Bioestadística (Master en Bioinformática)

Bloque: AMECP

Análisis de Modelos Estadísticos de Comparación y Predicción

Sesión: **Análisis de la varianza**

Manuel Franco
Dpto. Estadística e Investigación Operativa

Índice

1. Análisis de la varianza de un factor

El Análisis de la Varianza (ANOVA) es una técnica estadística que se encarga de determinar el comportamiento de una variable de interés en nuestro campo de estudio experimental a través sus observaciones o datos registrados en los distintos grupos disponibles en la población, analizando por término medio si se detectan o no diferencias entre los observaciones de cada grupo o nivel, es decir, el ANOVA consiste en un contraste de medias múltiple.

Para realizar este contraste, esta técnica introducida por Fisher, se basa en la descomposición de la variabilidad del experimento en diferentes causas independientes, de donde proviene el término de Análisis de la Varianza.

En este sentido, se llama ANOVA simple o ANOVA de un factor cuando la partición en grupos o niveles de la población bajo estudio viene dada por un solo factor, es decir, una variable cualitativa o categórica.

1.1. Test de análisis de la varianza

Para analizar el comportamiento de una variable Y en los distintos grupos/niveles del factor A, las observaciones de la muestra se agrupan identificando cada nivel del factor, es decir, y_{ij} es la j-ésima observación obtenida en el nivel i del factor A (formado por k niveles o tratamientos).

Así, Y_{ij} representa a la variable Y cada vez que se mide la j-ésima realización en el i-ésimo nivel de A, cuyo valor obtenido en la muestra observada ha sido y_{ij}.

En este contexto, la técnica del ANOVA se enfoca desde dos puntos de vista diferentes pero similares, bien a través del análisis del comportamiento medio en cada grupo, o bien a través del efecto de cada nivel o tratamiento sobre el comportamiento medio de la variable de interés; lo que conlleva a dos modelizaciones del problema.

Modelización I

$$Y_{ij} = E(Y_{ij}) + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}, \text{ para } i = 1, ..., k \text{ y } j = 1, ..., n_i,$$

siendo $E(Y_{ij}) = \mu_i$ para $j = 1, ..., n_i$ y $n = \sum_{i=1}^{k} n_i$.

Observación 1.1 *Denotando por Y_i la v.a. Y en el nivel i del factor A, $i = 1, ..., k$, las condiciones iniciales sobre $Y_1, ..., Y_k$ para el desarrollo del ANOVA son:*

- *Independencia (variables independientes)*
- *Homogeneidad de varianzas (igualdad de varianzas u homoscedasticidad)*
- *Normalidad (distribuciones normales)*

es decir, $Y_i \sim N(\mu_i, \sigma^2)$ e independientes, y asumiendo aleatoriedad en la realización del experimento o toma de observaciones en cada nivel.

Por tanto, para una m.a.s. de la variable Y en el nivel i, es decir, $Y_{i1}, ..., Y_{in_i}$ de Y_i, tenemos que:

$$Y_{ij} \sim N\left(\mu_i, \sigma^2\right) (\Leftrightarrow \varepsilon_{ij} \sim N(0, \sigma^2)) \text{ e independientes}, j = 1, ..., n_i$$

Interpretación gráfica de las condiciones

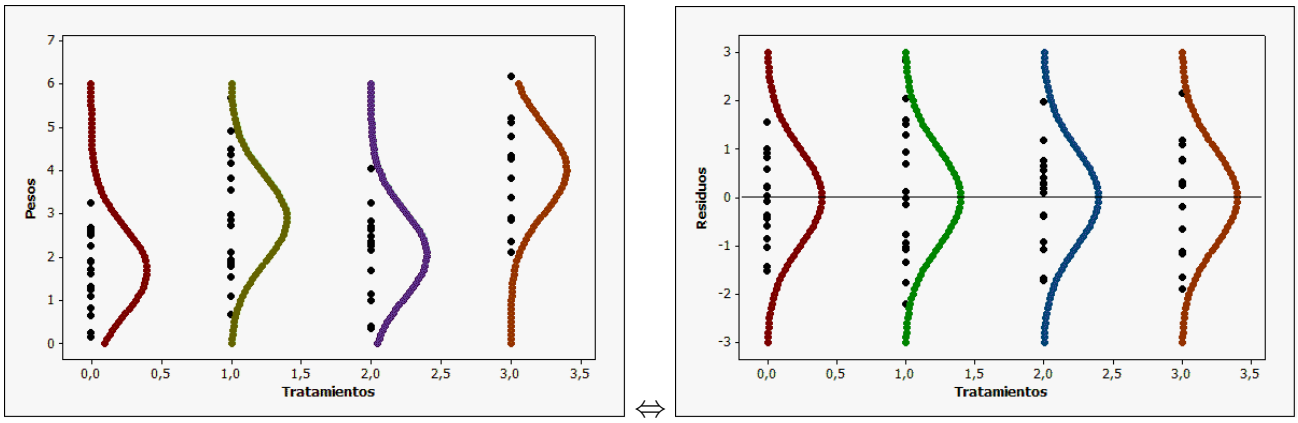
Para ilustrar la interpretación de las condiciones del ANOVA, consideramos el siguiente ejemplo.

Un grupo de investigadores de la Universidad del Estado de Pennsylvania han identificado un clon de álamo que produce árboles fuertes y de rápido crecimiento. Los clones son células genéticamente idénticas obtenidas del mismo individuo.

En una fase experimental de este estudio, plantaron el Clon de álamo 252 en dos sitios diferentes: uno en un terreno rico y húmedo cerca de una ensenada y el otro en un terreno seco y arenoso en una colina. Midieron el diámetro en centímetros, la altura en metros y el peso de la madera en seco en kilogramos de una muestra de árboles de tres años de edad. En un esfuerzo por maximizar el rendimiento del desarrollo de los árboles, incluyeron otro factor con los cuatro tipos de tratamientos utilizados, los cuales están determinados por el uso o no de fertilizante y por la aplicación o no de irrigación. La base *alamos* contiene los datos observados del peso ponderado junto con los niveles de estos factores.

En este caso, $Y_0 \sim N(\mu_0, \sigma^2)$ es el peso ponderado para el tratamiento 0 (no se aplica fertilizante ni irrigación), $Y_1 \sim N(\mu_1, \sigma^2)$ para el tratamiento 1 (sólo fertilizante), $Y_2 \sim N(\mu_2, \sigma^2)$ para el tratamiento

2 (solo irrigación), e $Y_3 \sim N(\mu_3, \sigma^2)$ en el tratamiento 3 (fertilizante e irrigación):



Objetivo del test ANOVA

En este contexto, el objetivo del ANOVA es contrastar si el valor medio de Y es el mismo en todos los niveles o tratamientos del factor A, es decir,

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_k \\ H_1 : \text{no todas las medias iguales} \end{cases}$$

Modelización II

$$Y_{ij} = E(Y_{ij}) + \varepsilon_{ij} = \mu + A_i + \varepsilon_{ij}, \text{ para } i = 1, ..., k \text{ y } j = 1, ..., n_i,$$

siendo A_i el efecto del nivel i del factor A sobre la media global $\mu = E(Y)$, e.d. $\mu_i = \mu + A_i$.

Objetivo del test ANOVA

Teniendo en cuenta la modelización 1, el mismo comportamiento de Y en los niveles del factor equivale a que los niveles del factor no afectan en el comportamiento provocando diferencias en la media de Y, es decir, los efectos del factor sobre cada nivel son los mismos.

Así, el objetivo del ANOVA es equivalente al anterior, pero en este contexto, el contraste se expresa como:

$$\begin{cases} H_0 : A_1 = A_2 = ... = A_k \\ H_1 : \text{no todos los efectos iguales} \end{cases}$$

Observación 1.2 *Si el factor es aleatorio, las observaciones corresponden a una muestra de niveles, por lo que H_0 se establece como la variación nula debida al factor, siendo el desarrollo del test similar.*

Terminología del diseño de experimentos

En el análisis de los modelos ANOVA resulta habitual el uso de términos específicos del diseño de experimentos que describen el tipo de factor y el diseño de obtención de las observaciones experimentales, algunos de ellos se indican a continuación.

Factor fijo: Si incluye todos los niveles del factor que son objeto de estudio.

Factor aleatorio: Si está formado por una muestra de niveles elegidos al azar.

Diseño completo: Si se ha muestreado en todos los niveles del factor.

Diseño balanceado: Si el número de observaciones es el mismo en todos los niveles.
Diseño no balanceado: Si se ha obtenido una muestra de tamaño diferente en algún nivel.
Bloque: Si su efecto sobre la variable no es directamente de interés, pero se considera en el modelo para controlar sus efectos (para comparaciones homogéneas y distinguir diferencias).
Factor anidado: Si el diseño contempla muestras de niveles del factor en cada nivel de un factor principal.
Factores cruzados: Si se contemplan todos los cruces entre sus niveles.

Desarrollo del test ANOVA

El test ANOVA se basa en la descomposición de la variación del experimento:

"Variación total = Variación entre grupos + Variación dentro de grupos"

dado que se verifica que

$$SS_T = SS_E + SS_D$$

donde la suma de cuadrados total representa la variabilidad de Y en el experimento, dada por:

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2$$

la suma de cuadrados dentro representa la componente de la variabilidad de Y dentro de cada grupo o nivel del factor, y la suma de cuadrados entre representa la componente de la variabilidad entre los distintos grupos o niveles del factor, ambas se expresan por:

$$SS_D = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2,\ SS_E = \sum_{i=1}^{k} n_i\left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2$$

siendo $\overline{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ la media muestral en el nivel i del factor, $\overline{Y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$ la media muestral total (con todos los niveles).

A partir de las condiciones iniciales, se derivan las siguientes distribuciones muestrales para el desarrollo de la tabla ANOVA y la estimación de la varianza poblacional:

- Para cada nivel i, $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2$ es la cuasivarianza muestral de $Y_i \sim N(\mu_i, \sigma^2)$, siendo $\frac{n_i-1}{\sigma^2}S_i^2 \sim \chi^2_{n_i-1}$
- Sumando las variaciones dentro de todos los niveles

  $$\frac{1}{\sigma^2}SS_D = \sum_{i=1}^{k}\frac{n_i-1}{\sigma^2}S_i^2 \sim \sum_{i=1}^{k}\chi^2_{n_i-1} = \chi^2_{n-k}$$

  por lo que, $MS_D =$ estimador insesgado para σ^2
- Bajo H_0, $S^2 = \frac{1}{n-1}SS_T$ es la cuasivarianza muestral de $Y \sim N(\mu, \sigma^2)$, siendo $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$
- Bajo H_0, para cada nivel i, $\overline{Y}_{i\bullet} \sim N\left(\mu, \frac{\sigma^2}{n_i}\right)$ independientes, por lo que $\frac{1}{\sigma^2}SS_E \sim \chi^2_{k-1}$

Presentación del contraste: Tabla ANOVA

Fuente	S.Cuadrados	G.Libertad	Medias Cuadrados	Estadístico	P-valor
Entre grupos:	SS_E	$k-1$	$MS_E = \frac{SS_E}{k-1}$	$F = \frac{MS_E}{MS_D}$	$P(F_{k-1,n-k} > F_{exp})$
Dentro:	SS_D	$n-k$	$MS_D = \frac{SS_D}{n-k}$		
Total:	SS_T	$n-1$			

en donde:

- El estadístico F, bajo H_0, sigue una distribución F-Snedecor de grados de libertad $(k-1, n-k)$, $F \sim F_{k-1,n-k}$ bajo H_0.
- La región crítica o de rechazo al nivel de significación α es

$$M_1^* = \{F > F_{k-1,n-k,\alpha}\}$$

- A partir de la muestra, el p-valor es $p = P\left(F_{k-1,n-k} > F_{\text{experimental}}\right)$

1.2. Diagnóstico del modelo ANOVA

En primer lugar, puede asumirse que partimos de un diseño de experimentos adecuado, es decir, que las observaciones experimentales se han realizado aleatoriamente dentro de cada grupo o nivel de la partición establecida por el factor en la población, y que no hay relación entre las muestras obtenidas en cada grupo dado que están tomadas de distintas subpoblaciones de individuos. No obstante, pueden aplicarse pruebas de aleatoriedad e independencia para garantizar la idoneidad del diseño empleado.

Sin embargo, las condiciones de normalidad e igualdad de varianza no están avaladas por el diseño experimental, debiéndose comprobar estas condiciones iniciales.

Test de normalidad

Hay diferentes tipos de contrastes de normalidad, es decir, que contrastan la hipótesis nula de que las observaciones provienen de una población normal, frente a la alternativa que la población no es normal:

$$\begin{cases} H_0 & : & Y_i \text{ son normales} \Leftrightarrow \varepsilon_i \text{ son normales}, i = 1, ..., k \\ H_1 & : & \text{no son normales} \end{cases}$$

Las técnicas gráficas, llamadas *Q-Q plot*, consisten en transformar el modelo normal en una recta representando la línea que correspondería a la normalidad de la muestra obtenida, junto con la correspondiente transformación de la distribución de la muestra, comparando su aproximación a la recta.

Las técnicas analíticas consisten en determinar un estadístico del contraste, por ejemplo los tests de Shapiro-Wilk o de Kolmogorov-Smirnov, y la región crítica correspondiente, tomando la decisión de rechazar o no la normalidad a través del p-valor.

Test de homogeneidad de varianzas: Homoscedasticidad

Los contrastes de homogeneidad de varianzas consisten en detectar si existen diferencias significativas entre las varianzas o las desviaciones típicas de las diferentes muestras (grupos o niveles del factor)

$$\begin{cases} H_0 & : & \sigma_1^2 = ... = \sigma_k^2 \\ H_1 & : & \text{no todas iguales} \end{cases}$$

Test de Bartlett Bajo la suposición de normalidad, el test de homogeneidad de varianzas se establece mediante una región crítica de una cola a partir del estadístico de Bartlett:

$$B = \frac{1}{C}\left((n-k)\ln MS_D - \sum_{i=1}^{k}(n_i-1)\ln S_i^2\right) \sim \chi^2_{k-1} \text{ (apróx.)}$$

donde $C = 1 + \frac{\frac{1}{n_1-1}+\ldots+\frac{1}{n_k-1}-\frac{1}{n-k}}{3(k-1)}$.

Test de Levene En caso de poblaciones no normales, el contraste de homoscedasticidad está determinado por una región crítica de una cola basado en el estadístico de Levene:

$$F = \frac{\frac{\sum_{i=1}^{k} n_i\left(\overline{Z}_{i\bullet}-\overline{Z}_{\bullet\bullet}\right)^2}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij}-\overline{Z}_{i\bullet}\right)^2}{n-k}} \sim F_{k-1,n-k},$$

donde $Z_{ij} = \mid Y_{ij} - \overline{Y}_{i\bullet} \mid, \overline{Z}_{i\bullet} = \sum_{j=1}^{n_i} Z_{ij}/n_i$ y $\overline{Z}_{\bullet\bullet} = \sum_{i=1}^{k} /n_i\overline{Z}_{i\bullet}n$.

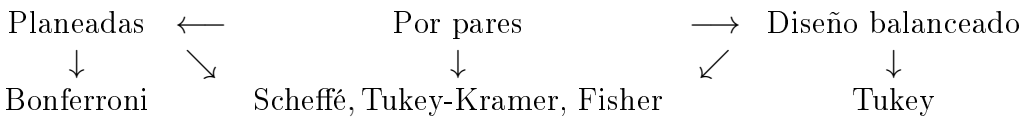
1.3. Comparaciones múltiples

Como hemos comentado, la técnica ANOVA es un contraste múltiple de medias cuyo objetivo es detectar si hay diferencias en el comportamiento medio de la variable de interés, o equivalentemente, si los niveles del factor en la población provocan efectos diferentes sobre dicho comportamiento de la variable.

En este contexto, cuando el ANOVA detecta que existen diferencias en el comportamiento medio de la variable de interés debidas a los efectos del factor, se plantea la cuestión de qué nivel o tratamiento provoca dichas diferencias, o entre qué subconjuntos de niveles hay o no diferencias en los efectos sobre la media de la variable objeto de estudio. Esta discusión se realiza a través de las comparaciones múltiples por pares.

Las comparaciones múltiples permiten detectar posibles diferencias entre los distintos niveles cuando el ANOVA detecta diferencias significativas, es decir, qué niveles provocan diferencias y qué conjuntos de niveles se pueden considerar homogéneos (influyen del mismo modo).

Las comparaciones se basan en la obtención de intervalos de confianza para las diferencias de medias entre niveles del factor, para detectar si tales intervalos contienen o no al cero como valor admisible, y pueden clasificarse de la siguiente forma:



Observar que existen diversas técnicas de comparación, entre las anteriores una de las más usuales en la práctica es la técnica de Tukey. También existen otras técnicas de comparaciones múltiples planeadas parciales, y por tanto que reducen la cantidad de pruebas simultáneas, entre otras, el método de Dunnett específico en situaciones de un grupo o tratamiento de control.

1.4. Caso práctico

Veamos un caso práctico de aplicación del ANOVA de un factor a través del programa R sobre un conjunto de datos utilizado por Sir Ronald Fisher, quien introdujo esta técnica estadística.

Ejemplo de análisis de la varianza de un factor

El conjunto de datos *Iris* es un conjunto de datos multivariantes relativos a la variación morfológica de flores de iris de tres especies relacionadas. El conjunto de datos consta de 50 muestras de cada una de las tres especies de Iris (setosa, virginica y versicolor). Para cada flor observada de cada muestra se midieron cuatro características: la longitud y la anchura de los sépalos y pétalos, en centímetros. El archivo **iris** contiene los datos observados en este experimento y están incluidos en el paquete base de R.

En este estudio, cabe preguntarse si alguna de las características morfológicas de las flores se ve afectada por la especie a la que pertenece, es decir, si hay diferencias entre estas especies de Iris. Por tanto, la aplicación del ANOVA permitirá analizar los comportamientos medios de estas medidas con respecto a los tres niveles del factor (especies de Iris). En particular, consideraremos la longitud de los sépalos como medida de interés en nuestro ejemplo.

Test ANOVA

Para analizar la longitud de los sépalos de las tres especies de Iris, primero cargamos el fichero y obtenemos un resumen descriptivo preliminar que nos proporciona una primera visión sobre las condiciones iniciales del ANOVA:

```
> data(iris)
> View(iris)
> attach(iris)
> tapply(Sepal.Length,Species,summary)
$setosa
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   4.800   5.000   5.006   5.200   5.800

$versicolor
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.900   5.600   5.900   5.936   6.300   7.000

$virginica
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.900   6.225   6.500   6.588   6.900   7.900
```

y siguientes gráficos descriptivos relativos a la dispersión de la longitud del sépalo en cada especie y la distribución de sus observaciones (Figura 1):

```
> plot(Sepal.Length ~ Species)
> stripchart(Sepal.Length ~ Species,method="stack")
```

Suponiendo que se satisfacen las condiciones de iniciales de normalidad y homoscedasticidad de la longitud de los sépalos en las tres especies de Iris, aplicamos el test ANOVA mediante la instrucción *aov*, cuyos argumentos es el modelo lineal formado por la variable de interés (longitud de los sépalos) y el factor (especies de Iris):

```
> sepalo <- aov(Sepal.Length ~ Species)
```

Los resultados del ANOVA para la longitud de los sépalos según las especies de Iris, es decir, el contraste de igualdad de longitud media en las tres especies, o equivalentemente, contraste de hipótesis nula que la especie de Iris no significa diferencia de longitud de los sépalos, son los siguientes:

```
> summary(sepalo)
             Df Sum Sq Mean Sq F value Pr(>F)
```

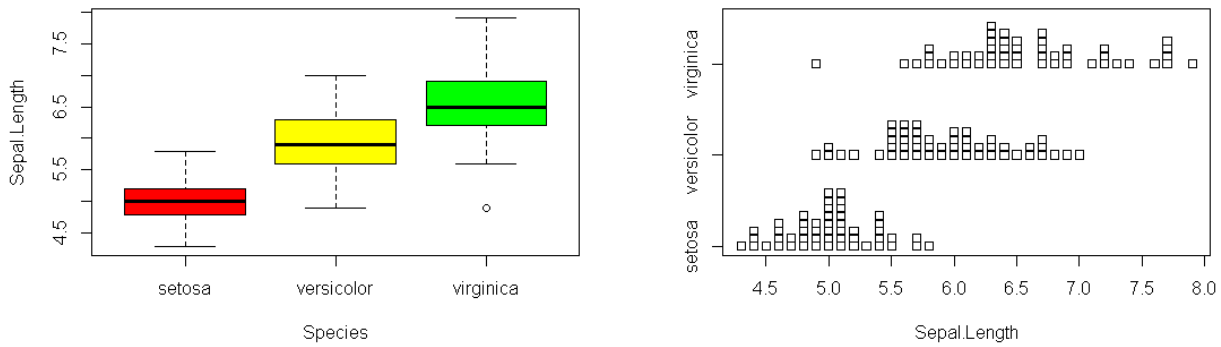


Figura 1: Boxplot y diagrama de caracteres de longitudes de sépalo por especie

```
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A partir del p-valor del test ANOVA, $P \simeq 0$, rechazamos la hipótesis nula de igualdad de medias de longitudes de los sépalos para las tres especies, siendo muy significativa la diferencia, dado que es prácticamente nula la probabilidad de que sea errónea la decisión de rechazar la igualdad.

Otra forma de realizar el ANOVA es mediante la función *lm* (linear model), que también utilizamos en la sección de análisis de regresión lineal. Si aplicamos esta opción, la tabla de resultados se obtiene con la función *anova*:

```
> forma2 <- lm(Sepal.Length ~ Species)
> anova(forma2)
Analysis of Variance Table

Response: Sepal.Length
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 63.212  31.606  119.26 < 2.2e-16 ***
Residuals 147 38.956   0.265
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

y como se observa ambas formas de ejecutar un ANOVA son iguales, así como los resultados que calcula en cada una de ellas y que puede comprobarse con la función *names* en cada uno de ellos:

```
> names(sepalo)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "contrasts"     "xlevels"
[11] "call"          "terms"         "model"
> names(forma2)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "contrasts"     "xlevels"
[11] "call"          "terms"         "model"
```

Por ejemplo, utilizando las función *model.tables* se muestran las estimaciones de los efectos sobre la longitud media de los sépalos, e incluyendo argumentos en esta función, entre otros, puede especificarse las estimaciones medias en cada especie de Iris y el error estándar de los efectos o niveles:

```
> model.tables(sepalo)
Tables of effects

 Species
Species
    setosa versicolor  virginica
   -0.8373     0.0927     0.7447

> model.tables(sepalo,type="means",se=TRUE)
Tables of means
Grand mean

5.843333

 Species
Species
    setosa versicolor  virginica
     5.006      5.936      6.588

Standard errors for differences of means
        Species
          0.103
replic.      50
```

como se observa en estos resultados, se ha utilizado el argumento `type="means"` en la función *model.tables* para indicar la tabla de estimaciones de las longitudes de medias en cada especie, aunque por defecto presenta la tabla de estimaciones de los efectos de cada nivel (especie) sobre la longitud media global, esto puede indicarse mediante el argumento `type="effects"`. El argumento `se=TRUE` incluye en el resultado el error estándar para las diferencias de las medias. Asimismo, la librería *gplots* de R, permite representar gráficamente las longitudes medias en cada especie mediante la función *plotmeans*:

```
> library(gplots)
> plotmeans(Sepal.Length~Species)
```

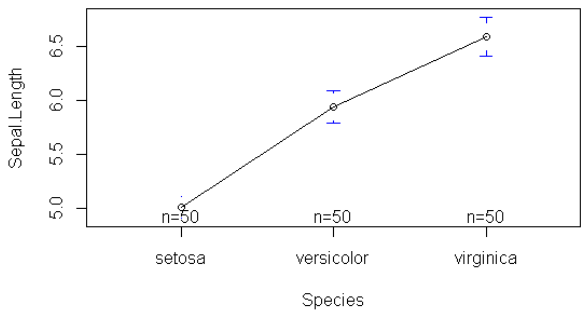


Figura 2: Gráfica de longitudes medias de sépalo por especie

Observar que el error cuadrático medio o estimación insesgada de la varianza del modelo lineal se encuentra en los resultados incluidos en la tabla ANOVA anterior. No obstante, también puede obtenerse de forma separada a través de la función *deviance* para la suma de cuadrados del error como para su media:

```
> deviance(sepalo)
[1] 38.9562
> deviance(sepalo)/sepalo$df.residual
[1] 0.2650082
```

Diagnóstico del modelo ANOVA

Para la validez de la conclusión del test ANOVA, es necesario comprobar las condiciones iniciales requeridas para su desarrollo.

En primer lugar, realizamos el análisis gráfico de los residuos que incluye por defecto el desarrollo del test ANOVA mostrado en la Figura 3:

```
> plot(sepalo)
```

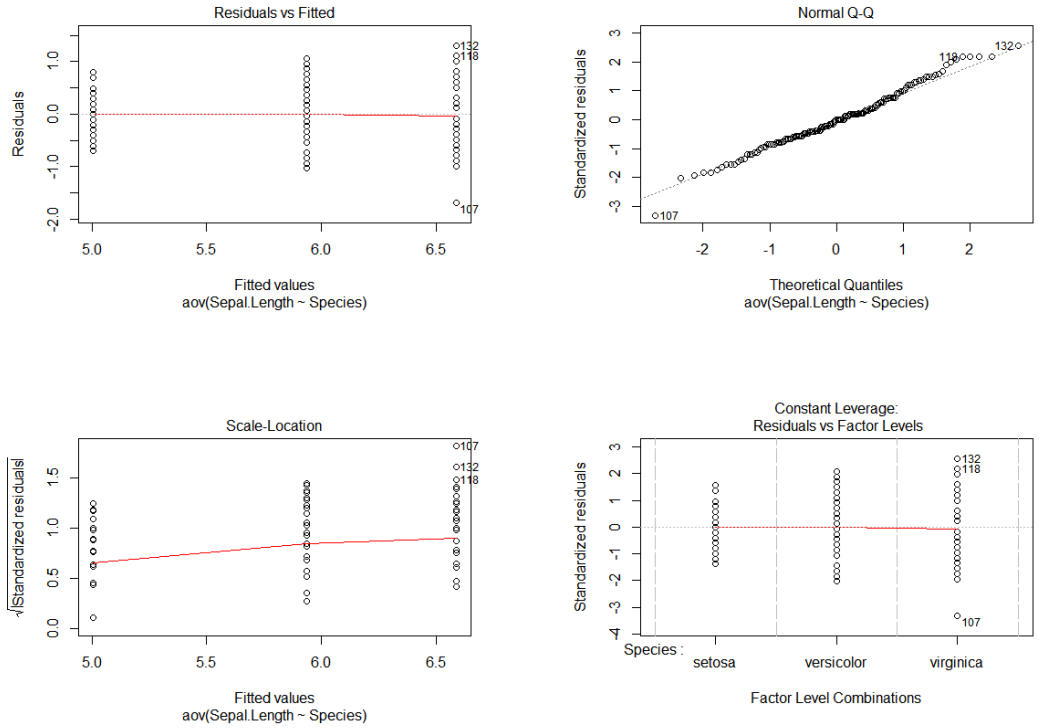


Figura 3: Gráficas de los residuos

así como el gráfico Boxplot para los residuos del ANOVA, observando que es equivalente al de la Figura 1 pero centrados sobre el eje de $x = 0$:

```
> plot(sepalo$model$Species,sepalo$residuals,main="Residuals vs.Species")
```

Observar que en los gráficos anteriores de los residuos, la aproximación a la normalidad se realiza para el total de la población, en la aplicación práctica del análisis de la varianza, la falta de información suficiente en cada grupo o nivel del factor provoca de forma habitual que se limite al diagnóstico de la normalidad de toda la población. No obstante, debería comprobarse la normalidad para cada nivel o tratamiento del factor, al igual que se estima la variabilidad en cada nivel y se contrasta su igualdad. En nuestro caso, estas gráficas de normalidad para cada especie de Iris, se obtienen como sigue:

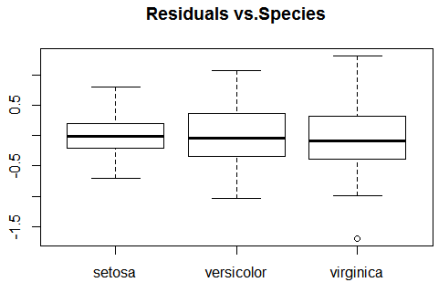


Figura 4: Gráfica Boxplot para los residuos

```
> qqnorm(sepalo$residuals[Species=="setosa"],
+ main="Normal Q-Q Plot: setosa")
> qqline(sepalo$residuals[Species=="setosa"])
> qqnorm(sepalo$residuals[Species=="versicolor"],
+ main="Normal Q-Q Plot: versicolor")
> qqline(sepalo$residuals[Species=="versicolor"])
> qqnorm(sepalo$residuals[Species=="versicolor"],
+ main="Normal Q-Q Plot: virginica")
> qqline(sepalo$residuals[Species=="versicolor"])
```

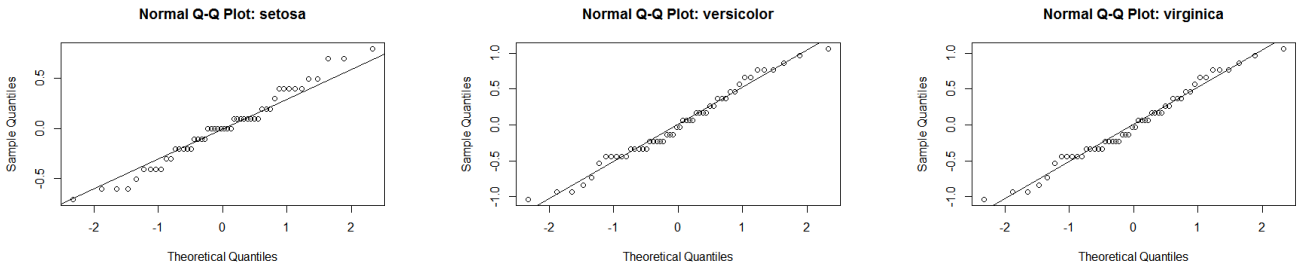


Figura 5: Gráficas de normalidad por niveles

En cualquier caso, debemos realizar los contrastes correspondientes que nos permitan decidir sobre las condiciones iniciales midiendo la seguridad de las mismas con sus respectivos p-valores. Así, para la condición de homoscedasticidad, es decir, igualdad de varianzas, aplicamos los tests de Bartlett y de Levene, observar que el test de Levene es válido en ausencia de normalidad y necesitamos cargar el paquete *car* para ejecutarlo. Los resultados mostrados a continuación evidencian una falta de homogeneidad de varianzas en la longitud de los sépalos entre las tres especies de Iris. Para la normalidad, por ejemplo aplicamos el test de Shapiro-Wilk a cada especie de Iris, aunque también incluimos el contraste de normalidad de toda la población bajo estudio, siendo en todos los casos no rechazable la normalidad de la longitud de los sépalos.

```
> bartlett.test(residuals(sepalo),Species)

Bartlett test of homogeneity of variances

data:  residuals(sepalo) and Species
```

```
Bartlett's K-squared = 16.0057, df = 2, p-value = 0.0003345

> leveneTest(residuals(sepalo),Species)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   2  6.3527 0.002259 **
      147
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> shapiro.test(sepalo$residuals[Species=="setosa"])

Shapiro-Wilk normality test

data:  sepalo$residuals[Species == "setosa"]
W = 0.9777, p-value = 0.4595

> shapiro.test(sepalo$residuals[Species=="versicolor"])

Shapiro-Wilk normality test

data:  sepalo$residuals[Species == "versicolor"]
W = 0.9778, p-value = 0.4647

> shapiro.test(sepalo$residuals[Species=="virginica"])

Shapiro-Wilk normality test

data:  sepalo$residuals[Species == "virginica"]
W = 0.9712, p-value = 0.2583

> shapiro.test(sepalo$residuals)

Shapiro-Wilk normality test

data:  sepalo$residuals
W = 0.9879, p-value = 0.2189
```

Usar también `by(sepalo$residuals, Species, shapiro.test)`. Otro test de normalidad conocido es el de Kolmogorov-Smirnov, que puede realizarse mediante la función *ks.test(residuls(sepalo),"pnorm")*, o bien mediante la librería *fBasics* que utiliza las funciones: *shapiroTest*, *ksnormTest*, *shapiroTest*, *adTest*, o en general mediante *normalTest* indicando en los argumentos el método a utilizar "sw", "ks", ...

Por otro lado, en este ejemplo de aplicación del ANOVA, con el test de Bartlett hemos observado una desviación significativa de la igualdad de varianzas, es decir, no se cumple una de las condiciones para la aplicación del test de análisis de la varianza. No obstante, en el programa R hay disponible una función para el análisis de la varianza de una vía (un factor) aplicable en situaciones de no homoscedasticidad, *oneway.test* para aplicar el test de Welch. Los resultados de este test, mostrados debajo, proporcionan un $P \simeq 0$, deduciéndose que hay diferencias muy significativas entre las longitudes medias de los sépalos de los tres tipos de especies de Iris.

```
> oneway.test(Sepal.Length ~ Species)
```

```
One-way analysis of means (not assuming equal variances)

data:  Sepal.Length and Species
F = 138.9083, num df = 2.000, denom df = 92.211, p-value < 2.2e-16
```

Comparaciones múltiples

Para completar el estudio de la longitud de los sépalos entre las tres especies de Iris, aplicamos un proceso de comparación múltiple, dado que se ha observado que hay diferencias significativas entre las longitudes medias de estas especies, es decir, la especie de pertenencia de las flores influye significativamente en la longitud de sus sépalos. Con el objetivo de detectar qué especie provoca estas diferencias controlando el error de tipo I (probabilidad de equivocarse al tomar la decisión de que dos especies tienen longitudes medias diferentes) al realizar simultáneamente varias comparaciones, tendremos que utilizar algoritmos de comparaciones simultáneas por pares.

Por ejemplo, utilizando uno de los métodos más usuales en la práctica, el método la diferencia significativa honesta de Tukey, en R está disponible la función *TukeyHSD* que se ejecuta directamente sobre el modelo:

```
> TukeyHSD(sepalo)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Sepal.Length ~ Species)

$Species
                      diff       lwr       upr p adj
versicolor-setosa    0.930 0.6862273 1.1737727     0
virginica-setosa     1.582 1.3382273 1.8257727     0
virginica-versicolor 0.652 0.4082273 0.8957727     0
```

Además, R proporciona diferentes funciones para poder presentar gráficamente estos intervalos de confianza simultáneos (véase Figura 6), por ejemplo, *plot(TukeyHSD(sepalo))*, así como los contrastes múltiples de igualdad de medias por pares de niveles, para una mejor toma de decisiones y fácil interpretación. Los siguientes resultados son la salida de este procedimiento, siendo necesario tener cargados los paquetes *multcomp*, *abind*, *Rcmdr*, *class* y *e1071*. Otros métodos de comparación dos a dos se pueden hallar en los paquetes *multcomp* y *agricolae*.

```
> library("multcomp");library("abind");library("Rcmdr")
> numSummary(iris$Sepal.Length , groups=iris$Species, statistics=c("mean","sd"))
            mean        sd data:n
setosa     5.006 0.3524897     50
versicolor 5.936 0.5161711     50
virginica  6.588 0.6358796     50
> pares <- glht(sepalo,linfct=mcp(Species="Tukey"))
> summary(pares)

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = Sepal.Length ~ Species)
```

```
Linear Hypotheses:
                           Estimate Std. Error t value Pr(>|t|)
versicolor - setosa == 0       0.930      0.103   9.033   <1e-08 ***
virginica - setosa == 0        1.582      0.103  15.366   <1e-08 ***
virginica - versicolor == 0    0.652      0.103   6.333   <1e-08 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> par(oma=c(0,5,0,0))
> plot(pares)
```

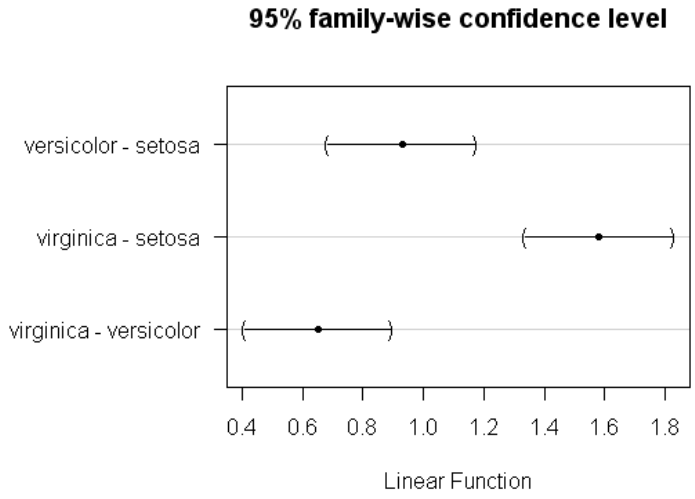


Figura 6: Intervalos simultáneos para diferencias de medias de Tukey

Asimismo, en las situaciones de comparaciones por pares con un nivel o tratamiento de control, se reduce sustancialmente la cantidad de intervalos simultáneos, como se observa al aplicar el método de Dunnett al conjunto de datos de las especies de Iris, y por defecto asigna como grupo de control la primera categoría o nivel del factor, mostrándose en la Figura 7 su representación gráfica:

```
> paresDunnett <- glht(sepalo,linfct=mcp(Species="Dunnett"))
> summary(paresDunnett)

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts


Fit: aov(formula = Sepal.Length ~ Species)

Linear Hypotheses:
                         Estimate Std. Error t value Pr(>|t|)
versicolor - setosa == 0    0.930      0.103   9.033   <1e-10 ***
virginica - setosa == 0     1.582      0.103  15.366   <1e-10 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

```
> plot(paresDunnett)
```

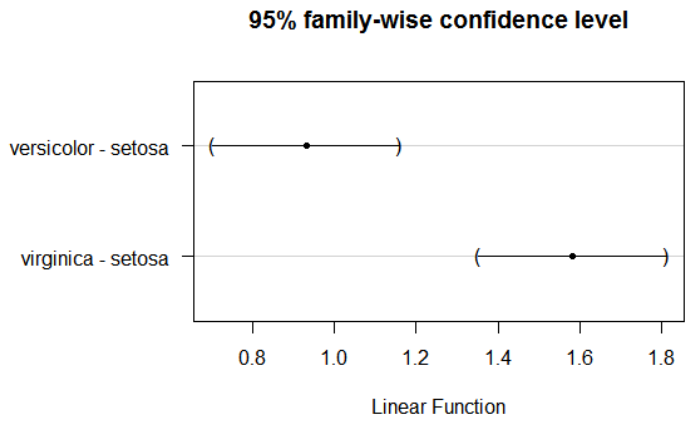


Figura 7: Intervalos simultáneos para diferencias de medias de Dunnett

Para cambiar el grupo de control o referencia utilizado por defecto, se utiliza la función *relevel*, por ejemplo:

```
> iris$Species <- relevel(iris$Species, ref="virginica")
```

Ejercicios prácticos

Ejercicio 1.1 *A partir del archivo de datos iris incluido en R, estudiar la medida morfológica de longitud de los pétalos de flores de las tres especies de Iris, resolviendo los siguientes puntos:*

(1) Presentar la tabla ANOVA para el contraste de igualdad de longitudes medias:
```
petalo <- aov(Petal.Length ~ Species)
```

(2) Obtener las estimaciones de los efectos sobre la longitud de los pétalos debidos al tipo de especie, así como las estimaciones de sus longitudes medias y su error estándar:
```
model.tables(petalo)
model.tables(petalo, type="means", se=TRUE)
```

(3) Representar los gráficos de los residuos e interpretarlos para el diagnóstico del modelo:
```
plot(petalo)
plot(petalo$model$Species, petalo$residuals)
qqnorm(petalo$residuals[Species=="setosa"])
qqline(petalo$residuals[Species=="setosa"])
```

(4) Realizar el diagnóstico de los residuos del modelo ANOVA, y comentar la validez de las conclusiones del test ANOVA:
```
bartlett.test(residuals(petalo),Species)
leveneTest(residuals(petalo),Species)
shapiro.test(petalo$residuals[Species=="setosa"])
```

(5) Analizar, si procede, el análisis de una vía para desigualdad de varianzas de las longitudes de los pétalos en cada especie:
```
oneway.test(Petal.Length ~ Species)
```

(6) Estudiar e interpretar, si procede, las comparaciones por pares de las diferencias de longitudes medias de los pétalos entre cada especie de Iris:

```
TukeyHSD(petalo)
paresTuk <- glht(petalo,linfct=mcp(Species="Tukey"))
summary(paresTuk);plot(paresTuk)
paresDun <- glht(petalo,linfct=mcp(Species="Dunnett"))
summary(paresDun);plot(paresDun)
```

Ejercicio 1.2 *Estudiar si hay similitudes o se detectan diferencias entre las tres especies de Iris, siguiendo el esquema del ejercicio anterior, en relación a:*

(1) La característica morfológica de anchura de los sépalos de las flores.

(2) La característica morfológica de anchura de los pétalos de las flores.

2. Análisis de la varianza de dos factores

Como se dice al inicio, el análisis de la varianza es una técnica estadística que se encarga de estudiar el comportamiento de una variable de interés en nuestro campo experimental a través de las observaciones procedentes de distintos grupos o tratamientos en la población, que nos informan si existen o no diferencias entre cada grupo o nivel. En el apartado anterior se considera el caso de que la partición en grupos (niveles o tratamientos) de la población objeto de estudio, se debe a un factor o variable cualitativa. En general, la variable de interés sobre la población puede verse afectada por diversos factores que determinan diversos tipos de particiones o clasificaciones de los individuos de la población, así como la estructura de diseño de experimentos llevado a cabo para la extracción de las observaciones según las diferentes cualidades o factores que pueden influir en el comportamiento de nuestra variable de interés, es decir, particiones establecidas por uno de los factores, particiones definidas por el cruce de factores, por la anidación de factores, ..., todo lo cual establecerá el modelo lineal para analizar la variable de interés.

En particular, como ejemplo del análisis de la varianza con más de un factor, nos limitaremos al estudio del ANOVA de dos factores o ANOVA doble, es decir, dos factores A y B y sus efectos sobre la variable de interés Y, teniendo en cuenta los tipos o componentes entre ambos que forman parte del modelo lineal.

2.1. Análisis de la varianza doble sin interacción

En primer lugar, consideramos el análisis de la varianza de dos factores sin interacción, es decir, cuando los dos factores A y B provocan particiones en niveles o grupos de la población y asumiendo que la interacción de ambos no es de interés en el problema, o al menos que la información disponible para el análisis no es suficiente para que dicha interacción pueda ser considerada en el modelo.

En este sentido, el objetivo del ANOVA es un contraste doble, contraste del efecto de cada factor sobre la variable Y, esto es, contrastar si se detectan diferencias significativas en el comportamiento medio de Y con respecto a los diferentes niveles del factor A, y análogamente, con respecto a los diferentes tratamientos del factor B.

Modelización en el ANOVA doble sin interacción

$$Y_{ij} = E(Y_{ij}) + \varepsilon_{ij} = \mu_{ij} + \varepsilon_{ij} = \mu + A_i + B_j + \varepsilon_{ij}$$

con $i = 1, ..., a$, y $j = 1, ..., b$, donde Y_{ij} representa la variable respuesta de la observación en los niveles (i, j) de los factores (A, B), A_i es el efecto que provoca sobre la media la pertenencia al nivel i del factor A (formado por a niveles), y B_j representa el efecto sobre la media del nivel j del otro factor B (formado por b niveles).

Observación 2.1 *Denotando por Y_{ij} la variable aleatoria Y en el nivel (i, j) del factor (A, B), $i = 1, ..., a$ y $j = 1, ..., b$, las condiciones iniciales sobre las Y_{ij} para el desarrollo del ANOVA son:*

- *Independencia*
- *Homoscedasticidad*
- *Normalidad*

es decir, $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ e independientes, $i = 1, ..., a$ y $j = 1, ..., b$; o equivalentemente, $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Objetivo del ANOVA doble sin interacción

Estudiar la influencia de cada factor sobre Y a través de los tests de hipótesis

$$\text{Test } A \begin{cases} H_0 & : & A_1 = A_2 = ... = A_a \text{ (efectos iguales de los niveles del factor } A) \\ H_1 & : & \text{no todos los efectos iguales} \end{cases}$$

$$\text{Test } B \begin{cases} H_0 & : & B_1 = B_2 = ... = B_b \text{ (efectos iguales de los niveles del factor } B) \\ H_1 & : & \text{no todos los efectos iguales} \end{cases}$$

Observación 2.2 *La diferencia entre el tipo de factores (fijos, aleatorios o mixtos) radica en considerar la variación nula entre los efectos correspondientes al factor aleatorio.*

Desarrollo del test ANOVA doble sin interacción

El test ANOVA doble sin interacción se basa en la descomposición de la variación del experimento:

$$\text{"Var. total} = \text{Var. entre grupos de } A + \text{Var. entre grupos de } B + \text{Variación error"}$$

dado que se cumple la siguiente igualdad de sumas de cuadrados:

$$SS_T = SS_A + SS_B + SS_R$$

donde las sumas de cuadrados de todas las variaciones (SS_T), de las variaciones sólo entre niveles de A (SS_A), de las variaciones sólo entre niveles de B (SS_B) y de las variaciones restantes (SS_R), vienen dadas por las siguientes expresiones:

$$\begin{aligned} SS_T &= \sum_{i=1}^{a} \sum_{j=1}^{b} \left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2 \\ SS_A &= \sum_{i=1}^{a} b \left(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet}\right)^2 \\ SS_B &= \sum_{j=1}^{b} a \left(\overline{Y}_{\bullet j} - \overline{Y}_{\bullet\bullet}\right)^2 \\ SS_R &= \sum_{i=1}^{a} \sum_{j=1}^{b} \left(Y_{ij} - \overline{Y}_{i\bullet} - \overline{Y}_{\bullet j} + \overline{Y}_{\bullet\bullet}\right)^2 \end{aligned}$$

siendo $\overline{Y}_{i\bullet} = \frac{1}{b} \sum_{j=1}^{b} Y_{ij}$ la media muestral en el nivel i del factor A, $\overline{Y}_{\bullet j} = \frac{1}{a} \sum_{i=1}^{a} Y_{ij}$ la media muestral en el nivel j del factor B, $\overline{Y}_{\bullet\bullet} = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} Y_{ij}$ la media muestral total.

Presentación de los contrastes: Tabla ANOVA doble

Fuente	S.Cuadrados	G.Libertad	Medias Cuadrados	Estadísticos	P-valor
Factor A:	SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_A = \frac{MS_A}{MS_R}$	p_A
Factor B:	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$F_B = \frac{MS_B}{MS_R}$	p_B
Error:	SS_R	$(a-1)(b-1)$	$MS_R = \frac{SS_R}{(a-1)(b-1)}$		
Total:	SS_T	$n-1 = ab-1$			

donde las distribuciones muestrales de estos estadísticos son las siguientes:

- Test A: el estadístico $F_A \sim F_{a-1,(a-1)(b-1)}$, bajo su hipótesis nula H_0. La región crítica al nivel α es $M_1^* = \left\{F_A > F_{a-1,(a-1)(b-1),\alpha}\right\}$, y el p-valor $p_A = \Pr\left(F_{a-1,(a-1)(b-1)} > F_{A,\text{experimental}}\right)$.
- Test B: el estadístico $F_B \sim F_{b-1,(a-1)(b-1)}$, bajo su hipótesis nula H_0. La región crítica al nivel α es $M_1^* = \left\{F_B > F_{b-1,(a-1)(b-1),\alpha}\right\}$, y el p-valor $p_B = \Pr\left(F_{b-1,(a-1)(b-1)} > F_{B,\text{experimental}}\right)$.

Por último, recordar que en el estudio de un modelo de análisis de la varianza de dos factores sin interacción, tiene que comprobarse las condiciones iniciales, es decir, un diagnóstico de los residuos de este modelo, para lo que disponemos de las mismas técnicas utilizadas en el caso de ANOVA simple. Análogamente, cuando los resultados del análisis indican que existen diferencias significativas con respecto a los niveles de un factor, pueden aplicarse, del mismo modo, las técnicas de comparación múltiple incluidas en el apartado de ANOVA de un factor.

2.2. Análisis de la varianza doble con interacción

En este caso, consideramos el análisis de la varianza de dos factores con interacción, es decir, cuando ambos factores A y B y la interacción entre ambos $C = AB$, provocan particiones en niveles o grupos de la población, siendo de interés en el problema, el estudio de las posibles diferencias en el comportamiento de la variable Y provocadas por cualquiera de estos tres términos, lo que habitualmente se llaman efectos principales de los factores y efectos del cruce de ambos factores.

Así, el objetivo del ANOVA es un contraste triple, contraste del efecto de cada término del modelo sobre la variable Y, esto es, contrastar si se detectan diferencias significativas en el comportamiento medio de Y con respecto a los diferentes niveles del factor A, con respecto a los diferentes tratamientos del factor B, y con respecto a los diferentes grupos de la partición formada por el intersección de ambos factores AB.

Modelización en el ANOVA doble con interacción

Para simplificar la notación, supondremos un diseño balanceado, $n_{ij} = m$, esto es el mismo tamaño para todas las muestras registradas en cada interacción o cruce de niveles de ambos factores, y los dos factores fijos:

$$Y_{ijk} = E(Y_{ijk}) + \varepsilon_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}$$

para $i = 1, ..., a$, $j = 1, ..., b$ y $k = 1, ..., m$, donde Y_{ijk} representa la variable respuesta de la k-ésima observación en los niveles (i, j) de los factores (A, B), A_i es el efecto que provoca sobre la media la pertenencia al nivel i del factor A (formado por a niveles), B_j representa el efecto sobre la media del nivel j del factor B (formado por b niveles), y AB_{ij} representa el efecto sobre la media del cruce de los niveles (i, j) de la interacción AB (formada por ab grupos).

Observación 2.3 *Representando por Y_{ij} la variable aleatoria Y en el nivel (i, j) del cruce de factores AB, $i = 1, ..., a$ y $j = 1, ..., b$, las condiciones iniciales sobre las variables Y_{ij} para el desarrollo del son:*

- *Independencia*
- *Homoscedasticidad*
- *Normalidad*

es decir, $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$ *e independientes.*

Por tanto, para una m.a.s. $Y_{ij1}, ..., Y_{ijm}$ de Y_{ij}, es decir, m.a.s. de la variable Y en el cruce de niveles (i, j) con $i = 1, ..., a$ y $j = 1, ..., b$, tenemos que:

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2) \Leftrightarrow \varepsilon_{ijk} \sim N(0, \sigma^2) \text{ e independientes, } k = 1, ..., m.$$

Objetivo del ANOVA doble con interacción

En este marco, el objetivo del ANOVA de dos factores con interacción es contrastar si el valor medio de Y es el mismo en todos los niveles o tratamientos del factor A, si es el mismo en todos los niveles del factor B, y si es el mismo en todos los grupos de la interacción de los factores AB, es decir,

$$\text{Test } A \begin{cases} H_0 & : & A_1 = A_2 = ... = A_a \\ H_1 & : & \text{no todos los efectos iguales del factor } A \end{cases}$$

$$\text{Test } B \begin{cases} H_0 & : & B_1 = B_2 = ... = B_b \\ H_1 & : & \text{no todos los efectos iguales del factor } B \end{cases}$$

$$\text{Test } AB \begin{cases} H_0 & : & AB_{11} = ... = AB_{1b} = ... = AB_{a1} = ... = AB_{ab} \\ H_1 & : & \text{no todos los efectos de la interacción son iguales} \end{cases}$$

Desarrollo del test ANOVA doble con interacción

En este tipo de diseño de experimentos, el test ANOVA doble con interacción se basa en la descomposición de la variación del experimento en las siguientes componentes independientes de variabilidad:

$$\begin{aligned} \text{"Var. total} = \text{ Var. entre grupos de } A + \text{ Var. entre grupos de } B \\ + \text{ Var. entre interacciones de } AB + \text{ Var. error"} \end{aligned}$$

y que se corresponde con la igualdad entre sus sumas de cuadrados

$$SS_T = SS_A + SS_B + SS_{AB} + SS_R$$

donde las sumas de cuadrados de todas las variaciones (SS_T), de las variaciones sólo entre los niveles de A (SS_A), de las variaciones sólo entre los niveles de B (SS_B), de las variaciones entre los cruces de niveles de A y B (SS_{AB}), y de las variaciones residuales, vienen dadas por las siguientes expresiones:

$$\begin{aligned} SS_T &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{m} \left(Y_{ijk} - \overline{Y}_{\bullet\bullet\bullet}\right)^2 \\ SS_A &= bm \sum_{i=1}^{a} \left(\overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}\right)^2 \\ SS_B &= am \sum_{j=1}^{b} \left(\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}\right)^2 \\ SS_{AB} &= m \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\bullet} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}\right)^2 \\ SS_R &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{m} \left(Y_{ijk} - \overline{Y}_{ij\bullet}\right)^2 \end{aligned}$$

siendo $\overline{Y}_{i\bullet\bullet} = \frac{1}{bm} \sum_{j=1}^{b} \sum_{k=1}^{m} Y_{ijk}$ la media muestral en el nivel i del factor A, $\overline{Y}_{\bullet j\bullet} = \frac{1}{am} \sum_{i=1}^{a} \sum_{k=1}^{m} Y_{ijk}$ la media muestral en el nivel j del factor B, $\overline{Y}_{ij\bullet} = \frac{1}{m} \sum_{k=1}^{m} Y_{ijk}$ la media muestral en el cruce de niveles i y j de los factores A y B, $\overline{Y}_{\bullet\bullet\bullet} = \frac{1}{abm} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{m} Y_{ijk}$ la media muestral total.

Presentación de los contrastes: Tabla ANOVA doble con interacción

Fuente	S. Cuadrados	G. libertad	Medias Cuadrados	Estadísticos	P-valor
Factor A	SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_A = \frac{MS_A}{MS_R}$	p_A
Factor B	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$F_B = \frac{MS_B}{MS_R}$	p_B
Interacción AB	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$F_{AB} = \frac{MS_{AB}}{MS_R}$	p_{AB}
Error	SS_R	$ab(m-1)$	$MS_R = \frac{SS_R}{ab(m-1)}$		
Total	SS_T	$abm-1$			

donde las distribuciones muestrales de estos estadísticos son las siguientes:

- Test A: $F_A \sim F_{a-1,ab(m-1)}$, bajo su hipótesis nula H_0. La región crítica al nivel de significación α es $M_1^* = \left\{F_A > F_{a-1,ab(m-1),\alpha}\right\}$, y el p-valor de este contraste es $p_A = \Pr\left(F_{a-1,ab(m-1)} > F_{A,\text{exp}}\right)$
- Test B: $F_B \sim F_{b-1,ab(m-1)}$, bajo su hipótesis nula H_0. La región crítica al nivel de significación α es $M_1^* = \left\{F_B > F_{b-1,ab(m-1),\alpha}\right\}$, y el p-valor de este contraste es $p_B = \Pr\left(F_{b-1,ab(m-1)} > F_{B,\text{exp}}\right)$
- Test AB: $F_{AB} \sim F_{(a-1)(b-1),ab(m-1)}$, bajo su hipótesis nula H_0. La región crítica al nivel de significación α es $M_1^* = \left\{F_{AB} > F_{(a-1)(b-1),ab(m-1),\alpha}\right\}$, y el p-valor de este contraste es $p_{AB} = \Pr\left(F_{(a-1)(b-1),ab(m-1)} > F_{AB,\text{exp}}\right)$.

Por último, recordar que en el estudio de un modelo de análisis de la varianza de dos factores con interacción, tiene que comprobarse las condiciones iniciales, es decir, un diagnóstico de los residuos de este modelo, para lo que disponemos de las mismas técnicas utilizadas en el caso de ANOVA simple. Análogamente, cuando los resultados del análisis indican que existen diferencias significativas con respecto a los niveles de un factor, pueden aplicarse, del mismo modo, las técnicas de comparación múltiple incluidas en el apartado de ANOVA de un factor.

2.3. Caso práctico

Veamos un caso práctico de aplicación del análisis de la varianza de dos factores a través del programa R. En primer lugar, utilizaremos un ejemplo de ANOVA doble sin interacción en un estudio de producción agrícola según el tratamiento de fertilizante aplicado teniendo en cuenta los tipos de fincas homogéneas sobre las que se experimenta. Posteriormente, consideraremos un ejemplo de ANOVA doble con interacción en un estudio del rendimiento de un proceso bajo determinados niveles de temperatura y de PH.

Ejemplo de ANOVA doble sin interacción

El conjunto de datos *cultivofert* contiene las medidas de producción que se registraron en un experimento agrícola. El experimento consistió en evaluar los efectos de cuatro tratamientos fertilizantes sobre la producción, para lo que se aplicó cada tratamiento en fincas de cinco zonas diferentes, las fincas tenían características similares y los tratamientos se aplicaron al azar entre las cuatro fincas seleccionadas de la misma zona.

Es este caso, la variable *fert* es un factor que indica el tipo de tratamiento del que procede la medida de producción agrícola, y la variable *finca* es un bloque incluido en el experimento para controlar el posible efecto de la zona de cultivo sobre la producción. Por tanto, se pretende saber si la aplicación de los diferentes tratamientos fertilizantes provoca diferencias en la producción, así como el posible efecto de los diferentes enclaves de las fincas.

Test ANOVA doble

En primer lugar, cargamos el fichero y declaramos como factores los términos del modelo de análisis de la varianza que afectan a la producción agrícola. Observar que cuando el factor no es numérico no es necesario declararlo, pero cuando es numérico hay que declarar la variable cualitativa como factor para que no la se considere como variable cuantitativa. También puede obtenerse una descripción de la producción por factores y gráficos descriptivos (ver Figuras 8 y 9).

```
> attach(cultivofert)
> fert <- factor(fert, labels=c(1,2,3,4))
> finca <- factor(finca, labels=c(1,2,3,4,5))
> plot(produc ~ finca + fert)
> tapply(produc, finca, summary)
> tapply(produc, fert, summary)
> stripchart(produc ~ finca, method="stack")
> stripchart(produc ~ fert, method="stack")
> interaction.plot(finca, fert, produc)
> interaction.plot(fert, finca, produc)
```

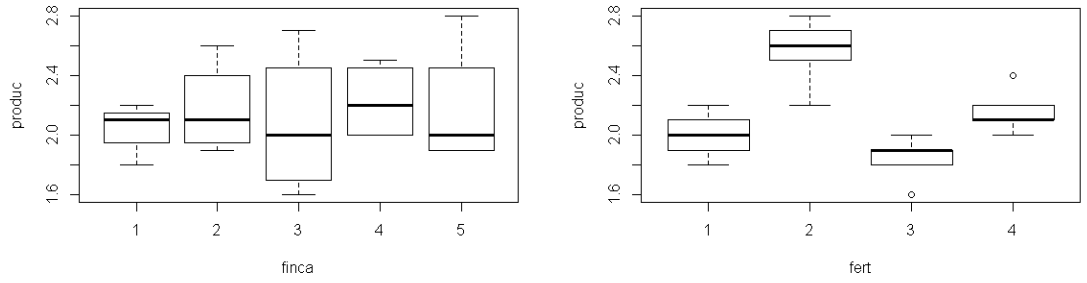


Figura 8: Gráficos Q-Q plot de la producción agrícola

```
> tapply(produc, finca, summary)
$`1`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     1.800   2.025   2.100   2.050   2.125   2.200
$`2`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     1.900   1.975   2.100   2.175   2.300   2.600
$`3`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     1.600   1.750   2.000   2.075   2.325   2.700
$`4`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     2.000   2.000   2.200   2.225   2.425   2.500
$`5`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     1.900   1.900   2.000   2.175   2.275   2.800

> tapply(produc,fert,summary)
$`1`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
       1.8     1.9     2.0     2.0     2.1     2.2
$`2`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2.20    2.50    2.60    2.56    2.70    2.80
$`3`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1.60    1.80    1.90    1.84    1.90    2.00
$`4`  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2.00    2.10    2.10    2.16    2.20    2.40
```

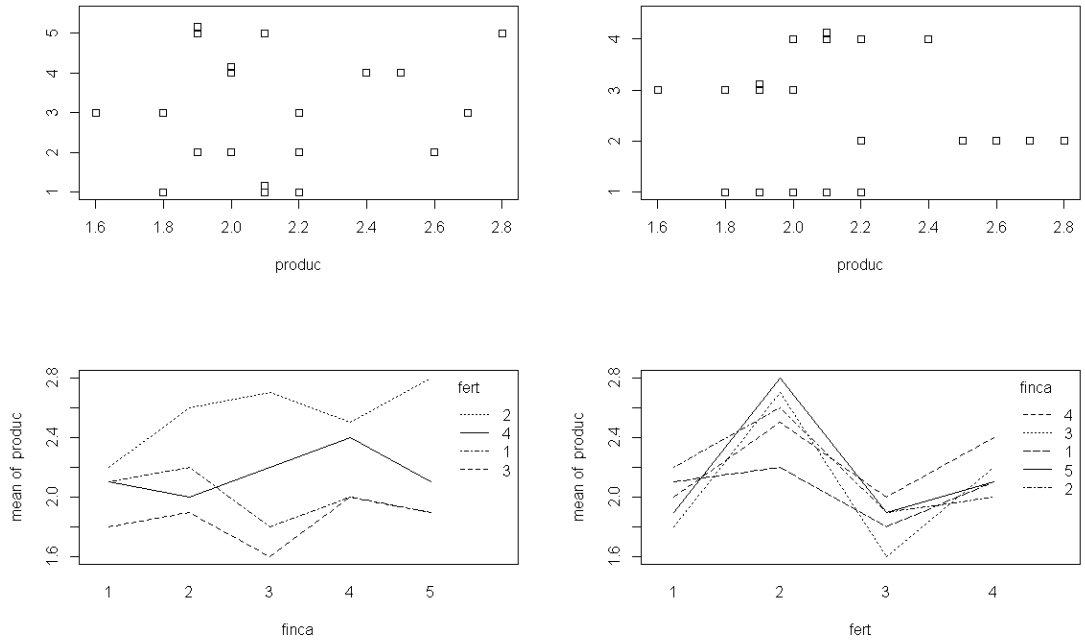


Figura 9: Gráficos Q-Q plot de la producción agrícola

Suponiendo que se mantienen las condiciones de normalidad y homoscedasticidad, las cuales deberán ser comprobadas con los residuos, aplicamos el test de ANOVA para los dos factores mediante la instrucción *aov*, obteniendo los resultados siguientes:

```
> doble <- aov(produc ~ finca + fert)
> summary(doble)
            Df Sum Sq Mean Sq F value   Pr(>F)
finca        4  0.088  0.0220   0.647 0.639572
fert         3  1.432  0.4773  14.039 0.000314 ***
Residuals   12  0.408  0.0340
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Además, de forma similar al ANOVA simple, obtenemos las estimaciones de los efectos de las ubicaciones de las fincas y de los tratamientos fertilizantes aplicados, las estimaciones de las producciones medias y de los errores estándar de las diferencias de medias:

```
> model.tables(doble)
Tables of effects

finca        1      2      3      4      5
       -0.090  0.035 -0.065  0.085  0.035

fert        1     2     3     4
       -0.14  0.42 -0.30  0.02
> model.tables(doble,type="means",se=TRUE)
Tables of means
Grand mean

2.14
```

```
finca       1     2     3     4     5
        2.050 2.175 2.075 2.225 2.175

fert       1    2    3    4
        2.00 2.56 1.84 2.16

Standard errors for differences of means
         finca   fert
        0.1304 0.1166
replic.      4      5
```

Comparaciones múltiples

A partir de los resultados, no se detecta que la ubicación de la finca afecte en la producción agrícola, ya que con un p-valor 0.6395, se tendría un 64 % de error en la decisión de que la zona de cultivo provoca diferencias en la producción media. Sin embargo, el tratamiento fertilizante si resulta significativo en la producción agrícola. Así, sería de interés conocer que tratamiento provoca estas diferencias, y sobretodo el tratamiento o tratamientos fertilizantes que provocan una mayor producción media. Para ello, representamos las producciones medias por cada tratamiento fertilizante y aplicamos las comparaciones múltiples de Tukey para discutir entre los cuatro tratamientos:

```
> library("gplots")
> plotmeans(produc ~ fert)
```

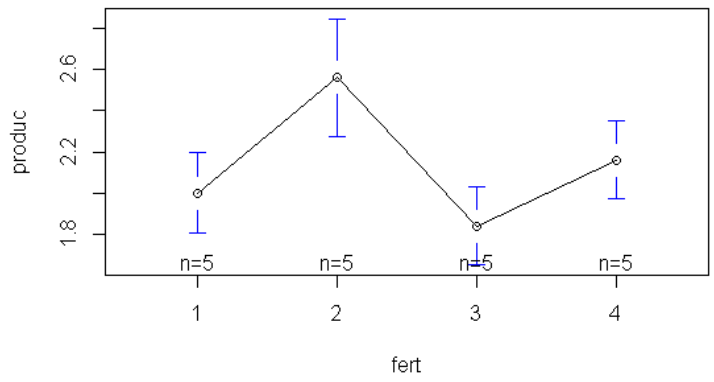


Figura 10: Gráfico de producciones medias por tratamiento fertilizante

```
> TukeyHSD(doble,"fert")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = produc ~ finca + fert)

$fert
     diff         lwr         upr     p adj
2-1  0.56  0.21376961  0.90623039 0.0021078
```

```
3-1 -0.16 -0.50623039  0.18623039 0.5385596
4-1  0.16 -0.18623039  0.50623039 0.5385596
3-2 -0.72 -1.06623039 -0.37376961 0.0002414
4-2 -0.40 -0.74623039 -0.05376961 0.0223807
4-3  0.32 -0.02623039  0.66623039 0.0734811

> library("multcomp")
> summary(glht(doble,linfct=mcp(fert="Tukey")))

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

Fit: aov(formula = produc ~ finca + fert)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0   0.5600     0.1166   4.802  0.00215 **
3 - 1 == 0  -0.1600     0.1166  -1.372  0.53848
4 - 1 == 0   0.1600     0.1166   1.372  0.53858
3 - 2 == 0  -0.7200     0.1166  -6.174  < 0.001 ***
4 - 2 == 0  -0.4000     0.1166  -3.430  0.02238 *
4 - 3 == 0   0.3200     0.1166   2.744  0.07369 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)

> plot(TukeyHSD(doble,"fert"))
```

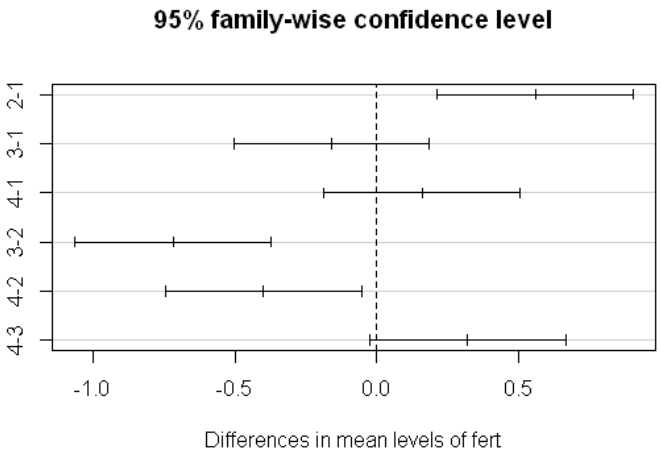


Figura 11: Intervalos de diferencias de producciones medias de Tukey

Observamos que el segundo tratamiento de fertilizante proporciona la mejor estimación de la producción media, y en la Figura 11 se aprecia como los intervalos de las diferencias medias de producciones entre este tratamiento y los restantes se sitúan completamente por encima o por debajo del 0 (este gráfico también se puede realizar mediante la función *plot(glht(...))*). En cualquier caso, a partir de los p-valores de las comparaciones anteriores, el efecto del segundo tratamiento es significativamente

diferente al primero (0.002), al tercero (<0.001) y al cuarto (0.02), aunque con este último tratamiento fertilizante la diferencia no es muy significativa.

Para finalizar este análisis de la varianza de dos factores, observar que en cada cruce de niveles del factor y el bloque sólo hay una observación, por lo que no se dispone de información suficiente para aplicar los contrastes de los residuos. No obstante, dejamos como ejercicio la comprobación de estos contrastes de normalidad y homoscedasticidad entre los niveles de cada factor, y en especial, en el factor de los tratamientos de fertilizante aplicados que afecta de forma significativa en la producción agrícola.

Ejemplo de ANOVA doble con interacción

En este ejemplo se analiza el rendimiento de un proceso en un experimento, evaluando el efecto de los niveles de temperatura y de PH bajo los que se realiza.

El experimento se realizó en 5 ocasiones bajo cada uno de los niveles de PH y de temperatura en las que se registraron el rendimiento del proceso. Los tres niveles utilizados de PH fueron básico, neutro y ácido, y para temperaturas de 30^{o}C, 35^{o}C y 40^{o}C. Las observaciones del experimento se encuentran en el conjunto de datos *ph*.

Es este problema, las variables *pH* y *Temp* son dos factores que indican el tipo de tratamiento aplicado a través del cruce de sus niveles para medir el rendimiento del proceso. Por tanto, se pretende saber si la aplicación de los diferentes niveles de ambos factores produce diferencias en el rendimiento, es decir, la interacción entre ambos, así como la posible influencia individual de un factor principal sobre este rendimiento.

Test ANOVA doble con interacción

En primer lugar, cargamos el fichero *ph* y declaramos como factores los términos *pH* y *Temp* del modelo de análisis de la varianza que afectan al rendimiento del proceso. Observar que si el factor está definido numéricamente en R es necesario declararlo como factor, en otro caso no será necesario.

Al igual que en el caso de ANOVA sin interacción, puede mostrarse previamente una descripción del rendimiento del proceso de acuerdo a los factores y su interacción (ver Figuras 12 y 13).

```
> attach(ph)
> plot(y ~ pH * Temp)
> tapply(y, pH, summary)
> tapply(y, Temp, summary)
> interaction.plot(pH, Temp, y)
> interaction.plot(Temp, pH, y)
> plotmeans(y ~ pH)
> plotmeans(y ~ Temp)


$Acido    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
           97.8   103.0   104.7   104.2   106.2   108.6

$Basico   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
          94.10   97.55   99.90   99.37  101.60  102.30

$Neutro   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
           98.6   100.2   101.8   101.8   102.8   105.6

$`30C`    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
          94.10   97.05  100.30   99.83  102.60  105.60
```

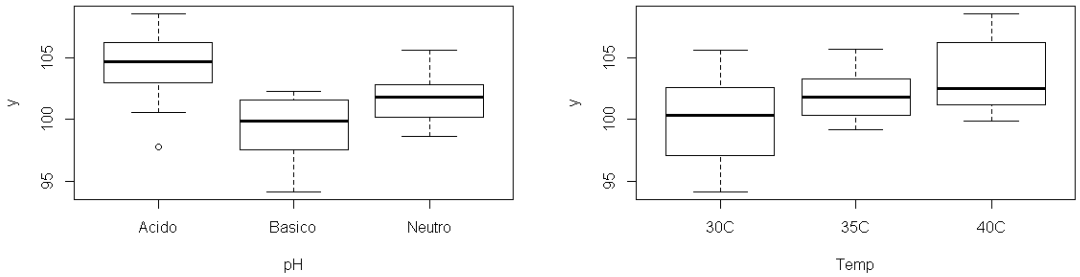


Figura 12: Gráficos Q-Q plot del rendimiento del proceso

```
$'35C'    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
          99.2   100.4   101.8   102.1   103.3   105.7

$'40C'    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
          99.9   101.2   102.5   103.5   106.2   108.6
```

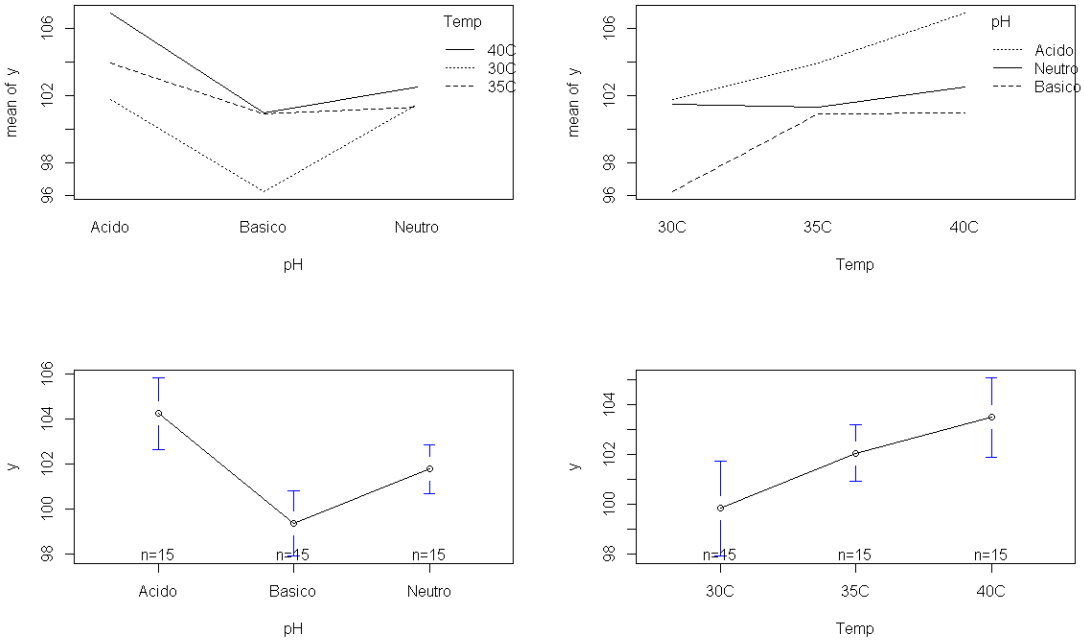


Figura 13: Gráficos de interacciones y de medias en el rendimiento del proceso

Además, asumiendo las condiciones iniciales de normalidad y homoscedasticidad, las cuales deberán ser comprobadas con los residuos, aplicamos el test de ANOVA para los dos factores incluyendo la interacción mediante la instrucción *aov*, obteniendo los resultados siguientes:

```
> doblecon <- aov(y ~ pH + Temp + pH:Temp)
> summary(doblecon)
            Df Sum Sq Mean Sq F value   Pr(>F)
pH           2 177.64   88.82  26.806 7.43e-08 ***
Temp         2 101.65   50.82  15.339 1.52e-05 ***
```

```
pH:Temp      4  43.58   10.90   3.288   0.0214 *
Residuals   36 119.28    3.31
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

de donde se deduce que los tres términos son significativos, es decir, ambos factores tienen un efecto muy significativo en el rendimiento del proceso, y la interacción entre ambos también produce diferencias significativas (p-valor 0.0214), lo que conlleva a que el rendimiento medio del proceso no es igual en todos los tratamientos establecidos por los cruces entre PH y temperatura.

Análogamente, vemos las estimaciones de los efectos de lo niveles de PH y temperatura, así como de las estimaciones de los rendimientos medios y sus errores estándar de las diferencias de medias:

```
> model.tables(doblecon)
Tables of effects

pH    Acido  Basico  Neutro
     2.4422 -2.4244 -0.0178

Temp    30C     35C     40C
    -1.9578  0.2622  1.6956

 pH:Temp
        Temp
pH       30C     35C     40C
  Acido  -0.4956 -0.5356  1.0311
  Basico -1.1689  1.2511 -0.0822
  Neutro  1.6644 -0.7156 -0.9489
> model.tables(doblecon, type="means",se=TRUE)
Tables of means
Grand mean

101.7911

pH   Acido Basico Neutro
    104.23  99.37 101.77

Temp   30C    35C    40C
     99.83 102.05 103.49

 pH:Temp
        Temp
pH       30C    35C    40C
  Acido  101.78 103.96 106.96
  Basico  96.24 100.88 100.98
  Neutro 101.48 101.32 102.52

Standard errors for differences of means
            pH   Temp pH:Temp
        0.6647 0.6647  1.1513
replic.     15     15       5
```

Comparaciones múltiples

Como hemos visto, los resultados del ANOVA detectan que la interacción entre los factores de niveles de PH y temperatura influye en las diferencias del rendimiento del proceso, por lo que el análisis se completa con las comparaciones para detectar qué cruce de PH y temperatura afecta en el rendimiento del proceso. Para ello, aplicamos el método de Tukey:

```
> TukeyHSD(doblecon)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ pH + Temp + pH:Temp)

$pH
                    diff       lwr        upr     p adj
Basico-Acido  -4.866667 -6.491329 -3.2420040 0.0000000
Neutro-Acido  -2.460000 -4.084663 -0.8353373 0.0020162
Neutro-Basico  2.406667  0.782004  4.0313294 0.0025225

$Temp
            diff        lwr      upr     p adj
35C-30C 2.220000  0.5953373 3.844663 0.0054333
40C-30C 3.653333  2.0286706 5.277996 0.0000096
40C-35C 1.433333 -0.1913294 3.057996 0.0927672

$`pH:Temp`
                       diff        lwr        upr     p adj
Basico:30C-Acido:30C  -5.54 -9.3357765 -1.7442235 0.0008074
Neutro:30C-Acido:30C  -0.30 -4.0957765  3.4957765 0.9999992
Acido:35C-Acido:30C    2.18 -1.6157765  5.9757765 0.6222068
Basico:35C-Acido:30C  -0.90 -4.6957765  2.8957765 0.9967275
Neutro:35C-Acido:30C  -0.46 -4.2557765  3.3357765 0.9999766
Acido:40C-Acido:30C    5.18  1.3842235  8.9757765 0.0020072
Basico:40C-Acido:30C  -0.80 -4.5957765  2.9957765 0.9985630
Neutro:40C-Acido:30C   0.74 -3.0557765  4.5357765 0.9991776
Neutro:30C-Basico:30C  5.24  1.4442235  9.0357765 0.0017269
Acido:35C-Basico:30C   7.72  3.9242235 11.5157765 0.0000027
Basico:35C-Basico:30C  4.64  0.8442235  8.4357765 0.0075195
Neutro:35C-Basico:30C  5.08  1.2842235  8.8757765 0.0025752
Acido:40C-Basico:30C  10.72  6.9242235 14.5157765 0.0000000
Basico:40C-Basico:30C  4.74  0.9442235  8.5357765 0.0059177
Neutro:40C-Basico:30C  6.28  2.4842235 10.0757765 0.0001186
Acido:35C-Neutro:30C   2.48 -1.3157765  6.2757765 0.4550931
Basico:35C-Neutro:30C -0.60 -4.3957765  3.1957765 0.9998242
Neutro:35C-Neutro:30C -0.16 -3.9557765  3.6357765 1.0000000
Acido:40C-Neutro:30C   5.48  1.6842235  9.2757765 0.0009409
Basico:40C-Neutro:30C -0.50 -4.2957765  3.2957765 0.9999557
Neutro:40C-Neutro:30C  1.04 -2.7557765  4.8357765 0.9913684
Basico:35C-Acido:35C  -3.08 -6.8757765  0.7157765 0.1926494
Neutro:35C-Acido:35C  -2.64 -6.4357765  1.1557765 0.3724635
Acido:40C-Acido:35C    3.00 -0.7957765  6.7957765 0.2195735
Basico:40C-Acido:35C  -2.98 -6.7757765  0.8157765 0.2267097
Neutro:40C-Acido:35C  -1.44 -5.2357765  2.3557765 0.9388247
Neutro:35C-Basico:35C  0.44 -3.3557765  4.2357765 0.9999834
Acido:40C-Basico:35C   6.08  2.2842235  9.8757765 0.0002000
```

```
Basico:40C-Basico:35C  0.10 -3.6957765  3.8957765 1.0000000
Neutro:40C-Basico:35C  1.64 -2.1557765  5.4357765 0.8808403
Acido:40C-Neutro:35C   5.64  1.8442235  9.4357765 0.0006249
Basico:40C-Neutro:35C -0.34 -4.1357765  3.4557765 0.9999978
Neutro:40C-Neutro:35C  1.20 -2.5957765  4.9957765 0.9786583
Basico:40C-Acido:40C  -5.98 -9.7757765 -2.1842235 0.0002595
Neutro:40C-Acido:40C  -4.44 -8.2357765 -0.6442235 0.0120417
Neutro:40C-Basico:40C  1.54 -2.2557765  5.3357765 0.9128484
```

No obstante, las gráficas de los intervalos de confianza simultáneos de Tukey para las diferencias de rendimientos medios entre los niveles de los factores principales, incluso entre los tratamientos establecidos por la interacción, Figura 14, ayudan a detectar los grupos con un rendimiento del proceso similar y los que producen las diferencias.

```
> plot(TukeyHSD(doblecon,"pH"))
> plot(TukeyHSD(doblecon,"Temp"))
> plot(TukeyHSD(doblecon))
```

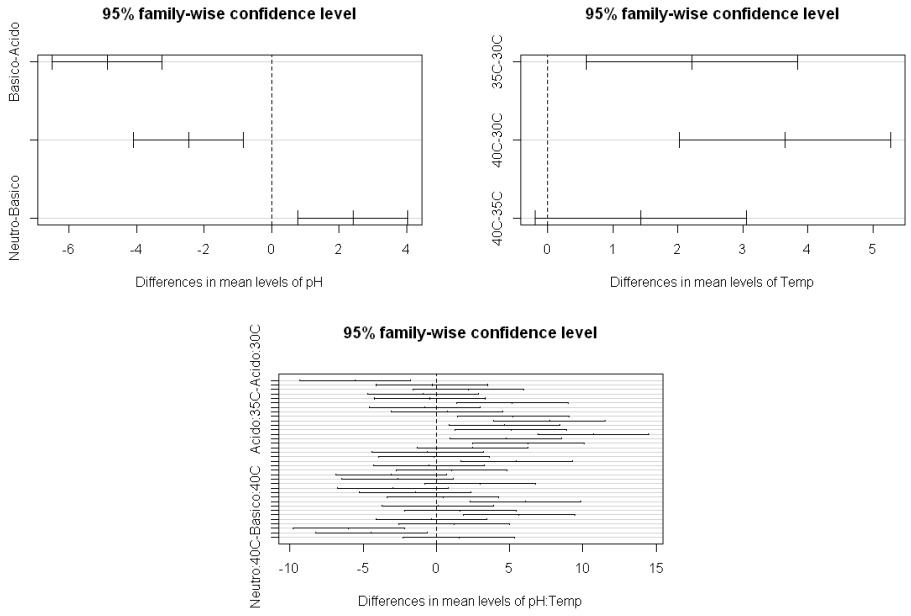


Figura 14: Intervalos para diferencias de rendimientos medios Tukey

Al igual que en el caso de ANOVA simple y doble sin interacción, se requieren unas condiciones iniciales de normalidad y homoscedasticidad, y por tanto, se deben realizar los contrastes de normalidad y de igualdad de varianzas entre los diferentes niveles, lo que dejamos como ejercicio.

Ejercicios prácticos

Ejercicio 2.1 *En el estudio de los efectos del fertilizante y zona de cultivo sobre la producción:*

(1) Comprobar que se mantienen las condiciones de normalidad y homoscedasticidad en el análisis de los efectos principales del factor fertilizante y del bloque zona de cultivo.

(2) Suponiendo que el tratamiento 3 es un fertilizante inocuo, analizar las diferencias entre los efectos de los factores mediante la comparación por pares de Dunnett.

Ejercicio 2.2 *En el estudio de los efectos de la temperatura y el PH sobre el rendimiento de un proceso experimental, realizar el diagnóstico de los residuos para la validez de los resultados.*

Ejercicio 2.3 *En un estudio sobre la radiación en una cámara experimental, un grupo de investigadores midieron la radiación utilizando cuatro dispositivos diferentes: filtros, membranas, vasos abiertos y placas. El experimento se basó en probar 20 dispositivos de cada tipo, y se registró la cantidad de radiación que midió cada dispositivo, cuyas observaciones están en el fichero radiacion.*

(1) Analizar si los dispositivos utilizados tienen algún efecto sobre las mediciones de radiación.

(2) ¿Cuál es el dispositivo que detecta mejor la radiación?

(3) ¿Qué dispositivos de medida de radiación no muestran diferencias?

(4) ¿Son fiables las respuestas a los apartados anteriores?

Ejercicio 2.4 *En el fichero de datos composite de la librería faraway, se recogen los datos de un experimento para probar la fuerza de un compuesto termoplástico en función de la potencia de un láser y la velocidad de una cinta. Las variables del fichero son strength la fuerza observada en el compuesto, laser es la potencia a 40W, 50W y 60W utilizada en el experimento y tape indica los niveles de velocidad de la cinta, slow=6.42 m/s, medium=13m/s y fast=27m/s.*

(1) Analizar si el efecto de los tratamientos sobre la fuerza del compuesto.

(2) ¿Qué nivel de potencia produce mayor diferencia en la fortaleza del compuesto?

(3) ¿Puede considerarse similitud de fortaleza del compuesto ante dos niveles de velocidad de la cinta?

(4) ¿Son fiables las respuestas a los apartados anteriores?

Ejercicio 2.5 *El archivo de datos falcon incluye las observaciones de un estudio sobre el efecto residual del pesticida DDT en halcones. La variable DDT mide el contenido de este pesticida en halcones que fueron capturados en tres áreas de anidamiento, la columna Site identifica cada área, y de tres grupos de edad, la columna Age incluye los tres grupos de edad.*

(1) Analizar si los efectos del sitio de captura y edad en la presencia del pesticida.

(2) ¿Qué sitio registra mayor influencia del pesticida en los halcones?

(3) Discutir la interacción de los factores sitio y edad en relación a las medidas del pesticida

(4) Debatir sobre las condiciones iniciales de la técnica empleada.

Ejercicio 2.6 *La base de datos chickwts disponible en R corresponde a las observaciones del peso de pollos criados con distintos alimentos.*

(1) Analizar si el tipo de alimentación tiene alguna influencia en el peso final de los pollos.

(2) ¿Es algún tratamiento alimenticio más productivo para el engorde de los pollos?

(3) Estimar las diferencias de pesos medios con los distintos tipos de alimentación de los pollos.

(4) ¿Son fiables las respuestas a los apartados anteriores?

Ejercicio 2.7 *Los estudiantes de una clase participaron en un sencillo experimento en el que cada estudiante registró su estatura, peso, sexo, preferencia para fumar, nivel de actividad usual y pulso en reposo. Luego, todos lanzaron una moneda y aquellos que obtuvieron cara corrieron durante un minuto, tras el que se registró de nuevo el pulso a toda la clase.*

El fichero pulso contiene los datos de este experimento realizado sobre 92 estudiantes, y sus variables se describen en la siguiente tabla:

Tabla 1: Descripción de las variables del archivo pulso

Nombre	Descripción
pulse1	Primera tasa de pulso
pulse2	Segunda tasa de pulso pasado un minuto
ran	Corrió un minuto =1, no corrió un minuto =2
smokes	Fuma regularmente =1, no fuma regularmente =2
sex	Masculino =1, femenino=2
weight	Peso del estudiante
height	Altura del estudiante
activity	Nivel usual de actividad física: 1=ligero, 2=moderado y 3=mucho

(1) Crear una variable con el incremento del pulso observado en el experimento transcurrido el minuto en el que han corrido los que obtuvieron cara. Incluir esta nueva variable en la tabla de datos pulso.

(2) Analizar los efectos del factor de actividad física practicada habitualmente y del bloque si es fumador o no sobre las diferencias registradas del incremento del pulso en el experimento.

(3) Discutir la posible influencia de la interacción de ambos factores (actividad y fumar) en el incremento medio del pulso.

Ejercicio 2.8 *A partir del fichero pulso analizado en el ejercicio anterior:*

(1) Analizar los efectos principales de los diferentes tipos de factores contemplados en el experimento sobre el incremento del pulso: si corrió un minuto o no (ran), si fuma o no (smokes), si es hombre o mujer (sex) y grado de actividad habitual (activity).

(2) Evaluar las diferencias en el aumento del pulso en relación a los efectos principales y cruces entre los factores del experimento.

(3) Discutir si puede identificarse algún grupo o grupos, mediante los factores registrados, con peor condiciones en cuanto al ritmo cardiaco medio.

Referencias

Bibliografía

- Draper, N.R.; Smith, H. (1998). Applied Regression Analysis, 3rd. John Wiley.
- Everitt, B.S.; Hothorn, T. (2010). A Handbook of Statistical Analysis Using R. Chapman Hall.
- Faraway, J. (2004). Linear Models with R. CRC Press.

- Faraway, J. (2005). Extending the Linear Model with R. CRC Press.
- García Pérez, A. (2008). Estadística aplicada con R. UNED.
- González Ortiz, F.J. (2007). Prácticas de Estadística con R (Parte I y Parte II). Universidad de Cantabria.
- Peña, D. (2002). Análisis de Datos Multivariantes. McGraw-Hill.

Recursos en Internet

- An Introduction to R: http://cran.r-project.org/doc/manuals/R-intro.pdf
- Comunidad R hispano: http://www.r-es.org/
- Curso introducción R: http://www.uv.es/conesa/CursoR/cursoR.html
- icebreakeR: http://www.ms.unimelb.edu.au/ andrewpr/r-users/icebreakeR.pdf
- Introduction to Data Technologies: http://www.stat.auckland.ac.nz/ paul/ItDT/itdt-2010-11-01.pdf
- Practical Regression and ANOVA in R: on CRAN, Faraway, J.
- Quick-R: http://www.statmethods.net/
- RStudio: http://rstudio.org/