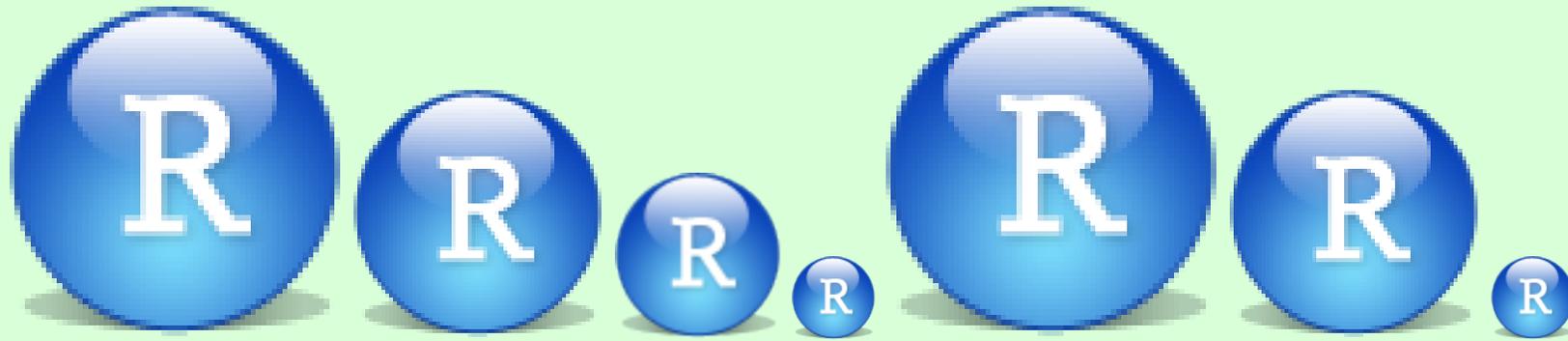


# Modelación Aplicada a las Ciencias Animales:

Diseño experimental, con implementación del programa R-project



**Mario Fernando Cerón-Muñoz**  
**Luis Fernando Galeano Vasco**  
**Luis Fernando Restrepo Betancur**

**Ciencias Animales**



Cerón-Muñoz, MF; Galeano Vasco, LF; Restrepo Betancur, LF

Modelación aplicada a las ciencias animales: diseño experimental,  
con implementación del programa R-project / Mario Fernando Cerón  
Muñoz, Luis Fernando Galeano, Luis Fernando Restrepo Bentancur.  
- Medellín, Colombia: Editorial Biogénesis, [2013]

165 p.: il; 27 cm

Incluye referencias bibliográficas e índice

ISBN: 978-958-8790-69-5

1. Modelación animal. 2. Diseño experimental

005

©Cerón-Muñoz MF; Galeano Vasco LF y Restrepo Betancur LF

Primera Edición: Junio de 2013

ISBN: 978-958-8790-69-5

Autores:

Mario Fernando Cerón-Muñoz, Zootecnista, Dr en Zootecnia

Luis Fernando Galeano Vasco, Zootecnista, cDr en Ciencias Animales

Luis Fernando Restrepo Betancur, Estadístico, esp en Estadística.

Facultad de Ciencias Agrarias

Universidad de Antioquia

Corrección de textos:

Diego García Sierra

Ligia Cecilia Alzate Suárez

Evaluación de contenido:

Diego Fernando Lemus Polanía

Ingeniero Industrial, MS en Estadística

Luis Gabriel González Herrera

Médico Veterinario Zootecnista, Dr en Genética y Mejoramiento Animal

Elkin Mauricio Arboleda Zapata

Zootecnista, MS en Ciencias Animales

Diseño y diagramación en *l<sup>A</sup>T<sub>E</sub>X*:

Mario Fernando Cerón-Muñoz

Asesoría en Diseño:

Sandra María Arango Mejía

Todos los derechos reservados, puede ser reproducido en todo o en parte y por cualquier medio, citando la fuente. Esta publicación contó con el apoyo de la Universidad de Antioquia y del CODI sostenibilidad 2011-2012 del grupo GAMMA.



©Fondo Editorial Biogénesis

Universidad de Antioquia

Facultad de Ciencias Agrarias

Ciudadela de Robledo, Carrera 75 No 65-87

Medellín, Colombia

## Prólogo

Este libro de diseño experimental incorpora ejemplos de las ciencias animales y su desarrollo en R-project. Los diseños que incluimos en el libro son: completamente aleatorizado, bloques aleatorizados y cuadrado latino. También incluimos los temas de arreglos factoriales, comparación de medias y transformación de datos. Para cada capítulo utilizamos paquetes del entorno R, siendo los más utilizados: *psych*, *agricolae*, *lattice*, *multcomp*, *lsmeans*, *gmodels*, *car*, *outliers* y *lmSupport*.

Esperamos que este libro sea de satisfactoria ayuda en el aprendizaje del diseño experimental y que sea utilizado en el desarrollo profesional de nuestros egresados del sector pecuario.

Estaremos atentos a cualquier sugerencia para mejorar esta publicación en próximas ediciones.

Este libro fue realizado en el sistema de composición de textos *L<sup>A</sup>T<sub>E</sub>X*, una gran herramienta para la diagramación de libros.

Mario Fernando Cerón-Muñoz  
Zootecnista, Dr.

# Contenido

<b>1. Generalidades</b>	<b>1</b>
1.1. Algunas definiciones importantes . . . . .	1
1.2. Diseños experimentales . . . . .	2
1.3. Medidas de posición y medidas de dispersión . . . . .	4
1.3.1. Medidas de tendencia central . . . . .	4
1.3.2. Medidas de dispersión o de variación . . . . .	7
1.3.3. Implementación en R-project . . . . .	9
1.4. Pruebas de significancia . . . . .	12
1.4.1. Prueba para dos muestras independientes con varianzas homogéneas .	13
1.4.2. Prueba para dos muestras independientes con varianzas heterogéneas	15
1.4.3. Prueba para muestras pareadas . . . . .	18
1.4.4. Prueba para más de dos muestras . . . . .	20
<b>2. Diseño completamente aleatorizado</b>	<b>21</b>
2.1. Generalidades . . . . .	21
2.2. Análisis de varianza . . . . .	23
2.2.1. Modelo de efecto fijo . . . . .	23
2.2.2. Modelo con efectos aleatorios . . . . .	24
2.3. Implementación en R-project . . . . .	25
2.3.1. Proceso de aleatorización . . . . .	25
2.3.2. Metodología para el desarrollo, análisis y diagnóstico del modelo . .	27
2.3.3. Análisis exploratorio de la hoja de datos . . . . .	28
2.3.4. Contraste de interés (efecto de los tratamientos) . . . . .	30
2.3.5. Validación de los supuestos del modelo . . . . .	32
<b>3. Pruebas de comparación de efectos promedios</b>	<b>41</b>
3.1. Generalidades . . . . .	41
3.2. Prueba de comparación de Tukey . . . . .	41
3.3. Mínima diferencia significativa: DMS o LSD . . . . .	48
3.4. Prueba de Duncan . . . . .	51
3.5. Prueba de Scheffé de comparaciones múltiples . . . . .	54
3.6. Contrastes ortogonales . . . . .	56

<b>4. Diseño en bloques aleatorizados</b>	<b>62</b>
4.1. Generalidades . . . . .	62
4.2. Supuestos estadísticos . . . . .	65
4.3. Proceso de aleatorización . . . . .	65
4.4. Modelo . . . . .	66
4.5. Análisis exploratorio de la hoja de datos . . . . .	67
4.6. Análisis de varianza . . . . .	68
4.7. Prueba de normalidad . . . . .	73
4.8. Homogeneidad de varianza de los errores . . . . .	75
4.9. Aditividad de efectos . . . . .	77
4.10. Parcelas perdidas . . . . .	79
4.11. Bloques incompletos balanceados . . . . .	81
<b>5. Diseño en cuadrado latino</b>	<b>84</b>
5.1. Generalidades . . . . .	84
5.2. Supuestos estadísticos . . . . .	85
5.3. Modelos de estructura experimental . . . . .	86
5.4. Proceso de aleatorización . . . . .	87
5.5. Modelo . . . . .	88
5.6. Análisis exploratorio de la hoja de datos . . . . .	89
5.7. Análisis de varianza . . . . .	91
5.8. Prueba de normalidad . . . . .	95
5.9. Efectos de arrastre de los tratamientos previos . . . . .	97
<b>6. Arreglos factoriales</b>	<b>110</b>
6.1. Generalidades . . . . .	110
6.2. Algunos tipos de arreglos factoriales . . . . .	111
6.3. Tipos de notación . . . . .	114
6.4. Modelos de estructura experimental . . . . .	114
6.5. Hipótesis en un arreglo factorial . . . . .	116
6.6. Proceso de aleatorización . . . . .	117
6.7. Análisis exploratorio de la hoja de datos . . . . .	119
6.8. Análisis de varianza . . . . .	122
6.9. Prueba de normalidad . . . . .	126
6.10. Pruebas de homogeneidad de varianza de los errores . . . . .	128
6.11. Interacciones significativas . . . . .	131
6.12. Arreglos factoriales con niveles vacíos . . . . .	133
<b>7. Transformación de datos</b>	<b>140</b>
7.1. Estandarización zeta ( $z$ ) y transformaciones lineal y logarítmica . . . . .	142
7.2. Transformación $y^2$ y $y^3$ . . . . .	147
7.3. Transformación <i>Arcoseno</i> $\sqrt[2]{y}$ y <i>PROBIT</i> . . . . .	147
7.4. Transformación $\sqrt[2]{y}$ . . . . .	152
7.5. Transformación <i>Box-Cox</i> . . . . .	156

## Capítulo 1

# Generalidades

### 1.1. Algunas definiciones importantes

- ♣ Aleatorización: Consiste en asignar los tratamientos a las distintas unidades experimentales, de manera que se asegure la imparcialidad. Existen diseños completamente aleatorizados, donde cualquier tratamiento puede ser asignado a cualquier unidad experimental, y diseños de aleatorización condicionada, donde la aleatorización se efectúa dentro de cada bloque de manera independiente.
- ♣ Ancova: Análisis de varianza con una sola covariable.
- ♣ Anova: Se conoce como análisis de la varianza, el cual se realiza con base en una variable respuesta y tiene como objetivo general establecer si existe diferencia estadística en el efecto promedio de los tratamientos. Este tipo de análisis está asociado con diferentes fuentes de variabilidad de acuerdo con el modelo experimental propuesto.
- ♣ Covariable: Es una variable que afecta a la variable respuesta, la cual se incluye en el modelo para ajustar los efectos promedios de los efectos.
- ♣ Error experimental: Se define como la cuantificación estadística de todos los factores no controlados que repercuten sobre las variables respuesta. El error experimental debe ser de naturaleza aleatoria, distribuirse en forma normal y de manera independiente entre unidades experimentales.
- ♣ Grados de libertad: Es un estimador del número de observaciones o categorías independientes en los problemas estadísticos que tienen restricciones en sus cálculos. En general, está dado por el número total de observaciones que se utiliza para estimar un parámetro, menos el número de estadísticos utilizados para tal fin.
- ♣ Mancova: Análisis multivariado de la varianza con una o más covariables, y donde se evalúa simultáneamente más de una variable respuesta.
- ♣ Manova: Análisis multivariado de la varianza; se efectúa teniendo en cuenta más de una variable respuesta de manera simultánea.
- ♣ Nivel de significancia (error tipo I): Se define como la probabilidad de tomar la decisión de rechazar la hipótesis nula cuando ésta es verdadera. Se denota con el símbolo  $\alpha$ .

- ♣ **Potencia de la prueba:** Es la probabilidad de que una prueba rechace la hipótesis nula, si es cierta la hipótesis alternativa.
- ♣ **Tratamiento testigo:** Tratamiento que usualmente se aplica y que sirve de base para comparar el efecto de los nuevos tratamientos propuestos con respecto a él.
- ♣ **Unidad experimental:** Material vivo o inerte al cual se le aplica un tratamiento para medir cada una de las respuestas que emite. La unidad experimental puede estar conformada por un animal o un conjunto de ellos, lo cual depende de los propósitos de la investigación.
- ♣ **Variable control:** Es cada uno de los factores que se pueden manipular experimentalmente y que repercuten sobre las variables respuesta.
- ♣ **Variable respuesta:** Variable que contiene los resultados encontrados al aplicar un tratamiento sobre la unidad experimental; se conoce también como variable dependiente. Puede tener naturaleza discreta, continua o cualitativa.

## 1.2. Diseños experimentales

A continuación se presenta la definición de algunos diseños experimentales que han sido aplicados en las ciencias animales.

- ♣ **Nulifactoriales:** Aquellos experimentos en los que no se tiene ningún tratamiento, simplemente se elige un material experimental con el fin de evaluar su dinámica. Se aplica para definir tamaños y formas de parcelas experimentales en campo; por ejemplo, permite determinar las distancias óptimas de siembra o el número ideal de animales que pueden cohabitar en una jaula.
- ♣ **Unifactoriales:** Consiste en elegir un único factor a distintas dosificaciones o niveles.
- ♣ **Bifactoriales:** Se evalúan dos factores de manera simultánea, con distintas concentraciones o niveles por factor. El interés central de este tipo de experimentos radica en poder medir la interacción de los factores, mediante un diseño de estructura preestablecido.
- ♣ **Multifactoriales:** Como su nombre lo indica, contienen más de dos factores controlados a distintas dosis o niveles por factor, donde se combinan las dosis de los distintos componentes controlados para generar los tratamientos. En este tipo de experimento se estudian las interacciones entre factores, como objetivo central.
- ♣ **Unidimensionales:** Consiste en evaluar una sola variable respuesta de tipo discreto, continuo o cualitativo, de acuerdo con un modelo experimental.
- ♣ **Multidimensionales:** Son aquellos experimentos donde se tienen dos o más variables respuesta o dependientes, las cuales se pueden evaluar conjuntamente mediante una combinación lineal, por medio del análisis multivariado de la varianza *MANOVA*, articulado a un diseño de clasificación experimental.

- ♣ Aleatorizados: Dícese de aquellos experimentos en los cuales se asignan los distintos tratamientos a las unidades experimentales por medio de una función probabilística, la cual conduce a resultados aleatorios en la repartición.
- ♣ No aleatorios: En estos experimentos se asignan los tratamientos de manera subjetiva al material experimental.
- ♣ Pre-experimentales: Son experimentos donde no existe aleatorización de las unidades experimentales y generalmente se realizan en ambientes no controlados (naturales). Se aplican principalmente en ensayos donde se busca medir la reacción de los individuos a determinado estímulo. Se caracterizan por un bajo nivel de control y validez. El inconveniente de estos diseños es que el investigador no puede saber con certeza, después de llevar a cabo su investigación, si los efectos producidos en la variable dependiente se deben exclusivamente a la variable independiente o al tratamiento.
- ♣ Experimentos de medidas repetidas: Consisten en analizar cada una de las variables respuesta varias veces en el tiempo, con el fin de evaluar la curva de crecimiento o la dinámica de cambio en el desempeño del animal.
- ♣ Experimentos compensados: Son aquellos en los que se aplican todos los tratamientos a todos los grupos de animales en distintos periodos de tiempo.
- ♣ Experimentos determinísticos: Son aquellos que utiliza el investigador para ratificar resultados de experimentos anteriores. Se asume que se conoce con antelación el resultado de las variables observables.
- ♣ Experimentos de optimización: Consisten en determinar la región óptima en la cual se maximiza o minimiza la variable dependiente, mediante las distintas combinaciones asociadas con las dosificaciones o niveles de los distintos factores presentes en el experimento.
- ♣ Experimentos estratificados: En este tipo de experimentos se impone una serie de condiciones previas en la asignación de los tratamientos, formando grupos de unidades experimentales más semejantes.
- ♣ Experimentos con efectos fijos: Experimentos en los cuales se elige una población de tratamientos que son los únicos de interés y las inferencias se efectúan sobre los efectos medios de los tratamientos.
- ♣ Experimentos con efectos aleatorios: Son aquellos donde el interés consiste en comparar el efecto de la variabilidad asociada a los tratamientos, y se se elige un subconjunto de ellos para tal fin.
- ♣ Experimentos mixtos: Son experimentos que contienen factores fijos y aleatorios en el modelo de clasificación experimental.
- ♣ Experimentos balanceados: Se caracterizan porque todos los tratamientos tienen el mismo número de replicaciones.

- ♣ Experimentos desbalanceados: Aquellos experimentos en los cuales uno o más tratamientos no poseen el mismo número de replicaciones que los demás.

### 1.3. Medidas de posición y medidas de dispersión

#### 1.3.1. Medidas de tendencia central

Este tipo de medidas se refieren a los parámetros y estadísticos que indican el centro de la distribución de un conjunto de datos; entre ellos se pueden destacar:

- ♣ Media aritmética: La media es el promedio de las observaciones de una variable  $y_i = (y_1, y_2, y_3, \dots, y_I)$ . La media aritmética estaría dada por:

En el caso de la población ( $\mu$ ):

$$\mu = \frac{\sum_{i=1}^I y_i}{N}, \text{ donde } N \text{ es el tamaño de la población } (N = I)$$

En el caso de la muestra ( $\bar{y}$ ):

$$\bar{y} = \frac{\sum_{i=1}^I y_i}{n}, \text{ donde } n \text{ es el tamaño de la muestra } (n = I)$$

Las principales características de la media aritmética son:

La sumatoria de los valores de las observaciones es igual al valor de la media multiplicado por el número de observaciones:

$$\bar{y} * n = \sum_{i=1}^I y_i$$

La suma de los desviaciones de los datos en relación con la media aritmética es cero:

$$\sum_{i=1}^I (y_i - \bar{y}) = 0$$

Si se suma o se resta una constante ( $c$ ) a los valores de las observaciones, la media aritmética estaría dada por la media de las observaciones más la constante:

$$\frac{\sum_{i=1}^I (y_i + c)}{n} = \bar{y} + c$$

$$\frac{\sum_{i=1}^I (y_i - c)}{n} = \bar{y} - c$$

En el caso de la resta, cuando la constante se aproxima a la media, la sumatoria del cuadrado de las desviaciones es mínima (teorema de König).

$$\frac{\sum_{i=1}^I (y_i - c)^2}{n} \text{ es mínimo, cuando } \bar{y} = c$$

Si se multiplica o se divide una constante ( $c$ ) a los valores de las observaciones, la media aritmética estaría dada por la media de las observaciones multiplicada o dividida por la constante:

$$\frac{\sum_{i=1}^I (y_i * c)}{n} = \bar{y} * c$$

$$\frac{\sum_{i=1}^I \frac{y_i}{c}}{n} = \frac{\bar{y}}{c}$$

La media aritmética asume un valor en el intervalo del valor mínimo y máximo de los datos y se caracteriza porque es sensible a valores muy extremos y su valor no será representativo de la centralización de la distribución de los datos.

Obtengamos la media aritmética del peso en kilogramos de una muestra de lechones:

$$y_i = (y_1, y_2, y_3, \dots, y_I) = (1.5, 1.6, 1.8, 1.2, 1.3, 1.3, 1.3, 1.1)$$

En este caso la muestra tiene 8 observaciones ( $n = I = 8$ ) y su media aritmética se calcula así:

$$\bar{y} = \frac{\sum_{i=1}^8 y_i}{8} = \frac{1.5+1.6+1.8+1.2+1.3+1.3+1.3+1.1}{8} = 1.3875$$

La suma de los desvíos es:

$$\sum_{i=1}^8 (y_i - \bar{y}) = (1.5 - 1.3875) + (1.6 - 1.3875) + \dots + (1.1 - 1.3875) = 0$$

- ♣ **Media geométrica:** Es recomendada principalmente para datos provenientes de una progresión geométrica, promedios de razones de variables, promedios de porcentajes y promedios de índices, y es menos sensible a valores extremos que la media aritmética.

Si una variable está dada por  $y_i = (y_1, y_2, y_3, \dots, y_I)$ , la media geométrica está dada por la siguiente expresión:

$$\bar{y}_G = \sqrt[I]{\prod_{i=1}^I y_i}$$

Para este ejemplo tendríamos:

$$\bar{y}_G = \sqrt[8]{\prod_{i=1}^8 y_i} = \sqrt[8]{1.5 * 1.6 * .. * 1.1} = 1.3716$$

- ♣ **Media armónica:** La media armónica es muy utilizada en datos que involucran la medición de velocidades, tiempos y rendimientos. También es utilizada en estudios de genética y crecimiento de poblaciones. Es sensible a datos con valores pequeños (ceranos a cero) y no permite datos iguales a cero.

Si una variable está dada por  $y_i = (y_1, y_2, y_3, \dots, y_I)$ , la media armónica tiene la siguiente expresión:

$$\bar{y}_H = \frac{I}{\sum_{i=1}^I \frac{1}{y_i}}$$

Para el ejemplo de los lechones tendríamos que el valor de la media armónica es:

$$\bar{y}_H = \frac{8}{\frac{1}{1.5} + \frac{1}{1.6} + \dots + \frac{1}{1.1}} = 1.3565$$

- ♣ **Moda:** En una variable aleatoria  $y_i = (y_1, y_2, y_3, \dots, y_I)$ , la moda del conjunto de datos estaría dada por el valor más repetido o de mayor frecuencia absoluta.

Para el ejemplo de los lechones anteriormente mencionado, tenemos que 1.3 kg es el valor más frecuente, entonces la moda es 1.3.

- ♣ **Mediana:** Si se tiene una variable  $y_i = (y_1, y_2, y_3, \dots, y_I)$ , con los  $y_i$  ordenados de menor a mayor ( $y_i = (y_{(1)}, y_{(2)}, y_{(3)}, \dots, y_{(I)})$ ), la mediana hace relación al dato que ocupa la posición central de ese ordenamiento.

En el caso de tamaño de muestra impar (por ejemplo  $n = 5$  o  $n = 7$ ), el valor que se encuentra en la mitad de los datos ordenados sería el que representa a la mediana. En el caso de un tamaño de muestra par (por ejemplo  $n = 6$  o  $n = 8$ ), la mediana está dada por el promedio de los dos valores centrales. Para el ejemplo de los pesos de los lechones tendríamos:

Datos originales:

$$y_i = (y_1, y_2, y_3, \dots, y_I) = (1.5, 1.6, 1.8, 1.2, 1.3, 1.3, 1.3, 1.1)$$

Datos ordenados:

$$y_i = (y_{(1)}, y_{(2)}, y_{(3)}, \dots, y_{(I)}) = (1.1_{(1)}, 1.2_{(2)}, 1.3_{(3)}, 1.3_{(4)}, 1.3_{(5)}, 1.5_{(6)}, 1.6_{(7)}, 1.8_{(8)}).$$

En este caso,  $I$  es un número par (8), entonces la mediana sería  $\frac{1.3_{(4)} + 1.3_{(5)}}{2} = 1.3$ .

### 1.3.2. Medidas de dispersión o de variación

Estas medidas se refieren a los parámetros que indican variabilidad de los datos en una distribución, entre ellas tenemos:

- ♣ **Rango ( $R$ ):** En una variable  $y_i = (y_1, y_2, y_3, \dots, y_I)$ , con los  $y_i$  ordenados de menor a mayor, el rango se define como la diferencia entre el valor mínimo y el valor máximo. Para el ejemplo, los datos ordenados son:

$$y_i = (1.1_1, 1.2_2, 1.3_3, 1.3_4, 1.3_5, 1.5_6, 1.6_7, 1.8_8)$$

En este caso, el valor mínimo es 1.1, el máximo es 1.8 y el rango sería 0.7.

- ♣ **Varianza:** La varianza mide la dispersión cuadrática de los datos con relación a la media aritmética.

En una variable  $y_i = (y_1, y_2, y_3, \dots, y_I)$ , la varianza estaría dada por:

En el caso de la población:

$$\sigma^2 = \frac{\sum_{i=1}^I (y_i - \mu)^2}{I} = \frac{\sum_{i=1}^I y_i^2 - \frac{\left(\sum_{i=1}^I y_i\right)^2}{I}}{I}$$

En el caso de la muestra:

$$S^2 = \frac{\sum_{i=1}^I (y_i - \bar{y})^2}{I - 1} = \frac{\sum_{i=1}^I y_i^2 - \frac{\left(\sum_{i=1}^I y_i\right)^2}{I}}{I - 1}$$

Donde  $I - 1$  hace relación a los grados de libertad (*g.l.*) para el cálculo de la varianza.

Esta medida es sensible a la presencia de datos extremos, sobre todo en muestras pequeñas.

En el caso del ejemplo se tendría:

$$S^2 = \frac{(1.1-1.3875)^2 + (1.2-1.3875)^2 + \dots + (1.8-1.3875)^2}{8-1} = 0.0527\text{kg}^2$$

- ♣ **Desviación estándar o típico:** Es la raíz cuadrada de la varianza y representa la magnitud de la dispersión que presentan los datos con respecto a la media. Si los datos se distribuyen normalmente, a una desviación por debajo ( $-$ ) y una desviación por encima ( $+$ ) de la media, se tiene el 68% de los datos, a dos desvíos ( $+/-$ ) se encuentra el 95% de los datos y a tres desvíos el 99%: La desviación estándar está dada por:

En el caso de la población:

$$\sigma = \sqrt{\sigma^2}$$

En el caso de la muestra:

$$S = \sqrt{S^2}$$

En el caso del ejemplo se tendría:

$$S = \sqrt{0.00527} = 0.2295\text{kg}$$

El valor anterior indica que el 68% de los pesos de los lechones ( $\bar{y} \pm S$ ) se encuentra entre  $1.3875 - 0.2295 = 1.158$  y  $1.3875 + 0.2295 = 1.617$  kg; el 95% de los pesos de los lechones ( $\bar{y} \pm 2S$ ) se encuentra entre  $1.3875 - (2 * 0.2295) = 0.9285$  y  $1.3875 + (2 * 0.2295) = 1.8465$  kg; y el 99% de los pesos de los lechones ( $\bar{y} \pm 3S$ ) se encuentra entre  $1.3875 - (3 * 0.2295) = 0.699$  y  $1.3875 + (3 * 0.2295) = 2.076$  kg.

- ♣ Coeficiente de variación: Es el cociente entre la desviación estándar y la media aritmética, expresada en porcentaje:

$$C.V. = \frac{S}{\bar{y}} * 100$$

En el caso del ejemplo se tendría:

$$C.V. = \frac{0.2295}{1.3875} * 100 = 16.5 \%$$

Si tuviéramos otra muestra de peso de lechones con un coeficiente de variación de 6%, podríamos decir que los pesos de nuestros lechones presentaron mayor dispersión que la segunda muestra.

- ♣ Error estándar o típico: Es una medida del error generado de las posibles medias muestrales ( $\bar{y}_s$ ) como estimativas de la media poblacional ( $\mu$ ). Entre mayor sea el tamaño de la muestra,  $\bar{y}$  es más representativo de  $\mu$ , por tanto:

$$S_{\bar{y}} = \frac{S}{\sqrt{I}}$$

El error estándar permite encontrar el intervalo de confianza para la media, donde este intervalo está dado por un conjunto de valores donde se encuentra el verdadero valor del parámetro.

Si una variable presenta distribución normal, entonces el intervalo de confianza, con una probabilidad de equivocación  $\alpha$  o nivel de confianza  $(1 - \alpha)$  estaría dado por:

$$\bar{y} - Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{I}} \leq \mu \leq \bar{y} + Z_{\frac{\alpha}{2}} \frac{S}{\sqrt{I}}$$

Donde  $Z$  es la notación que se le da a la distribución normal estándar, con media 0 y varianza 1.

El error estándar del peso de los lechones estaría dado por:

$$S_{\hat{y}} = \frac{0.2295}{\sqrt{8}} = 0.081$$

Si el nivel de confianza es del 95 % ( $Z_{\frac{\alpha}{2}} = 1.96$ ), el intervalo de confianza sería:

$$1.3875 - 1.96 * 0.081 \leq \mu \leq 1.3875 + 1.96 * 0.081$$

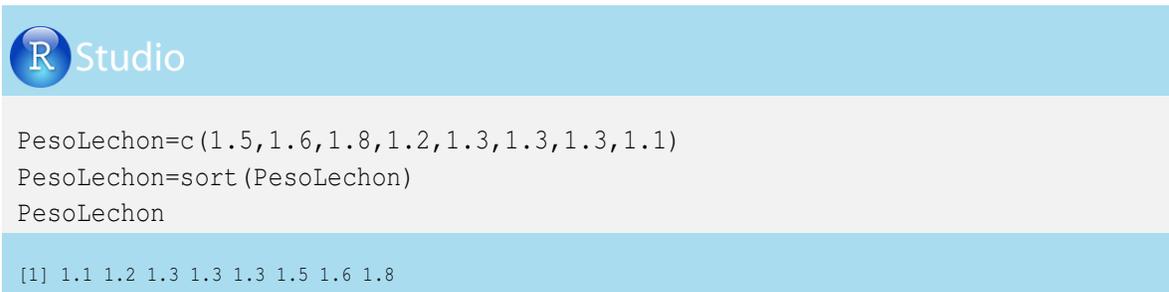
$$1.23 \leq \mu \leq 1.55$$

El intervalo de confianza de la media poblacional del peso de los lechones está entre 1.23 y 1.55 kg, con una confianza del 95 %.

### 1.3.3. Implementación en R-project

A continuación se presentan los comandos en R-project (Core Team, 2012) y su implementación para obtener las medidas de tendencia central y de dispersión mencionadas, mediante un ejemplo utilizado anteriormente, del peso de los lechones en kilogramos.

Para iniciar, ingresaremos los datos del peso y luego los organizaremos de menor a mayor con el comando *sort*, así:



```

R Studio

PesoLechon=c(1.5,1.6,1.8,1.2,1.3,1.3,1.3,1.1)
PesoLechon=sort(PesoLechon)
PesoLechon

[1] 1.1 1.2 1.3 1.3 1.3 1.5 1.6 1.8
  
```

Para estimar la media y la mediana del peso de los lechones en R-project, utilizaremos el comando *mean* y *median*, respectivamente:



```

R Studio

mean(PesoLechon)
median(PesoLechon)

> mean(PesoLechon)
[1] 1.3875
> median(PesoLechon)
[1] 1.3
  
```

En R-project no existe un comando específico para identificar la moda de una variable, pero podemos crear una función para tal fin, empleando los comandos *table* y *which.max*. El

comando *table* genera una tabla de frecuencias y el comando *which.max* permite escoger el valor máximo de los valores de frecuencia de la tabla, así:

```
R Studio

TablaDeFrecuencias=table (PesoLechon)
TablaDeFrecuencias
moda=which.max (TablaDeFrecuencias)
moda

> TablaDeFrecuencias
PesoLechon
 1.1 1.2 1.3 1.5 1.6 1.8
  1  1  3  1  1  1
>moda
 1.3
  3
```

En la tabla de frecuencias podemos ver que existen tres observaciones de peso de los lechones de 1.3 kg, los otros pesos no se repiten. Por tanto, la moda es 1.3 kg, que tiene una frecuencia de 3.

Para el cálculo de la media geométrica y armónica utilizaremos la librería *psych*, creada por Revelle (2012).

```
R Studio

install.packages("psych")
library (psych)
geometric.mean (PesoLechon)
harmonic.mean (PesoLechon)

> geometric.mean (PesoLechon)
[1] 1.371627
> harmonic.mean (PesoLechon)
[1] 1.356544
```

Para identificar los valores mínimos y máximos, y calcular el valor del rango, la varianza y la desviación estándar de un conjunto de datos en R-project, utilizaremos los comandos *min*, *max*, *range*, *var* y *sd*, respectivamente; veamos:



```

min(PesoLechon)
max(PesoLechon)
range(PesoLechon)
var(PesoLechon)
sd(PesoLechon)

> min(PesoLechon)
[1] 1.1
> max(PesoLechon)
[1] 1.8
> range(PesoLechon)
[1] 1.1 1.8
> var(PesoLechon)
[1] 0.05267857
> sd(PesoLechon)
[1] 0.2295181

```

El error estándar y el coeficiente de variación los calcularemos implementando las fórmulas vistas anteriormente, las cuales incluyen los comandos *sd* (desviación estándar) y *length* (para determinar el número de datos de la muestra). La programación sería:



```

error.standar= sd(PesoLechon)/sqrt(length(PesoLechon))
error.standar
c.v.= (sd(PesoLechon)/mean(PesoLechon))*100
c.v.

> error.standar
[1] 0.08114691
> c.v.= (sd(PesoLechon)/mean(PesoLechon))*100
> c.v.
[1] 16.54185

```

Si la variable a analizar presenta datos faltantes o perdidos (*NA*), es necesario utilizar el argumento *na.rm = TRUE*, el cual elimina los datos faltantes antes de realizar el cálculo computacional.

Para contextualizar al lector, vamos a suponer que en la variable que contiene los pesos de los lechones (*PesoLechon*), existe un dato *NA*. La programación para el cálculo de la media sería: *mean(PesoLechon, na.rm = TRUE)*.

Ahora veamos el comando *summary*, que genera una tabla resumen con algunas medidas de los datos:



```
summary(PesoLechon)

> summary(PesoLechon)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.100  1.275   1.300   1.388   1.525   1.800
```

En la programación encontramos que la media es de 1.388 kg, la mediana es de 1.3 kg, el peso mínimo fue de 1.1 kg y el peso máximo de 1.8 kg. Puede apreciarse también que existen dos valores adicionales (1st *Qu.* = 1.275 y 3st *Qu.* = 1.525), los cuales hacen referencia al primer y tercer cuartil, respectivamente.

Los cuartiles dividen la distribución de probabilidad empírica de los datos para ubicarlos en cuatro grupos de igual tamaño (25 % de los datos para cada grupo). En el ejemplo, el 25 % de los datos es menor o igual que 1.275 kg y el 75 % de los datos es menor o igual que 1.525 kg.

A continuación utilizaremos el comando *quantile* para generar la división de los datos en cuatro grupos de igual tamaño:



```
quantile(PesoLechon)

> quantile(PesoLechon)
  0%   25%   50%   75%  100%
 1.100 1.275 1.300 1.525 1.800
```

#### 1.4. Pruebas de significancia

El objetivo principal de un experimento es tomar decisiones a partir de la información muestral obtenida respecto a los factores que pueden afectar las variables de interés. Para llegar a tomar decisiones conviene hacer determinados supuestos o conjeturas sobre el proceso bajo estudio. Un investigador puede considerar inicialmente que los resultados obtenidos son iguales para cualquier factor (hipótesis nula,  $H_0$ ) o que existen diferencias significativas entre los niveles (hipótesis alternativa,  $H_a$ ). Esto se comprobará al final del experimento y una de estas hipótesis será aceptada.

Para escoger una de las hipótesis, es necesario hacer una prueba de contraste de hipótesis, siendo las pruebas *t* de Student y *F* de Snedecor las más empleadas en la estadística paramétrica.

Estas pruebas son válidas cuando se cumplen los supuestos teóricos, como:

- ♣ Las observaciones deben ser independientes
- ♣ Las observaciones provienen de poblaciones normalmente distribuidas
- ♣ Homocedasticidad de las varianzas

Presentaremos mayores detalles de estos supuestos en los próximos capítulos. En este capítulo nos concentraremos en las pruebas paramétricas  $t$  de Student y  $F$  de Snedecor.

#### 1.4.1. Prueba para dos muestras independientes con varianzas homogéneas

La prueba  $t$  de Student es utilizada para determinar diferencias entre dos medias muestrales de tamaño pequeño con media  $\bar{y}$  y varianza  $S^2$ . Si la variable  $y_i$  de dos muestras ( $a$  y  $b$ ) tomadas aleatoriamente en dos poblaciones son independientes, idénticamente distribuidas y se presume que las varianzas son homogéneas; las hipótesis para probar serían:

$$H_0 =: \mu_a = \mu_b$$

$$H_a =: \mu_a \neq \mu_b$$

El cálculo de la prueba  $t$  de Student para el anterior contraste está dado por:

$$t = \frac{\bar{y}_a - \bar{y}_b}{\sqrt{S_{y_a y_b}^2 \left( \frac{1}{I_a} + \frac{1}{I_b} \right)}}$$

Donde  $I_a$  y  $I_b$  es el tamaño de cada muestra  $a$  y  $b$ , respectivamente, y  $S_{y_a y_b}^2$  es la varianza conjunta de  $a$  y  $b$ , que está dada por:

$$S_{y_a y_b}^2 = \frac{(I_a - 1)S_{y_a}^2 + (I_b - 1)S_{y_b}^2}{I_a + I_b - 2}$$

Donde  $S_{y_a}^2$  y  $S_{y_b}^2$  son las varianzas de  $a$  y  $b$ , respectivamente.

Si las dos muestras son iguales, la comprobación estaría dada por la aproximación a una distribución  $t$  con determinados grados de libertad ( $g.l.t$ ), donde  $g.l.t$  estaría dado por:

$$g.l.t = I_a + I_b - 2$$

Se rechaza la hipótesis nula a determinado nivel de confiabilidad ( $\alpha$ ), si:

$$|t| > t_{(1-\frac{\alpha}{2});g.l.t}$$

Donde  $t_{(1-\frac{\alpha}{2});g.l.t}$  es un valor asociado a una distribución  $t$  de Student con  $g.l.t$  grados de libertad que deja a la derecha el  $\frac{\alpha}{2}$  de los datos.

En el siguiente ejemplo deseamos saber si existen diferencias en las medias de peso de lechones de dos muestras ( $\bar{y}_a$  y  $\bar{y}_b$ ), tomadas aleatoriamente de dos poblaciones, cuyas hipótesis serían:

$$H_0 =: \mu_a = \mu_b$$

$$H_a =: \mu_a \neq \mu_b$$

La muestra  $a$  tiene las siguientes observaciones:

$$y_{a_i} = (y_{a_1}, y_{a_2}, y_{a_3}, \dots, y_{a_l}) = (1.5, 1.6, 1.8, 1.2, 1.3, 1.3, 1.3, 1.1), \text{ con } I_a = 8$$

La muestra  $b$ :

$$y_{b_i} = (y_{b_1}, y_{b_2}, y_{b_3}, \dots, y_{b_l}) = (1.3, 1.4, 1.5, 1.5, 1.5, 1.8, 2), \text{ con } I_b = 7:$$

Las medias y las varianzas serían:

$$\bar{y}_a = 1.3875$$

$$\bar{y}_b = 1.5714$$

$$S_{y_a}^2 = 0.0527$$

$$S_{y_b}^2 = 0.0590$$

La varianza conjunta de  $a$  y  $b$  ( $S_{y_a y_b}^2$ ) sería:

$$S_{y_a y_b}^2 = \frac{(8-1)0.0527 + (7-1)0.059}{8+7-2} = 0.0556$$

El cálculo de la prueba  $t$  de Student sería:

$$t = \frac{1.3875 - 1.5714}{\sqrt{0.0556 * (\frac{1}{8} + \frac{1}{7})}} = -1.507$$

Para comprobar si las dos muestras son iguales se realiza la aproximación a una distribución  $t$  con  $g.l.t = 8 + 7 - 2 = 13$  grados de libertad y una confiabilidad del 95 % ( $\alpha = 0.05$ ).

Se tendría entonces un valor de la tabla de distribución  $t$  de Student:

$$t_{(1-\frac{0.05}{2});13} = 2.1604$$

La hipótesis nula estaría dada por:

$$H_0 : \bar{y}_a = \bar{y}_b, \text{ si } |t| < t_{(1-\frac{0.05}{2});13}.$$

En este caso sería:  $|-1.5069| < 2.1604$ ; por tanto, aceptamos la hipótesis nula (las medias de las muestras son iguales) y rechazamos la hipótesis alterna (las medias son diferentes).

Realicemos este ejercicio en R-project introduciendo inicialmente los datos:



```
PesoLechon=c(1.5,1.6,1.8,1.2,1.3,1.3,1.3,1.1)
PesoLechon2=c(1.3,1.4,1.5,1.5,1.5,1.8,2)
PesoLechon
PesoLechon2
```

```
> PesoLechon
[1] 1.5 1.6 1.8 1.2 1.3 1.3 1.3 1.1
> PesoLechon2
[1] 1.3 1.4 1.5 1.5 1.5 1.8 2.0
```

La implementación de esta prueba de hipótesis en R-project se realiza por medio del comando *t.student*, seguido de los nombres de las variables a comparar y del argumento *var.equal = TRUE*, con lo que suponemos que las varianzas son iguales; veamos:



```
t.test(PesoLechon, PesoLechon2, var.equal=TRUE)
```

```
Two Sample t-test
data: PesoLechon and PesoLechon2
t = -1.5069, df = 13, p-value = 0.1557
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4476149  0.0797578
sample estimates:
mean of x mean of y
 1.387500  1.571429
```

#### 1.4.2. Prueba para dos muestras independientes con varianzas heterogéneas

Cuando se presume que las muestras tienen diferente varianza y que las medias son diferentes, es necesario verificar inicialmente si las varianzas son iguales o diferentes. Esto es posible mediante la prueba de razón de varianzas de  $F$ , que sigue una distribución  $F$  de Snedecor.

Si dos muestras poblacionales  $a$  y  $b$  tienen varianzas  $S_a^2$  y  $S_b^2$ , respectivamente, las hipótesis para homogeneidad o heterogeneidad de varianzas serían:

$$H_o : S_a^2 = S_b^2$$

$$H_a : S_a^2 \neq S_b^2$$

La prueba  $F$  de razón de varianzas estaría dada por:

$$F = \frac{S_a^2}{S_b^2}$$

Teniendo en cuenta que  $S_a^2$  tiene un valor mayor que  $S_b^2$ , la hipótesis nula sigue una distribución  $F$  con grados de libertad en el numerador ( $g.l.n. = I_a - 1$ ) y en el denominador ( $g.l.d. = I_b - 1$ ). En el caso de rechazar  $H_0$ , las varianzas son heterogéneas.

El valor de  $F$  de la tabla estaría dado por el valor de  $\alpha$  requerido, de  $g.l.n.$  y de  $g.l.d.$ .

Rechazamos la hipótesis nula, si:

$$F > F_{(1-\alpha);(g.l.n.);(g.l.d)}$$

En el caso de presentar varianzas diferentes (varianzas heterogéneas), se debe utilizar la prueba  $t$  Student modificada por Welch ( $t$  de Welch) para probar la diferencia de medias de las muestras, la cual está dada por:

$$t = \frac{\bar{y}_a - \bar{y}_b}{\sqrt{2 \frac{S_{y_a}^2}{I_a} + \frac{S_{y_b}^2}{I_b}}}$$

Para comprobar si las dos muestras son iguales se realiza la aproximación a una distribución  $t$  de Student, con  $g.l.t$  grados de libertad:

$$g.l.t = \frac{\left( \frac{S_{y_a}^2}{I_a} + \frac{S_{y_b}^2}{I_b} \right)^2}{\frac{\left( \frac{S_{y_a}^2}{I_a} \right)^2}{I_a - 1} + \frac{\left( \frac{S_{y_b}^2}{I_b} \right)^2}{I_b - 1}}$$

Veamos el siguiente ejemplo: un frigorífico desea saber si existen diferencias en la cobertura de grasa de los lotes de animales que sacrifican. El administrador planteó un ensayo con dos lotes, con las siguientes hipótesis:

$$H_0 : \mu_a = \mu_b$$

$$H_a : \mu_a \neq \mu_b$$

La muestra  $a$  tiene las siguientes observaciones de espesor de grasa (mm):

$$y_{a_i} = (y_{a_1}, y_{a_2}, y_{a_3}, \dots, y_{a_8}) = (4.3, 5.4, 6.5, 5.5, 5.5, 6.8, 5.2, 5.3), \text{ con } I_a = 8$$

La muestra  $b$ :

$$y_{b_i} = (y_{b_1}, y_{b_2}, y_{b_3}, \dots, y_{b_9}) = (5.5, 5.6, 5.8, 5.2, 5.3, 5.3, 5.3, 5.1, 5.1), \text{ con } I_b = 9:$$

Las medias y las varianzas serían:

$$\bar{y}_a = 5.5625$$

$$\bar{y}_b = 5.3555$$

$$S_{y_a}^2 = 0.6055$$

$$S_{y_b}^2 = 0.0552$$

Podemos observar que existe una amplia diferencia en las varianzas de las dos muestras. Para confirmar estas diferencias, procederemos a verificar, con la prueba  $F$  de Snedecor, si las varianzas son iguales o diferentes:

$$F = \frac{S_{y_a}^2}{S_{y_b}^2} = \frac{0.6055}{0.552} = 10.95$$

El valor en la tabla de distribución de  $F$  de Snedecor, con un  $\alpha = 0.05$ , sería de  $F_{(1-0.05);(8-1);(9-1)} = 3.5$ , por tanto, rechazamos la hipótesis nula porque  $10.95 > 3.5$ , indicando que existen varianzas heterogéneas entre muestras. Por lo tanto, la comparación de medias sería realizada por la prueba  $t$  de Welch; veamos:

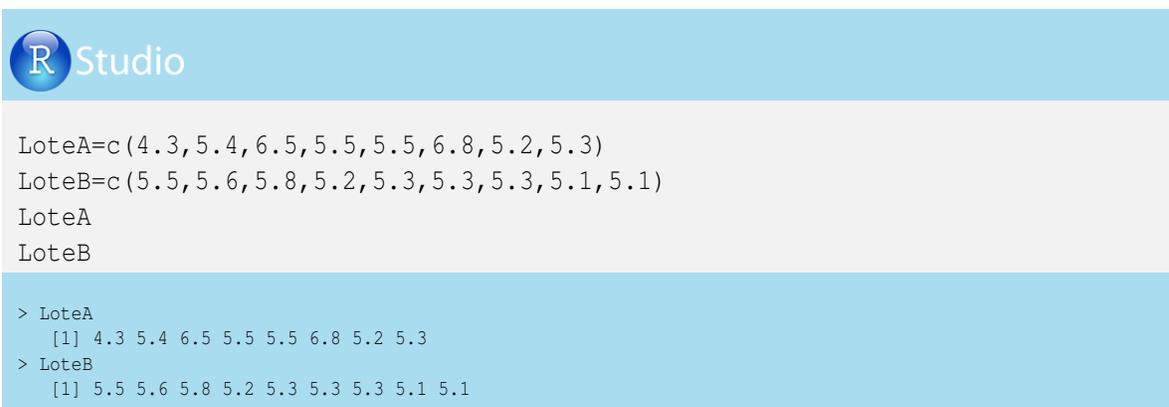
$$t = \frac{5.5625 - 5.3555}{\sqrt{\frac{0.6055}{8} + \frac{0.0553}{9}}} = 0.7234$$

El cálculo de los grados de libertad está dado por:

$$g.l.t = \frac{\left(\frac{0.6055}{8} + \frac{0.0553}{9}\right)^2}{\frac{\left(\frac{0.6055}{8}\right)^2}{7} + \frac{\left(\frac{0.0553}{9}\right)^2}{8}} = 8.135$$

Con un  $\alpha$  del 0.05 tendríamos que  $t_{(1-\frac{0.05}{2});8.135} \approx 2.31$

Aceptamos la hipótesis nula ( $H_0 : \bar{y}_a = \bar{y}_b$ ), porque  $|0.72| < 2.31$ . Por lo tanto, los lotes de animales presentaron igual media de espesor de grasa de cadera, con varianzas heterogéneas. A continuación ingresaremos los datos en R-project:



```

R Studio

LoteA=c(4.3,5.4,6.5,5.5,5.5,6.8,5.2,5.3)
LoteB=c(5.5,5.6,5.8,5.2,5.3,5.3,5.3,5.1,5.1)
LoteA
LoteB

> LoteA
[1] 4.3 5.4 6.5 5.5 5.5 6.8 5.2 5.3
> LoteB
[1] 5.5 5.6 5.8 5.2 5.3 5.3 5.3 5.1 5.1

```

Con el comando `var.test` seguido de los nombres de las muestras, realizaremos la prueba de igualdad de varianzas de  $F$  Snedecor.



```
var.test(LoteA, LoteB)

F test to compare two variances
data: LoteA and LoteB
F = 10.9544, num df = 7, denom df = 8,
p-value = 0.003073
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 2.418961 53.669410
sample estimates:
ratio of variances
 10.95441
```

Utilizaremos el comando *t.student* de R-project, asumiendo que las varianzas son diferentes con el argumento *var.equal = FALSE*:



```
t.test(LoteA, LoteB, var.equal=FALSE)

Welch Two Sample t-test
data: LoteA and LoteB
t = 0.7234, df = 8.135, p-value = 0.4897
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4508217 0.8647106
sample estimates:
mean of x mean of y
 5.562500 5.355556
```

### 1.4.3. Prueba para muestras pareadas

Cuando una misma muestra ha sido evaluada dos veces (antes y después), o cuando existen muestras pareadas, el estadístico del contraste *t* de Student para datos pareados sería:

$$t = \frac{(\bar{y}_c)}{\frac{S_{y_c}}{\sqrt{I}}}$$

Donde  $\bar{y}_c$  es la media de las diferencias entre los valores pareados y  $S_{y_c}$  es la desviación estándar de esas diferencias.

Por lo anterior, las hipótesis a contrastar estarían dadas por la diferencia de observaciones pareadas provenientes de una población normal con  $\mu_c$ , siendo:

$$H_0 =: \mu_c = 0$$

$$H_a =: \mu_c \neq 0$$

Se rechaza la hipótesis nula si:  $|t| > t_{(1-\frac{\alpha}{2}); g.l.t}$ , donde los grados de libertad estarían dados por  $g.l.t = I - 1$ :

A continuación desarrollaremos un ejemplo relacionado con la cantidad de semen producido (en ml) 6 búfalos (*Iterbio, Procedonio, Epitacio, Holmio, Coltan y Europio*) en dos colectas de semen realizadas en el mes de abril y agosto.

Las hipótesis planteadas son:

$$H_0 =: \mu_c = 0$$

$$H_a =: \mu_c \neq 0$$

Los datos de las colectas son:

$$y_a = (1.8, 1.7, 1.6, 1.5, 1.4, 1.2)$$

$$y_b = (1.4, 1.1, 1.1, 1.1, 1.2, 0.8)$$

La diferencia entre datos pareados es:

$$y_a - y_b = y_c = (0.4, 0.6, 0.5, 0.4, 0.2, 0.4)$$

$$\bar{y}_c = 0.4167$$

$$S_{y_c} = 0.1329$$

$$t = \frac{0.116}{\frac{0.1329}{\sqrt{6}}} = 7.6787$$

Los grados de libertad son  $6 - 1$  y el valor de  $t$  al 95 % de confianza sería:

$$t_{(1-\frac{0.05}{2}; 5)} = 2.571$$

En este caso,  $|7.6787| > 2.571$ , lo cual señala que existieron diferencias entre las medias de las dos colectas con una confianza del 95 %. Esto indica que los butoros sometidos a colectas sucesivas de semen eyaculan diferente cantidad.

Desarrollemos este ejercicio en R-project. Inicialmente realicemos la programación para ingresar los datos y así crear una hoja de datos llamada *REPRO*, con las columnas relacionadas con nombre del búfalo (*BUFALO*), cantidad de semen en el mes de abril (*Esperma1*) y cantidad de semen producida en una colecta en el mes de agosto (*Esperma2*):



```
Esperma1=c(1.8,1.7,1.6,1.5,1.4,1.2)
Esperma2=c(1.4,1.1,1.1,1.1,1.2,0.8)
BUFALO=c("Iterbio", "Procedonio","Epitacio", "Holmio", "Coltan","Europio")
REPRO=rbind (BUFALO, Esperma1, Esperma2)
REPRO
```

```
      [,1]      [,2]      [,3]      [,4]
BUFALO "Iterbio" "Procedonio" "Epitacio" "Holmio"
Esperma1 "1.8"    "1.7"    "1.6"    "1.5"
Esperma2 "1.4"    "1.1"    "1.1"    "1.1"
      [,5]      [,6]
BUFALO "Coltan" "Europio"
Esperma1 "1.4"    "1.2"
Esperma2 "1.2"    "0.8"
```

Para implementar la prueba  $t$  de Student para muestras pareadas, utilizaremos el comando `t.student` de R-project, indicando que las muestras son pareadas con el argumento `paired = TRUE`; veamos:



```
t.test(Esperma1, Esperma2, paired=TRUE)
```

```
Paired t-test
data: Esperma1 and Esperma2
t = 7.6787, df = 5, p-value = 0.0005971
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2771799 0.5561535
sample estimates:
mean of the differences
      0.4166667
```

#### 1.4.4. Prueba para más de dos muestras

Para probar las hipótesis si más de dos muestras son iguales o diferentes, se requiere la construcción del análisis de varianza, y la prueba que se utiliza para ello es la de  $F$  de Snedecor, descrita anteriormente, en la sección 1.4.1.

Para realizar esto último tendremos un valor de  $F_{calculado}$ , que está dado por la relación de factores de variación del modelo, y el  $F_{Tabla}$ , que está dado por los grados de libertad de las fuentes de variación comprometidas y el nivel de significancia para rechazar la hipótesis nula (mayores detalles en los próximos capítulos).

## Capítulo 2

### Diseño completamente aleatorizado

#### 2.1. Generalidades

El diseño completamente aleatorizado se caracteriza porque los tratamientos se asignan al azar en las unidades experimentales. En este diseño, la variable respuesta depende de un único factor. Si existen otras causas de variación, se consideran en el error experimental.

Este diseño se aplica cuando las unidades experimentales son homogéneas. En caso de utilizar animales se debe cumplir con un estricto patrón de semejanza (de la misma edad, del mismo sexo, de la misma raza o línea, del mismo lote, etc.).

En este diseño, el modelo lineal aditivo permite explicar la relación entre el factor y el error experimental, por medio de la siguiente expresión:

$$y_{ij} = \mu + \tau_i + \varepsilon_{j(i)}$$

Donde  $y_{ij}$  representa la variable respuesta o dependiente, asociada al  $i$ -ésimo nivel de tratamiento en la  $j$ -ésima replicación,

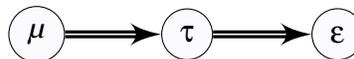
$\mu$  indica el efecto promedio del experimento,

$\tau_i$  indica el efecto promedio del  $i$ -ésimo nivel de tratamiento ( $i = 1, 2, \dots, I$ ), y

$\varepsilon_{j(i)}$  es el error experimental asociado con la  $j$ -ésima replicación ( $j = 1, 2, \dots, J$ ), anidada en el  $i$ -ésimo tratamiento.

Los  $j(i)$  errores son independientes e idénticamente distribuidos (iid) y tienen una distribución normal con media 0 y varianza  $\sigma_\varepsilon^2$ , cuya notación es:  $\varepsilon_{j(i)} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$ .

La estructura de este diseño es:



Existen dos variantes de este diseño unifactorial:

- ♣ Cuando el efecto del tratamiento es fijo ( $\tau$ ), donde los niveles del factor se seleccionan de modo determinista por el investigador. Esto conlleva a que los resultados y su interpretación solo se puedan aplicar a los niveles implicados en el experimento. Su

objetivo es comparar las respuestas promedio de los niveles de dicho factor, cuyas hipótesis nula y alterna son:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \dots = \tau_I$$

$H_a$  : los promedios de los niveles del tratamiento son diferentes

La hipótesis nula indica que ninguno de los niveles del factor tiene un efecto significativo sobre la variable respuesta, mientras que la hipótesis alterna expresa que existe al menos un nivel que tiene efecto notorio sobre los otros niveles.

- ♣ Cuando el efecto del tratamiento es considerado aleatorio ( $t$ ), los niveles de ese factor son una muestra aleatoria de una población y sus resultados se pueden generalizar a todos los elementos de la población. Las hipótesis a probar son:

$$H_0 : \sigma^2 = 0$$

$$H_a : \sigma^2 > 0$$

La hipótesis nula indica que todos los tratamientos son idénticos, mientras que la hipótesis alternativa indica que existe variabilidad entre ellos.

Las principales ventajas del diseño completamente aleatorizado son:

- ♣ Los experimentos pueden ser balanceados o desbalanceados (diferente número de unidades experimentales por nivel)
- ♣ Se puede utilizar un amplio número de tratamientos
- ♣ Maximiza los grados de libertad del error experimental
- ♣ Se puede estimar el error experimental
- ♣ Se pueden transformar los datos
- ♣ Permite realizar análisis con submuestras
- ♣ Permite la inclusión de covariables
- ♣ Permite generar arreglos factoriales, mixtos y confundidos
- ♣ Permite efectuar análisis multivariado de varianza (Manova)
- ♣ Genera la diferencia del efecto promedio estimado de los tratamientos
- ♣ Se pueden utilizar contrastes ortogonales
- ♣ Se pueden trabajar modelos de regresión con estructura de una vía
- ♣ Se puede particionar el experimento, si no hay homogeneidad de varianzas

Como desventaja se tiene que, si no es posible asegurar la homogeneidad del material experimental, este diseño no se puede utilizar.

Este tipo de diseño en experimentación de las ciencias animales puede emplearse en condiciones ambientales homogéneas y un control local adecuado, dadas principalmente en:

- ♣ Experimentos en laboratorios o galpones experimentales, minimizando efectos no controlados
- ♣ Experimentos acuícolas, controlando las condiciones físico-químicas
- ♣ Experimentos de invernadero con control climático
- ♣ Experimentos en bioterios

## 2.2. Análisis de varianza

El enfoque más conocido para probar las hipótesis de un diseño completamente al azar es el análisis de varianza (ANOVA), que busca descubrir cómo se reparte la variabilidad total de la variable respuesta. A continuación detallaremos la estructura del análisis de varianza para efecto fijo y efecto aleatorio:

### 2.2.1. Modelo de efecto fijo

Una posible medida de variabilidad total es la suma de cuadrados total (*SCT*) dada por:

$$SCT = \sum_{i=1}^I \sum_{j=1}^J (y_{ij})^2 - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J y_{ij} \right)^2}{n} = \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \bar{y}_{..})^2$$

Donde:

$$\bar{y}_{..} = \frac{\left( \sum_{i=1}^I \sum_{j=1}^J y_{ij} \right)}{n}$$

La suma de cuadrados relacionada con las diferencias de los promedios de los niveles del efecto de tratamiento, en relación con el promedio general, se conoce con el nombre de suma de cuadrados de los tratamientos (*SCTr*) y se obtiene así:

$$SCTr = \frac{\sum_{i=1}^I \tau_i^2}{J} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J y_{ij} \right)^2}{n}$$

La suma de cuadrados de las diferencias entre las observaciones dentro de cada nivel del efecto de tratamiento, respecto a su promedio, está relacionada con la variabilidad que no es

atribuida a los efectos considerados en el modelo, la cual se conoce como suma de cuadrados del error ( $SCE$ ) y se obtiene de esta manera:

$$SCE = SCT - SCTr$$

Si retomamos el concepto de grados de libertad del capítulo anterior, podemos entender que si tenemos  $n = IJ$  observaciones, entonces la  $SCT$  tiene  $n - 1$  grados de libertad, donde  $n$  es el número de observaciones, la  $SCTr$  tiene  $I - 1$  grados de libertad, donde  $I$  es el número de niveles del efecto, y la  $SCE$  tiene  $n - I$  grados de libertad.

Por tanto, los estimadores independientes de la variabilidad explicada por el efecto del tratamiento están dados por  $\frac{SCTr}{I-1}$  y los de la no explicada por el modelo están dados por  $\frac{SCE}{n-I}$ , conocidos comúnmente como cuadrados medios.

Si la hipótesis nula es cierta, se tiene que la razón entre las sumas de cuadrados medios se distribuye como una  $F$  de Snedecor (ver capítulo anterior) y que en este caso el valor de  $F$  calculado ( $F_c$ ) está dado por:

$$F_c = \frac{\frac{SCTr}{I-1}}{\frac{SCE}{n-I}} = \frac{CMT_r}{CME} \underset{\text{bajo } H_0}{\sim} F_{(I-1, n-I)}$$

Por lo tanto, si existen diferencias entre los promedios de los tratamientos, entonces el valor del estadístico  $F_c$  se hace mayor y se obtiene una región crítica superior, de modo que se rechaza la hipótesis nula si:

$$F_c > F_{\alpha, I-1, n-I}$$

Donde  $F_{\alpha, I-1, n-I}$  es el cuantil  $\alpha$  de la distribución de  $F$  con  $I - 1$  y  $n - I$  grados de libertad.

En resumen, el cuadro del análisis de varianza estaría conformado por:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$
Tratamiento	$I - 1$	$SCTr$	$\frac{SCTr}{I-1}$	$\frac{CMT_r}{CME}$
Error	$n - I$	$SCE$	$\frac{SCE}{n-I}$	
Total	$n - 1$	$SCT$		

### 2.2.2. Modelo con efectos aleatorios

El modelo aditivo lineal en este caso se expresa así:

$$y_{ij} = \mu + t_i + \varepsilon_{j(i)}$$

Pero en este caso  $t_i$  y  $\varepsilon_{j(i)}$  son aleatorios e independientes, con varianza  $\sigma_t^2$  y  $\sigma_\varepsilon^2$ , respectivamente, que se denominan componentes de varianza, donde:

$$\sigma_\varepsilon^2 = CME$$

$$\sigma_t^2 = \frac{CMTr - \sigma_\varepsilon^2}{n} = ECMTr$$

Donde  $ECMtr$  es la esperanza del cuadrado medio del efecto de tratamiento.

El cuadro del análisis de varianza estaría conformado por:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$	Esperanza del cuadrado medio
Tratamiento	$I - 1$	$SCTr$	$\frac{SCTr}{I-1}$	$\frac{CMtr}{CME}$	$\sigma_\varepsilon^2 + n\sigma_t^2$
Error	$n - I$	$SCE$	$\frac{SCE}{n-I}$		$\sigma_\varepsilon^2$
Total	$n - 1$	$SCT$			

### 2.3. Implementación en R-project

En esta sección procederemos a describir la forma correcta de implementar un diseño completamente aleatorizado de efecto fijo, iniciando con la etapa de aleatorización de los tratamientos en las unidades experimentales:

#### 2.3.1. Proceso de aleatorización

Como se enunció previamente, en el diseño completamente aleatorizado cualquier unidad experimental goza de igual probabilidad de ser asociada a un tratamiento, lo cual indica que el proceso de aleatorización utilizado tiene como base la distribución uniforme. Dicho proceso puede realizarse manualmente, sacando al azar balotas o papeles donde se tienen los tratamientos y réplicas, para ser asignados a cada unidad experimental o utilizando algoritmos computacionales que realicen la aleatorización de los tratamientos.

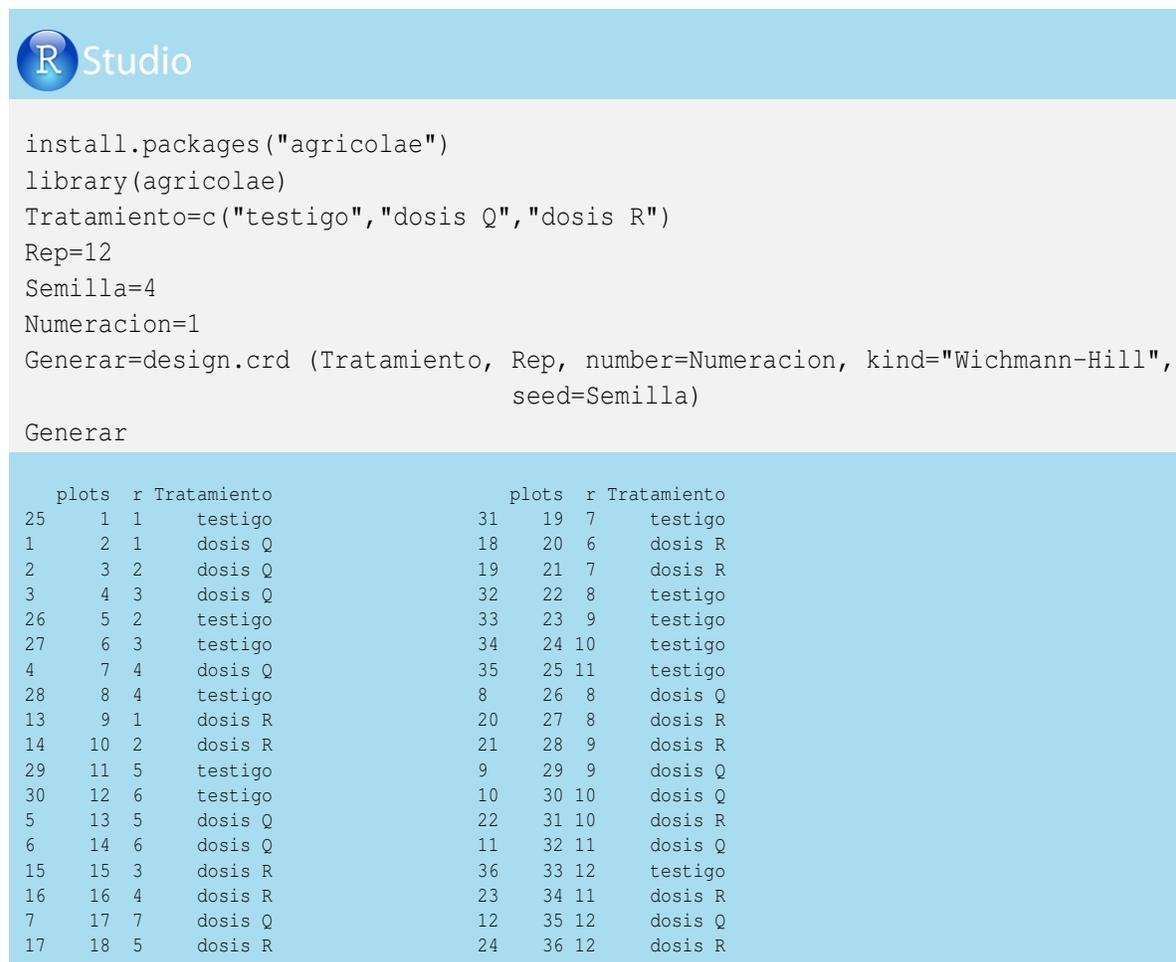
En este libro realizaremos el proceso de aleatorización empleando el algoritmo de Wichmann-Hill para la generación de números pseudoaleatorios, el cual está disponible en el comando *design.crd* de la librería *agricolae*. La librería *agricolae* fue construida por Felipe de Mendiburu (2012) con el objetivo de desarrollar experimentos de agricultura, y la utilizaremos ampliamente en este libro.

Mediante el comando *design.crd* procedemos a aleatorizar las réplicas en los tratamientos, utilizando los siguientes argumentos:

- ♣ El nombre de un vector que contenga los nombres de los niveles de los tratamientos.
- ♣ Número de réplicas. Si el diseño es desbalanceado, el usuario debe introducir un vector que contenga el número de réplicas de cada nivel del efecto de tratamiento.
- ♣ La asignación de un número para generar un vector que enumere secuencialmente las unidades experimentales.

- ♣ Especificación del método de aleatorización (Wichmann-Hill, Marsaglia-Multicarry, Super-Duper, entre otras)
- ♣ La semilla para generar la misma secuencia de números pseudoaleatorios, que utilizaremos para que el lector genere la misma aleatorización de los tratamientos del libro. En caso que el lector quiera otra secuencia, se puede cambiar el valor de la semilla. Si se desea que no mantenga el valor de semilla para generar una aleatorización diferente cada vez que se ejecute la programación, se debe incluir el argumento *seed* = 0.

En el siguiente ejemplo presentaremos la implementación del proceso de aleatorización de tres tratamientos (*testigo*, *dosis Q* y *dosis R*) y doce réplicas por tratamiento (*rep* = 12). Utilizaremos una semilla (*seed* = 4) y enumeraremos las unidades experimentales a partir de 1 (*numeracion* = 1). A continuación realizaremos la programación en R-project, creando la hoja de datos *Generar* que contiene la información resultante de la aleatorización; veamos:



```

install.packages("agricolae")
library(agricolae)
Tratamiento=c("testigo","dosis Q","dosis R")
Rep=12
Semilla=4
Numeracion=1
Generar=design.crd (Tratamiento, Rep, number=Numeracion, kind="Wichmann-Hill",
                    seed=Semilla)

Generar

```

plots	r	Tratamiento	plots	r	Tratamiento		
25	1	1	testigo	31	19	7	testigo
1	2	1	dosis Q	18	20	6	dosis R
2	3	2	dosis Q	19	21	7	dosis R
3	4	3	dosis Q	32	22	8	testigo
26	5	2	testigo	33	23	9	testigo
27	6	3	testigo	34	24	10	testigo
4	7	4	dosis Q	35	25	11	testigo
28	8	4	testigo	8	26	8	dosis Q
13	9	1	dosis R	20	27	8	dosis R
14	10	2	dosis R	21	28	9	dosis R
29	11	5	testigo	9	29	9	dosis Q
30	12	6	testigo	10	30	10	dosis Q
5	13	5	dosis Q	22	31	10	dosis R
6	14	6	dosis Q	11	32	11	dosis Q
15	15	3	dosis R	36	33	12	testigo
16	16	4	dosis R	23	34	11	dosis R
7	17	7	dosis Q	12	35	12	dosis Q
17	18	5	dosis R	24	36	12	dosis R

Al observar la tabla resultante podemos apreciar que la primera unidad experimental será la primera réplica del nivel *testigo*, la segunda unidad experimental será la primera réplica de la *dosis Q*, hasta la última unidad experimental, a la que le corresponde la *dosis R* y la réplica 12.

A continuación presentaremos un ejemplo de un experimento de peso al sacrificio de novillos sometidos a tres tratamientos para desarrollar el análisis y diagnóstico de un diseño completamente aleatorizado con efecto fijo en R-project.

### 2.3.2. Metodología para el desarrollo, análisis y diagnóstico del modelo

Tomaremos como ejemplo un experimento en un diseño completamente aleatorizado para evaluar el peso al sacrificio de novillos sometidos a tres tratamientos (*testigo*, *dosis Q* y *dosis R*). Los niveles del efecto de tratamiento fueron considerados de forma determinística (efecto fijo), cuyo modelo está representado por:

$$y_{ij} = \mu + \tau_i + \varepsilon_{j(i)}$$

Donde  $y_{ij}$  es el peso del  $j$ -ésimo novillo sometido al  $i$ -ésimo nivel del tratamiento,  $\mu$  indica el promedio del peso de los novillos,  $\tau_i$  indica el efecto promedio del  $i$ -ésimo nivel de tratamiento en el  $j$ -ésimo novillo, y  $\varepsilon_{j(i)}$  es el error experimental asociado a la  $j$ -ésima replicación, anidada en la  $i$ -ésima dosis.

Las hipótesis de este modelo son:

$$H_0 : \tau_{testigo} = \tau_{dosis Q} = \tau_{dosis R}, Y$$

$H_a$  : los promedios de los niveles del tratamiento son diferentes

Se consideró el proceso de aleatorización presentado anteriormente y los resultados encontrados fueron:

Información de machos al sacrificio							
Animal	Réplica	Tratamiento	Peso	Animal	Réplica	Tratamiento	Peso
1	1	testigo	450	19	7	testigo	450
2	1	dosis Q	449	20	6	dosis R	450
3	2	dosis Q	451	21	7	dosis R	449
4	3	dosis Q	453	22	8	testigo	453
5	2	testigo	450	23	9	testigo	452
6	3	testigo	449	24	10	testigo	451
7	4	dosis Q	451	25	11	testigo	450
8	4	testigo	453	26	8	dosis Q	449
9	1	dosis R	449	27	8	dosis R	450
10	2	dosis R	452	28	9	dosis R	448
11	5	testigo	452	29	9	dosis Q	450
12	6	testigo	451	30	10	dosis Q	448
13	5	dosis Q	450	31	10	dosis R	451
14	6	dosis Q	449	32	11	dosis Q	451
15	3	dosis R	450	33	12	testigo	453
16	4	dosis R	448	34	11	dosis R	452
17	7	dosis Q	450	35	12	dosis Q	447
18	5	dosis R	450	36	12	dosis R	453

Generemos la hoja de datos con la información anterior en R-project:

```

R Studio

Peso=c(450, 449, 451, 453, 450, 449, 451, 453, 449, 452, 452, 451, 450, 449,
      450, 448, 450, 450, 450, 450, 449, 453, 452, 451, 450, 449, 450, 448, 450,
      448, 451, 451, 453, 452, 447, 453)
Base= cbind(Generar, Peso)
Base

plots  r Tratamiento Peso
      1 1   testigo  450
      2 1   dosis Q  449
      .
      36 12  dosis R  453
    
```

### 2.3.3. Análisis exploratorio de la hoja de datos

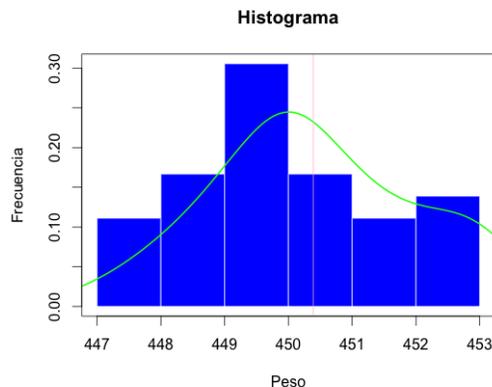
Inicialmente es necesario realizar una inspección general de los datos, para verificar incoherencias como errores en la digitación de datos o en la información recolectada en campo. Para tal fin, podemos utilizar histogramas, gráficos Boxplot, hojas y tallos, qqplot, entre otros.

A continuación se presenta la programación en R-project que permite generar un histograma de frecuencias para observar la distribución empírica de los datos mediante el comando *hist*. En este gráfico se incluirá la curva de densidad de probabilidad muestral de los datos y una línea vertical que indica el promedio aritmético del peso de los animales del experimento.

```

R Studio

hist(Base$Peso,probability=T,breaks="Sturges",col="blue", main="Histograma",
      xlab="Peso",ylab="Frecuencia",border="white")
lines(density(Base$Peso),lwd=2,col="green")
abline(v=mean(Base$Peso),col="pink")
    
```

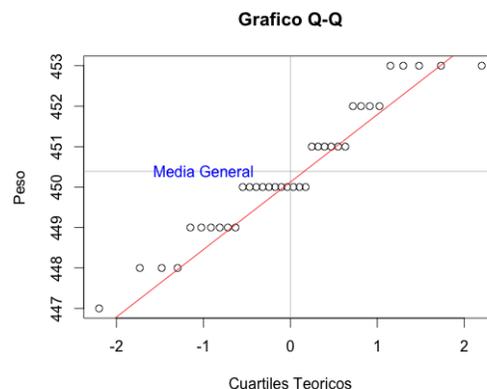


En el histograma de frecuencia se formaron seis grupos, con valor mínimo en el eje  $x$  de 447 kg y máximo de 453 kg y con una media de 450.4 kg (línea rosada del primer gráfico). Podemos observar que el pico de la línea verde está un poco a la izquierda de la línea roja (la media) y la campana que forma la línea verde descende más en el lado izquierdo que en el lado derecho, lo que demuestra que los datos tienen una ligera asimetría positiva (o a la derecha), lo cual no es característico de la función densidad de probabilidad de una distribución normal, pero esto se debe comprobar con una prueba de normalidad, como se indicará próximamente.

Un método gráfico más revelador para determinar si el peso de los animales sigue una distribución normal es el gráfico QQ de cuantiles, que permite observar el ajuste de los datos a la distribución normal, mediante el comando `qqnorm`; veamos:



```
qqnorm(Base$Peso,main="Grafico Q-Q",xlab="Cuartiles Teoricos", ylab="Peso")
abline(h=mean(Base$Peso), v=0,col="gray")
text(-1,mean(Base$Peso),"Media General",col="blue")
qqline(Base$Peso,col="red")
```

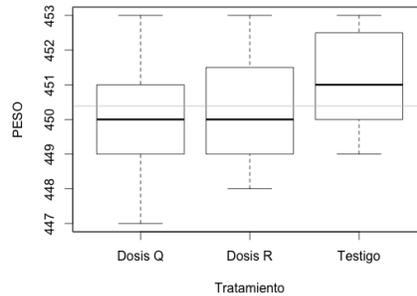


Al observar el gráfico anterior, se puede apreciar que los datos contenidos entre el primer y el tercer cuartil se ajustan adecuadamente a la línea roja que indica la normalidad de un conjunto de datos. Cabe resaltar que en los extremos se observa una ligera distorsión en las colas de la distribución. En estos casos se recomienda la implementación de pruebas de normalidad formales, como Shapiro-Wilk o Kolmogorov-Smirnov.

A continuación presentamos el código R-project para generar un gráfico Boxplot para tener una idea de la distribución de los datos de los diferentes tratamientos bajo estudio, mediante el comando `Boxplot`; veamos:



```
boxplot(Base$Peso~Base$Tratamiento,ylab="PESO",xlab="Tratamiento", data=Base)
abline(h=mean(Base$Peso),col="gray")
text(2,main=(Base$Peso),"Media General",col="blue",pos=2)
```



Podemos apreciar que la media de los tres tratamientos está cercana a los 450 kg, siendo el testigo con un valor promedio ligeramente superior. La *dosis Q* fue la que presentó la mayor variación (dispersión) en sus valores muestrales.

Uno de los principales objetivos de este análisis exploratorio es detectar valores extremos o sospechosos que no coincidan con la realidad biológica de la variable, como los pesos al sacrificio de 200 kg o de 1000 kg, o con lo observado por el investigador.

### 2.3.4. Contraste de interés (efecto de los tratamientos)

Ahora procederemos a realizar el análisis de varianza, donde la hipótesis a probar es:

$$H_0 : \tau_{testigo} = \tau_{dosis Q} = \tau_{dosis R}$$

Suma de cuadrados totales (*SCT*):

$$SCT = \sum_{i=1}^3 \sum_{j=1}^{12} (y_{ij})^2 - \frac{\left( \sum_{i=1}^3 \sum_{j=1}^{12} y_{ij} \right)^2}{36}$$

$$SCT = ((450)^2 + (449)^2 + \dots + (453)^2) - \frac{(450+449+\dots+453)^2}{36}$$

$$SCT = 7302694 - \frac{262893796}{36} = 7302694 - 7302605.44 = 88.55556$$

Suma de cuadrados de los tratamientos (*SCTr*):

$$SCTr = \frac{\sum_{i=1}^I \tau_i^2}{J} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J y_{ij} \right)^2}{n}$$

$$SCTr = \frac{(449+451+\dots+447)^2+(449+452+\dots+453)^2+(450+450+\dots+453)^2}{12} - \frac{(450+449+\dots+453)^2}{36}$$

$$SCTr = 7302617 - \frac{262893796}{36} = 7302617 - 7302605.44 = 11.55556$$

Suma de cuadrados del error (*SCE*):

$$SCE = SCT - SCTr$$

$$SCE = 88.555556 - 11.555556 = 77.00$$

El cuadro resumen del análisis de varianza sería:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Valor de F
Tratamiento	3 - 1 = 2	11.555556	$\frac{11.555556}{2}$	$\frac{5.7777778}{2.333333}$
Error	36 - 3 = 33	77.00	$\frac{77.00}{33}$	
Total	36 - 1 = 35	88.555556		

El análisis de varianza en R-project se puede realizar mediante el comando *aov* (Fit an Analysis of Variance Model), adaptado del paquete estadístico *S*, descrito por Chambers et al. (1992) o con el comando *lm* (fitting linear models), el cual fue implementado del paquete *S* descrito por Chambers, 1992. Los comandos *aov* y *lm* permiten realizar análisis de varianza para modelos balanceados y desbalanceados. Para este ejemplo utilizaremos el comando *aov*.

El comando *aov* requiere especificar en los argumentos el modelo a ser utilizado, incluyendo el símbolo  $\sim$  para relacionar la variable dependiente (en este caso *Peso*) con las variables independientes (en este caso *Tratamiento*). También se debe escribir el nombre de la hoja de datos antecedida de la palabra *data*.

Los resultados saldrán en una hoja de datos que llamaremos *Analisis*.

```

R Studio

Analisis=aov(Base$Peso~Base$Tratamiento,data=Base)
Analisis

Call:
  aov(formula = Base$Peso ~ Base$Tratamiento, data = Base)
Terms:
      Base$Tratamiento Residuals
Sum of Squares      11.55556  77.00000
Deg. of Freedom         2        33
Residual standard error: 1.527525
Estimated effects may be unbalanced
    
```

Solicitaremos un resumen para mirar los resultados con el comando *summary*:

```

R Studio

summary (Analisis)

      Df Sum Sq Mean Sq F value Pr(>F)
Base$Tratamiento  2  11.56   5.778   2.476 0.0995 .
Residuals        33  77.00   2.333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

De los valores obtenidos se puede observar que el valor de  $p$  es mayor al nivel significativo establecido del 5% (0.05), con una confianza del 95%. Aceptamos la hipótesis nula de que ninguno de los tratamientos tiene un efecto significativo sobre el peso de los animales al sacrificio, y por lo tanto no realizaremos pruebas de comparación de medias. Expondremos otro ejemplo en el siguiente capítulo, donde aplicaremos algunas de ellas.

### 2.3.5. Validación de los supuestos del modelo

En un diseño experimental completamente aleatorizado, los errores experimentales deben tener distribución normal, con media cero y varianza homogénea:  $\epsilon_{j(i)} \sim N(0, \sigma_{\epsilon}^2)$ .

Para verificar si el modelo cumple con los supuestos previamente enunciados, se procede a estimar los residuos:

Si:  $y_{ij} = \mu + \tau_i + \epsilon_{j(i)}$ , entonces:

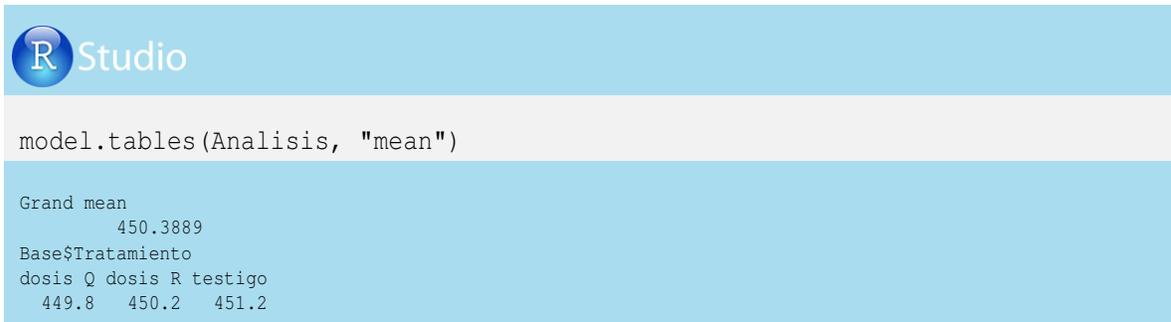
$$\epsilon_{j(i)} = y_{ij} - \mu - \tau_i$$

Al reemplazar el efecto promedio de cada tratamiento ( $\hat{\tau}_i$ ) y el efecto promedio general ( $\hat{\mu}$ ), se obtienen los residuales estimados del modelo ajustado:

$$\hat{\epsilon}_{j(i)} = y_{ij} - \hat{\mu} - (\hat{\tau}_i - \hat{\mu})$$

$$\hat{\epsilon}_{j(i)} = y_{ij} - \hat{\tau}_i$$

Para nuestro ejemplo, obtengamos los promedios de cada tratamiento mediante el comando `model.tables` y los argumentos: nombre de la hoja de datos que tiene el análisis de varianza (en nuestro caso *Analisis*) y la palabra *mean*, la cual genera el promedio general y los promedios de los tratamientos; veamos:



```

R Studio

model.tables(Analisis, "mean")

Grand mean
      450.3889
Base$Tratamiento
dosis Q dosis R testigo
  449.8   450.2   451.2

```

Los promedios estimados fueron:

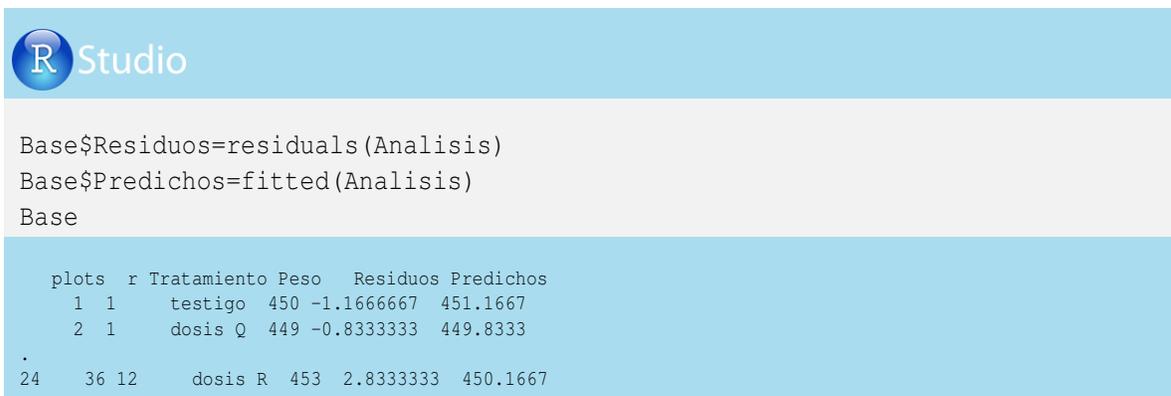
$$\hat{\mu} = 450,39 \text{ kg,}$$

$$\hat{\tau}_{testigo} = 451.1667 \text{ kg,}$$

$$\hat{\tau}_{dosis Q} = 449.833 \text{ kg, y}$$

$$\hat{\tau}_{dosis R} = 450.1667 \text{ kg}$$

Empleando el comando *residuals* de R-project, seguido del nombre de la hoja de datos con los resultados del modelo ajustado, podemos obtener los residuales, y con el comando *fitted* podemos obtener el valor esperado del peso de cada animal por tratamiento. En la programación de R-project que se presenta a continuación incluimos el vector de los residuales y de los valores ajustados en la hoja de datos *Base*; veamos:



```

R Studio

Base$Residuos=residuals(Analisis)
Base$Predichos=fitted(Analisis)
Base

plots r Tratamiento Peso  Residuos Predichos
  1  1   testigo  450 -1.1666667  451.1667
  2  1   dosis Q  449 -0.8333333  449.8333
.
24  36 12   dosis R  453  2.8333333  450.1667

```

El residuo de la primera observación, que corresponde al tratamiento *testigo*, es:

$$\hat{\epsilon}_{1(testigo)} = y_{testigo,1} - \hat{\tau}_{testigo}$$

$$\hat{\epsilon}_{1(testigo)} = 450 - 451.1667 = -1.166667$$

El residuo de la última observación, que corresponde al tratamiento *dosis R*, es:

$$\begin{aligned} \hat{\epsilon}_{36(dosis R)} &= y_{dosis R,36} - \hat{\tau}_{dosis R} \\ \hat{\epsilon}_{36(dosis R)} &= 453 - 450.1667 = 2.8333 \end{aligned}$$

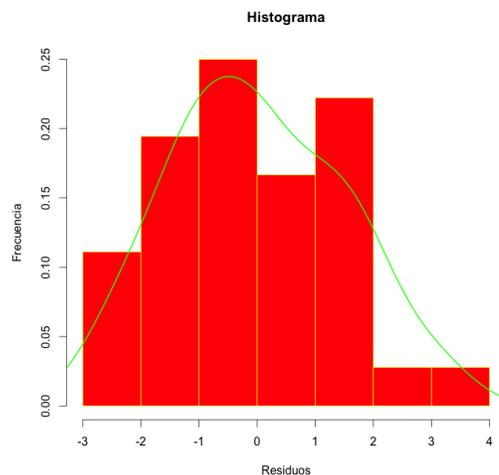
Los residuos de las otras observaciones se obtienen de la misma forma de cálculo; dejamos al lector esta tarea.

Para convalidar los supuestos estadísticos asociados al modelo de clasificación experimental, debemos realizar pruebas de normalidad y homogeneidad de los errores. Antes de efectuar dichas pruebas, realizaremos un análisis exploratorio de los residuales del modelo empleando un histograma de frecuencias para determinar la distribución empírica de los datos. Utilizaremos el comando *hist* y los argumentos en su orden: nombre de la variable que contiene los residuos, *probability = TRUE* para generar un histograma con las probabilidades, *breaks = "Sturges"* para el número de clases y otros argumentos relacionados con los colores y los títulos del gráfico. A este gráfico le incluiremos la curva de densidad de probabilidad empírica de los residuos con los comandos *lines* y *density* y los argumentos: nombre de la variable, *lwd* para el grosor de la línea y *col* para el color; veamos:

```

R Studio

hist(Base$Residuos,probability=T,breaks="Sturges",col="red",main="Histograma",
      xlab="Residuos",ylab="Frecuencia")
lines(density(Base$Residuos),lwd=2,col="green")
    
```



Los residuos estuvieron entre  $-3$  y  $4$ , con una mayor frecuencia de valores cercanos a  $0$ , y se aprecia una aparente asimetría en los residuos.

A continuación presentaremos algunas medidas resumen de los residuos:



```
summary(Base$Residuos)
var(Base$Residuos)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
> var(Base\$Residuos)	-2.8330	-1.1670	-0.1667	0.0000	1.1670	3.1670
[1]	2.2					

De los resultados presentados, es de interés observar que el promedio de los residuales es 0 y su varianza es 2.2. En otras palabras,  $\epsilon_{ij}$  tiene  $\mu_\epsilon = 0$  y  $\sigma_\epsilon^2 = 2.2$ .

Con el propósito de determinar si los residuos tienen distribución normal, es necesario realizar pruebas de normalidad, como Shapiro-Wilk, Kolmogorov-Smirnov o Anderson-Darling. Nosotros aplicaremos la prueba de Shapiro-Wilk con el comando *shapiro.test*; veamos:



```
ShapiroResiduos=shapiro.test(Base$Residuos)
ShapiroResiduos
```

```
Shapiro-Wilk normality test
data: Base$Residuos
W = 0.9699, p-value = 0.4215
```

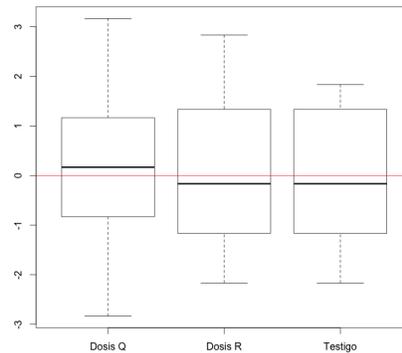
En la tabla anterior se puede observar que el valor de  $p$  es mayor al nivel de significancia del 5 %, que permite aceptar con una confianza del 95 % la hipótesis nula de que los residuos siguen una distribución normal.

Previamente se enunció que uno de los supuestos del modelo implementado es la homogeneidad de varianzas entre los tratamientos bajo estudio. Para validar dicho supuesto, existen pruebas de hipótesis formales, entre las cuales resaltan las de Cochran, Bartlett y Fligner-Killeen. Antes de implementar estas pruebas, realizaremos un análisis exploratorio de la distribución de los residuales condicionados por el tratamiento. Para tal fin, emplearemos un gráfico de cajas.

Utilizamos el comando *boxplot* con los argumentos: nombre de la variable a graficar, seguida del signo  $\sim$  y el nombre de la variable que agrupa, y el argumento *data* = con el nombre de la hoja de datos. También incluiremos una línea con el comando *abline* que horizontalmente muestre el valor medio de los residuos y que sea de color rojo; veamos:



```
boxplot(Base$Residuo~Base$Tratamiento, data=Base)
abline(h=mean(Base$Residuo), col="red")
```



De la gráfica resultante se puede observar que la media de los residuos por tratamiento fue cercana a 0. Otro detalle importante por revisar es el grado de similitud de la amplitud de cajas, lo cual nos indica si la varianza de los residuos por tratamiento es aproximadamente igual. De la figura anterior se puede apreciar que los residuales de la *dosis Q* presentaron una menor variabilidad en el 50% de los datos más centrales, pero también las máximas desviaciones en peso, respecto al valor esperado (límite inferior y límite superior).

A continuación presentaremos otros gráficos útiles para el análisis exploratorio. Emplearemos la librería *lattice* (Deepayan, 2008) para generar un histograma de frecuencias (mediante el comando *histogram*) y un gráfico de densidad muestral, clasificados por tratamientos con el comando *densityplot*.

Los argumentos incluidos en el comando *histogram* en esta oportunidad son:

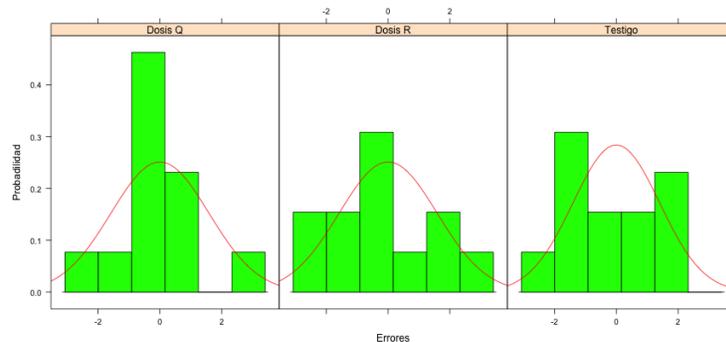
- ♣ El símbolo  $\sim$  seguido del nombre de la variable a graficar
- ♣ El signo  $|$  para indicar que la variable a graficar será agrupada (en este caso los residuos serán agrupados por tratamiento)
- ♣ El argumento *data* con el nombre de la hoja de datos
- ♣ Los argumentos *xlab* y *ylab* para los títulos de los ejes
- ♣ El argumento *col* para el color de las barras del gráfico
- ♣ El argumento *type* con la palabra *density* para generar un histograma de densidad de probabilidad de los datos

- ♣ Por último, el argumento *panel*, seguido de *= function* para crear una función que genera pantallas personalizadas del gráfico, como es el caso de *panel.histogram* y *panel.mathdensity*, que en este caso son utilizadas para generar un gráfico con los valores medios y una línea con la densidad de probabilidad para cada tratamiento.

La programación en R-project sería:



```
library(lattice)
histogram(~Base$Residuos|Base$Tratamiento,data=Base,col="green",xlab="Errores",
  ylab="Probabilidad", type="density",
  panel=function(x,...)
  {
    panel.histogram(x,...)
    panel.mathdensity(dmath=dnorm,col="red",
      args=list(mean=mean(x),sd=sd(x)))
  })
```

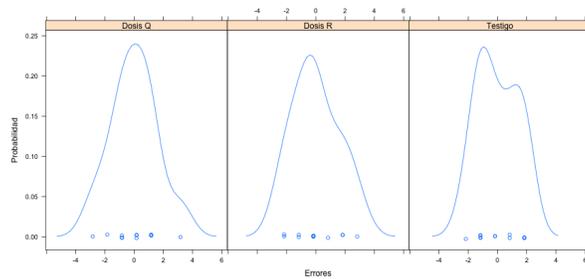


Se aprecia gráficamente que existen diferencias en la variabilidad de los residuos entre tratamientos, siendo la *dosis Q* visualmente diferente a los otros dos tratamientos.

Ahora veamos la implementación del comando *densityplot*, el cual permite también visualizar la variabilidad de los residuos entre tratamientos:



```
library(lattice)
densityplot(~Base$Residuos|Base$Tratamiento,ylab="Probabilidad",xlab="Errores")
```



La prueba de Cochran es utilizada para verificar homocedasticidad en experimentos balanceados y es útil cuando la varianza de algún tratamiento es mayor que otros, indicando si existe alguna varianza estadísticamente diferente (mayor) que la presentada por el resto de tratamientos. La prueba de hipótesis considerada es la siguiente:

$$H_0 : \sigma_{\epsilon_{tratamiento}}^2 = \sigma_{\epsilon_{dosis R}}^2 = \sigma_{\epsilon_{dosis Q}}^2, y$$

$H_a$  : Al menos una de las varianzas es diferente a las otras.

Aceptamos  $H_0$  cuando  $p > 0.05$ , lo que nos indica que las varianzas de los errores de los tratamientos son homogéneas.

Realicemos esta prueba en R-project con el comando `cochran.test` de la librería `outliers` de Komsta (2011). Este comando requiere que el acceso de los nombres de las variables sea directo, sin anteponer el nombre de la hoja de datos, por ejemplo `Residuos` y no `Base$Residuos`. Para esto utilizaremos el comando `attach`; veamos la programación para realizar la prueba de Cochran:

```

R Studio

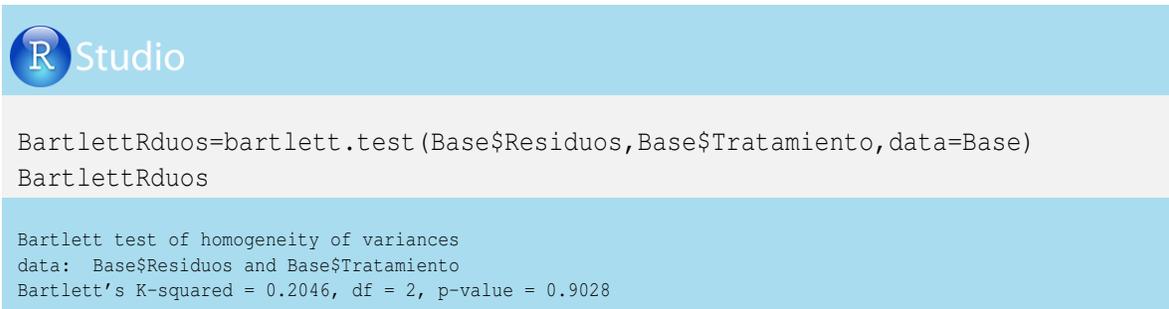
attach (Base)
library (outliers)
CochranRduos=cochran.test (Residuos~Tratamiento, data=Base)
CochranRduos

Cochran test for outlying variance
data: Residuos ~ Tratamiento
C = 0.3593, df = 12, k = 3, p-value = 1
alternative hypothesis:
Group dosis Q has outlying variance Group dosis R has outlying variance
sample estimates:
dosis Q dosis R Testigo
2.515152 2.515152 1.969697
    
```

Por presentar un  $p > 0.05$  (no significativo), aceptamos la hipótesis  $H_0$ . Las varianzas de los errores de los tratamientos son homogéneas, a pesar de visualizar que el tratamiento *dosis Q* presentó mayor varianza de los residuos. A continuación se presentan otras pruebas de homogeneidad de varianzas con el fin de confirmar el resultado de la prueba de Cochran.

La prueba de Bartlett se usa para verificar homocedasticidad en experimentos balanceados y desbalanceados. Se afirma que los tratamientos presentan homogeneidad de varianzas si el valor de  $p > 0.05$  (no significativo). Sin embargo, esta prueba puede estar influenciada por desajustes de las muestras a la distribución normal. Por lo tanto, antes de utilizarla es necesario verificar si los datos tienen distribución normal.

En R-project utilizaremos el comando `bartlett.test`, indicando en los argumentos la variable a analizar (en nuestro caso `Base$Residuos`, el efecto clasificatorio (`Base$Tratamiento`) y la hoja de datos donde se tienen estas variables). La programación en R-project es:



```

R Studio

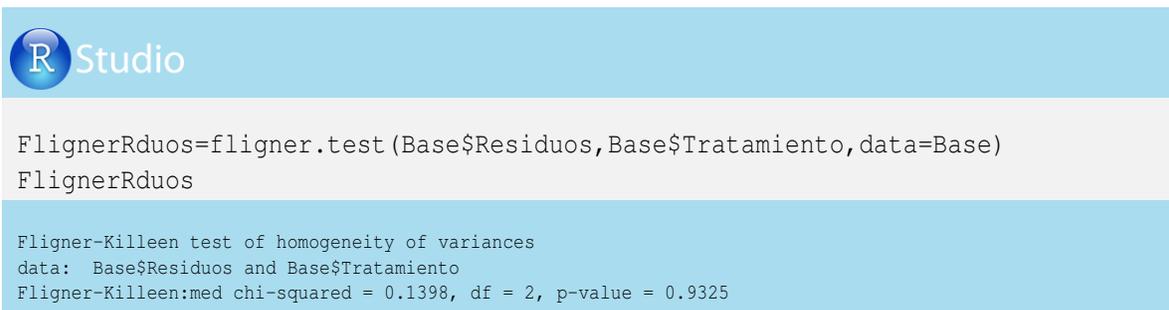
BartlettRduos=bartlett.test(Base$Residuos,Base$Tratamiento,data=Base)
BartlettRduos

Bartlett test of homogeneity of variances
data: Base$Residuos and Base$Tratamiento
Bartlett's K-squared = 0.2046, df = 2, p-value = 0.9028

```

Según la prueba de homogeneidad de varianzas de Bartlett, se observa que el  $p > 0.05$ , por lo cual no se rechaza la hipótesis nula (los residuos son homogéneos entre tratamientos).

La prueba de Fligner-Killeen es utilizada para verificar homocedasticidad y es recomendada cuando el supuesto de normalidad no se cumple. Se afirma que los tratamientos presentaron homogeneidad de varianzas si el valor de  $p > 0.05$  (no significativo).



```

R Studio

FlignerRduos=fligner.test(Base$Residuos,Base$Tratamiento,data=Base)
FlignerRduos

Fligner-Killeen test of homogeneity of variances
data: Base$Residuos and Base$Tratamiento
Fligner-Killeen:med chi-squared = 0.1398, df = 2, p-value = 0.9325

```

Al igual que las pruebas de Cochran y Bartlett, la prueba de Fligner-Killeen mostró que los residuos son homogéneos entre tratamientos.

En conclusión, se puede decir que los residuos del experimento presentaron una distribución normal y existió homogeneidad de los errores entre tratamientos; dicho de otra forma:  $\epsilon_{ij}$  tiene  $\mu_{\epsilon} = 0$  y  $\sigma_{\epsilon}^2 = 2.2$

En este ejemplo no fue necesario realizar transformaciones de los datos. En caso de necesitar la transformación de datos, puede realizarse con las transformaciones que se indican en el capítulo 7 de este libro.



## Capítulo 3

# Pruebas de comparación de efectos promedios

### 3.1. Generalidades

Cuando se lleva a cabo un análisis de la varianza y se detecta que existen diferencias estadísticas en un factor considerado en el modelo, se procede a realizar una prueba estadística de comparación de medias, con el fin de detectar dónde realmente está la diferencia entre los niveles de ese factor.

Existen diferentes pruebas de comparación de efectos promedio, dentro de las cuales se pueden citar: LSD (diferencia mínima significativa), Bonferroni, Scheffé, Students-Newman-Keuls, Tukey, Duncan, Dunnet, contrastes ortogonales, entre otras. A continuación veremos algunas de ellas:

### 3.2. Prueba de comparación de Tukey

Esta se conoce como la prueba de la honestidad y fue ideada por John Tukey; se clasifica dentro de las pruebas de comparación no planeadas o a posteriori. En el caso de un factor ( $\tau$ ), donde hemos rechazado la hipótesis nula ( $H_{0\tau} : \hat{\tau}_1 = \hat{\tau}_2 = \dots = \hat{\tau}_k$ ), se buscaría detectar la presencia de diferencias significativas entre las medias de pares de niveles de ese factor, previamente organizados de mayor a menor promedio ( $\hat{\tau}_{mayor} \dots \dots \hat{\tau}_{menor}$ ). Inicialmente se calcula el valor  $q$  para cada comparación de tratamientos, el cual está dado por:

$$q_{\tau_{mayor}:\tau_{menor}} = \frac{\hat{\tau}_{mayor} - \hat{\tau}_{menor}}{\sqrt{\frac{CM_{\epsilon}}{n}}}$$

Lo anterior indica que la significancia estadística de la comparación de niveles de un tratamiento está dada por la diferencia de promedios, el cuadrado medio del error ( $CM_{\epsilon}$ ) y el número de datos dentro de cada nivel ( $n$ ).

El valor resultante será comparado con la diferencia mínima necesaria o diferencia honestamente significativa ( $DHS$ ), calculado con el valor del rango studentizado ( $q_{tabla}$ ), el cuadrado medio del error ( $CM_{\epsilon}$ ) y el número de replicaciones dentro de cada grupo( $n$ ):

$$DHS = q_{tabla} \sqrt{\frac{CM_{\epsilon}}{n}}$$

El valor de  $q_{tabla}$  está determinado por el nivel de significancia deseado ( $\alpha$ ), los grados de libertad del error ( $g.l_e$ ) y el número de niveles del efecto ( $n$ ).

Si tenemos el valor de  $q_{\tau_{mayor}:\tau_{menor}}$  y el valor de DHS, decimos que existieron diferencias significativas entre los dos niveles comparados, cuando:

$$q_{\tau_{mayor}:\tau_{menor}} > DHS$$

En el caso de niveles con diferente número de replicaciones (pero no muy diferentes) se debería tener en cuenta la media armónica ( $n^*$ ), la cual está dada por:

$$n^* = \frac{K}{\frac{1}{n_{\tau_1}} + \frac{1}{n_{\tau_2}} + \dots + \frac{1}{n_{\tau_k}}}$$

Si el número de replicaciones es muy diferente, se utilizaría la media armónica del número de replicaciones de los niveles comparados:

$$n^* = \frac{2}{\frac{1}{n_{\tau_{mayor}}} + \frac{1}{n_{\tau_{menor}}}}$$

Cuando la diferencia del número de replicaciones es muy grande, se escoge como  $n^*$  el valor mínimo de replicaciones.

Retomemos el ejemplo de los tres tratamientos del diseño completamente aleatorizado (del capítulo anterior), con algunas modificaciones en los valores de peso de los animales y con el último dato perdido. La hoja de datos es:

Información de machos al sacrificio							
Animal	Réplica	Tratamiento	Peso	Animal	Réplica	Tratamiento	Peso
1	1	testigo	455	19	7	testigo	450
2	1	dosis Q	449	20	6	dosis R	450
3	2	dosis Q	451	21	7	dosis R	449
4	3	dosis Q	453	22	8	testigo	453
5	2	testigo	450	23	9	testigo	452
6	3	testigo	449	24	10	testigo	451
7	4	dosis Q	451	25	11	testigo	450
8	4	testigo	453	26	8	dosis Q	449
9	1	dosis R	449	27	8	dosis R	450
10	2	dosis R	452	28	9	dosis R	448
11	5	testigo	452	29	9	dosis Q	450
12	6	testigo	451	30	10	dosis Q	448
13	5	dosis Q	450	31	10	dosis R	451
14	6	dosis Q	449	32	11	dosis Q	451
15	3	dosis R	450	33	12	testigo	455
16	4	dosis R	448	34	11	dosis R	452
17	7	dosis Q	450	35	12	dosis Q	447
18	5	dosis R	450	36	12	dosis R	.

La variable a analizar es el peso al sacrificio de los animales sometidos a los tres tratamientos, cuyo modelo está representado por:

$$y_{ij} = \mu + \tau_i + \varepsilon_{j(i)}$$

Donde  $y_{ij}$  es el peso del  $j$ -ésimo novillo sometido a la  $i$ -ésima dosis,  $\mu$  indica el efecto promedio del peso de los novillos,  $\tau_i$  indica el efecto promedio de la  $i$ -ésima dosis y  $\varepsilon_{j(i)}$  es el error experimental asociado con la  $j$ -ésima replicación, anidada en la  $i$ -ésima dosis.

Este experimento es desbalanceado porque se perdió una unidad experimental a la cual se le aplicó la *dosis R*.

Las hipótesis de este modelo son:

$H_0 : \tau_{tratamiento} = \tau_{dosis Q} = \tau_{dosis R}, Y$

$H_a : \text{los promedios de los niveles del tratamiento son diferentes}$

Generamos la hoja de datos llamada *Hoja*, con las variables *Animal*, *Tratamiento*, *Replica* y *Peso*:

```

R Studio

Animal= c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,
          20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35,36)
Tratamiento=c(
"testigo", "dosis Q", "dosis Q", "dosis Q", "testigo", "testigo",
"dosis Q", "testigo", "dosis R", "dosis R", "testigo", "testigo",
"dosis Q", "dosis Q", "dosis R", "dosis R", "dosis Q", "dosis R",
"testigo", "dosis R", "dosis R", "testigo", "testigo", "testigo",
"testigo", "dosis Q", "dosis R", "dosis R", "dosis Q", "dosis Q",
"dosis R", "dosis Q", "testigo", "dosis R", "dosis Q", "dosis R")
Replica=c(1, 1, 2, 3, 2, 3, 4, 4, 1, 2, 5, 6, 5, 6, 3, 4, 7, 5, 7, 6, 7, 8, 9,
          10, 11, 8, 8, 9, 9, 10, 10, 11, 12, 11, 12, 12)
Peso=c(455,449,451,453,450,449,451,453,449,452,452,451,450,449,450,448,450,450,
        450,450,449,453,452,451,450,449,450,448,450,448,451,451,455,452,447, NA)
Hoja=data.frame (Animal, Tratamiento, Replica, Peso)

```

Eliminamos de la hoja de datos el registro que no tiene información de peso (*NA*), con la función *na.omit*:

```

R Studio

Base=na.omit (Hoja)
Base

  Animal Tratamiento Replica Peso
1      1      testigo         1 455
2      2      dosis Q         1 449
.
35     35      dosis Q        12 447

```

Realicemos el análisis de varianza con el comando *aov*, veamos un resumen del análisis con el comando *summary* y obtengamos los promedios con el comando *model.tables* y el argumento “*mean*”.

```

R Studio

Analisis=aov (Base$Peso~Base$Tratamiento,data=Base)
summary (Analisis)
model.tables (Analisis, "mean")

> summary (Analisis)
              Df Sum Sq Mean Sq F value Pr(>F)
Base$Tratamiento  2  27.92  13.959   5.029 0.0126 *
Residuals        32  88.83   2.776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables (Analisis, "mean")
Tables of means
Grand mean
450.5143
Base$Tratamiento
  dosis Q dosis R testigo
    449.8  449.9  451.7
rep    12.0   11.0   12.0

```

Observamos que el efecto del tratamiento fue significativo ( $p < 0.05$ ) y procedemos a aplicar la prueba de Tukey para los tres tratamientos. Tengamos en cuenta que el número de repeticiones fue diferente (11 para la *dosis R* y 12 para el *testigo* y la *dosis Q*); por lo tanto, debemos inicialmente calcular la media armónica ( $n^*$ ) del número de réplicas por tratamiento:

$$n^* = \frac{3}{\frac{1}{n_{\text{testigo}}} + \frac{1}{n_{\text{dosis R}}} + \frac{1}{n_{\text{dosis Q}}}} = \frac{3}{\frac{1}{12} + \frac{1}{11} + \frac{1}{12}} = 11.647$$

Luego procedemos a encontrar la diferencia honestamente significativa (DHS) para un nivel de significancia del 5% ( $\alpha = 0.05$ ), 3 tratamientos, 32 grados de libertad del error y un  $CM_{\epsilon} = 2.776$ :

$$DHS = q_{\text{tabla}} \sqrt{\frac{CM_{\epsilon}}{n^*}}$$

$$DHS = q_{(\alpha=0.05,3,32)} \sqrt{\frac{2.776}{11.647}} = 3.48 \sqrt{\frac{2.776}{11.647}} = 1.6965$$

Ahora realicemos el cálculo para verificar si existen diferencias estadísticas entre los tratamientos, teniendo en cuenta el orden según los promedios (*testigo* > *dosis R* > *dosis Q*).

Para el *testigo* y la *dosis R*

$$q_{\tau_{testigo}:\tau_{dosis R}} = \frac{\hat{\tau}_{testigo} - \hat{\tau}_{dosis R}}{\sqrt{\frac{CM_E}{n^*}}} = \frac{451.7 - 449.9}{\sqrt{\frac{2.776}{11.647}}} = 3.69$$

Para el *testigo* y la *dosis Q*

$$q_{\tau_{testigo}:\tau_{dosis Q}} = \frac{\hat{\tau}_{testigo} - \hat{\tau}_{dosis Q}}{\sqrt{\frac{CM_E}{n^*}}} = \frac{451.7 - 449.8}{\sqrt{\frac{2.776}{11.647}}} = 3.89$$

Para la *dosis R* y la *dosis Q*

$$q_{\tau_{dosis R}:\tau_{dosis Q}} = \frac{\hat{\tau}_{dosis R} - \hat{\tau}_{dosis Q}}{\sqrt{\frac{CM_E}{n^*}}} = \frac{449.9 - 449.8}{\sqrt{\frac{2.776}{11.647}}} = 0.20$$

Veamos si existieron diferencias entre los tratamientos:

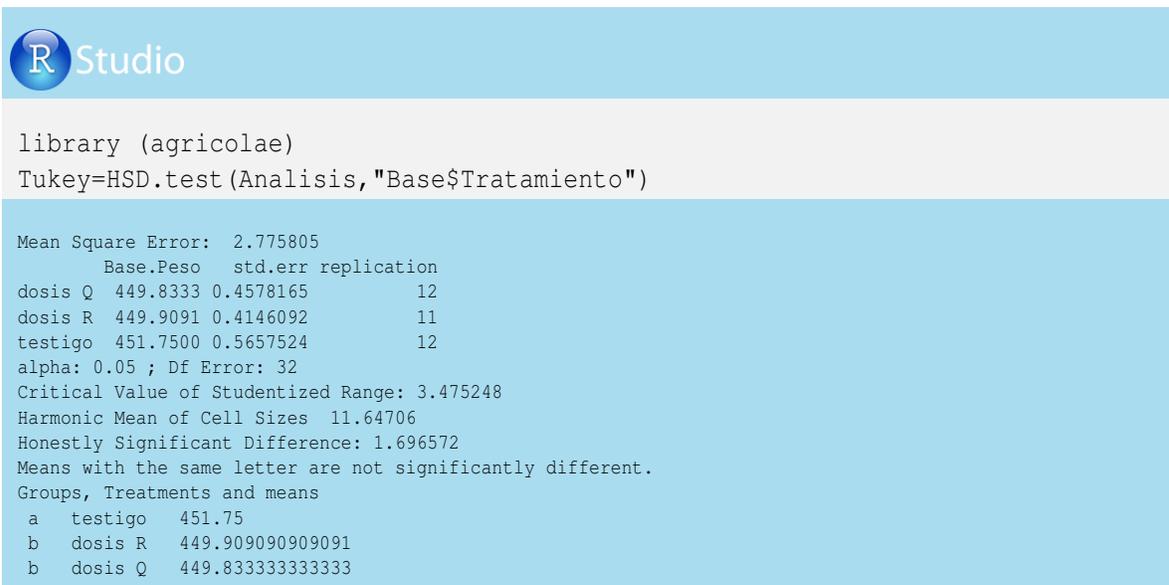
$$q_{\tau_{testigo}:\tau_{dosis R}} > DHS, (3.69 > 1.6965)$$

$$q_{\tau_{testigo}:\tau_{dosis Q}} > DHS, (3.89 > 1.6965)$$

$$q_{\tau_{dosis R}:\tau_{dosis Q}} < DHS, (0.20 < 1.6965)$$

Por lo tanto, el *testigo* fue diferente a la *dosis R* y *dosis Q* a un nivel de significancia del 5% en la prueba de Tukey. Las *dosis R* y la *dosis Q* no difirieron significativamente en la prueba de Tukey.

Veamos la programación de este ejercicio en R-project, con el comando *HSD.test* de la librería *agricolae* y comparemos los resultados con los anteriormente indicados:



```

R Studio

library (agricolae)
Tukey=HSD.test (Análisis, "Base$Tratamiento")

Mean Square Error:  2.775805
      Base.Peso  std.err replication
dosis Q  449.8333  0.4578165      12
dosis R  449.9091  0.4146092      11
testigo  451.7500  0.5657524      12
alpha: 0.05 ; Df Error: 32
Critical Value of Studentized Range: 3.475248
Harmonic Mean of Cell Sizes  11.64706
Honestly Significant Difference: 1.696572
Means with the same letter are not significantly different.
Groups, Treatments and means
a  testigo  451.75
b  dosis R  449.909090909091
b  dosis Q  449.833333333333

```

En resumen, tendríamos:

```

R Studio

Tukey

      trt  means  M  N  std.err
1 testigo 451.7500 a 12 0.5657524
2 dosis R 449.9091 b 11 0.4146092
3 dosis Q 449.8333 b 12 0.4578165
    
```

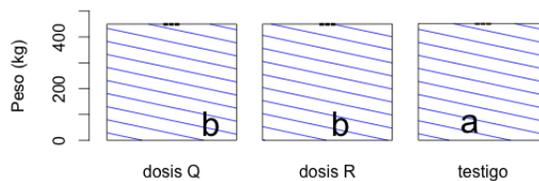
Realicemos un gráfico de comparación de promedios de tratamientos, incluyendo las letras que indican diferencias entre tratamientos según la prueba de Tukey, mediante el comando `bar.err` de la librería *agricolae*.

Iniciamos indicando la hoja de datos que contiene la información de la prueba de Tukey y los límites inferior y superior del eje *y*. Le incorporamos barras laterales con su respectiva densidad, angularidad y color y el nombre del eje *x*. El comando `text` incorpora un texto al gráfico, en este caso incluiremos las letras de significancia de la prueba de Tukey (desde el primer tratamiento hasta el tercero, 1:3); el número 66 indica la posición del texto dentro de cada barra y el argumento `expression` permite incorporar las letras de significancia para cada tratamiento en el orden que se presente en la figura:

```

R Studio

bar.err(Tukey$means, variation=c("SE"), ylim=c(0, 500), density=6, angle=168,
        col="blue", ylab="Peso (kg)")
text(1:3, 66, expression("b", "b", "a"), cex=2)
box()
    
```



En la parte superior de las barras de los tratamientos aparecen unas líneas de color negro, las cuales indican los errores estándar de los tratamientos. En este caso los errores estándar son muy pequeños en comparación con los promedios.

La prueba de Tukey también la podemos realizar con la librería *multcomp* (simultaneous inference in general parametric models) de Hothorn et al. (2008), mediante el comando `glht`

(general linear hypotheses). Este comando requiere de los argumentos: nombre de la hoja de datos que contiene el resultado del análisis de varianza, y el argumento *linfct* para especificar el tipo de hipótesis a probar, que en este caso es una comparación múltiple de clases (*mcp*), utilizando la prueba de Tukey.

Hay que tener en cuenta que el comando *glht* requiere que la hoja de datos sea un objeto para llamar directamente las variables mediante el comando *attach*:

```

R Studio

attach (Base)
Análisis=aov(Peso~Tratamiento, data=Base)
install.packages("multcomp")
library (multcomp)
a=glht (Análisis, linfct=mcp(Tratamiento= "Tukey"))

```

Cuando se utiliza el comando *summary* para obtener el resumen del análisis realizado con el comando *glht*, por defecto se presentan los valores de *p* ajustados por el método *single-step*. El lector podrá utilizar otros métodos de ajuste de los valores de *p*, como los de Westfall, Shaffer, Bonferroni, Holm y Hochberg (mayores detalles con comando *p.adjust*).

```

R Studio

summary(a)

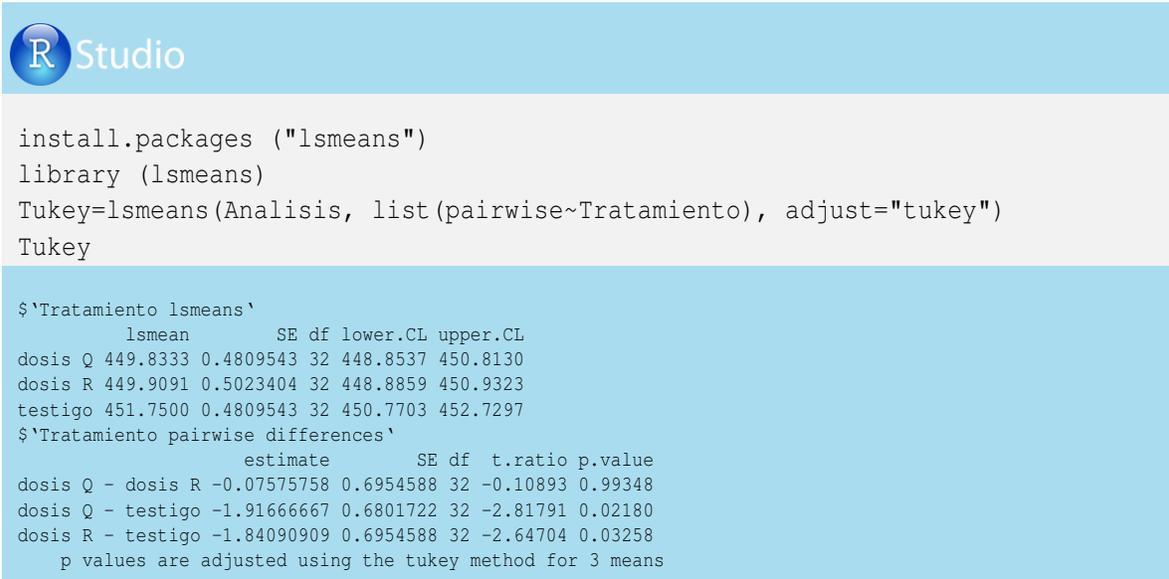
Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = Peso ~ Tratamiento, data = Base)
Linear Hypotheses:
      Estimate Std. Error t value Pr(>|t|)
dosis R - dosis Q == 0  0.07576   0.69546   0.109  0.9935
testigo - dosis Q == 0  1.91667   0.68017   2.818  0.0218 *
testigo - dosis R == 0  1.84091   0.69546   2.647  0.0326 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

```

Incluimos manualmente las letras de diferencia estadística en la comparación de medias. Vemos que la *dosis R* es semejante a la *dosis Q* y estas dos son diferentes al testigo, siendo el testigo el de mayor media. Por lo tanto, el testigo tendría la letra *a* y las dosis la letra *b*.

Otra librería muy útil para obtener las medias ajustadas (principalmente en modelos desbalanceados) y realizar comparaciones de medias es la librería *lsmeans* (least-squares means) de Lenth (2012). El comando que utilizaremos será el *lsmeans* con los argumentos: nombre de la hoja con el análisis realizado y el argumento *list* para indicar que utilizaremos la fórmula de comparación de medias por parejas de tratamientos (*pairwise*). Indicamos también que el

método de ajuste para los valores de  $p$  es la distribución de rango studentizado para la prueba de Tukey, mediante el argumento *ad just*.



```

install.packages ("lsmeans")
library (lsmeans)
Tukey=lsmeans (Análisis, list (pairwise~Tratamiento), adjust="tukey")
Tukey

$`Tratamiento lsmeans`
      lsmean      SE df lower.CL upper.CL
dosis Q 449.8333 0.4809543 32 448.8537 450.8130
dosis R 449.9091 0.5023404 32 448.8859 450.9323
testigo 451.7500 0.4809543 32 450.7703 452.7297
$`Tratamiento pairwise differences`
      estimate      SE df  t.ratio p.value
dosis Q - dosis R -0.07575758 0.6954588 32 -0.10893 0.99348
dosis Q - testigo -1.91666667 0.6801722 32 -2.81791 0.02180
dosis R - testigo -1.84090909 0.6954588 32 -2.64704 0.03258
p values are adjusted using the tukey method for 3 means
    
```

### 3.3. Mínima diferencia significativa: DMS o LSD

Consiste en llevar a cabo una prueba de hipótesis por parejas de niveles dentro de un efecto, comparando el efecto promedio de las mismas con base en la distribución  $t$ .

Este método debe aplicarse cuando se ratifique la significancia estadística en el análisis de la varianza asociado con la fuente de variabilidad del efecto. La comparación entre dos niveles sería significativa si:

$$|\hat{\tau}_{i_1} - \hat{\tau}_{i_2}| > LSD$$

El *LSD* (last significant difference) está determinado por un valor de la distribución de  $t$  de Student de doble cola, con un  $\alpha$  específico (generalmente es el 5%), el cuadrado medio del error del análisis de varianza ( $CM_\epsilon$ ), los grados de libertad del error ( $g.l_\epsilon$ ), y el número de réplicas dentro de cada nivel del efecto ( $n$ ).

Si el número de réplicas es diferente para cada tratamiento, se aplica la media armónica del número de réplicas ( $n^*$ ), así como se indicó para la prueba de Tukey; por tanto:

$$LSD = t_{(\alpha; g.l_\epsilon)} \sqrt{\frac{2 g.l_\epsilon CM_\epsilon}{n}}$$

Hay que tener cuidado en la aplicación de la prueba LSD cuando existe un número relativamente grande de niveles, porque se pueden presentar falsos rechazos de la hipótesis nula.

También se pueden presentar casos en los que se acepte que todas las posibles parejas de niveles sean iguales en la prueba LSD, a pesar de que el efecto en el análisis de varianza sea

significativo. Esto se debe a que la prueba de  $F$  considera simultáneamente todas las posibles combinaciones.

Para el ejemplo del peso al sacrificio de animales sometidos a tres tratamientos, tenemos:

$$LSD = t_{(\alpha_{5\%}; g.l_{\epsilon})} \sqrt{\frac{2g.l_{\epsilon} CM_{\epsilon}}{n^*}} = 2.036933 \sqrt{\frac{2*32*2.776}{11.647}} = 1.4063$$

Ahora realicemos el cálculo para verificar la existencia de diferencias estadísticas entre tratamientos, teniendo en cuenta el orden según los promedios de *testigo*, *dosis R* y *dosis Q*:

Para el *testigo* y la *dosis R*:

$$\tau_{testigo} : \tau_{dosis R} = |451.7 - 449.9| = 1.8$$

Para el *testigo* y la *dosis Q*:

$$\tau_{testigo} : \tau_{dosis Q} = |451.7 - 449.8| = 1.9$$

Para la *dosis R* y la *dosis Q*:

$$\tau_{dosis R} : \tau_{dosis Q} = |449.9 - 449.8| = 0.01$$

Veamos si existieron diferencias entre los tratamientos:

$$\tau_{testigo} : \tau_{dosis R} > LSD, (1.8 > 1.4063)$$

$$\tau_{testigo} : \tau_{dosis Q} > LSD, (1.9 > 1.4063)$$

$$\tau_{dosis R} : \tau_{dosis Q} < LSD, (0.01 < 1.4063)$$

Por lo tanto, el *testigo* fue diferente a la *dosis R* y la *dosis Q* a un nivel de significancia del 5% en la prueba de LSD.

Realicemos el ejercicio en R-project, con el comando *LSD.test* de la librería *agricolae* y comparemos los resultados con los anteriormente indicados:

```

R Studio

attach (Base)
LSD=LSD.test (Análisis, "Base", group=TRUE)

Mean Square Error: 2.775805
Base$Tratamiento, means and individual ( 95 %) CI
      Base.Peso  std.err replication      LCL      UCL
dosis Q  449.8333 0.4578165      12 448.9008 450.7659
dosis R  449.9091 0.4146092      11 449.0646 450.7536
testigo  451.7500 0.5657524      12 450.5976 452.9024
alpha: 0.05 ; Df Error: 32
Critical Value of t: 2.036933
Least Significant Difference 1.406301
Harmonic Mean of Cell Sizes 11.64706
Means with the same letter are not significantly different.
Groups, Treatments and means
a  testigo  451.75
b  dosis R  449.909090909091
b  dosis Q  449.833333333333
    
```

En resumen, tendríamos:

```

R Studio

LSD

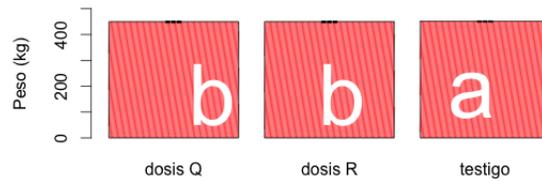
      trt  means  M  N  std.err      LCI      UCI
1 testigo 451.7500 a 12 0.5657524 450.5976 452.9024
2 dosis R 449.9091 b 11 0.4146092 449.0646 450.7536
3 dosis Q 449.8333 b 12 0.4578165 448.9008 450.7659
    
```

Apliquemos la rutina para generar un gráfico de promedios de los tratamientos y las letras que indican las diferencias estadísticas con la prueba de LSD, como se hizo en la prueba de Tukey, con algunas modificaciones en la presentación de las barras:

```

R Studio

bar.err(LSD$means, variation=c("SE"), ylim=c(0, 500), density=60, angle=268,
        col="red", ylab="Peso (kg)")
text(1:3, 166, expression("b", "b", "a"), cex=5, col="white")
    
```



En la parte superior de las barras de los tratamientos aparecen unas líneas de color negro, las cuales indican los errores estándar, que en este caso son muy pequeños en comparación con los valores promedio de los tratamientos.

### 3.4. Prueba de Duncan

El contraste de Duncan (al igual que el de Tukey) utiliza la distribución del recorrido studentizado. Se diferencia de esta prueba en que su aplicación es secuencial, en el sentido de no utilizar un único valor crítico para todas las diferencias de los promedios. Las medias son ordenadas previamente de mayor a menor ( $\hat{\tau}_{mayor}, \dots, \hat{\tau}_{menor}$ ) y el valor crítico depende del número de promedios comprendido entre las dos medias que se comparan. La diferencia significativa entre dos niveles estaría dada por:

$$\hat{\tau}_{i_1} - \hat{\tau}_{i_2} > RP$$

El  $RP$  (prueba de rangos mínimos) está determinado por el valor  $d$ . Este valor se encuentra en una tabla de Duncan y depende del nivel de significancia determinado ( $\alpha$ ), de los grados de libertad del error y del número de niveles que separan los niveles comparados, previamente ordenados ( $p$ ).

También está determinado por el cuadrado medio del error del análisis de varianza ( $CM_{\epsilon}$ ) y el número de réplicas dentro de cada nivel del efecto ( $n$ ). Si el número de réplicas es diferente para cada nivel, aplicamos la media armónica del número de réplicas ( $n^*$ ), como se muestra en la prueba de Tukey; por tanto:

$$RP = d_{(\alpha;p:g.l_{\epsilon})} \sqrt{\frac{CM_{\epsilon}}{n}}$$

Para el ejemplo del peso al sacrificio de animales sometidos a tres tratamientos, tenemos:

$$\hat{\tau}_{testigo} > \hat{\tau}_{dosis R} > \hat{\tau}_{dosis Q}$$

Por tanto, el número de niveles ( $p$ ) que separan los tratamientos sería:

$$p \text{ entre } \hat{\tau}_{testigo} : \hat{\tau}_{dosis R} = 2$$

$$p \text{ entre } \hat{\tau}_{testigo} : \hat{\tau}_{dosis Q} = 3$$

$$p \text{ entre } \hat{\tau}_{dosis R} : \hat{\tau}_{dosis Q} = 2$$

Los valores de  $d$  de la tabla de Duncan a un  $\alpha_{5\%}$ ,  $g.l_{\varepsilon} = 32$  y con  $p = 2$  o  $p = 3$  serían, respectivamente:

$$d_{(5\%:2:32)} = 2.88$$

$$d_{(5\%:3:32)} = 3.03$$

La prueba de rangos mínimos entre los tratamientos estaría dada por:

$$RP_{\tau_{testigo}:\tau_{dosisR}} = d_{(5\%:2:32)} \sqrt{\frac{2.776}{11.647}} = 2.88 \sqrt{\frac{2.776}{11.647}} = 1.406$$

$$RP_{\tau_{testigo}:\tau_{dosisQ}} = d_{(5\%:3:32)} \sqrt{\frac{2.776}{11.647}} = 3.03 \sqrt{\frac{2.776}{11.647}} = 1.48$$

$$RP_{\tau_{dosisQ}:\tau_{dosisR}} = d_{(5\%:2:32)} \sqrt{\frac{2.776}{11.647}} = 2.88 \sqrt{\frac{2.776}{11.647}} = 1.406$$

Ahora realicemos los anteriores cálculos con los datos de nuestro ejemplo, teniendo en cuenta el orden según los promedios de *testigo*, *dosis R* y *dosis Q*:

Para el *testigo* y la *dosis R*:

$$\tau_{testigo} : \tau_{dosisR} = 451.7 - 449.9 = 1.8, \text{ y } RP_{\tau_{testigo}:\tau_{dosisR}} = 1.406$$

Para el *testigo* y la *dosis Q*:

$$\tau_{testigo} : \tau_{dosisQ} = 451.7 - 449.8 = 1.9, \text{ y } RP_{\tau_{testigo}:\tau_{dosisQ}} = 1.478$$

Para la *dosis R* y la *dosis Q*:

$$\tau_{dosisR} : \tau_{dosisQ} = 449.9 - 449.8 = 0.01, \text{ y } RP_{\tau_{dosisQ}:\tau_{dosisR}} = 1.406$$

Veamos si existieron diferencias entre los tratamientos:

$$\tau_{testigo} : \tau_{dosisR} > RP, (1.8 > 1.406)$$

$$\tau_{testigo} : \tau_{dosisQ} > RP, (1.9 > 1.478)$$

$$\tau_{dosisR} : \tau_{dosisQ} < RP, (0.01 < 1.406)$$

Por lo tanto, el *testigo* fue diferente a la *dosis R* y la *dosis Q* a un nivel de significancia del 5% en la prueba de Duncan.

Realicemos el ejercicio, con el comando *duncan.test* de la librería *agricolae* y comparemos los resultados con los anteriormente indicados:



```
attach (Base)
Duncan=duncan.test (Analisis, "Tratamiento", group=TRUE)
```

```
Mean Square Error: 2.775805
      Base.Peso  std.err replication
dosis Q  449.8333 0.4578165      12
dosis R  449.9091 0.4146092      11
testigo  451.7500 0.5657524      12
alpha: 0.05 ; Df Error: 32
Critical Range
      2      3
1.406301 1.478079
Harmonic Mean of Cell Sizes 11.64706
Different value for each comparison
Means with the same letter are not significantly different.
Groups, Treatments and means
a  testigo  451.75
b  dosis R  449.909090909091
b  dosis Q  449.833333333333
```

En resumen, tendríamos:



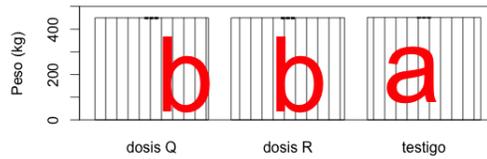
```
Duncan
```

```
      trt  means  M  N  std.err
1 testigo 451.7500 a 12 0.5657524
2 dosis R 449.9091 b 11 0.4146092
3 dosis Q 449.8333 b 12 0.4578165
```

Aplicaremos la rutina para generar un gráfico de promedios de los tratamientos y las letras que indican las diferencias estadísticas con la prueba de Duncan, como se hizo anteriormente, pero modificando la presentación:



```
bar.err(Duncan$means, variation=c("SE"), ylim=c(0, 500), density=6, angle=90,
        col="black", ylab="Peso (kg)")
text(1:3, 200, expression("b", "b", "a"), cex=7, col="red")
```



### 3.5. Prueba de Scheffé de comparaciones múltiples

En 1953, Scheffé propuso un método para elaborar un contraste ( $\Gamma$ ) entre medias asociadas a un conjunto de tratamientos.

Un contraste se define como una combinación lineal de efectos promedios ( $\hat{\tau}_i$ ) multiplicados por constantes ( $c$ ):

$$\hat{\Gamma} = c_1\hat{\tau}_1 + c_2\hat{\tau}_2 + \dots + c_I\hat{\tau}_I$$

Dicho de otra forma:

$$\hat{\Gamma} = \sum_{i=1}^I c_i\hat{\tau}_i$$

Las constantes deben obedecer a:

$$\sum_{i=1}^I c_i = 0$$

En el caso de dos niveles de un efecto, donde  $H_0 = \hat{\tau}_1 = \hat{\tau}_2$ , se tendría  $c_1 = 1$  y  $c_2 = -1$ , donde  $c_1 + c_2 = 1 + (-1) = 0$ , cuyo contraste sería:

$$\hat{\Gamma} = 1\hat{\tau}_1 + (-1)\hat{\tau}_2$$

En el caso de tres niveles de un efecto, donde  $H_0 = \hat{\tau}_1 = \hat{\tau}_2 = \hat{\tau}_3$ , se tendrían  $c_1 = 1$ ,  $c_2 = 1$  y  $c_3 = -2$ , cumpliéndose que  $c_1 + c_2 + c_3 = 1 + 1 + (-2) = 0$  o cualquier otra combinación posible de constantes que sume 0. En este caso se tendría:

$$\hat{\Gamma} = 1\hat{\tau}_1 + 1\hat{\tau}_2 + (-2)\hat{\tau}_3$$

Cada contraste ( $\hat{\Gamma}_1$  hasta  $\hat{\Gamma}_{I-1}$ ) tendrá su error estándar ( $\hat{S}$ ), que estaría dado por:

$$\hat{S} = \sqrt{\frac{CM_{\epsilon}}{n} \sum_{i=1}^I c_i^2}$$

En el caso de un número diferente de réplicas por nivel:

$$\hat{S} = \sqrt{CM_{\epsilon} \sum_{i=1}^I \frac{c_i^2}{n_i^*}}$$

Donde  $n_i^*$  es el número de observaciones del  $i$ -ésimo nivel.

Se tendrá  $\Gamma_1$  hasta  $\Gamma_{I-1}$  contrastes para un factor. Para verificar si existen diferencias para cada contraste, se requiere obtener un valor  $S_\alpha$ , el cual será comparado con el valor de  $|\hat{\Gamma}|$ .

El valor de  $S_\alpha$  tiene en cuenta el valor de  $\hat{S}$ , de los grados de libertad del efecto a analizar, y el valor de  $F$  de la distribución de Snedecor.

El valor de  $F$  requerido dependerá del nivel de significancia deseado ( $\alpha$ ), de los grados de libertad del error y del efecto ( $F_{\alpha,(I-1)g.l_\epsilon}$ ).

Por lo tanto, se tendría un  $\hat{S}_\alpha$ , dado por:

$$\hat{S}_\alpha = \hat{S} \sqrt{(I-1) F_{\alpha,(I-1)g.l_\epsilon}}$$

Si  $|\hat{\Gamma}| > \hat{S}_\alpha$ , hay diferencias estadísticas entre los niveles comparados y se rechaza  $H_0$ .

Retomando el ejemplo de este capítulo, en el cual tenemos tres tratamientos (*testigo*, *dosis R* y *dosis Q*), se generarían dos contrastes ( $I - 1 = 3 - 1 = 2$ ).

Consideremos el primer contraste, en el cual deseamos comparar el *testigo* con la *dosis R* y la *dosis Q*; entonces tendríamos:

$$\sum_{i=1}^I c_i = -2 + 1 + 1 = 0$$

El contraste estaría dado por:

$$\hat{\Gamma}_1 = -2\hat{\tau}_{testigo} + 1\hat{\tau}_{dosis Q} + 1\hat{\tau}_{dosis R} = (-2 * 451.7) + 449.83 + 449.90 = -3.67$$

El error estándar del contraste sería:

$$\hat{S}_1 = \sqrt{CM_\epsilon \sum_{i=1}^I \frac{c_i^2}{n_i^*}} = \sqrt{2.7758 \left[ \left(\frac{-2}{12}\right)^2 + \left(\frac{1}{12}\right)^2 + \left(\frac{1}{11}\right)^2 \right]} = 0.33$$

Las diferencias para los niveles de este contraste con un  $\alpha$  del 5% serían:

$$\hat{S}_\alpha = S_1 \sqrt{(I-1) F_{\alpha,(I-1)g.l_\epsilon}} = S_1 \sqrt{(3-1) F_{0.05,2,32}} = 0.33 \sqrt{2 * 3.31} = 0.856$$

Hay diferencias estadísticas para este contraste si  $|\hat{\Gamma}_1| > \hat{S}_\alpha$ . Para este caso, los valores fueron  $|-3.67| > 0.856$ , lo que indica que el tratamiento *testigo* es diferente a las dos dosis.

Consideremos ahora el segundo contraste, en el cual deseamos comparar la *dosis R* con la *dosis Q*; tendríamos:

$$\sum_{i=1}^I c_i = 0 + (-1) + 1 = 0$$

El contraste estaría dado por:

$$\hat{\Gamma}_2 = 0\hat{\tau}_{testigo} + (-1\hat{\tau}_{dosisQ}) + 1\hat{\tau}_{dosisR} = (0 * 451.7) + (-449.83) + 449.90 = 0.07$$

El error estándar del contraste sería:

$$\hat{S}_2 = \sqrt{CM_\epsilon \sum_{i=1}^I \frac{c_i^2}{n_i^*}} = \sqrt{2.7758 \left[ \left(\frac{0}{12}\right)^2 + \left(\frac{-1}{12}\right)^2 + \left(\frac{1}{11}\right)^2 \right]} = 0.198$$

Las diferencias para los niveles de este contraste con un  $\alpha$  del 5% serían:

$$\hat{S}_\alpha = \hat{S}_2 \sqrt{(I-1) F_{\alpha, (I-1)g, l_\epsilon}} = \hat{S}_2 \sqrt{(3-1) F_{0.05, 2, 32}} = 0.198 \sqrt{2 * 3.31} = 0.51$$

Para este contraste,  $|\hat{\Gamma}_1| < \hat{S}_\alpha$ , cuyos valores fueron: ( $|0.07| < 0.51$ ). Por lo tanto, no hay diferencias estadísticas entre las dosis comparadas.

En R-project no se tiene un comando satisfactorio para la prueba de Scheffé de comparaciones múltiples; sin embargo, se pueden utilizar funciones (*function*) ya existentes o construirlas específicamente para el análisis requerido. Dejamos al lector profundizar sobre el tema de programación de funciones (*function*).

### 3.6. Contrastes ortogonales

Los contrastes ortogonales son construidos antes de iniciar el experimento (a priori) y obedecen a una combinación lineal de los promedios de dos o más niveles de un efecto, como se indicó anteriormente en la prueba de Scheffé.

Si se tienen  $I$  niveles de un efecto, los contrastes ortogonales serían  $I - 1$ :

$$\hat{\Gamma}_1 = c_{1_1} \hat{\tau}_1 + c_{2_1} \hat{\tau}_2 \dots \dots + c_{I_1} \hat{\tau}_I$$

$$\hat{\Gamma}_2 = c_{1_2} \hat{\tau}_1 + c_{2_2} \hat{\tau}_2 \dots \dots + c_{I_2} \hat{\tau}_I$$

.

$$\hat{\Gamma}_{I-1} = c_{1_{I-1}} \hat{\tau}_1 + c_{2_{I-1}} \hat{\tau}_2 \dots \dots + c_{I_{I-1}} \hat{\tau}_I$$

Para que dos o más contrastes sean ortogonales, se requiere que la suma de las multiplicaciones de los coeficientes dentro de cada nivel de efecto sea cero.

$$(c_{1_1} * c_{1_2} \dots \dots * c_{1_{I-1}}) + (c_{2_1} * c_{2_2} \dots \dots * c_{2_{I-1}}) \dots \dots + (c_{I_1} * c_{I_2} \dots \dots * c_{I_{I-1}})$$

De forma general, se tendría:

$$\sum_{i=1}^I c_{i_1} c_{i_2} \dots c_{i_{I-1}} = 0$$

Veamos el ejemplo de este capítulo. Para ser ortogonales, se tendrían dos contrastes:

$$\hat{\Gamma}_1 = c_{11}\hat{\tau}_1 + c_{21}\hat{\tau}_2 + c_{31}\hat{\tau}_3$$

$$\hat{\Gamma}_2 = c_{12}\hat{\tau}_2 + c_{22}\hat{\tau}_2 + c_{32}\hat{\tau}_3$$

Una de las combinaciones de contrastes ortogonales puede ser el promedio de las dosis  $Q$  y  $R$  comparado con el *testigo* ( $\Gamma_1$ ), y la comparación entre la dosis  $Q$  y la dosis  $R$  ( $\Gamma_2$ ):

$$\hat{\Gamma}_1 = 0.5\hat{\tau}_{dosis\ Q} + 0.5\hat{\tau}_{dosis\ R} - 1\hat{\tau}_{testigo}$$

$$\hat{\Gamma}_2 = 1\hat{\tau}_{dosis\ Q} - 1\hat{\tau}_{dosis\ R} + 0\hat{\tau}_{testigo}$$

Los coeficientes para el primer contraste son:

$$c_{1\ dosis\ Q} = 0.5, c_{2\ dosis\ R} = 0.5 \text{ y } c_{3\ testigo} = -1$$

Los coeficientes para el segundo contraste son:

$$c_{1\ dosis\ Q} = 1, c_{2\ dosis\ R} = -1 \text{ y } c_{3\ testigo} = 0$$

Los contrastes serían ortogonales si:

$$\sum_{i=1}^3 c_{i1}c_{i2} = c_{11}c_{12} + c_{21}c_{22} + c_{31}c_{32} = (0.5 * 1) + (0.5 * -1) + (-1 * 0) = 0$$

Para estos dos contrastes se tendrían las siguientes hipótesis:

Para el primer contraste:

$$H_{01} : 0.5\hat{\tau}_{dosis\ Q} + 0.5\hat{\tau}_{dosis\ R} - 1\hat{\tau}_{testigo} = 0$$

$$H_{01} : \frac{\hat{\tau}_{dosis\ Q} + \hat{\tau}_{dosis\ R}}{2} = \hat{\tau}_{testigo}$$

Para el segundo contraste:

$$H_{02} : \hat{\tau}_{dosis\ Q} - \hat{\tau}_{dosis\ R} = 0$$

$$H_{02} : \hat{\tau}_{dosis\ Q} = \hat{\tau}_{dosis\ R}$$

Retomemos la información de los promedios estimados de los niveles del tratamiento:

$$\hat{\tau}_{dosis\ Q} = 449.8$$

$$\hat{\tau}_{dosis\ R} = 449.9$$

$$\hat{\tau}_{testigo} = 451.7$$

La estimación de los contrastes ( $\hat{\Gamma}$ ) del ejemplo estaría dada por:

$$\hat{\Gamma}_1 = (0.5 * 449.8) + (0.5 * 449.9) - 451.7 = -1.878$$

$$\hat{\Gamma}_2 = 449.8 - 449.9 = -0.075757576$$

Las varianzas de los contrastes estarían dadas por:

$$\hat{\sigma}_{\hat{\Gamma}}^2 = \sum_{i=1}^I \frac{c_i^2}{n_i} CM_{\epsilon}$$

En el caso de los contrastes del ejercicio, se tendrían:

Para el primer contraste:

$$\hat{\sigma}_{\hat{\Gamma}_1}^2 = \sum_{i=1}^3 \frac{c_i^2}{n_i} 2.776 = \left( \frac{0.5^2}{12} + \frac{0.5^2}{11} + \frac{-1^2}{12} \right) * 2.776 = 0.352257$$

Para el segundo contraste:

$$\hat{\sigma}_{\hat{\Gamma}_2}^2 = \sum_{i=1}^3 \frac{c_i^2}{n_i} 2.776 = \left( \frac{1^2}{12} + \frac{-1^2}{11} \right) * 2.776 = 0.48366$$

Los errores estándar de los contrastes serían:

Para el primer contraste:

$$\hat{\sigma}_{\hat{\Gamma}_1} = \sqrt{0.35226} = 0.5935$$

Para el segundo contraste:

$$\hat{\sigma}_{\hat{\Gamma}_2} = \sqrt{0.48366} = 0.695$$

El valor de  $t$  para cada contraste estaría dado por:

$$t = \frac{\hat{\Gamma}}{\hat{\sigma}_{\hat{\Gamma}}}$$

Para el primer contraste:

$$t_1 = \frac{\hat{\Gamma}_1}{\hat{\sigma}_{\hat{\Gamma}_1}} = \frac{-1.8787}{0.5935} = -3.1656$$

Para el segundo contraste:

$$t_2 = \frac{\hat{\Gamma}_2}{\hat{\sigma}_{\hat{\Gamma}_2}} = \frac{-0.075757576}{0.695} = -0.109$$

Se acepta  $H_0$  si  $|t| < t_{\alpha}$ . Para un  $t_{\alpha \frac{0.05}{2}; g.l.E} = 1.69$ , se tiene:

En el primer contraste:

$|t_1| > t_{1\alpha} = |-3.1656| > 1.69$ ; por lo tanto, el testigo difiere significativamente del promedio de las dosis  $Q$  y  $R$ .

Para el segundo contraste:

$|t_2| < t_{2\alpha} = |-0.109| < 1.69$ ; por lo tanto, los efectos medios de las dosis  $Q$  y  $R$  son iguales.

El lector podrá construir otros contrastes, lo importante es que existe la necesidad de comprobar si son ortogonales y deben realizarse antes de iniciar el experimento y que la interpretación de los resultados sea coherente.

Realicemos el ejercicio en R-project, iniciando con el análisis de varianza, usando el comando *lm* (fitting linear models), que es utilizado para realizar análisis en modelos lineales (implementado por Wilkinson y Rogers, 1973). También emplearemos el comando *summary.aov* para generar una tabla del análisis de varianza, y *levels* para ver el orden de aparición de los tratamientos. Este paso es necesario para construir la matriz de coeficientes de los contrastes, donde se tendrá en cuenta el ordenamiento alfanumérico de aparición de los niveles que hace

R-project. Cualquier error en la posición de los contrastes daría resultados inesperados:

```

R Studio

Analisis=lm(Base$Peso~Base$Tratamiento)
summary.aov(Analisis)
levels(Base$Tratamiento)

              Df Sum Sq Mean Sq F value Pr(>F)
Base$Tratamiento  2  27.92  13.959   5.029 0.0126 *
Residuals        32  88.83   2.776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> levels(Base$Tratamiento)
[1] "dosis Q" "dosis R" "testigo"

```

Vemos que el orden de posición de los niveles es *dosis Q*, *dosis R* y *testigo*, el cual se debe mantener para los contrastes.

Es necesario construir una matriz con los contrastes indicando los nombres y constantes:

```

R Studio

MatrizContraste= rbind (" Q, R vs testigo"=c(0.5,0.5,-1)," Q vs R"=c(1,-1,0))
MatrizContraste

              [,1] [,2] [,3]
Q, R vs testigo 0.5 0.5  -1
Q vs R          1.0 -1.0  0

```

En R-project existen diversas formas de realizar los contrastes. Nosotros utilizaremos la librería *gmodels* (programming tools for model fitting), elaborada por Warnes (2012). Esta librería tiene el comando *fit.contrast*, que genera las estimaciones de los contrastes, con sus respectivos errores estándar, los valores de  $t_c$ , las probabilidades de  $t$  y los intervalos de confianza, que en este caso probaremos con un nivel de confianza del 95%; veamos:

```

R Studio

install.packages("gmodels")
library(gmodels)
Contrastes=fit.contrast( Analisis,"Base$Tratamiento",MatrizContraste,conf=0.95)
Contrastes

              Estimate Std. Error  t value  Pr(>|t|)  lower CI  upper CI
Base$Tratamiento Q, R vs testigo -1.87878788  0.5934921 -3.1656496  0.003387426 -3.087692 -0.6698841
Base$Tratamiento Q vs R          -0.07575758  0.6954588 -0.1089318  0.913936815 -1.492361  1.3408457

```

Ahora utilizemos la librería *multcomp* de Hothorn et al. (2008) con el comando *glht*. Para esto, volveremos a realizar el análisis de varianza con la hoja de datos como objeto (*attach*).

```
R Studio  
attach (Base)  
resultados=lm(Peso~Tratamiento, data=Base)
```

Al comando *glht* le incluimos la hoja donde están los resultados del análisis y le indicamos el tipo de hipótesis a probar con el argumento *linfct*, con la comparación múltiple (*mcp*) generada por la matriz de contrastes, a la cual llamamos anteriormente *MatrizContraste*:

```
R Studio  
library (multcomp)  
a=glht(resultados, linfct=mcp(Tratamiento= MatrizContraste))  
summary (a, test=adjusted("none"))  
  
Simultaneous Tests for General Linear Hypotheses  
Multiple Comparisons of Means: User-defined Contrasts  
Fit: lm(formula = Peso ~ Tratamiento, data = Base)  
Linear Hypotheses:  


|                      | Estimate | Std. Error | t value | Pr(> t )   |
|----------------------|----------|------------|---------|------------|
| Q, R vs testigo == 0 | -1.87879 | 0.59349    | -3.166  | 0.00339 ** |
| Q vs R == 0          | -0.07576 | 0.69546    | -0.109  | 0.91394    |

  
---  
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
(Adjusted p values reported -- none method)
```

En la programación anterior, especificamos que no se ajusten los valores de *p* con el argumento *test = adjusted("none")*, para poder comparar los resultados de los contrastes obtenidos con el comando *fit.contrast* y los que veremos a continuación con el comando *lsmeans* de la librería del mismo nombre (*lsmeans*).

Los argumentos del comando *lsmeans* incluyen el nombre de la hoja de datos con los resultados del análisis, el nombre del factor que será contrastado y las especificaciones del contraste con el argumento *contr*, donde indicamos los nombres de los contrastes y las constantes a utilizar; veamos el ejemplo para un mejor entendimiento:



```
library (lsmeans)
lsmeans(resultados, Peso~Tratamiento,
        contr = list(Peso=list(QR_vs_testigo=c(0.5,0.5,-1), Q_vs_R=c(1,-1,0))))
```

```
$'Tratamiento lsmeans'
  Tratamiento  lsmean      SE df lower.CL upper.CL
  dosis Q  449.8333  0.4809543  32  448.8537  450.8130
  dosis R  449.9091  0.5023404  32  448.8859  450.9323
  testigo  451.7500  0.4809543  32  450.7703  452.7297

$'Tratamiento Peso'
      estimate      SE df  t.ratio p.value
QR_vs_testigo -1.87878788  0.5934921  32  -3.16565  0.00339
Q_vs_R        -0.07575758  0.6954588  32  -0.10893  0.91394
p values are not adjusted
```

Vemos que los resultados son iguales a los encontrados con los comandos *fit.contrast* y *glht*.

## Capítulo 4

### Diseño en bloques aleatorizados

#### 4.1. Generalidades

Este tipo de diseño se caracteriza porque el material experimental presenta un factor de heterogeneidad definido, siendo necesario formar grupos, clúster o bloques para que las unidades experimentales presentes dentro de cada bloque sean homogéneas entre sí.

En experimentación con animales es muy común realizar diseños en bloques aleatorizados, principalmente cuando se tienen fuentes de bloqueo, como es el caso de individuos de diferentes razas, diferente número de partos, diferente condición morfométrica o edad, o localizados en diferentes galpones, jaulas o estanques.

En experimentos agrícolas pueden hacerse bloques por el tipo de suelo, la variedad de la planta, la altitud sobre el nivel del mar, el tipo de pendiente o el tipo de arado, entre muchos otros factores.

Se recomienda que el número de unidades experimentales presentes dentro de cada bloque no sea demasiado grande, ya que a mayor número de tratamientos es más complejo formar los bloques y asegurar la homogeneidad de las unidades experimentales dentro del bloque. El modelo de clasificación es:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

En dicho modelo,  $y_{ij}$  representa la variable respuesta o dependiente, asociada al  $i$ -ésimo tratamiento en el  $j$ -ésimo bloque,  $\mu$  indica el efecto promedio del experimento,  $\tau_i$  indica el efecto promedio del  $i$ -ésimo tratamiento ( $i = 1, 2, \dots, I$ ),  $\beta_j$  representa el efecto del  $j$ -ésimo bloque ( $j = 1, 2, \dots, J$ ) y  $\varepsilon_{ij}$  es el error experimental asociado con el  $i$ -ésimo tratamiento y el  $j$ -ésimo bloque.

Los errores son independientes e idénticamente distribuidos (iid) y tienen una distribución normal con media 0 y varianza  $\sigma_\varepsilon^2$ , cuya notación es:

$$\varepsilon_{j(i)} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$$

El efecto de bloque ( $\beta$ ) es fijo y el tratamiento puede ser fijo ( $\tau$ ) o aleatorio ( $t$ ). Si es aleatorio, se entiende que los tratamientos provienen de una muestra aleatoria de una población de tratamientos, indicando que  $t_i \stackrel{\text{iid}}{\sim} N(0, \sigma_t^2)$ .

Las hipótesis para el efecto de tratamiento en un modelo fijo son:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \dots = \tau_I$$

$H_a$  : los promedios de los niveles del tratamiento son diferentes

La hipótesis nula indica que ninguno de los niveles del factor tiene un efecto significativo sobre la variable respuesta, mientras que la hipótesis alterna expresa que existe al menos un nivel que tiene efecto notorio sobre los otros niveles.

En el caso de un modelo aleatorio:

$$H_0 : \sigma_t^2 = 0$$

$$H_a : \sigma_t^2 > 0$$

La hipótesis nula indica que todos los tratamientos son idénticos, mientras que la hipótesis alternativa indica que existe variabilidad entre ellos.

Aunque las hipótesis del efecto de bloque parecen innecesarias, es conveniente plantearlas para el investigador, porque le indica si fue acertada la idea de utilizar bloques en su experimento. Principalmente cuando la suma de cuadrados de los bloques es parte importante de la suma de cuadrados total. Las hipótesis a plantearse serían:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \dots = \beta_I$$

$H_a$  : los promedios de los bloques son diferentes

La estructura de un diseño de bloques aleatorizados con efecto fijo es:



Las principales ventajas del diseño en bloques completamente aleatorizados son:

- ♣ Es uno de los diseños más utilizados para estimar superficies de respuesta
- ♣ Se puede estimar el error experimental asociado a cada unidad experimental
- ♣ Permite convalidar supuestos asociados con el modelo de clasificación experimental
- ♣ Puede tener componentes o factores de efecto fijo o aleatorio
- ♣ Permite generar arreglos experimentales especiales; por ejemplo: arreglos factoriales, arreglos mixtos, arreglos confundidos, arreglos parcialmente confundidos, arreglos doblemente confundidos, bloques en cadena, etc.
- ♣ Se puede utilizar transformación de datos sobre la variable respuesta
- ♣ Permite efectuar análisis multivariado de la varianza (Manova)

- ♣ Admite tener componentes o factores anidados
- ♣ Acepta efectuar análisis con submuestras
- ♣ Se puede estimar la diferencia en el efecto promedio de los tratamientos
- ♣ Reduce la magnitud del error experimental, incrementando el grado de precisión del experimento
- ♣ Permite la presencia de covariables
- ♣ Se pueden utilizar contrastes ortogonales
- ♣ Permite efectuar particiones experimentales, dependiendo del supuesto de homogeneidad de varianzas
- ♣ Puede ser diseño balanceado o desbalanceado

Si no se puede asegurar la similaridad del material experimental dentro de cada uno de los bloques, se ve seriamente afectado el análisis de la varianza, ya que se violarían los supuestos relacionados con el modelo de clasificación experimental.

En producción animal, el diseño de bloques aleatorizados se puede utilizar cuando:

- ♣ Se requiere saber y verificar si existe diferencia estadística en la producción animal cuando se tienen diferentes ambientes como efecto de bloqueo
- ♣ Se pretende establecer el punto óptimo de pmezclas, por medio de diseños de superficie de respuesta, en los cuales se puede incorporar el modelo en asocio con la técnica de regresión lineal
- ♣ Es aplicable en experimentos acuícolas cuando se pretende tener en cuenta la incidencia de la pendiente del estanque, la velocidad de entrada del agua bajo condiciones controladas, o el efecto de los estanques, entre múltiples aplicaciones
- ♣ Se utiliza en experimentación agrícola a fin de ver el efecto del tipo de suelo sobre el cultivo. También se puede hacer bloques por: tipo de variedad sembrada, tipo de manejo relacionado con factores culturales, tipo de riego, tipo de pendiente del suelo, profundidad de siembra, factores de altitud sobre el nivel del mar, densidad de siembra, distancia de siembra, entre otros
- ♣ Se utiliza en experimentos que deben considerar el efecto de la raza o la línea genética del animal sobre la producción
- ♣ En experimentación industrial se consideran bloques los factores como: jornada de trabajo, calificación del operario, tipo de máquina, tipo de materiales empleados, tiempos de procesos para el manejo de materiales, entre muchas otras.

## 4.2. Supuestos estadísticos

Los supuestos del diseño de bloques aleatorizados son:

- ♣ Debe existir aditividad entre los efectos de tratamiento y bloque.
- ♣ Se debe verificar que los errores experimentales se distribuyan en forma normal.
- ♣ Los errores deben ser independientes estadísticamente y aleatoriamente distribuidos.
- ♣ Las varianzas asociadas con los distintos tratamientos deben ser homogéneas.
- ♣ No debe existir relación entre los promedios de los tratamientos y las varianzas.

## 4.3. Proceso de aleatorización

La aleatorización se debe realizar dentro de cada bloque y de forma independiente, generando tantos números aleatorios como unidades experimentales existan en su interior. Al igual que en el diseño completamente aleatorizado, cualquier unidad experimental posee igual probabilidad de ser asociada a un tratamiento y, en este caso, dentro de cada bloque. Para el proceso de aleatorización se utiliza como base la distribución uniforme.

Se puede realizar manualmente, sacando al azar balotas o papeles donde se tienen los tratamientos dentro de cada bloque. Nosotros utilizaremos el algoritmo de Wichmann-Hill, para la generación de números pseudoaleatorios utilizando el comando *design.rcbd* de la librería *agricolae* (Mendiburu, 2012).

El comando *design.rcbd* permite aleatorizar los tratamientos dentro de cada bloque, y en los argumentos se requiere indicar el número de tratamientos, el número de bloques, la especificación de los métodos de aleatorización y la semilla (*seed*).

Veamos un ejemplo de aleatorización con tres tratamientos ( $T$ ,  $T1$  y  $T2$ ) y cuatro bloques ( $bloq = 4$ ). Utilizaremos una semilla ( $seed = 14$ ) y numeraremos las unidades experimentales a partir de 1.

```

library(agricolae)
Tratamiento=c("T", "T1", "T2")
Bloq=4
Semilla=14
Numeracion=1
Generar=design.rcbd (Tratamiento, Bloq, number=Numeracion,
                    kind="Wichmann-Hill", seed=Semilla)
Generar

```

```

plots block Tratamiento
1      1      T
2      1     T1
3      1     T2
4      2      T
5      2     T2
6      2     T1
7      3     T1
8      3     T2
9      3      T
10     4      T
11     4     T1
12     4     T2

```

En el siguiente esquema de la aleatorización de los tratamientos dentro de los bloques se puede observar cómo fueron distribuidos los tratamientos:



#### 4.4. Modelo

Para mostrar la construcción del modelo, utilizaremos como ejemplo un experimento de peso al sacrificio de cerdos sometidos a tres tratamientos, los cuales serán distribuidos en cuatro jaulas (bloques). El modelo para el diseño en bloques al azar con efecto fijo es:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

Donde  $y_{ij}$  es el peso al sacrificio del cerdo sometido al  $i$ -ésimo tratamiento en la  $j$ -ésima jaula,  $\mu$  indica el efecto promedio del experimento,  $\tau_i$  indica el efecto promedio del  $i$ -ésimo tratamiento ( $i = 1, 2$  y  $3$ ),  $\beta_j$  representa el efecto de la  $j$ -ésima jaula o factor de bloqueo ( $j = 1, 2, 3$  y  $4$ ), y  $\varepsilon_{ij}$  es el error experimental asociado al  $i$ -ésimo tratamiento y la  $j$ -ésima jaula y que se caracteriza porque  $\varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$ .

Las hipótesis a probar son:

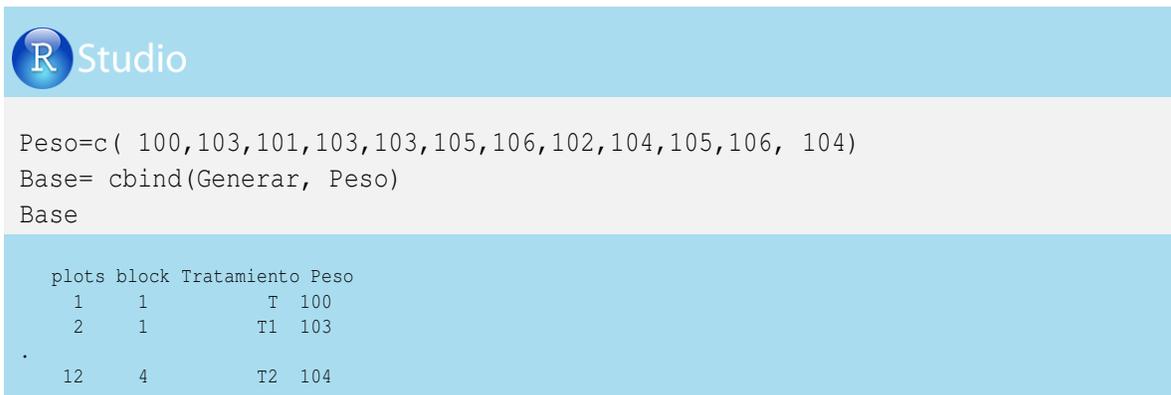
$$H_0 : \tau_T = \tau_{T1} = \tau_{T2}$$

$H_a$  : los promedios de los niveles del tratamiento son diferentes

Al terminarse el experimento con animales de cinco meses de edad, se encontraron los siguientes pesos al sacrificio:

Información del pesaje de los machos sacrificados (en kilogramos)			
Cerdo	Bloque	Tratamiento	Peso
1	1	T	100
2	1	T1	103
3	1	T2	101
4	2	T	103
5	2	T2	103
6	2	T1	105
7	3	T1	106
8	3	T2	102
9	3	T	104
10	4	T	105
11	4	T1	106
12	4	T2	104

Generemos la hoja de datos *Base* con las columnas del proceso de aleatorización de la hoja de datos *Generar* y una columna de peso de los animales en R-project:



```

R Studio

Peso=c( 100,103,101,103,103,105,106,102,104,105,106, 104)
Base= cbind(Generar, Peso)
Base

plots block Tratamiento Peso
1      1          T      100
2      1         T1     103
.
.
.
12     4         T2     104

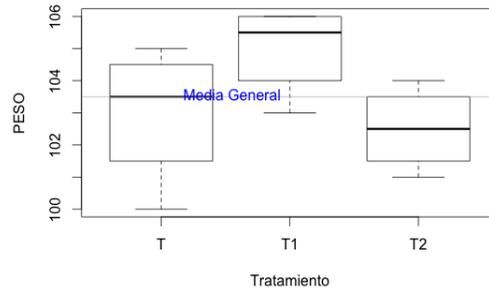
```

#### 4.5. Análisis exploratorio de la hoja de datos

Realizaremos gráficos de boxplot del peso en relación con los tratamientos. El objetivo es visualizar la distribución de los datos. Esta labor es importante para detectar errores de digitación y tener una idea general de los posibles resultados:



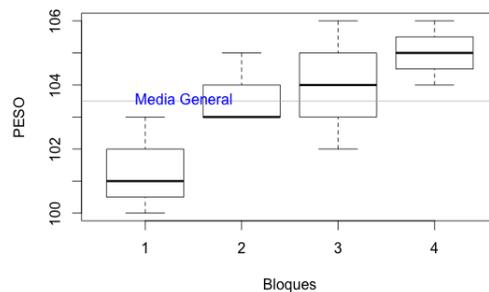
```
boxplot (Base$Peso~Base$Tratamiento,ylab="PESO",xlab="Tratamiento", data=Base)
abline (h=mean (Base$Peso) ,col="gray")
text (2,mean (Base$Peso) ,"Media General" ,col="blue" ,pos=2)
```



Vemos que el mayor promedio lo presentó el tratamiento  $\tau_{T_1}$  y la mayor variación la presentó el tratamiento  $t$ . Ahora veamos un boxplot en el caso de los bloques:



```
boxplot (Base$Peso~Base$block,ylab="PESO",xlab="Bloques", data=Base)
abline (h=mean (Base$Peso) ,col="gray")
text (2,mean (Base$Peso) ,"Media General" ,col="blue" ,pos=2)
```



Se nota que el primer bloque es diferente a los otros tres, mostrando que este efecto era importante en el diseño, pero esto se confirmará al realizar el análisis de varianza.

#### 4.6. Análisis de varianza

Ahora procederemos a realizar el análisis de varianza. Para esto, calcularemos las sumas de cuadrados total, de tratamientos, de bloques y del error, como se indica a continuación:

Suma de cuadrados totales (*SCT*):

$$SCT = \sum_{i=1}^I \sum_{j=1}^J (y_{ij})^2 - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J y_{ij} \right)^2}{n}$$

Suma de cuadrados de tratamientos (*SCTr*)

$$SCTr = \frac{\sum_{i=1}^I \tau_i^2}{J} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J y_{ij} \right)^2}{n}$$

Suma de cuadrados de bloques (*SCB*)

$$SCB = \frac{\sum_{j=1}^J \beta_j^2}{I} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J y_{ij} \right)^2}{n}$$

Suma de cuadrados del error (*SCE*)

$$SCE = SCT - SCTr - SCB$$

Para el ejemplo:

Suma de cuadrados totales (*SCT*):

$$SCT = \sum_{i=1}^3 \sum_{j=1}^4 (y_{ij})^2 - \frac{\left( \sum_{i=1}^3 \sum_{j=1}^4 y_{ij} \right)^2}{12}$$

$$SCT = (100)^2 + (103)^2 + \dots + (104)^2 - \frac{(100 + 103 + \dots + 104)^2}{12}$$

$$SCT = 128586 - \frac{(1242)^2}{12} = 39$$

Suma de cuadrados de tratamientos (*SCTr*)

$$SCTr = \frac{\sum_{i=1}^3 \tau_i^2}{4} - \frac{\left( \sum_{i=1}^3 \sum_{j=1}^4 y_{ij} \right)^2}{12}$$

$$SCTr = \frac{(412)^2 + (420)^2 + (410)^2}{4} - \frac{(100 + 103 + \dots + 104)^2}{12} = 14$$

Suma de cuadrados de bloques (*SCB*)

$$SCB = \frac{\sum_{j=1}^4 \beta_j^2}{3} - \frac{\left( \sum_{i=1}^3 \sum_{j=1}^4 y_{ij} \right)^2}{12}$$

$$SCB = \frac{(304)^2 + (311)^2 + (312)^2 + (315)^2}{3} - \frac{(100 + 103 + \dots + 104)^2}{12} = 21.6666667$$

Suma de cuadrados del error (*SCE*)

$$SCE = 39 - 14 - 21.6666667 = 3.333333$$

El cuadro del análisis de varianza estaría conformado por:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$
Tratamiento	$I - 1$	$SCTr$	$\frac{SCTr}{I-1}$	$\frac{CMtr}{CME}$
Bloque	$J - 1$	$SCB$	$\frac{SCB}{J-1}$	$\frac{CMB}{CME}$
Error	$(n-1)-(I-1)-(J-1)$	$SCE$	$\frac{SCE}{n-I}$	
Total	$n - 1$	$SCT$		

Para el ejemplo:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$
Tratamiento	$3 - 1$	14	$\frac{14}{2} = 7$	$\frac{7}{0.55556} = 12.6$
Bloques	$4 - 1$	21.667	$\frac{21.6667}{3} = 7.22$	$\frac{7.22222}{0.55556} = 13.0$
Error	$(12 - 1) - (3 - 1) - (4 - 1)$	3.333	$\frac{3.33}{6} = 0.556$	
Total	$12 - 1$	39		

Rechazamos la hipótesis nula de los tratamientos, si:  $F_c > F_{\alpha, I-1, ((n-1)-(I-1)-(J-1))}$

Donde  $F_{\alpha, I-1, ((n-1)-(I-1)-(J-1))}$  es el cuantil  $\alpha$  de la distribución de  $F$  con  $I - 1$  grados de libertad de tratamientos y  $(n - 1) - (I - 1) - (J - 1)$  grados de libertad del error.

En la tabla de distribución de  $F$  con 2 grados de libertad de los tratamientos y 6 grados de libertad del error y considerando un  $\alpha = 0.01$ , tenemos que el valor de  $F$  es de 10.925, entonces:

$$F_c > F_{0.01, 2, 6}$$

$$12.6 > 10.925$$

Por lo tanto, decimos que hay diferencias estadísticas altamente significativas ( $p < 0.01$ ) entre tratamientos.

Para saber si fue acertada la decisión de realizar el experimento en un diseño de bloques aleatorizados, rechazamos la hipótesis nula de los bloques, si:  $F_c > F_{\alpha, J-1, ((n-1)-(I-1)-(J-1))}$

Considerando un  $\alpha = 0.01$ , tenemos que el valor de  $F$  de la tabla es de 9.78, entonces:

$$F_c > F_{0.01, 3, 6}$$

$$13 > 9.78$$

Podemos decir que fue conveniente utilizar bloques en el experimento, porque existieron diferencias altamente significativas entre bloques ( $p < 0.01$ ). Adicionalmente, la suma de cuadrados de bloques fue mayor que la suma de cuadrados del error y de los tratamientos.

Realicemos el análisis de varianza en R-project. Inicialmente procederemos a convertir la hoja de datos *Base* en un objeto (*attach*) y desarrollaremos el análisis de varianza con el comando *aov*, cuyos resultados saldrán en la hoja de datos *Analisis*. Adicionalmente solicitaremos un resumen del análisis con el comando *summary*:



```
attach (Base)
Analisis=aov(Peso~Tratamiento + block, data=Base)
summary(Analisis)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tratamiento	2	14.000	7.000	12.6	0.00711 **
block	3	21.667	7.222	13.0	0.00492 **
Residuals	6	3.333	0.556		

---  
Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Observamos que el efecto del tratamiento y el efecto de los bloques fueron altamente significativos ( $p < 0.01$ ). Realizaremos la prueba de Tukey para comparar los tratamientos, mediante la librería *lsmeans* con el comando del mismo nombre, desarrollada por Lenth (2012):

```

R Studio

Tukey=lsmeans(Analisis, list(pairwise~Tratamiento), adjust="tukey")
Tukey

$`Tratamiento lsmeans`
  Tratamiento lsmean      SE df lower.CL upper.CL
      T      103.0 0.372678  6 102.0881 103.9119
     T1      105.0 0.372678  6 104.0881 105.9119
     T2      102.5 0.372678  6 101.5881 103.4119
$`Tratamiento pairwise differences`
      estimate      SE df  t.ratio p.value
T - T1      -2.0 0.5270463  6 -3.79473 0.02110
T - T2       0.5 0.5270463  6  0.94868 0.63237
T1 - T2       2.5 0.5270463  6  4.74342 0.00759
  p values are adjusted using the tukey method for 3 means

```

Las medias por tratamiento, de mayor a menor, fueron:

$$\hat{\tau}_{T1} = 105.0$$

$$\hat{\tau}_T = 103.0$$

$$\hat{\tau}_{T2} = 102.5$$

Al comparar los tratamientos tenemos:

$\hat{\tau}_{T1}$  es diferente de  $\hat{\tau}_T$  y de  $\hat{\tau}_{T2}$ , por lo tanto, le asignamos la letra *a*  
 $\hat{\tau}_T$  y  $\hat{\tau}_{T2}$ , son semejantes, por lo tanto, les asignamos la letra *b*, de modo que queda:

$$\hat{\tau}_{T1} = 105.0, a$$

$$\hat{\tau}_T = 103.0, b$$

$$\hat{\tau}_{T2} = 102.5, b$$

Ahora obtengamos las medias de los bloques mediante el comando *model.tables*, indicando la hoja donde está el análisis y el argumento "mean":

```

R Studio

model.tables(Analisis, "mean")

Grand mean
  103.5
Tratamiento
      T   T1  T2
  103.0 105.0 102.5
block
      1     2     3     4
  101.33 103.67 104.00 105.00

```

Con la información anterior, se obtiene:

$$\hat{\mu} = 103.5$$

En el caso de los bloques:

$$\hat{\beta}_1 = 101.33$$

$$\hat{\beta}_2 = 103.67$$

$$\hat{\beta}_3 = 104.00$$

$$\hat{\beta}_4 = 105.00$$

#### 4.7. Prueba de normalidad

Al igual que el modelo completamente aleatorizado, los errores experimentales deben tener distribución normal, con media cero y varianza homogénea:  $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

Para verificar lo anterior, se procede a estimar los residuos:

Si:  $y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$ , entonces:

$$\varepsilon_{ij} = y_{ij} - \mu - \tau_i - \beta_j$$

Al reemplazar el efecto promedio de cada tratamiento ( $\hat{\tau}$ ), el efecto de bloques ( $\hat{\beta}$ ) y el efecto promedio general ( $\hat{\mu}$ ), se tiene:

$$\varepsilon_{ij} = y_{ij} - \hat{\mu} - (\hat{\tau}_i - \hat{\mu}) - (\hat{\beta}_j - \hat{\mu})$$

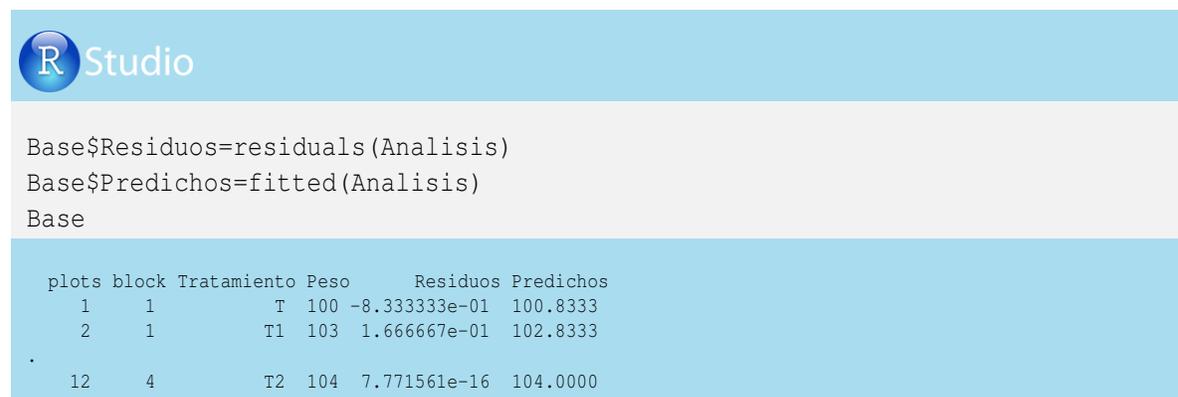
$$\varepsilon_{ij} = y_{ij} - \hat{\tau}_i - \hat{\beta}_j + \hat{\mu}$$

En el caso de la primera unidad experimental (tratamiento  $T$ , en el bloque 1), se tendrían:

$$\varepsilon_{i=T, j=1} = y_{i=T, j=1} - \hat{\tau}_T - \hat{\beta}_1 + \hat{\mu}$$

$$\varepsilon_{i=1, j=1} = 100 - 103 - 101.33 + 103.5 = -0.833$$

Para obtener los residuos y los valores predichos en R-project, utilizamos los comandos *residuals* y *fitted*, respectivamente, así:



```

R Studio

Base$Residuos=residuals(Analisis)
Base$Predichos=fitted(Analisis)
Base

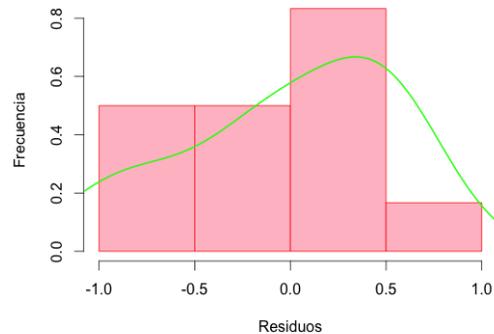
  plots block Tratamiento Peso      Residuos Predichos
1     1     1           T   100 -8.333333e-01  100.8333
2     1     1          T1   103  1.666667e-01  102.8333
.
12    4     4          T2   104  7.771561e-16  104.0000

```

Para convalidar el supuesto estadístico de normalidad de los errores, asociados al modelo de clasificación experimental, debemos realizar pruebas de normalidad, como es el caso de Shapiro-Wilk, pero antes realizaremos un histograma de frecuencias en R-project que nos dará un indicio de la distribución de los datos:



```
hist(Base$Residuos,probability=T,breaks="Sturges",
     col="pink", main=" ",xlab="Residuos",ylab="Frecuencia",border="red")
lines(density(Base$Residuos),lwd=2,col="green")
```



Aparentemente los residuos no tienen distribución normal, y el pico de la distribución de probabilidad (línea verde) está ligeramente a la derecha. Sin embargo, los valores están muy cercanos a 0 (entre  $-1$  y  $1$  kg). Veamos otros detalles de los residuos:



```
summary(Base$Residuos)
var(Base$Residuos)
mean(Base$Residuos)
```

```
> summary(Base$Residuos)
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.00000 -0.25000  0.08333  0.00000  0.50000  0.66670
> var(Base$Residuos)
[1] 0.3030303
> mean(Base$Residuos)
[1] -1.017757e-16
```

La media de los errores fue 0, con varianza 0.303030 ( $\epsilon_{ij}$  tiene  $\mu_{\epsilon} = 0$  y  $\sigma_{\epsilon}^2 = 0.303030$ ). Apliquemos la prueba de Shapiro-Wilk con el comando *shapiro.test*:



```
ShapiroResiduos=shapiro.test(Base$Residuos)
ShapiroResiduos
```

```
Shapiro-Wilk normality test
data: Base$Residuos
W = 0.9162, p-value = 0.2558
```

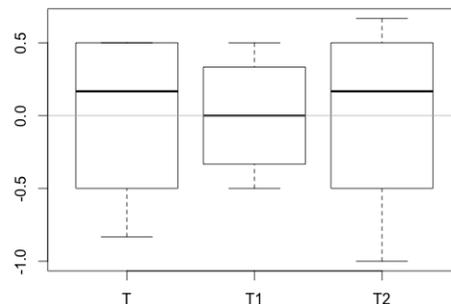
Por presentar un valor de  $p > 0.05$ , los errores se distribuyeron normalmente, por lo tanto:  $\epsilon_{ij} \sim N(0, 0.3030)$

#### 4.8. Homogeneidad de varianza de los errores

Adicionalmente, es necesario verificar si los errores experimentales presentan homogeneidad de varianzas en los tres tratamientos. Para esto aplicamos las pruebas de Cochran, Bartlett y Fligner-Killeen, pero antes visualicemos de forma gráfica la distribución de los errores en los tratamientos. Apliquemos el comando *boxplot* especificando que es por tratamiento:



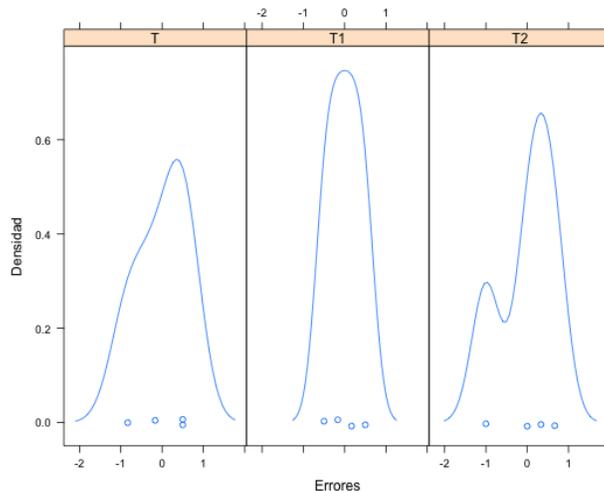
```
boxplot(Base$Residuo~Base$Tratamiento, data=Base)
abline(h=mean(Base$Residuo), col="gray")
```



Vemos que los residuos tienen una media cercana a cero dentro de cada tratamiento. Ahora utilizemos la librería *lattice*, la cual tiene el comando *densityplot*, para generar un gráfico mejor elaborado.



```
library (lattice)
densityplot (~Base$Residuos|Base$Tratamiento,ylab="Densidad", xlab="Errores")
```



Vemos que el tratamiento  $T_2$  se aprecia diferente a los otros dos tratamientos y el  $T_1$  presenta unos residuos más cercanos a 0. Ahora realicemos la prueba de Cochran, para confirmar si existe homogeneidad de los residuos en los tratamientos:



```
attach (Base)
library (outliers)
CochranRes=cochran.test (Residuos~Tratamiento,data=Base)
CochranRes
```

```
data: Residuos ~ Tratamiento
C = 0.4667, df = 4, k = 3, p-value = 0.7684
alternative hypothesis: Group T2 has outlying variance
sample estimates:
      T      T1      T2
0.4074074 0.1851852 0.5185185
```

Por presentar un  $p > 0.05$  (no significativo), no rechazamos la hipótesis  $H_0$ . Las varianzas de los errores de los tratamientos son homogéneas.

Realicemos la prueba de Bartlett con el comando *bartlett.test*, indicando en los argumentos, la variable a analizar, el efecto clasificatorio y la hoja de datos:



```
BartlettResiduos=bartlett.test(Base$Residuos,Base$Tratamiento, data=Base)
BartlettResiduos
```

```
Bartlett test of homogeneity of variances
data: Base$Peso and Base$Tratamiento
Bartlett's K-squared = 0.8375, df = 2, p-value = 0.6579
```

Según la prueba de Bartlett de homogeneidad de varianzas, los residuos son homogéneos entre tratamientos:

Para la prueba de Fligner-Killeen utilizaremos el comando *fligner-test* con los argumentos: nombre de la variable que tiene los residuos, nombre de la variable que agrupa los residuos y el nombre de la hoja de datos. Veamos:



```
FlignerResiduos=fligner.test(Base$Residuos,Base$Tratamiento,data=Base)
FlignerResiduos
```

```
Fligner-Killeen test of homogeneity of variances
data: Base$Residuos and Base$Tratamiento
Fligner-Killeen:med chi-squared = 0.3969, df = 2, p-value = 0.82
```

En general, los residuos del experimento se distribuyen normalmente y existió homogeneidad de los errores entre tratamientos. Por lo tanto, no es necesario realizar transformaciones de los datos; sin embargo, hace falta probar si existe aditividad entre los tratamientos y los bloques.

#### 4.9. Aditividad de efectos

En un diseño de bloques al azar puede existir interacción de los efectos principales con los bloques, lo que imposibilita la comparación de los niveles de estos efectos. La no aditividad causada por la presencia de interacción conlleva a que si existen diferencias significativas de los efectos, se daría una recomendación errónea acerca de cuál es el mejor tratamiento. En el caso contrario, si los efectos no son significativos, queda la duda si realmente la interacción es la causante de que no existieran diferencias de medias de los niveles del tratamiento o del bloque.

A pesar de que es muy difícil determinar la existencia de interacción en este diseño experimental, se puede aplicar la prueba de no aditividad de Tukey. En R-project podemos utilizar el comando *tukey.add.test* de la librería *asbio* de Aho (2012) con los siguientes argumentos en su orden: nombre de la variable dependiente, nombre de la variable que contiene el efecto

principal y nombre de la variable que contiene la descripción de los bloques. Si en el resultado el valor de  $p$  es no significativo ( $p > 0.05$ ), se cumple el supuesto de aditividad del análisis de varianza, como se indica en el siguiente ejemplo:

```

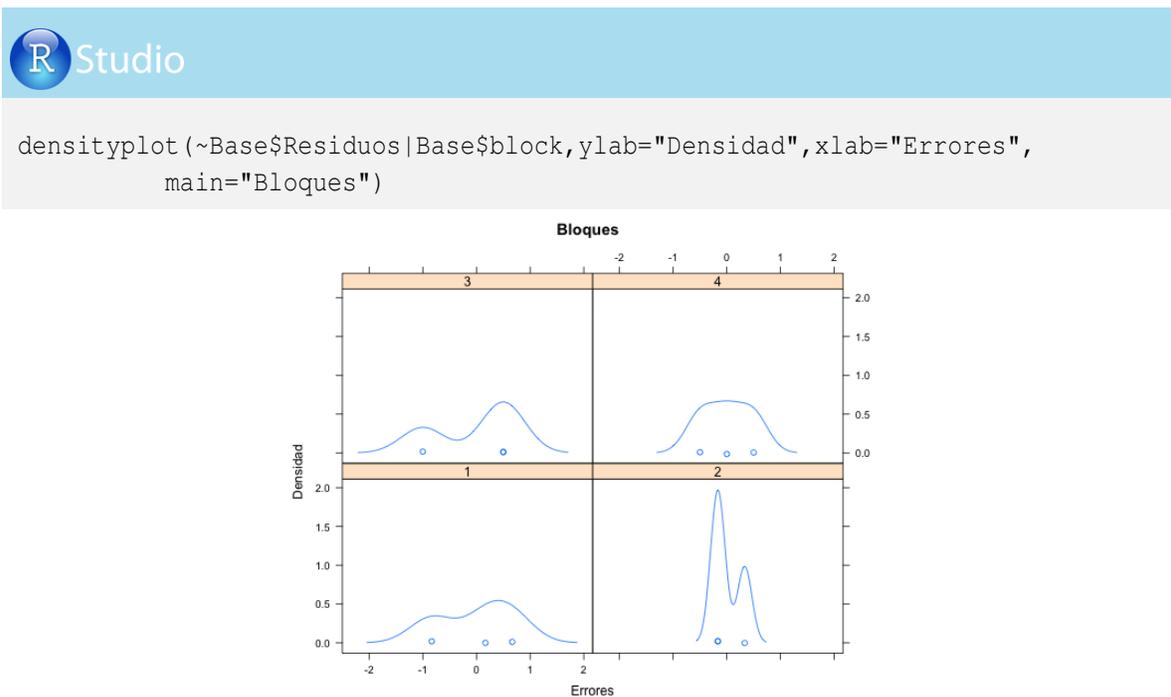
R Studio

install.packages("asbio")
library (asbio)
tukey.add.test (Base$Peso, Base$Tratamiento, Base$block)

Tukey's one df test for additivity
data: Base$Tratamiento and Base$block on Base$Peso
F = 0.0416, num.df = 1, denom.df = 5, p-value = 0.8465
    
```

Si la prueba de Tukey para la no aditividad es significativa, es necesario verificar si es causada por la presencia de valores *outliers* o por heterogeneidad de varianzas. De ser así, podríamos aplicar algún tipo de transformación para intentar solucionar el problema de no aditividad. Si realmente la causa es la interacción existente de los efectos principales y los bloques, los resultados encontrados estarían enmascarados por la presencia de la interacción.

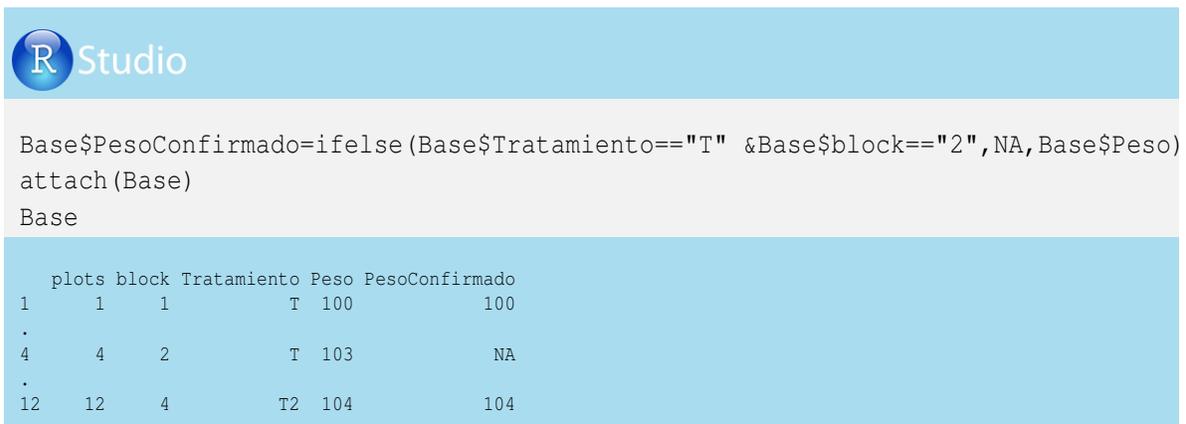
En nuestro ejemplo se cumplieron los supuestos de normalidad, homogeneidad de varianzas y aditividad. Adicionalmente, la decisión de tener en cuenta el efecto del bloque antes de iniciar el experimento fue acertada, porque se notó que las condiciones ambientales de las jaulas influyeron en el peso de los cerdos. Veamos el siguiente gráfico:



#### 4.10. Parcelas perdidas

En muchas ocasiones y por múltiples razones se pierden unidades experimentales, lo cual genera un desbalance. En estos casos la suma de cuadrados de los tratamientos y los bloques se ve afectada y es necesario realizar un análisis con suma de cuadrados que tenga en cuenta el desbalance, como es el caso de suma de cuadrados tipo III.

Retomemos el ejemplo del peso de los cerdos al sacrificio sometidos a tres tratamientos. No tendremos en cuenta el dato del tratamiento *T* del bloque 2 que es 103 kg porque confirmamos que existió un problema en la toma de información. Crearemos la variable *PesoConfirmado* incluyéndole el comando *ifelse* para ingresar un dato perdido (*NA*):



```

Base$PesoConfirmado=ifelse (Base$Tratamiento=="T" &Base$block=="2",NA,Base$Peso)
attach(Base)
Base

```

	plots	block	Tratamiento	Peso	PesoConfirmado
1	1	1	T	100	100
.					
4	4	2	T	103	NA
.					
12	12	4	T2	104	104

En esta oportunidad utilizaremos el comando *lm* en lugar de *aov* para realizar el análisis de varianza y crearemos la hoja de datos del análisis que se llama *Analisis1*. Solicitaremos un resumen del análisis con sumas de cuadrados tipo I, mediante el comando *anova*, y posteriormente pediremos un resumen con la suma de cuadrados tipo III con el comando *Anova* de la librería *car* de Fox y Weisberg (2011). Recordemos que el primer comando es con *a* (minúscula) y el segundo comando es con *A* (mayúscula).

```

R Studio

Analisis1=lm(PesoConfirmado~Tratamiento + block,data=Base)
anova(Analisis1)
install.packages("car")
library(car)
Anova(Analisis1, type=c("III"))

> anova(Analisis1)
  Response: PesoConfirmado
           Df Sum Sq Mean Sq F value Pr(>F)
Tratamiento 2 13.7273  6.8636  10.470 0.01631 *
block        3 21.7222  7.2407  11.045 0.01207 *
Residuals    5  3.2778  0.6556
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Anova(Analisis1, type=c("III"))
  Anova Table (Type III tests)
           Sum Sq Df  F value    Pr(>F)
(Intercept) 18321.4  1 27947.932 1.453e-10 ***
Tratamiento   13.4  2   10.212  0.01715 *
block         21.7  3   11.045  0.01207 *
Residuals     3.3  5
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Vemos que la suma de cuadrados de tratamientos de tipo I (13.7) fue diferente a la de tipo III (13.4), la cual indica la necesidad de obtener las sumas de cuadrados tipo III cuando existe un desbalance en los datos. También es necesario obtener las medias ajustadas para los tratamientos que tengan en cuenta el desbalance.

Utilizaremos la librería *lsmeans*, incluyendo el argumento *pairwise* para realizar comparaciones de medias entre tratamientos.

```

R Studio

library(lsmeans)
Dk=lsmeans(Analisis1, list(pairwise~Tratamiento))
Dk

$`Tratamiento lsmeans`
  lsmean      SE df lower.CL upper.CL
T  103.0833 0.4958158  5 101.8088 104.3579
T1 105.0000 0.4048319  5 103.9593 106.0407
T2 102.5000 0.4048319  5 101.4593 103.5407
$`Tratamiento pairwise differences`
  estimate      SE df  t.ratio p.value
T - T1  -1.9166667 0.6400955  5 -2.99434 0.06630
T - T2   0.5833333 0.6400955  5  0.91132 0.65727
T1 - T2  2.5000000 0.5725188  5  4.36667 0.01651
p values are adjusted using the tukey method for 3 means

```

Las diferencias significativas ( $p < 0.05$ ) se presentaron entre el tratamiento *T1* y el *T2*.

### 4.11. Bloques incompletos balanceados

En algunas ocasiones no es posible tener bloques que alberguen una gran cantidad de tratamientos. Para esto utilizamos los bloques incompletos balanceados. En este tipo de diseño, se busca que los tratamientos se distribuyan en los bloques, manteniendo el mismo número de réplicas por tratamiento (balanceado).

Veamos el siguiente ejemplo, donde necesitamos evaluar cuatro tipos de concentrado en peces. Los peces están en estanques experimentales (réplicas) los cuales están conectados a bocatomas (dos estanques por bocatoma). En este caso, el efecto de bloqueo está dado por las bocatomas, pero lastimosamente no se pueden tener cuatro estanques por bocatoma (uno para cada tratamiento), únicamente dos.

Para el desarrollo de este ejercicio utilizaremos la librería *agricolae* para generar la distribución aleatoria de las réplicas, con el comando *design.bib* (randomized balanced incomplete block) de Mendiburu (2012). En este comando especificamos el número de tratamientos, el número de bloques, el inicio de la numeración de observaciones (*number=1*, para que inicie en 1), la semilla (*seed = 1*) y el método de aleatorización (*supper-duper*). Recordemos que la semilla es un valor que siempre generará la misma aleatorización, la cual puede ser modificada o eliminada por el lector. Al finalizar los argumentos incluimos *\$book* con el objetivo de solicitar los datos generados por este comando o incluimos *\$parameters* si deseamos únicamente la lista de nombres de los parámetros. Veamos la programación en R-project:



```

library(agricolae)
Concentrado=c("Concentrado 1","Concentrado 2","Concentrado 3","Concentrado 4")
TanquesTomad=2
BloquesIncomp=design.bib(Concentrado,TanquesTomad,number=1,
                        seed =1,kinds ="Super-Duper")$book
BloquesIncomp

```

```

> BloquesIncomp
  plots block Concentrado
1     1     1 Concentrado 1
2     2     1 Concentrado 4
3     3     2 Concentrado 4
.
12    12     6 Concentrado 1

```

Veamos la cantidad de tratamientos en cada bloque, mediante el comando *table*.

```

R Studio

table (BloquesIncomp$Concentrado, BloquesIncomp$block)

BloquesIncomp$block)
      1 2 3 4 5 6
Concentrado 1 1 0 0 0 1 1
Concentrado 2 0 0 1 1 0 1
Concentrado 3 0 1 1 0 1 0
Concentrado 4 1 1 0 1 0 0
    
```

Podemos observar que el *Concentrado 1* está en tres bloques (1, 5 y 6), el *Concentrado 2* en los bloques 3, 4 y 6, el *Concentrado 3* en los bloques 2, 3 y 5 y el *Concentrado 4* en los bloques 1, 2 y 4. De esta forma, confirmamos que el diseño es de bloques incompletos (les falta un tratamiento) y balanceado (todos los tratamientos están en dos bloques).

Ahora creamos la hoja de datos con información del peso de los peces, elaboramos una nueva hoja de datos (llamada *Hoja*) y generamos unas nuevas variables con los nombres de *Tratamiento* y *Bocatoma*.

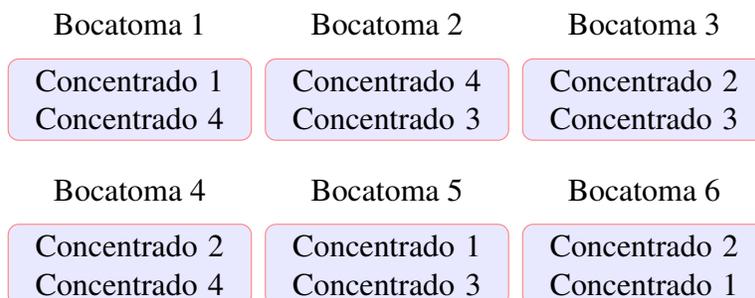
```

R Studio

Peso=c(420,390,380,360,360,410,355,390,420,360,356,412)
Hoja=data.frame(BloquesIncomp,Peso)
Hoja$Tratamiento=Hoja$Concentrado
Hoja$Bocatoma=Hoja$block
Hoja

plots block Concentrado Peso Tratamiento Bocatoma
1 1 1 Concentrado 1 420 Concentrado 1 1
2 2 1 Concentrado 4 390 Concentrado 4 1
3 3 2 Concentrado 4 380 Concentrado 4 2
4 4 2 Concentrado 3 360 Concentrado 3 2
.
12 12 6 Concentrado 1 412 Concentrado 1 6
    
```

El esquema del experimento sería:



El análisis de varianza y la comparación de medias entre tratamientos las realizaremos con el comando *BIB.test* (finding the variance analysis of the balanced incomplete block design) elaborado por Mendiburu (2012) en la librería *agricolae*. En la ejecución de este comando utilizamos las variables relacionadas con los bloques (*bocatoma*), los tratamientos (*Tratamiento*) y los pesos de los animales (*Peso*). Además, mencionamos en los argumentos que utilizaremos comparación de medias por la prueba de Tukey (*tukey*). Veamos:

```

R Studio

Salida= BIB.test(block=Hoja$Bocatoma,trt=Hoja$Tratamiento,Hoja$Peso,
               test="tukey")

Number of observations: 12
Analysis of Variance Table
Response: Hoja$Peso
      Df Sum Sq Mean Sq F value Pr(>F)
block.unadj 5 1610.4  322.08  1.7885 0.33532
trt.adj      3 5340.3 1780.08  9.8848 0.04595 *
Residuals   3  540.3  180.08
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
coefficient of variation: 3.5 %
Hoja$Peso Means: 384.4167

```

Veamos el resultado de la comparación de medias, mediante la prueba de Tukey:

```

R Studio

      means mean.adj StdError.adj
Concentrado 1 417.3333 420.9167    9.085053
Concentrado 2 357.0000 349.1667    9.085053
Concentrado 3 376.6667 376.9167    9.085053
Concentrado 4 386.6667 390.6667    9.085053
Tukey
Alpha      : 0.05
Std.err    : 9.489029
Means with the same letter are not significantly different.
Groups, Treatments and means
a  Concentrado 1  420.9
ab Concentrado 4  390.7
ab Concentrado 3  376.9
b  Concentrado 2  349.2

```

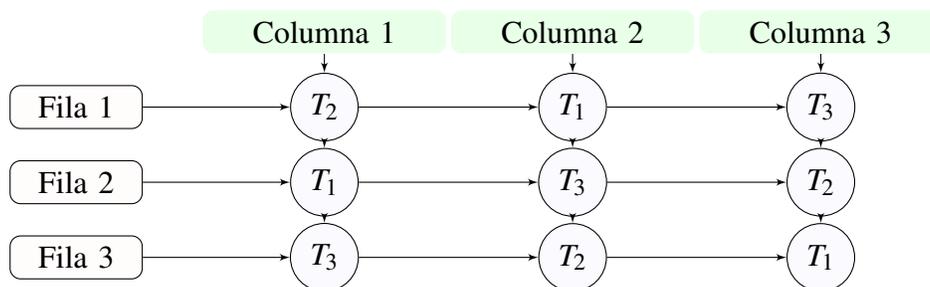
Podemos apreciar que el concentrado 1 presentó una media semejante a los concentrados 4 y 3 ( $p > 0.05$ ), pero diferente al concentrado 2 ( $p < 0.5$ ).

## Capítulo 5

### Diseño en cuadrado latino

#### 5.1. Generalidades

Este tipo de diseño se conoce también como diseño de doble bloqueo, y se caracteriza porque el número de filas, columnas y tratamientos es igual, formando un cuadrado perfecto. Adicionalmente, los tratamientos no se pueden repetir al interior de la fila o de la columna, como se indica a continuación en un arreglo de cuadrado latino con tres tratamientos:



El cuadrado latino es ampliamente utilizado en experimentos en ganado de leche, donde en las filas se ubica a las hembras en producción, y en las columnas a diversos periodos de lactancia.

Las principales ventajas del diseño en cuadrado latino son:

- ♣ Permite generar arreglos experimentales especiales, como: arreglos crossover, arreglos parcialmente confundidos, cuadrados latinos repetidos o cuadrados grecolatinos.
- ♣ Es aplicable en la estimación de superficies de respuesta.
- ♣ Se puede estimar el error experimental asociado con cada unidad experimental.
- ♣ Admite convalidar supuestos asociados con el modelo de clasificación experimental.
- ♣ Puede tener componentes o efectos fijos o aleatorios.
- ♣ Permite generar arreglos experimentales factoriales.
- ♣ Se puede utilizar transformación de datos sobre la variable respuesta.
- ♣ Admite efectuar análisis multivariado de la varianza Manova.
- ♣ Acepta componentes o factores anidados.

- ♣ Se puede efectuar análisis con submuestras.
- ♣ Acepta la presencia de covariables.
- ♣ Se pueden utilizar contrastes ortogonales.
- ♣ Permite efectuar análisis exploratorios de tipo unidimensional, sobre cada tratamiento.
- ♣ El diseño puede ser balanceado o desbalanceado.
- ♣ Se puede incorporar la técnica Mancova.

De no ser posible asegurar el efecto de doble bloqueo, se alteraría el análisis de la varianza.

Algunas aplicaciones en ciencias animales son:

- ♣ En experimentos que pretenden evaluar la dinámica del desempeño de un lote de animales cuando reciben cada uno de los tratamientos disponibles en periodos diferentes.
- ♣ Se recomienda aplicar este diseño a fin de evaluar el rendimiento de un cultivo ante la presencia de dos gradientes de fertilidad.
- ♣ Es aplicable en experimentos donde se desea evaluar la producción de operarios en turnos diferentes, ante distintos estímulos.
- ♣ Se emplea en experimentación animal para evaluar factores de producción ante la presencia de distintos tipos de manejo, condiciones de raza y suplementos suministrados.
- ♣ Se utiliza para establecer el efecto de las velocidades controladas de corrientes de agua, con diferentes pendientes y distintos tipos de alimentación en peces.

## 5.2. Supuestos estadísticos

Los supuestos son:

- ♣ Se debe verificar que los errores experimentales se distribuyan en forma normal.
- ♣ Los errores deben ser independientes estadísticamente y con distribución aleatoria.
- ♣ Las varianzas asociadas con los distintos tratamientos deben ser homogéneas.
- ♣ No debe existir relación entre los promedios de los tratamientos y las varianzas.

### 5.3. Modelos de estructura experimental

El modelo de clasificación está dado por:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \varepsilon_{k(ij)}$$

Donde:

$y_{ijk}$  es la variable respuesta, asociada al  $k$ -ésimo tratamiento en la  $j$ -ésima columna y en la  $i$ -ésima fila,

$\mu$  indica el efecto promedio del experimento,

$\tau_k$  indica el efecto promedio del  $k$ -ésimo tratamiento ( $i = 1, 2..K$ ),

$\beta_j$  representa el efecto de la  $j$ -ésima columna ( $j = 1, 2..J$ ),

$\alpha_i$  representa el efecto de la  $i$ -ésima fila ( $i = 1, 2..I$ ), y

$\varepsilon_{k(ij)}$  son los errores experimentales que se caracterizan por ser independientes e idénticamente distribuidos (iid) y tienen una distribución normal con media 0 y varianza  $\sigma_\varepsilon^2$ , cuya notación es:  $\varepsilon_{k(ij)} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$ .

El modelo es aplicable a tratamientos de carácter fijo ( $\tau$ ) o aleatorio ( $t$ ), entendiéndose este último cuando los niveles de un factor constituyen una muestra aleatoria de una población mayor de tratamientos. Las hipótesis asociadas a los tratamientos son:

En el caso de un modelo fijo:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \dots = \tau_K$$

$H_a$  : los promedios de los niveles del tratamiento son diferentes

La hipótesis nula indica que ninguno de los niveles del factor tiene un efecto significativo sobre la variable respuesta, mientras que la hipótesis alterna expresa que existe al menos un nivel que tiene efecto notorio sobre los otros niveles.

En el caso de un modelo aleatorio:

$$H_0 : \sigma_t^2 = 0$$

$$H_a : \sigma_t^2 > 0$$

Las hipótesis secundarias relacionadas con los efectos de filas y columnas son:

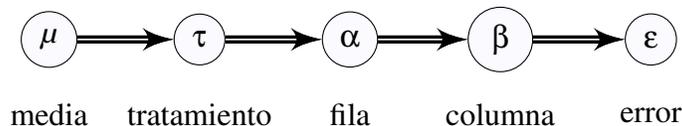
$$H_{0\alpha} : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \dots = \alpha_k$$

$H_a$  : los promedios de las filas son diferentes

$$H_{0\beta} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \dots = \beta_k$$

$H_a$  : los promedios de las columnas son diferentes

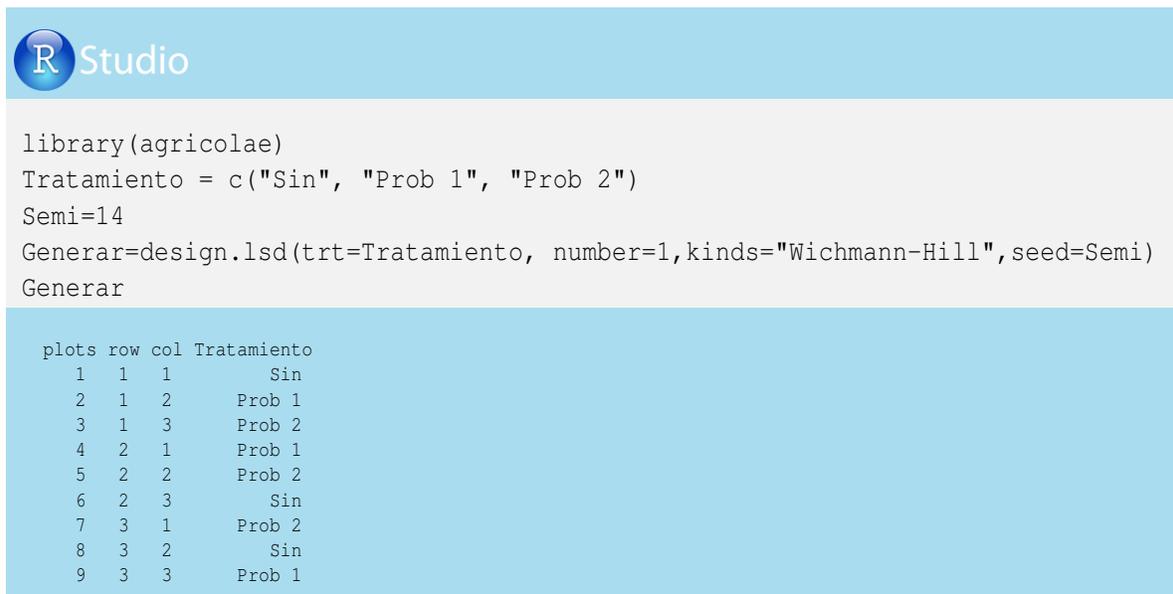
El diagrama de estructura se representa con el siguiente esquema:



## 5.4. Proceso de aleatorización

El comando *design.lsd* de la librería “*agricolae*” permite aleatorizar los tratamientos teniendo en cuenta las filas y las columnas. Los argumentos a tener en cuenta son: nombre de la hoja de datos que contiene los nombres de los tratamientos, el número inicial para enumerar secuencialmente las unidades experimentales, la especificación del método de aleatorización y la semilla (*seed*).

Veamos un ejemplo de aleatorización en R-project para un diseño de cuadrado latino con tres tratamientos (sin probiótico, probiótico 1 y probiótico 2). Las filas son las búfalas y las columnas son los periodos de lactancia (periodos de 20 días ).



```

library(agricolae)
Tratamiento = c("Sin", "Prob 1", "Prob 2")
Semi=14
Generar=design.lsd(trt=Tratamiento, number=1,kinds="Wichmann-Hill",seed=Semi)
Generar

plots row col Tratamiento
  1   1   1      Sin
  2   1   2   Prob 1
  3   1   3   Prob 2
  4   2   1   Prob 1
  5   2   2   Prob 2
  6   2   3      Sin
  7   3   1   Prob 2
  8   3   2      Sin
  9   3   3   Prob 1

```

Los tratamientos quedaron distribuidos de la siguiente forma:

	Columna	Columna	Columna
Fila	Sin	Prob 1	Prob 2
Fila	Prob 1	Prob 2	Sin
Fila	Prob 2	Sin	Prob 1

Ahora procedamos a identificar las filas con los respectivos nombres de las búfalas y las columnas con los periodos, mediante la utilización del comando *ifelse*; veamos:

```

R Studio

Generar$Bufala=ifelse(Generar$row==1, ("Neodimia"),
                      ifelse(Generar$row==2, ("Iterbia"), "Europaia"))
Generar$Periodo=ifelse(Generar$col==1, ("Primero"),
                       ifelse(Generar$col==2, ("Segundo"), "Tercero"))
Generar

plots row col Tratamiento Bufala Periodo
1 1 1 Sin Neodimia Primero
2 1 2 Prob 1 Neodimia Segundo
3 1 3 Prob 2 Neodimia Tercero
4 2 1 Prob 1 Iterbia Primero
5 2 2 Prob 2 Iterbia Segundo
6 2 3 Sin Iterbia Tercero
7 3 1 Prob 2 Europaia Primero
8 3 2 Sin Europaia Segundo
9 3 3 Prob 1 Europaia Tercero
    
```

El esquema del cuadrado latino quedó de la siguiente forma:

Búfala	Periodo		
	Primero	Segundo	Tercero
Neodimia	Sin	Prob 1	Prob 2
Iterbia	Prob 1	Prob 2	Sin
Europaia	Prob 2	Sin	Prob 1

### 5.5. Modelo

La característica a evaluar será la producción de leche:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{k(ij)}$$

Donde:  $y_{ijk}$  es la producción de leche/día,  $\mu$  es el efecto promedio del experimento,  $\alpha_i$  es el efecto promedio de la  $i$ -ésima búfala o fila ( $i = Neodimia, Iterbia$  y  $Europaia$ ),  $\beta_j$  es el efecto del  $j$ -ésimo periodo de lactancia o columna ( $j=1, 2$  y  $3$ ),  $\tau_k$  es el  $k$ -ésimo tratamiento ( $k=Sin, Probiótico 1$  y  $Probiótico 2$ ), y  $\epsilon_{k(ij)}$  son los errores experimentales ( $\epsilon_{k(ij)} \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$ ).

Las hipótesis principales asociadas a los tratamientos son:

$$H_0 : \tau_1 = \tau_2 = \tau_3$$

$H_a$  : los promedios de los niveles del tratamiento son diferentes

La hipótesis nula indica que ninguno de los tratamientos tiene un efecto significativo sobre la producción de leche.

Las hipótesis secundarias relacionadas con los efectos de filas y columnas son:

$$H_{0\alpha} : \alpha_1 = \alpha_2 = \alpha_3$$

$H_{a\alpha}$  : los promedios de las producciones de leche de las búfalas son diferentes

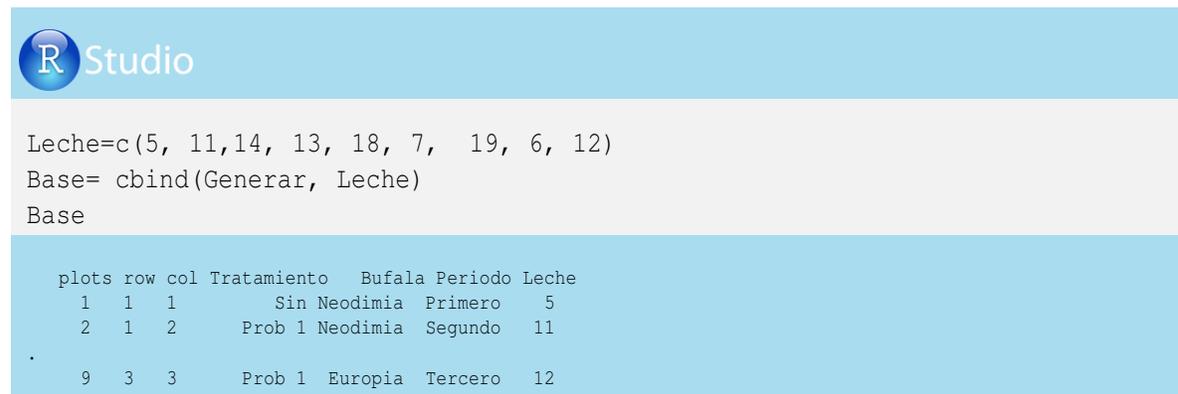
$$H_{0\beta} : \beta_1 = \beta_2 = \beta_3$$

$H_{a\beta}$  : los promedios de las producciones de leche en los periodos son diferentes

Los resultados de producción de leche/día durante los 60 días que duró el experimento fueron:

Información de producción de leche (kg/día)			
Búfala	Periodo	Tratamiento	Producción
Neodimia	Primero	Sin	5
Neodimia	Segundo	Probiótico 1	11
Neodimia	Tercero	Probiótico 2	14
Iterbia	Primero	Probiótico 1	13
Iterbia	Segundo	Probiótico 2	18
Iterbia	Tercero	Sin	7
Europia	Primero	Probiótico 2	19
Europia	Segundo	Sin	6
Europia	Tercero	Probiótico 1	12

A continuación crearemos la hoja de datos *Base* con la información de tratamiento asignado, nombre de la búfala, periodo de lactancia y producción de leche/día:



```

R Studio

Leche=c(5, 11,14, 13, 18, 7, 19, 6, 12)
Base= cbind(Generar, Leche)
Base

plots row col Tratamiento  Bufala Periodo Leche
1  1  1      Sin Neodimia  Primero  5
2  1  2      Prob 1 Neodimia  Segundo 11
.
9  3  3      Prob 1 Europia   Tercero 12

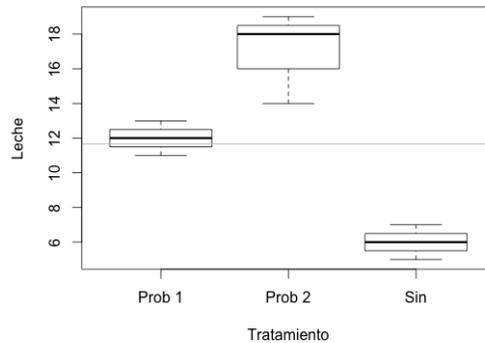
```

## 5.6. Análisis exploratorio de la hoja de datos

Realizaremos gráficos de boxplot de la producción de leche en relación con las filas (búfalas), columnas (periodos de lactancia) y tratamientos, con el objetivo de visualizar los datos y detectar errores en la toma o digitación de los datos, como se indicó en los capítulos anteriores. La programación para generar un gráfico boxplot para la producción de leche, según los tratamientos, es:



```
boxplot (Base$Leche~Base$Tratamiento,ylab="Leche",xlab="Tratamiento",data=Base)
abline (h=mean (Base$Leche), col="gray")
```

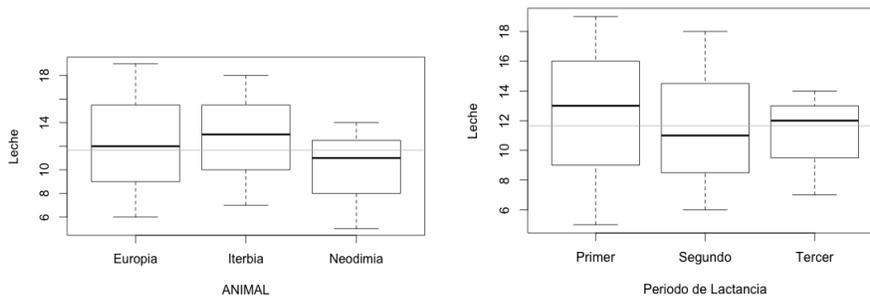


Este gráfico nos muestra que no se notaron datos extraños o que le generaran dudas al investigador, y que pueden existir diferencias entre los tratamientos, siendo el tratamiento *Prob 2* diferente al *Prob 1* y estos diferentes al tratamiento testigo. Estas diferencias serán verificadas en el análisis de varianza y en la prueba de comparación de medias.

Realicemos los gráficos de la producción de leche por búfala y por periodo de lactancia:



```
boxplot (Base$Leche~Base$Bufala,ylab="Leche",xlab="ANIMAL",data=Base)
abline (h=mean (Base$Leche), col="gray")
boxplot (Base$Leche~Base$Periodo,ylab="Leche",xlab="Periodo de Lactancia",
        data=Base)
abline (h=mean (Base$Leche), col="gray")
```



Podemos observar que el efecto del tratamiento tiene influencia mayor sobre la producción de leche (más que la diferencia de las hembras y el periodo); esto nos daría una idea de lo que esperamos encontrar en los análisis de varianza.

### 5.7. Análisis de varianza

Las sumas de cuadrados en un diseño de cuadrado latino básico son: totales, de filas, de columnas, de tratamientos y del error, y su cálculo es:

Suma de cuadrados totales ( $SCT$ ):

$$SCT = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (y_{ijk})^2 - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

Suma de cuadrados de filas  $\alpha$  ( $SC\alpha$ )

$$SC\alpha = \frac{\sum_{i=1}^I \alpha_{i..}^2}{K} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

Suma de cuadrados de columnas  $\beta$  ( $SC\beta$ )

$$SC\beta = \frac{\sum_{j=1}^J \beta_{.j.}^2}{K} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

Suma de cuadrados de tratamientos ( $SC\tau$ )

$$SC\tau = \frac{\sum_{k=1}^K \tau_{..k}^2}{K} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

Suma de cuadrados del error ( $SC\varepsilon$ )

$$SC\varepsilon = SCT - SC\alpha - SC\beta - SC\tau$$

Para nuestro ejemplo de producción de leche de búfalas sometidas a tres tratamientos, los cálculos de las sumas de cuadrados son:

Suma de cuadrados totales:

$$SCT = \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 (y_{ijk})^2 - \frac{\left( \sum_{i=1}^3 \sum_{j=1}^3 \sum_{k=1}^3 y_{ijk} \right)^2}{9}$$

$$SCT = (5)^2 + (11)^2 + \dots + (12)^2 - \frac{(5 + 11 + \dots + 12)^2}{9} = 200$$

Suma de cuadrados de filas  $\alpha$

$$SC\alpha = \frac{30^2 + 38^2 + 37^2}{3} - \frac{(5 + 11 + \dots + 12)^2}{9} = 1237.66 - 1225 = 12.66$$

Suma de cuadrados de columnas  $\beta$

$$SC\beta = \frac{37^2 + 35^2 + 33^2}{3} - \frac{(5 + 11 + \dots + 12)^2}{9} = 1227.67 - 1225 = 2.67$$

Suma de cuadrados de tratamientos ( $\tau$ )

$$SC\tau = \frac{18^2 + 36^2 + 51^2}{3} - \frac{(5 + 11 + \dots + 12)^2}{9} = 1407 - 1225 = 182$$

Suma de cuadrados del error ( $SC\epsilon$ )

$$SC\epsilon = 200 - 12.66666667 - 2.666666667 - 182 = 2.666674$$

El cuadro del análisis de varianza de un diseño de cuadrado latino básico es:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$
$\alpha$	$I - 1$	$SC\alpha$	$\frac{SC\alpha}{I-1}$	$\frac{CM\alpha}{CM\epsilon}$
$\beta$	$J - 1$	$SC\beta$	$\frac{SC\beta}{J-1}$	$\frac{CM\beta}{CM\epsilon}$
$\tau$	$K - 1$	$SC\tau$	$\frac{SC\tau}{K-1}$	$\frac{CM\tau}{CM\epsilon}$
$\epsilon$	$(K-2)(K-1)$	$SCE$	$\frac{SCE}{(K-2)(K-1)}$	$\frac{CM\epsilon}{CM\epsilon}$
Total	$n - 1$	$SCT$		

Para el ejemplo:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$
Búfalas (filas)	$3 - 1 = 2$	12.6667	6.3333	$\frac{6.333}{1.33} = 4.75$
Periodo (columnas)	$3 - 1 = 2$	2.6667	1.3333	$\frac{1.333}{1.33} = 1$
Tratamiento	$3 - 1 = 2$	182	91	$\frac{91}{1.33} = 68.25$
Error	$(3-2)(3-1)=2$	2.6667	1.3333	
Total	$9 - 1 = 8$	200		

Rechazamos la hipótesis nula de los tratamientos, si:

$$F_c > F_{\alpha, K-1, (K-2)(K-1)}$$

Donde  $F_{\alpha, K-1, (K-2)(K-1)}$  es el cuantil  $\alpha$  de la distribución de  $F$ , con grados de libertad  $(3 - 1)$  en el numerador y  $(3 - 2)(3 - 1)$  en el denominador.

Considerando un  $\alpha = 0.01$ , tenemos que el valor de  $F$  de la tabla es de 99; entonces:

$$F_c < F_{0.01,2,2}$$

$$68.25 < 99$$

Considerando un  $\alpha = 0.05$ , tenemos que el valor de F de la tabla es de 19; entonces:

$$F_c > F_{0.05,2,2}$$

$$68.25 > 19$$

Por lo tanto, decimos que hay diferencias estadísticas significativas ( $p < 0.05$ ) entre los tratamientos.

También es necesario saber si en los efectos de fila y columna se rechazan las hipótesis nulas:

En el caso de las filas:

$$F_c > F_{\alpha, I-1, (K-2)(K-1)}$$

Considerando un  $\alpha = 0.05$ , tenemos que el valor de F de la tabla es de 19, y obtenemos:

$$F_c < F_{0.01,2,2}$$

$$4.71 < 19$$

En el caso de las columnas:

$$F_c > F_{\alpha, J-1, (K-2)(K-1)}$$

Considerando un  $\alpha = 0.05$ , tenemos que el valor de F de la tabla es de 19, dando como resultado:

$$F_c < F_{0.01,2,2}$$

$$1 < 19$$

Por consiguiente, aceptamos la hipótesis nula para filas y columnas. No existieron diferencias significativas entre búfalas y entre periodos de lactancia ( $p > 0.05$ )

Para el desarrollo del análisis de varianza en R-project, utilizaremos el comando *aov*, cuyos resultados saldrán en la hoja de datos *Analisis*. Posteriormente solicitaremos un resumen del análisis de varianza para mirar los resultados con el comando *summary*; veamos:

```

R Studio

Analisis=aov(Base$Leche~Base$col + Base$row +Base$Tratamiento, data=Base)
summary (Analisis)

      Df Sum Sq Mean Sq F value Pr(>F)
Base$col  2  2.67   1.33   1.00 0.5000
Base$row  2 12.67   6.33   4.75 0.1739
Base$Tratamiento 2 182.00  91.00  68.25 0.0144 *
Residuals  2  2.67   1.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Observamos que el efecto del tratamiento fue significativo ( $p < 0.05$ ) y los efectos de filas (búfalas) y columnas (periodos) no fueron significativos ( $p > 0.05$ ).

A continuación obtendremos las soluciones y los promedios de la producción de leche para los niveles de los efectos considerados en el modelo:

```

R Studio

model.tables (Analisis)

Base$col
  1      2      3
0.6667 0.0000 -0.6667
Base$row
  1      2      3
-1.6667 1.0000 0.6667
Base$Tratamiento
Prob 1 Prob 2 Sin
0.333  5.333 -5.667
    
```

Los promedios de la producción de leche por nivel los obtendremos con el comando *model.tables*, con los argumentos: nombre de la hoja de datos que contiene el resultado del análisis y el argumento *mean*; veamos:



```

model.tables(Analysis, "mean")

Tables of means
Grand mean
11.66667
Base$col
  1     2     3
12.333 11.667 11.000
Base$row
  1     2     3
10.000 12.667 12.333
Base$Tratamiento
Prob 1 Prob 2 Sin
  12     17     6

```

Detallemos los resultados suministrados en las dos programaciones anteriores:

$$\hat{\mu} = 11.66667$$

En el caso de las filas:

$$\hat{\alpha}_{Europa} = 11.66667 - 1.66667 = 10$$

$$\hat{\alpha}_{Iterbia} = 11.66667 + 1 = 12.66667$$

$$\hat{\alpha}_{Neodimia} = 11.66667 + 0.66667 = 12.33333$$

En el caso de las columnas:

$$\hat{\beta}_{Primero} = 11.66667 + 0.66667 = 12.3333$$

$$\hat{\beta}_{Segundo} = 11.66667 + 0 = 11.66667$$

$$\hat{\beta}_{Tercero} = 11.66667 - 0.66667 = 11$$

En el caso de los tratamientos:

$$\hat{\beta}_{Prob1} = 11.66667 + 0.33333 = 12$$

$$\hat{\beta}_{Prob2} = 11.66667 + 5.33333 = 17$$

$$\hat{\beta}_{Sin} = 11.66667 - 5.66667 = 6$$

## 5.8. Prueba de normalidad

Comprobemos el supuesto de normalidad de los errores ( $\epsilon_{k(ij)} \sim N(0, \sigma_{\epsilon}^2)$ ). Para esto obtenemos los residuos de las unidades experimentales, teniendo en cuenta el modelo utilizado:

Si:  $y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \epsilon_{k(ij)}$ , entonces:

$$\epsilon_{k(ij)} = y_{ijk} - \mu - \alpha_i - \beta_j - \tau_k$$

Al reemplazar el efecto promedio de cada efecto ( $\hat{\alpha}$ ,  $\hat{\beta}$  y  $\hat{\tau}$ ) y el efecto promedio general ( $\hat{\mu}$ ), se tiene:

$$\epsilon_{k(ij)} = y_{ijk} - \hat{\mu} - (\hat{\alpha}_i - \hat{\mu}) - (\hat{\beta}_j - \hat{\mu}) - (\hat{\tau}_k - \hat{\mu})$$

El error de la observación asociado a la búfala Neodimia, en el primer periodo del experimento y al tratamiento testigo, es:

$$\epsilon_{Sin(Neodimia,Primero)} = 5 - 11.66667 - (-1.6667) - 0.66667 - (-5.6667) = 0.0$$

El error experimental de la observación asociado a Neodimia en el segundo periodo y al probiótico 1 es:

$$\epsilon_{Prob1(Neodimia,Segundo)} = 11 - 11.66667 - (-1.66667) - 0 - 0.33 = 0.067$$

El error experimental de la observación asociada a la búfala Europa en el tercer periodo y al probiótico 1 es:

$$\epsilon_{prob1(Europa,Tercero)} = 12 - 11.67 - 0.67 - (-0.67) - 0.33 = 0$$

Para obtener los residuos utilizamos el comando *residuals* y para los valores predichos utilizamos el comando *fitted*, veamos:

```
R Studio

Base$Residuos=residuals(Analisis)
Base$Predichos=fitted(Analisis)
Base

plots row col Tratamiento  Bufala Periodo Leche      Residuos Predichos
  1   1   1      Sin Neodimia  Primero    5 -5.412337e-16  5.000000
  2   1   2      Prob 1 Neodimia  Segundo   11  6.666667e-01 10.333333
  3   1   3      Prob 2 Neodimia  Tercero   14 -6.666667e-01 14.666667
  4   2   1      Prob 1 Iterbia   Primero   13 -6.666667e-01 13.666667
  5   2   2      Prob 2 Iterbia   Segundo   18  3.469447e-18 18.000000
  6   2   3      Sin Iterbia   Tercero    7  6.666667e-01  6.333333
  7   3   1      Prob 2 Europa    Primero   19  6.666667e-01 18.333333
  8   3   2      Sin Europa    Segundo    6 -6.666667e-01  6.666667
  9   3   3      Prob 1 Europa    Tercero   12 -2.602085e-16 12.000000
```

Encontremos la media y la varianza de los errores:

```
R Studio

summary(Base$Residuos)
var(Base$Residuos)
mean(Base$Residuos)

  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.6667 -0.6667  0.0000  0.0000  0.6667  0.6667
> var(Base$Residuos)
[1] 0.3333333
> mean(Base$Residuos)
[1] 1.001749e-17
```

La media de los errores fue 0 con varianza 0.33 ( $\epsilon_{k(ij)}$  tiene  $\mu_\epsilon = 0$  y  $\sigma_\epsilon^2 = 0.33$ ). Apliquemos la prueba de Shapiro-Wilk con el comando `shapiro.test`, para verificar la distribución de los residuos:

```

R Studio

ShapiroResiduos=shapiro.test(Base$Residuos)
ShapiroResiduos

data: Base$Residuos
W = 0.823, p-value = 0.03729

```

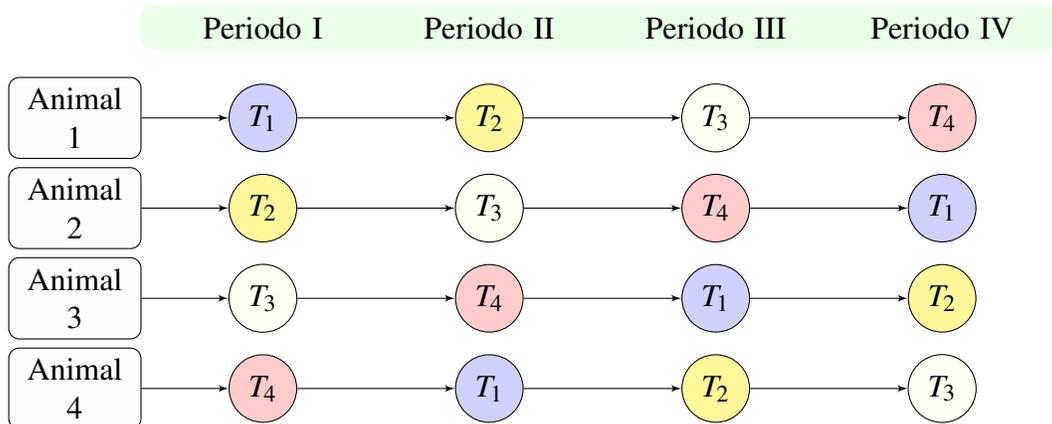
Por presentar un valor de  $p < 0.05$ , los  $\epsilon_{k(ij)}$  **no** se distribuyeron normalmente, por lo tanto, hay que transformar los datos (ver capítulo 7).

### 5.9. Efectos de arrastre de los tratamientos previos

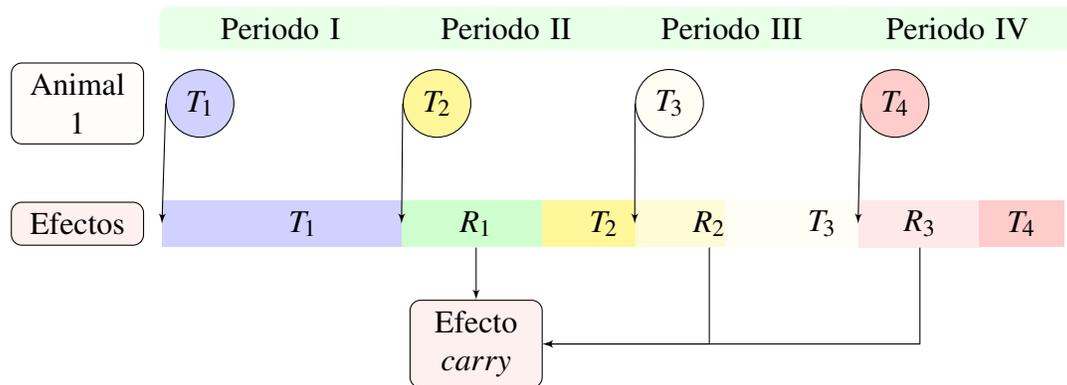
Cuando se realizan cuadrados latinos donde las filas son individuos y las columnas son periodos, se presume que el individuo puede estar afectado por el tratamiento aplicado anteriormente. Este efecto residual (secuelas del tratamiento previo, efecto de arrastre o efecto *carry*) puede ser detectado utilizando una modificación al cuadrado latino denominado “cross over” o de reversión, propuesto por Cochran et al. (1941), como una necesidad de experimentación en ganado lechero, y es muy útil cuando no es posible garantizar un adecuado periodo de descanso entre la aplicación de tratamientos.

Si no se considera el efecto *carry* en el análisis de varianza, se pueden obtener resultados enmascarados de un tratamiento. Incluso, un tratamiento puede tener un efecto de arrastre mayor que otro tratamiento o perdurar por más de un periodo.

Para explicar el efecto *carry* desarrollaremos a continuación un ejemplo con un arreglo de cuadrado latino de cuatro tratamientos, cuatro periodos y cuatro animales:



En el siguiente esquema, podemos observar el efecto *carry* de la aplicación de los cuatro tratamientos al primer animal.



Como se indicó anteriormente, el primer tratamiento aplicado al animal fue el  $T_1$ , el cual no fue afectado por un tratamiento previo, pero a partir del segundo tratamiento encontramos que existe el efecto *carry*, donde los tratamientos aplicados desde el segundo periodo se ven afectados por el tratamiento anterior. En el caso del  $T_2$ , se ve afectado por el  $T_1$ , el  $T_3$  por el  $T_2$  y el  $T_4$  por el  $T_3$ .

Con respecto a este tema, existen diversas formas de estructura del diseño de cuadrado latino que incluyen el efecto *carry*, entre ellas están:

- ♣ El cuadrado latino con algunas especificaciones (las veremos más adelante)
- ♣ Secuencias de cuadrados latinos (varios cuadrados latinos)
- ♣ Rectángulos latinos, los cuales presentan mayor número de individuos que tratamientos y el número de tratamientos es igual al número de columnas. En este libro nos centraremos en el cuadrado latino modificado, caracterizado por el número de columnas igual al número de filas y de tratamientos. También se debe garantizar que todos los tratamientos sean precedidos por otro tratamiento, el mismo número de veces.

También se encuentran diversas formas de realizar o incluir el efecto *carry* en el modelo y de comprobar su ortogonalidad; en este libro desarrollaremos una de ellas.

El ejemplo que desarrollaremos es cuadrado latino con las siguientes especificaciones:

- ♣ El número de tratamientos es par
- ♣ No incluiremos secuencias de cuadrados latinos (solo un cuadrado latino)
- ♣ El número de tratamientos es igual al número de periodos y de animales (cuatro)
- ♣ Tendremos en cuenta la secuela del tratamiento inmediatamente anterior (*carry* de primer orden)

Consideraremos el siguiente modelo:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \sum_{l=1}^{L=(K-1)} c\gamma_l + \varepsilon_{k(ij)}$$

Donde  $y_{ijk}$  es la variable respuesta,  $\mu$  indica el efecto promedio del experimento,  $\alpha_i$  representa el efecto de la  $i$ -ésima fila ( $i = 1, 2, \dots, I$ ),  $\beta_j$  representa el efecto de la  $j$ -ésima columna ( $j = 1, 2, \dots, J$ ),  $\tau_k$  indica el efecto promedio del  $k$ -ésimo tratamiento ( $k = 1, 2, \dots, K$ ), y  $\varepsilon_{k(ij)}$  es el error experimental, con distribución normal, media cero y varianza  $\sigma_\varepsilon^2$ .

El efecto *carry* estaría dado por:

$$\sum_{l=1}^{L=(K-1)} c\gamma_l = c\gamma_{l_1=k_1} + c\gamma_{l_2=k_2} + \dots + c\gamma_{L=(K-1)}$$

Donde  $c$  puede asumir los valores de  $-1, 0$  y  $1$ , cuando se cumplen las siguientes condiciones:

- ♣  $c = 1$ , Si ese tratamiento fue aplicado al animal en el periodo anterior
- ♣  $c = 0$ , Si ese tratamiento no fue aplicado al animal en el periodo anterior
- ♣  $c = -1$ , Si el tratamiento anterior es el último nivel del tratamiento ( $K$ )
- ♣  $c = 0$ , Si es el primer periodo

Desarrollemos un ejercicio de la aplicación de un suplemento protéico aplicado a cuatro vacas y cuatro periodos de diez días. Los tratamientos son suplementos alimenticios: dos dosis de un producto comercial ( $T_1$  y  $T_2$ ) y dos productos naturales ( $T_3$  y  $T_4$ ), cuyo modelo es:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_k + \sum_{l=1}^{L=(4-1)} c\gamma_l + \varepsilon_{kl(ij)}$$

Donde:  $y_{ijk}$  es el nivel de nitrógeno ureico en leche ( $\frac{mg}{dl}$ ),  $\mu$  indica el efecto promedio del experimento,  $\alpha_i$  representa el efecto de la  $i$ -ésima vaca ( $i=05-10, 14-10, 23-10, 33-10$ ),  $\beta_j$  representa el efecto del  $j$ -ésimo periodo ( $j=1, 2, 3$  y  $4$ ),  $\tau_k$  indica el efecto promedio del  $k$ -ésimo suplemento alimenticio ( $k=\tau_1, \tau_2, \tau_3$  y  $\tau_4$ ),  $\sum_{l=1}^{L=(4-1)} c\gamma_l = c\gamma_{\tau_1} + c\gamma_{\tau_2} + c\gamma_{\tau_3}$  es el efecto *carry* (será explicado más adelante), y  $\varepsilon_{k(ij)}$  es el error experimental, con distribución normal, media cero y varianza  $\sigma_\varepsilon^2$ .

Las hipótesis asociadas a los tratamientos son:

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4$$

$H_a$  : los promedios de los niveles del tratamiento son diferentes

Las hipótesis secundarias relacionadas con los efectos de filas y columnas son:

$$H_{0\alpha} : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$$

$H_a$  : los promedios de las vacas son diferentes

$$H_{0\beta} : \beta_1 = \beta_2 = \beta_3 = \beta_4$$

$H_a$  : los promedios de los periodos son diferentes

Las hipótesis para el efecto *carry* serían:

$$H_{0\gamma_{\tau_1}} : \gamma_{\tau_1} = 0$$

$$H_{0\gamma_{\tau_1}} : \gamma_{\tau_1} \neq 0$$

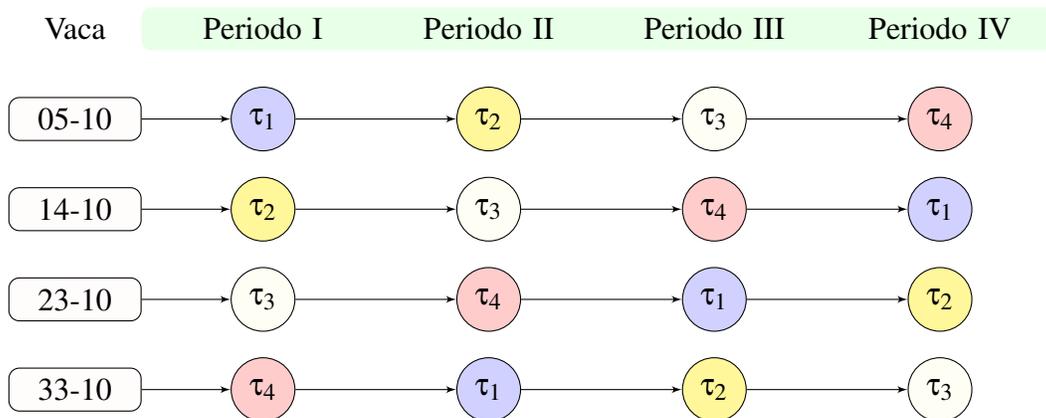
$$H_{0\gamma_{\tau_2}} : \gamma_{\tau_2} = 0$$

$$H_{0\gamma_{\tau_2}} : \gamma_{\tau_2} \neq 0$$

$$H_{0\gamma_{\tau_3}} : \gamma_{\tau_3} = 0$$

$$H_{0\gamma_{\tau_3}} : \gamma_{\tau_3} \neq 0$$

Veamos en la siguiente figura la distribución de los tratamientos por vacas y periodos:



A continuación asignaremos los coeficientes del efecto *carry* de los tratamientos para algunas vacas y periodos:

Si el efecto *carry* está dado por:  $\sum_{l=1}^{L=(4-1)} c\gamma_l = c\gamma_{\tau_1} + c\gamma_{\tau_2} + c\gamma_{\tau_3}$ , entonces:

En el caso de la vaca 05-10 en el primer periodo, no hay efecto *carry* porque no existió la aplicación de un tratamiento previo, por tanto:

$$0\gamma_{\tau_1} + 0\gamma_{\tau_2} + 0\gamma_{\tau_3}$$

En el caso de la vaca 05-10 en el periodo II, existe un efecto *carry* para  $\tau_1$ , porque este tratamiento fue aplicado en el periodo I, por tanto:

$$1\gamma_{\tau_1} + 0\gamma_{\tau_2} + 0\gamma_{\tau_3}$$

La vaca 05-10 en el periodo III tiene efecto *carry* para  $\tau_2$ :

$$0\gamma_{\tau_1} + 1\gamma_{\tau_2} + 0\gamma_{\tau_3}$$

La vaca 05-10 el periodo IV tiene efecto *carry* para  $\tau_3$ :

$$0\gamma_{\tau_1} + 0\gamma_{\tau_2} + 1\gamma_{\tau_3}$$

En resumen, los cuatro periodos de la vaca 05-10 quedaron así:

$$0\gamma_{\tau_1} + 0\gamma_{\tau_2} + 0\gamma_{\tau_3}$$

$$1\gamma_{\tau_1} + 0\gamma_{\tau_2} + 0\gamma_{\tau_3}$$

$$0\gamma_{\tau_1} + 1\gamma_{\tau_2} + 0\gamma_{\tau_3}$$

$$0\gamma_{\tau_1} + 0\gamma_{\tau_2} + 1\gamma_{\tau_3}$$

Cuando el tratamiento previo es el último nivel del tratamiento (en este caso  $\tau_4$ ), todos los coeficientes del efecto *carry* son  $-1$ , porque se tiene una restricción al último nivel para tenerlo en cuenta en los  $k - 1$  efectos *carry*. Es el caso de la vaca 14-10 en el periodo IV, donde el tratamiento aplicado en el periodo III fue el  $\tau_4$ , de modo que queda:

$$-1\gamma_{\tau_1} - 1\gamma_{\tau_2} - 1\gamma_{\tau_3}$$

Esto también ocurre en la vaca 23-10 en el periodo III y en la 33-10 en el periodo II.

Veamos la hoja de datos con los resultados del experimento, incluyendo el efecto *carry*:

Información de niveles de úrea en leche ( $\frac{mg}{dl}$ ) y efecto <i>carry</i>						
Vaca	Periodo	Tratamiento	Carry $\tau_1$	Carry $\tau_2$	Carry $\tau_3$	MUN
05-10	I	$T_1$	0	0	0	4.1
05-10	II	$T_2$	1	0	0	13.1
05-10	III	$T_3$	0	1	0	14.1
05-10	IV	$T_4$	0	0	1	23.1
14-10	I	$T_2$	0	0	0	12.5
14-10	II	$T_3$	0	1	0	14.3
14-10	III	$T_4$	0	0	1	23.2
14-10	IV	$T_1$	-1	-1	-1	7
23-10	I	$T_3$	0	0	0	12.9
23-10	II	$T_4$	0	0	1	23.1
23-10	III	$T_1$	-1	-1	-1	6.6
23-10	IV	$T_2$	1	0	0	13.3
33-10	I	$T_4$	0	0	0	21.3
33-10	II	$T_1$	-1	-1	-1	7
33-10	III	$T_2$	1	0	0	13.5
33-10	IV	$T_3$	0	1	0	14.3

Si observamos detalladamente la hoja de datos anterior, comprobaremos que los valores de  $c$  para *carry* de los tres tratamientos presentaron valores de 0, 1 y  $-1$  en 10, 3 y 3 oportunidades, respectivamente.

Construyamos la hoja de datos con las vacas, los periodos y los tratamientos en R-project:



```
Vaca=c("05-10","05-10","05-10","05-10",
      "14-10","14-10","14-10","14-10",
      "23-10","23-10","23-10","23-10",
      "33-10","33-10","33-10","33-10")
Trat=c("T1","T2","T3","T4",
      "T2","T3","T4","T1",
      "T3","T4","T1","T2",
      "T4","T1","T2","T3")
Peri=c("I","II","III","IV",
      "I","II","III","IV",
      "I","II","III","IV",
      "I","II","III","IV")
```

Ahora incluyamos los resultados de los niveles de úrea en leche:



```
Mun=c(4.1,13.1,14.1,23.1,12.5,14.3,23.2,7,12.9,23.1, 6.6,13.3,21.3,7,13.5,14.3)
Hoja=data.frame(Vaca, Peri, Trat, Mun)
Hoja
```

	Vaca	Peri	Trat	Mun
1	05-10	I	T1	4.1
2	05-10	II	T2	13.1
3	05-10	III	T3	14.1
4	05-10	IV	T4	23.1
5	14-10	I	T2	12.5
6	14-10	II	T3	14.3
7	14-10	III	T4	23.2
8	14-10	IV	T1	7.0
9	23-10	I	T3	12.9
10	23-10	II	T4	23.1
11	23-10	III	T1	6.6
12	23-10	IV	T2	13.3
13	33-10	I	T4	21.3
14	33-10	II	T1	7.0
15	33-10	III	T2	13.5
16	33-10	IV	T3	14.3

Verifiquemos la información y el tipo de variables de la hoja de datos, para comprobar que los datos estén completos y que la variable respuesta sea de tipo numérico y los efectos (vaca, periodo y tratamiento) sean de tipo factor, para no tener inconvenientes al realizar las posteriores programaciones en R-project. Veamos:



```
summary(Hoja)
str(Hoja)

> summary(Hoja)
  Vaca   Peri   Trat      Mun
05-10:4 I   :4   T1:4   Min.   : 4.10
14-10:4 II  :4   T2:4   1st Qu.:11.12
23-10:4 III:4   T3:4   Median :13.40
33-10:4 IV  :4   T4:4   Mean    :13.96
                3rd Qu.:16.05
                Max.    :23.20

> str(Hoja)
'data.frame': 16 obs. of  4 variables:
 $ Vaca: Factor w/ 4 levels "05-10","14-10",...: 1 1 1 1 2 2 2 3 3 ...
 $ Peri: Factor w/ 4 levels "I","II","III",...: 1 2 3 4 1 2 3 4 1 2 ...
 $ Trat: Factor w/ 4 levels "T1","T2","T3",...: 1 2 3 4 2 3 4 1 3 4 ...
 $ Mun : num  4.1 13.1 14.1 23.1 12.5 14.3 23.2 7 12.9 23.1 ...
```

Las variables *Vaca*, *Peri* y *Trat* son alfanuméricas (*Factor*) y *Mun* es numérica.

Ahora vamos a crear las variables *carry*. Para esto tenemos que comprobar que los datos estén ordenados por vaca y periodo. Utilizaremos la siguiente función y crearemos la *Hoja1* con los datos ordenados:



```
Hoja1=Hoja[do.call(order,Hoja[c("Vaca","Peri")]),]
```

El lector podrá utilizar otro tipo de comandos para ordenar los datos, como es el caso de *orderBy* de la librería *doBy* de Højsgard y Halekoh (2012).

Ahora creamos una función a la que la llamamos *LAG*, para dar origen a una variable que baje en una línea los datos que aparecen en la variable *Trat*:



```
FunLag=function(LAG)
{
  MenosUltimo=matrix(LAG[1:(length(LAG)-1)]);
  rbind(NA,MinusUltimo)
}
```

Aplicamos la función *LAG* a la hoja de datos:

```

R Studio

Hojal$LagTrat=FunLag(Hojal$Trat)
Hojal

```

	Vaca	Peri	Trat	Mun	LagTrat
1	05-10	I	T1	4.1	<NA>
2	05-10	II	T2	13.1	T1
3	05-10	III	T3	14.1	T2
4	05-10	IV	T4	23.1	T3
5	14-10	I	T2	12.5	T4
6	14-10	II	T3	14.3	T2
7	14-10	III	T4	23.2	T3
8	14-10	IV	T1	7.0	T4
9	23-10	I	T3	12.9	T1
10	23-10	II	T4	23.1	T3
11	23-10	III	T1	6.6	T4
12	23-10	IV	T2	13.3	T1
13	33-10	I	T4	21.3	T2
14	33-10	II	T1	7.0	T4
15	33-10	III	T2	13.5	T1
16	33-10	IV	T3	14.3	T2

A continuación le asignaremos la expresión *Nova* la variable *LagTrat*, cuando sea el primer periodo del experimento. Recordemos que no existe efecto *carry*:

```

R Studio

Hojal$LagTrat=ifelse (Hojal$Peri=="I", "No",Hojal$LagTrat)
Hojal

```

	Vaca	Peri	Trat	Mun	LagTrat
1	05-10	I	T1	4.1	No
2	05-10	II	T2	13.1	T1
3	05-10	III	T3	14.1	T2
4	05-10	IV	T4	23.1	T3
5	14-10	I	T2	12.5	No
6	14-10	II	T3	14.3	T2
7	14-10	III	T4	23.2	T3
8	14-10	IV	T1	7.0	T4
9	23-10	I	T3	12.9	No
10	23-10	II	T4	23.1	T3
11	23-10	III	T1	6.6	T4
12	23-10	IV	T2	13.3	T1
13	33-10	I	T4	21.3	No
14	33-10	II	T1	7.0	T4
15	33-10	III	T2	13.5	T1
16	33-10	IV	T3	14.3	T2

Ahora creamos la variable *CarryT1* con valores de  $-1$ ,  $0$  y  $1$ . El valor  $0$  será asignado cuando el tratamiento anterior no fue el  $\tau_1$ , el valor de  $1$  cuando el tratamiento anterior fue el  $\tau_1$  y de  $-1$  cuando el tratamiento anterior fue el  $\tau_4$ :



```

Hojal$CarryT1=ifelse (Hojal$LagTrat=="T1",1,0)
Hojal$CarryT1=ifelse (Hojal$LagTrat=="T4",-1,Hojal$CarryT1)
Hojal

```

	Vaca	Peri	Trat	Mun	LagTrat	CarryT1
1	05-10	I	T1	4.1	No	0
2	05-10	II	T2	13.1	T1	1
3	05-10	III	T3	14.1	T2	0
4	05-10	IV	T4	23.1	T3	0
5	14-10	I	T2	12.5	No	0
6	14-10	II	T3	14.3	T2	0
7	14-10	III	T4	23.2	T3	0
8	14-10	IV	T1	7.0	T4	-1
9	23-10	I	T3	12.9	No	0
10	23-10	II	T4	23.1	T3	0
11	23-10	III	T1	6.6	T4	-1
12	23-10	IV	T2	13.3	T1	1
13	33-10	I	T4	21.3	No	0
14	33-10	II	T1	7.0	T4	-1
15	33-10	III	T2	13.5	T1	1
16	33-10	IV	T3	14.3	T2	0

De igual forma creamos las variables *CarryT2* y *CarryT3*:



```

Hojal$CarryT2=ifelse (Hojal$LagTrat=="T2",1,0)
Hojal$CarryT2=ifelse (Hojal$LagTrat=="T4",-1,Hojal$CarryT2)
Hojal$CarryT3=ifelse (Hojal$LagTrat=="T3",1,0)
Hojal$CarryT3=ifelse (Hojal$LagTrat=="T4",-1,Hojal$CarryT3)
Hojal

```

	Vaca	Peri	Trat	Mun	LagTrat	CarryT1	CarryT2	CarryT3
1	05-10	I	T1	4.1	No	0	0	0
2	05-10	II	T2	13.1	T1	1	0	0
3	05-10	III	T3	14.1	T2	0	1	0
4	05-10	IV	T4	23.1	T3	0	0	1
5	14-10	I	T2	12.5	No	0	0	0
6	14-10	II	T3	14.3	T2	0	1	0
7	14-10	III	T4	23.2	T3	0	0	1
8	14-10	IV	T1	7.0	T4	-1	-1	-1
9	23-10	I	T3	12.9	No	0	0	0
10	23-10	II	T4	23.1	T3	0	0	1
11	23-10	III	T1	6.6	T4	-1	-1	-1
12	23-10	IV	T2	13.3	T1	1	0	0
13	33-10	I	T4	21.3	No	0	0	0
14	33-10	II	T1	7.0	T4	-1	-1	-1
15	33-10	III	T2	13.5	T1	1	0	0
16	33-10	IV	T3	14.3	T2	0	1	0

Verificamos que las variables *CarryT1*, *CarryT2* y *CarryT3* sean numéricas:

```

R Studio

str(Hojal)

'data.frame': 16 obs. of 8 variables:
 $ Vaca   : Factor w/ 4 levels "05-10","14-10",...: 1 1 1 1 2 2 2 2 3 3 ...
 $ Peri   : Factor w/ 4 levels "I","II","III",...: 1 2 3 4 1 2 3 4 1 2 ...
 $ Trat   : Factor w/ 4 levels "T1","T2","T3",...: 1 2 3 4 2 3 4 1 3 4 ...
 $ Mun    : num  4.1 13.1 14.1 23.1 12.5 14.3 23.2 7 12.9 23.1 ...
 $ LagTrat: chr  "No" "T1" "T2" "T3" ...
 $ CarryT1: num  0 1 0 0 0 0 0 -1 0 0 ...
 $ CarryT2: num  0 0 1 0 0 1 0 -1 0 0 ...
 $ CarryT3: num  0 0 0 1 0 0 1 -1 0 1 ...
    
```

Para verificar si es necesario incluir el efecto *carry*, realizaremos una comparación de dos análisis de varianza (sin incluir e incluyendo este efecto). A continuación presentamos la programación y la discusión del análisis sin la inclusión de este efecto. El análisis de varianza lo realizaremos con el comando *lm*:

```

R Studio

Resultado=lm(Mun~Vaca+Peri+Trat, data=Hojal)
summary.aov(Resultado)

      Df Sum Sq Mean Sq  F value    Pr(>F)
Vaca    3    0.9    0.29    1.918  0.2279
Peri    3    8.5    2.84   18.709  0.0019 **
Trat    3  549.2  183.07 1207.038 9.9e-09 ***
Residuals 6    0.9    0.15

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Tengamos en cuenta el valor de la suma de cuadrados del efecto del tratamiento (549.2) para compararlo posteriormente con el encontrado en el análisis de varianza incluyendo el efecto *carry*, que se presenta a continuación:



```
ResultadoCarry=aov(Mun~Vaca+Peri+Trat+CarryT1+CarryT2+CarryT3, data=Hoja1)
summary(ResultadoCarry)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Vaca	3	0.9	0.29	23.267	0.014025	*
Peri	3	8.5	2.84	227.000	0.000492	***
Trat	3	549.2	183.07	14645.400	9.58e-07	***
CarryT1	1	0.7	0.68	54.450	0.005146	**
CarryT2	1	0.1	0.11	8.817	0.059102	.
CarryT3	1	0.1	0.08	6.533	0.083502	.
Residuals	3	0.0	0.01			

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Vemos que el efecto *carry* del tratamiento  $\tau_1$  fue altamente significativo ( $p < 0.01$ ) y los tratamientos  $\tau_2$  y  $\tau_3$  fueron significativos ( $p < 0.05$ ).

Por la estructura del análisis de varianza, los efectos *carry* fueron corregidos (suma de cuadrados tipo I), pero también es necesario ajustar el efecto del tratamiento por los efectos *carry*, para lo cual utilizamos la suma de cuadrados tipo III. Para esto, utilizaremos el comando *Anova* de la librería *car* de Fox y Weisberg (2011) y le indicaremos que se obtenga la suma de cuadrados tipo III:



```
install.packages("car")
library(car)
ResultadoCarryIII=Anova(ResultadoCarry, type=c("III"))
ResultadoCarryIII
```

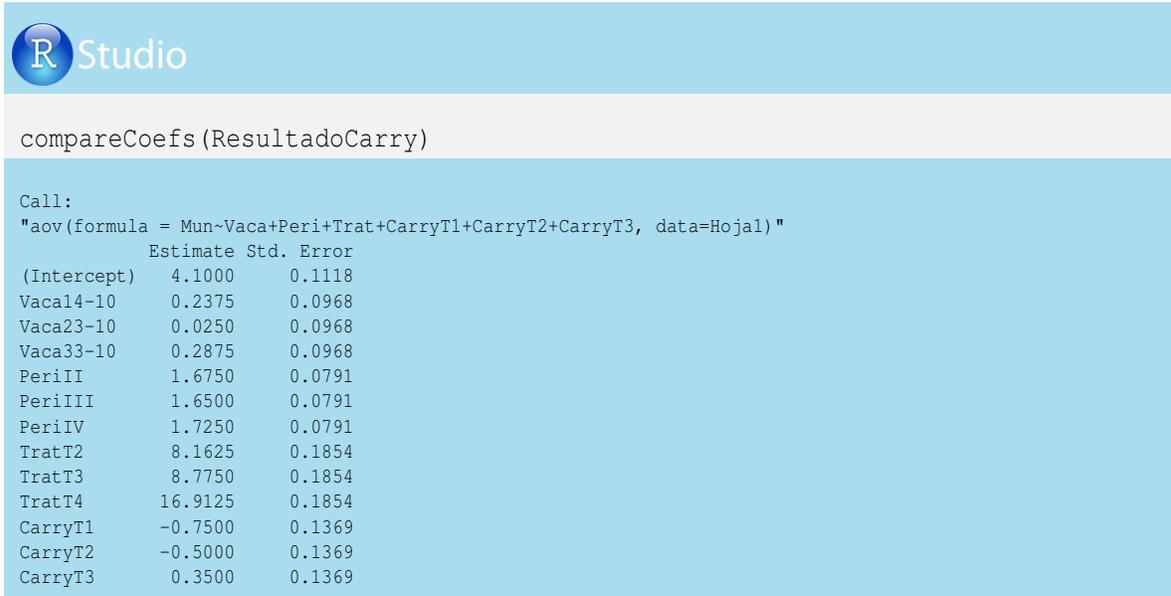
Anova Table (Type III tests)  
Response: Mun

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	16.810	1	1344.8000	4.460e-05	***
Vaca	0.171	3	4.5556	0.1224023	
Peri	8.512	3	227.0000	0.0004925	***
Trat	104.148	3	2777.2909	1.159e-05	***
CarryT1	0.375	1	30.0000	0.0119669	*
CarryT2	0.167	1	13.3333	0.0354555	*
CarryT3	0.082	1	6.5333	0.0835019	.
Residuals	0.038	3			

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Vemos que la suma de cuadrados del efecto del tratamiento se redujo de 549.2 a 104.15, por el ajuste del efecto *carry*. Lo que indica la importancia de incluir el efecto *carry* en el experimento, que estaba enmascarando el resultado del análisis de varianza.

Ahora obtengamos las estimaciones de los efectos residuales de los tratamientos (*CarryT1*, *CarryT2* y *CarryT3*), con el comando *compareCoeefs*:



```

R Studio
compareCoeefs (ResultadoCarry)

Call:
"aov(formula = Mun~Vaca+Peri+Trat+CarryT1+CarryT2+CarryT3, data=Hojal)"

      Estimate Std. Error
(Intercept)  4.1000    0.1118
Vaca14-10    0.2375    0.0968
Vaca23-10    0.0250    0.0968
Vaca33-10    0.2875    0.0968
PeriII       1.6750    0.0791
PeriIII      1.6500    0.0791
PeriIV       1.7250    0.0791
TratT2       8.1625    0.1854
TratT3       8.7750    0.1854
TratT4      16.9125    0.1854
CarryT1      -0.7500    0.1369
CarryT2      -0.5000    0.1369
CarryT3       0.3500    0.1369
    
```

Los valores estimados para los efectos *carry* de los tratamientos son:

$$\hat{\gamma}_{\tau_1} = -0.75$$

$$\hat{\gamma}_{\tau_2} = -0.5$$

$$\hat{\gamma}_{\tau_3} = 0.35$$

$$\hat{\gamma}_{\tau_4} = -\hat{\gamma}_{\tau_1} - \hat{\gamma}_{\tau_2} - \hat{\gamma}_{\tau_3} = 0.75 + 0.5 - 0.35 = 0.9$$

Los tratamientos  $\tau_1$  y  $\tau_2$  tienen un efecto residual *carry*, que reduce los niveles de úrea en la leche en el periodo siguiente, contrario a los tratamientos  $\tau_3$  y  $\tau_4$ .

Las medias ajustadas para los tratamientos en los dos modelos (sin y con efecto *carry*) las podemos obtener con la librería *lsmeans* (least-squares means) de Lenth et al. (2012), que requiere la instalación previa de la librería *multcomp* de Hothorn et al. (2012):



```
install.packages("multcomp")
library("multcomp")
install.packages("lsmeans")
library(lsmeans)
lsmeans(Resultado, ~Trat)
lsmeans(ResultadoCarry, ~Trat)
```

```
> lsmeans(Resultado, ~Trat)
$`Trat lsmeans`
      lsmean      SE df lower.CL upper.CL
T1  6.175 0.194722  6  5.698532  6.651468
T2 13.100 0.194722  6 12.623532 13.576468
T3 13.900 0.194722  6 13.423532 14.376468
T4 22.675 0.194722  6 22.198532 23.151468
> lsmeans(ResultadoCarry, ~Trat)
$`Trat lsmeans`
      lsmean      SE df lower.CL upper.CL
T1  5.5000 0.1169268  3  5.127887  5.872113
T2 13.6625 0.1169268  3 13.290387 14.034613
T3 14.2750 0.1169268  3 13.902887 14.647113
T4 22.4125 0.1169268  3 22.040387 22.784613
```

Vemos cómo cambiaron las medias de los tratamientos en los dos modelos, indicando la importancia de la inclusión del efecto *carry* para no cometer errores en los análisis.

## Capítulo 6

# Arreglos factoriales

### 6.1. Generalidades

Se llama arreglos factoriales a aquellos experimentos donde los tratamientos surgen de la combinación de niveles de dos o más factores. Los arreglos factoriales se ensamblan en los diseños base: completamente aleatorizado, bloques aleatorizados, cuadrado latino, entre otros.

Las principales ventajas de los arreglos factoriales son:

- ♣ Permiten evaluar el efecto de las interacciones entre factores
- ♣ Posibilitan estimar el error experimental asociado con cada replicación presente
- ♣ Permiten convalidar supuestos asociados con el modelo de clasificación experimental, de acuerdo con el diseño base
- ♣ Admiten efectos fijos o aleatorios
- ♣ Permiten generar arreglos experimentales especiales tales como: arreglos factoriales, arreglos mixtos o arreglos confundidos
- ♣ Permiten utilizar transformación de datos sobre la variable respuesta
- ♣ Permiten efectuar análisis multivariado de la varianza Manova
- ♣ Admiten componentes o factores anidados
- ♣ Permiten efectuar análisis con submuestras
- ♣ Permiten estimar la diferencia en el efecto promedio de los tratamientos derivada del arreglo factorial
- ♣ Permiten la presencia de una o más covariables
- ♣ Permiten utilizar contrastes ortogonales
- ♣ Posibilitan efectuar análisis exploratorios de tipo unidimensional, sobre cada arreglo o tratamiento

- ♣ Son aplicables en las estimaciones de superficie de respuesta
- ♣ Permiten efectuar particiones experimentales, de acuerdo con el supuesto de homogeneidad de varianzas
- ♣ Son aplicables a modelos de regresión con estructura experimental de una vía, dos vías o triple vía
- ♣ Pueden ser balanceados o desbalanceados

Si el número de dosificaciones por factor es grande, se generan demasiados tratamientos y por lo tanto es poco probable garantizar la homogeneidad del material experimental

En ciencias animales, se tienen las siguientes aplicaciones:

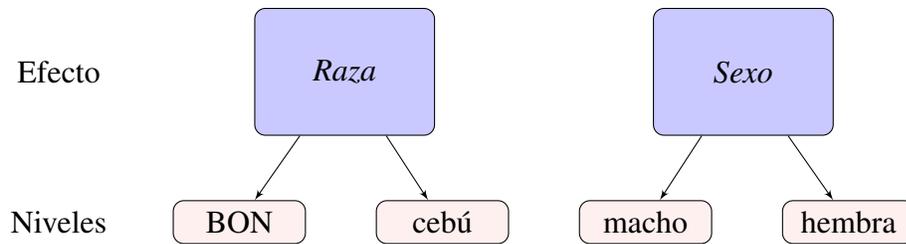
- ♣ Se utilizan para evaluar el efecto de los elementos químicos en las variables de desarrollo de cultivos utilizados en la alimentación animal.
- ♣ Se emplean para medir el efecto de premezclas o producción de alimento de origen animal en relación con las variables respuesta: sabor, textura, palatabilidad, apariencia, entre otras.
- ♣ Permiten evaluar el desarrollo morfométrico de animales que estén siendo alimentados con varios factores a distintas dosis; por ejemplo, la mezcla de materias primas.
- ♣ Se utilizan en el campo de la zootecnia para evaluar el efecto de diferentes factores en la producción animal.
- ♣ Se utilizan para evaluar el efecto de la combinación de niveles en la conservación de productos cárnicos o lácteos.

Los supuestos estadísticos de un arreglo factorial son:

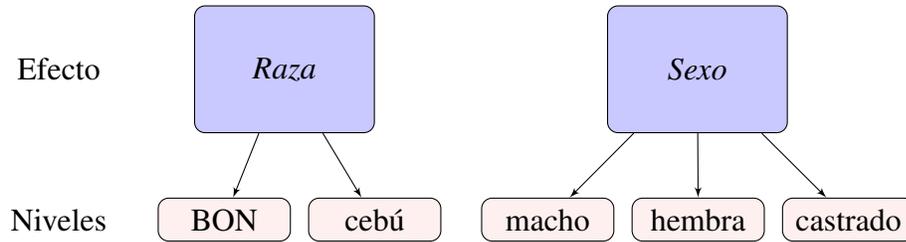
- ♣ Los errores experimentales deben tener una distribución normal.
- ♣ Los errores deben ser independientes estadísticamente y aleatoriamente distribuidos.
- ♣ Las varianzas asociadas con los distintos tratamientos deben ser homogéneas.
- ♣ No debe existir relación entre los promedios de los tratamientos y las varianzas.

## 6.2. Algunos tipos de arreglos factoriales

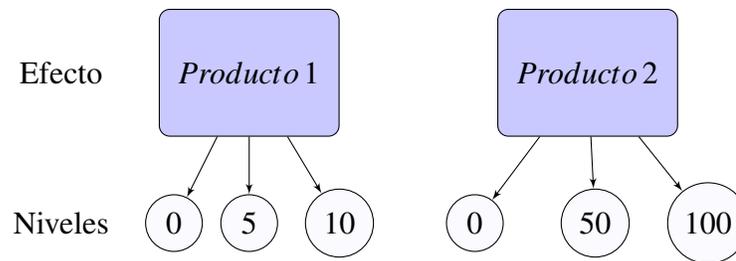
- ♣ Arreglos simétricos: Son aquellos en los cuales todos los factores tienen el mismo número de niveles. Por ejemplo, caso que se disponga de dos factores, *Raza* y *Sexo*, cada uno con dos niveles.



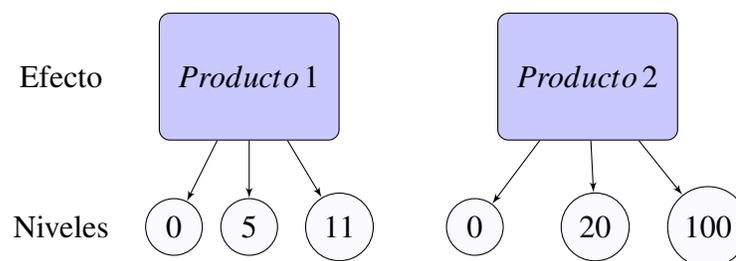
- ♣ Arreglos asimétricos: Se caracterizan porque los factores presentan diferente número de niveles.



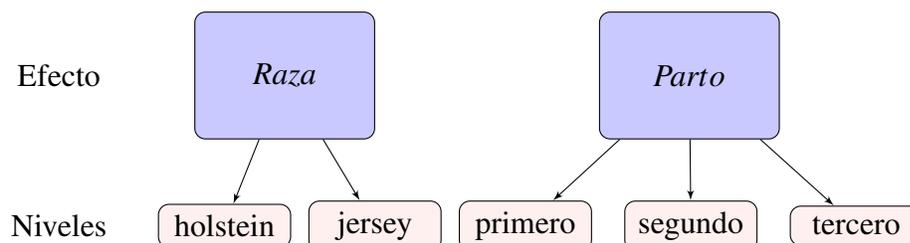
- ♣ Arreglos factoriales espaciados: Son aquellos experimentos en donde la distancia entre niveles al interior del factor es constante.



- ♣ Arreglos factoriales no espaciados: Cuando la distancia entre las dosificaciones de un factor no es igual.



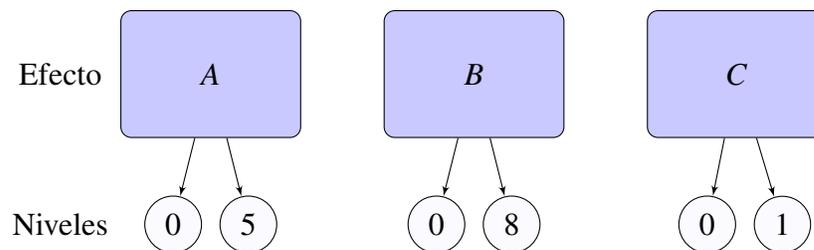
- ♣ Arreglos bifactoriales: Son aquellos en donde solo existen dos factores y por lo tanto hay una sola interacción.



La interacción estará dada por *Raza \* Parto*, originando seis tratamientos:

Tratamiento	Raza	Parto
1	holstein	primer
2	holstein	segundo
3	holstein	tercer
4	jersey	primer
5	jersey	segundo
6	jersey	tercer

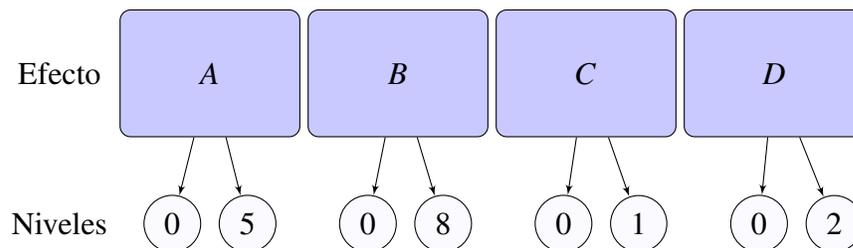
- ♣ Arreglos trifactoriales: En ellos se presentan tres factores (*A*, *B* y *C*), donde existen cuatro interacciones:  $A * B$ ,  $A * C$ ,  $B * C$  y  $A * B * C$ , como en el siguiente esquema:



En este caso, se generan ocho tratamientos ( $2^3$ ):

Tratamiento	A	B	C
1	0	0	0
2	0	0	1
3	0	8	0
4	0	8	1
5	5	0	0
6	5	0	1
7	5	8	0
8	5	8	1

- ♣ Arreglos multifactoriales: Cuando existen más de tres factores presentes.

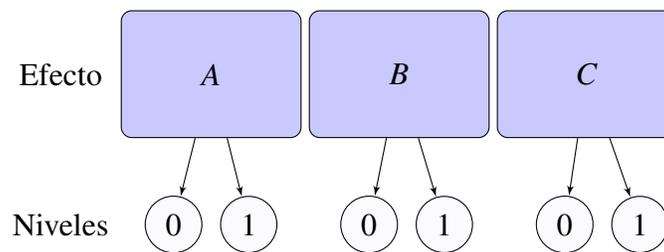


- ♣ Arreglos factoriales de efecto fijo: Cuando todos los factores presentan dosis específicas de interés.
- ♣ Arreglos factoriales aleatorios: Se definen todos los factores de naturaleza aleatoria.

- ♣ Arreglos factoriales mixtos: Son aquellos donde existen factores fijos y aleatorios.
- ♣ Arreglos factoriales balanceados: Cuando los tratamientos que se generan a partir del arreglo poseen igual número de replicaciones.
- ♣ Arreglos factoriales desbalanceados: Cuando los tratamientos que se generan a partir del arreglo no poseen igual número de replicaciones.
- ♣ Arreglos factoriales de naturaleza cualitativa: Cuando los factores son de tipo binomial o multinomial.

### 6.3. Tipos de notación

En los arreglos factoriales existen tres tipos de notación: numérica, alfabética (*tipo I*) y alfa numérica (*tipo II*). Veamos el siguiente ejemplo:



A continuación se presentan las formas de notación de los tratamientos en arreglos factoriales:

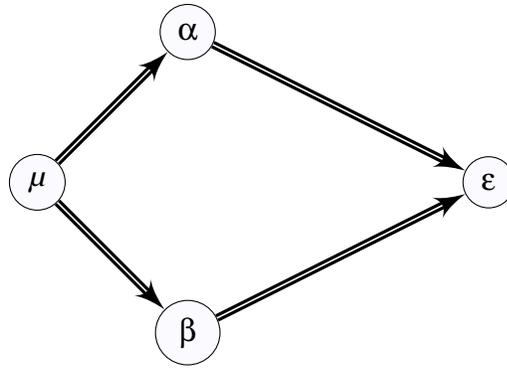
Tratamiento	Numérica	Alfa Numérica	
		Tipo I	Tipo II
T1	0 0 0	a0 b0 c0	1
T2	0 0 1	a0 b0 c1	c
T3	0 1 0	a0 b1 c0	b
T4	0 1 1	a0 b1 c1	bc
T1	1 0 0	a1 b0 c0	a
T2	1 0 1	a1 b0 c1	ac
T3	1 1 0	a1 b1 c0	ab
T4	1 1 1	a1 b1 c1	abc

Para los arreglos con tres o más factores se recomienda aplicar la notación numérica o la alfa numérica tipo I.

### 6.4. Modelos de estructura experimental

Los modelos de clasificación dependen del número de factores presentes y del diseño base elegido; a continuación se presenta una serie de modelos:

- ♣ En diseños completamente aleatorizados: Los factores se distribuyen aleatoriamente en las unidades experimentales, como se indicó en el capítulo 2. Veamos como ejemplo un diagrama de estructura de un arreglo factorial de dos factores ( $\alpha$  y  $\beta$ ):

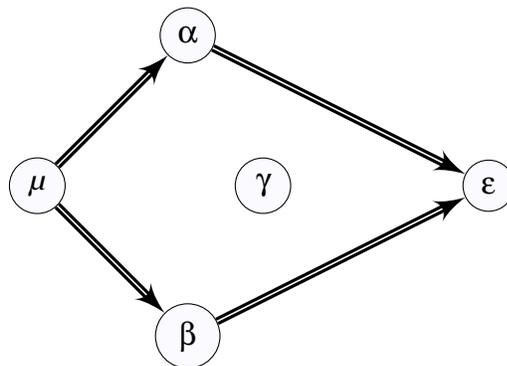


En un arreglo factorial, los factores pueden ser fijos ( $\alpha$  y  $\beta$ ) o aleatorios ( $A$  y  $B$ ). En el caso de modelo de efectos fijos, se tiene:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$$

Donde  $y_{ijk}$  es la  $k$ -ésima observación de la variable respuesta sometida al  $i$ -ésimo nivel del factor  $\alpha$  y al  $j$ -ésimo nivel del factor  $\beta$ ,  $\mu$  indica el efecto promedio del experimento,  $\alpha\beta_{ij}$  representa el efecto promedio causado por la interacción del  $i$ -ésimo nivel del factor  $\alpha$  con el  $j$ -ésimo nivel del factor  $\beta$ , y  $\epsilon_{ijk}$  es el error experimental.

- ♣ En diseños de bloques aleatorizados: Los factores se distribuyen aleatoriamente en los bloques (ver capítulo 4). El diagrama de estructura para dos factores ( $\alpha$  y  $\beta$ ) incluye el efecto de bloque ( $\gamma$ ), como se indica a continuación:



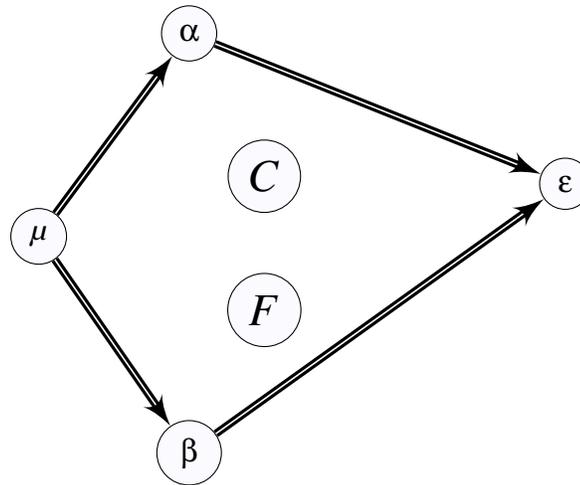
Los efectos pueden ser fijos ( $\alpha$  o  $\beta$ ) o aleatorios ( $A$  o  $B$ ). El modelo para dos factores fijos está dado por:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \epsilon_{ijk}$$

Donde:  $y_{ijk}$  representa la variable respuesta sometida al  $i$ -ésimo nivel del factor  $\alpha$  y al  $j$ -ésimo nivel del factor  $\beta$  y pertenece al  $k$ -ésimo bloque ( $\gamma$ ),  $\mu$  indica el efecto promedio del experimento,  $\alpha\beta_{ij}$  representa el efecto promedio causado por la interacción del  $i$ -ésimo nivel del factor  $\alpha$  con el  $j$ -ésimo nivel del factor  $\beta$ , y  $\epsilon_{ijk}$  es el error experimental.

- ♣ En diseños de cuadrado latino: Los factores se distribuyen en un diseño de cuadrado latino (ver capítulo 5).

El diagrama de estructura de un arreglo factorial de dos factores ( $\alpha$  y  $\beta$ ) en un diseño de cuadrado latino que incluye un doble boqueo de filas ( $F$ ) y columnas ( $C$ ) es:



Al igual que en los otros diseños, los efectos pueden ser fijos ( $\alpha$  y  $\beta$ ) o aleatorios ( $A$  y  $B$ ). En el caso de un modelo de efectos fijos de dos factores, el modelo es:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + F_k + C_l + \varepsilon_{ijkl}$$

Donde:  $y_{ijkl}$  representa la variable respuesta sometida al  $i$ -ésimo nivel del factor  $\alpha$  y al  $j$ -ésimo nivel del factor  $\beta$  y pertenece a la  $k$ -ésima fila ( $F$ ) y a la  $l$ -ésima columna ( $C$ ) en un arreglo de cuadrado latino,  $\mu$  indica el efecto promedio del experimento,  $\alpha\beta_{ij}$  representa el efecto promedio causado por la interacción del  $i$ -ésimo nivel del factor  $\alpha$  con el  $j$ -ésimo nivel del factor  $\beta$ , y  $\varepsilon_{ijkl}$  es el error experimental.

### 6.5. Hipótesis en un arreglo factorial

Las hipótesis principales a probar en un arreglo factorial de dos factores en un modelo de efectos fijos son:

Para el efecto  $\alpha$ :

$$H_{0\alpha} : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \dots = \alpha_I$$

$H_{a\alpha}$  : los promedios de los niveles del factor  $\alpha$  son diferentes

Para el efecto  $\beta$ :

$$H_{0\beta} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \dots = \beta_J$$

$H_{a\beta}$  : los promedios de los niveles del factor  $\beta$  son diferentes

Para la interacción:

$$H_{0\alpha\beta} : \alpha_1\beta_1 = \alpha_1\beta_2 = \alpha_1\beta_3 = \dots = \alpha_I\beta_J$$

$H_{a\alpha\beta}$  : los promedios de la relación de niveles de los factores  $\alpha\beta$  son diferentes

En el caso de efectos aleatorios, las hipótesis principales son:

Para el efecto  $a$ :

$$H_{0_a} : \sigma_a^2 = 0$$

$$H_{a_a} : \sigma_a^2 > 0$$

Para el efecto  $b$ :

$$H_{0_b} : \sigma_b^2 = 0$$

$$H_{a_b} : \sigma_b^2 > 0$$

Para el efecto de la interacción:

$$H_{0_{a*b}} : \sigma_{a*b}^2 = 0$$

$$H_{a_{a*b}} : \sigma_{a*b}^2 > 0$$

Las otras hipótesis a probar dependerán del diseño utilizado y de la existencia de covariables, como se indicó en capítulos anteriores.

## 6.6. Proceso de aleatorización

Este proceso se efectúa de acuerdo con el diseño base elegido: completamente aleatorizado, bloques aleatorizados, cuadrado latino, entre otros. En el siguiente ejemplo utilizaremos un diseño de bloques aleatorizados y dos factores: probiótico 1 (0g/ton y 100g/ton) y probiótico 2 (0g/ton y 300g/ton). Los bloques son los individuos provenientes de tres camadas. El modelo utilizado es:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \varepsilon_{ijk}$$

Donde  $y_{ijk}$  es el peso del lechón sometido a la  $i$ -ésima dosis del probiótico 1 y a la  $j$ -ésima dosis del probiótico 2 y pertenece a la  $k$ -ésima camada (ambiente materno),  $\mu$  indica el efecto promedio del experimento,  $\alpha_i$  indica el efecto promedio de la  $i$ -ésima dosis del probiótico 1 ( $i = 0$  y 100),  $\beta_j$  representa el efecto de la  $j$ -ésima dosis del probiótico 2 ( $j = 0$  y 300),  $\gamma_k$  es el efecto de bloqueo relacionado a la  $k$ -ésima camada ( $k = 1, 2$  y 3), y  $\varepsilon_{ijk}$  el error experimental.

Recordemos que los errores deben ser independientes e idénticamente distribuidos (iid) y que tienen una distribución normal con media 0 y varianza  $\sigma_\varepsilon^2$ , cuya notación es:  $\varepsilon_{j(i)} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$ .

Las hipótesis nulas de los efectos principales, de la interacción y de los bloques son:

$$H_{0_\alpha} : \alpha_0 = \alpha_{100}$$

$$H_{0_\beta} : \beta_0 = \beta_{300}$$

$$H_{0_{\alpha\beta}} : \alpha_0\beta_0 = \alpha_0\beta_{300} = \alpha_{100}\beta_0 = \alpha_{100}\beta_{300}$$

$$H_{0_\gamma} : \gamma_1 = \gamma_2 = \gamma_3$$

Ahora que ya están definidos los niveles de los factores y el uso de bloques (efecto de camada), procederemos a realizar la aleatorización de los tratamientos dentro de cada bloque. El comando *design.ab* de la librería *agricolae* permite aleatorizar los factores dentro de cada bloque. Este comando requiere en los argumentos: los nombres de los niveles para

cada efecto, el número de bloques o repeticiones (en este caso trabajaremos bloques), la especificación de los métodos de aleatorización y la semilla (*seed*), como se indicó en otros capítulos con otros comandos de esta librería.

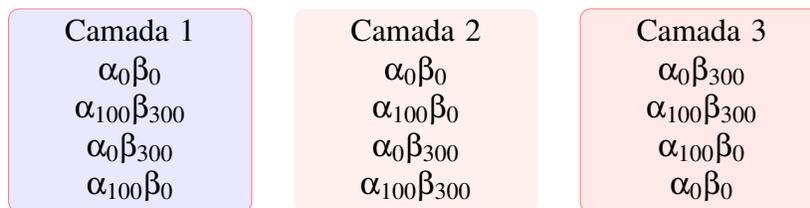
```

R Studio

library(agricolae)
Probiotico1 = c("0g/ton", "100g/ton")
Probiotico2 = c("0g/ton", "300g/ton")
Camada=3
Semi=14
Generar=design.ab (Probiotico1,Probiotico2,Camada,
                  kind="Wichmann-Hill",seed=Semi)
Generar

  plots block Probiotico1 Probiotico2
1      1      1      0g/ton      0g/ton
2      2      1     100g/ton     300g/ton
3      3      1      0g/ton     300g/ton
4      4      1     100g/ton      0g/ton
5      5      2      0g/ton      0g/ton
6      6      2     100g/ton      0g/ton
7      7      2      0g/ton     300g/ton
8      8      2     100g/ton     300g/ton
9      9      3      0g/ton     300g/ton
10     10     3     100g/ton     300g/ton
11     11     3     100g/ton      0g/ton
12     12     3      0g/ton      0g/ton
    
```

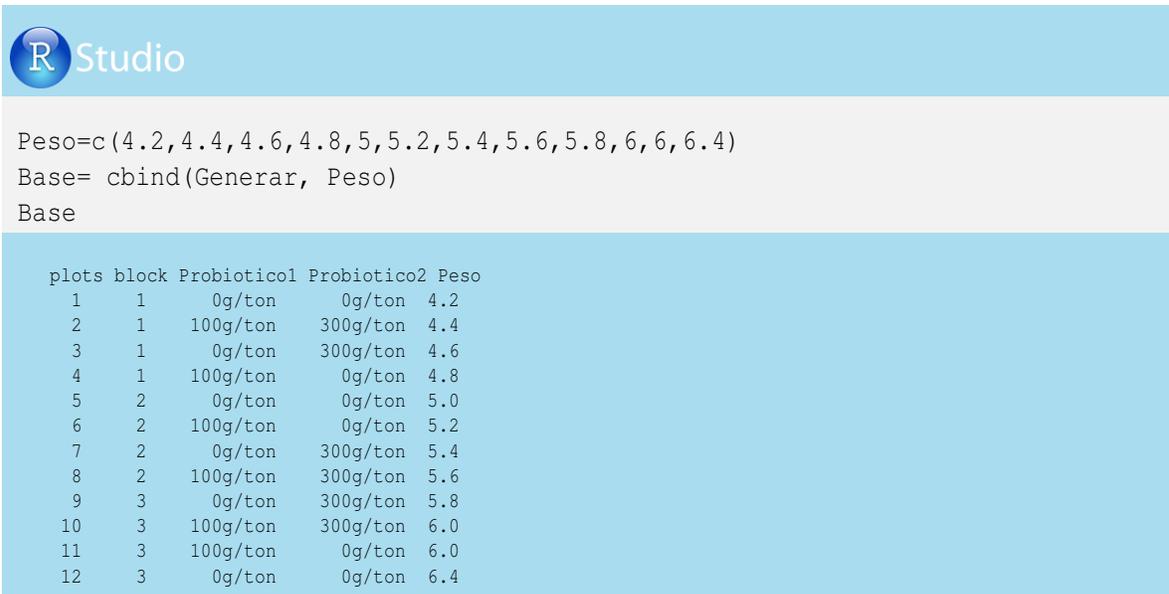
Veamos un esquema de los bloques y los tratamientos:



Los resultados encontrados fueron:

Información de lechones al destete				
Cerdo	Camada	Probiótico 1	Probiótico 2	Peso
1	1	0g/ton	0g/ton	4.2
2	1	100g/ton	300g/ton	4.4
3	1	0g/ton	300g/ton	4.6
4	1	100g/ton	0g/ton	4.8
5	2	0g/ton	0g/ton	5.0
6	2	100g/ton	0g/ton	5.2
7	2	0g/ton	300g/ton	5.4
8	2	100g/ton	300g/ton	5.6
9	3	0g/ton	300g/ton	5.8
10	3	100g/ton	300g/ton	6.0
11	3	100g/ton	0g/ton	6.0
12	3	0g/ton	0g/ton	6.4

Generemos la hoja de datos *Base* en R-project:



```

R Studio

Peso=c(4.2,4.4,4.6,4.8,5,5.2,5.4,5.6,5.8,6,6,6.4)
Base= cbind(Generar, Peso)
Base

plots block Probiotico1 Probiotico2 Peso
1      1      0g/ton      0g/ton 4.2
2      1     100g/ton    300g/ton 4.4
3      1      0g/ton    300g/ton 4.6
4      1     100g/ton      0g/ton 4.8
5      2      0g/ton      0g/ton 5.0
6      2     100g/ton      0g/ton 5.2
7      2      0g/ton    300g/ton 5.4
8      2     100g/ton    300g/ton 5.6
9      3      0g/ton    300g/ton 5.8
10     3     100g/ton    300g/ton 6.0
11     3     100g/ton      0g/ton 6.0
12     3      0g/ton      0g/ton 6.4

```

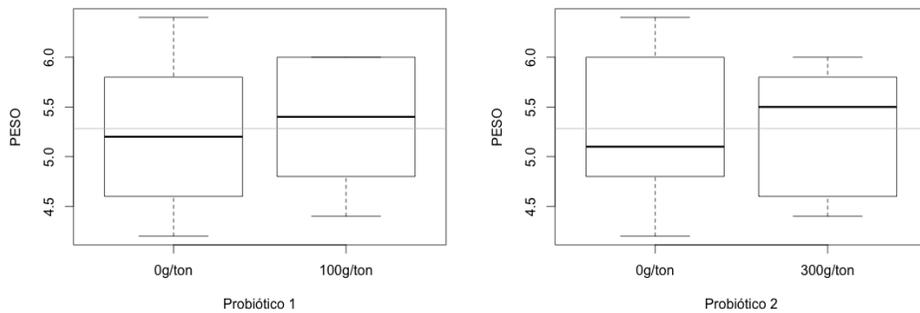
## 6.7. Análisis exploratorio de la hoja de datos

Realizemos unos gráficos de boxplot del peso con relación a los factores y a los bloques, con el objetivo de visualizar los datos. Es muy probable en el argumento (*xlab* =) al escribir la palabra Probiótico no aparezca la tilde en la vocal o. En este caso se puede utilizar la norma iso 8859-1 que define la codificación del alfabeto latino, donde la vocal o minúscula con acento agudo equivale a *f3*, quedando en R-project *Probi\uf3tico*.



```

boxplot (Base$Peso~Base$Probiotico1,ylab="PESO",
          xlab="Probi\uf3tico 1",data=Base)
abline (h=mean (Base$Peso) , col="gray")
boxplot (Base$Peso~Base$Probiotico2,ylab="PESO",
          xlab="Probi\uf3tico 2",data=Base)
abline (h=mean (Base$Peso) , col="gray")
    
```

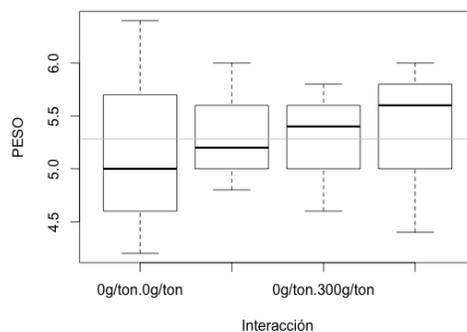


A continuación vemos un gráfico de cajas de la interacción utilizando el comando *boxplot* para tener una idea de la media y la variabilidad existente en la combinación de niveles de los dos factores.



```

boxplot (Base$Peso~Base$Probiotico1*Probiotico2,ylab="PESO",
          xlab="Interacci\uf3n", data=Base)
abline (h=mean (Base$Peso) , col="gray")
    
```

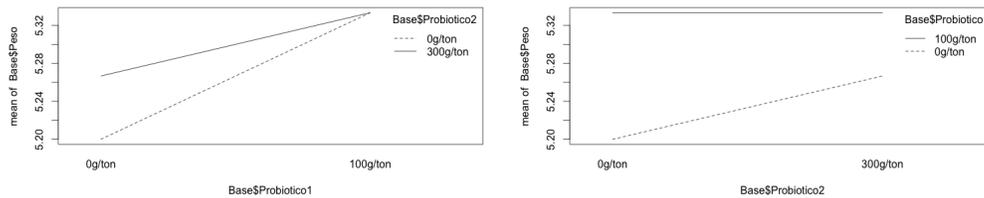


En el gráfico de las interacciones podemos observar que la combinación  $\alpha_0\beta_0$  presenta menor media y mayor variabilidad que los otros tratamientos. Veamos un gráfico que muestra la

relación de niveles de dos factores para la variable *peso* con el comando *interaction.plot* con los siguientes argumentos en su orden: nombre de la variable que contiene el primer factor, nombre de la variable que contiene el segundo factor y nombre de la variable respuesta; veamos:



```
interaction.plot(Base$Probiotico1, Base$Probiotico2, Base$Peso)
interaction.plot(Base$Probiotico2, Base$Probiotico1, Base$Peso)
```

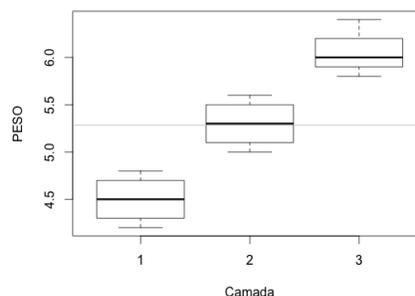


Los anteriores gráficos nos dan una idea de una posible interacción entre los dos factores analizados, porque se visualiza una diferencia de respuesta de un factor, cuando se aplican los niveles del otro factor.

Ahora realizaremos un gráfico *boxplot* del efecto de bloques (camadas):



```
boxplot(Base$Peso~Base$block, ylab="PESO", xlab="Camada", data=Base)
abline(h=mean(Base$Peso), col="gray")
```



Notamos que existe una diferencia de pesos causada por el ambiente materno de camada.

Para confirmar nuestra sospecha de significancia de la interacción y de los bloques, realizaremos el análisis de varianza para aceptar o rechazar las hipótesis nulas de los factores, de la interacción y de los bloques.

## 6.8. Análisis de varianza

Ahora procederemos a realizar el análisis de varianza, obteniendo las sumas de cuadrados del total, de los factores, de la interacción, de los bloques y del error:

La suma de cuadrados totales es ( $SCT$ ):

$$SCT = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (y_{ijk})^2 - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

La suma de cuadrados del efecto  $\alpha$  ( $SC\alpha$ ):

$$SC\alpha = \frac{\sum_{i=1}^I \alpha_{i..}^2}{JK} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

La suma de cuadrados del efecto  $\beta$  ( $SC\beta$ ):

$$SC\beta = \frac{\sum_{j=1}^J \beta_{.j.}^2}{IK} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

La suma de cuadrados de la interacción  $\alpha\beta$  ( $SC\alpha\beta$ ):

$$SC\alpha\beta = \frac{\sum_{i=1}^I \sum_{j=1}^J (y_{ij.})^2}{K} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n} - SC\alpha - SC\beta$$

La suma de cuadrados de bloques ( $SC\gamma$ ):

$$SC\gamma = \frac{\sum_{k=1}^K \gamma_{..k}^2}{IJ} - \frac{\left( \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K y_{ijk} \right)^2}{n}$$

La suma de cuadrados del error ( $SC\epsilon$ ):

$$SC\epsilon = SCT - SC\alpha - SC\beta - SC\alpha\beta - SC\gamma$$

Para el ejemplo:

Suma de cuadrados totales:

$$SCT = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^3 (y_{ijk})^2 - \frac{\left( \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^3 y_{ijk} \right)^2}{12}$$

$$SCT = (4.2)^2 + (4.4)^2 + \dots + (6.4)^2 - \frac{(4.2 + 4.4 + \dots + 6.4)^2}{12}$$

$$SCT = 340.36 - \frac{(63.4)^2}{12} = 5.396667$$

Suma de cuadrados del efecto  $\alpha$ :

$$SC\alpha = \frac{31.4^2 + 32^2}{2 \times 3} - \frac{(4.2 + 4.4 + \dots + 6.4)^2}{12} = 334.99333 - 334.9633 = 0.03$$

Suma de cuadrados del efecto  $\beta$

$$SC\beta = \frac{31.6^2 + 31.8^2}{2 \times 3} - \frac{(4.2 + 4.4 + \dots + 6.4)^2}{12} = 334.967 - 334.96 = 0.003$$

Suma de cuadrados de la interacción:

$$SC\alpha\beta = \frac{15.6^2 + 15.8^2 + 16^2 + 16^2}{3} - \frac{(4.2 + 4.4 + \dots + 6.4)^2}{12} - 0.03 - 0.00333$$

$$SC\alpha\beta = 335 - 334.9633 - 0.3 - 0.00333 = 0.00333$$

Suma de cuadrados de bloques (camadas):

$$SC\gamma = \frac{18^2 + 21.2^2 + 24.2^2}{2 \times 2} - \frac{(4.2 + 4.4 + \dots + 6.4)^2}{12} = 339.77 - 334.963 = 4.8067$$

Suma de cuadrados del error ( $SC\epsilon$ ):

$$SC\epsilon = 5.396667 - 0.03 - 0.00333 - 0.0033 - 4.80667 = 0.553$$

El cuadro del análisis de varianza está conformado por:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$
$\alpha$	$I - 1$	$SC\alpha$	$\frac{SC\alpha}{I-1}$	$\frac{CM\alpha}{CM\epsilon}$
$\beta$	$J - 1$	$SC\beta$	$\frac{SC\beta}{J-1}$	$\frac{CM\beta}{CM\epsilon}$
$\gamma$	$K - 1$	$SC\gamma$	$\frac{SC\gamma}{K-1}$	$\frac{CM\gamma}{CM\epsilon}$
$\alpha\beta$	$(I - 1)(J - 1)$	$SC\alpha\beta$	$\frac{SC\alpha\beta}{(I-1)(J-1)}$	$\frac{CM\alpha\beta}{CM\epsilon}$
$\epsilon$	$(n-1)-(I-1)-(J-1)-$ $(I-1)(J-1)-(K-1)$	$SCE$	$\frac{SCE}{g \cdot I_\epsilon}$	
Total	$n - 1$	$SCT$		

Para el ejemplo:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	$F_c$
$\alpha$	$2 - 1 = 1$	0.03	0.03	$\frac{0.03}{0.0922} = 0.325$
$\beta$	$2 - 1 = 1$	0.003	0.003	$\frac{0.003}{0.0922} = 0.036$
$\gamma$	$3 - 1 = 2$	4.80666	2.4033	$\frac{2.4033}{0.0922} = 26.60$
$\alpha\beta$	$1 \times 1 = 1$	0.0033	0.00333	$\frac{0.00333}{0.0922} = 0.036$
$\epsilon$	$(11)-(1)-(1)-(1 \times 1)-(3-1)=6$	0.553	0.0922	
Total	$12 - 1$	5.39667		

Para el desarrollo del análisis de varianza utilizaremos el comando *aov*, cuyos resultados saldrán en la hoja de datos *Analisis*. Miremos los resultados con el comando *summary*:

```

R Studio

Analisis=aov(Base$Peso~Base$Probiotico1*Base$Probiotico2+Base$block,data=Base)
summary (Analisis)

          Df Sum Sq Mean Sq F value Pr(>F)
Base$Probiotico1  1  0.030  0.0300   0.325 0.5891
Base$Probiotico2  1  0.003  0.0033   0.036 0.8555
Base$block        2  4.807  2.4033  26.060 0.0011 **
Base$Probiotico1:Base$Probiotico2  1  0.003  0.0033   0.036 0.8555
Residuals        6  0.553  0.0922

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Observamos que los probióticos y la interacción fueron no significativos ( $p > 0.05$ ). El efecto del bloque (camada) fue altamente significativo ( $p < 0.01$ ), dada la influencia de la habilidad de las madres para cuidar a sus crías.

Obtengamos las soluciones (comando *model.tables*) y las medias (*model.tables* y argumento "mean") de los niveles de los efectos incluidos en el modelo:



```

model.tables(Análisis)

Tables of effects
Base$Probiotico1
  0g/ton 100g/ton
   -0.05   0.05
Base$Probiotico2
  0g/ton 300g/ton
-0.016667 0.016667
Base$block
   1     2     3
-0.7833 0.0167 0.7667
Base$Probiotico1:Base$Probiotico2
      Base$Probiotico2
Base$Probiotico1 0g/ton 300g/ton
  0g/ton -0.016667 0.016667
 100g/ton 0.016667 -0.016667

```

También podemos generar las medias de los niveles de cada efecto incluido en el modelo:



```

model.tables(Análisis, "mean")

Tables of means
Grand mean
      5.283333
Base$Probiotico1
  0g/ton 100g/ton
   5.233   5.333
Base$Probiotico2
  0g/ton 300g/ton
   5.267   5.300
Base$block
   1     2     3
4.50 5.30 6.05
Base$Probiotico1:Base$Probiotico2
      Base$Probiotico2
Base$Probiotico1 0g/ton 300g/ton
  0g/ton 5.200 5.267
 100g/ton 5.333 5.333

```

Con la información anterior, se obtiene:

$$\hat{\mu} = 5.283$$

En el caso de los efectos:

$$\hat{\alpha}_0 = 5.283 - 0.05 = 5.233$$

$$\hat{\alpha}_{100} = 5.283 + 0.05 = 5.333$$

$$\hat{\beta}_0 = 5.283 - 0.016667 = 5.267$$

$$\hat{\beta}_{300} = 5.283 + 0.016667 = 5.300$$

En el caso de los bloques:

$$\hat{\gamma}_1 = 5.283 - 0,7833 = 4.50$$

$$\hat{\gamma}_2 = 5.283 + 0,0167 = 5.30$$

$$\hat{\gamma}_3 = 5.283 + 0,7667 = 6.05$$

En el caso de la interacción:

$$\hat{\alpha}_0\hat{\beta}_0 = 5.283 - 0.05 - 0.016667 - 0.016667 = 5.200$$

$$\hat{\alpha}_0\hat{\beta}_{300} = 5.283 - 0.05 + 0.016667 - 0.016667 = 5.333$$

$$\hat{\alpha}_{100}\hat{\beta}_0 = 5.283 + 0.05 - 0.016667 + 0.016667 = 5.267$$

$$\hat{\alpha}_{100}\hat{\beta}_{300} = 5.283 + 0.05 + 0.016667 - 0.016667 = 5.333$$

## 6.9. Prueba de normalidad

Los errores experimentales deben tener distribución normal, con media cero ( $\mu_\varepsilon = 0$ ) y varianza  $\sigma_\varepsilon^2$ , de modo que  $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$ .

Para verificar lo anterior, se procede a estimar los residuos:

$$\text{Si: } y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \varepsilon_{ijk}$$

Entonces:

$$\varepsilon_{ijk} = y_{ijk} - \mu - \alpha_i - \beta_j - \alpha\beta_{ij} - \gamma_k$$

Al reemplazar el efecto promedio de cada efecto ( $\hat{\alpha}$  y  $\hat{\beta}$ ), el efecto de bloques ( $\hat{\gamma}$ ), la interacción ( $\hat{\alpha\beta}$ ) y el efecto promedio general ( $\hat{\mu}$ ), se tiene:

$$\varepsilon_{ijk} = y_{ijk} - \hat{\mu} - (\hat{\alpha}_i - \hat{\mu}) - (\hat{\beta}_j - \hat{\mu}) - (\hat{\alpha\beta}_{ij} - \hat{\mu}) - (\hat{\gamma}_k - \hat{\mu})$$

$$\varepsilon_{ijk} = y_{ijk} - (\hat{\alpha}_i) - (\hat{\beta}_j) - (\hat{\alpha\beta}_{ij}) - \hat{\gamma}_k - \hat{\mu}$$

En nuestro ejercicio, tendríamos que el error asociado al primer cerdo de la lista de resultados del experimento que no estuvo sometido a las dosis de los dos probióticos ( $\alpha_0\beta_0$ ) y que permaneció en la jaula 1 ( $\gamma_1$ ) es:

$$\varepsilon_{\alpha_0\beta_0\gamma_1} = 4.2 - (-0.05) - (-0.0167) - (-0.0167) - (-0.783) - 5.283 = -0.2163$$

El error asociado al último cerdo de la lista de resultados del experimento que no estuvo sometido a las dosis de los dos probióticos ( $\alpha_0\beta_0$ ) y que permaneció en la jaula 3 ( $\gamma_3$ ) es:

$$\varepsilon_{\alpha_0\beta_0\gamma_1} = 6.4 - (-0.05) - (-0.0167) - (-0.0167) - (-0.7667) - 5.283 = 0.4333$$

Para obtener los residuos y los valores predichos de cada unidad experimental en R-project, utilizamos los comandos *residuals* y *fitted*; veamos:

```

R Studio

Base$Residuos=residuals(Analisis)
Base$Predichos=fitted(Analisis)
Base

plots block Probiotico1 Probiotico2 Peso Residuos Predichos
1 1 0g/ton 0g/ton 4.2 -0.2166667 4.416667
2 1 100g/ton 300g/ton 4.4 -0.1500000 4.550000
3 1 0g/ton 300g/ton 4.6 0.1166667 4.483333
.
12 3 0g/ton 0g/ton 6.4 0.4333333 5.966667

```

El resumen general (comando *summary*), la media (comando *mean*) y la varianza (comando *var*) de los residuos son:

```

R Studio

summary(Base$Residuos)
var(Base$Residuos)
mean(Base$Residuos)

> summary(Base$Residuos)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.2333 -0.1667 -0.1000  0.0000  0.1500  0.4333
> var(Base$Residuos)
[1] 0.05030303
> mean(Base$Residuos)
[1] -2.197091e-17

```

La media de los errores fue 0 y la varianza fue 0.5030 ( $\epsilon_{ijk}$  tiene  $\mu_\epsilon = 0$ ,  $\sigma_\epsilon^2 = 0.5030$ ).

Apliquemos la prueba de Shapiro-Wilk con el comando *shapiro.test*, para verificar si los residuos siguen una distribución normal:

```

R Studio

ShapiroResiduos=shapiro.test(Base$Residuos)
ShapiroResiduos

Shapiro-Wilk normality test
w = 0.877, p-value = 0.08025

```

Por presentar un valor de  $p > 0.05$ , se concluye que los errores se distribuyeron normalmente, y por tanto:  $\epsilon_{ijk} \sim N(0, 0.5030)$ .

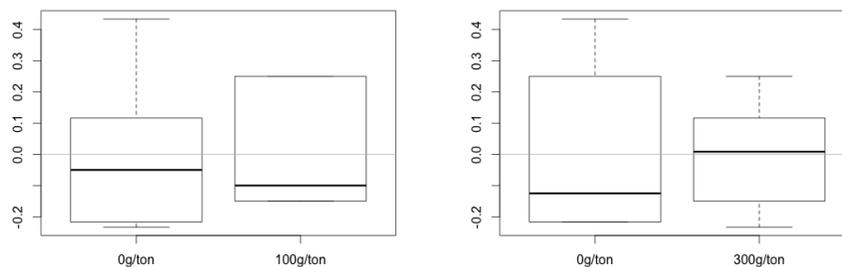
## 6.10. Pruebas de homogeneidad de varianza de los errores

Para tener una idea de los residuos por nivel de efecto, utilicemos el comando *boxplot* y realizaremos la prueba Cochran para verificar la homogeneidad de los residuos, como se indicó en los capítulos anteriores. En el caso de los dos probióticos se tendría:

```

R Studio

boxplot(Base$Residuo~Base$Probiotico1, data=Base)
abline(h=mean(Base$Residuo), col="gray")
boxplot(Base$Residuo~Base$Probiotico2, data=Base)
abline(h=mean(Base$Residuo), col="gray")
    
```



Se observa mayor dispersión de los errores cuando no se aplicó el probiótico 1 y el probiótico 2. Apliquemos la prueba de Cochran para los residuos agrupados por los niveles del probiótico 1:

```

R Studio

library(outliers)
CochranA=cochran.test(Residuos~Probiotico1,data=Base)
CochranA

Cochran test for outlying variance
data: Residuos ~ Probiotico1
C = 0.6566, df = 6, k = 2, p-value = 0.4939
alternative hypothesis: Group 0g/ton has outlying variance
sample estimates:
 0g/ton  100g/ton
0.07266667 0.03800000
    
```

Por presentar un  $p > 0.05$  (no significativo), no rechazamos la hipótesis  $H_0$  de homogeneidad de varianzas: Los errores presentaron varianzas homogéneas en los niveles del probiótico 1. Ahora realicemos la prueba de Cochran para los niveles del probiótico 2:



```
CochranB=cochran.test(Residuos~Probiotico2,data=Base)
CochranB
```

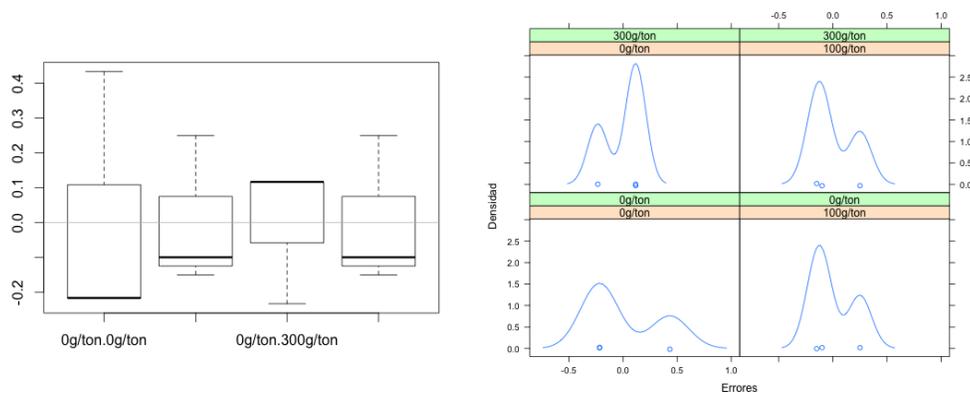
```
data: Residuos ~ Probiotico2
C = 0.6807, df = 6, k = 2, p-value = 0.4257
alternative hypothesis: Group 0g/ton has outlying variance
sample estimates:
  0g/ton  300g/ton
0.07533333 0.03533333
```

La prueba de Cochran indicó que los errores presentaron varianzas homogéneas en los niveles del probiótico 2.

En el caso de la interacción, el gráfico de dispersión de la interacción con el comando *boxplot* y *densityplot* son:



```
boxplot(Base$Residuo~Base$Probiotico1*Base$Probiotico2,data=Base)
abline(h=mean(Base$Residuo),col="gray")
library(lattice)
densityplot(~Base$Residuos|Base$Probiotico1*
  Base$Probiotico2,ylab="Densidad",xlab="Errores")
```



Según los gráficos anteriores, el tratamiento que no tuvo aplicación de los dos probióticos presentó mayor variabilidad. Realicemos la prueba de Cochran para verificar si existe homogeneidad de varianzas. Para esto, es necesario crear otra variable que aglutine las variables *Probiotico1* y *Probiotico2*. Utilizaremos el comando *paste* con los argumentos: nombres de las variables y *sep = "\*" para que la separación de las variables sea un asterisco:*

```

R Studio

Base$Interaccion=paste(Base$Probiotico1,Base$Probiotico2, sep="*")
attach (Base)
CochranResiduos=cochran.test (Residuos~(Interaccion) , data=Base)
CochranResiduos

data: Residuos ~ (Interaccion)
C = 0.509, df = 3, k = 4, p-value = 0.4734
alternative hypothesis: Group 0g/ton*0g/ton has outlying variance
sample estimates:
  0g/ton*0g/ton  0g/ton*300g/ton  100g/ton*0g/ton 100g/ton*300g/ton
  0.14083333    0.04083333    0.04750000    0.04750000
    
```

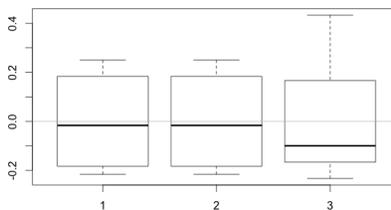
La prueba de Cochran mostró que la variabilidad de los residuos de los niveles de la interacción es la misma (homogeneidad de varianzas).

Ahora veamos la dispersión de los residuos por bloque con el comando *boxplot*:

```

R Studio

boxplot(Base$Residuo~Base$block, data=Base)
abline(h=mean(Base$Residuo) , col="gray")
    
```



Según el gráfico anterior, el bloque 3 presentó mayor variabilidad; realicemos la prueba de Cochran para verificar si existe homogeneidad de varianzas:



```
CochranResiduos=cochran.test(Residuos~block,data=Base)
CochranResiduos
```

```
C = 0.4739, df = 4, k = 3, p-value = 0.7411
alternative hypothesis: Group 3 has outlying variance
sample estimates:
      1      2      3
0.04851852 0.04851852 0.08740741
```

La prueba de Cochran mostró que los residuos son homogéneos por bloques.

### 6.11. Interacciones significativas

Cuando en el análisis de varianza alguna de las interacciones es significativa, se entiende que existe un factor que está afectado por otro. En estos casos, se realizan pruebas de comparación de medias, fijando un factor y observando las diferencias entre niveles del otro factor.

En el caso anterior (peso al destete) no encontramos diferencias significativas en la interacción, por eso trabajaremos otra variable (peso a los tres meses), donde el efecto de la interacción fue significativo. Veamos la hoja de datos:

Información de peso de cerdos (kg) a los 3 meses de edad				
Cerdo	Camada	Probiótico 1	Probiótico 2	Peso
1	1	0g/ton	0g/ton	59.2
2	1	100g/ton	300g/ton	43.2
3	1	0g/ton	300g/ton	41.5
4	1	100g/ton	0g/ton	48.0
5	2	0g/ton	0g/ton	59.6
6	2	100g/ton	0g/ton	42.5
7	2	0g/ton	300g/ton	50.1
8	2	100g/ton	300g/ton	50.7
9	3	0g/ton	300g/ton	45.0
10	3	100g/ton	300g/ton	41.5
11	3	100g/ton	0g/ton	44.4
12	3	0g/ton	0g/ton	59.0

Creamos la hoja de datos *Base* en R-project:

```

R Studio

Peso3= c(59.2,43.2,41.5,48.0,59.6,42.5,50.1,50.7,45.0,41.5,44.4,59.0)
Base= cbind(Generar, Peso3)
attach (Base)
Base

plots block Probiotico1 Probiotico2 Peso3
1 1 0g/ton 0g/ton 59.2
2 1 100g/ton 300g/ton 43.2
3 1 0g/ton 300g/ton 41.5
4 1 100g/ton 0g/ton 48.0
5 2 0g/ton 0g/ton 59.6
6 2 100g/ton 0g/ton 42.5
7 2 0g/ton 300g/ton 50.1
8 2 100g/ton 300g/ton 50.7
9 3 0g/ton 300g/ton 45.0
10 3 100g/ton 300g/ton 41.5
11 3 100g/ton 0g/ton 44.4
12 3 0g/ton 0g/ton 59.0
    
```

El análisis de varianza es:

```

R Studio

Analisis=aov(Base$Peso3~Probiotico1*Probiotico2+block,data=Base)
anova (Analisis)

Response: Peso3
          Df Sum Sq Mean Sq F value Pr(>F)
Probiotico1  1 162.068  162.068  12.6902  0.01189 *
Probiotico2  1 138.041  138.041  10.8088  0.01666 *
block        2  24.500   12.250   0.9592  0.43505
Probiotico1:Probiotico2  1 144.908  144.908  11.3465  0.01507 *
Residuals    6  76.627   12.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Observamos que el efecto del bloque (camada) fue no significativo ( $p > 0.05$ ), indicando que el efecto materno y ambiente común de la camada se pierde después del destete.

A pesar de que los efectos (los dos probióticos) fueron significativos ( $p < 0.05$ ), la discusión se debe centrar en el efecto de la interacción, porque también fue significativa.

Utilicemos el comando *lsmeans* y realicemos una prueba de comparación de medias pareadas (argumento *pairwise*) mediante la prueba de Tukey para el efecto de la interacción:



```
lsmeans (Analysis, pairwise~Probiotico1*Probiotico2, adjust="tukey")
```

```
$`Probiotico1:Probiotico2 lsmeans`
  Probiotico1 Probiotico2  lsmean      SE df lower.CL upper.CL
    0g/ton      0g/ton  59.26667  2.063259  6  54.21805  64.31528
   100g/ton      0g/ton  44.96667  2.063259  6  39.91805  50.01528
    0g/ton      300g/ton  45.53333  2.063259  6  40.48472  50.58195
   100g/ton      300g/ton  45.13333  2.063259  6  40.08472  50.18195

$`Probiotico1:Probiotico2 pairwise differences`
              estimate      SE df  t.ratio p.value
0g/ton, 0g/ton - 100g/ton, 0g/ton    14.3000000  2.917889  6  4.90080  0.01073
0g/ton, 0g/ton - 0g/ton, 300g/ton    13.7333333  2.917889  6  4.70660  0.01302
0g/ton, 0g/ton - 100g/ton, 300g/ton   14.1333333  2.917889  6  4.84368  0.01136
100g/ton, 0g/ton - 0g/ton, 300g/ton   -0.5666667  2.917889  6 -0.19420  0.99712
100g/ton, 0g/ton - 100g/ton, 300g/ton -0.1666667  2.917889  6 -0.05712  0.99993
0g/ton, 300g/ton - 100g/ton, 300g/ton  0.4000000  2.917889  6  0.13709  0.99898
  p values are adjusted using the tukey method for 4 means
```

Ordenando las medias de los niveles de interacción de mayor a menor, tendríamos:

$$\hat{\alpha}_0\hat{\beta}_0 = 59.27$$

$$\hat{\alpha}_0\hat{\beta}_{300} = 45.53$$

$$\hat{\alpha}_{100}\hat{\beta}_{300} = 45.13$$

$$\hat{\alpha}_{100}\hat{\beta}_0 = 44.97$$

Según la prueba de comparación de medias del análisis anterior,  $\hat{\alpha}_0\hat{\beta}_0$  fue diferente a los otros tres niveles (le asignaremos la letra *a*). Los otros tres niveles no fueron diferentes, y por tanto; le asignaremos la letra *b*, de modo que queda:

$$\hat{\alpha}_0\hat{\beta}_0 = 59.27, a$$

$$\hat{\alpha}_0\hat{\beta}_{300} = 45.53, b$$

$$\hat{\alpha}_{100}\hat{\beta}_{300} = 45.13, b$$

$$\hat{\alpha}_{100}\hat{\beta}_0 = 44.97, b$$

## 6.12. Arreglos factoriales con niveles vacíos

En muchas ocasiones, en experimentación animal no se pueden tener todas las combinaciones posibles de niveles de los factores para ser considerado como un diseño factorial y probar las interacciones entre factores.

Veamos el siguiente ejemplo de tres líneas de gallinas (leghorn, white rock y plymouth) y tres sistemas de producción de huevo, con las siguientes especificaciones de hábitat: aves en piso todo el tiempo, aves en jaula todo el tiempo y un sistema mixto (piso y jaula).

Lastimosamente, por dificultades en la consecución de los galpones, no se pueden tener todas las líneas en todos los sistemas. Por eso se decidió diseñar un experimento donde no se tienen

galpones experimentales en jaula con la línea plymouth y en piso con white rock. Veamos el esquema de número de galpones propuesto por el investigador para su experimento:

Número de galpones por sistema

Línea comercial	jaula	piso	mixto
leghorn	6	3	3
plymouth		4	3
white rock	4		3

En estos casos, las medias de los niveles vacíos no son estimables y se deben aplicar restricciones adicionales. Es el caso de considerar como fuente de variación las combinaciones de los efectos y realizar contrastes adecuados a esas combinaciones posibles de efectos y al interés del investigador. Para un mejor entendimiento, realizaremos a continuación una descripción de asignación de las combinaciones posibles y contrastes con base en el ejemplo del experimento de las gallinas en diferentes sistemas de hábitat. Veamos:

Nombre	Línea y sistema	Número de galpones
Lj	leghorn y jaula	6
Lp	leghorn y piso	3
Lq	leghorn y mixto	3
Pp	plymouth y piso	4
Pq	plymouth y mixto	3
Wj	white rock y jaula	4
Wq	white rock y mixto	3

En un modelo factorial de dos factores con todos los niveles se tiene:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

El modelo propuesto para evaluar el experimento sería:

$$y_{ij} = \mu + \gamma_i + \varepsilon_{ij}$$

Donde  $\gamma_i$  sería el efecto de la  $i$ -ésima combinación de línea genética y sistema de producción en el peso del huevo.

Construyamos las hojas de datos necesarias con la información de peso del huevo/ave en gramos de las líneas genéticas leghorn (L), plymouth (P) y white (W), sometidas a tres sistemas de producción: piso (p), jaula (j) y mixto (q):

Información de peso de los huevos de las aves				
Galpón	Línea	Sistema	Combinación	Peso (gramos)
1	L	p	Lp	58
2	L	p	Lp	61
3	L	p	Lp	59
4	L	j	Lj	65
5	L	j	Lj	64
6	L	j	Lj	61
7	L	j	Lj	62
8	L	j	Lj	61
9	L	j	Lj	63
10	L	q	Lq	64
11	L	q	Lq	67
12	L	q	Lq	67
13	P	p	Pp	57
14	P	p	Pp	55
15	P	p	Pp	63
16	P	p	Pp	61
17	P	q	Pq	68
18	P	q	Pq	67
19	P	q	Pq	62
20	W	j	Wj	60
21	W	j	Wj	64
22	W	j	Wj	66
23	W	j	Wj	65
24	W	q	Wq	65
25	W	q	Wq	66
26	W	q	Wq	66

Generemos la hoja de datos con el nombre *hoja* en R-project:

```

R Studio

Linea=c("L","L","L","L","L","L","L","L","L","L","L","L",
        "P","P","P","P","P","P","P",
        "W","W","W","W","W","W","W")
siste=c("p","p","p","j","j","j","j","j","j","q","q","q",
        "p","p","p","p","q","q","q",
        "j","j","j","j","q","q","q")
peso =c(58,61,59,65,64,61,62,61,63,64,67,67,
        57,55,63,61,68,67,62,
        60,64,66,65,65,66,66)
hoja=data.frame(Linea,siste,peso)
table(hoja$Linea, hoja$siste)

  j p q
L 6 3 3
P 0 4 3
W 4 0 3

```

Utilicemos el comando *as.factor* con el argumento *paste* para concatenar las variables: línea genética y sistema de producción y crear la variable *gc*:

```

R Studio

hoja$gc=as.factor (paste(hoja$Linea, hoja$siste,sep=" " ))
levels(hoja$gc)
table(hoja$gc)

> levels(hoja$gc)
[1] "Lj" "Lp" "Lq" "Pp" "Pq" "Wj" "Wq"
> table(hoja$gc)
  Lj Lp Lq Pp Pq Wj Wq
  6  3  3  4  3  4  3

```

Ahora realicemos el análisis de varianza de acuerdo con el modelo considerado en este caso con la fuente de variación de grupo que contempla las líneas genéticas y los sistemas de producción incluidos en el experimento:



```

 analisis=aov(peso~gc, data=hoja)
 summary(analisis)

```

```

      Df Sum Sq Mean Sq F value Pr(>F)
gc      6  176.9   29.480    5.28 0.00237 **
Residuals 19  106.1    5.583
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

El efecto combinado de líneas genéticas y sistemas de producción fue altamente significativo ( $p < 0.05$ ). Rechazamos la hipótesis nula y concluimos que existen diferencias de medias entre los grupos analizados. Por consiguiente, procederemos a realizar los contrastes respectivos:

Los contrastes estarían dados por la comparación de niveles y factores completos, o por necesidades del investigador. Adicionalmente, el número de contrastes debe ser menor que el número de combinaciones de factores (en nuestro caso, máximo seis contrastes). Los contrastes pueden ser (entre otros):

- ♣ Los contrastes de la comparación de leghorn y plymouth en los sistemas de producción de piso y mixto son:

	Lj	Lp	Lq	Pp	Pq	Wj	Wq
leghorn vs plymouth	0	0.5	0.5	-0.5	-0.5	0	0
piso vs mixto	0	-0.5	0.5	-0.5	0.5	0	0
Interacción	0	-0.5	0.5	0.5	-0.5	0	0

- ♣ Los contrastes de la comparación de leghorn y white rock en jaula y mixto son:

	Lj	Lp	Lq	Pp	Pq	Wj	Wq
leghorn vs white	0.5	0	0.5	0	0	-0.5	-0.5
jaula vs mixto	1	0	-0.5	0	0	0.5	-0.5
Interacción	0.5	0	-0.5	0	0	-0.5	0.5

- ♣ El contraste de la comparación de leghorn en los tres sistemas es:

	Lj	Lp	Lq	Pp	Pq	Wj	Wq
jaula vs piso y mixto	0.5	-0.25	-0.25	0	0	0	0

- ♣ El contraste de la comparación de plymouth en piso y mixto es:

	Lj	Lp	Lq	Pp	Pq	Wj	Wq
piso vs mixto	0	0	0	-0.5	0.5	0	0

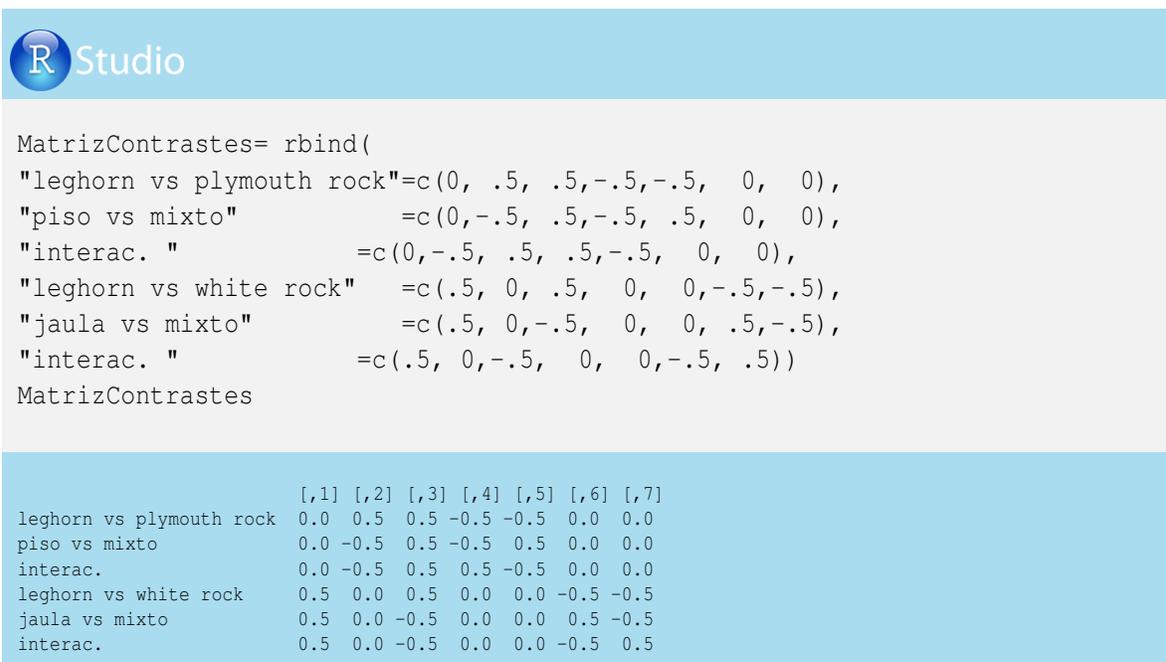
♣ El contraste de la comparación de leghorn y white rock en jaula y mixto es:

	Lj	Lp	Lq	Pp	Pq	Wj	Wq
jaula vs mixto	0	0	0	0	0.5	0	-0.5

El investigador decidió escoger los seis primeros contrastes, donde evaluaremos las dos interacciones posibles del arreglo factorial propuesto:

Contraste	Lj	Lp	Lq	Pp	Pq	Wj	Wq
leghorn vs plymouth	0	0.5	0.5	-0.5	-0.5	0	0
piso vs mixto	0	-0.5	0.5	-0.5	0.5	0	0
Interacción	0	-0.5	0.5	0.5	-0.5	0	0
leghorn vs white	0.5	0	0.5	0	0	-0.5	-0.5
jaula vs mixto	0.5	0	-0.5	0	0	0.5	-0.5
Interacción	0.5	0	-0.5	0	0	-0.5	0.5

Para realizar los contrastes en R-project, debemos construir una matriz con los contrastes a evaluar. La llamaremos *MatrizContrastes*, como se describe a continuación:



```

R Studio

MatrizContrastes= rbind(
"leghorn vs plymouth rock"=c(0, .5, .5,-.5,-.5, 0, 0),
"piso vs mixto"           =c(0,-.5, .5,-.5, .5, 0, 0),
"interac. "              =c(0,-.5, .5, .5,-.5, 0, 0),
"leghorn vs white rock"  =c(.5, 0, .5, 0, 0,-.5,-.5),
"jaula vs mixto"         =c(.5, 0,-.5, 0, 0, .5,-.5),
"interac. "              =c(.5, 0,-.5, 0, 0,-.5, .5))
MatrizContrastes

      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
leghorn vs plymouth rock 0.0 0.5 0.5 -0.5 -0.5 0.0 0.0
piso vs mixto             0.0 -0.5 0.5 -0.5 0.5 0.0 0.0
interac.                  0.0 -0.5 0.5 0.5 -0.5 0.0 0.0
leghorn vs white rock     0.5 0.0 0.5 0.0 0.0 -0.5 -0.5
jaula vs mixto            0.5 0.0 -0.5 0.0 0.0 0.5 -0.5
interac.                  0.5 0.0 -0.5 0.0 0.0 -0.5 0.5
    
```

Para analizar los contrastes utilizaremos la librería *gmodels* de Warnes (2012), mediante el comando *fit.contrast*. Los argumentos utilizados son: el nombre de la hoja donde está el análisis, el efecto que contiene los niveles para contrastar, la hoja donde está la matriz de contrastes y el nivel de confianza de la prueba de comparación de los contrastes:



```
library(gmodels)
Contrastes=fit.contrast(analisis,"gc", MatrizContrastes, conf=0.95)
Contrastes
```

	Estimate	Std. Error	t value	Pr(> t )	lower CI	upper CI
gcleghorn vs plymouth rock	3.333333e-01	1.320906	2.523521e-01	8.034787e-01	-2.431354	3.09802058
gcpiso vs mixto	6.666667e+00	1.320906	5.047042e+00	7.157971e-05	3.901979	9.43135391
gcinterac.	1.326556e-14	1.320906	1.004278e-14	1.000000e+00	-2.764687	2.76468724
gcleghorn vs white rock	-3.750000e-01	1.229696	-3.049534e-01	7.637170e-01	-2.948784	2.19878375
gcjaula vs mixto	-2.625000e+00	1.229696	-2.134674e+00	4.603144e-02	-5.198784	-0.05121625
gcinterac.	1.326556e-14	1.320906	1.004278e-14	1.000000e+00	-2.764687	2.76468724

Los análisis de los contrastes indicaron que la comparación de los sistemas piso y mixto en las líneas leghorn y plymouth presentó diferencias altamente significativas ( $p < 0.01$ ) a favor del sistema de piso (+6.6g) y la comparación de jaula y mixto en las líneas leghorn y white presentó diferencias significativas ( $p < 0.05$ ) a favor del sistema mixto en 2.625g. Los contrastes de las interacciones y los que relacionan la línea genética no fueron significativos ( $p > 0.5$ ).

## Capítulo 7

### Transformación de datos

Los métodos estadísticos de tipo paramétrico parten de supuestos teóricos como la normalidad, la homogeneidad de varianzas y la no correlación de errores, o la aditividad. Al no cumplirse uno o más de ellos, se puede invalidar el análisis estadístico de los resultados. Por tanto, es necesario utilizar transformaciones a los datos o aplicar métodos estadísticos no paramétricos.

Es muy frecuente utilizar transformaciones de los datos en análisis de varianzas empleando escala logarítmica, raíz cuadrada, función arco seno o familias de transformaciones que son aplicadas a los datos referentes a la variable respuesta. Estos datos transformados permiten efectuar análisis de varianza buscando asegurar los supuestos estadísticos, porque muchas de las pruebas estadísticas, como ancova, prueba  $t$ , análisis de regresión, análisis de varianza, manova o mancova, requieren que las observaciones cumplan con las siguientes características:

- ♣ Las observaciones deben provenir de una población con distribución normal
- ♣ Diferentes grupos de observaciones deben pertenecer a poblaciones que tengan varianzas homogéneas
- ♣ El análisis de varianza debe generar errores experimentales aleatorios, independientes, con distribución normal y varianzas homogéneas (no asociadas a los tratamientos)
- ♣ Adicional a los anteriores, no debe existir correlación entre los promedios y sus varianzas de los distintos tratamientos (por ejemplo: mayores promedios o mayores varianzas)

El análisis de varianza presenta robustés en el supuesto de normalidad, por lo que este supuesto no es tan importante como el de heterogeneidad de varianzas (Peña, 1994). Aunque la no presencia de normalidad afecta la homogeneidad de varianzas.

Veamos un ejemplo de un análisis de varianza donde no se cumple el supuesto de normalidad de los errores, el cual será utilizado para aplicar algunas de las transformaciones utilizadas en este capítulo.

El ejemplo consiste en evaluar el peso de cerdas sometidas a tres tratamientos ( $T$ ,  $C$  y  $D$ ) en un diseño completamente aleatorizado con efecto fijo, cuyo modelo es:

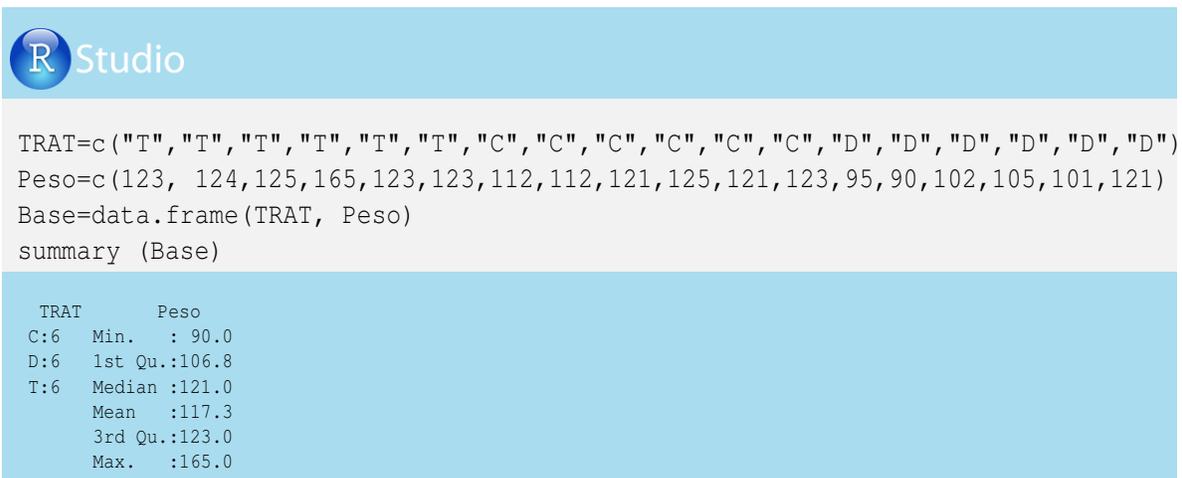
$$y_{ij} = \mu + \tau_i + \varepsilon_{j(i)}$$

En dicho modelo  $y_{ij}$  representa el peso de las cerdas, asociada al  $i$ -ésimo tratamiento ( $i = T, C$  y  $D$ ) en la  $j$ -ésima replicación  $\mu$  indica el efecto promedio del experimento,  $\tau_i$  indica el efecto promedio del  $i$ -ésimo tratamiento, y  $\varepsilon_{j(i)}$  es el error experimental asociado con la  $j$ -ésima replicación ( $j = 1, 2, \dots, J$ ), anidada en el  $i$ -ésimo tratamiento. Para tener validez, el análisis de varianza  $\varepsilon_{j(i)} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$ .

Los resultados obtenidos en el experimento fueron los siguientes:

Información de cerdas (kg)								
Cerda	Tratamiento	Peso	Cerda	Tratamiento	Peso	Cerda	Tratamiento	Peso
1	T	123	7	C	112	13	D	95
2	T	124	8	C	112	14	D	90
3	T	125	9	C	121	15	D	102
4	T	165	10	C	125	16	D	105
5	T	123	11	C	121	17	D	101
6	T	123	12	C	123	18	D	121

La construcción de la hoja de datos en R-project es:



```

R Studio

TRAT=c("T", "T", "T", "T", "T", "T", "C", "C", "C", "C", "C", "C", "D", "D", "D", "D", "D", "D")
Peso=c(123, 124, 125, 165, 123, 123, 112, 112, 121, 125, 121, 123, 95, 90, 102, 105, 101, 121)
Base=data.frame(TRAT, Peso)
summary (Base)

  TRAT      Peso
C:6  Min.   : 90.0
D:6  1st Qu.:106.8
T:6  Median :121.0
     Mean   :117.3
     3rd Qu.:123.0
     Max.   :165.0

```

Los datos anteriores fueron sometidos al análisis de varianza, en el cual los errores no presentaron una distribución normal. Realicemos primero el análisis de varianza y posteriormente aplicaremos la prueba de Shapiro-Wilk para normalidad:

```

R Studio

Res=lm(Base$Peso~ (Base$TRAT), data=Base)
summary.aov (Res)

          Df Sum Sq Mean Sq F value Pr(>F)
Base$TRAT  2   2407   1203.4    8.385 0.00359 **
Residuals 15    2153    143.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

A pesar de encontrar diferencias estadísticas altamente significativas ( $p < 0.05$ ), es necesario comprobar la existencia de normalidad y homogeneidad de varianzas. Obtendremos los residuos con el comando *residuals* y les aplicaremos la prueba de Shapiro-Wilk con el comando *shapiro.test*:

```

R Studio

Base$Rdos=residuals(Res)
ShapiroPeso=shapiro.test(Base$Rdos)
ShapiroPeso

Shapiro-Wilk normality test
data:  Base$Rdos
W = 0.7796, p-value = 0.0007933

```

La prueba de Shapiro-Wilk fue altamente significativa ( $p < 0.01$ ), indicando que los residuos no presentaron una distribución normal. Por este motivo hay que analizar si alguna transformación permitirá que estos residuos sean normales y que presenten homogeneidad de varianzas.

En las páginas siguientes presentamos algunas transformaciones que pueden ser utilizadas:

### 7.1. Estandarización zeta ( $z$ ) y transformaciones lineal y logarítmica

A continuación presentaremos tres transformaciones:

- ♣ La estandarización  $z$  convierte todos los valores de la población o muestra en medidas de desviación estándar por encima o por debajo de la media ( $\mu$ ). Estos valores transformados se llaman scores estándar o *z-scores*. Un score estándar es la distancia de un valor a la media de la distribución. La distribución resultante de este proceso tiene una media de cero ( $\mu = 0$ ) y una desviación estándar de uno ( $\sigma = 1$ ) y estaría dada por:

$$y_z = \frac{y - \mu}{\sigma}$$

La estandarización  $z$  no normaliza una distribución. La ventaja de este proceso consiste en hacer variables con diferentes unidades o escalas de medida comparables.

- ♣ La transformación lineal traslada los datos a un nuevo origen por expansión o contracción de la escala de medida. Tiene la siguiente forma general:

$$y_l = \beta_1 y + \beta_0$$

Donde:  $\beta_0$  es la translación o constante aditiva, y  $\beta_1$  es la constante multiplicativa.

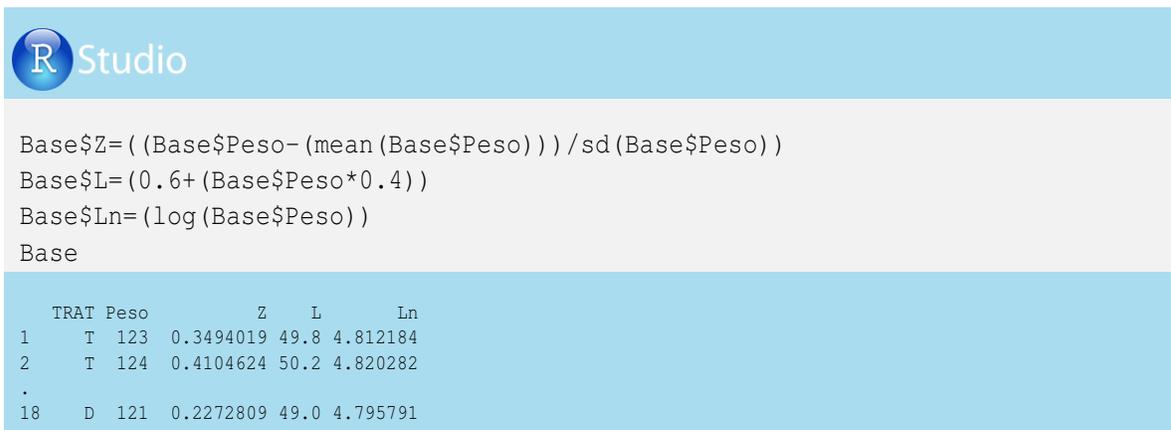
La constante de translación solamente agrega un valor constante a cada medida. Tiene el efecto de cambiar el origen del eje  $y$ , y la constante multiplicativa expande o contrae la escala, por consiguiente tiene un efecto de escala.

- ♣ La transformación logarítmica se usa para normalizar una distribución que tenga un sesgo positivo extremo. También es apropiada para alguna población cuando la varianza es proporcional a la media. Esta transformación generalmente es aplicable a variables de naturaleza continua, donde no existan valores negativos asociados con la variable respuesta: se pueden utilizar logaritmos naturales como en cualquier otra expresión logarítmica.

Ahora apliquemos las transformaciones anteriores a los datos de peso de las cerdas, con la siguiente descripción:

- ♣ En el caso de la transformación  $z$  utilizaremos  $y_{z_{ij}} = \frac{y_{ij} - \mu}{\sigma}$
- ♣ En el caso de la transformación lineal, con  $\beta_0 = 0.6$  y  $\beta_1 = 0.4$  (valores arbitrarios) utilizaremos  $y_{L_{ij}} = 0.6 + 0.4 y_{ij}$
- ♣ En el caso de la transformación con logaritmo natural utilizaremos  $y_{Ln_{ij}} = \ln(y_{ij})$

La hoja de datos en R-project con los datos originales y los transformados es:



```

R Studio

Base$Z= ( (Base$Peso- (mean (Base$Peso) ) ) /sd (Base$Peso) )
Base$L= ( 0.6+ (Base$Peso*0.4) )
Base$Ln= (log (Base$Peso) )
Base

  TRAT  Peso      Z      L      Ln
1     T   123  0.3494019 49.8 4.812184
2     T   124  0.4104624 50.2 4.820282
.
18    D   121  0.2272809 49.0 4.795791

```

Obtengamos el resumen de las variables *peso* y los pesos transformados utilizando el comando *summary*:



```
summary (Base)
```

TRAT	Peso	Z	L	Ln
C:6	Min. : 90.0	Min. :-1.6656	Min. :36.60	Min. :4.500
D:6	1st Qu.:106.8	1st Qu.: -0.6428	1st Qu.:43.30	1st Qu.:4.670
T:6	Median :121.0	Median : 0.2273	Median :49.00	Median :4.796
	Mean :117.3	Mean : 0.0000	Mean :47.51	Mean :4.756
	3rd Qu.:123.0	3rd Qu.: 0.3494	3rd Qu.:49.80	3rd Qu.:4.812
	Max. :165.0	Max. : 2.9139	Max. :66.60	Max. :5.106

Los promedios fueron:

$$\bar{y} = 117.3$$

$$\bar{y}_z = 0.0$$

$$\bar{y}_L = 47.51$$

$$\bar{y}_{Ln} = 4.756$$

Convertimos la hoja de datos en objeto (*attach*) y realizamos los análisis de varianza para cada una de las variables (peso y las transformaciones) con el efecto de tratamiento (*TRAT*). Utilicemos la opción *lm* anteriormente mencionada.



```
attach(Base)
Res=lm(Peso~ (TRAT), data=Base)
ResZ=lm(Z~ (TRAT), data=Base)
ResL=lm(L~ (TRAT), data=Base)
ResLn=lm(Ln~ (TRAT), data=Base)
```

No es necesario explicar en este momento los resultados de los análisis de varianza. Nos centraremos en la obtención de los residuos para realizar las pruebas de normalidad y homogeneidad de varianzas, con el comando *residuals*.



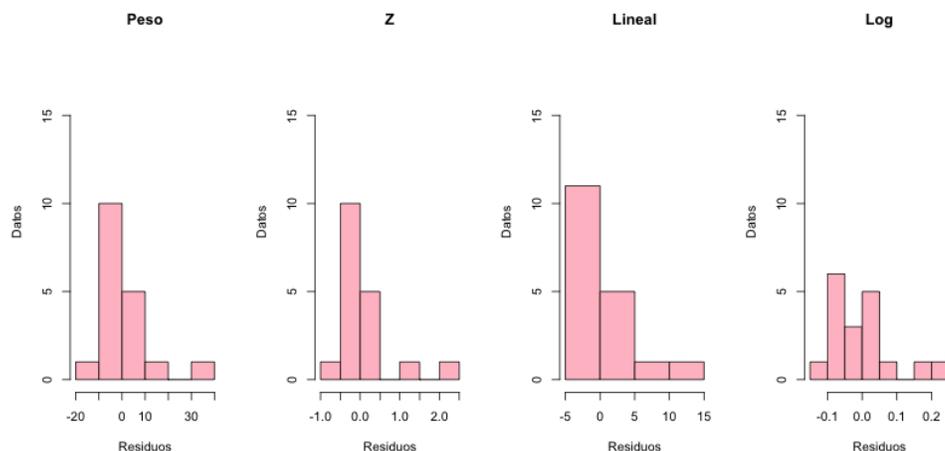
```
Base$Rdos =residuals(Res)
Base$RdosZ=residuals(ResZ)
Base$RdosL=residuals(ResL)
Base$RdosLn=residuals(ResLn)
attach(Base)
Base
```

	TRAT	Peso	Z	L	Ln	Rdos	RdosZ	RdosL	RdosLn
1	T	123	0.3494019	49.8	4.812184	-7.5000000	-0.45795396	-3.0000000	-0.052997952
2	T	124	0.4104624	50.2	4.820282	-6.5000000	-0.39689344	-2.6000000	-0.044900742
.									
18	D	121	0.2272809	49.0	4.795791	18.6666667	1.13979653	7.4666667	0.171868748

Ahora realicemos los histogramas de frecuencia de los residuos para la variable *peso* y las transformaciones, utilizando el comando *hist*. Los cuatro gráficos saldrán en una sola figura de cuatro columnas y una fila porque incluiremos el comando *par* y argumento *mfrow*):



```
par(mfrow=c(1,4))
hist(Rdos,breaks="Sturges",col="pink",main="Peso",
      xlab="Residuos",ylab="Datos",ylim=c(0,18))
hist(RdosZ,breaks="Sturges",col="pink",main="Z",
      xlab="Residuos",ylab="Datos",ylim=c(0,18))
hist(RdosL,breaks="Sturges",col="pink",main="Lineal",
      xlab="Residuos",ylab="Datos",ylim=c(0,18))
hist(RdosLn,breaks="Sturges",col="pink",main="Log",
      xlab="Residuos",ylab="Datos",ylim=c(0,18))
```



Como podemos apreciar, existe una semejanza visual de los residuos de peso y del peso estandarizado con la normal, lo que no ocurrió con la transformación lineal y la transformación logarítmica. Sin embargo, al presentar una aparente distribución normal, es necesario realizar la prueba de Shapiro-Wilk para verificar la distribución normal de los errores:



```
ShapiroPeso =shapiro.test(Rdos)
ShapiroZ=shapiro.test(RdosZ)
ShapiroL=shapiro.test(RdosL)
ShapiroLn=shapiro.test(RdosLn)
Shapiro= cbind (ShapiroPeso, ShapiroZ, ShapiroL, ShapiroLn)
Shapiro
```

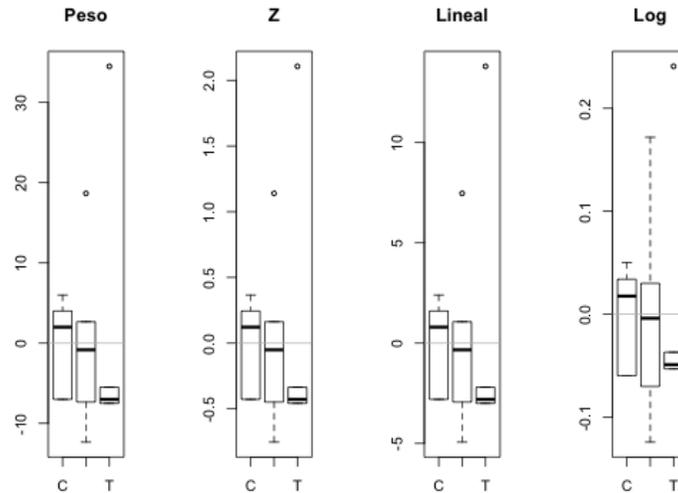
	ShapiroPeso	ShapiroZ
statistic	0.7795948	0.7795948
p.value	0.000793335	0.000793335
	ShapiroL	ShapiroLn
statistic	0.7795948	0.8466685
p.value	0.000793335	0.007507161

Según la prueba de Shapiro-Wilk, los errores de las variables *peso* y las transformaciones no tienen una distribución normal, porque los valores de  $p$  fueron menores que 0.01 ( $p < 0.01$ ). Por lo tanto, las anteriores transformaciones no lograron la normalidad de los residuos.

Ahora, mediante el comando *boxplot*, realizaremos un gráfico de los residuos por tratamiento para visualizar la dispersión de los residuos con la variable *peso* y el peso transformado:



```
par(mfrow=c(1,4))
boxplot(Rdos~TRAT, data=Base, main="Peso")
abline(h=mean(Rdos), col="gray")
boxplot(RdosZ~TRAT, data=Base, main="Z")
abline(h=mean(RdosZ), col="gray")
boxplot(RdosL~TRAT, data=Base, main="Lineal")
abline(h=mean(RdosL), col="gray")
boxplot(RdosLn~TRAT, data=Base, main="Log")
abline(h=mean(RdosLn), col="gray")
```



## 7.2. Transformación $y^2$ y $y^3$

La transformación  $y^2$  se usa para corregir sesgo negativo en la distribución de los errores y la transformación  $y^3$  para reducir sesgo negativo extremo. El sesgo negativo hace relación a una mayor concentración de datos a la derecha de la media. La forma de obtener  $y_{ij}^2$  y  $y_{ij}^3$  en R-project es:

```

R Studio

ycuadrado=y^2
ycubo=y^3

```

## 7.3. Transformación $\sqrt[2]{y}$ y *PROBIT*

- ♣ La transformación *Arcoseno*  $\sqrt[2]{y}$  o transformación angular es utilizada para variables relacionadas con porcentajes y proporciones. Dependiendo de los datos, se pueden aplicar las siguientes alternativas:

Cuando los porcentajes son menores que el 30% y hay presencia de ceros:

$$\text{Arcoseno} \sqrt{\frac{y+0.5}{100}}$$

Cuando los porcentajes son mayores que el 70%:

$$\text{Arcoseno} \sqrt{\frac{100-y}{100}}$$

Casos diferentes a los anteriores:

$$\text{Arcoseno} \sqrt{\frac{y}{100}}$$

- ♣ La transformación *PROBIT* se aplica a datos provenientes de una distribución binomial. Tiene como base la inversa de la tabla normal tipificada ( $N \sim (\mu = 0, \sigma^2 = 1)$ ) y estaría dada por:

$$\text{PROBIT} = 5 + \frac{y - \mu}{\sigma}$$

Veamos un ejemplo con las transformaciones anteriores aplicadas a la variable proporción de espermatozoides vivos en pajillas de cerdos sometidos a tres métodos de criopreservación, en escala de 0 a 1, mediante un diseño completamente aleatorizado, efecto fijo:

$$y_{ij} = \mu + \tau_i + \varepsilon_{j(i)}$$

Donde:  $y_{ij}$  representa la proporción de espermatozoides vivos por pajilla de cerdos, asociada al  $i$ -ésimo tratamiento ( $i = A, B \text{ o } C$ ) en la  $j$ -ésima replicación,  $\mu$  indica el efecto promedio del experimento,  $\tau_i$  indica el efecto promedio del  $i$ -ésimo tratamiento, y  $\varepsilon_{j(i)}$  es el error experimental asociado con la  $j$ -ésima replicación ( $j = 1, 2, \dots, J$ ), anidada en el  $i$ -ésimo tratamiento, garantizando que  $\varepsilon_{j(i)} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$ .

Los resultados encontrados fueron:

Proporción (y) de espermatozoides vivos por pajilla de cerdos								
Pajilla	Método	y	Pajilla	Método	y	Pajilla	Método	y
1	A	0.11	7	B	0.14	13	C	0.13
2	A	0.14	8	B	0.34	14	C	0.21
3	A	0.15	9	B	0.21	15	C	0.12
4	A	0.12	10	B	0.31	16	C	0.14
5	A	0.12	11	B	0.21	17	C	0.25
6	A	0.25	12	B	0.13	18	C	0.13

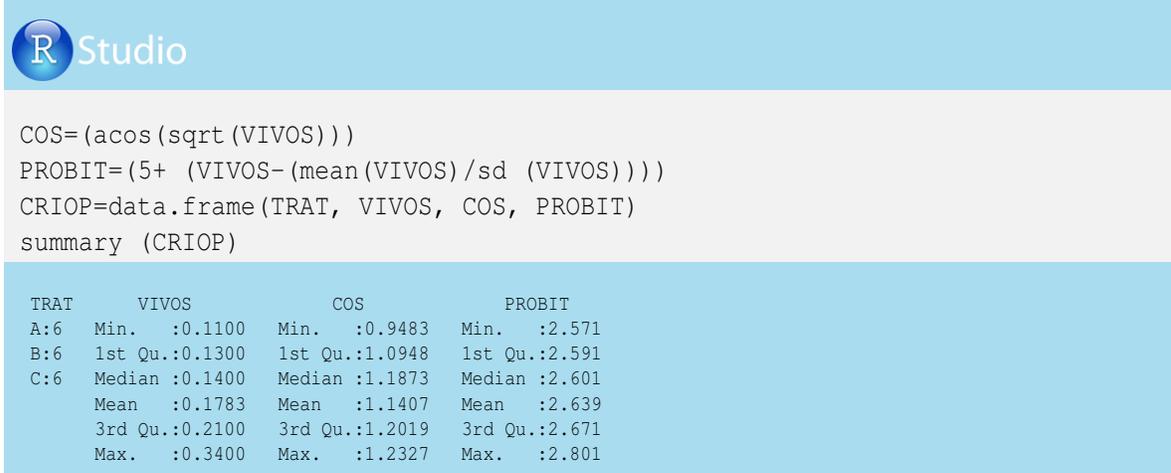
La construcción de la hoja de datos en R-project es:

```

R Studio
TRAT=c("A", "A", "A", "A", "A", "A", "B", "B", "B", "B", "B", "B", "C", "C", "C", "C", "C", "C")
VIVOS=c(0.11, 0.14, 0.15, 0.12, 0.12, 0.25, 0.14, 0.34, 0.21, 0.31, 0.21, 0.13, 0.13, 0.21,
        0.12, 0.14, 0.25, 0.13)
    
```

- ♣ La transformación arcoseno sería  $y_{arcij} = \text{Arcoseno} \sqrt{y_{ij}}$
- ♣ La transformación PROBIT sería  $y_{probij} = 5 + \frac{y_{ij} - \mu}{\sigma}$

La hoja de datos sería:



```

R Studio

COS=(acos(sqrt(VIVOS)))
PROBIT=(5+(VIVOS-(mean(VIVOS)/sd(VIVOS))))
CRIOP=data.frame(TRAT, VIVOS, COS, PROBIT)
summary(CRIOP)

```

TRAT	VIVOS	COS	PROBIT
A:6	Min. :0.1100	Min. :0.9483	Min. :2.571
B:6	1st Qu.:0.1300	1st Qu.:1.0948	1st Qu.:2.591
C:6	Median :0.1400	Median :1.1873	Median :2.601
	Mean :0.1783	Mean :1.1407	Mean :2.639
	3rd Qu.:0.2100	3rd Qu.:1.2019	3rd Qu.:2.671
	Max. :0.3400	Max. :1.2327	Max. :2.801

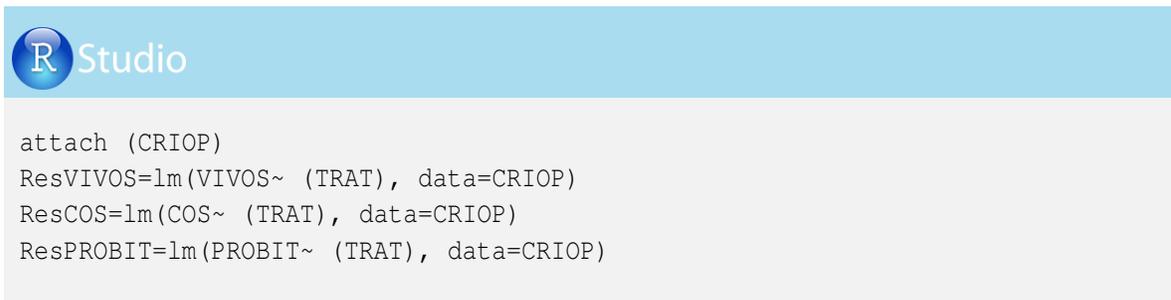
Los promedios y las medianas fueron:

$\bar{y} = 0.178$ , con mediana de 0.14

$\bar{y}_{arc} = 1.14$ , con mediana de 1.18

$\bar{y}_{prob} = 2.64$ , con mediana de 2.60

Convertimos la hoja de datos en objeto (*attach*) y realizamos los análisis de varianza para cada una de las variables (proporción de espermatozoides vivos y las dos transformaciones). Recordemos que es un diseño completamente al azar con tres métodos de criopreservación de semen:



```

R Studio

attach(CRIOP)
ResVIVOS=lm(VIVOS~(TRAT), data=CRIOP)
ResCOS=lm(COS~(TRAT), data=CRIOP)
ResPROBIT=lm(PROBIT~(TRAT), data=CRIOP)

```

El usuario puede deducir e interpretar los siguientes resultados de los análisis de varianza, principalmente en las sumas de cuadrados, en los cuadrados medios y en los valores de  $F$ :

```

R Studio

summary.aov (ResVIVOS)
summary.aov (ResCOS)
summary.aov (ResPROBIT)

> summary.aov (ResVIVOS)
      Df Sum Sq Mean Sq F value Pr(>F)
TRAT    2  0.01890  0.00945    2.182  0.147
Residuals 15  0.06495  0.00433
> summary.aov (ResCOS)
      Df Sum Sq Mean Sq F value Pr(>F)
TRAT    2  0.02962  0.014811    2.184  0.147
Residuals 15  0.10173  0.006782
> summary.aov (ResPROBIT)
      Df Sum Sq Mean Sq F value Pr(>F)
TRAT    2  0.01890  0.00945    2.182  0.147
Residuals 15  0.06495  0.00433

```

Obtengamos los residuos para realizar las pruebas de normalidad y homogeneidad de varianzas, con el comando *residuals*.

```

R Studio

CRIOP$RdosVIVOS=residuals (ResVIVOS)
CRIOP$RdosCOS=residuals (ResCOS)
CRIOP$RdosPROBIT=residuals (ResPROBIT)

```

Convirtamos nuevamente la hoja de datos en objeto (*attach*) y veamos los valores de los residuos de la variable sin transformar y transformada:

```

R Studio

attach (CRIOP)
CRIOP

  TRAT VIVOS      COS PROBIT  RdosVIVOS  RdosCOS  RdosPROBIT
1    A  0.11  1.2327311  2.57075 -0.038333333  0.053658689 -0.038333333
2    A  0.14  1.1872993  2.60075 -0.008333333  0.008226940 -0.008333333
.
18   C  0.13  1.2019333  2.59075 -0.033333333  0.043569878 -0.033333333

```

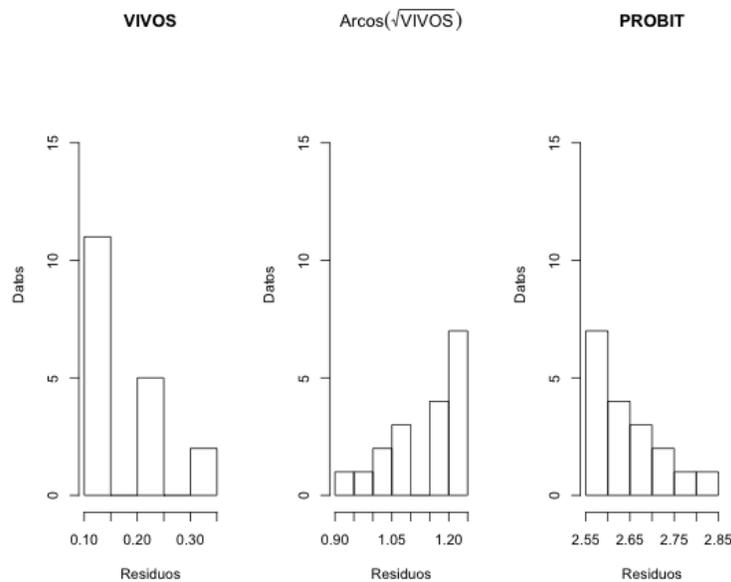
Ahora realicemos los histogramas de frecuencia de los residuos, utilizando el comando *hist*. Los tres gráficos saldrán en una sola figura (comando *par* y argumento *mflow*). En el caso de la gráfica de la transformación Arcoseno, incluimos una programación para generar un título que tenga el símbolo de la raíz cuadrada, al cual llamaremos *TITULOcos*:



```

par (mfrow=c(1,3))
TITULOcos=(expression(Arcos(sqrt(VIVOS))))
hist (VIVOS,breaks="Sturges",main="VIVOS", xlab="Residuos",
      ylab="Datos", ylim=c(0,18))
hist (COS,breaks="Sturges",main=TITULOcos,xlab="Residuos",
      ylab="Datos", ylim=c(0,18))
hist (PROBIT,breaks="Sturges",main=" PROBIT ",xlab="Residuos",
      ylab="Datos",ylim=c(0,18))

```



Como podemos apreciar en los gráficos anteriores, existe una semejanza visual de la distribución de los residuos para  $y$ ,  $y_{arc}$  y  $y_{prob}$ . Realicemos la prueba de Shapiro-Wilk para verificar la distribución normal de los errores:



```

ShapiroVIVOS=shapiro.test (RdosVIVOS)
ShapiroCOS=shapiro.test (RdosCOS)
ShapiroPROBIT=shapiro.test (RdosPROBIT)
Shapiro= cbind (ShapiroVIVOS, ShapiroCOS, ShapiroPROBIT)
Shapiro

```

	ShapiroVIVOS	ShapiroCOS	ShapiroPROBIT
statistic	0.8925088	0.9009177	0.8925088
p.value	0.04261842	0.05961322	0.04261842

Según la prueba de Shapiro-Wilk, los errores de las variables  $y$  y  $y_{prob}$  fueron significativos ( $p < 0.05$ ), indicando que los residuos no se distribuyeron normalmente, caso contrario a la variable  $y_{arc}$ . Verifiquemos si las variables anteriores presentan homogeneidad de varianzas de los errores en los tratamientos, mediante la prueba de Cochran de la librería *outliers* de Komsta (2011):



```

library(outliers)
CochranVIVOS=cochran.test(RdosVIVOS~TRAT,data=CRIOP)
CochranCOS=cochran.test(RdosCOS~TRAT,data=CRIOP)
CochranPROBIT=cochran.test(RdosPROBIT~TRAT,data=CRIOP)
Cochran= cbind (CochranVIVOS, CochranCOS, CochranPROBIT)
Cochran

      CochranVIVOS
statistic 0.5717218
alternative "Group B has outlying variance"
p.value   0.2623571
      CochranCOS
statistic 0.5307062
alternative "Group B has outlying variance"
p.value   0.3822808
      CochranPROBIT
statistic 0.5717218
alternative "Group B has outlying variance"
p.value   0.2623571
    
```

Según la prueba de Cochran, las tres variables presentaron homogeneidad de varianzas de los tratamientos, porque tuvieron un  $p > 0.05$ . Queda en manos del lector realizar las pruebas de Bartlett y Fligner con los comandos *bartlett.test* y *fligner.test*, respectivamente.

En conclusión para este ejemplo: Es necesario realizar la transformación de los datos utilizando la transformación arcoseno.

#### 7.4. Transformación $\sqrt[3]{y}$

La transformación  $\sqrt[3]{y}$  es aplicable a datos de conteo cuya distribución probabilística es el modelo de Poisson. Cuando existen ceros en la variable de conteo se ajusta la transformación adicionándole un valor constante 0.5 o 1.

Veamos un ejemplo del número de lechones obtenidos por camada en hembras sometidas a tres tratamientos en un diseño completamente al azar:

$$y_{ij} = \mu + \tau_i + \varepsilon_{j(i)}$$

Donde:  $y_{ij}$  es el número de lechones por parto de la  $j$ -ésima cerda, sometida al  $i$ -ésimo tratamiento ( $i = T, C o D$ ),  $\mu$  indica el efecto promedio del experimento, y  $\varepsilon_{j(i)}$  es el error experimental ( $\varepsilon_{j(i)} \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$ ).

Los resultados encontrados fueron:

Número de lechones por parto								
Pajilla	Método	y	Pajilla	Método	y	Pajilla	Método	y
1	T	13	7	C	12	13	B	14
2	T	11	8	C	13	14	B	11
3	T	13	9	C	14	15	B	11
4	T	11	10	C	10	16	B	11
5	T	11	11	C	10	17	B	13
6	T	12	12	C	11	18	B	10

La construcción de la hoja de datos en R-project es:



```
TRAT=c("T","T","T","T","T","T","C","C","C","C","C","C","D","D","D","D","D","D")
CRIAS=c(13,11,13,11,11,12,12,13,14,10,10,11,14,11,11,11,13,10)
```

La transformación raíz cuadrada ( $y_{rij} = \sqrt[2]{y_{ij}}$ ) se realiza de la siguiente forma:



```
RAIZ=(sqrt(CRIAS))
BRZ=data.frame(TRAT, CRIAS, RAIZ)
summary(BRZ)
```

TRAT	CRIAS	RAIZ
C:6	Min. :10.00	Min. :3.162
D:6	1st Qu.:11.00	1st Qu.:3.317
T:6	Median :11.00	Median :3.317
	Mean :11.72	Mean :3.419
	3rd Qu.:13.00	3rd Qu.:3.606
	Max. :14.00	Max. :3.742

Los promedios y las medianas fueron:

$\bar{y} = 11.72$ , con mediana de 11

$\bar{y}_r = 3.419$ , con mediana de 3.316

Convertimos la hoja de datos en objeto (*attach*) y realizamos los análisis de varianza para cada una de las variables: número de lechones y la transformación por raíz cuadrada:



```
ResCRIAS=lm(CRIAS~ (TRAT), data=BRZ)
ResRAIZ=lm(RAIZ~ (TRAT), data=BRZ)
```

Queda en manos del lector interpretar los siguientes resultados:



```
summary.aov (ResCRIAS)
summary.aov (ResRAIZ)
```

```
> summary.aov (ResCRIAS)
      Df Sum Sq Mean Sq F value Pr(>F)
TRAT    2  0.111  0.0556   0.028  0.972
Residuals 15 29.500  1.9667
> summary.aov (ResRAIZ)
      Df Sum Sq Mean Sq F value Pr(>F)
TRAT    2  0.0032  0.00159   0.039  0.962
Residuals 15  0.6194  0.04129
```

Obtengamos los residuos para realizar las pruebas de normalidad y homogeneidad de varianzas con el comando *residuals*, posteriormente convirtamos la hoja de datos en objeto con el comando *attach*.

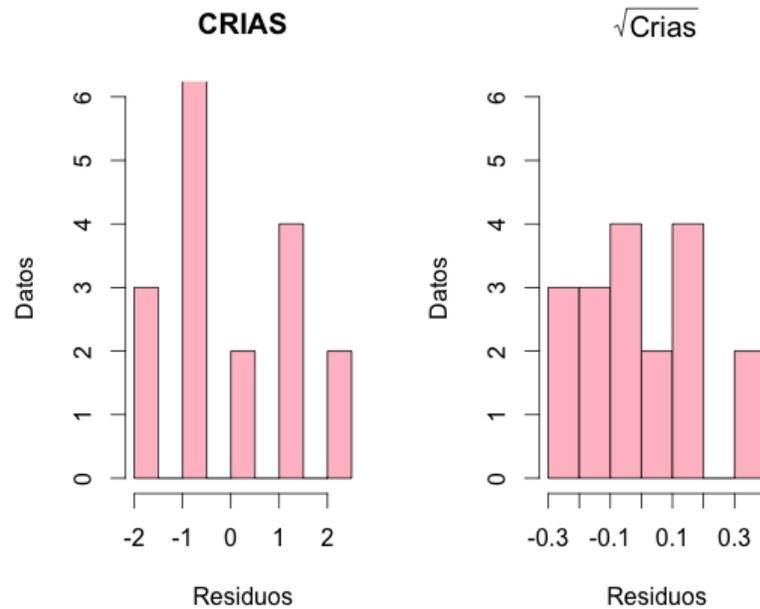


```
BRZ$RdosCRIAS=residuals(ResCRIAS)
BRZ$RdosRAIZ=residuals(ResRAIZ)
attach (BRZ)
```

Ahora realicemos los histogramas de frecuencia de los residuos, utilizando el comando *hist*, como se indicó en las anteriores transformaciones:



```
par (mfrow=c(1,2))
TITULO=(expression(sqrt(Crias)))
hist(RdosCRIAS, breaks="Sturges",col="pink", main="CRIAS",
      xlab="Residuos",ylab="Datos", ylim=c(0,6))
hist(RdosRAIZ, breaks="Sturges",col="pink", main=TITULO,
      xlab="Residuos", ylab="Datos", ylim=c(0,6))
```



Los residuos en la variable  $\sqrt[3]{y}$  presentaron menor variación que la variable  $y$ ; realicemos la prueba de Shapiro-Wilk para verificar la normalidad de los errores:



```
ShapiroCRIAS=shapiro.test(RdosCRIAS)
ShapiroRAIZ=shapiro.test(RdosRAIZ)
Shapiro= cbind (ShapiroCRIAS, ShapiroRAIZ)
Shapiro
```

	ShapiroCRIAS	ShapiroRAIZ
statistic	0.89646	0.9030983
p.value	0.0498701	0.06507801

Según la prueba de Shapiro-Wilk, los errores de la variable  $\sqrt[3]{y}$  presentaron una distribución normal ( $p < 0.05$ ), lo que no ocurrió con la variable  $y$  (número de lechones por parto). Verifiquemos si las variables anteriores presentan homogeneidad de varianzas de los errores en los tratamientos, mediante la prueba de Cochran:

```

R Studio

CochranCRIAS=cochran.test(RdosCRIAS~TRAT,data=BRZ)
CochranRAIZ=cochran.test(RdosRAIZ~TRAT,data=BRZ)
Cochran= cbind(CochranCRIAS, CochranRAIZ)
Cochran

      CochranCRIAS      CochranRAIZ
statistic 0.4519774      0.4562662
alternative "Group C has outlying variance" "Group C has outlying variance"
p.value    0.7017855      0.6812793
    
```

Según la prueba de Cochran,  $y$  y  $\sqrt[3]{y}$  presentaron homogeneidad de varianzas, porque tuvieron un valor de  $p > 0.05$ . Para este caso, se justificó la transformación de la variable.

### 7.5. Transformación *Box-Cox*

La transformación *Box-Cox* busca un valor (denominado  $\lambda$ ) que sirve como exponente de una variable ( $y$ ), dando origen a una variable  $y^\lambda$ .

En un rango de valores de  $\lambda$  se escoge el que, al ser aplicado a  $y_i$  y realizado el análisis de varianza, genere el menor valor de la suma de cuadrados del residuo.

*Box-Cox* es considerada una familia de transformaciones porque con ella se pueden generar varias formas de transformar los datos, entre ellas:

- ♣ Si los valores de  $y_i > 0$

$$y_i(\lambda) = \begin{cases} \frac{(y_i^\lambda - 1)}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(y_i) & \text{si } \lambda = 0 \end{cases}$$

- ♣ Si alguno de los valores de  $y$  es menor que cero, se procede a sumar un valor constante ( $m$ ) hasta que  $(y_i + m) > 0$

$$y_i(\lambda) = \begin{cases} \frac{((y_i + m)^\lambda - 1)}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(y_i + m) & \text{si } \lambda = 0 \end{cases}$$

- ♣ Usando la media geométrica de  $y$ , con  $y > 0$ :

$$y_i(\lambda) = \begin{cases} \frac{(y_i^\lambda - 1)}{\lambda \check{y}_i^{(\lambda-1)}} & \text{si } \lambda \neq 0 \\ \check{y}_i(\lg(y_i)) & \text{si } \lambda = 0 \end{cases}$$

Donde  $\check{y}$  es la media geométrica de  $y$ , y está dada por:

$$\check{y} = (y_1 * y_2 * \dots * y_n)^{1/n}$$

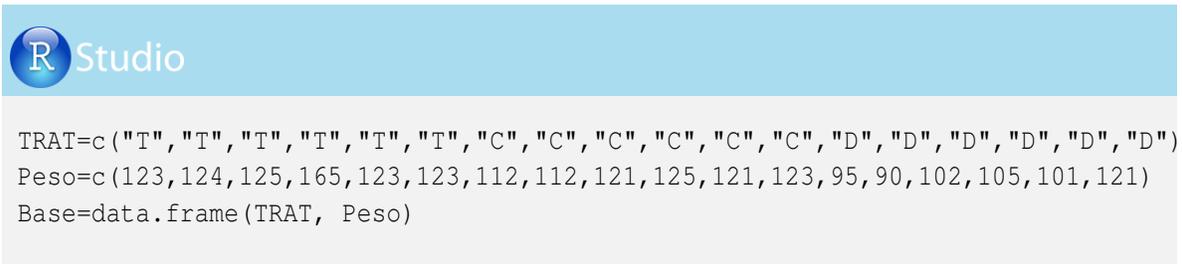
La familia de transformaciones de *Box-Cox* corrige problemas de normalidad y en la mayoría de los casos permite que las variables presenten homogeneidad de varianzas.

Ahora observemos un ejemplo en R-project para realizar una transformación *Box-Cox*. Utilizaremos el ejemplo del inicio de este capítulo, donde tenemos la variable dependiente peso (kg) de cerdas sometidas a tres tratamientos (T, C y D) en un diseño completamente aleatorizado, con efecto fijo:

$$y_{ij} = \mu + \tau_i + \varepsilon_{j(i)}$$

En dicho modelo,  $y_{ij}$  representa el peso de las cerdas, asociado al  $i$ -ésimo tratamiento ( $i = T, C$  y  $D$ ) en la  $j$ -ésima replicación,  $\mu$  indica el efecto promedio del experimento,  $\tau_i$  indica el efecto promedio del  $i$ -ésimo tratamiento, y  $\varepsilon_{j(i)}$  es el error experimental asociado con la  $j$ -ésima replicación ( $j = 1, 2, \dots, J$ ), anidada en el  $i$ -ésimo tratamiento, la cual tiene que tener distribución normal y homogeneidad de los errores.

Generemos la hoja de datos *Base* en R-project:



```

R Studio

TRAT=c("T", "T", "T", "T", "T", "T", "C", "C", "C", "C", "C", "C", "D", "D", "D", "D", "D", "D")
Peso=c(123, 124, 125, 165, 123, 123, 112, 112, 121, 125, 121, 123, 95, 90, 102, 105, 101, 121)
Base=data.frame(TRAT, Peso)

```

Para realizar la transformación *Box-Cox* se pueden utilizar varias librerías del R-project. En este capítulo utilizaremos la librería *lmSupport*, desarrollada por Curtin (2012).

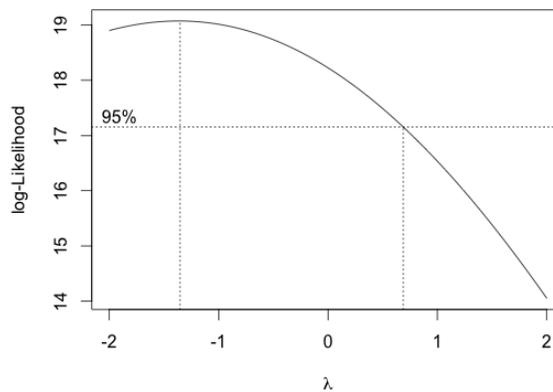
Inicialmente se realiza el análisis de varianza con la opción *lm*, y luego se utiliza el comando *Box-Cox*, el cual genera el mejor  $\lambda$  para la transformación del peso, según el modelo propuesto para el análisis de varianza.

La opción *Box-Cox* también genera un gráfico de diferentes valores de  $\lambda$  y su respectivo valor del logaritmo de máxima verosimilitud. En este gráfico se puede apreciar el punto máximo que corresponde al mejor  $\lambda$  (*Bestlambda* =). Veamos:



```
Res=lm(Peso~ (TRAT), data=Base)
install.packages("lmSupport")
library (lmSupport)
BoxCox=lm.boxCox(Res, seq(-2, 2, by =0.1))
```

```
[1] "Best Lambda= -1.35"
[1] "Chi-square (df=1)= 5.12"
[1] "p-value= 0.02361"
```



El mejor  $\lambda$  fue  $-1.35$ , por lo tanto  $y_{ij}(\lambda) = \frac{(y_{ij}^{-1.35}-1)}{-1.35}$

A continuación desarrollamos el análisis de varianza con el  $\lambda$  adecuado, pero a manera de ejemplo utilizaremos otros valores de  $\lambda$ , uno menor ( $\lambda = -2$ ) y otro mayor ( $\lambda = 2$ ), con el objetivo de comparar la capacidad de ajuste con diferente  $\lambda$ :

- ♣  $\lambda$  menor:  $y_{mr_{ij}} = \frac{(y_{ij}^{-2}-1)}{-2}$
- ♣  $\lambda$  adecuado  $y_{ad_{ij}} = \frac{(y_{ij}^{-1.35}-1)}{-1.35}$
- ♣  $\lambda$  mayor  $y_{my_{ij}} = \frac{(y_{ij}^2-1)}{2}$

Crearemos en R-project las tres variables con las transformaciones: *Lambmr*, *Lambad* y *Lambmy* para los  $\lambda = -2$ ,  $\lambda = -1.35$  y  $\lambda = 2$ , respectivamente:



```
Base$Lambmr = (Base$Peso^(-2.00)-1)/(-2.00)
Base$Lambad = (Base$Peso^(-1.35)-1)/(-1.35)
Base$Lambmy = (Base$Peso^( 2.00)-1)/( 2.00)
attach (Base)
summary (Base)
```

TRAT	Peso	Lambmr	Lambad	Lambmy
C:6	Min. : 90.0	Min. : 0.4999	Min. : 0.7390	Min. : 4050
D:6	1st Qu.: 106.8	1st Qu.: 0.5000	1st Qu.: 0.7394	1st Qu.: 5702
T:6	Median : 121.0	Median : 0.5000	Median : 0.7396	Median : 7320
	Mean : 117.3	Mean : 0.5000	Mean : 0.7395	Mean : 7003
	3rd Qu.: 123.0	3rd Qu.: 0.5000	3rd Qu.: 0.7396	3rd Qu.: 7564
	Max. : 165.0	Max. : 0.5000	Max. : 0.7400	Max. : 13612

Ahora realicemos los análisis de varianza para las variables *peso* y las tres transformaciones, teniendo en cuenta el efecto fijo de tratamiento (*TRAT*), y generemos los residuos:



```
Res =lm(Base$Peso ~ (Base$TRAT), data=Base)
Resmr=lm(Base$Lambmr~ (Base$TRAT), data=Base)
Resad=lm(Base$Lambad~ (Base$TRAT), data=Base)
Resmy=lm(Base$Lambmy~ (Base$TRAT), data=Base)
```

Por el momento no nos preocupemos del análisis de varianza; comprobemos si los residuos se distribuyen normalmente y presentan igualdad de varianzas para los tratamientos.

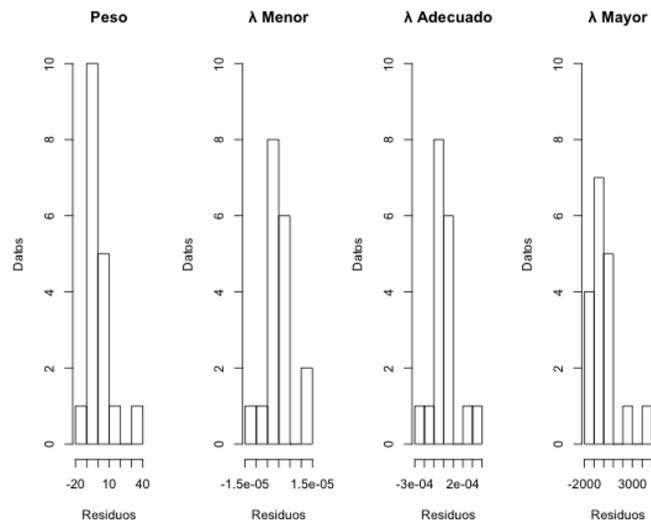


```
Base$Rdos=residuals(Res)
Base$Rdosmr=residuals(Resmr)
Base$Rdosad=residuals(Resad)
Base$Rdosmy=residuals(Resmy)
attach (Base)
```

Los histogramas de los residuos de las variables *peso* y las transformaciones *Box-Cox* con diferente  $\lambda$  son:



```
par(mfrow=c(1,4))
hist(Rdos,breaks="Sturges",main="Peso",xlab="Residuos",
      ylab="Datos",ylim=c(0,10))
hist(Rdosmr,breaks="Sturges",main="λ Menor",
      xlab="Residuos",ylab="Datos",ylim=c(0,10))
hist(Rdosad,breaks="Sturges",main="λ Adecuado",
      xlab="Residuos",ylab="Datos",ylim=c(0,10))
hist(Rdosmy,breaks="Sturges",main="λ Mayor",
      xlab="Residuos",ylab="Datos",ylim=c(0,10))
```



Ahora realicemos las pruebas de Shapiro-Wilk, Cochran, Bartlett y Fligner para ver la normalidad de los residuos y la heterogeneidad de varianzas de los errores:



```
ShapiroPeso=shapiro.test(Rdos)
ShapiroMr=shapiro.test(Rdosmr)
ShapiroAd=shapiro.test(Rdosad)
ShapiroMy=shapiro.test(Rdosmy)
Shapiro=cbind(ShapiroPeso,ShapiroMr,ShapiroAd,ShapiroMy)
Shapiro
```

	ShapiroPeso	ShapiroMr
statistic	0.7795946	0.9262246
p.value	0.0007933304	0.166596
	ShapiroAd	ShapiroMy
statistic	0.8997604	0.7219877
p.value	0.0569078	0.0001455385

Según la prueba de Shapiro-Wilk, los residuos de las variables  $y$  y la transformación *Box-Cox* con  $\lambda = 2$  no presentaron una distribución normal al igual que las transformaciones del inicio de este capítulo (*Z*, Lineal y Logaritmo). Los residuos del peso transformado por *Box-Cox* con  $\lambda = -2$  y con  $\lambda = -1.35$  presentaron una distribución normal.

Ahora realicemos las pruebas de Cochran y Bartlett para  $\lambda = -2$  y  $\lambda = -1.35$ :



```

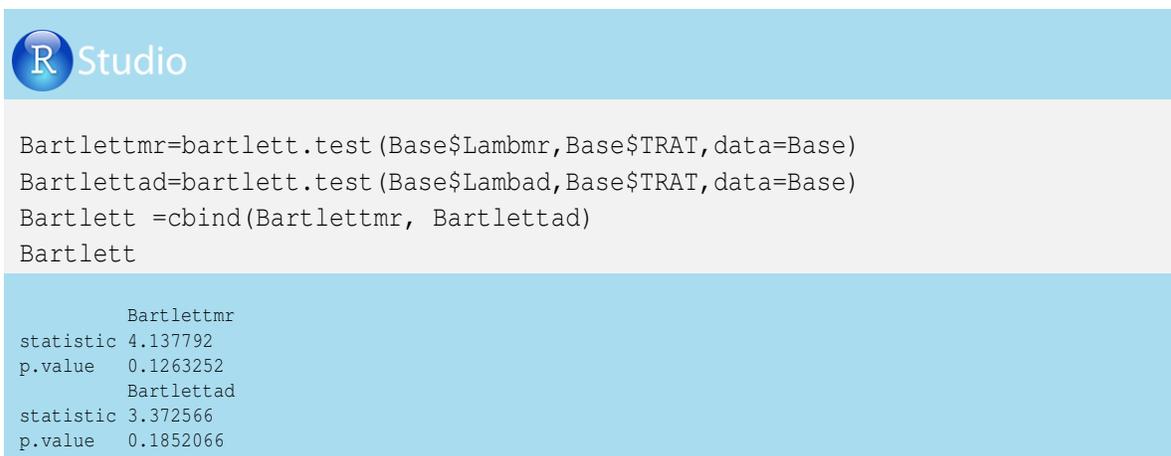
R Studio

library(outliers)
Cochranmr=cochran.test(Rdosmr~TRAT,data=Base)
Cochranad=cochran.test(Rdosad~TRAT,data=Base)
Cochran =cbind (Cochranmr, Cochranad)
Cochran

      Cochranmr
statistic 0.6527679
alternative "Group D has outlying variance"
p.value   0.1066035
      Cochranad
statistic 0.5696935
alternative "Group D has outlying variance"
p.value   0.2675832

```

Por la prueba de Cochran, las dos transformaciones presentaron homogeneidad de varianzas de los residuos entre tratamientos ( $p > 0.05$ ), indicando que la transformación *Box-Cox* fue eficiente para transformar los datos de peso de cerdas sometidas a tres tratamientos. A manera de ejercicio, realicemos la prueba de Bartlett:



```

R Studio

Bartlettmr=bartlett.test(Base$Lambmr,Base$TRAT,data=Base)
Bartlettad=bartlett.test(Base$Lambad,Base$TRAT,data=Base)
Bartlett =cbind(Bartlettmr, Bartlettad)
Bartlett

      Bartlettmr
statistic 4.137792
p.value   0.1263252
      Bartlettad
statistic 3.372566
p.value   0.1852066

```

Al igual que la prueba de Cochran, la prueba de Bartlett también indicó que las dos transformaciones presentaron homogeneidad de varianzas de los residuos entre tratamientos.

## Bibliografía

La siguiente bibliografía fue citada o consultada para la elaboración del libro:

**Aho K 2012** asbio. A collection of statistical tools for biologists. Contributed from Winston V, Maintainer R D. R package versión 0.4. <http://CRAN.R-project.org/package=asbio>

**Amézquita M 1990** Diseño y análisis de ensayos para evaluación de pasturas en fincas: Séptima reunión del Comité Asesor de la RIEPT-CIAT, Agosto 27-29 de 1990. <http://201.234.78.28:8080/jspui/bitstream/123456789/3460/1/39666.pdf>

**Arnau G.J. Viader JM sf** Diseños cross-over (alternativos): Aspectos metodológicos y analíticos. [http://www.quadernsdigitals.net/datos\\_web/articles/quriculum/quriculum1/qrlidisenyoscrossover.pdf](http://www.quadernsdigitals.net/datos_web/articles/quriculum/quriculum1/qrlidisenyoscrossover.pdf)

**Balluerka N, Vergara AI 2002** Diseños de investigación experimental en Psicología. Pearson Educación. Madrid. ISBN:84-205-3447-1.

**Bartlett MS 1947** The use of transformations. *Biometrics*. 3 (1): 39-52

**Beall G 1940** The transformation of data from entomological field experiments. *Canadian Entomologist* 72: 168.

**Box GEP, Cox DR 1964** An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B*. 26: 211-252

**Calzadilla J, Guerra W, Torres V. 2002** El uso y abuso de transformaciones matemáticas. Aplicaciones en modelos de análisis de varianza. *Revista Cubana de Ciencia Agrícola*. 36(2): 103-106. <http://redalyc.uaemex.mx/redalyc/pdf/1930/193018119002.pdf>

**Chambers JM 1992** Linear models. Chapter 4 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

**Chambers JM, Freeny A, Heiberger RM 1992** Analysis of variance; designed experiments. Chapter 5 of *Statistical Models in S* eds JM Chambers and TJ Hastie, Wadsworth and Brooks/Cole.

**Cochran WG 1947** Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*. 3(1): 22-38

**Cochran WG, Autrey KM, Canon CY 1941** A double change-over design for dairy cattle feeding experiments. *Journal of Dairy Science*. 24(11): 937-951

**Cochran WG, Cox GM 1957** *Experimental Designs*. Segunda Edición. New York, John Wiley and Sons. Wiley Classics Library Edition published 1992. ISBN:0-471-5467-8. 611p

**Correa JC, Iral R, Rojas L 2006** Estudio de potencia de pruebas de homogeneidad de varianza. *Revista Colombiana de Estadística*. 29(1):57-76. <http://www.scielo.org.co/pdf/rce/v29n1/v29n1a04.pdf>

**Curtin J 2012** lmSupport: Support for Linear Models. R package version 1.07. <http://CRAN.R-project.org/package=lmSupport>

**Deepayan s 2008** Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5

**Di Rienzo JA, Casanoves F, González LA, Tablada EM, Díaz MP, Robledo CW, Balzarini MG 2005** Estadística para las ciencias agropecuarias. sexta edición (edición electrónica). Editorial las Brujas. Córdoba, Argentina. <http://es.scribd.com/doc/52816014/201/Unidad-experimental>

**Everitt BS, Hothorn T** *sf* A Handbook of statistical analyses using R. Analysis of Variance: Chapter 4. Weight Gain, Foster Feeding in Rats, Water Hardness and Male Egyptian Skulls. [http://cran.r-project.org/web/packages/HSAUR/vignettes/Ch\\_analysis\\_of\\_variance.pdf](http://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_analysis_of_variance.pdf)

**Finney DJ 1941** On the distribution of a variate whose logarithm is normally distributed. *Journal of the Royal Statistical Society*. 7(2) supplement:155-161.

**Fox J, Weisberg S 2011** An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. <http://socserv.socsci.mcmaster.ca/~jfox/Books/Companion>

**Galbiati RJ** Diseño de experimentos factoriales aplicados a procesos industriales. [http://www.jorgegalbiati.cl/enero\\_07/VariaCompleto.pdf](http://www.jorgegalbiati.cl/enero_07/VariaCompleto.pdf)

**Gao L 2005** Latin squares in experimental design. Michigan State University. <http://www.mth.msu.edu/~jhall/classes/MTH880-05/Projects/latin.pdf>

**Godfrey A 2010** Drugs demonstrating using graphics and statistics. Institute of Fundamental Sciences, Massey University. New Zealand. ISBN: 978-0-473-17651-8. <http://r-resources.massey.ac.nz/drugs2-14-1.pdf>

**Heitlinger EG 2010** Transcript of Mick Crawley's R course 2010. Imperial College London, Silwood Park. <http://www.docstoc.com/docs/32130429/Transcript-of-Mick-Crawleys-R-course-2010-Imperial-College>

**Højsgaard S, Halekoh U 2012** doBy: doBy - Groupwise summary statistics, general linear contrasts, population means (least-squares-means), and other utilities. Contributed by Robison-Cox J, Wright K, Leidi A (2012). Version 4.5-5. <http://CRAN.R-project.org/package=doBy>

**Hothorn T, Bretz F, Westfall P 2008** Simultaneous Inference in General Parametric Models. *Biometrical Journal* 50(3): 346-363

**Jonh JA, Draper NR 1980** An alternative family of transformations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(2):190-197

**Kempthorne O, Folks L 1961** Probability Statistics and Data Analysis. Iowa State University Press. ISBN: 978-08-813824703.

**Komsta L 2011** outliers: Tests for outliers. R package version 0.14. <http://CRAN.R-project.org/package=outliers>

**Lane DM, Scott D, Hebl M, Guerra R, Osherson D, Zimmer H** *sf* Online statistics education: A multimedia course of study. [http://onlinestatbook.com/Online\\_Statistics\\_Education.pdf](http://onlinestatbook.com/Online_Statistics_Education.pdf)

**Lenth RV 2012** lsmeans: Least-squares means. R package version 1.05-00. <http://CRAN.R-project.org/package=lsmeans>

**Macchiavelli RE sf** Agro 6600 Biometría Avanzada. Notas de clase. <http://academic.uprm.edu/zmacchia/agro6600/agro6600.pdf>

**Maindonald JH 2008** Using R for Data Analysis and Graphics: Introduction, Code and Commentary. Australian National University. <http://cran.r-project.org/doc/contrib/usingR.pdf>

**Mendiburu F 2012** agricolae: Statistical Procedures for Agricultural Research. R package version 1.1-2. <http://CRAN.R-project.org/package=agricolae>

**Mendoza H, Bautista G 2002** Diseño Experimental. Universidad Nacional de Colombia. Bogotá Colombia. Licencia Creative Commons BY-NC-ND. <http://www.virtual.unal.edu.co/cursos/ciencias/2000352/>

**Montgomery D 2009** Design and analysis of experiments. 7 edition. John Wiley and Sons, Inc., NY. ISBN: 978-0-470-12866-4. [http://www.knovel.com/web/portal/browse/display?\\_EXT\\_KNOVEL\\_DISPLAY\\_bookid=3168&VerticalID=0](http://www.knovel.com/web/portal/browse/display?_EXT_KNOVEL_DISPLAY_bookid=3168&VerticalID=0)

**Murison RD 2008** STAT404 - Mixed Models Lecture Notes. School of Mathematics, Statistics and Computer Science. University of New England. <http://mcs.une.edu.au/~stat504/StudyGuide/UNEnotes.pdf>

**Nonyane B 2004** Design and analysis for subjective assessment of visual and taste stimuli. Tesis de Doctorado. University of Edinburgh. 103p. <http://www.maths.ed.ac.uk/pg/thesis/nonyane.pdf>

**Obispo NE, Espinoza Y, Gil JL 2004** El diseño cruzado: un diseño para la experimentación con vacas lecheras. Zootecnia Tropical. Maracay. 22(4) 371-385. [http://www.scielo.org.ve/scielo.php?pid=S0798-72692004000400007&script=sci\\_arttext](http://www.scielo.org.ve/scielo.php?pid=S0798-72692004000400007&script=sci_arttext)

**Parsad R, Srivastava R, Gupta VK 2008** Design and analysis of agricultural experiments. Módulo 5. Change-over designs. Autor. Varghese C. Library Avenue. New Delhi. <http://www.iasri.res.in/iasriwebsite/DESIGNOFEXPAPPLICATION/Electronic-Book/module5/12changeover.pdf>

**Peña S 1994** Estadística Modelos y Métodos. 2. Modelos lineales y series temporales. Editorial Alianza, SA. ISBN:978-84-206-8110-8.

**Provete DB, da Silva FR, Souza TG 2011** Estatística aplicada à ecologia usando o R. Posgraduacao em biologia animal. Universidade Estadual Paulista. Sao Jose do Rio Preto. [http://cran.r-project.org/doc/contrib/Provete-Estatistica\\_aplicada.pdf](http://cran.r-project.org/doc/contrib/Provete-Estatistica_aplicada.pdf)

**R Core Team 2012** R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org/>

**Revelle W 2012** Psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.2.8. <http://personality-project.org/r/>

**S.A. 2012** The personality project: Using R for psychological research. A simple guide to an elegant package. Versión 4. [http://personality-project.org/r/r\\_guide.html](http://personality-project.org/r/r_guide.html)

**S.A. sf** Cross over studies. Introduction and explanation. [http://www.stattools.net/Cover\\_Exp.php](http://www.stattools.net/Cover_Exp.php)

**S.A. sf** Crossover designs and latin squares. <http://stat.ethz.ch/education/semesters/as2010/anova/Slid9.pdf>

**S.A. sf** Working with orthogonal contrasts in R. <http://wwwuser.gwdg.de/~cscherb1/content/Statistics%20Course%20files/Working%20with%20orthogonal%20contrasts%20in%20R.pdf>

**Schlesselman J 1971** Power families: a note on the Box and Cox Transformation. Journal of the Royal Statistical Society. Serie B. 33(2) 307-311.

**Snedecor GW, Cochran WG 1994** Statistical Methods, Octava Edición, Iowa State University Press. ISBN:0-8138-1561-4.

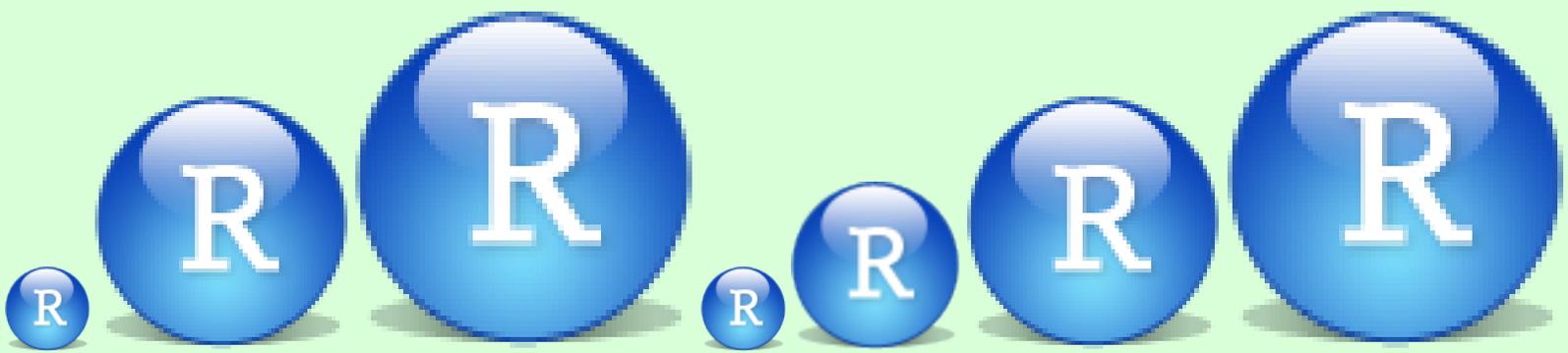
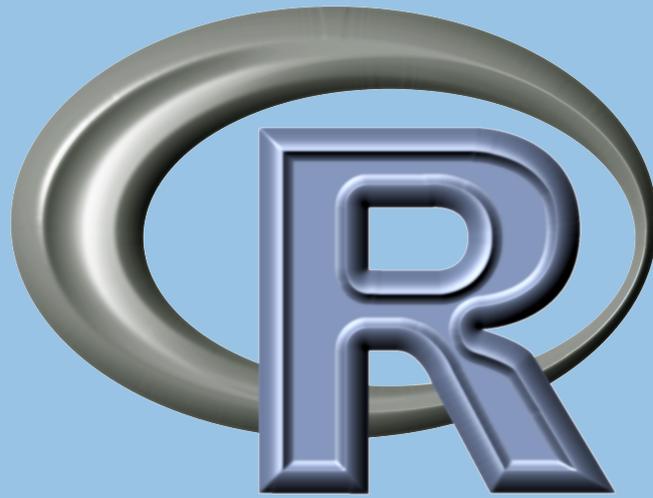
**Steel RGD, Torrie JH 1988** Bioestadística: Principios y Procedimientos. Segunda Edición, McGraw-Hill. Interamericana de México. Naucalpan de Juárez. México.

**Tukey JW 1957** On the Comparative Anatomy of Transformations. Annals of Mathematical Statistics. 28: 602-632.

**Vikneswaran 2005** An R companion to “Experimental Design”. [http://cran.r-project.org/doc/contrib/Vikneswaran-ED\\_companion.pdf](http://cran.r-project.org/doc/contrib/Vikneswaran-ED_companion.pdf)

**Warnes GR 2012** Includes R source code and/or documentation contributed by Bolker B, Lumley T, Johnson RC. Contributions from Randall C. Johnson are Copyright SAIC-Frederick, Inc. Funded by the Intramural Research Program, of the NIH, National Cancer Institute and Center for Cancer Research under NCI Contract NO1-CO-12400 (2012). gmodels: Various programming tools for model fitting. R package version 2.15.3. <http://CRAN.R-project.org/package=gmodels>

**Yaque PM sf** Análisis exploratorio de datos con R y Minitab. Universidad Complutense de Madrid. ISBN: 978-84-692-6255-9. <http://www.mat.ucm.es/~palomam/aed.pdf>



Mario Fernando Cerón-Muñoz



Luis Fernando Galeano Vasco



Luis Fernando Restrepo Betancour