

## Lección 7

# ANOVA básico

En esta lección explicamos cómo efectuar con R los análisis de la varianza básicos que se han estudiado en el curso. El tema central es el estudio de los aspectos técnicos del ANOVA con R y los tests posteriores de comparación de pares de medias; en la última sección presentamos además algunas instrucciones que permiten contrastar las codiciones necesarias sobre los datos para que un contraste ANOVA tenga sentido.

### 7.1. Los modelos del ANOVA en R

Los modelos a los que se aplica un ANOVA u otras muchas funciones, como por ejemplo la función `lm` para calcular la recta de regresión lineal, se especifican en R mediante *fórmulas*. El operador básico para construir una fórmula es la tilde `~`. Las fórmulas suelen tener la forma `Y~modelo`, donde la `Y` es un vector y el `modelo` es una combinación de vectores o factores que representa el modelo con el que queremos explicar el vector `Y` (en palabras técnicas, al que queremos *ajustar* los datos del vector `Y`). Por ejemplo, para calcular la recta de regresión por mínimos cuadrados de un vector `Y` respecto de un vector `X`, usamos `lm(Y~X)`. Esto significa que aplicamos la función `lm` a la fórmula `Y~X` que indica que queremos explicar `Y` como una función lineal de `X`.

En el ANOVA básico que se estudia en cursos introductorios de estadística inferencial, se usan cuatro tipos de fórmulas; si `X` es una variable numérica y `F1` y `F2` son dos factores de una cierta tabla de datos:

- La fórmula `X~F1` se usa para indicar el ANOVA de un factor, `F1`, de la variable `X`.
- La fórmula `X~F1+F2` se usa para indicar el ANOVA de dos factores, `F1` y `F2`, de la variable `X`, sin tener en cuenta la interacción entre los factores; es decir, suponiendo que sus efectos se acumulan, sin que haya interacción entre los mismos. Es el tipo de fórmula que se usa en los ANOVA de bloques.

- La fórmula  $X \sim F1 * F2$  se usa para indicar el ANOVA de dos factores, F1 y F2, de la variable X, teniendo en cuenta además la interacción entre estos factores.
- La fórmula  $X \sim F1:F2$  se usa para indicar el ANOVA de un factor que tiene como niveles los pares de niveles de F1 y F2.

La función básica de R para realizar un ANOVA es `aov`. Su sintaxis genérica es

```
aov(formula, data=...),
```

con los argumentos siguientes:

- **formula**: Una fórmula que especifique un modelo de ANOVA.
- **data**: Opcional, sirve para especificar, si es necesario, el *data frame* al que pertenecen las variables utilizadas en la fórmula.

Así, por ejemplo, si tenemos un *data frame* llamado DF, con una variable numérica X y un factor Fact, para realizar el ANOVA de la variable X respecto del factor Fact con la función `aov` podríamos entrar

```
aov(X ~ Fact, data=DF) o aov(DF$X ~ DF$Fact).
```

Otra posibilidad, que por ahora no usaremos pero sí más adelante, es aplicar la función `anova` (no confundirla con `aov`) al resultado de una función que calcule algún tipo de ajuste de una variable según un modelo; por ejemplo, una `lm`. La sintaxis sería entonces

```
anova(lm(formula, data=...)).
```

## 7.2. ANOVA de un factor

Para ilustrar el ANOVA de un factor con R utilizaremos un experimento en el que se quiso determinar si cuatro dietas concretas tenían alguna influencia en el tiempo de coagulación de la sangre en mamíferos.<sup>1</sup> Para ello se escogieron 24 animales, se repartieron de manera aleatoria en 4 grupos de 6 ejemplares cada uno, y a cada grupo se le asignó de manera aleatoria una de las 4 dietas bajo estudio. Al cabo de un cierto tiempo se midió el tiempo de coagulación de la sangre en estos animales. Los resultados (redondeados a enteros) se muestran en la Tabla 7.1. Observad que estamos ante un diseño experimental de un solo factor (la dieta) y de efectos fijos, ya que los niveles del factor (las dietas usadas) han sido escogidos por el experimentador.

Para contrastar si los tiempos medios de coagulación son los mismos para las cuatro dietas o no, vamos a realizar un ANOVA de estos datos. Para ello, lo primero que tenemos que hacer es recoger los datos en un *data frame*, formado por una variable numérica con los valores de la tabla y un factor de niveles las diferentes dietas, de manera que a cada valor

---

<sup>1</sup> Véase *Statistics for Experimenters* (2a edición), de G. P. Box, W. G. Hunter y J. S. Hunte (Wiley, 2005), p. 133.

Dieta			
A	B	C	D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
63	65	68	63
59	66	68	64

Tabla 7.1. Tabla con los tiempos de coagulación según la dieta.

en la variable numérica le corresponda la dieta con la que se obtuvo. Nosotros entraremos los datos por filas, y entonces el factor tendrá que ser

A, B, C, D, A, B, C, D, A ...

Otra opción sería entrar los datos por columnas, y entonces entraríamos como factor

A, A, A, A, A, A, B, B, B, ...

```
> coag=c(62,63,68,56,60,67,66,62,63,71,71,60,59,64,67,61,63,
65,68,63,59,66,68,64) #Los datos de la tabla, por filas
> diet=rep(c("A","B","C","D"), times=6) #6 copias de la fila A B
C D
> coagulacion=data.frame(coag,diet)
> str(coagulacion)
'data.frame': 24 obs. of 2 variables:
 $ coag: num 62 63 68 56 60 67 66 62 63 71 ...
 $ diet: Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3 4 1 2
...
> head(coagulacion)
  coag diet
1   62    A
2   63    B
3   68    C
4   56    D
5   60    A
6   67    B
```

Para analizar la igualdad de los tiempos medios de coagulación bajo las cuatro dietas, realizaremos un ANOVA de la variable `coag` separándola según (*ajustándola a*) el factor `diet`. Antes de empezar, es conveniente visualizar los datos para hacernos una idea de su distribución; por ejemplo, por medio de un diagrama de cajas con una caja por cada nivel:

```
> boxplot(coag~diet, data=coagulacion)
```

(Observad que hemos empleado una fórmula para indicar que queremos los diagramas de cajas de la variable `coag` separada por la variable `diet` del *data frame* `coagulacion`.) Obtenemos la Figura 7.1, donde vemos que las medias para las dietas *A*, *B* y *C* tienen el aspecto de ser diferentes dos a dos, y en cambio seguramente no podremos rechazar que las de las dietas *A* y *D* sean iguales. Por lo tanto, el resultado que esperamos del ANOVA es que nos permita rechazar la hipótesis nula de que los cuatro tiempos medios de coagulación son iguales.

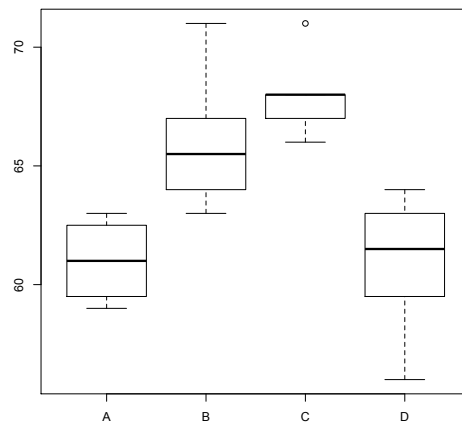


Figura 7.1. Diagrama de cajas de los tiempos de coagulación según las diferentes dietas.

Como hemos comentado, para realizar el ANOVA deseado, entramos:

```
> aov(coag~diet, data=coagulacion)
Call:
  aov(formula = coag ~ diet, data = coagulacion)

Terms:
              diet Residuals
Sum of Squares    228      112
Deg. of Freedom     3       20

Residual standard error: 2.366432
Estimated effects may be unbalanced
```

El resultado no es la tabla del ANOVA. Para obtenerla, hay que aplicar `summary` al resultado de `aov`:

```
> summary(aov(coag~diet, data=coagulacion))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)			
diet	3	228	76.0	13.571	4.658e-05 ***			
Residuals	20	112	5.6					
---								
Signif. codes:	0	'***'	0.001	'**'	0.01	'*' 0.05	'.' 0.1	' ' 1

El resultado de esta función `summary` es la tabla ANOVA usual:

- En la primera columna, dos etiquetas: el nombre del factor, en este caso `diet`, y `Residuals`, que representa los errores o residuos del ANOVA.
- La segunda columna, etiquetada `Df`, nos da los grados de libertad correspondientes al factor (su número de niveles menos 1) y a los residuos (el número de individuos en la tabla, menos el número de niveles del factor).
- La tercera columna, `Sum Sq`, nos muestra las sumas de los cuadrados del factor,  $SS_{Tr}$ , y de los residuos,  $SS_E$ .
- La cuarta columna, `Mean Sq`, contiene las medias de los cuadrados del factor,  $MS_{Tr}$ , y de los residuos,  $MS_E$ .
- La quinta columna, `F value`, nos da el valor del estadístico de contraste.
- En la sexta columna, `Pr(>F)`, aparece el p-valor del contraste.
- La séptima columna, sin etiqueta, indica el nivel de significación del p-valor según el código usual, explicado en la última línea del resultado. A mayor número de asteriscos, más significativo es el p-valor y por lo tanto hay más evidencia para rechazar la hipótesis nula de que las medias comparadas son iguales.

En nuestro caso, hemos obtenido un p-valor del orden de  $5 \cdot 10^{-5}$ . Esto nos permite rechazar la hipótesis nula y concluir que no todos los tiempos medios de coagulación para las diferentes dietas consideradas son iguales, como esperábamos. Recordad que esto no significa que todos los tiempos medios de coagulación sean diferentes, solo que hay pares de dietas que dan tiempos medios diferentes. Si ahora queremos determinar de qué pares de dietas se trata, tendremos que realizar algún test de comparaciones de pares de medias: véase la Sección 7.6.

Por razones de ahorrar espacio vertical, en lo que queda de lección vamos a eliminar los símbolos que marcan los niveles de significación de los p-valores. Para ello entramos la siguiente instrucción:

```
> options(show.signif.stars=FALSE)
```

A partir de ahora, y mientras no cerremos la sesión, estos símbolos no aparecerán más en las tablas ANOVA.

```
> summary(aov(coag~diet, data=coagulacion))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diet	3	228	76.0	13.571	4.658e-05
Residuals	20	112	5.6		

Si en algún momento de la sesión queréis volver a ver los códigos de significación, basta que entréis

```
options(show.signif.stars=TRUE)
```

Como hemos comentado, una manera alternativa de realizar un ANOVA es mediante `anova(lm(...))`. Con esta construcción obtenemos directamente la tabla, sin necesidad de aplicar `summary`.

```
> anova(lm(coag~diet, data=coagulacion))
```

Analysis of Variance Table

Response: coag

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diet	3	228	76.0	13.571	4.658e-05
Residuals	20	112	5.6		

Para extraer los datos de la tabla ANOVA obtenida con la función `summary(aov(...))`, y así poder operar directamente con ellos, hay que añadirle un sufijo adecuado.

```
> tabla=summary(aov(coag~diet, data=coagulacion))
> str(tabla)
```

List of 1

\$ :Classes 'anova' and 'data.frame': 2 obs. of 5 variables:

```
..$ Df      : num [1:2] 3 20
..$ Sum Sq  : num [1:2] 228 112
..$ Mean Sq: num [1:2] 76 5.6
..$ F value: num [1:2] 13.6 NA
..$ Pr(>F)  : num [1:2] 4.66e-05 NA
- attr(*, "class")= chr [1:2] "summary.aov" "listof"
```

```
> tabla[[1]]$"Sum Sq"
[1] 228 112
> tabla[[1]]$"Mean Sq"
[1] 76.0 5.6
> tabla[[1]]$"F value"
[1] 13.57143 NA
> tabla[[1]]$"Pr(>F)"
[1] 4.658471e-05 NA
> tabla[[1]]$"Pr(>F)" [1]
[1] 4.658471e-05
```

El resultado de `str(tabla)` nos indica que `summary(aov(...))` produce una *list* formada por un solo objeto (*List of 1*), un *data frame*, y de aquí que para extraer sus valores tengamos que usar la construcción mostrada: el sufijo `[[1]]` extrae el único objeto de la *list*, el *data frame*, y a continuación el sufijo `$columna` extrae el contenido de la columna de la tabla ANOVA: por ejemplo, la columna de las sumas de los cuadrados es `tabla[[1]]$"Sum Sq"`, y, como el p-valor es el primer elemento de la columna `Pr(>F)`, lo obtenemos con `tabla[[1]]$"Pr(>F)"[1]`. Fijaos en que, como los nombres de estas columnas contienen espacios en blanco u otros símbolos no aceptados en los nombres de objetos de R (recordad la Lección 2 del primer volumen), hay que especificarlos entre comillas.

El resultado de `anova(lm(...))` es un *data frame*, por lo que, para extraer los valores de la tabla ANOVA que produce, podemos usar la sintaxis usual de los *data frames*.

```
> tabla2=anova(lm(coag~diet, data=coagulacion))
> str(tabla2)
Classes 'anova' and 'data.frame': 2 obs. of 5 variables:
 $ Df      : int  3 20
 $ Sum Sq : num  228 112
 $ Mean Sq: num   76 5.6
 $ F value: num   13.6 NA
 $ Pr(>F)  : num  4.66e-05 NA
 - attr(*, "heading")= chr  "Analysis of Variance Table\n" "
   Response: coag"
> tabla2$"F value"
[1] 13.57143      NA
> tabla2$"Pr(>F)"[1]      #El p-valor
[1] 4.658471e-05
```

Veamos otro ejemplo de ANOVA de un factor.

**Ejemplo 7.1.** En un experimento, se estudió el efecto de seis dietas sobre el crecimiento de crías de conejos domésticos.<sup>2</sup> Los datos obtenidos están recogidos en la tabla de datos `rabbit` del paquete `faraway`.

```
> #Instalamos y cargamos el paquete faraway
...
> str(rabbit)
'data.frame': 30 obs. of 3 variables:
 $ treat: Factor w/ 6 levels "a","b","c","d",...: 6 2 3 3 1 2 3 6
   4 1 ...
 $ gain : num  42.2 32.6 35.2 40.9 40.1 38.1 34.6 34.3 37.5 44.9
   ...
 $ block: Factor w/ 10 levels "b1","b10","b2",...: 1 1 1 3 3 3 4
   4 4 5 ...
> help(rabbit)
```

---

<sup>2</sup> Véase *Experimental Design and Analysis*, de M. Lentner y T. Bishop (Valley Book Co. 1986), p. 428.

La variable `treat` indica la dieta, sus niveles son  $a, b, \dots, f$ , y la variable `gain` indica el aumento de peso; la variable `block`, que indica la camada, es irrelevante en este análisis concreto.

Si dibujamos el diagrama de cajas de los crecimientos para cada dieta,

```
> boxplot(gain~treat, data=rabbit)
```

obtenemos la Figura 7.2, donde no observamos grandes diferencias en los crecimientos medios.

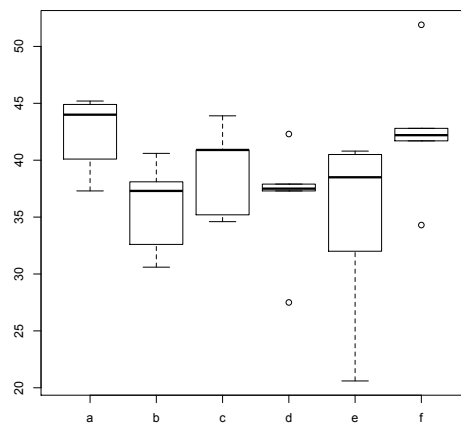


Figura 7.2. Diagrama de cajas de los crecimientos según las diferentes dietas.

Para determinar si hay diferencia en los aumentos medios de peso bajo las seis dietas, realizaremos un ANOVA de la variable `gain` ajustándola al factor `treat`.

```
> summary(aov(gain~treat, data=rabbit))
      Df Sum Sq Mean Sq F value Pr(>F)
treat    5  293.4    58.68   1.886  0.134
Residuals 24  746.5    31.10
```

El p-valor es 0.134, lo que indica que, efectivamente, no hay evidencia que permita afirmar que los aumentos medios no son todos iguales. Si hubiéramos querido obtener sólo el p-valor, podríamos haber entrado:

```
> summary(aov(gain~treat, data=rabbit))[[1]]$"Pr(>F)"[1]
[1] 0.1342124
```



Al usar las instrucciones `aov` o `anova(lm(...))` para realizar un ANOVA, se ha de emplear un factor (o varios, en las próximas secciones) para separar la variable numérica en subpoblaciones. Veamos un ejemplo de lo que pasa si nos descuidamos en este punto.

**Ejemplo 7.2.** En un experimento, se estudió el efecto de la vitamina C en el crecimiento de los dientes.<sup>3</sup> Se trató, en diferentes momentos, a cada ejemplar de un grupo de 10 cobayas con todas las combinaciones consideradas en el experimento de dosis de vitamina C (0.5, 1 o 2 mg) y métodos de suministro (mediante zumo de naranja o como ácido ascórbico) durante 6 semanas, y se cuantificó el crecimiento de sus dientes durante dicho período (más en concreto, se midió la longitud media de sus odontoblastos al final del mismo). El resultado es una tabla de 60 datos, que aparecen recogidos en el fichero `ToothGrowth` del paquete `UsingR`.

```
> Instalamos y cargamos el paquete "UsingR"
...
> str(ToothGrowth)
'data.frame': 60 obs. of 3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
> help(ToothGrowth)
```

La variable `len` contiene la longitud media final de los odontoblastos del animal, la variable `supp` el método de suministro (OJ indica zumo de naranja, *orange juice*, y VC indica ácido ascórbico puro, *vitamine C*), y la variable `dose` la dosis.

Nos vamos a olvidar por el momento de la variable `supp`, y vamos a contrastar si la dosis de vitamina C tiene influencia en el crecimiento de los dientes. Para ello realizamos un ANOVA de un factor.

```
> summary(aov(len~dose, data=ToothGrowth))
          Df Sum Sq Mean Sq F value    Pr(>F)
dose       1   2224   2224.3    105.1 1.23e-14
Residuals 58   1228     21.2
```

¿Veis algo raro? Hemos comentado que se usaron tres dosis, por lo que el número de grados de libertad en la fila del factor, `dose`, tendría que ser 2, y no 1. ¿Qué ha pasado? Muy sencillo: `dose` es una variable numérica, y para usar `aov` el factor ha de ser eso, un factor. Así que lo primero que vamos a hacer es convertir esta variable en un factor.

```
> ToothGrowth2=ToothGrowth #Realizamos un duplicado de la tabla
original
> ToothGrowth2$dose=as.factor(ToothGrowth2$dose)
> str(ToothGrowth2)
'data.frame': 60 obs. of 3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
```

<sup>3</sup> Véase *The Statistics of Bioassay* de C. I. Bliss (Academic Press, 1952), p. 499–501.

```

$ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
$ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1
...
> summary(aov(len~dose, data=ToothGrowth2))
              Df Sum Sq Mean Sq F value    Pr(>F)
dose           2   2426    1213    67.42 9.53e-16
Residuals     57   1026      18

```

Ahora está bien. El p-valor prácticamente 0 nos permite rechazar la hipótesis nula y concluir que diferentes dosis de vitamina C dan lugar a diferentes crecimientos medios de los odontoblastos.

### 7.3. ANOVA de bloques completos aleatorios

Para ilustrar este tipo de ANOVA, analizaremos un experimento sobre producción de penicilina.<sup>4</sup> En dicho experimento, se evaluaron cuatro procesos diferentes para determinar si había diferencias en su efectividad. Los cuatro procesos usaban una técnica de cultivo sumergido y empleaban agua de macerado de maíz como fuente de nitrógeno orgánico. Como la composición de este líquido puede afectar la producción final de penicilina, para evaluar los cuatro procesos se prepararon 5 mezclas diferentes de agua de macerado de maíz, de cada mezcla se tomaron 4 muestras y se asignaron de manera aleatoria a los cuatro procesos de producción. Los resultados obtenidos son los de la Tabla 7.2, y aparecen recogidos en la tabla de datos `penicillin` del paquete `faraway`.

Mezcla	Proceso de prod.			
	A	B	C	D
1	89	88	97	94
2	84	77	92	79
3	81	87	87	85
4	87	92	89	84
5	79	81	80	88

Tabla 7.2. Tabla de la producción de penicilina según el proceso de producción y la mezcla.

Observad que estamos ante un diseño experimental de bloques completos aleatorios. Los bloques son las mezclas de agua de macerado de maíz, los tratamientos (los procesos de producción) se han asignado de manera aleatoria a las unidades experimentales (las muestras) de cada bloque, de tal manera que cada bloque contiene exactamente una unidad experimental para cada tratamiento. Este es un ejemplo paradigmático de uso de bloques: para evitar la influencia de la variable «extraña» dada por la composición del agua de macerado de maíz, que puede influir en la producción, se escogen al azar unas composiciones y se prueban todos los procesos de manera independiente sobre cada composición.

<sup>4</sup> Véase *Practical Regression and Anova using R*, de J. Faraway, p. 186. Este texto se puede descargar de la página web [cran.r-project.org/doc/contrib/Faraway-PRA.pdf](http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf).

Demos un vistazo a esta tabla de datos.

```
> str(penicillin)
'data.frame': 20 obs. of 3 variables:
 $ treat: Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3 4 1
 2 ...
 $ blend: Factor w/ 5 levels "Blend1","Blend2",...: 1 1 1 1 2 2 2
 2 3 3 ...
 $ yield: num 89 88 97 94 84 77 92 79 81 87 ...
> head(penicillin)
  treat blend yield
1     A Blend1   89
2     B Blend1   88
3     C Blend1   97
4     D Blend1   94
5     A Blend2   84
6     B Blend2   77
> help(penicillin)
```

La variable `treat` contiene el proceso de producción, con valores A, B, C y D; la variable `blend` contiene la mezcla usada (los bloques), con valores `Blend1`, `Blend2`, ..., `Blend5`; y la variable numérica `yield` contiene un valor que cuantifica la producción de penicilina.

Veamos cómo son los diagramas de cajas de la producción de penicilina separada por procesos de producción y por mezclas.

```
> boxplot(yield~treat, data=penicillin)
> boxplot(yield~blend, data=penicillin)
```

Obtenemos los gráficos de la Figura 7.3. Vemos que no hay mucha diferencia entre las producciones para los diferentes procesos (sin tener en cuenta los bloques), y que sí que hay algunas diferencias en las producciones según la composición del agua de macerado de maíz; ya habíamos comentado que es bien sabido que su composición influye en la producción.

Si realizamos un ANOVA de un factor para cada uno de estos dos factores por separado, obtenemos resultados consistentes con esta observación visual:

```
> summary(aov(yield~treat, data=penicillin))
          Df Sum Sq Mean Sq F value Pr(>F)
treat      3     70    23.33   0.762  0.532
Residuals 16    490    30.62
> summary(aov(yield~blend, data=penicillin))
          Df Sum Sq Mean Sq F value Pr(>F)
blend      4    264    66.00   3.345  0.038
Residuals 15    296    19.73
```

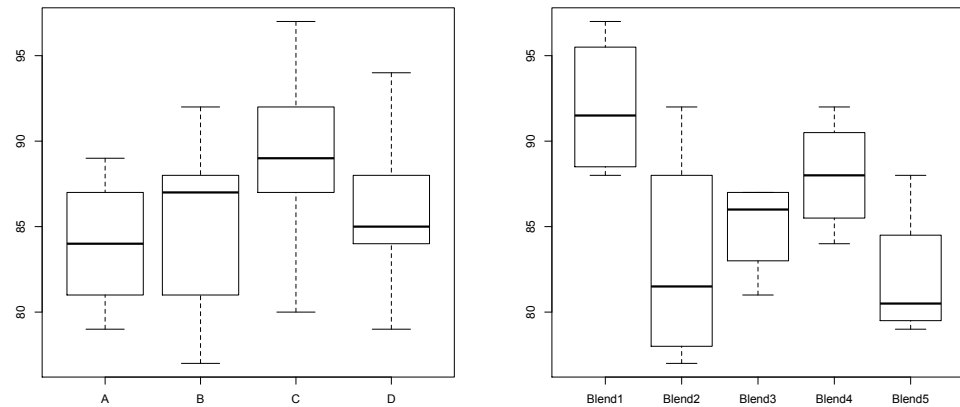


Figura 7.3. Diagramas de cajas de las producciones de penicilina según los diferentes métodos (izquierda) y según las diferentes composiciones del agua de macerado de maíz (derecha).

El p-valor del ANOVA separando por procesos de producción es 0.532, lo que indica que no podemos rechazar la hipótesis nula de que los procesos de producción tengan igual productividad media (sin tener en cuenta las mezclas), y el p-valor del ANOVA separando por mezclas es 0.038, lo que, a un nivel de significación del 5 %, nos permite rechazar que todas las mezclas produzcan la misma cantidad media de penicilina. Pero precisamente, el hecho de que la mezcla influya en la producción es lo que hace el primer ANOVA no fiable: a lo mejor el efecto de las mezclas enmascara las diferencias en las productividades de los procesos bajo estudio. Para determinarlo, vamos a realizar un ANOVA de bloques, es decir, de dos factores y efectos acumulados.

```
> summary(aov(yield~treat+blend, data=penicillin))
              Df Sum Sq Mean Sq F value Pr(>F)
treat          3      70    23.33   1.239  0.3387
blend          4     264    66.00   3.504  0.0407
Residuals     12     226    18.83
```

El p-valor que nos interesa en esta tabla es el de la fila **treat**: es 0.3387, por lo que en este experimento de bloques tampoco detectamos evidencia de que haya diferencias en la productividad media de los procesos estudiados. El segundo p-valor de la tabla, 0.0407, es el del ANOVA de bloques que resulta de intercambiar los bloques y el tratamiento, tomando las mezclas como el factor a analizar, y los procesos de producción como los bloques.

```
> summary(aov(yield~blend+treat, data=penicillin))
```

	Df	Sum Sq	Mean Sq	F	value	Pr(>F)
blend	4	264	66.00	3.504	0.0407	
treat	3	70	23.33	1.239	0.3387	
Residuals	12	226	18.83			

Podemos extraer los resultados de una tabla de un ANOVA de bloques generada por `summary(aov(...))` añadiendo los mismos sufijos que en el caso de un factor:

```
> tabla=summary(aov(yield~treat+blend, data=penicillin))
> tabla[[1]]$"Mean Sq"
[1] 23.33333 66.00000 18.83333
> tabla[[1]]$"Pr(>F)"
[1] 0.33865812 0.04074617 NA
> tabla[[1]]$"Pr(>F)"[1]
[1] 0.3386581
```

Veamos otro ejemplo de ANOVA de bloques completos aleatorios.

**Ejemplo 7.3.** Para estudiar si las diferentes actividades de evaluación llevadas a cabo en la asignatura de Matemáticas I tienen una dificultad similar (o mejor dicho, para mirar de confirmar nuestras sospechas de que no es así), escogimos una muestra aleatoria de 15 estudiantes de Biología o Bioquímica que se hubieran presentado al control 2 de dicha asignatura en el curso 2012-2013, y anotamos las notas obtenidas por estos estudiantes en los apartados de Tests, Talleres, Ejercicios de Casa, Control 1 y Control 2. Si obtenemos evidencia de que no todas las medias de las notas de las diferentes actividades fueron iguales, podremos concluir que no todas las actividades tienen la misma dificultad. Los datos obtenidos son los de la Tabla 7.3.

Como podemos observar, se trata de un experimento de bloques completos aleatorios: hemos escogido de manera aleatoria unos bloques (los estudiantes) y para cada estudiante hemos apuntado el valor de cada uno de los niveles del factor a estudiar (las notas en las diferentes actividades). Este diseño es el adecuado para este problema, puesto que hay una gran variabilidad en las notas obtenidas por estudiantes diferentes, desde matrículas a suspensos. Al tomar bloques, es decir, al considerar todas las notas de un conjunto aleatorio fijo de estudiantes, eliminamos el efecto de esta variabilidad.

Vamos a construir un *data frame* con estos datos. Este *data frame* tendrá tres variables: **notas**, con las notas obtenidas por los estudiantes; **acts**, con los tipos de actividades de evaluación realizados; y **bloques**, con el indicador de cada estudiante. Entraremos las notas siguiendo las filas de la tabla anterior, y por lo tanto tenemos que construir estos dos factores entrando sus niveles en el orden adecuado: el factor **acts** ha de estar formado por 15 copias de la fila de actividades, y el factor **bloques** ha de estar formado por 5 copias de 1, 5 copias de 2, etc. Y recordad que **bloques** ha de ser un factor, puesto que queremos usarlo en un ANOVA.

```
> notas=c(48,42,85,31,20,94,86,100,52,48,98,94,93,93,90,60,58,
71,66,46,52,56,79,64,24,78,73,84,80,95,84,84,94,83,70,83,76,
```

Estudiante	Actividad				
	Tests	Casa	Talleres	Control 1	Control 2
1	48	42	85	31	20
2	94	86	100	52	48
3	98	94	93	93	90
4	60	58	71	66	46
5	52	56	79	64	24
6	77	73	84	80	95
7	84	84	94	83	70
8	83	76	99	51	55
9	75	55	70	85	48
10	24	49	57	19	55
11	47	47	64	44	35
12	53	42	78	50	8
13	82	93	92	66	30
14	66	62	74	24	56
15	49	40	74	35	21

Tabla 7.3. Tabla de las notas obtenidas por 15 estudiantes en las diferentes actividades de evaluación.

```

99,51,55,75,55,70,85,48,24,49,57,19,55,47,47,64,44,35,53,42,
78,50,8,82,93,92,66,30,66,62,74,24,56,49,40,74,35,21)
> acts=rep(c("Tests","Casa","Talleres","Control1","Control2"),
times=15)
> bloques=as.factor(rep(1:15,each=5)) #Los bloques han de ser un
factor
> notas.bloques=data.frame(notas,acts,bloques)
> str(notas.bloques)
'data.frame': 75 obs. of 3 variables:
 $ notas : num 48 42 85 31 20 94 86 100 52 48 ...
 $ acts : Factor w/ 5 levels "Casa","Control1",...: 5 1 4 2 3 5
 1 4 2 3 ...
 $ bloques: Factor w/ 15 levels "1","2","3","4",...: 1 1 1 1 1 2
 2 2 2 2 ...
> head(notas.bloques)
  notas      acts bloques
1    48    Tests      1
2    42     Casa      1
3    85 Talleres      1
4    31 Control1      1
5    20 Control2      1
6    94    Tests      2

```

Vamos a dibujar los diagramas de cajas de las notas de las diferentes actividades y de los diferentes estudiantes:

```
> boxplot(notas.bloques$notas~notas.bloques$acts)
> boxplot(notas.bloques$notas~notas.bloques$bloques)
```

Obtenemos los gráficos de la Figura 7.4. Se observan diferencias en las notas de algunas actividades: por ejemplo, las notas del control 2 son muy inferiores a las de los talleres. Asimismo, como nos temíamos, vemos una gran variabilidad en las notas de los estudiantes.

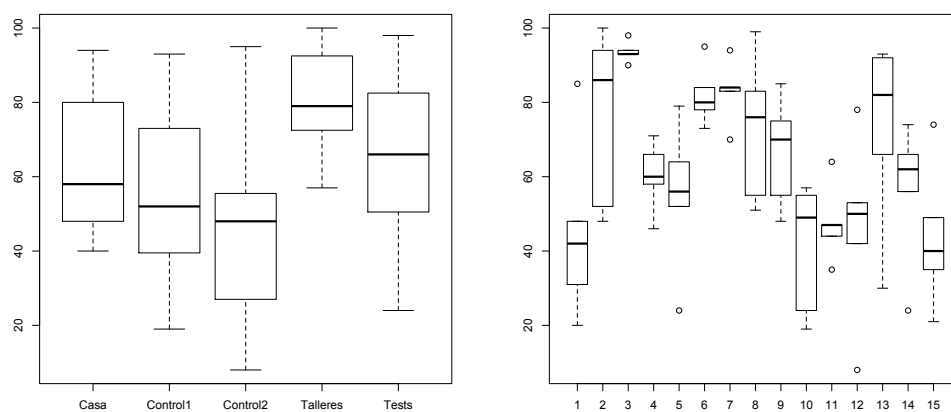


Figura 7.4. Diagramas de cajas de las notas según las diferentes actividades (izquierda) y según los diferentes estudiantes (derecha).

Realicemos ahora el ANOVA.

```
> summary(aov(notas~acts+bloques, data=notas.bloques))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
acts	4	9646	2411.5	13.133	1.31e-07
bloques	14	19430	1387.9	7.558	1.47e-08
Residuals	56	10283	183.6		

El p-valor que nos interesa es el de la primera fila, etiquetada con el factor cuyos niveles queremos comparar, en este caso **acts**. Este p-valor es muy pequeño, del orden de  $10^{-7}$ , lo que es evidencia de que, como nos temíamos, no todas las notas medias de las actividades de evaluación fueron iguales.

## 7.4. ANOVA de dos vías

Para ilustrar el ANOVA de dos vías, es decir, de dos factores con interacción, analizaremos los resultados de un experimento sobre el efecto de venenos y antídotos.<sup>5</sup> En este experimento se usaron tres venenos y cuatro antídotos, y cada combinación de veneno y antídoto se administró a cuatro ratas elegidas de manera aleatoria e independiente. A continuación, se anotó el tiempo de supervivencia de cada animal, en unidades de 10 horas. Se trata, pues, de un diseño experimental de dos factores, en el que cada individuo ha sido asignado al azar a cada nivel de los dos factores. El objetivo del experimento era contrastar la igualdad de los tiempos medios de supervivencia según el veneno, según el antídoto, y según la combinación veneno-antídoto. Los datos de este experimento están contenidos en la tabla `rats` del paquete `faraway`. Vamos a explorarlos.

Veneno	Antídoto			
	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

Tabla 7.4. Tabla de tiempos de supervivencia según el veneno y el antídoto.

```
> str(rats)
'data.frame': 48 obs. of 3 variables:
 $ time : num 0.31 0.82 0.43 0.45 0.45 1.1 0.45 0.71 0.46 0.88
 ...
 $ poison: Factor w/ 3 levels "I","II","III": 1 1 1 1 1 1 1 1 1 1
 1 ...
 $ treat : Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3 4 1
 2 ...
> head(rats)
  time poison treat
1 0.31      I      A
2 0.82      I      B
```

<sup>5</sup> Véase «An analysis of transformations», de G. Box y D. Cox, *J. Roy. Stat. Soc. Series B* 26 (1964), pp. 211–252. Véase también *Practical Regression and Anova using R*, de J. Faraway, p. 182.



```

3 0.43      I      C
4 0.45      I      D
5 0.45      I      A
6 1.10      I      B
> help(rats)

```

La variable `treat` contiene el antídoto, con valores A, B, C y D; la variable `poison` contiene el veneno, con valores I, II y III; y la variable numérica `time` contiene el tiempo de supervivencia de las ratas. Veamos cómo son los diagramas de cajas de estos tiempos, separados por venenos y por antídotos.

```

> boxplot(time~poison, data=rats)
> boxplot(time~treat, data=rats)

```

Obtenemos los gráficos de la Figura 7.5. Parece que hay diferencias en los tiempos de supervivencia tanto para los diferentes venenos como para los diferentes antídotos. Podemos también dibujar un diagrama de cajas de los tiempos de supervivencia separándolos por combinaciones de veneno y antídoto. La instrucción para hacerlo es la siguiente (observad la fórmula del argumento) y el resultado es la Figura 7.6.

```

> boxplot(time~poison:treat, data=rats)

```

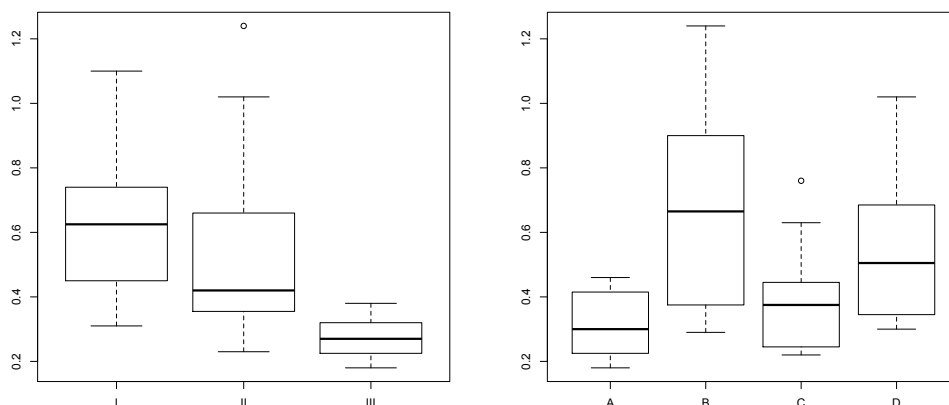


Figura 7.5. Diagramas de cajas de los tiempos de supervivencia según los diferentes venenos (izquierda) y según los diferentes antídotos (derecha).

El código para efectuar el ANOVA de dos vías del tiempo de supervivencia de las ratas bajo los efectos combinados de veneno y antídoto es el siguiente:

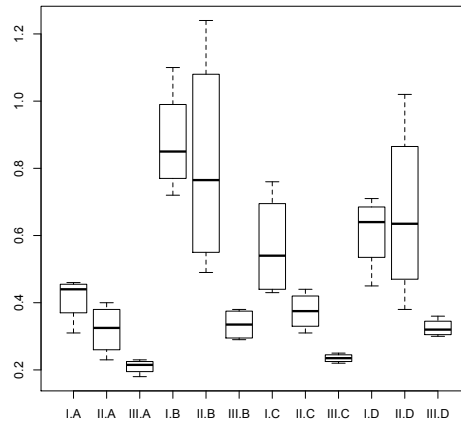


Figura 7.6. Diagramas de cajas de tiempos de supervivencia según diferentes combinaciones de veneno y antídoto.

```
> summary(aov(time~poison*treat, data=rats))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
poison	2	1.0330	0.5165	23.222	3.33e-07
treat	3	0.9212	0.3071	13.806	3.78e-06
poison:treat	6	0.2501	0.0417	1.874	0.112
Residuals	36	0.8007	0.0222		

Obtenemos la tabla de un ANOVA de dos vías. En la última columna, el primer p-valor es el del contraste de la igualdad de tiempos medios de supervivencia según el veneno (**poison**): su valor,  $3.3 \cdot 10^{-7}$ , indica que estos tiempos medios no son todos iguales. El segundo p-valor es el del contraste de la igualdad de tiempos medios de supervivencia según el antídoto (**treat**): que valga  $3.8 \cdot 10^{-6}$  indica que estos tiempos medios tampoco son todos iguales. Finalmente, el tercer p-valor es el del contraste de la interacción entre veneno y antídoto (**poison:treat**): como vale 0.112, no obtenemos evidencia de dicha interacción.

Observaréis que la tabla del ANOVA que produce R con esta fórmula no contiene la fila que permite contrastar si hay diferencia entre las medias de las poblaciones definidas por combinaciones de niveles, uno por cada factor: en este ejemplo, por las combinaciones de veneno y antídoto. Si se desea obtener esta fila, basta realizar un ANOVA de un factor que combine los dos factors del experimento.

```
> summary(aov(time~poison:treat, data=rats))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
poison:treat	11	2.2044	0.20040	9.01	1.99e-07

```
Residuals      36  0.8007  0.02224
```

Vemos que también hay diferencias significativas entre las combinaciones veneno—antídoto.

En el caso del ANOVA de dos vías, es interesante dibujar el *gráfico de interacción* entre factores. En un gráfico de interacción del factor  $F_1$  respecto del factor  $F_2$  en el ANOVA de una variable numérica  $X$  respecto de estos factores, se dibuja una línea quebrada para cada nivel del factor  $F_1$ . Esta línea une, mediante segmentos, los valores medios que toma la variable  $X$  para cada nivel de  $F_2$  en el nivel de  $F_1$  correspondiente. Si no hay ninguna interacción entre estos factores, las líneas resultantes serán paralelas. Cuanto más se alejen de ser paralelas, más evidencia de interacción habrá entre estos dos factores.

Estos gráficos de interacción se dibujan en R con la instrucción `interaction.plot` aplicada, por este orden, a los dos factores y la variable numérica. Así, pues, para obtener el gráfico de interacción de la variable antídoto respecto de la variable veneno en el ANOVA anterior, entraríamos

```
> interaction.plot(rats$treat, rats$poison, rats$time)
```

y obtendríamos el gráfico de la izquierda en la Figura 7.7. En él observamos que los tres niveles del veneno producen líneas casi paralelas, aunque se observa una ligera interacción: la pendiente de la recta correspondiente al veneno II entre los valores medios de C y D es mucho mayor que la de las rectas correspondientes a los otros dos venenos.

Este diagrama no ha quedado muy bonito, ya que las variables aparecen en él con sus nombres completos, y los factores no aparecen en la leyenda ordenados alfabéticamente. Una posibilidad para mejorarlo es cambiar las etiquetas de los ejes y la leyenda (la etiqueta del factor en la leyenda se define con `trace.label`) y reordenar los factores en dicha leyenda (con `fixed=TRUE`). Ya que estamos, dibujaremos algo más gruesas las líneas.

```
> interaction.plot(rats$treat, rats$poison, rats$time,
  xlab="Antídoto", ylab="Tiempo medio de supervivencia",
  trace.label="Veneno", lwd=c(2,2,2), fixed=TRUE)
```

Obtenemos el gráfico de la derecha en la Figura 7.7.

Veamos otro ejemplo de ANOVA de dos vías.

**Ejemplo 7.4.** En el Ejemplo 7.2 hablábamos de un cierto experimento sobre el efecto de la vitamina C en el crecimiento de los dientes, y de la tabla de datos `ToothGrowth` del paquete `UsingR` que recoge los datos obtenidos en ese experimento. Ahora vamos a realizar un ANOVA de dos vías sobre esta tabla de datos para contrastar la influencia en dicho crecimiento de la dosis de vitamina C y del método de suministrarla, teniendo en cuenta que sus efectos pueden interaccionar. Recordad de aquel ejemplo que, si queremos llevar a cabo un ANOVA sobre esta tabla que involucre la variable `dose` que contiene la dosis, primero hay que convertir esta variable en un factor. Para ello, copiamos la tabla `ToothGrowth` en una nueva tabla `ToothGrowth2`, y en esta última modificamos dicha variable.

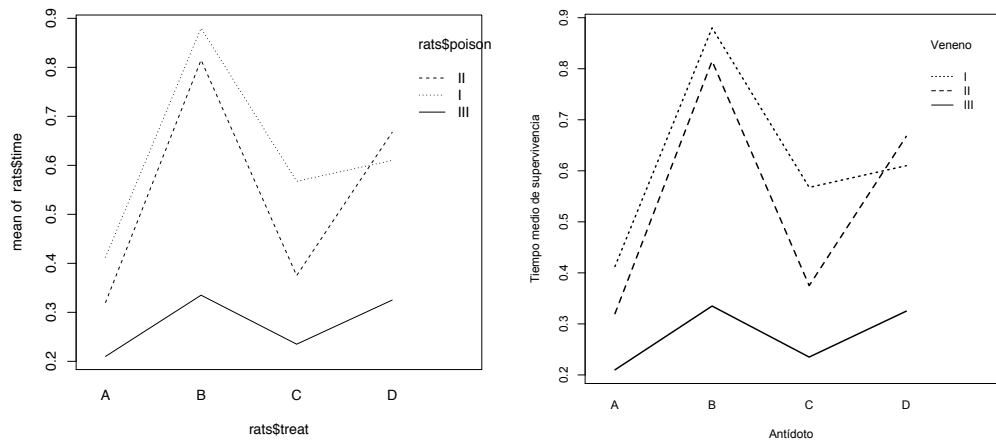


Figura 7.7. Gráficos de interacción del veneno y el antídoto.

```
> ToothGrowth2=ToothGrowth
> ToothGrowth2$dose=as.factor(ToothGrowth2$dose)
> str(ToothGrowth2)
'data.frame': 60 obs. of 3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

Demos ahora una ojeada a los diagramas de cajas de esta tabla, separando las longitudes de los odontoblastos por dosis, por suministro y por ambos.

```
> boxplot(len~dose, data=ToothGrowth2)
> boxplot(len~supp, data=ToothGrowth2)
> boxplot(len~dose:supp, data=ToothGrowth2)
```

Obtenemos los diagramas de la Figura 7.8. Se observan diferencias según la dosis y también según el método de suministro. ¿Serán significativas?

```
> summary(aov(len~supp*dose, data=ToothGrowth2))
          Df Sum Sq Mean Sq F value    Pr(>F)
supp      1  205.4    205.4   15.572 0.000231
dose      2 2426.4   1213.2   92.000 < 2e-16
supp:dose  2   108.3     54.2    4.107 0.021860
Residuals 54   712.1     13.2
> summary(aov(len~supp:dose, data=ToothGrowth2))
          Df Sum Sq Mean Sq F value    Pr(>F)
```

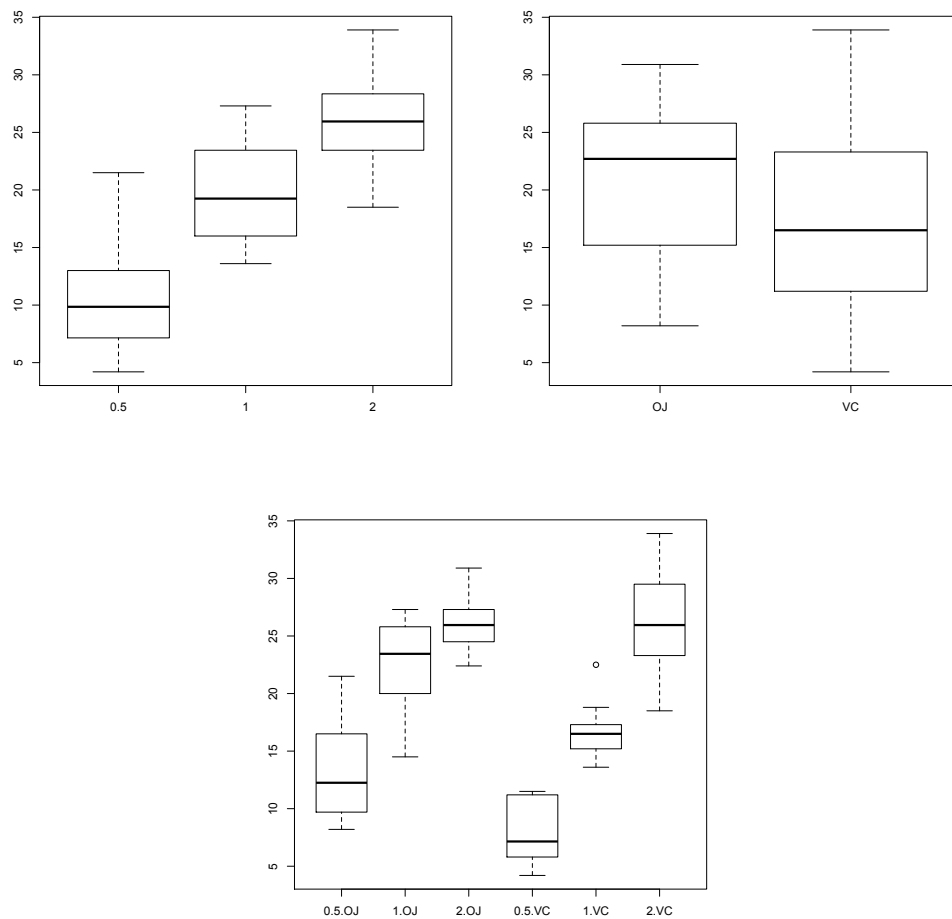


Figura 7.8. Diagramas de cajas del crecimiento de los dientes según la dosis de vitamina C (arriba, izquierda), según el método de suministro (arriba, derecha), y según la combinación de ambos (abajo).

```

supp:dose    5 2740.1    548.0    41.56 <2e-16
Residuals   54  712.1     13.2
> interaction.plot(ToothGrowth2$supp, ToothGrowth2$dose,
  ToothGrowth2$len, xlab="Método de suministro",
  ylab="Crecimiento medio", trace.label="Dosis", lwd=c(2,2,2),
  fixed=TRUE)

```

El resultado muestra que, efectivamente, hay diferencias significativas en el crecimiento de los dientes tanto según la dosis (p-valor prácticamente 0) como según el método de suministro (p-valor 0.000231) o la combinación de ambos (p-valor prácticamente 0), y también que hay evidencia significativa de interacción entre los dos factores (p-valor 0.02186). El gráfico de interacción producido con la última instrucción (Figura 7.9) refleja esta interacción: la línea correspondiente a la dosis de 2 mg no es paralela a las otras.

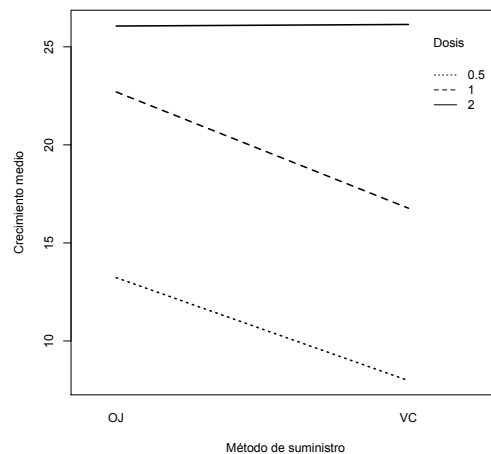


Figura 7.9. Gráfico de interacción de las dosis de vitamina C y los métodos de suministro.

El hecho de haber detectado interacción entre los dos factores hace que nos interese estudiar el efecto de los niveles de uno en el otro. Esto se puede llevar a cabo mediante varios ANOVA de un factor.

En primer lugar, vamos estudiar, para cada método de suministro de vitamina C, si hay diferencias entre los crecimientos medios de los odontoblastos según la dosis cuando la vitamina C se suministra mediante ese método. Para ello vamos a realizar dos ANOVA de un factor, la dosis, sobre dos *dataframes* extraídos de `ToothGrowth2`, uno con las filas donde `supp` toma el valor `OJ` y otro con las filas donde `supp` toma el valor `VC`. Vamos extraer estas subtablas de datos usando la instrucción `subset`. Recordad que la construcción

```
subset(df, condición)
```

define un *dataframe* con las filas del *data frame* `df` que cumplen la condición.

```
> TG.OJ=subset(ToothGrowth2, supp=="OJ")
> TG.VC=subset(ToothGrowth2, supp=="VC")
```

Ahora realizamos los dos ANOVA:

```
> summary(aov(len~dose, data=TG.OJ))
      Df Sum Sq Mean Sq F value    Pr(>F)
dose      2   885.3    442.6   31.44 8.89e-08
Residuals 27   380.1     14.1
> summary(aov(len~dose, data=TG.VC))
      Df Sum Sq Mean Sq F value    Pr(>F)
dose      2  1650    824.7   67.07 3.36e-11
Residuals 27   332     12.3
```

En ambos casos el p-valor es prácticamente nulo, lo que indica que para cada tipo de suministro de vitamina C, el crecimiento medio de los dientes varía con la dosis.

En este tipo de análisis, hay que tener en cuenta que si realizamos  $N$  contrastes independientes en un mismo experimento, cada uno de ellos con un nivel de significación  $\alpha$ , la probabilidad máxima de rechazar en alguno de ellos la hipótesis nula si en todos ellos es verdadera es

$$1 - (1 - \alpha)^N \approx N\alpha.$$

Por lo tanto, si queremos garantizar un nivel de significación global  $\alpha$ , hay que realizar cada contraste con nivel de significación  $\alpha/N$ .<sup>6</sup> En este ejemplo concreto, si hubiéramos deseado un nivel de significación global 0.05, deberíamos haber realizado cada ANOVA con un nivel de significación 0.025, lo que no afecta a las conclusiones, ya que los p-valores han sido realmente muy pequeños.

También podemos estudiar, para cada dosis, si hay diferencias entre las medias de crecimiento según el método de suministro. Como sólo hay dos métodos de suministro, para cada dosis tanto da realizar un ANOVA o un simple t-test suponiendo que las varianzas son iguales (ya que es una de las condiciones para poder realizar el ANOVA).

```
> TG.0.5=subset(ToothGrowth2, dose==0.5)
> summary(aov(len~supp, data=TG.0.5))
      Df Sum Sq Mean Sq F value    Pr(>F)
supp      1   137.8    137.81   10.05 0.0053
Residuals 18   246.9     13.72
> TG.1=subset(ToothGrowth2, dose==1)
> summary(aov(len~supp, data=TG.1))
      Df Sum Sq Mean Sq F value    Pr(>F)
```

<sup>6</sup> Para ser exactos, si llamamos  $\alpha_c$  al nivel de significación de cada contraste necesario para obtener un nivel de significación global  $\alpha$ , se cumple que  $1 - (1 - \alpha_c)^N = \alpha$ , de donde podemos despejar  $\alpha_c$  y obtenemos  $\alpha_c = 1 - \sqrt[N]{1 - \alpha}$ , que es un poco mayor que  $\alpha/N$ . Por ejemplo, para  $\alpha = 0.05$  y  $N = 20$  tenemos que  $\alpha_c = 0.03406$  mientras que  $\alpha/N = 0.025$ .

```

supp      1  175.8  175.82  16.26 0.000781
Residuals 18  194.6   10.81
> TG.2=subset(ToothGrowth2, dose==2)
> summary(aov(len~supp, data=TG.2))
              Df Sum Sq Mean Sq F value Pr(>F)
supp           1   0.03   0.032   0.002  0.964
Residuals     18 270.61  15.034

```

En este caso, para garantizar un nivel de significación global de 0.05, hay que realizar cada contraste con un nivel de significación de  $0.05/3 = 0.017$ . Obtenemos evidencia de que hay diferencia en el crecimiento medio de los odontoblastos según el tipo de suministro sólo para las dosis de 0.5 y 1 mg, que es donde hemos obtenido p-valores inferiores al nivel de significación.

Si la variable `dose` hubiera tenido, pongamos, veinte niveles en vez de tres, definir una subtabla para cada nivel y efectuar el ANOVA correspondiente hubiera sido muy largo. Para automatizar cálculos como estos, lo mejor es definir una función que calcule el p-valor del ANOVA de una variable respecto de un factor sobre la subtabla definida por un nivel variable «*x*» de un segundo factor, y luego aplicar esta función al vector de los niveles de este segundo factor. En este ejemplo concreto usaríamos el código siguiente.

```

> Anova_dosis=function(x){
  summary(aov(len~supp,
    data=subset(ToothGrowth2, dose==x)))[[1]]$"Pr(>F)"[1]
}
> round(sapply(levels(ToothGrowth2$dose), FUN=Anova_dosis),4)
  0.5      1      2
0.0053 0.0008 0.9637

```

## 7.5. Condiciones del ANOVA

Como sabemos, los resultados de un ANOVA nos permiten extraer conclusiones sólo si las poblaciones a las que se aplica cumplen una serie de condiciones; las dos básicas son la normalidad y la igualdad de varianzas. La condición de normalidad requiere que cada variable de cada factor defina una distribución normal, es decir, que cada muestra aleatoria simple de un tratamiento provenga de una población normal. La condición de igualdad de varianzas requiere que todas estas distribuciones tengan la misma varianza.

En la Lección 5 ya estudiamos algunos tests específicos de normalidad: los tests de Kolmogorov-Smirnov-Lilliefors, de Anderson-Darling, de Shapiro-Wilk y de D'Agostino-Pearson. Cada uno tiene sus ventajas y sus inconvenientes: el primero es muy popular, por ser una variante del test de Kolmogorov-Smirnov; el segundo, detecta mejor las discrepancias en los extremos; el test de Shapiro-Wilk es más potente para muestras grandes y menos sensible a repeticiones que los dos anteriores; y el test de D'Agostino-Pearson permite repeticiones de datos, pero no se puede usar con muestras de tamaño inferior a 20.



Veamos algunos ejemplos.

**Ejemplo 7.5.** Queremos contrastar si podemos aceptar que las cuatro dietas en el ejemplo de la coagulación de la Sección 7.2 provienen de distribuciones normales. Como hay algunas repeticiones de valores en las columnas, pero no llegan a los 20 valores cada una, usaremos el test de Shapiro-Wilk que, recordemos, está implementado en la función `shapiro.test`.

```
> coagA=coagulacion[coagulacion$diet=="A",]$coag
> shapiro.test(coagA)$p.value
[1] 0.1130298
> coagB=coagulacion[coagulacion$diet=="B",]$coag
> shapiro.test(coagB)$p.value
[1] 0.5227052
> coagC=coagulacion[coagulacion$diet=="C",]$coag
> shapiro.test(coagC)$p.value
[1] 0.2375366
> coagD=coagulacion[coagulacion$diet=="D",]$coag
> shapiro.test(coagD)$p.value
[1] 0.5227052
```

Los cuatro p-valores que obtenemos son grandes, mayores que 0.05, lo que quiere decir que en ninguno de los cuatro tests podemos rechazar la hipótesis nula: no hay evidencia de que los tiempos de coagulación bajo alguna de las cuatro dietas no sigan una ley normal, por lo que aceptamos que sí la siguen. Recordad además que, como hemos efectuado cuatro contrastes, para garantizar un nivel de significación global de 0.05 hemos de realizar cada uno con un nivel de significación de  $0.05/4 = 0.0125$ : es decir, que con p-valores por encima de este nivel de significación no podríamos rechazar ninguna hipótesis nula.

¿Cómo podríamos haber automatizado el cálculo de estos p-valores, para no tener que definir «a mano» cada vector de tiempos de coagulación? Con el código siguiente.

```
> Test.SW=function(x){round(shapiro.test(x)$p.value,4)}
> aggregate(coag~diet, data=coagulacion, FUN=Test.SW)
  diet  coag
1    A 0.1130
2    B 0.5227
3    C 0.2375
4    D 0.5227
```

La columna `coag` de este *data frame* contiene, para cada nivel de `diet`, el p-valor del test de Shapiro-Wilk para los tiempos de coagulación de la dieta correspondiente.

**Ejemplo 7.6.** Vamos a comprobar si el ANOVA de dos vías del ejemplo de venenos y antídotos explicado en la Sección 7.4 satisface las condiciones de normalidad. En los ANOVA de dos vías, se requiere que cada combinación de niveles de los dos factores bajo consideración defina una población normal; en este ejemplo, cada combinación de veneno y antídoto. Para determinarlo, deberíamos repetir 12 veces, una para cada combinación de veneno y antídoto, un par de instrucciones como las siguientes:

```
> time.I.A=rats[rats$poison=="I" & rats$treat=="A", ]$time
> shapiro.test(time.I.A)$p.value
[1] 0.07414486
```

Para evitarnos trabajo, vamos a usar la estrategia explicada al final del ejemplo anterior: aplicaremos la función `Test.SW` que acabamos de definir a la variable numérica `time` separándola por las combinaciones de veneno y antídoto.

```
> aggregate(time~poison+treat, data=rats, FUN=Test.SW)
  poison treat   time
1      I     A 0.0741
2     II     A 0.8476
3     III     A 0.5774
4      I     B 0.6998
5     II     B 0.7008
6     III     B 0.1706
7      I     C 0.4050
8     II     C 0.9209
9     III     C 0.9719
10     I     D 0.4274
11     II     D 0.9065
12     III     D 0.6889
```

Todos los p-valores que obtenemos son mayores que 0.05 (y de hecho, para garantizar un nivel de significación global de 0.05, deberíamos realizar cada test de Shapiro-Wilk con un nivel de significación  $0.05/12 = 0.0042$ ), por lo que no podemos rechazar que todas las combinaciones de veneno y antídoto definan unos tiempos de supervivencia que sigan leyes normales.

Pasemos al contraste de igualdad de varianzas (su nombre técnico es *contraste de homocedasticidad*), en el que usaremos el llamado *test de Bartlett*. Supongamos, para fijar ideas, que tenemos  $k$  vectores numéricos

$$x_i = (x_{i1}, x_{i2}, \dots, x_{in_i}), \quad i = 1, \dots, k.$$

Suponemos que cada uno de ellos es una muestra aleatoria simple de una variable aleatoria normal (tenemos que haberlo comprobado previamente) de varianza  $\sigma_i^2$ . El contraste que queremos realizar es:

$$\begin{cases} H_0 : \sigma_1^2 = \dots = \sigma_k^2 \\ H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ para algunos } i, j \end{cases}$$

Para realizar este contraste, se usa el *estadístico de Bartlett*:

$$K^2 = \frac{(N - k) \ln(\tilde{s}_p^2) - \sum_{i=1}^k (n_i - 1) \ln(\tilde{s}_i^2)}{1 + \frac{1}{3(k-1)} \left( \sum_{i=1}^k \left( \frac{1}{n_i - 1} \right) - \frac{1}{N - k} \right)},$$

donde

$$N = \sum_{i=1}^k n_i, \quad \tilde{s}_p^2 = \frac{\sum_{i=1}^k (n_i - 1) \tilde{s}_i^2}{N - k}, \quad \tilde{s}_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1}.$$

Este estadístico sigue aproximadamente una distribución  $\chi_{k-1}^2$  y, si la hipótesis nula es cierta, su valor esperado es pequeño (puesto que, en este caso, los valores esperados de  $\tilde{s}_1^2, \dots, \tilde{s}_k^2$  y  $\tilde{s}_p^2$  son todos iguales). El  $p$ -valor del contraste es entonces  $P(\chi_{k-1}^2 > K^2)$ .

R tiene una función que efectúa este test, `bartlett.test`, y puede aplicarse a una fórmula. Por ejemplo, para contrastar si las tres dietas en el ejemplo de la coagulación de la Sección 7.2 definen poblaciones con la misma varianza, podemos entrar

```
> bartlett.test(coag~diet, data=coagulacion)

Bartlett test of homogeneity of variances

data:  coag by diet
Bartlett's K-squared = 1.668, df = 3, p-value = 0.6441
```

Como el  $p$ -valor es alto, concluimos que las varianzas de los tiempos de coagulación bajo las tres dietas son la misma.

En los ANOVA de dos vías, se ha de comprobar la igualdad de varianzas para las combinaciones de niveles de los dos factores. En la fórmula que se entra al `bartlett.test` esta combinación no se puede especificar con `:`, como hemos hecho hasta ahora, sino con `interaction`. Por ejemplo, para contrastar por medio del test de Bartlett si las 12 combinaciones veneno–dieta en el ejemplo de venenos y antídotos explicado en la Sección 7.4 tienen la misma varianza, tenemos que entrar:

```
> bartlett.test(time~interaction(poison,treat), data=rats)

Bartlett test of homogeneity of variances

data:  time by interaction(poison, treat)
Bartlett's K-squared = 45.1369, df = 11, p-value = 4.59e-06
```

Obtenemos un  $p$ -valor de  $4.6 \cdot 10^{-6}$ , por lo tanto, hemos de rechazar la igualdad de varianzas: las conclusiones de ese ANOVA de dos vías no tienen en principio ninguna validez, puesto que no hay evidencia de que se cumpla una de las condiciones para realizarlo.

Otros tests de igualdad de varianzas son el de Fligner–Killeen, implementado en la función `fligner.test` (y que tiene la misma particularidad que `bartlett.test` a la hora de especificar combinaciones de factores) y el test de Levene, implementado en la función `leveneTest` del paquete `car` (donde las combinaciones de factores sí que se pueden especificar con `:`). En cada caso, el test produce un  $p$ -valor con el significado usual. El test de Levene es muy popular, porque no requiere que las poblaciones sean normales, pero esto en el contexto del ANOVA no tiene ningún interés; además, cuando las poblaciones son normales, el test de Bartlett es más potente que el de Levene.

## 7.6. Comparaciones de pares de medias

Si en un ANOVA hemos rechazado la hipótesis nula de la igualdad de todas las medias, puede que nos interese determinar qué medias son diferentes. Para ello podemos llevar a cabo algunos tests. En esta sección explicamos las instrucciones de R para tres de ellos.

### 7.6.1. t-tests por parejas

En los t-tests por parejas, realizamos un contraste de medias para cada par de niveles del factor, teniendo en cuenta que, para garantizar un nivel de significación global  $\alpha$ , se ha de ajustar el nivel de significación de cada test. Hay diversos métodos para llevar a cabo este ajuste: los más populares son el método de Bonferroni, el de Holm, que es más potente, y el de Hochberg, que, para contrastes ANOVA que no sean de bloques, aún tiene mayor potencia.

La función que efectúa t-tests por parejas es `pairwise.t.test`. Su sintaxis es

```
pairwise.t.test(var.numérica, factor, paired=..., p.adjust.method="...")
```

El valor de `p.adjust.method` es el método de ajuste de p-valores que deseamos usar, y se ha de entrar entrecomillado. El valor por defecto, que no es necesario especificar, es el de Holm. Para usar el de Bonferroni, hay que especificar `p.adjust.method="bonferroni"`, y para usar el de Hochberg, `p.adjust.method="hochberg"`. Para conocer qué otros métodos se pueden usar, sus ventajas e inconvenientes, consultad `help(p.adjust)`.

Por lo que refiere al parámetro `paired`, sirve para indicar si las muestras son independientes (igualándolo a `FALSE`, que es su valor por defecto) o emparejadas (igualándolo a `TRUE`); cuando el ANOVA es de bloques, las muestras son emparejadas, y por lo tanto se ha de especificar `paired=TRUE`.

Veamos algunos ejemplos de su uso.

**Ejemplo 7.7.** Para determinar cuáles de las tres dietas en el ejemplo de la coagulación de la Sección 7.2 dan tiempos medios de coagulación diferentes, vamos a efectuar un t-test de Bonferroni.

```
> pairwise.t.test(coagulacion$coag, coagulacion$diet,
  p.adjust.method="bonferroni")
      Pairwise comparisons using t tests with pooled SD
data:  coagulacion$coag and coagulacion$diet
      A      B      C
B 0.02282 -      -
C 0.00108 0.95266 -
D 1.00000 0.00518 0.00014

P value adjustment method: bonferroni
```

Los p-valores del resultado ya han sido ajustados, y por lo tanto se han de comparar directamente con el nivel de significación  $\alpha$  global. Por ejemplo, con un nivel de significación  $\alpha = 0.05$ , obtenemos evidencia de que los tiempos medios para los pares de dietas (A,B), (A,C), (B,D), (C,D) son diferentes, y en cambio no hay evidencia de que los tiempos medios para los pares (A,D) y (B,C) sean diferentes.

**Ejemplo 7.8.** Para determinar qué actividades de evaluación de las consideradas en el Ejemplo 7.3 dan notas medias diferentes, vamos a realizar un t-test por parejas de Holm. Primero vamos a comprobar que dichas actividades de evaluación definen poblaciones normales con la misma varianza, condición necesaria para que tanto el ANOVA como los t-tests por parejas tengan sentido.

Contrastaremos la normalidad de las notas para cada actividad de evaluación con el test de Shapiro-Wilk, y para no trabajar más de la cuenta, usaremos la técnica explicada en el Ejemplo 7.5.

```
> Test.SW=function(x){round(shapiro.test(x)$p.value,4)}
> aggregate(notas~acts, data=notas.bloques, FUN=Test.SW)
      acts  notas
1   Casa 0.1384
2 Control1 0.7054
3 Control2 0.5515
4 Talleres 0.7612
5   Tests 0.5984
```

Todos los p-valores son altos, los que nos permite aceptar que las notas de cada actividad provienen de una distribución normal. Contrastemos ahora la igualdad de varianzas.

```
> bartlett.test(notas~acts, data=notas.bloques)

Bartlett test of homogeneity of variances

data:  notas by acts
Bartlett's K-squared = 6.1529, df = 4, p-value = 0.188
```

El p-valor superior a 0.1 indica que podemos aceptar que las varianzas son todas iguales. Por lo tanto, podemos realizar el t-test de Holm. En este caso tenemos que usar `paired=TRUE`, puesto que se trata de un ANOVA de bloques.

```
> pairwise.t.test(notas.bloques$notas, notas.bloques$acts,
  paired=TRUE)

Pairwise comparisons using paired t tests

data:  notas.bloques$notas and notas.bloques$acts

      Casa    Control1 Control2 Talleres
Control1 0.4562 -          -          -
```

Control2	0.0271	0.4562	-	-
Talleres	0.0012	0.0038	0.0012	-
Tests	0.4562	0.1640	0.0271	0.0038

P value adjustment method: holm

Los p-valores superiores a 0.05 son los de los pares de actividades «Ejercicios de casa»–«Control 1», «Control 1»–«Control 2», «Ejercicios de casa»–«Tests» y «Control 1»–«Tests». Por lo tanto, a un nivel de significación global de 0.05, no podemos rechazar que las notas medias de estos pares de actividades sean iguales, mientras que en los otros tenemos evidencia de que son diferentes.

### 7.6.2. Test de Duncan

Otro método para comparar los pares de medias es el test de Duncan. Con R se puede efectuar con la instrucción `duncan.test` del paquete `agricolae`. Su sintaxis básica es

```
duncan.test(aov, "factor", alpha=..., group=...)$sufijo,
```

donde

- `aov` es el resultado de la función `aov` (sin `summary`) con la que hemos calculado el ANOVA de partida.
- El `factor` es el factor del ANOVA, y se ha de entrar entrecomillado y con el mismo nombre que se ha usado en la fórmula del `aov`.
- El parámetro `alpha` sirve para entrar el nivel de significación  $\alpha$ : por defecto vale 0.05.
- `group` puede valer `TRUE` o `FALSE`, y hace que el resultado se presente de forma diferente.
- El sufijo tiene que ser `groups` si `group=TRUE` y `comparison` si `group=FALSE`.

Este test no se puede usar en el ANOVA de bloques, puesto que supone que las muestras de los tratamientos son independientes.

**Ejemplo 7.9.** Vamos a usar el test de Duncan para determinar qué pares de dietas en el ejemplo de la coagulación de la Sección 7.2 tienen tiempos medios de coagulación diferentes, con un nivel de significación global del 5%. En primer lugar, guardamos en un objeto el ANOVA calculado con `aov`, y luego aplicamos la función `duncan.test` a este objeto. Aquí lo haremos dos veces, una para cada opción para `group`, para poder comparar y comentar cómo se muestran los resultados. No entraremos el parámetro `alpha`, puesto que usamos su valor por defecto.

```
> anova.coag=aov(coag~diet, data=coagulacion)
> #Instalamos y cargamos el paquete "agricolae", si no lo está
```

```

...
> duncan.test(anova.coag, "diet", group=FALSE)$comparison
      Difference    pvalue sig.      LCL      UCL
A - B          -5 0.002007  **   -7.937680 -2.0623197
A - C          -7 0.000104  ***  -10.083577 -3.9164227
A - D           0 1.000000      -2.937680  2.9376803
B - C          -2 0.170968      -4.937680  0.9376803
B - D           5 0.002722  **    1.916423  8.0835773
C - D           7 0.000132  ***    3.823702 10.1762981
> duncan.test(anova.coag, "diet", group=TRUE)$groups
  trt means M
1   C     68 a
2   B     66 a
3   A     61 b
4   D     61 b

```

Observemos los resultados:

- Con `group=FALSE` y sufijo `comparison`, obtenemos una tabla donde, para cada pareja de niveles del factor, se da, entre otros datos, un p-valor (la columna `p value`). Estos p-valores tienen el significado usual. En este ejemplo, de los p-valores se desprende que no podemos rechazar que las medias para los pares A–D y B–C sean iguales (los p-valores son grandes), mientras que, en cambio, hay evidencia significativa de que el resto de los pares de dietas tienen medias diferentes (los p-valores son pequeños).
- Con `group=TRUE` y sufijo `groups`, obtenemos una tabla donde se agrupan los niveles según la igualdad de medias. En este caso, asigna las dietas B y C al grupo `a` y las dietas A y D al grupo `b`: dietas en el mismo grupo tienen medias iguales, dietas en grupos diferentes tienen medias diferentes.

### 7.6.3. Método de Tukey

El *método HSD de Tukey* (las siglas HSD refieren a *Honestly Significant Difference*) es el método más preciso de comparación de parejas de medias para contrastes ANOVA de un factor en los que cada nivel tenga el mismo número de observaciones. Es un test similar al t-test, pero usa otro estadístico con una distribución diferente. Con R se efectúa aplicando la función `TukeyHSD` al resultado de `aov`.

Para explicar qué información obtenemos con esta función, veamos un ejemplo. Vamos a aplicar el método de Tukey al ANOVA del ejemplo sobre tiempos de coagulación de la Sección 7.2.

```

> anova.coag=aov(coag~diet, data=coagulacion)
> TukeyHSD(anova.coag)
Tukey multiple comparisons of means
 95% family-wise confidence level

```

```
Fit: aov(formula = coag ~ diet, data = coagulacion)

$diet
      diff      lwr      upr      p adj
B-A      5  0.7245544  9.275446 0.0183283
C-A      7  2.7245544 11.275446 0.0009577
D-A      0 -4.0560438  4.056044 1.0000000
C-B      2 -1.8240748  5.824075 0.4766005
D-B     -5 -8.5770944 -1.422906 0.0044114
D-C     -7 -10.5770944 -3.422906 0.0001268
```

De esta manera, para cada par de niveles obtenemos:

- La diferencia de sus medias muestrales, en la columna `diff`.
- Los extremos inferior y superior de un intervalo de confianza del 95% para la diferencia de medias poblacionales, en las columnas `lwr` y `upr`, respectivamente. El nivel de confianza se puede ajustar con el parámetro `conf.level` dentro de la función.
- El p-valor del contraste de igualdad de medias correspondiente, en la columna `p adj`.

En nuestro ejemplo, los intervalos de confianza para las diferencias de medias de los pares A-B, A-C, B-D y C-D no contienen el valor 0 (y los p-valores correspondientes son pequeños), lo que indica que podemos rechazar que estas medias sean iguales. En cambio, los intervalos de confianza para las diferencias de medias de los pares A-D y B-C sí que contienen el 0 (y los p-valores correspondientes son grandes), lo que indica que podemos aceptar que estas medias son iguales. En este caso obtenemos, por lo tanto, la misma conclusión que con el t-test de Bonferroni y el test de Duncan.

## 7.7. Guía rápida

- Fórmulas:
  - `X~F1` ajusta la variable `X` al factor `F1`.
  - `X~F1+F2` ajusta la variable `X` a los factores `F1` y `F2` sin tener en cuenta la interacción entre los factores.
  - `X~F1*F2` ajusta la variable `X` a los factores `F1` y `F2` teniendo en cuenta además la interacción entre los factores.
  - `X~F1:F2` ajusta la variable `X` a un factor que tiene como niveles los pares de niveles de `F1` y `F2`.
- `options(show.signif.stars=...)` sirve para desactivar (usando `FALSE`) o activar (con `TRUE`, su valor por defecto) los códigos de significación.



- `aov(formula)` efectúa el ANOVA indicado por la *formula*. El parámetro adicional *data* sirve para indicar el *data frame* del que se extraen los vectores y factores de la *formula*.
- `summary(aov(...))` muestra la tabla del ANOVA efectuado con la función `aov`. Con el sufijo `[[1]]$"Pr(>F)"`, nos da directamente la columna de p-valores de la tabla.
- `anova(lm(formula, data=...))` también produce la tabla del ANOVA indicado por la *formula*. Con el sufijo `$"Pr(>F)"`, nos da directamente la columna de p-valores de la tabla.
- `interaction.plot(factor1, factor2, vector)` produce el gráfico de interacción del ANOVA de dos vías del *vector* respecto de los dos factores.
- `subset(data frame, condición)` define un *dataframe* con las filas del *data frame* que cumplen la *condición*.
- `shapiro.test(x)$p.value` calcula el p-valor del test de normalidad de Shapiro-Wilk aplicado al vector numérico *x*.
- `aggregate` sirve para aplicar una función a una o varias variables de un *data frame* agrupando sus entradas por los niveles de uno o varios factores.
- `bartlett.test`, con la misma sintaxis que `aov`, efectúa el test de Bartlett de homocedasticidad.
- `pairwise.t.test(x, factor)` efectúa un t-test por parejas de la variable *x* separada por los niveles del *factor*. Dos parámetros importantes:
  - `paired`: se ha de igualar a `TRUE` en los ANOVA de bloques.
  - `p.adjust.method`: sirve para indicar el método de ajuste de los p-valores.
- `duncan.test(aov, "factor")$sufijo`, del paquete "`agricolae`", efectúa el test de Duncan del ANOVA *aov*. Dos parámetros importantes:
  - `alpha`: sirve para indicar el nivel de significación.
  - `group`: sirve para indicar cómo queremos que muestre el resultado.

El *sufijo* ha de ser `groups` si `group=TRUE` y `comparison` si `group=FALSE`.
- `TukeyHSD` efectúa el test HSD de Tukey.