Min.   : 1.000   EBa:206   H:904
##  1st Qu.: 4.600   ES1:669   M:874
##  Median : 5.600   ES2:412
##  Mean   : 5.725   EPS:145
##  3rd Qu.: 7.000   EU :346
##  Max.   :10.000
```

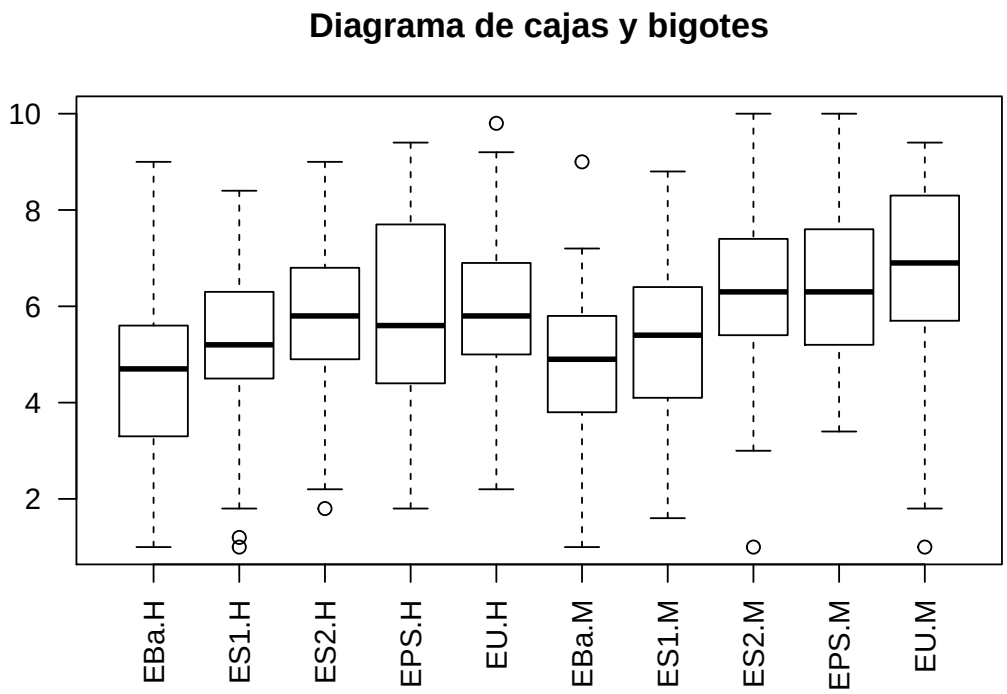
Análogamente al resto de apartados, pasemos a obtener dicha muestra:

```
summary(Muestra2factores)
```

```
##    NotaMedia      Estudios Sexo
##  Min.   : 1.000   EBa:80   H:200
##  1st Qu.: 4.550   ES1:80   M:200
##  Median : 5.600   ES2:80
##  Mean   : 5.676   EPS:80
##  3rd Qu.: 6.800   EU :80
##  Max.   :10.000
```

A continuación, pasemos a la representación de estos resultados a través del diagrama de cajas y bigotes de la *Nota Media* del alumno, para cada uno de los tratamientos conjuntos entre el *Nivel de estudios máximo de los padres o tutores* y *Sexo del alumno.*

```
par(mfrow=c(1,1))
boxplot(Muestra2factores$NotaMedia~
          Muestra2factores$Estudios*Muestra2factores$Sexo,las=2,
          main="Diagrama de cajas y bigotes")
```



2.3.4. Muestra balanceada: Nota en matemáticas del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

Por último, obtendremos una muestra balanceada de tamaño 36 donde aparezcan observaciones de los alumnos con su correspondiente *Nota en Matemáticas*, *Nivel de estudios máximo de los padres o tutores* y *Lugar de residencia del alumno.*

```
DatosBAC<-cbind(NotaMates,Estudios,Provincia)
DatosBAC<- as.data.frame(DatosBAC)
DatosBAC$NotaMates <-as.numeric(as.character(NotaMates))
```

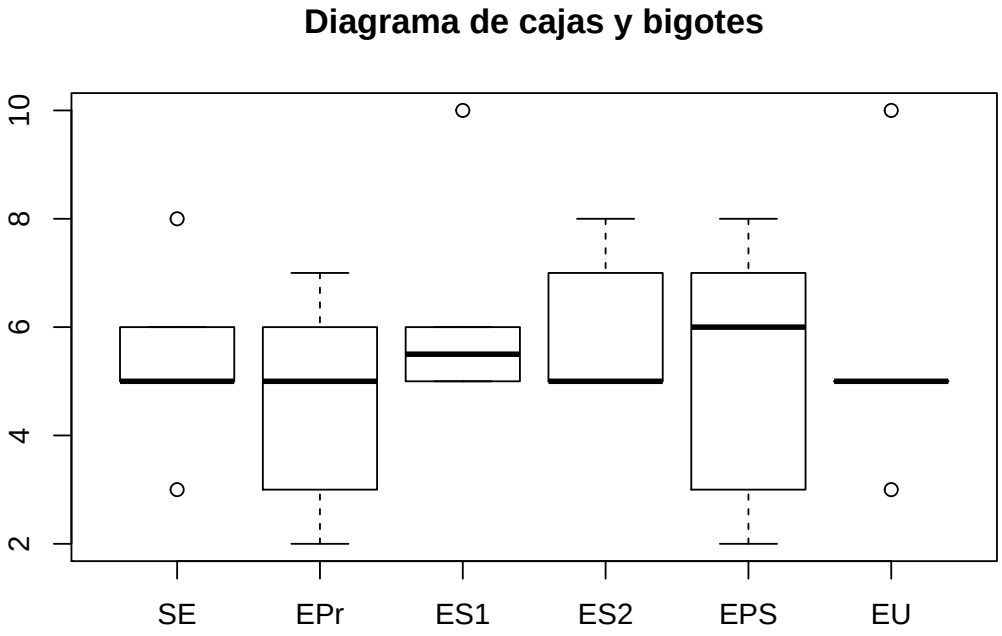
Como realizamos anteriormente con el resto de apartados, pasaremos a obtener dicha muestra. Análogamente:

```
summary(MuestraBAC)
```

```
##    NotaMates      Estudios   Provincia
##  Min.   : 2.000   SE :6    Alm    :6
##  1st Qu.: 5.000   EPr:6    Cad    :6
##  Median : 5.000   ES1:6    CoryJae:6
##  Mean   : 5.472   ES2:6    Gra    :6
##  3rd Qu.: 6.250   EPS:6    Mal    :6
##  Max.   :10.000   EU :6    SevyHue:6
```

Para finalizar, pasaremos a la representación de estos resultados a través del diagrama de cajas y bigotes de la *Nota en Matemáticas* del alumno, para cada uno de los tratamientos del *Nivel de estudios máximo de los padres o tutores*, de modo que:

```
par(mfrow=c(1,1))
boxplot(MuestraBAC$NotaMates~MuestraBAC$Estudios,
          main="Diagrama de cajas y bigotes")
```



2.4. Diagnosis de los modelos anteriores.

En esta sección, pasaremos a comprobar cada una de las hipótesis vistas en el capítulo 1 para cada una de las muestras. Para ello, nos basaremos en las correspondientes gráficas, así como de los debidos test cuando esto sea posible.

2.4.1. Normalidad

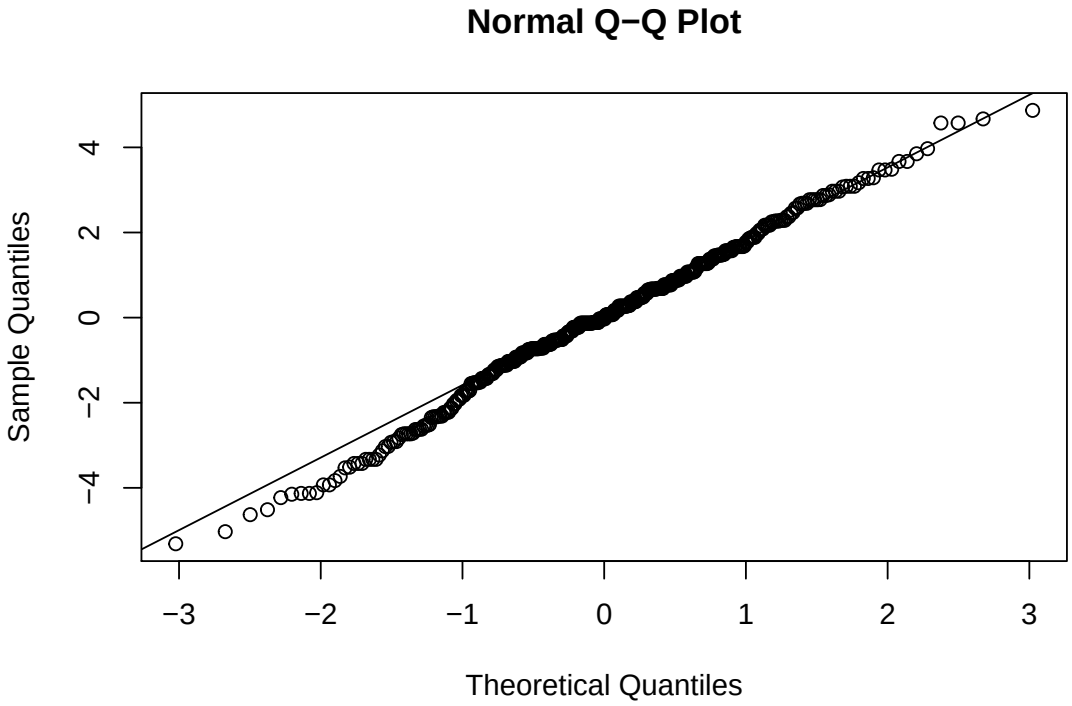
Para comprobar que los datos de los modelos anteriores cumplen la hipótesis de Normalidad de manera gráfica, recordamos que si representamos los pares $\left(e_{(r)}, \Phi^{-1}\left(\frac{r-0.5}{n}\right)\right)$ con $r = 1, \ldots, n$ para cada modelo, éstos deberían estar en torno a la recta del tipo $y = \sigma x$. Para realizar estos gráficos en R, es necesario realizar previamente el ANOVA correspondiente a través de la función: *aov*. Dado que el estudio de los resultados del ANOVA corresponde a capítulos posteriores, simplemente ejecutaremos el comando correspondiente sin obtener aún conclusión alguna.

Por lo que a continuación pasaremos a la representación de dichos pares, así como la realización del Test de Cramér-Von Mises, para cada uno de los casos:

Para la muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres.

```
model1factor<-aov(Muestra1factor$Nota~Muestra1factor$Estudios,
                  data=Muestra1factor)
```

```
par(mfrow=c(1,1))
qqnorm(model1factor$residuals)
qqline(model1factor$residuals)
```



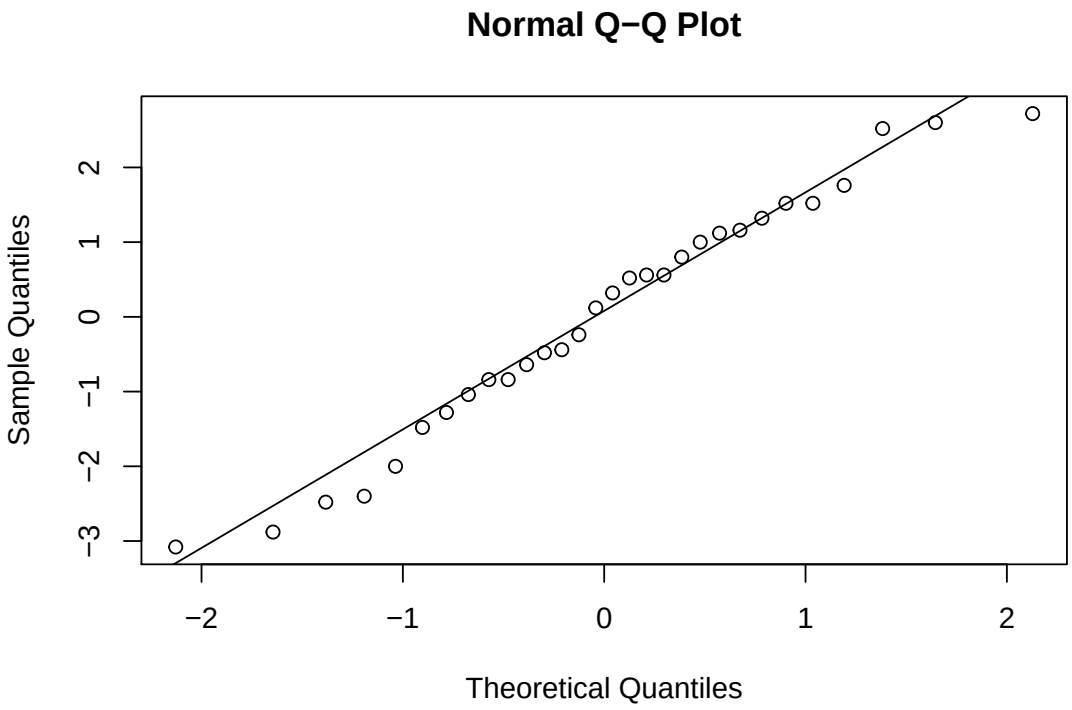

```
cvm.test(model1factor$residuals)
```

```
##
##  Cramer-von Mises normality test
##
## data:  model1factor$residuals
## W = 0.11697, p-value = 0.06567
```

Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
modelBIB<-aov(MuestraBIB$Nota~MuestraBIB$Estudios, data=MuestraBIB)
```

```
par(mfrow=c(1,1))
qqnorm(modelBIB$residuals)
qqline(modelBIB$residuals)
```



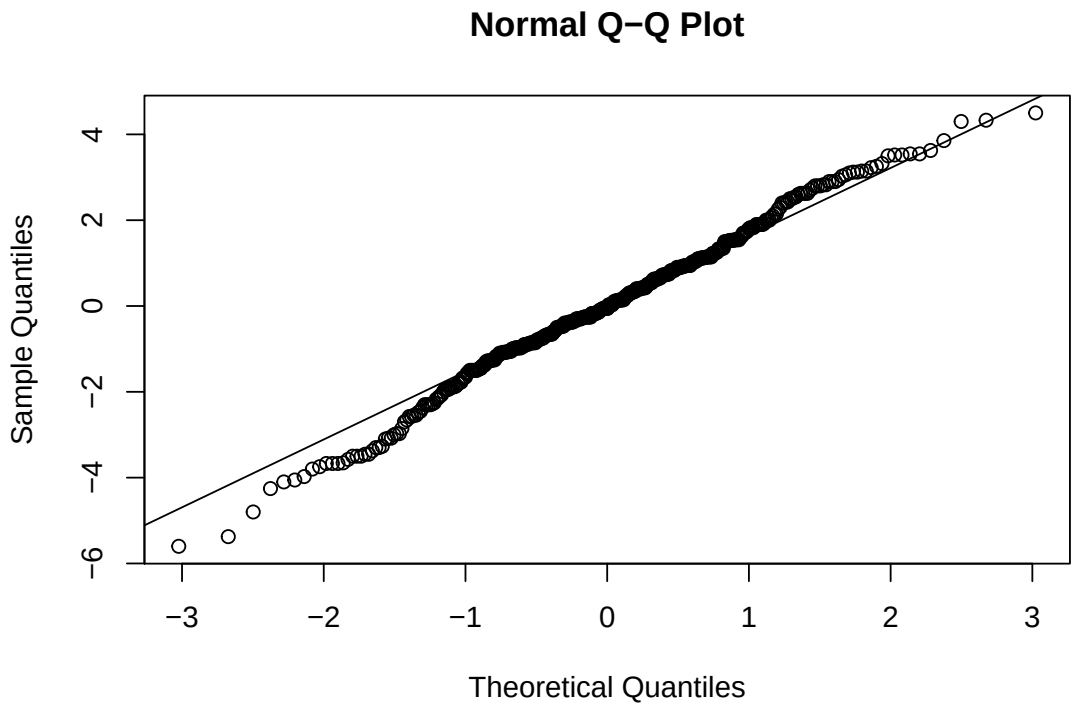

```
cvm.test(modelBIB$residuals)
```

```
##
##  Cramer-von Mises normality test
##
## data:  modelBIB$residuals
## W = 0.031724, p-value = 0.8152
```

Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Sexo del alumno.

```
model2factores<-aov(Muestra2factores$Nota~
                      Muestra2factores$Estudios*Muestra2factores$Sexo,
                    data=Muestra2factores)
```

```
par(mfrow=c(1,1))
qqnorm(model2factores$residuals)
qqline(model2factores$residuals)
```




```
cvm.test(model2factores$residuals)
```

```
##
##  Cramer-von Mises normality test
##
## data:  model2factores$residuals
## W = 0.10077, p-value = 0.1093
```

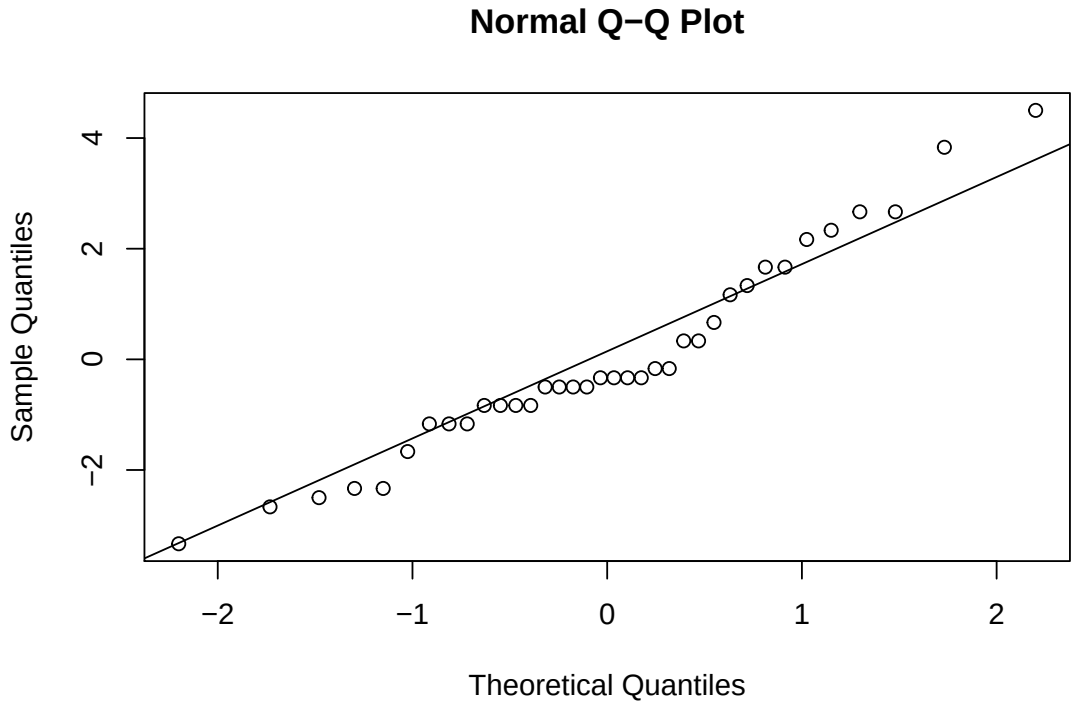
***Conclusión**: Aparentemente, vemos que para los 3 primeros casos los puntos que contienen los residuos ordenados de cada muestra se encuentran en torno a una recta del tipo $y = \sigma x$, luego tiene sentido aceptar como cierta la hipótesis de normalidad para cada caso.*

Además, una prueba más fiable de la hipótesis de normalidad es a través del Test de Cramér-Von Mises. Los p-valores correspondientes son respectivamente: 0.06567, 0.8152 y 0.1093, por lo que no tenemos evidencias necesarias para rechazar la normalidad de los errores en ninguna de las 3 primeras muestras.

Para la muestra balanceada: Nota en matemáticas del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
modelBAC<-aov(MuestraBAC$NotaMates~MuestraBAC$Estudios,
                    data=MuestraBAC)
```

```
par(mfrow=c(1,1))
qqnorm(modelBAC$residuals)
qqline(modelBAC$residuals)
```




```
cvm.test(modelBAC$residuals)
```

```
##
##  Cramer-von Mises normality test
##
## data:  modelBAC$residuals
## W = 0.14103, p-value = 0.02952
```

***Conclusión**: Aparentemente, vemos que los puntos que contienen los residuos ordenados de esta muestra no parecen estar en torno a una recta del tipo $y = \sigma x$; por lo que, en principio, parece que no deberíamos aceptar como cierta la hipótesis de normalidad para este caso.*

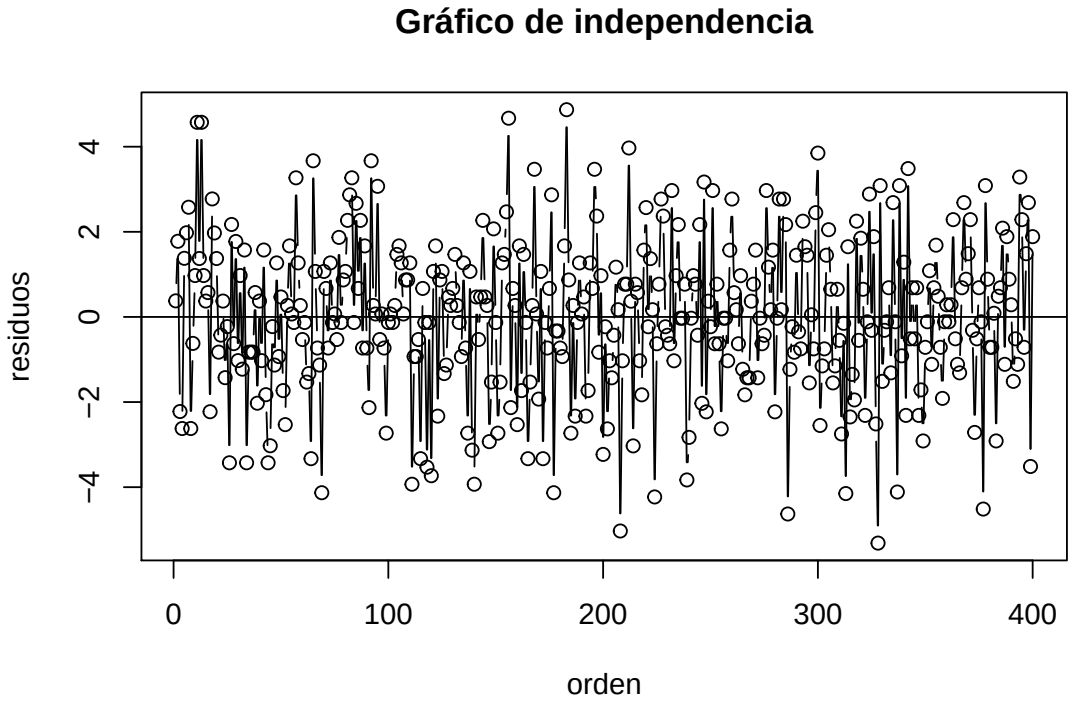
Una prueba más fiable de esto, es a través del Test de Cramér-Von Mises, cuyo p-valor para este caso ha sido de: 0.02952. Por lo tanto, para un nivel de significación $\alpha = 0.05$, tenemos que rechazar la hipótesis de normalidad de los errores en esta muestra.

2.4.2. Independencia

Para comprobar que los datos de un modelo cumplen la hipótesis de Independencia de manera gráfica, había que representar los residuos en función del tiempo. Si estos no mostraban patrón alguno, significaba que podemos asumir dicha hipótesis. Así pues, pasemos a la representación de éstos para cada uno de los casos:

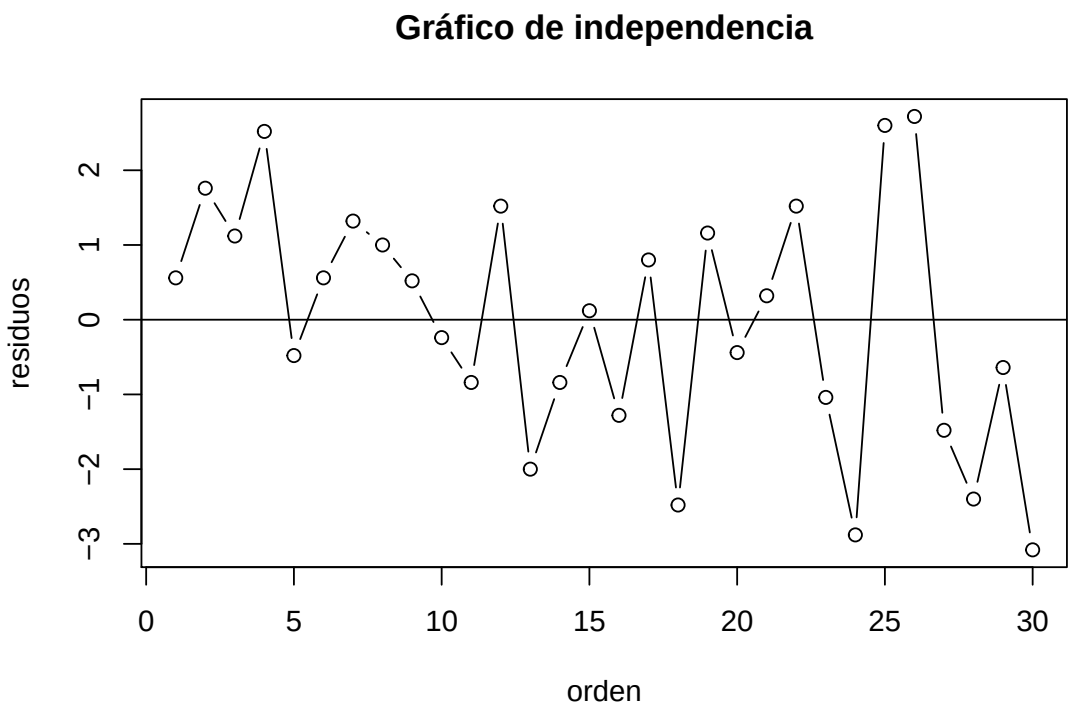
Para la muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres.

```
plot(c(1:400),model1factor$residuals, type="b",xlab="orden",
     ylab="residuos", main="Gráfico de independencia")
abline(h=0)
```



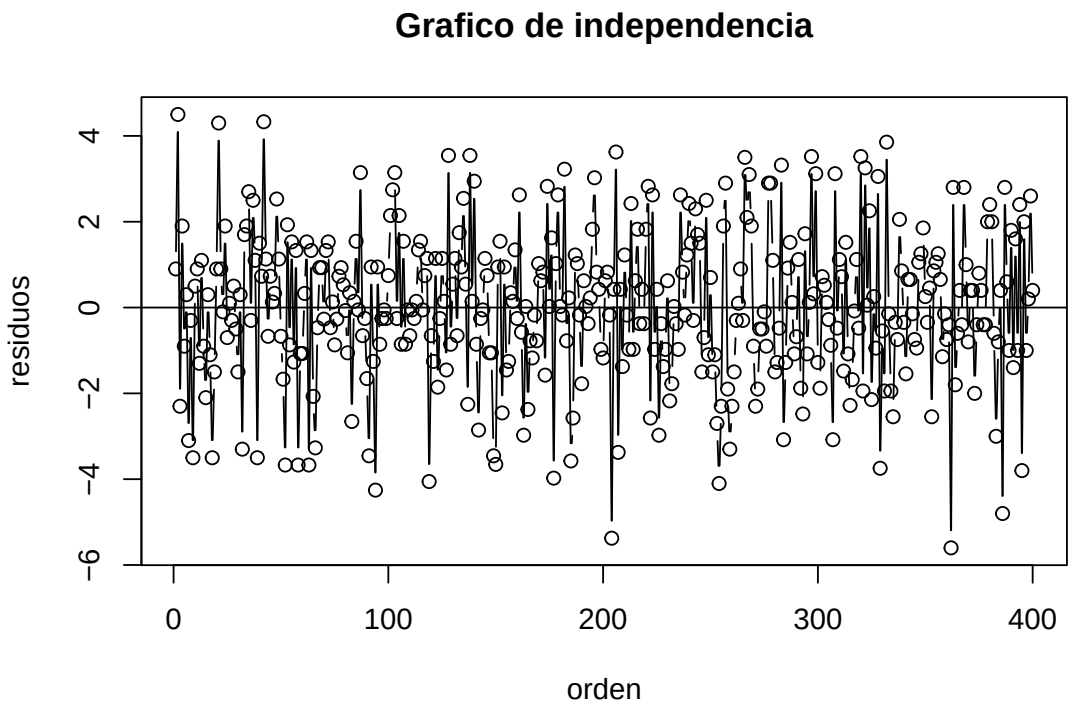
Para la muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
plot(c(1:30),modelBIB$residuals, type="b",xlab="orden",
     ylab="residuos", main="Gráfico de independencia")
abline(h=0)
```



Para la muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Sexo del alumno.

```
plot(c(1:400),model2factores$residuals, type="b",xlab="orden",
     ylab="residuos", main="Grafico de independencia") #ambos
abline(h=0)
```



Para la muestra balanceada: Nota en matemáticas del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
plot(c(1:36),modelBAC$residuals, type="b",xlab="orden",
     ylab="residuos", main="Gráfico de independencia")
abline(h=0)
```

***Conclusión**: Aparentemente, podemos observar que la representación de los puntos con los residuos respecto al tiempo, para cada una de las muestras, no parecen mostrar patrón alguno. Por lo tanto, asumimos la hipótesis de Independencia.*

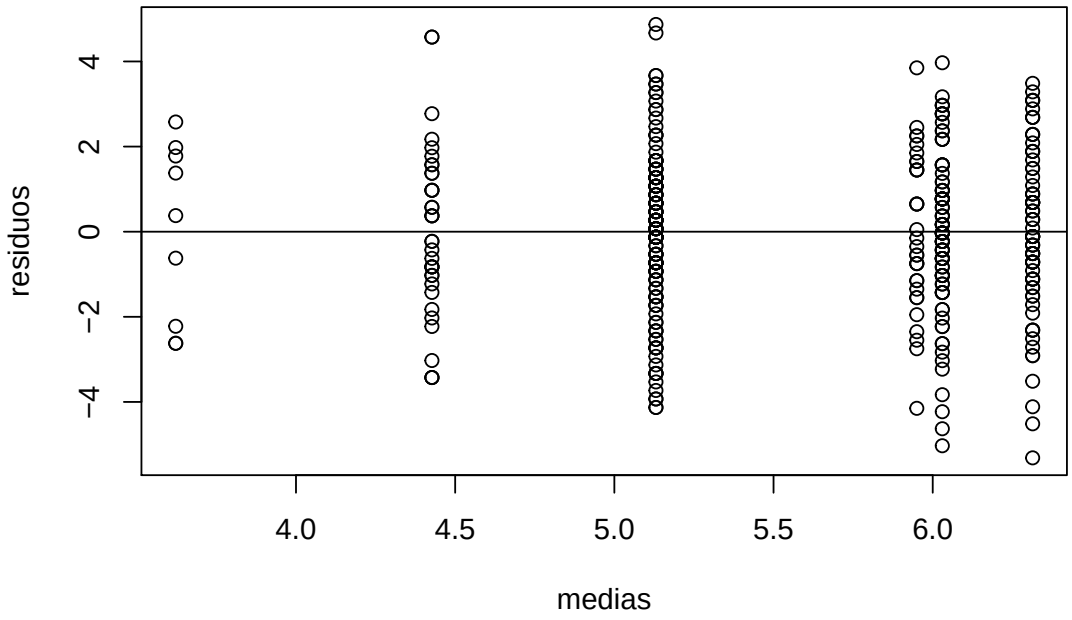
2.4.3. Homocedasticidad

Para comprobar que los datos de los modelos anteriores cumplen la hipótesis de Homocedasticidad de manera gráfica, recordamos que, si el aspecto de la nube de puntos al representar los residuos frente a las medias de cada grupo se distribuye de forma aleatoria, esto es señal de que se cumple la igualdad de varianzas.

Por lo que a continuación pasaremos a la representación de éstos, así como la realización del Test de Levene, para cada uno de los casos:

Para la muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres.

```
plot(model1factor$fitted.values, model1factor$residuals, xlab="medias",
     ylab="residuos")
abline(h=0)
```

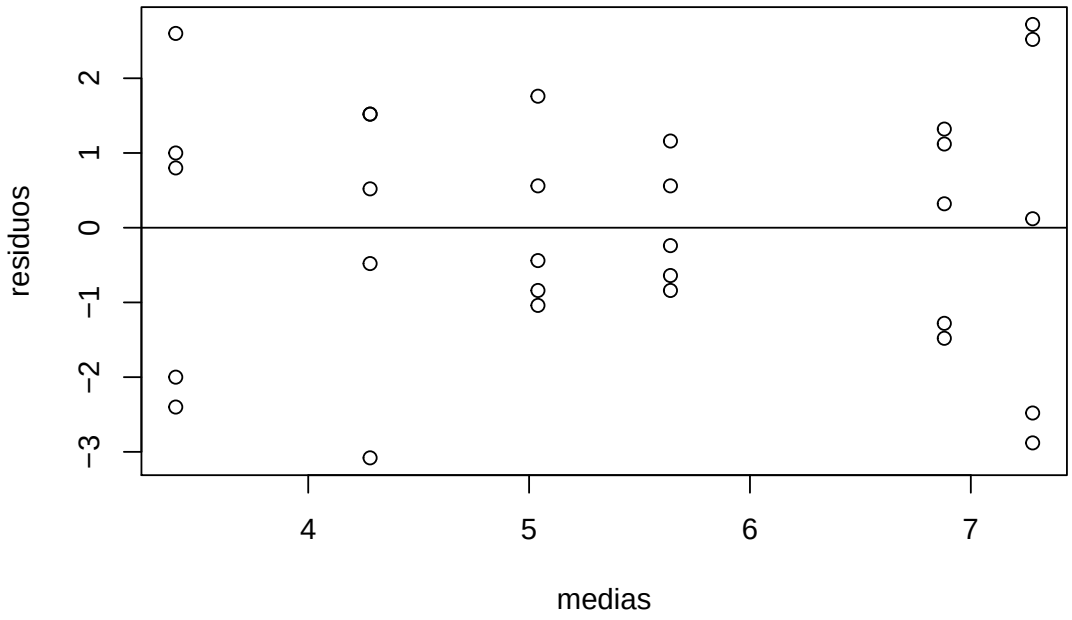



```
levene.test(Muestra1factor$NotaMedia,Muestra1factor$Estudios)
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  Muestra1factor$NotaMedia
## Test Statistic = 0.33447, p-value = 0.892
```

Para la muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
plot(modelBIB$fitted.values, modelBIB$residuals, xlab="medias",
     ylab="residuos")
abline(h=0)
```

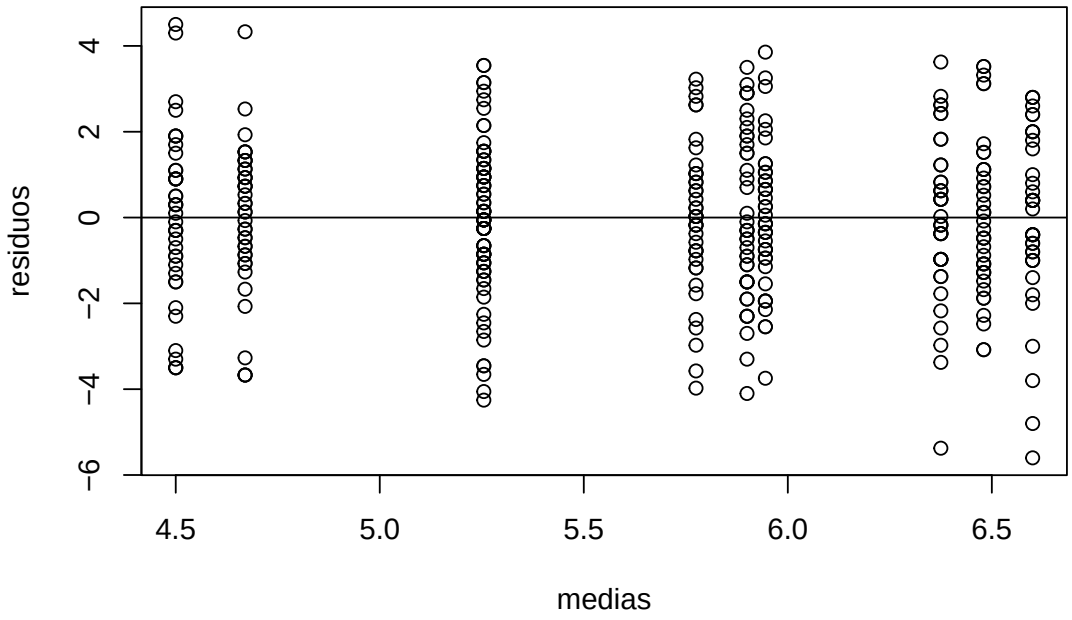



```
levene.test(MuestraBIB$NotaMedia,MuestraBIB$Estudios)
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  MuestraBIB$NotaMedia
## Test Statistic = 1.2787, p-value = 0.3055
```

Para la muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Sexo del alumno.

```
plot(model2factores$fitted.values, model2factores$residuals,
     xlab="medias", ylab="residuos")
abline(h=0)
```



Realizamos el test de Levene para los dos factores por separado:

```
levene.test(Muestra2factores$NotaMedia,Muestra2factores$Sexo)
```

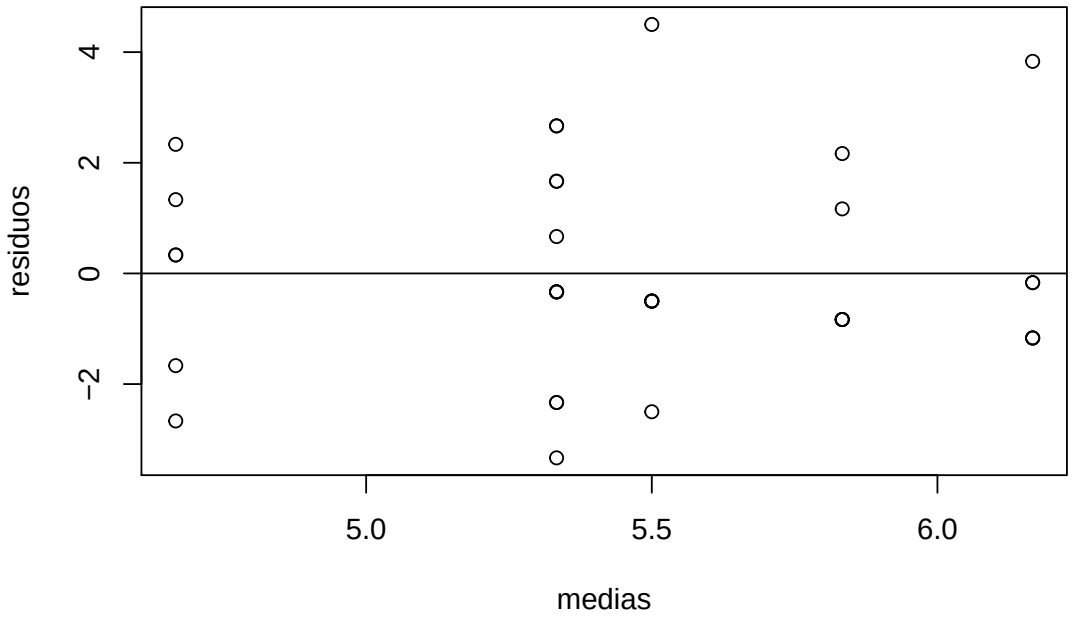
```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  Muestra2factores$NotaMedia
## Test Statistic = 0.13637, p-value = 0.7121
```

```
levene.test(Muestra2factores$NotaMedia,Muestra2factores$Estudios)
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  Muestra2factores$NotaMedia
## Test Statistic = 0.57046, p-value = 0.6842
```

Para la muestra balanceada: Nota en matemáticas del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
plot(modelBAC$fitted.values, modelBAC$residuals, xlab="medias",
     ylab="residuos")
abline(h=0)
```




```
levene.test(MuestraBAC$NotaMates,MuestraBAC$Estudios)
```

```
##
##  modified robust Brown-Forsythe Levene-type test based on the
##  absolute deviations from the median
##
## data:  MuestraBAC$NotaMates
## Test Statistic = 0.44472, p-value = 0.8136
```

***Conclusión**: Aparentemente, vemos que las nubes de puntos que contienen los residuos frente a las medias para cada una de las muestras, se distribuye de forma aleatoria, esto es señal de que se cumple la igualdad de varianzas. Además, al realizar el Test de Levene hemos obtenido los p-valores correspondientes: 0.892, 0.3055, 0.7121 y 0.6842 (para el tercer modelo) y por último 0.8136; por lo que no tenemos evidencias necesarias para rechazar la igualdad de varianzas en ninguna de las muestras.*

Capítulo 3

Diseño de experimentos.

En este capítulo, pasaremos a explicar los distintos tipos de diseños de experimentos más utilizados actualmente. Además, a modo de ejemplo, nos apoyaremos en las muestras obtenidos en el anterior capítulo para generar el análisis de algunos de los diseños. El proceso de selección de los tratamientos de los factores que veremos durante la memoria, será fijado al principio. Es decir, realizaremos en todo momento el diseño de experimentos para los modelos conocidos como: *modelo de efectos fijos.*

Hay que recalcar que el objetivo de esta memoria no es la obtención de los desarrollos teóricos ni la explicación de todos los modelos existentes que forman parte del inmenso mundo de los diseños de experimentos, sino más bien, crear una memoria dónde el lector pueda ayudarse a partir de ésta y poder así elaborar su propio diseño de experimentos. Dicho esto, durante todo el capítulo nos basaremos en los desarrollos teóricos realizados por los autores [3], [8] y [15].

3.1. Experimentos con un único factor: Experimento completamente aleatorizado

Vamos a suponer que seleccionamos un conjunto de datos con **un factor** compuesto por diferentes tratamientos o niveles, pongamos **k**, que son aplicados a **N** unidades experimentales (u.e.), aplicándose un tratamiento sobre cada u.e., y obteniéndose tras ello, una observación o respuesta **y**.

Supondremos que las u.e. son asignadas a los tratamientos de manera totalmente aleatoria. Un experimento así se dice que es un experimento completamente aleatorizado.

Supongamos que se dispone de k muestras independientes procedentes diferentes poblaciones, o de la misma población, pero cada una extraída de subpoblaciones afectadas por diferentes niveles de un factor.

Vamos a suponer además que se cumple la hipótesis de Normalidad: cada una de estas k poblaciones se distribuye según una ley normal $N(\mu_i, \sigma^2)$ es decir, la varianza es la misma en las k poblaciones, y pueden diferir en la media.

Bajo estas hipótesis, las observaciones pueden expresarse como:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad con \quad \begin{array}{l} i = 1, 2, ..., k. \\ j = 1, 2, ..., n_i. \\ \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2). \end{array} \tag{3.1}$$

donde:

$\boldsymbol{y_{ij}}$ representa la j-ésima observación en el i-ésimo tratamiento del factor.

$\boldsymbol{\mu}$ representa la media global de todas las observaciones.

$\boldsymbol{\alpha_i}$ es el efecto del i-ésimo tratamiento.

$\boldsymbol{\varepsilon_{ij}}$ es el error aleatorio de la i,j-ésima observación.

$\boldsymbol{n_i}$ es el número de observaciones en la muestra i.

Tanto las medias, μ_i, $i = 1, 2, \cdots, k$, con la varianza, σ^2, son constantes desconocidas.

Matricialmente, obtenemos el modelo:

$$Y = X\beta + \varepsilon \tag{3.2}$$

donde:

$$\begin{aligned} \boldsymbol{Y^t} &= (y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}, y_{k1}, y_{k2}, \ldots, y_{kn_k}) \\ \boldsymbol{\varepsilon^t} &= (\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{1n_1}, \varepsilon_{21}, \varepsilon_{22}, \ldots, \varepsilon_{2n_2}, \varepsilon_{k1}, \varepsilon_{k2}, \ldots, \varepsilon_{kn_k}) \\ \boldsymbol{\beta^t} &= (\mu, \alpha_1, \alpha_2, \ldots, \alpha_k) \\ X &= \begin{pmatrix} 1_{n_1} & 1_{n_1} & 0 & \ldots & 0 \\ 1_{n_2} & 0 & 1_{n_2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1_{n_k} & 0 & 0 & \ldots & 1_{n_k} \end{pmatrix} \end{aligned}$$

La matriz X, denominada **matriz de diseño**, es una matriz $N \times (k+1)$, donde $N = \sum_{i=1}^{k} n_i$, con rango $rg(X) = k$, pues la suma de las columnas correspondientes a los efectos de los tratamientos es igual a la primera columna, que corresponde a la media global.

Para diseñar un experimento completamente aleatorizado en el programa estadístico *R*, nos apoyaremos en la función *design.crd* del paquete *Agricolae* [10]. A modo de ejemplo, durante esta sección usaremos la primera muestra obtenida en el capítulo 2: *muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres*, para ver y analizar cómo realizar un diseño de este tipo.

Lo primero que veremos es como plantear la estructura de un posible diseño, de manera que:

```
str(design.crd)
```

```
## function (trt, r, serie = 2, seed = 0, kinds = "Super-Duper", randomization = TRUE
```

```
trt <- c("Sin educacion formal o inferior primaria", "Educación primaria",
"Enseñanza secundaria de 1era etapa",  "Educación secundaria de 2a etapa",
"Educación secundaria pero no terciaria", "Estudios universitarios")
repeticion <- c(9, 37, 150, 93, 32, 79)
outdesignCA <- design.crd(trt,r=repeticion,seed=777,serie=0)
DisComplAlea <- outdesignCA$book
```

Luego, las primeras observaciones del diseño de nuestro experimento completamente aleatorizado para este caso serían:

```
head(DisComplAlea)
```

```
##   plots r                                  trt
## 1     1 1     Enseñanza secundaria de 1era etapa
## 2     2 1       Educación secundaria de 2a etapa
## 3     3 2       Educación secundaria de 2a etapa
## 4     4 1 Educación secundaria pero no terciaria
## 5     5 1                 Estudios universitarios
## 6     6 2     Enseñanza secundaria de 1era etapa
```

Donde la primera columna nos indica el número de la observación, la segunda cuantas veces ha salido repetido anteriormente un tratamiento de un determinado factor, y en la tercera el tratamiento que debe seguir la observación.

3.1.0.1. Estimación de los parámetros del modelo.

La hipótesis de normalidad sobre los términos de error conlleva el hecho de que las variables y_{ij} sean normales e independientes, por lo que es inmediato construir la función de verosimilitud asociada a la muestra $Y^t = (y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}, y_{k1}, y_{k2}, \ldots, y_{kn_k})$:

$$\mathbb{L}(\mu, \alpha_i, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}[y_{ij} - \mu - \alpha_i]^2\right) \tag{3.3}$$

Los estimadores máximo-verosímiles para los parámetros μ, α_i y σ^2 son los valores para los cuales la función de verosimilitud alcanza su máximo. Para determinarlos habrá que obtener los puntos donde la derivada de la función (3.3) se anule. Entonces, si tomamos logaritmos (ya que conserva los puntos críticos por ser una función creciente) e igualando las derivadas parciales respecto de los parámetros del modelo, se obtiene un sistema de ecuaciones que proporciona los estimadores máximo verosímiles. Dichos estimadores vienen dados por las expresiones:

$$\hat{\mu} = \bar{y}_{..}, \qquad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..} \tag{3.4}$$

donde:

$$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}, \qquad \bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

Como la matriz de diseño tiene rango k, un estimador insesgado de la varianza viene dado por:

$$\hat{\sigma}^2 = \frac{SC_\varepsilon}{N-k}, \qquad con\ SC_\varepsilon = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

3.1.0.2. Análisis de la varianza: descomposición de la variabilidad total.

Como dijimos en el Capítulo 1, para comparar los efectos de los distintos niveles de un factor se emplea la técnica estadística denominada análisis de la varianza, abreviadamente ANOVA, que está basada en la descomposición de la variabilidad total de los datos en distintas componentes.

Así pues, a partir de la siguiente identidad:

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{ij} - \bar{y}_{i.}) \tag{3.5}$$

podemos obtener la siguiente descomposición:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k} n_i\left(y_{i.} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i.}\right)^2 \tag{3.6}$$

que representa la ecuación básica del análisis de la varianza, que simbólicamente podemos escribir:

$$SC_{tot} = SC_\alpha + SC_\varepsilon$$

donde hemos desglosado la variabilidad total de los datos en dos partes, la suma de cuadrados de las desviaciones de las medias de los tratamientos respecto de la media general (denominada suma de cuadrados entre tratamientos o variabilidad explicada) y la suma de cuadrados de las desviaciones de las observaciones de cada nivel respecto de su media (denominada suma de cuadrados dentro de los tratamientos, no-explicada o residual).

Los grados de libertad de estas formas cuadráticas son:

$$(N-1) = (k-1) + (N-k)$$

de donde, bajo la hipótesis de normalidad, se tiene que:

$$\frac{SC_\alpha}{\sigma^2} \sim \chi^2_{k-1,\alpha}, \qquad \frac{SC_\varepsilon}{\sigma^2} \sim \chi^2_{N-k,\alpha}$$

y además son independientes.

A partir de las sumas de cuadrados anteriores se pueden construir los denominados cuadrados medios, definidos como los cocientes entre dichas sumas de cuadrados y sus correspondientes grados de libertad:

$$CM_\alpha = \frac{SC_\alpha}{k-1} \qquad CM_\varepsilon = \frac{SC_\varepsilon}{N-k}$$

Además, podemos obtener los valores esperados de los cuadrados medios, que son:

$$E(CM_\alpha) = \sigma^2 + \frac{1}{k-1}\sum_{i=1}^{k} n_i \alpha_i^2, \qquad E(CM_\varepsilon) = \sigma^2$$

3.1.0.3. Contraste fundamental.

En el modelo considerado, es de interés comenzar con el contraste de igualdad de medias:

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ &H_1 : \mu_i \neq \mu_j, \quad \textit{para algunos } i \neq j \end{aligned}$$

o equivalentemente, el contraste de igualdad de efectos:

$$\begin{aligned} &H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k \\ &H_1 : \alpha_i \neq \alpha_j, \quad \textit{para algunos } i \neq j \end{aligned}$$

El estadístico del test de razón de verosimilitud del contraste anterior viene dado por:

$$F = \frac{SC_\alpha}{SC_\varepsilon}\frac{(N-k)\sigma^2}{(k-1)\sigma^2} = \frac{CM_\alpha}{CM_\varepsilon}$$

que bajo H_0 sigue una distribución F-Snedecor con $k-1$ y $N-k$ grados de libertad, luego los puntos críticos se obtienen cuando el valor de dicho estadístico sea mayor que el correspondiente valor teórico de la distribución F con $k-1$ y $N-k$ grados de libertad al nivel de significación α, es decir:

$$\textit{Rechazar } H_0 \textit{ si } F \geq \mathcal{F}_{k-1,N-k,1-\alpha}$$

Los resultados obtenidos se resumen en la denominada **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Factor	$SC_\alpha = \sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$	$k-1$	$CM_\alpha = \frac{SC_\alpha}{k-1}$	$F = CM_\alpha / CM_\varepsilon$
Error	$SC_\varepsilon = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2$	$N-k$	$CM_\varepsilon = \frac{SC_\alpha}{N-k}$	
Total	$SC_{tot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2$	$N-1$		

Cuadro 3.1: Tabla Anova de un experimento completamente aleatorizado.

A continuación, para ilustrar estos resultados, realizamos a través de *R* el análisis de la varianza a la *muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres*. El ANOVA se construye a través de la función *aov* y, además, realizaremos una regresión lineal a través de la función *lm* para obtener las estimaciones de los parámetros correspondientes:

```
model1factor<-aov(Muestra1factor$Nota~Muestra1factor$Estudios,
                  data=Muestra1factor)
model1factor2<-lm(Muestra1factor$Nota~Muestra1factor$Estudios,
                  data=Muestra1factor)
```

TABLA ANOVA:

```
summary(model1factor)
```

```
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## Muestra1factor$Estudios     5  179.1   35.82    10.4 2.21e-09 ***
## Residuals                 394 1357.6    3.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model1factor2)
```

```
##
## Call:
## lm(formula = Muestra1factor$Nota ~ Muestra1factor$Estudios, data = Muestra1factor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3139 -1.0301 -0.0301  1.2693  4.8693
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     3.6222     0.6188   5.854 1.01e-08 ***
## Muestra1factor$EstudiosEPr      0.8048     0.6899   1.167 0.244109
## Muestra1factor$EstudiosES1      1.5084     0.6370   2.368 0.018373 *
## Muestra1factor$EstudiosES2      2.4079     0.6480   3.716 0.000232 ***
## Muestra1factor$EstudiosEPS      2.3278     0.7004   3.324 0.000972 ***
## Muestra1factor$EstudiosEU       2.6917     0.6531   4.122 4.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.856 on 394 degrees of freedom
## Multiple R-squared:  0.1165, Adjusted R-squared:  0.1053
## F-statistic:  10.4 on 5 and 394 DF,  p-value: 2.211e-09
```

Conclusión: Rechazamos la hipótesis nula de la igualdad de los efectos, por lo tanto, tiene sentido realizar las correspondientes comparaciones múltiples.

3.1.0.4. Comparaciones múltiples.

Una vez aceptada la existencia de diferencias entre los efectos del factor, el interés reside en conocer cuáles son los tratamientos diferentes entre sí o qué tratamientos concretos producen mayor efecto. Es decir, nos interesa contrastar cualquier hipótesis de la forma:

$$\begin{aligned} &H_0 : \mu_i = \mu_j \\ &H_1 : \mu_i \neq \mu_j, \quad \textit{para algunos } i \neq j \end{aligned}$$

Podemos realizar estos contrastes mediante múltiples comparaciones dos a dos, o mediante comparaciones en las que intervienen combinaciones de varios niveles. Sin embargo, habría que realizar todas las comparaciones dos a dos, que son:

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

lo que conllevaría un incremento del error de tipo I global. Para solucionar esto, se han desarrollado una serie de técnicas denominadas procedimientos de "comparaciones múltiples", que se enmarcan dentro de los métodos de "Inferencia Estadística Simultánea". Las pruebas estadísticas para comparaciones múltiples más frecuentemente utilizadas se basan en la distribución t de Student. Veamos dos de los métodos más importantes:

Método de la mínima diferencia significativa

La técnica más antigua y popular para efectuar estas comparaciones múltiples es el procedimiento LSD, (*Least Significant Difference*). Este procedimiento fue sugerido por Fisher en 1935 y es el primer método de comparaciones múltiples que vamos a utilizar.

Dicho procedimiento consiste en una prueba de hipótesis por parejas basada en la distribución t-student. Para ello, se determina el siguiente estadístico:

$$t = \frac{\bar{y}_{i.} - \bar{y}_{j.}}{\sqrt{CM_\varepsilon \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

por las hipótesis del modelo, sigue una distribución t de Student, con $N - k$ grados de libertad.

Por lo tanto, se concluye que la pareja de medias μ_i y μ_j son estadísticamente diferentes y por tanto la hipótesis H_0 es rechazada si:

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq t_{N-k,1-\frac{\alpha}{2}} \sqrt{CM_\varepsilon \left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \equiv LSD$$

donde la expresión de la derecha de la desigualdad, es la denominada *mínima diferencia significativa* (LSD), por ser la cantidad más pequeña a partir de la cual, diferencias entre medias son consideradas significativas. Además, si el diseño está balanceado, es decir, $n_1 = n_2 = \cdots = n_k = n$, entonces:

$$LSD \equiv t_{N-k,1-\frac{\alpha}{2}}\sqrt{\frac{2CM_\varepsilon}{n}}$$

El procedimiento LSD es sencillo de utilizar y se puede aplicar tanto en modelos balanceados como no balanceados. Además, proporciona también intervalos de confianza para diferencias de medias. Dichos intervalos son de la forma:

$$((\bar{y}_{i.} - \bar{y}_{j.}) - LSD, (\bar{y}_{i.} - \bar{y}_{j.}) + LSD)$$

Sin embargo, este método presenta el inconveniente de que el nivel de significación α puede incrementarse de forma considerable a medida que aumenta el número de grupos.

Volviendo al ejemplo con nuestra muestra, al haber sido rechazada la hipótesis nula de la igualdad de los efectos, podemos realizar las comparaciones múltiples a través del método de la mínima diferencia significativa, para así conocer qué tratamientos del factor: *Nivel de estudios máximo de los padres*, influyen de mejor manera en el rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz. Luego aplicando este método, obtenemos que:

```
LSD1factor <-LSD.test(model1factor, "Muestra1factor$Estudios",
                      group=FALSE)
print(LSD1factor$comparison)
```

```
##            difference pvalue signif.        LCL         UCL
## EPr - EPS -1.52297297 0.0007     *** -2.4039682 -0.64197771
## EPr - ES1 -0.70363964 0.0396       * -1.3735230 -0.03375628
## EPr - ES2 -1.60308050 0.0000     *** -2.3124201 -0.89374088
## EPr - EU  -1.88689702 0.0000     *** -2.6139055 -1.15988856
## EPr - SE   0.80480480 0.2441         -0.5515761  2.16118574
## EPS - ES1  0.81933333 0.0239       *  0.1087092  1.52995747
## EPS - ES2 -0.08010753 0.8333         -0.8280418  0.66782672
## EPS - EU  -0.36392405 0.3500         -1.1286359  0.40078782
## EPS - SE   2.32777778 0.0010     ***  0.9508202  3.70473536
## ES1 - ES2 -0.89944086 0.0003     *** -1.3811012 -0.41778049
## ES1 - EU  -1.18325738 0.0000     *** -1.6905791 -0.67593568
## ES1 - SE   1.50844444 0.0184       *  0.2560055  2.76088337
## ES2 - EU  -0.28381652 0.3183         -0.8422020  0.27456891
## ES2 - SE   2.40788530 0.0002     ***  1.1339064  3.68186421
## EU - SE    2.69170183 0.0000     ***  1.4078012  3.97560247
```

```
options(digits=2)

LSD1factor <-LSD.test(model1factor, "Muestra1factor$Estudios",
                      group=TRUE)
print(LSD1factor$group)
```

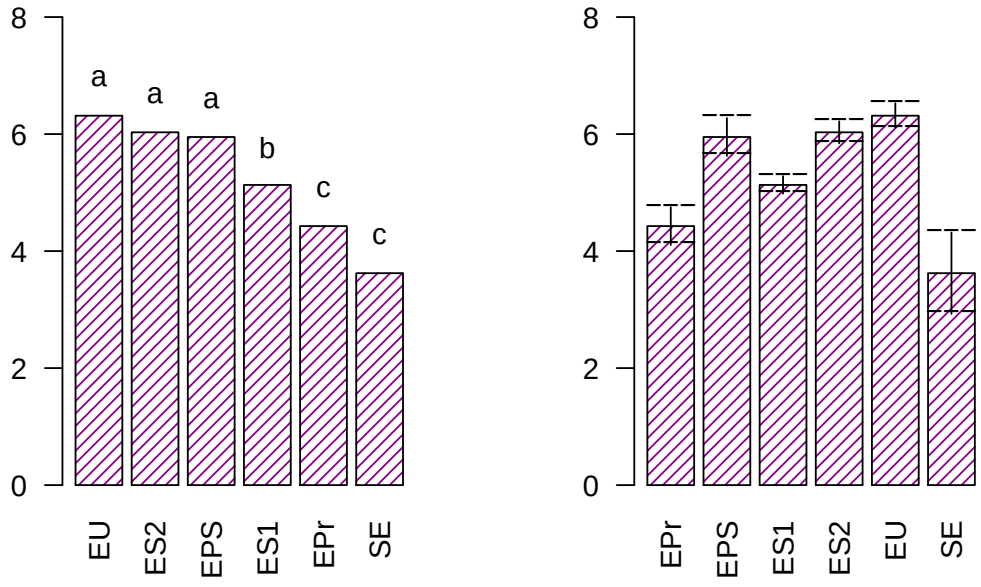
```
##     Muestra1factor$Nota groups
## EU                  6.3      a
## ES2                 6.0      a
## EPS                 6.0      a
## ES1                 5.1      b
## EPr                 4.4      c
## SE                  3.6      c
```

Conclusión: Podemos afirmar que a mayor nivel de estudios máximo de los padres existe mayor rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz.

Al haber diferencias entre los niveles del factor, podemos representar qué tratamientos del factor podrían agruparse, puesto que entre ellos no existen diferencias significativas en el rendimiento escolar del alumno. Realizaremos, además, una gráfica que nos indique visualmente el porqué de esa agrupación.

Dicho esto, pasemos a representar lo mencionado anteriormente:

```
par(mfrow=c(1,2))
bar.group (LSD1factor$group, col=colors()[84],ylim = c(0,8),
           frequency=1,las=2,density=20)
bar.err(LSD1factor$means, col=colors()[84], ylim = c(0, 8),frequency=1,
        las=2,density=20)
```



Método de Bonferroni

En este procedimiento se fija un nivel de significación α que se reparte entre cada una de las comparaciones consideradas y se utiliza la desigualdad de Bonferroni:

$$P\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} P\left(A_m\right)$$

Consideremos que queremos realizar estimación por intervalos para las m $= m = \frac{k(k-1)}{2}$ comparaciones posibles. Entonces, si definimos $\bar{A}_i$ el suceso de rechazar la igualdad de dos medias μ_i y μ_j, cuando realmente sí son iguales. Supongamos que las comparaciones entre media se hacen con:

$$P\left(\bar{A}_i\right)$$

Por otro lado, llamemos B al suceso de rechazar uno o más contraste de igualdad de medias, cuando todas son iguales. Entonces, B serás de la forma:

$$B = \bar{A}_1 \cup \bar{A}_2 \cup \ldots \bar{A}_m$$

que, al no ser los sucesos $\bar{A}_i$ mutuamente excluyente, tenemos:

$$P\left(B\right) = P\left(\bar{A}_1 \cup \bar{A}_2 \cup \ldots \bar{A}_m\right) \leq \sum_{i=1}^{m} P\left(A_i\right) = m\alpha$$

Por lo tanto, si consideramos cada contraste individual a un nivel $\alpha = \alpha_T/m$, garantizamos un error de tipo I total para el conjunto de contrastes α_T, ya que la probabilidad del suceso B sería cómo máximo α_T.

Luego, se concluye que la pareja de medias μ_i y μ_j son estadísticamente diferentes y por tanto la hipótesis H_0 es rechazada si:

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq t_{N-k,1-\frac{\alpha}{2m}} \sqrt{CM_\varepsilon \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Este método presenta el inconveniente de que cuando m es muy grande, pueden hacer falta niveles de significación α tan pequeños, que no detecten diferencias que de otra forma sí serían significativas.

De manera análoga al método de la mínima diferencia significativa, pasemos a realizar las comparaciones múltiples a través del Método de Bonferroni:

```
Bonf1factor2<-LSD.test(model1factor,"Muestra1factor$Estudios",group=FALSE,
               p.adj= "bonferroni")
print(Bonf1factor2$comparison)
```

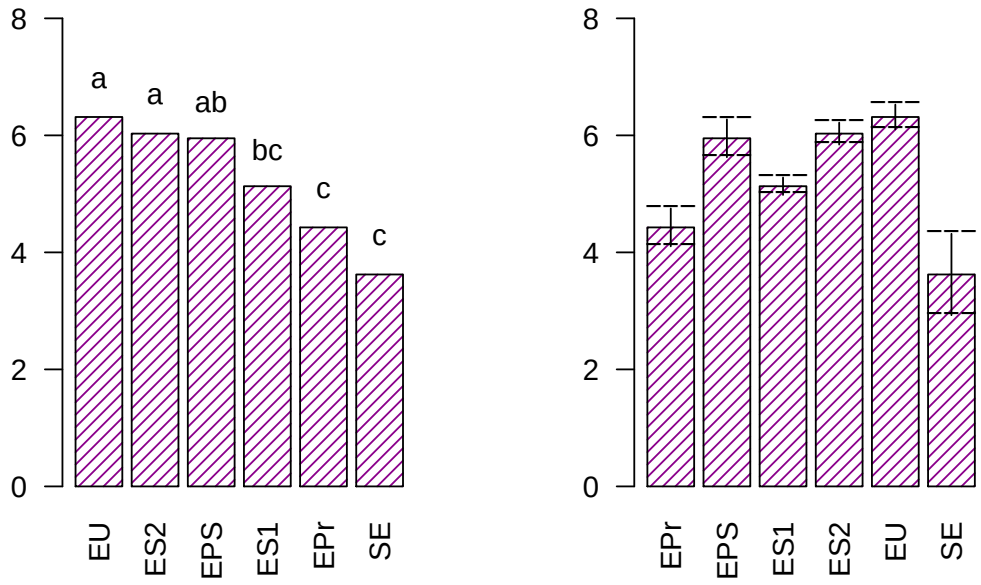
```
##           difference pvalue signif.   LCL   UCL
## EPr - EPS      -1.52 0.0112       * -2.85 -0.20
## EPr - ES1      -0.70 0.5935         -1.71  0.30
## EPr - ES2      -1.60 0.0002     *** -2.67 -0.54
## EPr - EU       -1.89 0.0000     *** -2.98 -0.79
## EPr - SE        0.80 1.0000         -1.23  2.84
## EPS - ES1       0.82 0.3592         -0.25  1.89
## EPS - ES2      -0.08 1.0000         -1.20  1.04
## EPS - EU       -0.36 1.0000         -1.51  0.78
## EPS - SE        2.33 0.0146       *  0.26  4.40
## ES1 - ES2      -0.90 0.0041      ** -1.62 -0.18
## ES1 - EU       -1.18 0.0001     *** -1.95 -0.42
## ES1 - SE        1.51 0.2756         -0.37  3.39
## ES2 - EU       -0.28 1.0000         -1.12  0.55
## ES2 - SE        2.41 0.0035      **  0.49  4.32
## EU - SE         2.69 0.0007     ***  0.76  4.62
```

```
Bonf1factor1<-LSD.test(model1factor, "Muestra1factor$Estudios", group=TRUE,
               p.adj= "bonferroni")
print(Bonf1factor1$group)
```

```
##     Muestra1factor$Nota groups
## EU                  6.3      a
## ES2                 6.0      a
## EPS                 6.0     ab
## ES1                 5.1     bc
## EPr                 4.4      c
## SE                  3.6      c
```

Pasemos a la representación de las agrupaciones de los tratamientos del factor que ha proporcionado este método:

```
par(mfrow=c(1,2))
bar.group (Bonf1factor1$group, col=colors()[84],ylim = c(0, 8),
           frequency=1,las=2,density=20)
bar.err(Bonf1factor1$means, col=colors()[84], ylim = c(0, 8),
        frequency=1,las=2,density=20)
```



Conclusión: Aunque existen diferencias entre la agrupación de ciertos tratamientos del factor, la conclusión es la misma. Así pues, podemos afirmar que a mayor nivel de estudios máximo de los padres existe mayor rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz.

Estos dos métodos para realizar comparaciones múltiples no son ni mucho menos los únicos. Existen diversos métodos que, aunque la mayoría sean modificaciones leves o variaciones de éstos, pueden llegar a dar resultados distintos. Sin embargo, hay que recalcar que la mayoría darán resultados análogos. Debido a la limitación de esta memoria, dejaremos a opción del lector la investigación de estos otros métodos. Algunos de ellos son:

```
p.adjust.methods
```

```
## [1] "holm"       "hochberg"   "hommel"     "bonferroni" "BH"
## [6] "BY"         "fdr"        "none"
```

donde:

– **"holm":** Es el método de Holm (1979).

– **"hochberg":** Es el método de Hochberg (1988).

– **"hommel":** Es el método de Hommel (1988).

– **"BH" o "fdr":** Es el método de Benjamini & Hochberg (1995) o su alias que proviene de las siglas en ingles de *False Discovery Rate.*

– **"BY":** Es el método de Benjamini & Yekutieli (2001).

3.2. Diseño en bloques

Hasta ahora hemos supuesto que la variabilidad de las observaciones procedía, principalmente, del efecto del tratamiento y del error aleatorio (o ruido) de la propia observación, es decir, que existía bastante homogeneidad entre las unidades experimentales.

Sin embargo, puede suceder que dichas u.e. sean distintas y contribuyan a la variabilidad observada.

Si en esta situación utilizamos un diseño completamente aleatorizado, no sabremos si las diferencias entre los resultados de dos u.e. sometidas a distintos tratamientos se deben a una diferencia real entre los efectos de los tratamientos o a la heterogeneidad de dichas unidades.

Por lo tanto, para evitar esta situación y conseguir que el error experimental sea lo más pequeño posible podemos considerar 2 opciones:

- Hacer el estudio solo con unidades experimentales que consideremos muy homogéneas.
- Formar bloques de u.e. de manera que la u.e. de cada bloque sea lo más homogéneo posible; y los bloques sean entre sí heterogéneos.

Por tanto, podemos definir **bloque** como aquella unidad experimental que es homogéneas con respecto a cierto factor o fuente de variación.

3.2.1. Diseño en bloque aleatorizados completos.

Supongamos ahora un diseño donde se tiene **I** tratamientos y **J** bloques. Se realiza una observación por tratamiento en cada bloque, y el orden en el que los tratamientos son asignados o ensayados en las unidades experimentales que componen cada bloque se determina aleatoriamente. De una forma más esquemática, tenemos que:

Bloque 1		Bloque j		Bloque J
y_{11}	$\cdots$	y_{1j}	$\cdots$	y_{1J}
$\vdots$		$\vdots$		$\vdots$
y_{i1}		y_{ij}		y_{iJ}
$\vdots$		$\vdots$		$\vdots$
y_{I1}	$\cdots$	y_{Ij}	$\cdots$	y_{IJ}

El modelo estadístico para este diseño es:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad con \quad \begin{array}{l} i = 1, 2, ..., I. \\ j = 1, 2, ..., J. \\ \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2). \end{array} \tag{3.7}$$

donde:

$\boldsymbol{y_{ij}}$ representa la observación del i-ésimo tratamiento en el j-ésimo bloque.

$\boldsymbol{\mu}$ representa la media global de todas las observaciones.

$\boldsymbol{\alpha_i}$ es el efecto del i-ésimo tratamiento.

$\boldsymbol{\beta_j}$ es el efecto del j-ésimo bloque.

$\boldsymbol{\varepsilon_{ij}}$ es el error aleatorio de la i,j-ésima observación.

Para diseñar un experimento en bloque aleatorizados completos a través de *R*, utilizaremos la función *design.rcbd*. Un ejemplo sencillo para este caso, sería obtener un diseño a través de la muestra obtenida en el capítulo 2: *Muestra balanceada: Nota en matemáticas del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno*. Para obtener un posible esquema del diseño, haremos lo siguiente:

```
str(design.rcbd)
```

```
## function (trt, r, serie = 2, seed = 0, kinds = "Super-Duper", first = TRUE,
##     continue = FALSE, randomization = TRUE)
```

```
trt <-c("SE","EPr","ES1","ES2","EPS","EU")
repeticion <- 6
outdesignBAC <- design.rcbd(trt,r=repeticion, seed=-513, serie=2)
print(outdesignBAC$sketch)
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
## [1,] "ES2" "EPr" "EU"  "EPS" "ES1" "SE"
## [2,] "SE"  "EU"  "ES1" "ES2" "EPr" "EPS"
## [3,] "ES2" "SE"  "EPS" "EU"  "ES1" "EPr"
## [4,] "EPS" "EU"  "ES2" "ES1" "SE"  "EPr"
## [5,] "SE"  "EPr" "ES2" "EPS" "ES1" "EU"
## [6,] "EPr" "ES1" "ES2" "SE"  "EPS" "EU"
```

NOTA: La realización del posterior análisis se desarrollará, debido a la ausencia de normalidad de la muestra, en el apartado "Comparaciones no paramétricas" que veremos más adelante.

3.2.1.1. Estimación de parámetros

Al igual que en modelo completamente aleatorizado se construye la función de verosimilitud asociada a la muestra $Y^t = (y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}, y_{k1}, y_{k2}, \ldots, y_{kn_k})$:

$$\mathbb{L}(\mu, \alpha_i, \beta_j, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J}[y_{ij} - \mu - \alpha_i - \beta_j]^2\right) \tag{3.8}$$

tomando logaritmos e igualando las derivadas parciales respecto de los parámetros del modelo, se obtiene un sistema de ecuaciones que proporciona los estimadores máximo verosímiles.

Dichos estimadores vienen dados por las expresiones:

$$\hat{\mu} = \bar{y}_{..}, \qquad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}, \qquad \hat{\beta}_i = \bar{y}_{j.} - \bar{y}_{..} \tag{3.9}$$

donde:

$$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}, \qquad \bar{y}_{i.} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}, \qquad \bar{y}_{.j} = \frac{1}{I}\sum_{i=1}^{I} y_{ij}$$

Como la matriz de diseño tiene rango $I + J - 1$, un estimador insesgado de la varianza viene dado por:

$$\hat{\sigma}^2 = \frac{SC_\varepsilon}{IJ - J - I + 1}, \qquad con \qquad SC_\varepsilon = \sum_{i=1}^{I}\sum_{j=1}^{J} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

3.2.1.2. Análisis de la varianza: descomposición de la variabilidad total.

En los diseños en bloques aleatorizados también se emplea la técnica estadística análisis de la varianza para comparar globalmente los efectos de los distintos niveles de un factor. Ésta se basaba en la descomposición de la variabilidad total de los datos en distintas componentes. Para ello consideramos la siguiente identidad:

$$y_{ij} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) \tag{3.10}$$

podemos obtener la siguiente descomposición:

$$\sum_{i=1}^{I}\sum_{j=1}^{J} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{I} J\,(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{j=1}^{J} I\,(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

que simbólicamente podemos escribir:

$$SC_{tot} = SC_\alpha + SC_\beta + SC_\varepsilon$$

donde hemos desglosado la variabilidad total de los datos en tres componentes:

– **La suma de cuadrados entre tratamientos:** que es la suma de cuadrados de las diferencias entre las medias de los tratamientos y la media general, que expresa la variabilidad explicada por los tratamientos.

– **La suma de cuadrados entre bloques:** que es la suma de cuadrados de las diferencias entre las medias de los bloques y la media general, que expresa la variabilidad explicada por los bloques.

– **La suma de cuadrados del error:** que es la suma de cuadrados de los residuos, que expresa la variabilidad no explicada.

Los grados de libertad de estas formas cuadráticas son:

$$(IJ - 1) = (I - 1) + (J - 1) + (I - 1)(J - 1)$$

por lo que, bajo la hipótesis de normalidad, se tiene que:

$$\frac{SC_\alpha}{\sigma^2} \sim \chi^2_{I-1,\lambda_\alpha}, \qquad \frac{SC_\beta}{\sigma^2} \sim \chi^2_{J-1,\lambda_\beta}, \qquad \frac{SC_\varepsilon}{\sigma^2} \sim \chi^2_{(I-1)(J-1),\lambda_\varepsilon}$$

con:

$$\lambda_\alpha = \frac{J}{\sigma^2}\sum_i^I \alpha_i^2, \qquad \lambda_\beta = \frac{I}{\sigma^2}\sum_j^J \beta j^2$$

y además son independientes.

A partir de las sumas de cuadrados anteriores y sus grados de libertad, obtenemos los cuadrados medios:

$$CM_\alpha = \frac{SC_\alpha}{I-1}, \qquad CM_\beta = \frac{SC_\beta}{J-1}, \qquad CM_\varepsilon = \frac{SC_\varepsilon}{(I-1)(J-1)}$$

Los valores esperados de los cuadrados medios son:

$$E(CM_\alpha) = \sigma^2 + \frac{J}{I-1}\sum_{i=1}^I \alpha_i^2, \qquad E(CM_\beta) = \sigma^2 + \frac{I}{J-1}\sum_{j=1}^J \beta j^2, \qquad E(CM_\varepsilon) = \sigma^2$$

3.2.1.3. Contraste fundamental.

El contraste estadístico de más interés en este modelo, como mencionamos anteriormente, es el que tiene como hipótesis nula la igualdad de efectos:

$$\begin{aligned} &H_{0\alpha} : \alpha_1 = \alpha_2 = \cdots = \alpha_I \\ &H_{1\alpha} : \alpha_i \neq \alpha_j, \quad \textit{para algunos } i \neq j \end{aligned}$$

También es interesante contrastar la igualdad de medias de los bloques:

$$\begin{aligned} &H_{0\beta} : \beta_1 = \beta_2 = \cdots = \beta_J \\ &H_{1\beta} : \beta_i \neq \beta_j, \quad \textit{para algunos } i \neq j \end{aligned}$$

Por consiguiente, bajo las hipótesis de igualdad de efectos de los tratamientos y los bloques, se verifica:

$$F_\alpha = \frac{SC_\alpha}{SC_\varepsilon}\frac{(I-1)(J-1)\sigma^2}{(I-1)\sigma^2} = \frac{CM_\alpha}{CM_\varepsilon} \sim F_{I-1,(I-1)(J-1)}$$

y

$$F_\beta = \frac{SC_\beta}{SC_\varepsilon}\frac{(I-1)(J-1)\sigma^2}{(J-1)\sigma^2} = \frac{CM_\beta}{CM_\varepsilon} \sim F_{J-1,(I-1)(J-1)}$$

luego la regiones críticas de ambos test son, respectivamente:

$$Rechazar\ H_{0\alpha} \quad si \quad F_\alpha \geq \mathcal{F}_{I-1,(I-1)(J-1),1-\alpha}$$

y

$$Rechazar\ H_{0\beta} \quad si \quad F_\beta \geq \mathcal{F}_{J-1,(I-1)(J-1),1-\alpha}$$

Los resultados obtenidos se resumen en la siguiente **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Factor	SC_α	$I-1$	CM_α	$F_\alpha = CM_\alpha/CM_\varepsilon$
Bloques	SC_β	$J-1$	CM_β	$F_\beta = CM_\beta/CM_\varepsilon$
Error	SC_ε	$(I-1)(J-1)$	CM_ε	
Total	SC_{tot}	$IJ-1$		

Cuadro 3.2: Tabla Anova para el modelo de bloques aleatorizado.

NOTA: Si del análisis se concluye que existe diferencia significativa entre los efectos de los tratamientos, será de interés realizar comparaciones entre medias de tratamientos. Para ello, pueden aplicarse los métodos ya estudiados en el experimento completamente aleatorizado. Además, en el caso de que no haya diferencia significativa entre los efectos de los bloques, para ocasiones posteriores, resultará de interés plantearse un nuevo diseño sin la inclusión de dichos bloques en el estudio.

3.2.2. Diseño en cuadrado latino.

En el diseño por bloques completos aleatorizados se supone que las unidades experimentales presentaban una fuente de variación (ajena al factor principal).

Supongamos ahora que las unidades experimentales presentan dos fuentes de variación (ambas ajenas al factor principal). Si el factor principal tiene K tratamientos, y cada factor secundario tiene también K tratamientos, entonces un experimento donde cada tratamiento sea ensayado en todas las combinaciones de niveles de los factores secundarios, requeriría K^3 pruebas, que puede resultar un número muy elevado incluso para un K moderado.

Un modo de reducir el número de ensayos es considerar un esquema experimental donde cada tratamiento sea ensayado una sola vez en cada nivel de los factores secundarios, para lo que puede utilizarse un diseño en cuadrado latino.

Un cuadrado latino de lado se define como una matriz pxp cuyos elementos son letras (latinas), cada una repetida K veces, de modo que cada letra aparece exactamente una vez en cada fila y en cada columna. Un ejemplo de cuadrado latino con $K = 4$ sería:

$$\begin{pmatrix} A & B & C & D \\ B & C & D & A \\ C & D & A & B \\ D & A & B & C \end{pmatrix}$$

Para realizar un diseño en cuadrado latino se selecciona un cuadrado latino, y se asigna de manera aleatoria cada fila con los niveles de uno de los factores secundarios, cada columna con los niveles del otro factor secundario, y cada letra con los tratamientos del factor principal.

El modelo estadístico para este diseño es:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk} \qquad con \quad \begin{array}{l} i, j = 1, 2, \ldots, K \\ k = 1, 2, \ldots, K \\ \varepsilon_{ijk} \text{ i.i.d. } N(0, \sigma^2). \end{array} \tag{3.11}$$

donde:

$\boldsymbol{y_{ijk}}$ Es la observación correspondiente a la i-ésima fila, la j-ésima columna y la k-ésima letra.

$\boldsymbol{\mu}$ representa la media global de todas las observaciones.

$\boldsymbol{\alpha_i}$ es el efecto del i-ésima fila.

$\boldsymbol{\beta_j}$ es el efecto del j-ésima columna.

$\boldsymbol{\gamma_k}$ es el efecto del k-ésima letra latina.

$\boldsymbol{\varepsilon_{ijk}}$ es el error aleatorio de la i,j-ésima observación.

Nótese que se han utilizado tres índices para describir el modelo, aunque realmente sólo son necesarios dos, ya que una vez fijadas la fila y la columna, el tratamiento está determinado.

Para diseñar un experimento en cuadrado latino a través de *R*, utilizaremos la función *design.lsd*. Análogamente al apartado anterior, podríamos plantear un diseño similar al proporcionado por la *muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno*, aunque sin realizar el análisis de éste. A modo de ejemplo y solo como ilustración, veremos un posible diseño para 64 observaciones:

```
str(design.lsd)
```

```
## function (trt, serie = 2, seed = 0, kinds = "Super-Duper", first = TRUE,
##     randomization = TRUE)
```

```
trt <-c("SE","EPr","ES1","ES2","EPS","EU")
outdesignDCL <- design.lsd(trt, seed=543, serie=2)
print(outdesignDCL$sketch)
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
## [1,] "EPS" "ES1" "ES2" "SE"  "EPr" "EU"
## [2,] "SE"  "EPS" "EU"  "ES1" "ES2" "EPr"
## [3,] "EU"  "ES2" "EPS" "EPr" "ES1" "SE"
## [4,] "EPr" "EU"  "SE"  "ES2" "EPS" "ES1"
## [5,] "ES2" "EPr" "ES1" "EU"  "SE"  "EPS"
## [6,] "ES1" "SE"  "EPr" "EPS" "EU"  "ES2"
```

3.2.2.1. Estimación de parámetros

Se construye la función de verosimilitud asociada a la muestra para las $N = K^2$ observaciones:

$$\mathbb{L}(\mu, \alpha_i, \beta_j, \gamma_k, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{K}[y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_k]^2\right) \tag{3.12}$$

tomamos logaritmos e igualamos las derivadas parciales respecto de los parámetros del modelo, para obtener un sistema de ecuaciones que proporciona los estimadores máximo verosímiles.

Dichos estimadores vienen dados por las expresiones:

$$\hat{\mu} = \bar{y}_{...}, \qquad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \qquad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...} \qquad \hat{\gamma}_k = \bar{y}_{..k} - \bar{y}_{...} \tag{3.13}$$

donde:

$$\bar{y}_{...} = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{ij.}, \qquad \bar{y}_{i..} = \frac{1}{K}\sum_{j=1}^{K} y_{ij.}, \qquad \bar{y}_{.j.} = \frac{1}{K}\sum_{i=1}^{K} y_{ij.}, \qquad \bar{y}_{..k} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{ijk}$$

Un estimador insesgado de la varianza viene dado por:

$$\hat{\sigma}^2 = \frac{SC_\varepsilon}{(K-1)(K-2)}, \qquad con \qquad SC_\varepsilon = \sum_{i=1}^{K}\sum_{j=1}^{K}(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{..k} + 2\bar{y}_{...})^2$$

3.2.2.2. Análisis de la varianza: descomposición de la variabilidad total.

Siguiendo el mismo procedimiento que en modelo en bloques aleatorizados, se comprueba que la descomposición de la variabilidad viene dada por:

$$\begin{aligned}\sum_{i=1}^{K}\sum_{j=1}^{K}(y_{ijk} - \bar{y}_{...})^2 = &\sum_{i=1}^{K} K\,(\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_{j=1}^{K} K\,(\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_{k=1}^{K} K\,(\bar{y}_{..k} - \bar{y}_{...})^2 \\ &+ \sum_{i=1}^{K}\sum_{j=1}^{K}(y_{ijk} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{..k} + 2\bar{y}_{...})^2\end{aligned} \tag{3.14}$$

que simbólicamente podemos escribir:

$$SC_{tot} = SC_\alpha + SC_\beta + SC_\gamma + SC_\varepsilon$$

Los grados de libertad de estas formas cuadráticas son:

$$(K^2-1) = (K-1) + (K-1) + (K-1) + (K-1)(K-2)$$

por lo que, bajo la hipótesis de normalidad, se tiene que:

$$\frac{SC_\alpha}{\sigma^2} \sim \chi^2_{K-1}, \qquad \frac{SC_\beta}{\sigma^2} \sim \chi^2_{K-1}, \qquad \frac{SC_\gamma}{\sigma^2} \sim \chi^2_{K-1}, \qquad \frac{SC_\varepsilon}{\sigma^2} \sim \chi^2_{(K-1)(K-2)}$$

y además son independientes.

A partir de las sumas de cuadrados anteriores y sus grados de libertad, obtenemos los cuadrados medios:

$$CM_\alpha = \frac{SC_\alpha}{K-1}, \qquad CM_\beta = \frac{SC_\beta}{K-1}, \qquad CM_\gamma = \frac{SC_\gamma}{K-1}, \qquad CM_\varepsilon = \frac{SC_\varepsilon}{(K-1)(K-2)}$$

Los valores esperados de los cuadrados medios son:

$$E(CM_\alpha) = \sigma^2 + \frac{K}{K-1}\sum_{i=1}^{K}\alpha_i^2, \qquad E(CM_\beta) = \sigma^2 + \frac{K}{K-1}\sum_{j=1}^{K}\beta_j^2,$$
$$E(CM_\gamma) = \sigma^2 + \frac{K}{K-1}\sum_{k=1}^{K}\gamma_k^2, \qquad E(CM_\varepsilon) = \sigma^2$$

3.2.2.3. Contraste fundamental.

Como en el caso del modelo de bloques aleatorizado, los contrastes estadísticos de más interés en este modelo, son:

$$\begin{aligned} H_{0\alpha} &: \alpha_1 = \alpha_2 = \cdots = \alpha_K, \qquad \forall i \\ H_{0\beta} &: \beta_1 = \beta_2 = \cdots = \beta_K, \qquad \forall j \\ H_{0\gamma} &: \gamma_1 = \gamma_2 = \cdots = \gamma_K, \qquad \forall k \end{aligned}$$

Por consiguiente, los estadísticos de contrastes son:

$$F_\alpha = \frac{SC_\alpha}{SC_\varepsilon}\frac{(K-1)(K-2)\sigma^2}{(K-1)\sigma^2} = \frac{CM_\alpha}{CM_\varepsilon} \sim F_{K-1,(K-1)(K-2)}$$

$$F_\beta = \frac{SC_\beta}{SC_\varepsilon}\frac{(K-1)(K-2)\sigma^2}{(K-1)\sigma^2} = \frac{CM_\beta}{CM_\varepsilon} \sim F_{K-1,(K-1)(K-2)}$$

y

$$F_\gamma = \frac{SC_\gamma}{SC_\varepsilon}\frac{(K-1)(K-2)\sigma^2}{(K-1)\sigma^2} = \frac{CM_\gamma}{CM_\varepsilon} \sim F_{K-1,(K-1)(K-2)}$$

Y se rechazará cualquiera de los H_0, al nivel de significación $1-\alpha$, cuando el valor experimental del respectivo estadístico sea mayor que el valor crítico de la distribución F con $K-1$ y $(K-1)(K-2)$ grados de libertad, es decir:

$$Rechazar\ H_0 \quad si \quad F_{\alpha,\beta\,ó\,\gamma} \geq \mathcal{F}_{K-1,(K-1)(K-2),1-\alpha}$$

Los resultados obtenidos se resumen en la siguiente **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Filas	SC_α	$K-1$	CM_α	$F_\alpha = CM_\alpha/CM_\varepsilon$
Columnas	SC_β	$K-1$	CM_β	$F_\beta = CM_\beta/CM_\varepsilon$
Letras	SC_γ	$K-1$	CM_γ	$F_\gamma = CM_\gamma/CM_\varepsilon$
Error	SC_ε	$(K-1)(K-2)$	CM_ε	
Total	SC_{tot}	K^2-1		

Cuadro 3.3: Tabla Anova para el modelo de cuadrado latino.

NOTA: Si del análisis se concluye que existe diferencia significativa entre los efectos de los tratamientos, será de interés realizar comparaciones entre medias de tratamientos. Para ello, pueden aplicarse los métodos ya estudiados en el experimento completamente aleatorizado. Además, en el caso de que no haya diferencia significativa entre los efectos de los bloques, para ocasiones posteriores, resultará de interés plantearse un nuevo diseño sin la inclusión de dichos bloques en el estudio.

3.2.3. Diseño en cuadrado greco-latino.

Una extensión del diseño en cuadrado latino es el diseño en cuadrado greco-latino.

Supongamos que tenemos un cuadrado latino y superponemos sobre él un segundo cuadrado latino con los tratamientos denotados mediante letras griegas. Si los dos cuadrados latinos tienen la propiedad que cada letra latina coincide exactamente una vez con cada letra griega, se dice entonces que son ortogonales. La superposición de dos cuadrados latinos ortogonales, uno de ellos con los tratamientos denotados mediante letras griegas, se denomina cuadrado greco-latino.

Los cuadrados greco-latinos pueden ser utilizados para el análisis de un factor cuando se desean eliminar tres fuentes de variación en las unidades experimentales.

Un ejemplo de cuadrado latino con $K=4$ sería:

$$\begin{pmatrix} A\alpha & B\beta & C\gamma & D\delta \\ D\gamma & C\delta & B\alpha & A\beta \\ B\delta & A\gamma & D\beta & C\alpha \\ C\beta & D\alpha & A\delta & B\gamma \end{pmatrix}$$

El modelo estadístico para este diseño es:

$$y_{ijkp} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_p + \varepsilon_{ijkp} \qquad con \quad \begin{array}{l} i,j = 1,2,\ldots,K \\ k,p = 1,2,\ldots,K \\ \varepsilon_{ijkp} \text{ i.i.d. } N(0,\sigma^2). \end{array} \tag{3.15}$$

donde:

$\boldsymbol{y_{ijkp}}$ Es la observación correspondiente a la i-ésima fila, la j-ésima columna y la k-ésima letra.

$\boldsymbol{\mu}$ representa la media global de todas las observaciones.

$\boldsymbol{\alpha_i}$ es el efecto del i-ésima fila.

$\boldsymbol{\beta_j}$ es el efecto del j-ésima columna.

$\boldsymbol{\gamma_k}$ es el efecto del k-ésima letra latina.

$\boldsymbol{\delta p}$ es el efecto del k-ésima letra griega.

$\boldsymbol{\varepsilon_{ijkp}}$ es el error aleatorio de la i,j-ésima observación.

Nótese al igual que en el caso de diseño latino, aunque la notación tenga cuatro subíndices, k y p toman valores que dependen de las celdillas (i, j).

Para diseñar un experimento en cuadrado greco-latino a través de R, utilizaremos la función design.graeco. Esta función solo admite modelos para $K = 4$, 8, 10 y 12. Por lo que, a modo de ejemplo, obtendremos el esquema del diseño para un estudio con $K = 8$ tratamientos para cada factor:

```
str(design.graeco)
```

```
## function (trt1, trt2, serie = 2, seed = 0, kinds = "Super-Duper",
##     randomization = TRUE)
```

```
trt1 <-c("A", "B", "C", "D","E", "F", "G", "H")
trt2 <- 1:8
outdesignDCGL <- design.graeco(trt1,trt2, seed=543, serie=2)
print(outdesignDCGL$sketch)
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
## [1,] "B 2" "A 6" "H 7" "F 1" "G 4" "C 8" "D 5" "E 3"
## [2,] "H 6" "D 2" "B 5" "G 8" "F 3" "E 1" "A 7" "C 4"
## [3,] "F 7" "C 5" "G 2" "B 4" "H 1" "A 3" "E 6" "D 8"
## [4,] "C 1" "F 8" "E 4" "A 2" "D 7" "B 6" "G 3" "H 5"
## [5,] "A 4" "B 3" "D 1" "C 7" "E 2" "F 5" "H 8" "G 6"
## [6,] "E 8" "G 1" "C 3" "D 6" "A 5" "H 2" "F 4" "B 7"
## [7,] "G 5" "E 7" "F 6" "H 3" "B 8" "D 4" "C 2" "A 1"
## [8,] "D 3" "H 4" "A 8" "E 5" "C 6" "G 7" "B 1" "F 2"
```

3.2.3.1. Estimación de parámetros

Siguiendo el mismo proceso que en los diseños anteriores, construimos la función de verosimilitud asociada a la muestra para las $N = K^2$ observaciones:

$$\mathbb{L}(\mu, \alpha_i, \beta_j, \gamma_k, \delta_p, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{K}[y_{ijkp} - \mu - \alpha_i - \beta_j - \gamma_k - \delta_p]^2\right)$$

obteniéndose los siguientes estimadores máximo verosimilitudes de los parámetros del modelo:

$$\hat{\mu} = \bar{y}_{....}, \qquad \hat{\alpha}_i = \bar{y}_{i...} - \bar{y}_{....}, \qquad \hat{\beta}_j = \bar{y}_{.j..} - \bar{y}_{....}, \qquad \hat{\gamma}_k = \bar{y}_{..k.} - \bar{y}_{....}, \qquad \hat{\delta}_p = \bar{y}_{...p} - \bar{y}_{....}$$

donde:

$$\bar{y}_{....} = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{ij..}, \qquad \bar{y}_{i...} = \frac{1}{K}\sum_{j=1}^{K} y_{ij..}, \qquad \bar{y}_{.j..} = \frac{1}{K}\sum_{i=1}^{K} y_{ij..},$$
$$\bar{y}_{..k.} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{ijk.}, \qquad \bar{y}_{...p} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K} y_{ij.p}$$

Un estimador insesgado de la varianza viene dado por:

$$\hat{\sigma}^2 = \frac{SC_\varepsilon}{(K-1)(K-3)}, \qquad con \qquad SC_\varepsilon = \sum_{i=1}^{K}\sum_{j=1}^{K}(y_{ijkp} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...p} + 3\bar{y}_{....})^2$$

3.2.3.2. Análisis de la varianza: descomposición de la variabilidad total.

Siguiendo el mismo procedimiento que anteriormente, comprobamos que la ecuación básica del análisis de la varianza es:

$$\sum_{i=1}^{K}\sum_{j=1}^{K}(y_{ijkp} - \bar{y}_{....})^2 = \sum_{i=1}^{K} K(\bar{y}_{i...} - \bar{y}_{....})^2 + \sum_{j=1}^{K} K(\bar{y}_{.j..} - \bar{y}_{....})^2 + \sum_{k=1}^{K} K(\bar{y}_{..k.} - \bar{y}_{....})^2$$
$$+ \sum_{p=1}^{K} K(\bar{y}_{...p} - \bar{y}_{....})^2 + \sum_{i=1}^{K}\sum_{j=1}^{K}(y_{ijkp} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} - \bar{y}_{...p} + 3\bar{y}_{....})^2$$

que simbólicamente podemos escribir:

$$SC_{tot} = SC_\alpha + SC_\beta + SC_\gamma + SC_\delta + SC_\varepsilon$$

Los grados de libertad de estas formas cuadráticas son:

$$(K^2-1) = (K-1) + (K-1) + (K-1) + (K-1)(K-3)$$

por lo que, bajo la hipótesis de normalidad, se tiene que:

$$\frac{SC_\alpha}{\sigma^2} \sim \chi^2_{K-1}, \quad \frac{SC_\beta}{\sigma^2} \sim \chi^2_{K-1}, \quad \frac{SC_\gamma}{\sigma^2} \sim \chi^2_{K-1}, \quad \frac{SC_\delta}{\sigma^2} \sim \chi^2_{K-1}, \quad \frac{SC_\varepsilon}{\sigma^2} \sim \chi^2_{(K-1)(K-3)}$$

y además son independientes.

A partir de las sumas de cuadrados anteriores y sus grados de libertad, obtenemos los cuadrados medios:

$$\begin{aligned} CM_\alpha &= \frac{SC_\alpha}{K-1}, \qquad CM_\beta = \frac{SC_\beta}{K-1}, \\ CM_\gamma &= \frac{SC_\gamma}{K-1}, \qquad CM_\delta = \frac{SC_\delta}{K-1}, \\ CM_\varepsilon &= \frac{SC_\varepsilon}{(K-1)(K-3)} \end{aligned}$$

Los valores esperados de los cuadrados medios son:

$$\begin{aligned} E(CM_\alpha) &= \sigma^2 + \frac{K}{K-1}\sum_{i=1}^{K}\alpha_i^2, \qquad E(CM_\beta) = \sigma^2 + \frac{K}{K-1}\sum_{j=1}^{K}\beta_j^2, \\ E(CM_\gamma) &= \sigma^2 + \frac{K}{K-1}\sum_{k=1}^{K}\gamma_k^2, \quad E(CM_\delta) = \sigma^2 + \frac{K}{K-1}\sum_{k=1}^{K}\delta_p^2, \\ E(CM_\varepsilon) &= \sigma^2 \end{aligned}$$

3.2.3.3. Contraste fundamental.

Los contrastes estadísticos de más interés en este modelo, son:

$$\begin{aligned} H_{0\alpha} &: \alpha_1 = \alpha_2 = \cdots = \alpha_K, \qquad \forall i \\ H_{0\beta} &: \beta_1 = \beta_2 = \cdots = \beta_K, \qquad \forall j \\ H_{0\gamma} &: \gamma_1 = \gamma_2 = \cdots = \gamma_K, \qquad \forall k \\ H_{0\delta} &: \delta_1 = \delta_2 = \cdots = \delta_K, \qquad \forall p \end{aligned}$$

Por consiguiente, los estadísticos de contrastes serían:

$$F_\alpha = \frac{SC_\alpha}{SC_\varepsilon}\frac{(K-1)(K-3)\sigma^2}{(K-1)\sigma^2} = \frac{CM_\alpha}{CM_\varepsilon} \sim F_{K-1,(K-1)(K-3)}$$

$$F_\beta = \frac{SC_\beta}{SC_\varepsilon}\frac{(K-1)(K-3)\sigma^2}{(K-1)\sigma^2} = \frac{CM_\beta}{CM_\varepsilon} \sim F_{K-1,(K-1)(K-3)}$$

$$F_\gamma = \frac{SC_\gamma}{SC_\varepsilon}\frac{(K-1)(K-3)\sigma^2}{(K-1)\sigma^2} = \frac{CM_\gamma}{CM_\varepsilon} \sim F_{K-1,(K-1)(K-3)}$$

y

$$F_\delta = \frac{SC\ _{delta}}{SC_\varepsilon}\frac{(K-1)(K-3)\sigma^2}{(K-1)\sigma^2} = \frac{CM\ _{delta}}{CM_\varepsilon} \sim F_{K-1,(K-1)(K-3)}$$

Y se rechazará cualquiera de los H_0, al nivel de significación $1-\alpha$, cuando el valor experimental del respectivo estadístico sea mayor que el valor crítico de la distribución F con $K-1$ y $(K-1)(K-3)$ grados de libertad, es decir:

$$Rechazar\ H_0 \quad si \quad F_{\alpha,\,\beta\,\acute{o}\,\gamma} \geq \mathcal{F}_{K-1,(K-1)(K-3),1-\alpha}$$

Los resultados obtenidos se resumen en la siguiente **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Filas	SC_α	$K-1$	CM_α	$F_\alpha = CM_\alpha/CM_\varepsilon$
Columnas	SC_β	$K-1$	CM_β	$F_\beta = CM_\beta/CM_\varepsilon$
Letras Latinas	SC_γ	$K-1$	CM_γ	$F_\gamma = CM_\gamma/CM_\varepsilon$
Letras Griegas	SC_δ	$K-1$	CM_δ	$F_\delta = CM_\delta/CM_\varepsilon$
Error	SC_ε	$(K-1)(K-3)$	CM_ε	
Total	SC_{tot}	K^2-1		

Cuadro 3.4: Tabla Anova para el modelo de cuadrado greco-latino.

NOTA: Si del análisis se concluye que existe diferencia significativa entre los efectos de los tratamientos, será de interés realizar comparaciones entre medias de tratamientos. Para ello, pueden aplicarse los métodos ya estudiados en el experimento completamente aleatorizado. Además, en el caso de que no haya diferencia significativa entre los efectos de los bloques, para ocasiones posteriores, resultará de interés plantearse un nuevo diseño sin la inclusión de dichos bloques en el estudio.

3.2.4. Diseño por bloques incompletos balanceado.

En el diseño en bloques aleatorizados completos cada tratamiento es ensayado en todos los bloques. Un problema relacionado con este diseño es que puede ocurrir que no todos los tratamientos puedan ser ensayados en todos los bloques debido, por ejemplo, a escasez de recursos. En este caso pueden utilizarse diseños en los que cada tratamiento no esté presente en todos los bloques. A tales diseños se les denomina diseños en bloques incompletos. En este apartado se estudiará un tipo particular de diseño en bloques incompletos, el diseño por bloques incompletos balanceado

Un diseño balanceado (o equilibrado) por bloques incompletos es un diseño en bloques incompletos en el que cualquier par de tratamientos aparecen juntos en un mismo bloque igual número de veces, y ningún tratamiento aparece más de una vez en cualquier bloque. Abreviadamente los denotaremos **DBIB** o diseño **BIB**.

Supongamos que se tienen I tratamientos de los cuales sólo se pueden experimentar K con $(K < I)$ tratamientos en cada bloque. Se puede construir un diseño tomando $\binom{I}{K}$ bloques de forma que a cada bloque se le asigne una de las $\binom{I}{K}$ combinaciones de tratamientos posibles.

Los parámetros que caracterizan este modelo son los siguientes:

- **I**, número de tratamientos o niveles del factor principal.
- **J**, número de bloques.
- **K**, número de tratamientos por bloque.
- **R**, número de veces que cada tratamiento se presenta en el diseño, es decir el número de réplicas de un tratamiento dado.
- $\boldsymbol{\lambda}$, número de bloques en los que un par de tratamientos ocurren juntos.
- **N**, número total de observaciones.

Estos 6 parámetros, I, J, K, R λ y N, no son independientes. Son enteros no negativos verificando:

1. $N = IR = JK$
2. $\lambda = R\dfrac{K-1}{I-1}$
3. $J \geq I$ (Cuando $J = I$ el diseño recibe el nombre de simétrico.)

El modelo es el mismo que el diseño por bloques completos con la siguiente modificación:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad si \quad n_{ij} = 1 \qquad con \quad \begin{array}{l} i = 1, 2, ..., I. \\ j = 1, 2, ..., J. \\ \varepsilon_{ij} \text{ i.i.d. } N(0, \sigma^2). \end{array} \tag{3.16}$$

Para diseñar un experimento por bloques incompletos balanceado en *R*, nos apoyaremos en la función *design.bib*. A modo de ejemplo, durante esta sección usaremos la muestra obtenida en el capítulo 2: *Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno*, para ver y analizar cómo realizar un diseño de este tipo. Lo primero que veremos es cómo plantear la estructura de un posible diseño, así pues:

```
str(design.bib)
```

```
## function (trt, k, r = NULL, serie = 2, seed = 0, kinds = "Super-Duper",
##     maxRep = 20, randomization = TRUE)
```

```
trt <- c("SE","EPr","ES1","ES2","EPS","EU")
outdesignBIB <- design.bib(trt,k=5,maxRep=400, seed=543, serie=2)
```

```
##
## Parameters BIB
## ==============
## Lambda     : 4
## treatmeans : 6
## Block size : 5
## Blocks     : 6
## Replication: 5
##
## Efficiency factor 0.96
##
## <<< Book >>>
```

```
DisBIB <- outdesignBIB$book
print(outdesignBIB$sketch)
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]
## [1,] "EPS" "SE"  "ES2" "EU"  "ES1"
## [2,] "SE"  "ES2" "EPr" "ES1" "EPS"
## [3,] "EPS" "ES1" "EPr" "SE"  "EU"
## [4,] "ES2" "EPr" "EU"  "EPS" "SE"
## [5,] "ES2" "ES1" "SE"  "EU"  "EPr"
## [6,] "EU"  "ES2" "EPr" "EPS" "ES1"
```

3.2.4.1. Estimación de parámetros

En el modelo BIB, las estimaciones de los parámetros del modelo vienen dadas por:

$$\hat{\mu} = \bar{y}_{..}, \qquad \hat{\alpha}_i = \frac{KQ_i}{\lambda I}, \qquad \hat{\beta}_j = \frac{RQ_j}{\lambda J}$$

donde:

$$\bar{y}_{..} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}n_{ij}, \quad Q_i = T_i - \frac{1}{K}\sum_{j=1}^{J} n_{ij}B_j, \quad Q_j = B_j - \frac{1}{R}\sum_{i=1}^{I} n_{ij}T_i,$$

$$B_j = \sum_{i=1}^{I} y_{ij}n_{ij}, \qquad T_i = \sum_{j=1}^{J} y_{ij}n_{ij} \qquad n_{ij} = \begin{cases} 1 & \text{si el trat. } i \text{ ocurre en el bloque } j \\ 0 & \text{en caso contrario} \end{cases}$$

3.2.4.2. Análisis de la varianza y Contraste fundamental.

En este diseño la variabilidad total SC_{tot} se descompone en:

$$SC_{tot} = SC_{\alpha^*} + SC_{\beta} + SC_{\varepsilon}$$

donde hemos desglosado la variabilidad total de los datos en tres componentes:

- SC_{tot} tiene la misma expresión que en el diseño en bloques completos aleatorizados, es decir
$$SC_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J} (y_{ij} - \bar{y}_{..})^2$$
- La suma de cuadrados de tratamientos ajustada, que tiene la siguiente expresión:
$$SC_{\alpha^*} = \sum_{i=1}^{I} \frac{KQ_i^2}{\lambda I}$$
- Como en este diseño se realizan K de los I tratamientos en cada bloque, la suma de cuadrados correspondiente a los bloques tiene la siguiente expresión:
$$SC_{\beta} = \frac{1}{K}\sum_{j=1}^{J} B_j^2 - \frac{1}{N}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}n_{ij}\right)^2$$
- SC_{ε} se calcula a partir de las otras sumas de cuadrados, es decir:
$$SC_{\varepsilon} = SC_{tot} - SC_{\alpha^*} - SC_{\beta}$$

Los grados de libertad de estas formas cuadráticas son:

$$(N-1) = (I-1) + (J-1) + (N-I-J+1)$$

Los cuadrados medios tienen las siguientes expresiones:

$$CM_{\alpha^*} = \frac{SC_{\alpha^*}}{I-1}, \qquad CM_{\varepsilon} = \frac{SC_{\varepsilon}}{(N-I-J+1)}$$

Bajo las hipótesis de igualdad de efectos de los tratamientos y los bloques, se verifica:

$$F_{\alpha} = \frac{SC_{\alpha^*}}{SC_{\varepsilon}}\frac{(N-I-J+1)\sigma^2}{(I-1)\sigma^2} = \frac{CM_{\alpha^*}}{CM_{\varepsilon}} \sim F_{I-1,(N-I-J+1)}$$

Los resultados obtenidos se resumen en la siguiente **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Factor	SC_{α^*}	$I-1$	CM_{α^*}	$F_\alpha = CM_{\alpha^*}/CM_\varepsilon$
Bloques	SC_β	$J-1$		
Error	SC_ε	$(N-I-J+1)$	CM_ε	
Total	SC_{tot}	$IJ-1$		

Cuadro 3.5: Tabla Anova para el modelo de BIB (Ajustando la suma de cuadrados de tratamientos.

En algunas ocasiones puede resultar de interés contrastar también la igualdad de efectos de los bloques, para ello la suma de cuadrados total se debe descomponer de la siguiente forma:

$$SC_{tot} = SC_\alpha + SC_{\beta^*} + SC_\varepsilon$$

donde:

- SC_{tot} tiene la misma expresión, es decir:

$$SC_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{..})^2$$

- SC_α La suma de cuadrados de tratamientos no-ajustada:

$$SC_\alpha = \frac{1}{R}\sum_{i=1}^{I} T_i^2 - \frac{1}{N}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}n_{ij}\right)^2$$

- SC_{β^*} es la suma de cuadrados ajustada de los bloques, que en el caso del diseño en bloques incompletos balanceado tiene la siguiente expresión:

$$SC_{\beta^*} = \sum_{j=1}^{J}\frac{RQ_j^2}{\lambda J}$$

- SC_ε se calcula a partir de las otras sumas de cuadrados, es decir:

$$SC_\varepsilon = SC_{tot} - SC_\alpha - SC_{\beta^*}$$

Los cuadrados medios tienen las siguientes expresiones:

$$CM_{\beta^*} = \frac{SC_{\beta^*}}{J-1}, \qquad CM_\varepsilon = \frac{SC_\varepsilon}{(N-I-J+1)}$$

Bajo las hipótesis de igualdad de efectos de los tratamientos y los bloques, se verifica:

$$F_\beta = \frac{SC_{\beta^*}}{SC_\varepsilon}\frac{(N-I-J+1)\sigma^2}{(J-1)\sigma^2} = \frac{CM_{\beta^*}}{CM_\varepsilon} \sim F_{J-1,(N-I-J+1)}$$

Los resultados obtenidos se resumen en la siguiente **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Factor	SC_α	$I-1$		
Bloques	SC_{β^*}	$J-1$	CM_{β^*}	$F_\beta = CM_{\beta^*}/CM_\varepsilon$
Error	SC_ε	$(N-I-J+1)$	CM_ε	
Total	SC_{tot}	$N-1$		

Cuadro 3.6: Tabla Anova para el modelo de BIB (Ajustando la suma de cuadrados de los bloques).

NOTA: Si del análisis se concluye que existe diferencia significativa entre los efectos de los tratamientos, será de interés realizar comparaciones entre medias de tratamientos. Para ello, pueden aplicarse los métodos ya estudiados en el experimento completamente aleatorizado. Además, en el caso de que no haya diferencia significativa entre los efectos de los bloques, para ocasiones posteriores, resultará de interés plantearse un nuevo diseño sin la inclusión de dichos bloques en el estudio.

A continuación, para ilustrar los resultados obtenidos en esta sección, vamos a realizar a través de *R* el análisis de la varianza y las posteriores comparaciones múltiples de la *Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.*

```
LSDBib<-BIB.test(MuestraBIB$Provincia,MuestraBIB$Estudios,
              MuestraBIB$NotaMedia, test="lsd", group=TRUE, console=TRUE)
```

```
##
## ANALYSIS BIB:  MuestraBIB$NotaMedia
## Class level information
##
## Block:  Alm Cad CoryJae Gra Mal SevyHue
## Trt  :  EPS SE ES2 EU ES1 EPr
##
## Number of observations:  30
##
## Analysis of Variance Table
##
## Response: MuestraBIB$NotaMedia
##             Df Sum Sq Mean Sq F value Pr(>F)
## block.unadj  5   18.3    3.67    1.15  0.371
## trt.adj      5   52.7   10.55    3.29  0.026 *
## Residuals   19   60.9    3.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## coefficient of variation: 33 %
## MuestraBIB$NotaMedia Means: 5.4
##
## MuestraBIB$Estudios,  statistics
```

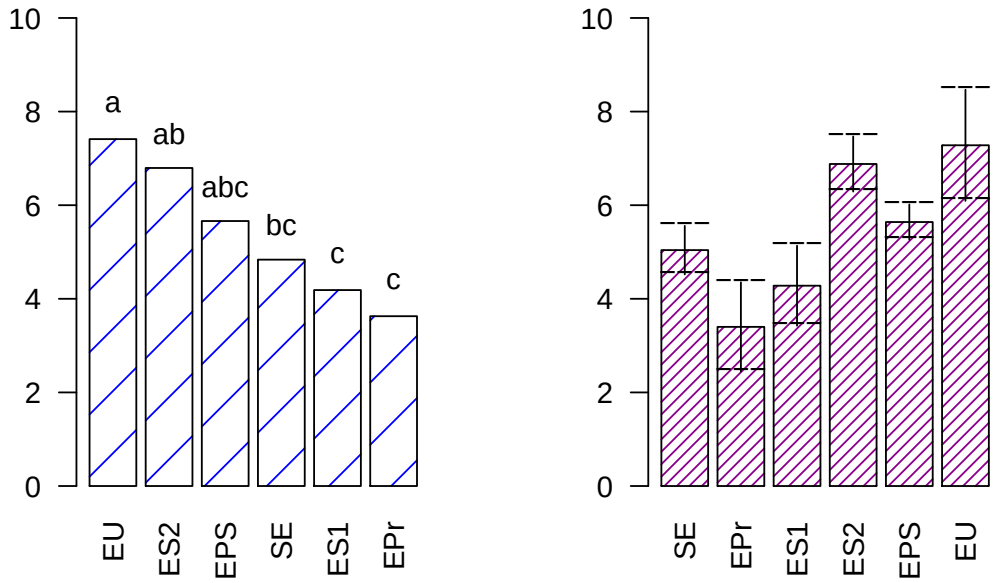
```
##
##     MuestraBIB$NotaMedia mean.adj   SE r  std Min  Max
## SE                   5.0      4.8 0.81 5 1.16 1.0  6.0
## EPr                  3.4      3.6 0.81 5 2.13 4.8  6.8
## ES1                  4.3      4.2 0.81 5 1.91 1.2  5.8
## ES2                  6.9      6.8 0.81 5 1.32 5.4  8.2
## EPS                  5.6      5.7 0.81 5 0.84 4.4 10.0
## EU                   7.3      7.4 0.81 5 2.66 4.0  6.8
##
## LSD test
## Std.diff   : 1.2
## Alpha      : 0.05
## LSD        : 2.4
## Parameters BIB
## Lambda     : 4
## treatmeans : 6
## Block size : 5
## Blocks     : 6
## Replication: 5
##
## Efficiency factor 0.96
##
## <<< Book >>>
##
## Comparison between treatments means
##           Difference pvalue sig.
## SE - EPr        1.21  0.309
## SE - ES1        0.65  0.580
## SE - ES2       -1.96  0.106
## SE - EPS       -0.83  0.484
## SE - EU        -2.58  0.038    *
## EPr - ES1      -0.56  0.634
## EPr - ES2      -3.17  0.013    *
## EPr - EPS      -2.03  0.094    .
## EPr - EU       -3.78  0.004   **
## ES1 - ES2      -2.61  0.036    *
## ES1 - EPS      -1.48  0.217
## ES1 - EU       -3.23  0.012    *
## ES2 - EPS       1.13  0.339
## ES2 - EU       -0.62  0.600
## EPS - EU       -1.75  0.146
##
## Treatments with the same letter are not significantly different.
##
##     MuestraBIB$NotaMedia groups
## EU                   7.4      a
## ES2                  6.8     ab
## EPS                  5.7    abc
```

```
## SE                    4.8     bc
## ES1                   4.2      c
## EPr                   3.6      c
```

Conclusión: Rechazamos la hipótesis nula de la igualdad de los efectos de los diferentes tratamientos del Nivel de estudios máximo de los padres/tutores, por lo tanto, tiene sentido realizar las correspondientes comparaciones múltiples. Así pues, realizando éstas vemos que los alumnos que mejor rendimiento han obtenido son aquellos con al menos un padre o tutor *universitario* y, por el contrario, los que peores resultados académicos tienen son los alumnos con padres o tutores con nivel máximo de estudio *Enseñanza secundaria de 1era etapa* o *Educación primaria*.

Podemos representar las diferentes agrupaciones entre las cuales, los tratamientos que formen parte del mismo grupo no tengan diferencias significativas en el rendimiento escolar del alumno a través de:

```
par(mfrow=c(1,2))
bar.group(LSDBib$groups,col="blue",density=4,las=2,ylim=c(0,10))
bar.err(LSDBib$means, col=colors()[84], ylim = c(0, 10),frequency=1,
        las=2,density=20)
```



3.3. Comparaciones no paramétricas

Hasta ahora hemos supuesto que se cumplían las hipótesis iniciales que vimos en el Capítulo 1. Sin embargo, será frecuente que nuestro modelo no sea el adecuado ya que puede que no cumpla alguna de estas hipótesis. Existen una serie de técnicas estadísticas que podremos usar en este caso. Esta rama de la estadística inferencial se conoce como **estadística no paramétrica**. Estas técnicas resultan de gran interés, debido a que no necesitan que se especifique la forma de la distribución de la población.

A continuación, veremos algunas pruebas no paramétricas más importantes.

3.3.1. Prueba de Kruskall-wallis.

Como alternativa no paramétrica al Análisis de la Varianza para el caso del **experimento completamente aleatorizado** para el contraste:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k$$
$$H_1 : \alpha_i \neq \alpha_j, \quad \textit{para algunos } i \neq j$$

se utiliza el test de Kruskall-wallis de igualdad de k poblaciones continuas con muestras independientes, ya que no asume normalidad en los datos, en oposición al tradicional ANOVA.

Supongamos que aplicamos a un grupo de N individuos k tratamientos diferentes, en distintos momentos. Sea R_i la suma de los rangos del i-ésimo tratamiento. Entonces el procedimiento consiste en:

– Ordenar conjuntamente (de menor a mayor) las observaciones muestrales.

– Asignar los rangos (de forma natural) a dichas observaciones.

– Calcular para cada muestra la suma R_i.

Si H_0 es cierta, la suma total de rangos deberá estar repartida proporcionalmente entre las k muestras en función de sus tamaños, por lo que el estadístico propuesto es:

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

El estadístico H se distribuye asintóticamente, bajo H_0 , como una chi-cuadrado con $k-1$ grados de libertad. De forma que, dado un nivel de significación α, si el estadístico evaluado H supera el valor crítico, $\chi_{k-1,1-\alpha}$, rechazaremos la hipótesis nula de que las muestras provienen de poblaciones iguales.

NOTA: En el caso de que la hipótesis nula sea rechazada, se debería de recurrir a alguna técnica de comparaciones múltiples como, por ejemplo, el *test de Kruskal-Nemenyi*, para poder obtener más información sobre los diferentes tratamientos. Sin embargo, por la limitación de esta memoria, no profundizaremos en el análisis y obtención de éste test.

Dado que en el estudio de la *Muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres* que realizamos en el capítulo 2, dio como resultado que no había evidencias necesarias para rechazar la Hipótesis de Normalidad para dicha muestra, sólo realizaremos una simulación de cómo habría que actuar en caso de rechazar esta hipótesis. Es decir, solo expondremos los códigos necesarios para realizar la prueba de Kruskall-wallis de nuestra muestra, pero sin la ejecución de éstos. Una posible forma, sería la siguiente:

```
str(kruskal)
```

```
## function (y, trt, alpha = 0.05, p.adj = c("none", "holm", "hommel",
##     "hochberg", "bonferroni", "BH", "BY", "fdr"), group = TRUE, main = NULL,
##     console = FALSE)
```

```
outKruskal<-with(Muestra1factor,kruskal(Muestra1factor$NotaMedia,
                                        Muestra1factor$Estudios,group=TRUE,
                                        console=TRUE))
```

```
outKruskal2<-with(Muestra1factor,kruskal(Muestra1factor$NotaMedia,
                                 Muestra1factor$Estudios,group=TRUE,
                                 p.adj="holm"))
print(outKruskal2$group)
```

3.3.2. Prueba de Friedman.

La prueba de Friedman es una prueba no paramétrica desarrollado por el economista Milton Friedman y, al igual que en la prueba de Kruskall-wallis, es equivalente a la prueba ANOVA, pero para el caso del **experimento en bloques aleatorizados completos**.

Supongamos que aplicamos a un grupo de N individuos I tratamientos diferentes, en distintos momentos. Sea R_i la suma de los rangos del i-ésimo tratamiento. Lo que queremos es contrastar la Hipótesis nula de igualdad entre esos tratamientos. Ahora la igualdad no será de medias, como en el ANOVA en el caso paramétrico, sino que será igualdad de medianas o de distribuciones.

El método consiste en ordenar conjuntamente (de menor a mayor) las observaciones muestrales dentro de cada bloque, se asignan los rangos (de forma natural) a dichas observaciones y por último, se calcula para cada muestra la suma de los rangos correspondientes.

El estadístico propuesto en este test de Friedman es el siguiente:

$$F = \frac{12}{NI(I+1)} \sum_{i=1}^{I} R_i^2 - 3N(I+1) \sim \chi_{I-1,1-\alpha}$$

Luego la región crítica del test sería:

$$Rechazar\ H_0 \quad si \quad F \geq \chi_{I-1,1-\alpha}$$

Como ya adelantamos en la sección de *Diseño en bloque aleatorizados completos*, al no cumplir la *Muestra balanceada: Nota en matemáticas del alumno, Nivel de estudios*

máximo de los padres/tutores y Lugar de residencia del alumno la hipótesis de Normalidad, debemos realizar el análisis de la varianza a través de la prueba de Friedman. Podremos realizar también, en el caso de rechazar la igualdad de los efectos, las correspondientes comparaciones múltiples, de manera que:

```
str(friedman)
```

```
## function (judge, trt, evaluation, alpha = 0.05, group = TRUE, main = NULL,
##     console = FALSE)
```

```
outFried<-with(MuestraBAC,friedman(MuestraBAC$Provincia,MuestraBAC$Estudios,
                                   MuestraBAC$NotaMates,alpha=0.05, group=FALSE,
                                   console=TRUE))
```

```
##
## Study: MuestraBAC$NotaMates ~ MuestraBAC$Provincia + MuestraBAC$Estudios
##
## MuestraBAC$Estudios,  Sum of the ranks
##
##     MuestraBAC.NotaMates r
## EPr                   20 6
## EPS                   18 6
## ES1                   25 6
## ES2                   22 6
## EU                    20 6
## SE                    21 6
##
## Friedman's Test
## ===============
## Adjusted for ties
## Critical Value: 2.1
## P.Value Chisq: 0.83
## F Value: 0.38
## P.Value F: 0.86
##
## Post Hoc Analysis
##
## Comparison between treatments
## Sum of the ranks
##
##           difference pvalue signif.   LCL  UCL
## EPr - EPS        2.5   0.67          -9.6 14.6
## EPr - ES1       -5.0   0.40         -17.1  7.1
## EPr - ES2       -2.5   0.67         -14.6  9.6
## EPr - EU         0.0   1.00         -12.1 12.1
## EPr - SE        -1.0   0.87         -13.1 11.1
## EPS - ES1       -7.5   0.21         -19.6  4.6
## EPS - ES2       -5.0   0.40         -17.1  7.1
## EPS - EU        -2.5   0.67         -14.6  9.6
## EPS - SE        -3.5   0.56         -15.6  8.6
```

```
## ES1 - ES2        2.5   0.67         -9.6 14.6
## ES1 - EU         5.0   0.40         -7.1 17.1
## ES1 - SE         4.0   0.50         -8.1 16.1
## ES2 - EU         2.5   0.67         -9.6 14.6
## ES2 - SE         1.5   0.80        -10.6 13.6
## EU - SE         -1.0   0.87        -13.1 11.1
```

Conclusión: No hay evidencias suficientes para rechazar la hipótesis nula de la igualdad de los efectos entre los distintos tratamientos del *Nivel de estudios máximo de los padres o tutores*, es decir, parece ser que el *Nivel de estudios máximo de los padres o tutores* no influyen en la *Nota en matemáticas* de los alumnos del territorio andaluz. Por lo tanto, no tiene sentido realizar las correspondientes comparaciones múltiples.

NOTA: Al igual que en la prueba de Kruskal-Wallis, en el caso de que la hipótesis nula sea rechazada, se debería de recurrir a alguna técnica de comparaciones múltiples como, por ejemplo, el *test de Kruskal-Nemenyi*.

3.3.3. Prueba de Durbin.

La prueba de Durbin es un test no paramétrico que puede utilizarse en el caso del **experimento por bloques incompletos balanceado (BIB)**. Este test se reduce al test de Friedman si el número de tratamientos es igual al número de unidades experimentales por bloque.

Sea y_{ij} el resultado del tratamiento i en el bloque j, si el tratamiento i aparece en el bloque j y sea $R_{ij} = R(y_{ij})$ el rango de y_{ij} dentro de su bloque. Estos rangos toman valores entre 1 y K. Sea R_i la suma de los rangos correspondientes al tratamiento i. Al igual que en las técnicas anteriores, lo que queremos es contrastar la Hipótesis nula de igualdad entre esos tratamientos.

Por lo que, el estadístico usado en el test de Durbin es el siguiente:

$$T = \frac{12(I-1)}{RI(K^2+1)} \sum_{i=1}^{I} \left(R_i - \frac{R(K+1)}{2} \right)^2 \sim \chi_{I-1,1-\alpha}$$

Luego la región crítica viene dada por:

$$Rechazar\ H_0 \quad si \quad F \geq \chi_{I-1,1-\alpha}$$

Al igual que en el apartado de Kruskall-wallis, solo expondremos los códigos necesarios para realizar la prueba de Durbin de la *Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Sexo del alumno*, pero sin la ejecución de éstos. Una posible forma, sería la siguiente:

```
str(durbin.test)
```

```
## function (judge, trt, evaluation, alpha = 0.05, group = TRUE, main = NULL,
##     console = FALSE)
```

```
outDurb<-durbin.test(MuestraBIB$Provincia,MuestraBIB$Estudios,
                     MuestraBIB$NotaMedia,group=FALSE,console=TRUE)
```

NOTA: Al igual que en la prueba de Kruskal-Wallis, en el caso de que la hipótesis nula sea rechazada, se debería de recurrir a alguna técnica de comparaciones múltiples como, por ejemplo, el *test de Kruskal-Nemenyi*.

3.4. Experimentos con dos factores

Si bien hasta ahora nuestro objetivo ha sido estudiar la influencia de un único factor, muchos experimentos se llevan a cabo para estudiar el efecto que sobre una variable respuesta Y tienen dos o más factores.

Por **diseño factorial** se entiende aquel en el que se investigan todas las posibles combinaciones de los niveles de los factores en cada ensayo o réplica del experimento.

Por ejemplo, en un diseño con dos factores, llamémosles A y B, con a y b niveles respectivamente, cada réplica del experimento consiste en observar una respuesta en cada una de las ab combinaciones de los tratamientos o niveles de ambos factores. También se dice que los factores están cruzados. El **efecto de un factor** se define como el cambio esperado en la respuesta producido por un cambio en el nivel del factor, donde el cambio en la respuesta se obtiene promediando sobre todas las combinaciones de niveles del resto de los factores. También se le denomina **efecto principal**. En algunos experimentos puede ocurrir que la diferencia de respuesta entre dos niveles de un factor no sea la misma para todos los niveles del otro (u otros) factor (o factores). Cuando esto ocurre, se dice que existe **interacción** entre los factores.

A continuación, se estudiará el diseño factorial más simple, el diseño con dos factores.

3.4.1. Diseño factorial con dos factores.

Supongamos un experimento con dos factores cruzados A y B, con a y b niveles respectivamente. Supongamos que se realizan $n\,(n \geq 2)$ réplicas del experimento, es decir, estamos considerando un diseño de 2 factores balanceado como se muestra en la siguiente tabla:

		Factor B			
		1	2	...	b
Factor A	1	$y_{111}, y_{112}, \ldots, y_{11n}$	$y_{121}, y_{122}, \ldots, y_{12n}$	$\cdots$	$y_{1b1}, y_{1b2}, \ldots, y_{1bn}$
	2	$y_{211}, y_{212}, \ldots, y_{21n}$	$y_{221}, y_{222}, \ldots, y_{22n}$	$\cdots$	$y_{2b1}, y_{2b2}, \ldots, y_{2bn}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	a	$y_{a11}, y_{a12}, \ldots, y_{a1n}$	$y_{a21}, y_{a22}, \ldots, y_{a2n}$	$\cdots$	$y_{ab1}, y_{ab2}, \ldots, y_{abn}$

Hay un total de $N = nab$ observaciones. Supondremos que el experimento es completamente aleatorizado, es decir, que la asignación de combinaciones de tratamientos a las unidades experimentales se realiza de manera aleatoria.

Las observaciones pueden describirse mediante el siguiente modelo lineal:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \quad con \quad \begin{array}{l} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, n \\ \varepsilon_{ijk} \text{ i.i.d. } N(0, \sigma^2). \end{array} \tag{3.17}$$

donde:

$\boldsymbol{y_{ijk}}$ representa la k-ésima observación en el i-ésimo nivel de A y en el j-ésimo nivel de B.

$\boldsymbol{\mu}$ representa la media global de todas las observaciones.

$\boldsymbol{\alpha_i}$ es el efecto principal del i-ésimo nivel del factor A.

$\boldsymbol{\beta_j}$ es el efecto principal del j-ésimo nivel del factor B.

$\boldsymbol{\alpha\beta_{ij}}$ es el efecto de interacción del i-ésimo nivel del factor A y del j-ésimo nivel del factor B.

$\boldsymbol{\varepsilon_{ijk}}$ es el efecto aleatorio que recoge todas las posibles causas restantes de variabilidad del experimento.

Matricialmente, el modelo puede expresarse:

$$Y = X\theta + \varepsilon$$

donde la matriz X tiene dimensiones $N \times (1+a+b+ab)$, con rango $rg(X) = ab$, pues las ab columnas de X correspondientes a las ab interacciones son linealmente independientes,la columna correspondiente a α_i es la suma en j de las columnas correspondientes a $\alpha\beta ij$, la columna correspondiente a βj es la suma en i de las columnas correspondientes a $\alpha\beta ij$, y la columna correspondiente a μ es la suma de todas las columnas correspondientes a las interacciones.

Para diseñar un experimento factorial con dos factores en *R*, nos apoyaremos en la libería *Agricolae*[10] a través de la función *design.bib*. A modo de ejemplo, durante esta sección usaremos la muestra obtenida en el capítulo 2: *Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno*, para ver y analizar cómo realizar un diseño de este tipo. Lo primero que veremos es cómo plantear la estructura de un posible diseño, de manera que:

```
str(design.ab)
```

```
## function (trt, r = NULL, serie = 2, design = c("rcbd", "crd", "lsd"),
##     seed = 0, kinds = "Super-Duper", first = TRUE, randomization = TRUE)
```

```
trt<-c(5,2) # factorial 5x2 (5 y 2 niveles respectivamente)
outdesign2fact <-design.ab(trt, r=40, serie=2)
Dis2Fact <- outdesign2fact$book
head(Dis2Fact)
```

```
##   plots block A B
## 1   101     1 1 2
## 2   102     1 1 1
## 3   103     1 2 2
## 4   104     1 2 1
## 5   105     1 4 1
## 6   106     1 5 1
```

Recordemos que los niveles de los dos factores para este estudio son:

```
levels(Muestra2factores$Estudios)
```

```
## [1] "EBa" "ES1" "ES2" "EPS" "EU"
```

```
levels(Muestra2factores$Sexo)
```

```
## [1] "H" "M"
```

3.4.1.1. Estimación de parámetros

Construimos la función de verosimilitud asociada a la muestra:

Los estimadores máximo verosímiles de los parámetros del modelo son:

$$\mathbb{L}(\mu, \alpha_i, \beta_j, \alpha\beta_{ij}, \sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[y_{ijk} - \mu - \alpha_i - \beta_j - \alpha\beta_{ij}]^2\right)$$

tomamos logaritmos e igualamos las derivadas parciales respecto de los parámetros del modelo, para obtener un sistema de ecuaciones que proporciona los estimadores máximo verosímiles.

Los estimadores máximo verosímiles de los parámetros del modelo son:

$$\hat{\mu} = \bar{y}_{...}, \qquad \hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}, \qquad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...} \qquad \hat{\alpha\beta}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{...} \tag{3.18}$$

donde:

$$\bar{y}_{...} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}, \qquad \bar{y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}, \qquad \bar{y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}, \qquad \bar{y}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}$$

Un estimador insesgado de la varianza viene dado por:

$$\hat{\sigma}^2 = \frac{SC_\varepsilon}{ab(n-1)}, \qquad con \qquad SC_\varepsilon = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$$

3.4.1.2. Análisis de la varianza: descomposición de la variabilidad total.

A partir de la siguiente identidad:

$$y_{ijk} - \bar{y}_{...} = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (y_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (\bar{y}_{ijk} - \bar{y}_{ij.}) \tag{3.19}$$

podemos obtener la siguiente descomposición:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 = bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$$

que simbólicamente podemos escribir:

$$SC_{tot} = SC_\alpha + SC_\beta + SC_{\alpha\beta} + SC_\varepsilon$$

Los grados de libertad de estas formas cuadráticas son:

$$abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1)$$

por lo que, bajo la hipótesis de normalidad, se tiene que:

$$\frac{SC_\alpha}{\sigma^2} \sim \chi^2_{a-1}, \qquad \frac{SC_\beta}{\sigma^2} \sim \chi^2_{b-1}, \qquad \frac{SC_{\alpha\beta}}{\sigma^2} \sim \chi^2_{(a-1)(b-1)}, \qquad \frac{SC_\varepsilon}{\sigma^2} \sim \chi^2_{ab(n-1)}$$

y además son independientes.

A partir de las sumas de cuadrados anteriores y sus grados de libertad, obtenemos los cuadrados medios:

$$CM_\alpha = \frac{SC_\alpha}{a-1}, \qquad CM_\beta = \frac{SC_\beta}{b-1}, \qquad CM_{\alpha\beta} = \frac{SC_{\alpha\beta}}{(a-1)(b-1)}, \qquad CM_\varepsilon = \frac{SC_\varepsilon}{ab(n-1)}$$

Los valores esperados de los cuadrados medios son:

$$E(CM_\alpha) = \sigma^2 + \frac{bn}{a-1}\sum_{i=1}^{a}\alpha_i^2, \qquad E(CM_\beta) = \sigma^2 + \frac{an}{b-1}\sum_{j=1}^{b}\beta_j^2,$$

$$E(CM_{\alpha\beta}) = \sigma^2 + \frac{n}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\alpha\beta_{ij}^2, \qquad E(CM_\varepsilon) = \sigma^2$$

3.4.1.3. Contraste fundamental.

El objetivo del análisis es realizar los contrastes de hipótesis nula:

1. $H_{0\alpha} : \alpha_1 = \alpha_2 = \cdots = \alpha_K, \quad \forall i.$ Es decir, considerando la presencia de las interacciones con el factor β, contrastar si los efectos de los niveles del factor α son nulos. El estadístico de contraste es:

$$F_\alpha = \frac{SC_\alpha}{SC_\varepsilon}\frac{ab(n-1)\sigma^2}{(a-1)\sigma^2} = \frac{CM_\alpha}{CM_\varepsilon} \sim F_{(a-1),ab(n-1)}$$

   Se rechaza $H_{0\alpha}$ al nivel $1-\alpha$ si: $F_\alpha \geq \mathcal{F}_{a-1,ab(n-1),1-\alpha}$

2. $H_{0\beta} : \beta_1 = \beta_2 = \cdots = \beta_K, \quad \forall j.$ Es decir, considerando la presencia de las interacciones con el factor A, contrastar si los efectos de los niveles del factor B son nulos. El estadístico de contraste es:

$$F_\beta = \frac{SC_\beta}{SC_\varepsilon}\frac{ab(n-1)\sigma^2}{(b-1)\sigma^2} = \frac{CM_\beta}{CM_\varepsilon} \sim F_{b-1,ab(n-1)}$$

   Se rechaza $H_{0\beta}$ al nivel $1-\alpha$ si: $F_\beta \geq \mathcal{F}_{b-1,ab(n-1),1-\alpha}$

3. $H_{0\alpha\beta} : \alpha\beta_1 = \alpha\beta_2 = \cdots = \alpha\beta_K, \quad \forall k.$ Es decir, contrastar si los efectos de las interacciones entre los factores A y B son nulos. Este contraste es quizás el más importante, ya que si resulta que no existe interacción entre los factores, lo más inteligente sería estudiar la influencia de los factores en la variable respuesta por separado.

   El estadístico de contraste es

$$F_{\alpha\beta} = \frac{SC_{\alpha\beta}}{SC_\varepsilon}\frac{ab(n-1)\sigma^2}{(a-1)(b-1)\sigma^2} = \frac{CM_{\alpha\beta}}{CM_\varepsilon} \sim F_{(a-1)(b-1),ab(n-1)}$$

   Se rechaza $H_{0\alpha\beta}$ al nivel $1-\alpha$ si: $F_\alpha \geq \mathcal{F}_{a-1,ab(n-1),1-\alpha\beta}$

Los resultados obtenidos se resumen en la siguiente **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
Factor A	SC_α	$a-1$	CM_α	$F_\alpha = CM_\alpha/CM_\varepsilon$
Factor B	SC_β	$b-1$	CM_β	$F_\beta = CM_\beta/CM_\varepsilon$
Interacción	$SC_{\alpha\beta}$	$(a-1)(b-1)$	$CM_{\alpha\beta}$	$F_{\alpha\beta} = CM_{\alpha\beta}/CM_\varepsilon$
Error	SC_ε	$ab(n-1)$	CM_ε	
Total	SC_{tot}	$nab-1$		

Cuadro 3.7: Tabla Anova para el modelo factorial con dos factores.

NOTA: La diagnosis y validación del modelo se realiza igual que en los modelos anteriores.

A continuación, para ilustrar estos resultados, realizamos a través de *R* el análisis de la varianza a la *Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno*. Además, realizaremos una regresión lineal para obtener las estimaciones de los parámetros correspondientes:

```
model2factores<-aov(Muestra2factores$Nota~
                      Muestra2factores$Estudios*Muestra2factores$Sexo,
                    data=Muestra2factores)
model2factores2<-lm(Muestra2factores$Nota~
                      Muestra2factores$Estudios*Muestra2factores$Sexo,
                    data=Muestra2factores)
```

Obtenemos los grados de libertad y el valor esperado de los cuadrados medios para el error:

```
(df<-df.residual(model2factores))
```

```
## [1] 390
```

```
(MSerror<-deviance(model2factores)/df)
```

```
## [1] 3.3
```

TABLA ANOVA:

```
summary(model2factores)
```

```
##                                                Df Sum Sq Mean Sq F value
## Muestra2factores$Estudios                       4    172    42.9   13.09
## Muestra2factores$Sexo                           1     16    16.1    4.90
## Muestra2factores$Estudios:Muestra2factores$Sexo  4      7     1.8    0.53
## Residuals                                     390   1279     3.3
##                                                  Pr(>F)
## Muestra2factores$Estudios                      5.2e-10 ***
## Muestra2factores$Sexo                            0.027 *
## Muestra2factores$Estudios:Muestra2factores$Sexo   0.711
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model2factores2)
```

```
##
## Call:
## lm(formula = Muestra2factores$Nota ~ Muestra2factores$Estudios *
##     Muestra2factores$Sexo, data = Muestra2factores)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.600 -1.014 -0.055  1.120  4.500
##
## Coefficients:
##                                                Estimate Std. Error
```

```
## (Intercept)                                          4.500      0.286
## Muestra2factores$EstudiosES1                         0.755      0.405
## Muestra2factores$EstudiosES2                         1.275      0.405
## Muestra2factores$EstudiosEPS                         1.400      0.405
## Muestra2factores$EstudiosEU                          1.445      0.405
## Muestra2factores$SexoM                               0.170      0.405
## Muestra2factores$EstudiosES1:Muestra2factores$SexoM -0.170      0.573
## Muestra2factores$EstudiosES2:Muestra2factores$SexoM  0.430      0.573
## Muestra2factores$EstudiosEPS:Muestra2factores$SexoM  0.410      0.573
## Muestra2factores$EstudiosEU:Muestra2factores$SexoM   0.485      0.573
##                                                    t value Pr(>|t|)
## (Intercept)                                          15.72  < 2e-16 ***
## Muestra2factores$EstudiosES1                          1.86  0.06301 .
## Muestra2factores$EstudiosES2                          3.15  0.00177 **
## Muestra2factores$EstudiosEPS                          3.46  0.00061 ***
## Muestra2factores$EstudiosEU                           3.57  0.00040 ***
## Muestra2factores$SexoM                                0.42  0.67486
## Muestra2factores$EstudiosES1:Muestra2factores$SexoM  -0.30  0.76674
## Muestra2factores$EstudiosES2:Muestra2factores$SexoM   0.75  0.45320
## Muestra2factores$EstudiosEPS:Muestra2factores$SexoM   0.72  0.47447
## Muestra2factores$EstudiosEU:Muestra2factores$SexoM    0.85  0.39758
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 390 degrees of freedom
## Multiple R-squared:  0.132,  Adjusted R-squared:  0.112
## F-statistic:  6.6 on 9 and 390 DF,  p-value: 8.64e-09
```

Conclusión: Rechazamos la hipótesis nula de la igualdad de ambos efectos. Sin embargo, dado que no existe interacción entre los factores *Nivel de estudios máximo de los padres/tutores* y *Sexo del alumno*, para futuros estudios se aconsejaría estudiar el efecto que tiene ambos factores en la *Nota media del alumno* por separado.

3.4.1.4. Comparaciones múltiples

Si la interacción resulta ser significativa, entonces no tiene sentido aplicar los métodos de comparaciones múltiples a los niveles de cada uno de los factores ya que, las medias entre las diferencias de un factor pueden ser ocultadas por la interacción.

Una opción es considerar las medias de las ab celdillas para determinar entre cuáles hay diferencias significativas, es decir, considerar el modelo:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

En este análisis, las diferencias entre las celdas incluyen tanto los efectos principales como las interacciones.

Al haber sido rechazado la existencia de interacción entre los factores *Nivel de estudios máximo de los padres/tutores* y *Sexo del alumno*, a modo de ejemplo, explicaremos a continuación cómo debería realizarse las comparaciones múltiples en el caso de que sí existiera dicha interacción. El código de *R* empleado sería el siguiente:

```
#1.- Metodo de la minima diferencia significativa:
 LSD2factores <-LSD.test(Muestra2factores$NotaMedia,
                  Muestra2factores$Estudios:Muestra2factores$Sexo,
                  df,MSerror,group=FALSE)
print( LSD2factores$comparison)
options(digits=2)

 LSD2factores <-LSD.test(Muestra2factores$NotaMedia,
                  Muestra2factores$Estudios:Muestra2factores$Sexo,
                  df,MSerror, group=TRUE)
print( LSD2factores$group)


par(mfrow=c(1,2))
bar.group ( LSD2factores$group, col=colors()[84],ylim = c(0,8),
           frequency=1,las=2,density=20)
bar.err( LSD2factores$means, col=colors()[84], ylim = c(0,8),
        frequency=1,las=2,density=20)
```

NOTA: No obstante, en el caso de que la interacción no resulte ser significativa, podremos realizar las comparaciones múltiples a través de los métodos ya explicados en el apartado *Experimento completamente aleatorizado* para cada uno de los factores por separado. Veámoslo:

1.- Para el factor: Nivel de estudios máximo de los padres/tutores.

```
LSD2factores1 <-LSD.test(Muestra2factores$NotaMedia,
                  Muestra2factores$Estudios,
                  df,MSerror,group=FALSE)
print( LSD2factores1$comparison)
```

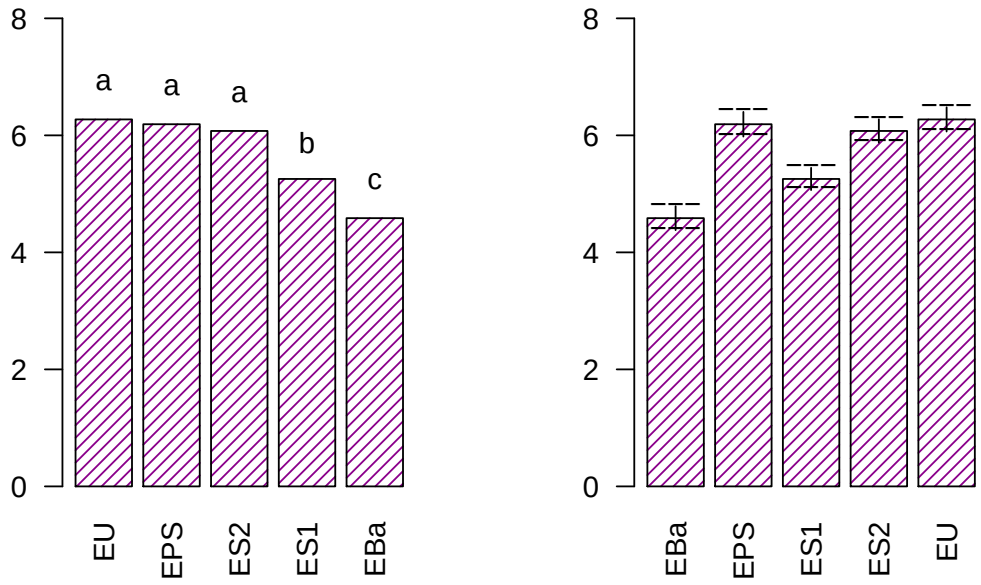
```
##           difference pvalue signif.   LCL   UCL
## EBa - EPS     -1.605 0.0000     *** -2.17 -1.04
## EBa - ES1     -0.670 0.0198       * -1.23 -0.11
## EBa - ES2     -1.490 0.0000     *** -2.05 -0.93
## EBa - EU      -1.688 0.0000     *** -2.25 -1.12
## EPS - ES1      0.935 0.0012      **  0.37  1.50
## EPS - ES2      0.115 0.6882          -0.45  0.68
## EPS - EU      -0.082 0.7734          -0.65  0.48
## ES1 - ES2     -0.820 0.0044      ** -1.38 -0.26
## ES1 - EU      -1.018 0.0004     *** -1.58 -0.45
## ES2 - EU      -0.197 0.4908          -0.76  0.37
```

```
options(digits=2)

LSD2factores1 <-LSD.test(Muestra2factores$NotaMedia,
```

```
                  Muestra2factores$Estudios,
                  df,MSerror, group=TRUE)

print( LSD2factores1$group)
```

```
##     Muestra2factores$NotaMedia groups
## EU                         6.3      a
## EPS                        6.2      a
## ES2                        6.1      a
## ES1                        5.3      b
## EBa                        4.6      c
```

```
par(mfrow=c(1,2))
bar.group ( LSD2factores1$group, col=colors()[84],ylim = c(0,8),
           frequency=1,las=2,density=20)
bar.err( LSD2factores1$means, col=colors()[84], ylim = c(0,8),
         frequency=1,las=2,density=20)
```



Conclusión: Podemos afirmar que a mayor nivel de estudios máximo de los padres existe mayor rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz.

2.- Sexo del alumno.

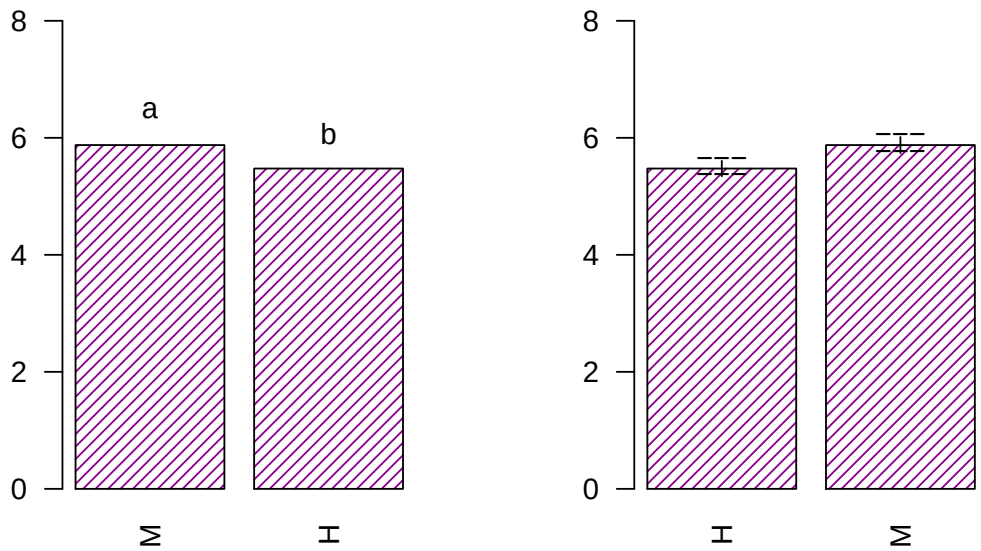
```
LSD2factores2 <-LSD.test(Muestra2factores$NotaMedia,
                  Muestra2factores$Sexo,
                  df,MSerror,group=FALSE)
print( LSD2factores2$comparison)
```

```
##       difference pvalue signif.   LCL    UCL
## H - M       -0.4  0.027       * -0.76 -0.045
```

```
options(digits=2)

LSD2factores2 <-LSD.test(Muestra2factores$NotaMedia,
                  Muestra2factores$Sexo,
                  df,MSerror, group=TRUE)

print( LSD2factores2$group)
```

```
##   Muestra2factores$NotaMedia groups
## M                        5.9      a
## H                        5.5      b
```

```
par(mfrow=c(1,2))
bar.group ( LSD2factores2$group, col=colors()[84],ylim = c(0,8),
           frequency=1,las=2,density=20)
bar.err( LSD2factores2$means, col=colors()[84], ylim = c(0,8),
        frequency=1,las=2,density=20)
```



Conclusión: Podemos afirmar que las alumnas que asisten a centros educativos del territorio andaluz, obtienen mejor rendimiento escolar respecto de los alumnos.

3.5. Experimentos multifactoriales

Como generalización al diseño de experimentos con 2 factores analizados anteriormente, podríamos analizar cualquier diseño multifactorial para L factores, con $L \geq 2$. A título informativo, se recogerá a continuación el caso factorial con 3 factores.

3.5.1. Experimento con tres factores completo

Consideremos ahora un experimento completo con tres factores: A, con a niveles, B, con b niveles, y C, con c niveles. Supongamos que se realizan n ($n \geq 2$) réplicas del mismo, teniéndose por tanto un total de $N = nabc$ observaciones.

Las observaciones pueden ser descritas según el siguiente modelo:

$$y_{ijkm} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkm} \quad con \quad \begin{array}{l} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots, c \\ m = 1, 2, \ldots, n \\ \varepsilon_{ijkm} \text{ i.i.d. } N(0, \sigma^2). \end{array}$$

donde:

$\boldsymbol{y_{ijkm}}$ representa la m-ésima observación en el i-ésimo nivel de A, en el j-ésimo nivel de B y en el k-ésimo nivel de C.

$\boldsymbol{\mu}$ representa la media global de todas las observaciones.

$\boldsymbol{\alpha_i}$ es el efecto principal del i-ésimo nivel del factor A.

$\boldsymbol{\beta_j}$ es el efecto principal del j-ésimo nivel del factor B.

$\boldsymbol{\gamma_k}$ es el efecto principal del k-ésimo nivel del factor C.

$\boldsymbol{\alpha\beta_{ij}}$ es el efecto de interacción del i-ésimo nivel del factor A y del j-ésimo nivel del factor B.

$\boldsymbol{\alpha\gamma_{ik}}$ es el efecto de interacción del i-ésimo nivel del factor A y del k-ésimo nivel del factor C.

$\boldsymbol{\beta\gamma_{jk}}$ es el efecto de interacción del j-ésimo nivel del factor B y del k-ésimo nivel del factor C.

$\boldsymbol{\alpha\beta\gamma_{ijk}}$ es el efecto de interacción del i-ésimo nivel del factor A, del j-ésimo nivel del factor B y del k-ésimo nivel del factor C.

$\boldsymbol{\varepsilon_{ijkm}}$ es el efecto aleatorio que recoge todas las restante posibles causas de variabilidad del experimento.

3.5.1.1. Estimación de parámetros

El análisis del modelo de efectos fijos es similar al del modelo con dos factores.

Los estimadores máximo verosímiles de los parámetros del modelo son:

– El E.M.V. de μ es:

$$\hat{\mu} = \bar{y}_{....},$$

– Los E.M.V. de los efectos principales son:

$$\hat{\alpha}_i = \bar{y}_{i...} - \bar{y}_{....}, \qquad \hat{\beta}_j = \bar{y}_{.j..} - \bar{y}_{....} \qquad \hat{\gamma}_k = \bar{y}_{..k.} - \bar{y}_{....}$$

– Los E.M.V. de las interacciones de segundo orden son:

$$\widehat{\alpha\beta}_{ij} = \bar{y}_{ij.} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....} \qquad \widehat{\alpha\gamma}_{ik} = \bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....} \qquad \widehat{\beta\gamma}_{jk} = \bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....}$$

– El E.M.V. de la interacción de tercer orden es:

$$\widehat{\alpha\beta\gamma}_{ijk} = \bar{y}_{ijk.} - \bar{y}_{ij..} - \bar{y}_{i.k.} - \bar{y}_{.jk.} + \bar{y}_{i...} + \bar{y}_{.j..} + \bar{y}_{..k.} - \bar{y}_{....}$$

donde:

$$\bar{y}_{....} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n} y_{ijkm}, \quad \bar{y}_{i...} = \frac{1}{bcn}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n} y_{ijkm}, \quad \bar{y}_{.j..} = \frac{1}{acn}\sum_{i=1}^{a}\sum_{k=1}^{c}\sum_{m=1}^{n} y_{ijkm},$$

$$\bar{y}_{..k.} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{m=1}^{n} y_{ijkm}, \quad \bar{y}_{ij..} = \frac{1}{cn}\sum_{k=1}^{c}\sum_{m=1}^{n} y_{ijkm}, \quad \bar{y}_{i.k.} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{m=1}^{n} y_{ijkm},$$

$$\bar{y}_{.jk.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{m=1}^{n} y_{ijkm}, \qquad \bar{y}_{ijk.} = \frac{1}{n}\sum_{m=1}^{n} y_{ijkm}$$

Un estimador insesgado de la varianza viene dado por:

$$\hat{\sigma}^2 = \frac{SC_\varepsilon}{abc(n-1)}, \qquad con \qquad SC_\varepsilon = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n} (y_{ijkm} - \bar{y}_{ijk.})^2$$

3.5.1.2. Análisis de la varianza: descomposición de la variabilidad total.

En este modelo la variabilidad total se descompone en

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}(y_{ijkm}-\bar{y}_{....})^2 = bcn\sum_{i=1}^{a}\hat{\alpha}_i^2 &+ acn\sum_{j=1}^{b}\hat{\beta}_j^2 + abn\sum_{k=1}^{c}\hat{\gamma}_k^2\\ + cn\sum_{i=1}^{a}\sum_{j=1}^{b}\widehat{\alpha\beta}_{ij}^2 &+ bn\sum_{i=1}^{a}\sum_{k=1}^{c}\widehat{\alpha\gamma}_{ik}^2 + an\sum_{j=1}^{b}\sum_{k=1}^{c}\widehat{\beta\gamma}_{jk}^2\\ + n\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\widehat{\alpha\beta\gamma}_{ijk}^2 &+ \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{m=1}^{n}(y_{ijkm}-\bar{y}_{ijk.})^2\end{aligned}$$

que simbólicamente podemos escribir:

$$SC_{tot} = SC_\alpha + SC_\beta + SC_\gamma + SC_{\alpha\beta} + SC_{\alpha\gamma} + SC_{\beta\gamma} + SC_{\alpha\beta\gamma} + SC_\varepsilon$$

Los grados de libertad de estas formas cuadráticas son:

$$\begin{aligned}abcn-1 = &\ (a-1)\ +\ (b-1)\ +\ (c-1)\ +\ (a-1)(b-1)\ +\ (a-1)(c-1)\\ &+\ (b-1)(c-1)\ +\ (a-1)(b-1)(c-1)\ +\ abc(n-1)\end{aligned}$$

por lo que, bajo la hipótesis de normalidad, se tiene que:

$$\begin{array}{llll}\dfrac{SC_\alpha}{\sigma^2}\sim\chi^2_{a-1}, & \dfrac{SC_\beta}{\sigma^2}\sim\chi^2_{b-1}, & \dfrac{SC_\gamma}{\sigma^2}\sim\chi^2_{c-1}, & \dfrac{SC_{\alpha\beta}}{\sigma^2}\sim\chi^2_{(a-1)(b-1)}\\ \dfrac{SC_{\alpha\gamma}}{\sigma^2}\sim\chi^2_{(a-1)(c-1)}, & \dfrac{SC_{\beta\gamma}}{\sigma^2}\sim\chi^2_{(b-1)(c-1)}, & \dfrac{SC_{\alpha\beta\gamma}}{\sigma^2}\sim\chi^2_{(a-1)(b-1)(c-1)}, & \dfrac{SC_\varepsilon}{\sigma^2}\sim\chi^2_{abc(n-1)}\end{array}$$

y además son independientes.

A partir de las sumas de cuadrados anteriores y sus grados de libertad, obtenemos los cuadrados medios:

$$\begin{array}{lll}CM_\alpha = \dfrac{SC_\alpha}{a-1}, & CM_\beta = \dfrac{SC_\beta}{b-1}, & CM_\gamma = \dfrac{SC_\gamma}{(c-1)},\\ CM_{\alpha\beta} = \dfrac{SC_{\alpha\beta}}{(a-1)(b-1)} & CM_{\alpha\gamma} = \dfrac{SC_{\alpha\gamma}}{(a-1)(c-1)}, & CM_{\beta\gamma} = \dfrac{SC_{\beta\gamma}}{(b-1)(c-1)},\\ CM_{\alpha\beta\gamma} = \dfrac{SC_{\alpha\beta\gamma}}{(a-1)(b-1)(c-1)}, & CM_\varepsilon = \dfrac{SC_\varepsilon}{ab(n-1)} & \end{array}$$

Los valores esperados de los cuadrados medios son:

$$E(CM_\alpha) = \sigma^2 + \frac{bcn}{a-1}\sum_{i=1}^{a}\alpha_i^2, \quad E(CM_\beta) = \sigma^2 + \frac{acn}{b-1}\sum_{j=1}^{b}\beta_j^2, \quad E(CM_\gamma) = \sigma^2 + \frac{abn}{c-1}\sum_{k=1}^{c}\gamma_k^2,$$

$$E(CM_{\alpha\beta}) = \sigma^2 + \frac{cn}{(a-1)(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\alpha\beta_{ij}^2, \quad E(CM_{\alpha\gamma}) = \sigma^2 + \frac{bn}{(a-1)(c-1)}\sum_{i=1}^{a}\sum_{k=1}^{c}\alpha\gamma_{ij}^2,$$

$$E(CM_{\beta\gamma}) = \sigma^2 + \frac{an}{(b-1)(c-1)}\sum_{j=1}^{b}\sum_{k=1}^{c}\beta\gamma_{ij}^2, \qquad E(CM_\varepsilon) = \sigma^2,$$

$$E(CM_{\alpha\beta\gamma}) = \sigma^2 + \frac{n}{(a-1)(b-1)(c-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\alpha\beta\gamma_{ijk}^2,$$

3.5.1.3. Contraste fundamental.

En este modelo, el objetivo del análisis es realizar los contrastes de hipótesis nula que, junto al estadístico de contraste, se muestran a continuación:

1. $H_{0\alpha} : \alpha_1 = \alpha_2 = \cdots = \alpha_a, \quad \forall i : F_\alpha = \frac{CM_\alpha}{CM_\varepsilon} \sim F_{(a-1),abc(n-1)}$
   Se rechaza $H_{0\alpha}$ al nivel $1-\alpha$ si: $F_\alpha \geq \mathcal{F}_{a-1,abc(n-1),1-\alpha}$
2. $H_{0\beta} : \beta_1 = \beta_2 = \cdots = \beta_b, \quad \forall j : F_\beta = \frac{CM_\beta}{CM_\varepsilon} \sim F_{c-1,abc(n-1)}$
   Se rechaza $H_{0\beta}$ al nivel $1-\alpha$ si: $F_\beta \geq \mathcal{F}_{b-1,abc(n-1),1-\alpha}$
3. $H_{0\gamma} : \gamma_1 = \gamma_2 = \cdots = \gamma_c, \quad \forall k : F_\gamma = \frac{CM_\gamma}{CM_\varepsilon} \sim F_{c-1,abc(n-1)}$
   Se rechaza $H_{0\gamma}$ al nivel $1-\alpha$ si: $F_\gamma \geq \mathcal{F}_{c-1,abc(n-1),1-\alpha}$
4. $H_{0\alpha\beta} : \alpha\beta_{ij} = 0, \quad \forall ij : F_{\alpha\beta} = \frac{CM_{\alpha\beta}}{CM_\varepsilon} \sim F_{(a-1)(b-1),abc(n-1)}$
   Se rechaza $H_{0\alpha\beta}$ al nivel $1-\alpha$ si: $F_\alpha \geq \mathcal{F}_{(a-1)(b-1),abc(n-1),1-\alpha}$
5. $H_{0\alpha\gamma} : \alpha\gamma_{ik} = 0, \quad \forall ik : F_{\alpha\gamma} = \frac{CM_{\alpha\gamma}}{CM_\varepsilon} \sim F_{(a-1)(c-1),abc(n-1)}$
   Se rechaza $H_{0\alpha\gamma}$ al nivel $1-\alpha$ si: $F_{\alpha\gamma} \geq \mathcal{F}_{(a-1)(c-1),ab(n-1),1-\alpha}$
6. $H_{0\beta\gamma} : \beta\gamma_{jk} = 0, \quad \forall jk : F_{\beta\gamma} = \frac{CM_{\beta\gamma}}{CM_\varepsilon} \sim F_{(b-1)(c-1),abc(n-1)}$
   Se rechaza $H_{0\beta\gamma}$ al nivel $1-\alpha$ si: $F_{\beta\gamma} \geq \mathcal{F}_{(b-1)(c-1),abc(n-1),1-\alpha}$
7. $H_{0\alpha\beta\gamma} : \alpha\beta\gamma_{ij} = 0, \quad \forall ijk : F_{\alpha\beta\gamma} = \frac{CM_{\alpha\beta\gamma}}{CM_\varepsilon} \sim F_{(a-1)(b-1)(c-1),abc(n-1)}$
   Se rechaza $H_{0\alpha\beta\gamma}$ al nivel $1-\alpha$ si: $F_{\alpha\beta\gamma} \geq \mathcal{F}_{(a-1)(b-1)(c-1),abc(n-1),1-\alpha}$

Los resultados obtenidos se resumen en la siguiente **tabla ANOVA**:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	F
A	SC_α	$a-1$	CM_α	$F_\alpha = CM_\alpha/CM_\varepsilon$
B	SC_β	$b-1$	CM_β	$F_\beta = CM_\beta/CM_\varepsilon$
C	SC_γ	$c-1$	CM_γ	$F_\gamma = CM_\gamma/CM_\varepsilon$
AB	$SC_{\alpha\beta}$	$(a-1)(b-1)$	$CM_{\alpha\beta}$	$F_{\alpha\beta} = CM_{\alpha\beta}/CM_\varepsilon$
AC	$SC_{\alpha\gamma}$	$(a-1)(c-1)$	$CM_{\alpha\gamma}$	$F_{\alpha\gamma} = CM_{\alpha\gamma}/CM_\varepsilon$
BC	$SC_{\beta\gamma}$	$(b-1)(c-1)$	$CM_{\beta\gamma}$	$F_{\beta\gamma} = CM_{\beta\gamma}/CM_\varepsilon$
ABC	$SC_{\alpha\beta\gamma}$	$(a-1)(b-1)(c-1)$	$CM_{\alpha\beta\gamma}$	$F_{\alpha\beta\gamma} = CM_{\alpha\beta\gamma}/CM_\varepsilon$
Error	SC_ε	$abc(n-1)$	CM_ε	
Total	SC_{tot}	$nabc-1$		

Cuadro 3.8: Tabla Anova para el modelo factorial con tres factores.

NOTA: La diagnosis y validación del modelo se realiza igual que en los modelos anteriores. Cabe añadir que el objetivo principal de este estudio es analizar las posibles interacciones entre los distintos factores.

3.5.2. Diseños factoriales con más de tres factores

Las ideas anteriores se extienden inmediatamente para modelos factoriales con cualquier número de factores. Para más de tres factores, las interacciones superiores a tres suelen suponerse nulas, lo que permite obtener una estimación del error experimental. Consideremos un diseño con cuatro factores a niveles $N1, N2, N3, N4$. Las $N1 \times N2 \times N3 \times N4$ observaciones permiten estimar:

- La media general μ.
- $\sum_{i=1}^{4}(N_i - 1) = \sum_{i=1}^{4} N_i - 4$
- $(N_i - 1)(N_j - 1)$ interacciones de segundo orden para cada una de las $\binom{4}{2}$ parejas de interacciones de segundo orden.
- $(N_i - 1)(N_j - 1)(N_k - 1)$ interacciones de tercer orden para cada una de las $\binom{4}{3}$ parejas de interacciones de tercer orden.
- Si suponemos que las iteracciones de cuarto orden son cero, tendremos: $(N_1 - 1)(N_2 - 1)(N_3 - 1)(N_4 - 1)$ grados de libertad para calcular los residuos y efectuar los contrastes.

Capítulo 4

Conclusiones globales.

Para finalizar, pasaremos a desarrollar las conclusiones globales de los diferentes resultados obtenidos a lo largo de la memoria. Es decir, compararemos los distintos diseños de manera conjunta, para poder obtener conclusiones sobre la idoneidad de los modelos utilizados. Además, en la mayoría de casos, al haber usado tamaños de muestras y diseños distintos para un mismo estudio conjunto: Saber cómo influye el *Nivel de estudios máximo de los padres/tutores* en el *rendimiento escolar de los alumnos del territorio andaluz*; podremos valorar si con menos recursos podemos obtener resultados análogos a los que obtendríamos en el caso de realizar el estudio con un tamaño muestral más grande.

Lo primero que veremos es ver si cada una de las muestras obtenidas son representativas, es decir, si las observaciones seleccionadas durante el proceso de la extracción de la muestra no se alejan demasiado, debida a la presencia de valores *outliers*, de la población real. Es por ello que pasaremos a recordar brevemente, un resumen de cada una de las muestras:

Para la variable Nota media.

```
#Nota media de la población:
summary(NotaMedia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       1       5       6       6       7      10     805
```

```
#Nota media del experimento completamente aleatorizado:
summary(Muestra1factor$NotaMedia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     4.4     5.6     5.5     6.8    10.0
```

```
#Nota media del experimento por bloques incompletos balanceados:
summary(MuestraBIB$NotaMedia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     4.4     5.4     5.4     6.6    10.0
```

```
#Nota media del experimento con dos factores:
summary(Muestra2factores$NotaMedia)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     4.5     5.6     5.7     6.8    10.0
```

Para la variable Nota en Matemáticas.

```
#Nota en Matemáticas de la población:

summary(NotaMates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       1       5       5       5       7      10     583
```

```
#Nota en Matemáticas del experimento en bloque aleatorizados completos:
summary(MuestraBAC$NotaMates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     5.0     5.0     5.5     6.2    10.0
```

Si nos fijamos en las medias, aparentemente, observamos que en todos los casos la media muestral se asemeja mucho a la media poblacional con un valor de 6 en el caso de la *Nota Media*, y de 5 en el caso de la *Nota en Matemáticas*. Por lo que, en un principio, no parece que hayamos obtenido ninguna muestra con demasiados valores atípicos. En relación a la *Nota en Matemáticas* hemos aplicado un Experimento en bloque aleatorizados completos, cuyos resultados se recogen en el apartado **3.2.1.**

En cuanto a la Nota media de las asignaturas: *Ciencias Naturales, Inglés, Lengua Castellana y Literatura, Matemáticas y Ciencias Sociales Geografía e Historia*; es de interés realizar comparaciones entre los distintos resultados obtenidos para cada uno de los diseños, puesto que se han usado metodologías y tamaños muestrales distintos. Así pues, pasemos a realizar comparativas de los diseños para los que hemos obtenido resultados de las mismas variables:

Experimento Completamente Aleatorizado, Experimento con dos factores y Experimento por Bloques Incompletos Balanceados.

Lo primero de todo, es conocer las diferencias principales que residen entre los tres experimentos que hemos realizado. Mientras que en los dos primeros casos, el *Experimento Completamente Aleatorizado* y el *Experimento con dos factores*, hemos partido de una muestra de tamaño 400, para el caso del *Experimento por Bloques Incompletos Balanceados* partimos de una muestra de tamaño de solo 30. Es decir, en los dos primeros casos hemos partido de un tamaño muestral muy superior.

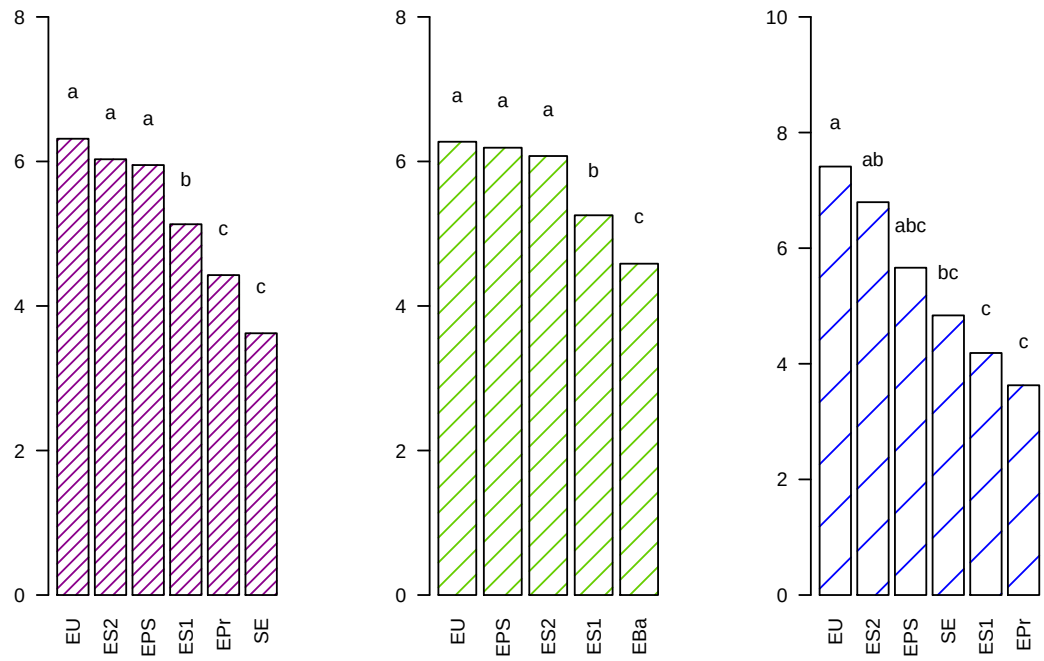
Por otra parte, para el *Experimento Completamente Aleatorizado* hemos empleado una muestra más representativa de la población ya que ésta era ponderada, mientras que en los casos del *Experimento con dos factores* y del *Experimento por Bloques Incompletos Balanceados* las muestras eran balanceadas. Para realizar un experimento real como en el primer caso, sería necesario conocer previamente cómo está distribuida la población, hecho que se consigue a través de la elaboración de un estudio o partiendo de alguno ya realizado por cuenta ajena.

Sin embargo, aun existiendo estas diferencias, en los tres diseños llegamos a la misma conclusión: Rechazamos la hipótesis nula de la igualdad de los efectos de los diferentes tratamientos del *Nivel de estudios máximo de los padres/tutores* en la variable respuesta *Nota Media* de los alumnos. Es decir, el *Nivel de estudios máximo de los padres/tutores* influye en el rendimiento escolar de los alumnos y alumnas que asisten a centros educativos del territorio andaluz. Luego en el caso de realizar la obtención de los datos de manera

real, hubieramos tenido que gastar una cantidad de recursos y tiempo mucho mayor en los dos primeros experimentos para llegar a la misma conclusión que en el tercero.

No obstante, si nuestro interés también reside en conocer además qué tratamientos del factor *Nivel de estudios máximo de los padres/tutores* obtienen mayor o peor rendimiento en la *Nota Media* del alumnos, sí que podemos encontrar ciertas diferencias. En el primer y segundo caso, afirmamos que a mayor nivel de estudios máximo de los padres existe mayor rendimiento escolar de los alumnos, sin embargo, para el segundo caso vimos que esto no tenía por qué ser exactamente así. Si representamos de manera conjunta la agrupación de los tratamientos del factor *Nivel de estudios máximo de los padres/tutores* para cada uno de los experimentos, podremos observar lo dicho anteriormente:

```
par(mfrow=c(1,3))
bar.group (LSD1factor$group, col=colors()[84],ylim = c(0,8),
           frequency=1,las=2,density=20)
bar.group (LSD2factores1$group, col="chartreuse3",ylim = c(0,8),
           frequency=1,las=2,density=10)
bar.group( LSDBib$groups,col="blue",density=4,las=2,ylim=c(0,10))
```



Cabe añadir que, aunque no coincidieron en la totalidad de los casos, para la mayoría de los tratamientos sí que obtuvimos los mismos resultados. Luego aquí entra en juego las prioridades del investigador, ya que con muchos menos recursos hemos obtenido unos resultados en el *Experimento por Bloques Incompletos Balanceados* casi tan acertados como en los casos del *Experimento Completamente Aleatorizado* y del *Experimento con dos factores*.

Por lo tanto, la conclusión final obtenida por lo dicho anteriormente es que, dependerá de los recursos disponibles o de los condicionantes ambientales, sociales y/o económicos, la realización de un experimento u otro.

Apéndice A

Obtención de las muestras.

En este apéndice explicaremos más detenidamente cómo obtuvimos las muestras en el Capítulo 2, así como los códigos usados durante todo el proceso.

A.1. Librerías utilizadas.

Lo primero que necesitaremos es instalar y cargar cada una de las siguientes librerías de *R*:

```
#install.packages("agricolae")
library(agricolae)
#install.packages("nortest")
library(nortest)
#install.packages("MASS")
library(MASS)
#install.packages("nlme")
library(nlme)
#install.packages("klaR")
library(klaR)
#install.packages("cluster")
library(cluster)
#install.packages("spDataLarge")
library(spDataLarge)
#install.packages("spdep")
library(spdep)
#install.packages("AlgDesign")
library(AlgDesign)
#install.packages("foreign)
library(foreign)
#install.packages("lawstat")
library(lawstat)
```

Una vez instaladas y cargadas cada una de las librerías, pasaremos a continuación a explicar y desarrollar la obtención de cada una de las muestras. A modo de ejemplo, iremos comentando cada uno de los pasos para el primer caso y realizaremos, de forma análoga, el resto de muestras.

A.2. Muestra ponderada: Nota media del alumno y Nivel de estudios máximo de los padres.

Nuestro objetivo es crear una muestra de tamaño 400 dónde aparezcan observaciones de los alumnos con su correspondiente nota media y el correspondiente nivel de estudios máximo de los padres o tutores. Lo primero que haremos es crear el *Dataframe* con los datos de la Nota media del alumno y Nivel de estudios máximo de los padres para cada una de las observaciones posibles:

```
Datos1factor<-cbind(Estudios,NotaMedia)
Datos1factor<- as.data.frame(Datos1factor)
Datos1factor$NotaMedia <-as.numeric(as.character(NotaMedia))
summary(Datos1factor)
```

```
##  Estudios    NotaMedia
##  EPr:306   Min.   : 1
##  EPS:182   1st Qu.: 5
##  ES1:969   Median : 6
##  ES2:559   Mean   : 6
##  EU :470   3rd Qu.: 7
##  NA :  2   Max.   :10
##  SE : 96   NA's   :805
```

Para la aleatoriedad de la muestra, añadiremos a nuestro *Dataframe* una tercera columna formada por valores aleatorios de una Uniforme con parámetros 0 y 1 (i.e U(0,1)), y lo ordenaremos según éstos.

Por otro lado, como expusimos al principio del capítulo 2, lo más óptimo es descartar aquellas variables que tengan valores perdidos. Por lo que también procederemos a la eliminación de estos.

Para el caso del Nivel de estudios máximo de los padres, al haber solo dos observaciones con valores perdidos, lo más rápido es localizar dichas observaciones y quitarlas directamente:

```
M1<-Datos1factor[-1635,]
M1<-M1[-2517,]
any(is.na(Datos1factor$Estudios))
```

```
## [1] FALSE
```

Como vemos, la columna correspondiente al Nivel de estudios máximo de los padres ya no contiene valores perdidos.

Para la aleatoriedad de la muestra:

```
set.seed(62401)
M1$Unif01<-runif(dim(M1)[1],0,1)
M2<-M1[order(M1$Unif01),]
```

Quitamos los valores perdidos de la variable Nota media:

```
colSums(is.na(M2))
```

```
##  Estudios NotaMedia    Unif01
##         0       804         0
```

```
M3<-na.omit(M2)
any(is.na(M3))
```

```
## [1] FALSE
```

Por último, arreglaremos nuestro *Dataframe*, quitando primero la columna con los valores de la U(0,1). A continuación, volvemos a poner en orden creciente los valores y quitamos el nivel sobrante.

```
M4<-M3[-3]
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[c(7,1,3,4,2,5,6)])
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[-7])
summary(M4)
```

```
##  Estudios    NotaMedia
##  SE : 42   Min.   : 1.0
##  EPr:164   1st Qu.: 4.6
##  ES1:669   Median : 5.6
##  ES2:412   Mean   : 5.7
##  EPS:145   3rd Qu.: 7.0
##  EU :346   Max.   :10.0
```

Una vez depurado nuestro *Dataframe*, pasaremos a crear nuestra muestra ponderada. Lo primero que haremos, será obtener las distintas ponderaciones para cada uno de los niveles de la variable Nivel de estudios máximo de los padres/tutores, y la cantidad de veces que deberá aparecer en la muestra.

```
n1<-summary(M4$Estudios)[[1]]
n2<-summary(M4$Estudios)[[2]]
n3<-summary(M4$Estudios)[[3]]
n4<-summary(M4$Estudios)[[4]]
n5<-summary(M4$Estudios)[[5]]
n6<-summary(M4$Estudios)[[6]]
ntotal=dim(M4)[1]
#Porcentaje de: Sin educación formal o inferior primaria:
p1<-n1/ntotal; (A<-400*p1)
```

```
## [1] 9.4
```

```
#Porcentaje de: Educación primaria:
p2<-n2/ntotal; (B<-400*p2)
```

```
## [1] 37
```

```
#Porcentaje de: Enseñanza secundaria de 1era etapa:
p3<-n3/ntotal; (C<-400*p3)
```

```
## [1] 151
```

```
#Porcentaje de: Enseñanza secundaria de 2a etapa:
p4<-n4/ntotal; (D<-400*p4)
```

```
## [1] 93
```

```
#Porcentaje de: Educación post secundaria pero no terciaria:
p5<-n5/ntotal; (E<-400*p5)
```

```
## [1] 33
```

```
#Porcentaje de: Estudios universitarios:
p6<-n6/ntotal; (F<-400*p6)
```

```
## [1] 78
```

Arreglamos la cantidad de apariciones de cada nivel para que nos dé números enteros:

```
A=9;B=37;C=150;D=93;E=32;F=79
#Comprobamos:
p1+p2+p3+p4+p5+p6;A+B+C+D+E+F
```

```
## [1] 1
```

```
## [1] 400
```

```
set.seed(62401)
```

A continuación, seleccionamos las observaciones dependiendo de la cantidad de apariciones correspondientes:

```
#Para: Sin educación formal o inferior primaria
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1])
Casos1<-M4[casospornivel,]
seleccionados1<-Casos1[1:A,]
#Para: educación primaria
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2])
Casos1<-M4[casospornivel,]
seleccionados2<-Casos1[1:B,]
#Para: Enseñanza secundaria de 1era etapa
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3])
Casos1<-M4[casospornivel,]
seleccionados3<-Casos1[1:C,]
#Para: Enseñanza secundaria de 2a etapa
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4])
Casos1<-M4[casospornivel,]
```

```
seleccionados4<-Casos1[1:D,]
#Para: educación post secundaria pero no terciaria
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5])
Casos1<-M4[casospornivel,]
seleccionados5<-Casos1[1:E,]
#Para: Estudios universitarios
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6])
Casos1<-M4[casospornivel,]
seleccionados6<-Casos1[1:F,]
```

Por último, creamos el *Dataframe* con los datos seleccionados anteriormente:

```
Muestra1factor <- rbind(seleccionados1,seleccionados2,seleccionados3,
                  seleccionados4,seleccionados5,seleccionados6)
Muestra1factor$NotaMedia<-as.numeric(as.character(Muestra1factor$NotaMedia))
```

A.3. Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
#Creación del Dataframe:
DatosBIB<-cbind(NotaMedia,Estudios,Provincia)
DatosBIB<- as.data.frame(DatosBIB)
DatosBIB$NotaMedia <-as.numeric(as.character(NotaMedia))
summary(DatosBIB)
```

```
##    NotaMedia   Estudios    Provincia
##  Min.   : 1    EPr:306   Alm    :273
##  1st Qu.: 5    EPS:182   Cad    :390
##  Median : 6    ES1:969   CoryJae:507
##  Mean   : 6    ES2:559   Gra    :215
##  3rd Qu.: 7    EU :470   Mal    :448
##  Max.   :10    NA :  2   SevyHue:751
##  NA's   :805   SE : 96
```

```
#Eliminación NA's de la variable Estudios:
M1<-DatosBIB[-1635,]
M1<-M1[-2517,]
any(is.na(DatosBIB$Estudios))
```

```
## [1] FALSE
```

```
#Cargamos la semilla de números aleatorios:
set.seed(123456)

#Creamos columna de U(0,1) y ordenamos:
M1$Unif01<-runif(dim(M1)[1],0,1)
```

```
M2<-M1[order(M1$Unif01),]

#Eliminamos el resto de NA's de la muestra:
colSums(is.na(M2))
```

```
## NotaMedia  Estudios Provincia    Unif01
##       804         0         0         0
```

```
M3<-na.omit(M2)
colSums(is.na(M3))
```

```
## NotaMedia  Estudios Provincia    Unif01
##         0         0         0         0
```

```
#Eliminamos la columna U(0,1):
M4<-M3[-4]

#Reordenamos los niveles y eliminamos el sobrante:
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[c(7,1,3,4,2,5,6)])
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[-7])
summary(M4)
```

```
##    NotaMedia    Estudios    Provincia
##  Min.   : 1.0   SE : 42   Alm    :179
##  1st Qu.: 4.6   EPr:164   Cad    :275
##  Median : 5.6   ES1:669   CoryJae:387
##  Mean   : 5.7   ES2:412   Gra    :165
##  3rd Qu.: 7.0   EPS:145   Mal    :298
##  Max.   :10.0   EU :346   SevyHue:474
```

Para este caso concreto, nos fijaremos en el esquema de un diseño balanceado incompleto acorde con nuestra muestra, de modo que:

```
trt <- c("SE","EPr","ES1","ES2","EPS","EU")

outdesign <- design.bib(trt,k=5,maxRep=400, seed=543, serie=2)
```

```
##
## Parameters BIB
## ==============
## Lambda     : 4
## treatmeans : 6
## Block size : 5
## Blocks     : 6
## Replication: 5
##
## Efficiency factor 0.96
##
## <<< Book >>>
```

```
book5 <- outdesign$book
outdesign$statistics
```

```
##        lambda treatmeans blockSize blocks r Efficiency
## values      4          6         5      6 5       0.96
```

```
print(outdesign$sketch)
```

```
##      [,1]  [,2]  [,3]  [,4]  [,5]
## [1,] "EPS" "SE"  "ES2" "EU"  "ES1"
## [2,] "SE"  "ES2" "EPr" "ES1" "EPS"
## [3,] "EPS" "ES1" "EPr" "SE"  "EU"
## [4,] "ES2" "EPr" "EU"  "EPS" "SE"
## [5,] "ES2" "ES1" "SE"  "EU"  "EPr"
## [6,] "EU"  "ES2" "EPr" "EPS" "ES1"
```

```
#Seleccionamos los casos para la creación de la muestra:

#Para: Almeria
#A) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[1])
Casos1<-M4[casospornivel,]
seleccionados11<-Casos1[1,]
#B) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[1])
Casos1<-M4[casospornivel,]
seleccionados12<-Casos1[1,]
#C) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[1])
Casos1<-M4[casospornivel,]
seleccionados13<-Casos1[1,]
#D) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[1])
Casos1<-M4[casospornivel,]
seleccionados14<-Casos1[1,]
#E) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[1])
Casos1<-M4[casospornivel,]
seleccionados15<-Casos1[1,]
#Para: Cadiz
#A) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[2])
Casos1<-M4[casospornivel,]
```

```
seleccionados21<-Casos1[1,]
#B) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[2])
Casos1<-M4[casospornivel,]
seleccionados22<-Casos1[1,]
#C) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[2])
Casos1<-M4[casospornivel,]
seleccionados23<-Casos1[1,]
#D) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[2])
Casos1<-M4[casospornivel,]
seleccionados24<-Casos1[1,]
#E) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[2])
Casos1<-M4[casospornivel,]
seleccionados25<-Casos1[1,]
#Para: Cordoba y Jaen
#A) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[3])
Casos1<-M4[casospornivel,]
seleccionados31<-Casos1[1,]
#B) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[3])
Casos1<-M4[casospornivel,]
seleccionados32<-Casos1[1,]
#C) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[3])
Casos1<-M4[casospornivel,]
seleccionados33<-Casos1[1,]
#D) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[3])
Casos1<-M4[casospornivel,]
seleccionados34<-Casos1[1,]
#E) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[3])
Casos1<-M4[casospornivel,]
seleccionados35<-Casos1[1,]
```

```
#Para: Granada
#A) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[4])
Casos1<-M4[casospornivel,]
seleccionados41<-Casos1[1,]
#B) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[4])
Casos1<-M4[casospornivel,]
seleccionados42<-Casos1[1,]
#C) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[4])
Casos1<-M4[casospornivel,]
seleccionados43<-Casos1[1,]
#D) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[4])
Casos1<-M4[casospornivel,]
seleccionados44<-Casos1[1,]
#E) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[4])
Casos1<-M4[casospornivel,]
seleccionados45<-Casos1[1,]
#Para: Sevilla y Huelva
#A) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[5])
Casos1<-M4[casospornivel,]
seleccionados51<-Casos1[1,]
#B) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[5])
Casos1<-M4[casospornivel,]
seleccionados52<-Casos1[1,]
#C) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[5])
Casos1<-M4[casospornivel,]
seleccionados53<-Casos1[1,]
#D) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[5])
Casos1<-M4[casospornivel,]
seleccionados54<-Casos1[1,]
```

```
#E) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[5])
Casos1<-M4[casospornivel,]
seleccionados55<-Casos1[1,]
#Para: Malaga
#A) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[6])
Casos1<-M4[casospornivel,]
seleccionados61<-Casos1[1,]
#B) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[6])
Casos1<-M4[casospornivel,]
seleccionados62<-Casos1[1,]
#C) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[6])
Casos1<-M4[casospornivel,]
seleccionados63<-Casos1[1,]
#D) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[6])
Casos1<-M4[casospornivel,]
seleccionados64<-Casos1[1,]
#E) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[6])
Casos1<-M4[casospornivel,]
seleccionados65<-Casos1[1,]

#Creamos la muestra balanceada MuestraBIB:
MuestraBIB <- rbind(seleccionados11,seleccionados12,seleccionados13,
                    seleccionados14,seleccionados15,seleccionados21,
                    seleccionados22,seleccionados23,seleccionados24,
                    seleccionados25,seleccionados31,seleccionados32,
                    seleccionados33,seleccionados34,seleccionados35,
                    seleccionados41,seleccionados42,seleccionados43,
                    seleccionados44,seleccionados45,seleccionados51,
                    seleccionados52,seleccionados53,seleccionados54,
                    seleccionados55,seleccionados61,seleccionados62,
                    seleccionados63,seleccionados64,seleccionados65)
MuestraBIB$NotaMedia<-as.numeric(as.character(MuestraBIB$NotaMedia))
```

A.4. Muestra balanceada: Nota media del alumno, Nivel de estudios máximo de los padres/tutores y Sexo del alumno.

```
#Creación del Dataframe:
Datos2factores<-cbind(NotaMedia,Estudios,Sexo)
Datos2factores<- as.data.frame(Datos2factores)
Datos2factores$NotaMedia <-as.numeric(as.character(NotaMedia))
summary(Datos2factores)
```

```
##    NotaMedia   Estudios  Sexo
##  Min.   : 1    EPr:306   H:1355
##  1st Qu.: 5    EPS:182   M:1229
##  Median : 6    ES1:969
##  Mean   : 6    ES2:559
##  3rd Qu.: 7    EU :470
##  Max.   :10    NA :  2
##  NA's   :805   SE : 96
```

```
#Eliminación NA's de la variable Estudios:
M1<-Datos2factores[-1635,]
M1<-M1[-2517,]
any(is.na(Datos2factores$Estudios))
```

```
## [1] FALSE
```

```
#Cargamos la semilla de números aleatorios
set.seed(62401)

#Creamos columna de U(0,1) y ordenamos
M1$Unif01<-runif(dim(M1)[1],0,1)
M2<-M1[order(M1$Unif01),]

#Eliminamos el resto de NA's de la muestra:
colSums(is.na(M2))
```

```
## NotaMedia  Estudios      Sexo    Unif01
##       804         0         0         0
```

```
M3<-na.omit(M2)
any(is.na(M3))
```

```
## [1] FALSE
```

```
#Eliminamos la columna U(0,1):
M4<-M3[-4]

#Reordenamos los niveles y eliminamos el sobrante:
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[c(7,1,3,4,2,5,6)])
```

```
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[-7])
summary(M4)
```

```
##    NotaMedia    Estudios  Sexo
##  Min.   : 1.0   SE : 42   H:904
##  1st Qu.: 4.6   EPr:164   M:874
##  Median : 5.6   ES1:669
##  Mean   : 5.7   ES2:412
##  3rd Qu.: 7.0   EPS:145
##  Max.   :10.0   EU :346
```

Para el caso de esta muestra, por falta de observaciones para alguno de los cruces, tendremos que anexionar alguno de los tratamientos de la variable *Nivel de estudios máximo de los padres/tutores* para crear así el nivel: *"Estudios Básicos"*:

```
levels(M4$Estudios)<-c("EBa","EBa","ES1","ES2","EPS","EU")
```

```
#Seleccionamos los casos para la creación de la muestra:

#Para: Educación Básica
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                      M4$Sexo==levels(M4$Sexo)[1])
Casos1<-M4[casospornivel,]
seleccionados11<-Casos1[1:40,]

casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                      M4$Sexo==levels(M4$Sexo)[2])
Casos1<-M4[casospornivel,]
seleccionados12<-Casos1[1:40,]
#Para: Enseñanza secundaria de 1era etapa
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                      M4$Sexo==levels(M4$Sexo)[1])
Casos1<-M4[casospornivel,]
seleccionados21<-Casos1[1:40,]

casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                      M4$Sexo==levels(M4$Sexo)[2])
Casos1<-M4[casospornivel,]
seleccionados22<-Casos1[1:40,]
#Para: Enseñanza secundaria de 2a etapa
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                      M4$Sexo==levels(M4$Sexo)[1])
Casos1<-M4[casospornivel,]
seleccionados31<-Casos1[1:40,]

casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                      M4$Sexo==levels(M4$Sexo)[2])
Casos1<-M4[casospornivel,]
```

```
seleccionados32<-Casos1[1:40,]
#Para: Educacion post secundaria pero no terciaria
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Sexo==levels(M4$Sexo)[1])
Casos1<-M4[casospornivel,]
seleccionados41<-Casos1[1:40,]

casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Sexo==levels(M4$Sexo)[2])
Casos1<-M4[casospornivel,]
seleccionados42<-Casos1[1:40,]
#Para: Estudios universitarios
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Sexo==levels(M4$Sexo)[1])
Casos1<-M4[casospornivel,]
seleccionados51<-Casos1[1:40,]
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Sexo==levels(M4$Sexo)[2])
Casos1<-M4[casospornivel,]
seleccionados52<-Casos1[1:40,]

#Creamos la muestra balanceada Muestra2factores:
Muestra2factores <- rbind(seleccionados11,seleccionados12,seleccionados21,
                          seleccionados22,seleccionados31,seleccionados32,
                          seleccionados41,seleccionados42,seleccionados51,
                          seleccionados52)
Muestra2factores$NotaMedia<-as.numeric(as.character(
                          Muestra2factores$NotaMedia))
```

A.5. Muestra balanceada: Nota en matemáticas del alumno, Nivel de estudios máximo de los padres/tutores y Lugar de residencia del alumno.

```
#Creación del Dataframe:
DatosBAC<-cbind(NotaMates,Estudios,Provincia)
DatosBAC<- as.data.frame(DatosBAC)
DatosBAC$NotaMates <-as.numeric(as.character(NotaMates))
summary(DatosBAC)
```

```
##    NotaMates    Estudios     Provincia
##  Min.   : 1    EPr:306   Alm    :273
##  1st Qu.: 5    EPS:182   Cad    :390
##  Median : 5    ES1:969   CoryJae:507
##  Mean   : 5    ES2:559   Gra    :215
##  3rd Qu.: 7    EU :470   Mal    :448
##  Max.   :10    NA :  2   SevyHue:751
##  NA's   :583   SE : 96
```

```
#Eliminación NA's de la variable Estudios:
M1<-DatosBAC[-1635,]
M1<-M1[-2517,]

#Cargamos la semilla de números aleatorios:
set.seed(2401)

#Creamos columna de U(0,1) y ordenamos:
M1$Unif01<-runif(dim(M1)[1],0,1)
M2<-M1[order(M1$Unif01),]

#Eliminamos el resto de NA's de la muestra:
colSums(is.na(M2))
```

```
## NotaMates  Estudios Provincia    Unif01
##       583         0         0         0
```

```
Dim(M2)[1]
```

```
## [1] 2582
```

```
M3<-na.omit(M2)
colSums(is.na(M3))
```

```
## NotaMates  Estudios Provincia    Unif01
##         0         0         0         0
```

```
#Eliminamos la columna U(0,1):
M4<-M3[-4]

#Reordenamos los niveles y eliminamos el sobrante:
```

```
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[c(7,1,3,4,2,5,6)])
M4$Estudios <- factor(M4$Estudios,
                      levels = levels(M4$Estudios)[-7])
summary(M4)
```

```
##    NotaMates    Estudios    Provincia
##  Min.   : 1.0   SE : 49   Alm    :204
##  1st Qu.: 5.0   EPr:210   Cad    :306
##  Median : 5.0   ES1:759   CoryJae:424
##  Mean   : 5.3   ES2:454   Gra    :182
##  3rd Qu.: 7.0   EPS:155   Mal    :339
##  Max.   :10.0   EU :372   SevyHue:544
```

```
#Seleccionamos los casos para la creación de la muestra:

#Para: Almeria
#B) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                      M4$Provincia==levels(M4$Provincia)[1])
casos1<-M4[casospornivel,]
seleccionados11<-Casos1[1,]

#B) Epr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                      M4$Provincia==levels(M4$Provincia)[1])
casos1<-M4[casospornivel,]
seleccionados12<-Casos1[1,]
#E) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                      M4$Provincia==levels(M4$Provincia)[1])
casos1<-M4[casospornivel,]
seleccionados13<-Casos1[1,]
#C) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                      M4$Provincia==levels(M4$Provincia)[1])
casos1<-M4[casospornivel,]
seleccionados14<-Casos1[1,]
#A) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                      M4$Provincia==levels(M4$Provincia)[1])
casos1<-M4[casospornivel,]
seleccionados15<-Casos1[1,]
#D) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                      M4$Provincia==levels(M4$Provincia)[1])
casos1<-M4[casospornivel,]
seleccionados16<-Casos1[1,]
```

```
#Para: Cadiz
#A) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[2])
casos1<-M4[casospornivel,]
seleccionados21<-Casos1[1,]
#C) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[2])
casos1<-M4[casospornivel,]
seleccionados22<-Casos1[1,]
#D) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[2])
casos1<-M4[casospornivel,]
seleccionados23<-Casos1[1,]
#B) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[2])
casos1<-M4[casospornivel,]
seleccionados24<-Casos1[1,]
#E) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[2])
casos1<-M4[casospornivel,]
seleccionados25<-Casos1[1,]
#E) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[2])
casos1<-M4[casospornivel,]
seleccionados26<-Casos1[1,]

#Para: Cordoba y Jaen
#D) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[3])
casos1<-M4[casospornivel,]
seleccionados31<-Casos1[1,]
#C) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[3])
casos1<-M4[casospornivel,]
seleccionados32<-Casos1[1,]
#B) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[3])
casos1<-M4[casospornivel,]
```

```
seleccionados33<-Casos1[1,]
#B) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[3])
casos1<-M4[casospornivel,]
seleccionados34<-Casos1[1,]
#A) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[3])
casos1<-M4[casospornivel,]
seleccionados35<-Casos1[1,]
#E) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[3])
casos1<-M4[casospornivel,]
seleccionados36<-Casos1[1,]

#Para: Granada
#E) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[4])
casos1<-M4[casospornivel,]
seleccionados41<-Casos1[1,]
#B) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[4])
casos1<-M4[casospornivel,]
seleccionados42<-Casos1[1,]
#A) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[4])
casos1<-M4[casospornivel,]
seleccionados43<-Casos1[1,]
#A) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[4])
casos1<-M4[casospornivel,]
seleccionados44<-Casos1[1,]
#D) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[4])
casos1<-M4[casospornivel,]
seleccionados45<-Casos1[1,]
#C) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[4])
casos1<-M4[casospornivel,]
```

```
seleccionados46<-Casos1[1,]

#Para: Sevilla y Huelva
#C) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[5])
casos1<-M4[casospornivel,]
seleccionados51<-Casos1[1,]
#E) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[5])
casos1<-M4[casospornivel,]
seleccionados52<-Casos1[1,]
#B) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
                       M4$Provincia==levels(M4$Provincia)[5])
casos1<-M4[casospornivel,]
seleccionados53<-Casos1[1,]
#A) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                       M4$Provincia==levels(M4$Provincia)[5])
casos1<-M4[casospornivel,]
seleccionados54<-Casos1[1,]
#D) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                       M4$Provincia==levels(M4$Provincia)[5])
casos1<-M4[casospornivel,]
seleccionados55<-Casos1[1,]
#D) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                       M4$Provincia==levels(M4$Provincia)[5])
casos1<-M4[casospornivel,]
seleccionados56<-Casos1[1,]

#Para: Malaga
#C) SE
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[1]&
                       M4$Provincia==levels(M4$Provincia)[6])
casos1<-M4[casospornivel,]
seleccionados61<-Casos1[1,]
#C) EPr
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[2]&
                       M4$Provincia==levels(M4$Provincia)[6])
casos1<-M4[casospornivel,]
seleccionados62<-Casos1[1,]
#E) ES1
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[3]&
```

```
                        M4$Provincia==levels(M4$Provincia)[6])
casos1<-M4[casospornivel,]
seleccionados63<-Casos1[1,]
#B) ES2
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[4]&
                        M4$Provincia==levels(M4$Provincia)[6])
casos1<-M4[casospornivel,]
seleccionados64<-Casos1[1,]
#D) EPS
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[5]&
                        M4$Provincia==levels(M4$Provincia)[6])
casos1<-M4[casospornivel,]
seleccionados65<-Casos1[1,]
#A) EU
casospornivel<-which(M4$Estudios==levels(M4$Estudios)[6]&
                        M4$Provincia==levels(M4$Provincia)[6])
casos1<-M4[casospornivel,]
seleccionados66<-Casos1[1,]
```

```
#Creamos la muestra balanceada MuestraBAC:
MuestraBAC <- rbind(seleccionados11,seleccionados12,seleccionados13,
                    seleccionados14,seleccionados15,seleccionados16,
                    seleccionados21,seleccionados22,seleccionados23,
                    seleccionados24,seleccionados25,seleccionados26,
                    seleccionados31,seleccionados32,seleccionados33,
                    seleccionados34,seleccionados35,seleccionados36,
                    seleccionados41,seleccionados42,seleccionados43,
                    seleccionados44,seleccionados45,seleccionados46,
                    seleccionados51,seleccionados52,seleccionados53,
                    seleccionados54,seleccionados55,seleccionados56,
                    seleccionados61,seleccionados62,seleccionados63,
                    seleccionados64,seleccionados65,seleccionados66)

MuestraBAC$NotaMates<-as.numeric(as.character(MuestraBAC$NotaMates))
```

Bibliografía

[1] Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J. and Chang, W. 2018. *Rmarkdown: Dynamic documents for r.*

[2] Diazaraque, J. *Diseño de experimentos.*

[3] García, A., J. & Lara Porras 1998. *Diseño estadístico de experimentos. análisis de la varianza.* Grupo Editorial Universitario.

[4] García, I., J.A. & Barranco *Ampliación de inferencia estadística.*

[5] García, J.M., J.A. & Muñoz *Modelos lineales y diseño de experimentos.*

[6] Hochberg, A., Y. & TAMHANE 1987. *Multiple comparison procedures.* Wiley.

[7] Jiménez, M.D. *Diseño de experimentos.*

[8] Lara Porras, A. 2000. *Diseño estadístico de experimentos, análisis de la varianza y temas relacionados: Tratamiento informático mediante spss.* Proyecto Sur de Ediciones.

[9] Luque-Calvo, P.L. 2017. *Escribir un trabajo fin de estudios con r markdown.* Disponible en http://destio.us.es/calvo.

[10] Mendiburu, F. de 2017. *Agricolae tutorial.*

[11] Montgomery, D. 1991. *Diseño y análisis de experimentos.* Grupo Editorial Iberoamericano.

[12] R Core Team 2016. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

[13] RStudio Team 2015. *RStudio: Integrated development environment for r.* RStudio, Inc.

[14] Seber, G. 1977. *Linear regression analysis.* Wiley.

[15] Tamhane, A. 2009. *Statistical analysis of designed experiments: Theory and applications.* Wiley.

[16] Tanco, E.&.P., M. & Viles *Diferentes enfoques del diseño de experimentos.* Disponible en http://www.um.edu.uy/_upload/_descarga/web_descarga_178_DiferentesenfoquesDiseoexperimentosDOE.-Tanco_Viles_Pozueta.pdf.

[17] Weber, J., D.C. & Skillings 2000. *A first course in the design of experiments.* CRC Press.