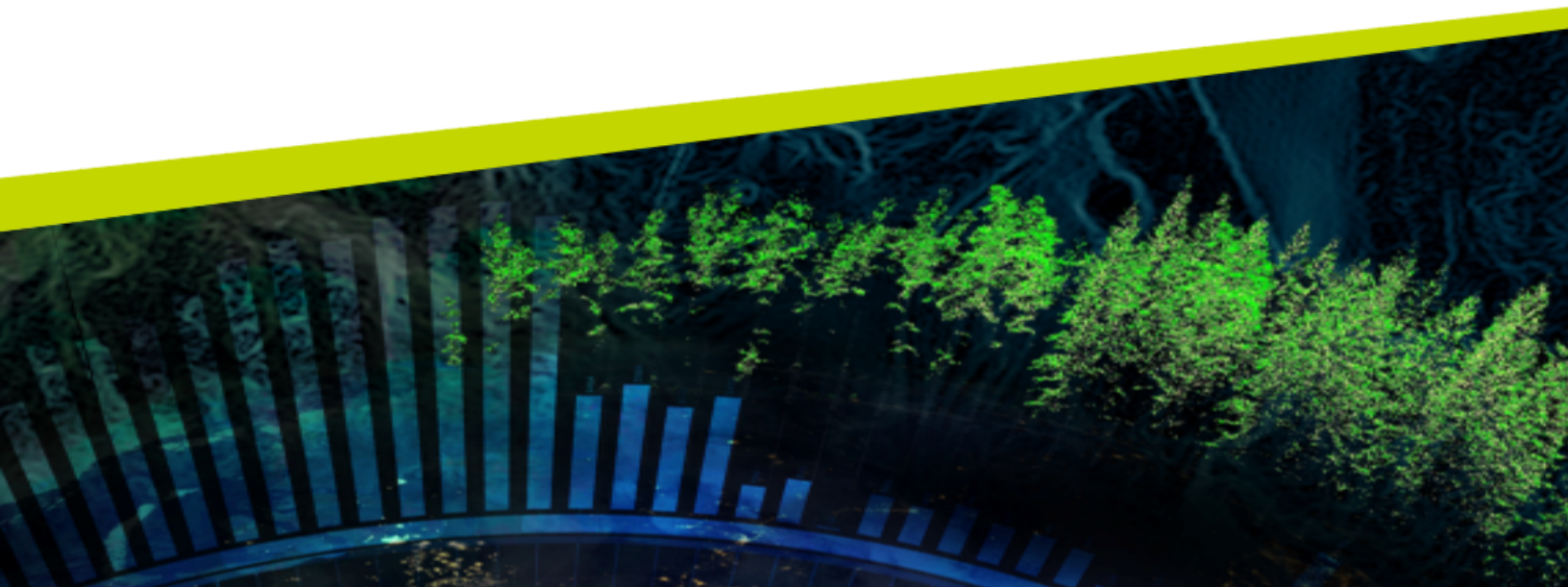
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 9. ANOVA DE UNA VÍA PARA MUESTRAS INDEPENDIENTES

En el capítulo 5 conocimos la prueba t de Student que permite, entre otras funciones, inferir acerca de la diferencia entre las medias de dos poblaciones a partir de dos muestras. Sin embargo, muchas veces necesitaremos realizar un procedimiento similar para $k \geq 3$ grupos (tratamientos). En el capítulo 8 nos enfrentamos a un escenario similar para cuando trabajamos con proporciones, para lo cual conocimos las pruebas chi-cuadrado de Pearson y Q de Cochran. Aprendimos que estas son pruebas de tipo **ómnibus**: es decir, comprueba la igualdad de todos los tratamientos globalmente, pero en el caso de encontrar diferencias significativas, no nos indica dónde están. En consecuencia, conocimos los **procedimientos post-hoc** que permiten identificar entre qué grupos existen estas diferencias, que además consideran **ajustes** en los valores p para evitar inflar la tasa de errores de tipo I o o la tasa de falsos positivos.

En el caso de las medias, cuando tenemos más de dos grupos también podemos usar una prueba ómnibus y algunos procedimientos post-hoc, para lo cual nos basaremos en las ideas presentadas por Lowry (1999, caps. 13-14); Glen (2021); IBM (1989); Meier (2021, p. 4) y Berman (2000).

9.1 SURGIMIENTO

Cuando se necesita inferir sobre las medias de múltiples **muestras independientes**, intuitivamente podríamos recurrir a la idea de aplicar pruebas t de Student para diferencia de medias a cada par de grupos con un nivel de significación α.

Como se discutió en la sección 8.3, esta idea trae problemas. Para ilustrar, con tres tratamientos se tendrían que comparar tres pares, por lo que la probabilidad de no cometer un error de tipo I en estas tres pruebas sería $(1 - alpha)^3$. En consecuencia, la probabilidad de declarar que existen diferencias significativas entre las medias de los grupos sería $(1 - (1 - alpha)^3)$[1] Si, por ejemplo, nominalmente $\alpha = 0{,}05$, tasa de errores de tipo I se vería inflada a aproximadamente 0,143.

El famosos estadístico Sir Ronald A. Fisher, al verse enfrentado a esta dificultad para analizar datos provenientes de experimentos agrícolas a principios del siglo XX, diseñó el método de **análisis de varianza**, comúnmente conocido como **ANOVA** o AoV (del inglés *analysis of variance*), que permite evaluar la hipótesis nula ómnibus (no hay diferencia entre las medias de los grupos). Si bien Fisher propuso el procedimiento en 1918, su popularidad vino al ser incluido en su libro de 1925, por lo esta prueba ya tiene ¡más de 100 años!

De manera análoga, existe el procedimiento ANOVA para muestras correlacionadas (que se aborda en el capítulo siguiente), semejante a la prueba t de Student con muestras apareadas. Los procedimientos ANOVA para muestras independientes y muestras correlacionadas que estudiaremos corresponden al **análisis de varianza de una vía**, pues solo consideran una única variable independiente (de tipo categórica, un **factor**) cuyos niveles definen los grupos (o tratamientos) que se están comparando.

Existe además el **análisis de varianza de dos vías**, no abordado en el presente texto, el cual permite examinar simultáneamente los efectos de dos variables independientes e, incluso, determinar si ambas **interactúan**. De hecho, existen métodos para el análisis con más factores, que también están fuera del alcance de este curso.

Para explicar en detalle el procedimiento ANOVA de una vía para muestras independientes, consideremos el siguiente ejemplo: una ingeniera se enfrenta a un problema de logística en una empresa de reparto de último kilómetro, el que ha logrado modelar como instancias del problema de la mochila 0-1 con capacidad 1.000 y objetos con pesos que se distribuyen según $\mathcal{N}(50, 16)$. Buscando en la literatura, ella ha conseguido implementar tres algoritmos (A, B y C) propuestos para resolver el problema de la mochila 0-1, y desea

[1] Si las pruebas fueran independientes, que no lo son.

comparar su eficiencia. Para cada algoritmo generó, de forma aleatoria, cinco instancias, cada una con 100 elementos. Los tiempo de ejecución registrados (en milisegundos) fueron los siguientes:

- Algoritmo A: 23, 19, 25, 23, 20
- Algoritmo B: 26, 24, 28, 23, 29
- Algoritmo C: 19, 24, 20, 21, 17

De esta forma, la ingeniera necesita contrastar las siguientes hipótesis:

H_0 : el tiempo de ejecución promedio necesitado para resolver instancias del problema enfrentado es igual para los tres algoritmos implementados. Matemáticamente: si denotamos como $\{t_{\text{A}}\}, \{t_{\text{B}}\}, \{t_{\text{C}}\}$ a los tiempos requeridos respectivamente por los algoritmos A, B y C en resolver instancias del problema de la mochila 0-1 con capacidad 1.000 y objetos con pesos $\sim \mathcal{N}(50, 16)$, entonces $\mu_{t_{\text{A}}} = \mu_{t_{\text{B}}} = \mu_{t_{\text{C}}}$.

H_A : el tiempo de ejecución promedio necesitado para resolver instancias del problema enfrentado de tamaño 100 es diferente para al menos un algoritmo. Matemáticamente: $\exists\, i, j \in \{\text{A}, \text{B}, \text{C}\}, i \neq j \mid \mu_{t_i} \neq \mu_{t_j}$.

Puede verse que la hipótesis nula de la prueba ANOVA para muestras independientes se refiere a la igualdad de todas las medias (una hipótesis ómnibus).

9.2 CONDICIONES

Al igual que otras pruebas estudiadas en capítulos anteriores, el procedimiento ANOVA requiere que se cumplan algunas condiciones:

1. La escala con que se mide la variable dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las k muestras son obtenidas de manera aleatoria e independiente desde la(s) población(es) de origen.
3. Se puede suponer razonablemente que la(s) población(es) de origen sigue(n) una distribución normal.
4. Si las muestras provienen de más de una población, estas poblaciones tienen la misma varianza.

En nuestro ejemplo con los algoritmos, la primera condición se verifica, puesto que si para una instancia i un algoritmo tarda 20 [ms] mientras que otro algoritmo tarda 30 [ms], esa es la misma diferencia (10 milisegundos) que se presenta para una instancia j en que uno tarda 35 [ms] y el otro 45 [ms]. A su vez, el enunciado señala que el proceso seguido por la ingeniera garantiza el cumplimiento de la segunda condición.

La figura 9.1 (creada mediante el script 9.1, líneas 20–29) muestra gráficos Q-Q para cada muestra. Como se observa un par de valores en la muestra C que podrían están un poco apartados de la línea teórica y, sobretodo porque, las muestras son pequeñas, es mejor que procedamos con cautela y usemos un nivel de significación $\alpha = 0{,}025$.

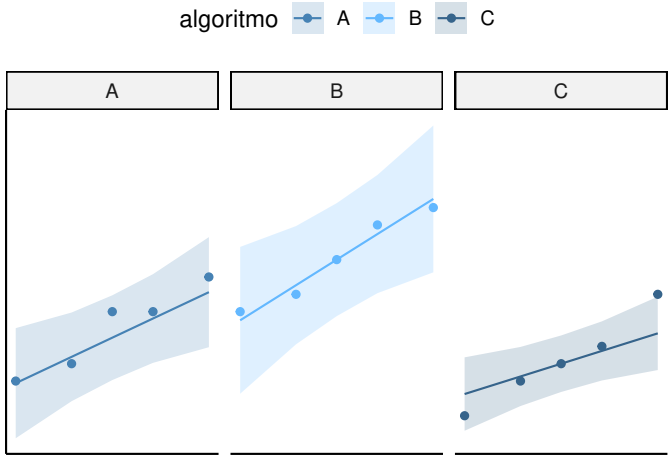


Figura 9.1: gráfico para comprobar el supuesto de normalidad en las tres muestras del ejemplo.

Una regla sencilla para comprobar la cuarta condición, llamada **homogeneidad de las varianzas** o también **homocedasticidad**, es comprobar que la razón entre la máxima y la mínima varianza muestral de los grupos no es superior a 1,5. Comencemos obteniendo las medias de cada grupo:

$$\begin{aligned} \overline{x}_{\text{A}} &= \frac{23+19+25+23+20}{5} = 22{,}0 \\ \overline{x}_{\text{B}} &= \frac{26+24+28+23+29}{5} = 26{,}0 \\ \overline{x}_{\text{C}} &= \frac{19+24+20+21+17}{5} = 20{,}2 \end{aligned} \tag{9.1}$$

Ahora, podemos obtener las varianzas muestrales:

$$\begin{aligned} s^2_{\text{A}} &= \frac{(23-22{,}0)^2+(19-22{,}0)^2+(25-22{,}0)^2+(23-22{,}0)^2+(20-22{,}0)^2}{5-1} &&= \frac{24{,}0}{4} = 6{,}0 \\ s^2_{\text{B}} &= \frac{(26-26{,}0)^2+(24-26{,}0)^2+(28-26{,}0)^2+(23-26{,}0)^2+(29-26{,}0)^2}{5-1} &&= \frac{26{,}0}{4} = 6{,}5 \\ s^2_{\text{C}} &= \frac{(19-20{,}2)^2+(24-20{,}2)^2+(20-20{,}2)^2+(21-20{,}2)^2+(17-20{,}2)^2}{5-1} &&= \frac{26{,}8}{4} = 6{,}7 \end{aligned} \tag{9.2}$$

Así, en el ejemplo la muestra obtenida para el algoritmo A tiene la menor varianza, mientras que la muestra del algoritmo C tiene la mayor. La razón de estas varianzas sería:

$$\frac{s^2_{\text{C}}}{s^2_{\text{A}}} = \frac{6{,}7}{6{,}0} \approx 1{,}117 \tag{9.3}$$

En consecuencia, aplicando este criterio simple, podemos considerar que la condición de homocedasticidad se verifica para el ejemplo.

Script 9.1: verificación gráfica de la condición de normalidad para un procedimiento ANOVA de una vía para muestras independientes.

```r
library(DescTools)
library(ez)
library(ggpubr)
library(tidyverse)

# Crear la matriz de datos en formato ancho
A <- c(23, 19, 25, 23, 20)
B <- c(26, 24, 28, 23, 29)
C <- c(19, 24, 20, 21, 17)
datos_anchos <- data.frame(A, B, C)

# Llevar la matriz a formato largo
datos_largos <- datos_anchos |> pivot_longer(c("A", "B", "C"),
                                             names_to = "Algoritmo",
                                             values_to = "Tiempo")
datos_largos[["Algoritmo"]] <- factor(datos_largos[["Algoritmo"]])
datos_largos[["Instancia"]] <- factor(1:nrow(datos_largos))

# Comprobar la condición de normalidad con gráficos Q-Q
g <- ggqqplot(datos_largos, x = "Tiempo", y = "Algoritmo", color = "Algoritmo",
              palette = c("steelblue", "steelblue1", "steelblue4"))
g <- g + facet_wrap(~ Algoritmo)
g <- g + rremove("x.ticks") + rremove("x.text")
g <- g + rremove("y.ticks") + rremove("y.text")
g <- g + rremove("axis.title")
print(g)
```

Se ha encontrado que ANOVA es una prueba **robusta**, que resiste razonablemente bien cierta desviaciones en las condiciones de normalidad o de homocedasticidad, especialmente cuando las muestras tienen el mismo tamaño. Pero estas suposiciones sí están detrás de la lógica y matemática de la prueba, por lo que **no debemos ignorar** violaciones importantes a estas condiciones.

9.3 VARIABILIDAD

Como su nombre indica, ANOVA se centra en la **variabilidad** observada en los datos, una generalización de la varianza, en base a la suma de los cuadrados de las desviaciones, más conocida como **suma de cuadrados** y usualmente denotada *SS* del inglés *sum of squares*. La fórmula genérica es mostrada en la ecuación 9.4, donde n corresponde al número de observaciones consideradas $(x_1, x_2, \ldots, x_n)$, y $\overline{x}$ a su media.

$$SS = \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{9.4}$$

Combinando las observaciones de las muestras correspondientes a los diferentes grupos, se puede calcular la media combinada, también llamada media global o **media total**, denotada $\overline{x}_T$ Luego, usando la ecuación 9.4, se puede calcular la variabilidad total SS_T considerando todas las observaciones. Para el ejemplo, esto sería:

$$\begin{aligned}\overline{x}_T &= \frac{23+19+25+23+20+26+24+28+23+29+19+24+20+21+17}{15} \\ &\approx 22{,}733\end{aligned}$$

$$\begin{aligned}SS_T \approx\ & (23-22{,}733)^2 + (19-22{,}733)^2 + (25-22{,}733)^2 + (23-22{,}733)^2 + (20-22{,}733)^2 + \\ & (26-22{,}733)^2 + (24-22{,}733)^2 + (28-22{,}733)^2 + (23-22{,}733)^2 + (29-22{,}733)^2 + \\ & (19-22{,}733)^2 + (24-22{,}733)^2 + (20-22{,}733)^2 + (21-22{,}733)^2 + (17-22{,}733)^2 \\ \approx\ & 164{,}933\end{aligned}$$

La variabilidad total puede descomponerse en dos partes: una de ellas corresponde a la variabilidad existente al interior de cada uno de los grupos llamada **variabilidad intra-grupos** y usualmente denotada SS_{wg} del inglés *within groups*; la otra corresponde a la **variabilidad entre grupos**, denotada como SS_{bg} del inglés *between groups*. La ecuación 9.5 muestra una **identidad importante** que relaciona ambas componentes:

$$SS_T = SS_{bg} + SS_{wg} \tag{9.5}$$

La variabilidad entre grupos mide de manera agregada la magnitud de las diferencias entre las distintas medias muestrales. Si se tienen k grupos o tratamientos, y cada i-ésimo grupo cuenta con n_i observaciones en su muestra, con media $\overline{x}_i$, entonces la variabilidad entre grupos se calcula como muestra la ecuación 9.6:

$$SS_{bg} = \sum_{i=1}^{k} n_i \left(\overline{x}_i - \overline{x}_T\right)^2 \tag{9.6}$$

Esta medida entonces corresponde a la suma de cuadrados de las medias de cada grupo con respecto a la media total, donde cada desviación cuadrada se pondera por la cantidad de observaciones que incluye, a fin de mantener la representatividad de cada grupo. Así, mide el grado en que los grupos difieren unos de otros.

Usando los valores calculados en la ecuación 9.1, para el ejemplo tenemos:

$$\begin{aligned}SS_{bg} &\approx 5 \cdot (22{,}0 - 22{,}733)^2 + 5 \cdot (26{,}0 - 22{,}733)^2 + 5 \cdot (20{,}2 - 22{,}733)^2 \\ &\approx 88{,}133\end{aligned}$$

La variabilidad intra-grupos se calcula de acuerdo a la ecuación 9.7:

$$SS_{wg} = \sum_{i=1}^{k} SS_i \tag{9.7}$$

donde SS_i corresponde a la variabilidad del i-ésimo grupo, calculada mediante la ecuación 9.4.

Luego, esta medida corresponde al total de las sumas de cuadrados al interior de cada grupo, por lo que representa una medida global de la variabilidad aleatoria (natural) de los diferentes grupos. Para el ejemplo, podemos reconocer que la suma de cuadrados de cada grupo corresponden a los numeradores calculados en la ecuación 9.2, por lo que el valor sería:

$$\begin{aligned} SS_{wg} &= SS_{\text{A}} + SS_{\text{B}} + SS_{\text{C}} \\ &= 24{,}0 + 26{,}0 + 26{,}8 \\ &= 76{,}8 \end{aligned}$$

9.4 EL ESTADÍSTICO DE PRUEBA F

En el cálculo de las varianzas muestrales para el ejemplo (ecuación 9.2) utilizamos la ecuación 2.2, que reproducimos aquí:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\overline{x}_i - \overline{x})^2$$

Si nos fijamos, esto corresponde a la suma de cuadrados normalizada por sus grados de libertad (que, recordemos, denotamos ν). Es decir, la varianza muestral corresponde a la desviación cuadrada media de la muestra.

Podemos formalizar este estadístico, que llamaremos **cuadrado medio** y denotaremos MS, del inglés *mean square*, por medio de la ecuación 9.8:

$$MS = \frac{SS}{\nu} \tag{9.8}$$

En el contexto de análisis de varianza, para el caso de la variabilidad entre grupos, se tienen $\nu_{bg} = k - 1$ grados de libertad, donde k corresponde a la cantidad de grupos o tratamientos. Con ello, el cuadrado medio entre grupos queda dada por la ecuación 9.9:

$$MS_{bg} = \frac{SS_{bg}}{\nu_{bg}} = \frac{SS_{bg}}{k-1} \tag{9.9}$$

De manera similar, los grados de libertad para la variabilidad al interior de los grupos está dada por la suma de los grados de libertad en cada grupo, por lo que su cuadrado medio queda dado por la ecuación 9.10:

$$MS_{wg} = \frac{SS_{wg}}{\nu_{wg}} = \frac{SS_{wg}}{\sum_{i=1}^{k}(n_i - 1)} \tag{9.10}$$

En ocasiones resulta útil conocer también la cantidad total de grados de libertad, que puede obtenerse mediante la ecuación 9.11, donde $n_{_T}$ es el tamaño de la muestra combinada. Podemos notar que los grados de libertad cumplen la misma relación de descomposición que las sumas de cuadrados.

$$\nu_{_T} = n_{_T} - 1 = \nu_{bg} + \nu_{wg} \tag{9.11}$$

Para el ejemplo, entonces, se tiene que $\nu_{bg} = 3 - 1 = 2$ y $\nu_{wg} = (5-1) + (5-1) + (5-1) = 12$, luego:

$$\begin{aligned} MS_{bg} &= \frac{88{,}133}{2} \approx 44{,}066 \\ MS_{wg} &= \frac{8{,}12}{2} \approx 6{,}4 \end{aligned}$$

Si bien la relación entre MS_{bg} y MS_{wg} es compleja, en general se cumple que:

- Si la hipótesis nula es verdadera (igualdad de las medias), MS_{bg} tiende a ser menor o igual que MS_{wg}
- Si la hipótesis nula es falsa, MS_{bg} tiende a ser mayor que MS_{wg}.

Representamos esta relación mediante la **razón** F, que es el estadístico de prueba para ANOVA, que se calcula como muestra la ecuación 9.12, donde MS_{efecto} corresponde a una estimación de la varianza del efecto que se desea medir y MS_{error}, a la variabilidad natural (puramente aleatoria) asociada al fenómeno. Así, si efectivamente las medias de los grupos son iguales, se esperaría observar un estadlisto $F \approx 1$.

$$F = \frac{MS_{\text{efecto}}}{MS_{\text{error}}} \tag{9.12}$$

En el ejemplo queremos estudiar si existe diferencia entre las medias de los grupos, por lo que $MS_{\text{efecto}} = MS_{bg}$. Asimismo, la variabilidad aleatoria está dada por la variabilidad al interior de los grupos, por lo que $MS_{\text{error}} = MS_{wg}$. Así, en este caso:

$$F = \frac{44{,}066}{6{,}4} \approx 6{,}885$$

Vemos que el estadístico de prueba es bastante alto. De manera similar a lo que hemos visto en otras pruebas, podemos obtener un valor p para este valor que corresponde al área bajo la cola superior de la distribución F mayor o igual al estadístico F obtenido.

De esta forma para el ejemplo, que cuenta con 2 y 12 grados de libertad, el valor p puede calcularse en R mediante la llamada `pf(6.885, 2, 12, lower.tail = FALSE)`, obteniéndose $p \approx 0{,}010$.

9.5 RESULTADO

El resultado del procedimiento ANOVA suele representarse en una forma tabular que es bastante estándar. Para el ejemplo, el resultado se organiza como muestra la tabla 9.1.

Fuente	ν	SS	MS	F	p
Entre grupos (efecto)	2	88,133	44,067	6,885	0,010
Intra-grupos (error)	12	76,800	6,400		
TOTAL	14	164,933			

Tabla 9.1: resultado del procedimiento ANOVA.

Como es usual, la conclusión de esta prueba se efectúa comparando el valor p con el nivel de significación. En el ejemplo, $\alpha = 0{,}025$ y $p < \alpha$, por lo que rechazamos la hipótesis nula en favor de la hipótesis alternativa. En consecuencia, debemos concluir con 97,5 % de confianza que el tiempo de ejecución promedio **es diferente** para al menos uno de los algoritmos comparados.

Una observación importante que debemos tener en cuenta es que, si usamos ANOVA para casos con **solo dos grupos** (en su correspondientes versiones pareada o independiente), los resultados son equivalentes a los que obtendríamos con una **prueba t de Student**, y el estadístico F sería igual al cuadrado del estadístico t. No obstante, la prueba t puede ser unilateral o bilateral, mientras que el análisis de varianza es intrínsecamente unidireccional, pues la distribución F solo está definida para valores no negativos.

9.6 PROCEDIMIENTO EN R

Desde luego, R nos ofrece (muchas) funciones para realizar la prueba ANOVA de una vía para muestras independientes. La primera alternativa que conoceremos es la función `aov(formula, data)` (del paquete base), donde:

- `formula`: se escribe de la forma `<variable_dependiente>~<variable_independiente>`.

- `data`: `data.frame` que contiene las variables especificadas en la fórmula.

Se deben hacer tres observaciones. Primero, la función `aov()` está diseñada para grupos balanceados (del mismo tamaño); si esto no se cumple, los resultados que entrega pueden interpretarse erróneamente. Segundo, la fórmula que se le entrega como argumento permite obtener procedimientos ANOVA de distintos "sabores" (con diferentes diseños lineales), incluyendo ANOVA con más de un factor (dos o más vías). Para la prueba estudiada en las secciones anteriores, la fórmula debe ser como se acaba de especificar, con un único factor (variable independiente). Y tercero, la función retorna un objeto, que para este caso es de clase `"aov"`, que guarda toda la información del procedimiento ANOVA aplicado, para que luego pueda ser consultado, resumido en pantalla, graficado, etc.

El script 9.2 muestra la parte del código, a continuación del mostrado en el script 9.1, que aplica ANOVA de una vía para muestras independientes usando esta función con los datos del ejemplo que hemos estado utilizando. La figura 9.2 muestra la salida que se obtiene ejecutando esta parte del script, donde podemos apreciar que sigue la estructura tabular presentada en la tabla 9.1 y que los resultados son los mimos del cálculo manual excepto por algunos cambios en el redondeo de los valores reportados.

```
Procedimiento ANOVA usando aov
------------------------------
           Df Sum Sq Mean Sq F value Pr(>F)
Algoritmo   2  88.13   44.07   6.885 0.0102 *
Residuals  12  76.80    6.40
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 9.2: resultado obtenido con la función `aov()`.

Debemos notar que este resumen tabular se obtiene aplicando la función base `summary(object)` al objeto `"aov"` devuelto por la función `aov()`. También que este resumen usa el nombre de la variable independiente (el factor) en la matriz de datos al reportar el efecto entre los grupos ("Algoritmo" en el ejemplo), y que para referirse al error intra-grupos usa el término "Residuos" (más bien `Residuals`), que es más común en el contexto de regresiones lineales. Esto se debe a que, nuevamente usando una descripción en palabras muy simples y omitiendo un montón de detalles, una prueba ANOVA de una vía para muestras independientes puede entender como una regresión lineal entre una variable dependiente numérica (el tiempo de ejecución en el ejemplo) y una variable independiente categórica (el algoritmo en el ejemplo).

Script 9.2: (continuación del script 9.1) procedimiento ANOVA de una vía para muestras independientes usando la función `aov()` de R.

```r
# Realizar y mostrar el procedimiento ANOVA con aov()
prueba <- aov(Tiempo ~ Algoritmo, data = datos_largos)
cat("Procedimiento ANOVA usando aov\n")
cat("------------------------------\n")
print(summary(prueba))
```

Otra opción es usar la función `ezANOVA(data, dv, wid, between, return_aov)` del paquete `ez`, donde:

- `data`: `data.frame` con los datos.
- `dv`: variable dependiente (columna en la matriz de datos).
- `wid`: variable con el identificador de cada instancia (un factor en la matriz de datos).
- `between`: variable independiente (factor en la matriz de datos que identifica los grupos).
- `return_aov`: si es verdadero, devuelve un objeto de tipo `"aov"` para uso posterior (toma valor `FALSE` por omisión).

El script 9.3 muestra la llamada a la función `ezANOVA()`, a continuación de los scripts anteriores aplicada al ejemplo de los algoritmos.

La figura 9.3 muestra la salida que entrega esta llamada. Se observa que es un tanto distinta a la que entrega la función `aov()`. Primero, la función da un mensaje de advertencia que indica que está utilizando la función

`hccm()` del paquete `car`, que intenta ajustar las varianzas a violaciones de homogeneidad de la varianza. Si bien el mensaje es molesto, esto es normal.

```
Coefficient covariances computed by hccm()

Procedimiento ANOVA usando ezANOVA
----------------------------------
$ANOVA
     Effect DFn DFd        F          p p<.05       ges
1 Algoritmo   2  12 6.885417 0.01019334     * 0.5343573

$`Levene's Test for Homogeneity of Variance`
  DFn DFd       SSn  SSd          F        p p<.05
1   2  12 0.1333333 29.6 0.02702703 0.973394

$aov
Call:
   aov(formula = formula(aov_formula), data = data)

Terms:
                Algoritmo Residuals
Sum of Squares   88.13333  76.80000
Deg. of Freedom         2        12

Residual standard error: 2.529822
Estimated effects may be unbalanced
```

Figura 9.3: resultado obtenido con la función `ezANOVA()`.

Script 9.3: (continuación del script 9.2) procedimiento ANOVA de una vía para muestras independientes usando la función `ezANOVA()` de R.

```
34 # Realizar y mostrar el procedimiento ANOVA con ezANOVA()
35 prueba2 <- ezANOVA(data = datos_largos, dv = Tiempo, between = Algoritmo,
36                    wid = Instancia, return_aov = TRUE)
37 cat("\nProcedimiento ANOVA usando ezANOVA\n")
38 cat("----------------------------------\n")
39 print(prueba2)
```

Luego, bajo el título `$ANOVA`, vemos que solo se reporta el efecto entre grupos, que es la misma información obtenida anteriormente (con más decimales), y se omite la información sobre el error aleatorio.

A continuación, la función `ezANOVA()` nos reporta el resultado de una **prueba de Levene de homogeneidad de varianza** (NIST/SEMATECH, 2013). Si bien no estudiaremos esta prueba en detalle, es pertinente mencionar que es una prueba robusta a desviaciones de normalidad que sirve para verificar si la condición de homocedasticidad se cumple para poder aplicar el procedimiento ANOVA de una vía para muestras independientes. Las hipótesis detrás de esta prueba son:

H_0: las varianzas de las k poblaciones desde donde se obtuvieron las muestras son iguales. Matemáticamente, $\sigma_1^2 = \sigma_2^2 = \cdots \sigma_k^2$.

H_A: al menos una de las poblaciones de origen tiene una varianza diferente a alguna de las otras poblaciones. Es decir, $\exists\, i, j \in \{1, 2, \ldots, k\}, i \neq j \mid \sigma_i^2 \neq \sigma_j^2$.

Podemos notar que la hipótesis nula de la prueba de Levene es de tipo ómnibus. En la figura 9.3 vemos que esta prueba no resultó significativa al nivel de significación $\alpha = 0{,}025$ ($F_{2,12} = 0{,}027$; $p = 0{,}973$), por lo que no podemos rechazar la hipótesis nula y, en consecuencia, concluir que no existe evidencia para dudar que la condición de homocedasticidad se está cumpliendo.

Como la función fue llamada indicando `return_aov = TRUE`, en este caso se reporta la existencia de un objeto de tipo `"aov"` con los resultados de la prueba, que puede ser utilizada posteriormente, indicando las sumas

de cuadrados y los grados de libertad entre e intra-grupos (`Residuals`), además de la desviación estándar observada en los residuos.

Una nota importante para **el futuro**: un argumento adicional de la función `ezANOVA()` que no hemos mencionado es `type`, el cual no estudiaremos en detalle porque escapa a los alcances de este libro, que indica qué **tipo de suma de cuadrados** la función debe realizar. Existen tres posibilidades en esta implementación: valor 1 para usar el tipo secuencial, valor 2 para el tipo marginal sin interacciones, o el valor 3 para que se usen sumas marginales considerando las interacciones.

Cuando se realiza un procedimiento ANOVA de una sola vía, el tipo de *SS* es irrelevante, pues la varianza no debe ser repartida entre varios factores, por lo que se obtienen los mismos valores con cualquiera de ellos. Pero si existen dos o más vías, entonces el tipo 1 normalmente no es adecuado, el tipo 2 puede usarse en ciertos casos y el tipo 3 suele ser el más conveniente, aunque no está exento de críticas. Por omisión, `type` toma el valor 2 que entrega el mismo valor que el tipo 1 cuando los grupos están balanceados, pero no es apropiado si hay interacciones entre los factores (resultados cambiantes) o los grupos no están balanceados.

La función `aov()` usa el tipo 1 y no funciona (da error) si los grupos no tienen el mismo tamaño, por lo que suele no ser de utilidad para modelos más complejos. Luego, para datos desbalanceados o varios factores, solo podemos usar `ezANOVA()` (y otras funciones más completas en otros paquetes).

9.7 TAMAÑO DEL EFECTO

Una vez que tenemos una medida de la varianza aleatoria que engloba a todos los grupos, es decir MS_{wg}, podemos seguir los procedimientos estudiados en el capítulo 4 para obtener intervalos de confianza para cada grupo. Recordemos que para las medias:

- El error estándar es la raíz cuadrada de la varianza dividida por el tamaño de la muestra.
- Usando el error estándar, podemos obtener un margen de error multiplicándolo por el quantil determinado por el nivel de significación.
- El intervalo de confianza correspondiente se obtiene sumando y/o restando el margen de error sumando y restando el margen de error a la media muestral.

Extendiendo estos principios al caso de ANOVA de una vía para muestras independientes, para intervalos de confianza bilaterales, podemos escribir:

$$\mathrm{IC}_{(1-\alpha)} = \overline{x}_i \pm t^*_{1-\alpha/2} \sqrt{\frac{MS_{wg}}{n_T}}$$

Para el ejemplo, tendríamos que el error estándar es $\sqrt{6{,}4/15} \approx 0{,}653$ y $t^*_{0{,}9875} \approx 2{,}56$, con lo que el margen de error sería aproximadamente 1,672. Así, los intervalos con 97,5 % confianza para los algoritmos A, B y C serían, respectivamente, [20,3; 23,7], [24,3; 27,7] y [18,5; 21,9].

El paquete `ez` también proporciona la función `ezPlot(data, dv, wid, between, x)`, la cual nos permite ver gráficamente estos intervalos de confianza. En general, los argumentos son los mismos que para `ezANOVA()`, con la salvedad del nuevo argumento `x`, que señala la variable que va en el eje horizontal del gráfico. La función tiene varios argumentos opcionales para controlar algunos parámetros del gráfico, entre otras cosas. Sin embargo, si se quiere cambiar los colores de sus elementos, hay que cambiarlos "a mano" en el gráfico que devuelve la función, como se muestra en el script 9.4. El gráfico que genera este trozo de código para el ejemplo que hemos seguido es presentado en la figura 9.4.

El propósito de mostrar los intervalos de confianza en un gráfico como el la figura 9.4 es proporcionar una representación visual del **tamaño del efecto** encontrado en el análisis.

Para el ejemplo, se puede observar a simple vista que el intervalo de confianza de los tiempos de ejecución del algoritmo B está más arriba que los intervalos de los algoritmos A y C, por lo que parecen ser **significativamente peores**. También se podría notar que los tiempos del algoritmo A **parecieran** más altos que los algoritmo C, pero no es tan claro si esta diferencia es real o debida al azar (es decir, si la diferencia es estadísticamente significativa o no). Obviamente este tipo de evaluación visual se complica cuando se tienen factores con más niveles que para este ejemplo.

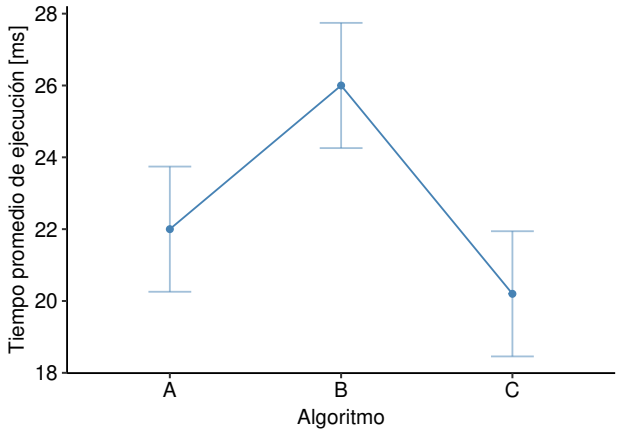


Figura 9.4: tamaño del efecto encontrado en el análisis de varianza del ejemplo.

Script 9.4: (continuación del script 9.3) procedimiento ANOVA de una vía para muestras independientes usando la función `ezANOVA()` de R.

```
# Obtener el gráfico del tamaño del efecto
g2 <- ezPlot(data = datos_largos, dv = Tiempo,
             wid = Instancia, between = Algoritmo,
             y_lab = "Tiempo promedio de ejecución [ms]", x = Algoritmo)
# Cambiar sus colores
g2 <- g2 + theme_pubr()
g2[["layers"]][[1]][["aes_params"]][["colour"]] <- "steelblue"
g2[["layers"]][[2]][["aes_params"]][["colour"]] <- "steelblue"
g2[["layers"]][[3]][["aes_params"]][["colour"]] <- "steelblue"
# Y mostrar el gráfico
print(g2)
```

Aquí vale la pena mencionar otro dato importante que nos reporta la función `ezANOVA()` para el efecto entre grupos, que se encuentra bajo el título `ges` en la figura 9.3. Este acrónimo significa *generalized eta squared*, denotada η^2_G en los libros, que corresponde a una medida estandarizada del tamaño del efecto que resumen en un único valor las diferencias encontradas en pruebas ANOVA, que fue propuesta por Olejnik y Algina (2003).

Si bien existen múltiples medidas estandarizadas con este fin, como η^2 (eta cuadrado) y η^2_p (eta cuadrado parcial), todas entregan el mismo resultado para procedimientos ANOVA de una vía (ecuación 9.13), pero η^2_G permite comparaciones más justas con procedimientos más complejos o con muestras correlacionadas (Bakeman, 2005).

$$\eta^2_G = \eta^2 = \eta^2_p = \frac{SS_{bg}}{SS_T} \tag{9.13}$$

En nuestro ejemplo, $\eta^2_G = 0{,}534$. Si se considera las recomendaciones de Cohen (1992), deberíamos considerar valores mayores a 0,01, 0,06 y 0,14 como efectos pequeños, medianos y grandes, respectivamente. Así podemos concluir estamos observando un efecto bastante grande en el ejemplo.

Debemos mencionar que las medidas estandarizadas basadas en η^2 están siendo criticadas en los últimos años, porque no son estimadores insesgados y tienden a sobrestimar el tamaño del efecto en la población (Lakens, 2013). Actualmente están apareciendo llamados a usar otras medidas como ω^2 (ómega cuadrado), ϵ^2 (épsilon cuadrado, también llamado eta cuadrado ajustado) que son mejores estimadores (Kroes & Finley, 2023; Lakens, 2013). Estas medidas son accesibles en R a través de funciones proporcionadas por el paquete `effectsize` (y graficables por medio del paquete `see`), que también permite obtener intervalos de confianza para estas medidas y para las medidas del tamaño del efecto basadas en η^2 (Ben-Shachar et al., 2020).

9.8 POTENCIA ESTADÍSTICA

Como vimos en capítulos anteriores, el poder estadístico de una prueba depende de varios factores: tamaño(s) de la(s) muestra(s), el nivel de significación definido, y el tamaño del efecto que debe detectarse.

La lógica es la misma que estudiamos en el capítulo 7, e implementadas en las funciones del paquete `pwr` de R. Para el procedimiento ANOVA de una vía para muestras independientes, este paquete provee la función `pwr.anova.test(k, n, f, sig.level, power)`, donde:

- `k`: número de grupos.
- `n`: número de observaciones en cada grupo.
- `f`: tamaño del efecto medido con la **f de Cohen**.
- `sig.level`: nivel de significación de la prueba.
- `power`: poder de la prueba.

Notemos, además de que obviamente no hay alternativas bilaterales o menores, que esta función solo funciona cuando los grupos están balanceados.

La f de Cohen es una medida estandarizada del efecto, que puede estimarse como:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \tag{9.14}$$

Que, para el caso de simple de ANOVA de una vía para muestras independientes corresponde a:

$$f = \sqrt{\frac{SS_{bg}}{SS_{wg}}}$$

Así, revisando la tabla 9.1, para el ejemplo de los algoritmos se tiene que $f = \sqrt{88{,}133/76{,}800} \approx 1{,}071$. Entonces, la llamada `pwr.anova.test(k = 3, n = 5, f = 1.071, sig.level = 0.05)` entrega la salida mostrada en la figura 9.5, donde conseguimos una estimación de la potencia de la prueba realizada. Vemos que esta es bastante alta, sobre 90 %, lo que es consistente con un efecto grande como el observado en este ejemplo.

```
     Balanced one-way analysis of variance power calculation

              k = 3
              n = 5
              f = 1.071247
      sig.level = 0.05
          power = 0.9125067

NOTE: n is number in each group
```

Figura 9.5: potencia estadística para ejemplo obtenida con la función `pwr.anova.test(k)`

.

Como estudiamos para otras funciones del paquete `pwr`, dejando sin definir otro de los argumentos de la función (pues por omisión toman valor `NULL`), y definiendo un poder estadístico deseado, podemos estimar otros factores al momento de diseñar el estudio, como el nivel de significación que podríamos conseguir o, más frecuentemente, la cantidad de observaciones requerida en cada grupo.

9.9 PROCEDIMIENTO POST-HOC

Al aplicar el procedimiento ANOVA de una vía para muestras independientes al ejemplo de los algoritmos, pudimos concluir que existe al menos uno de ellos cuyo tiempo promedio de ejecución es significativamente diferente al de otro. Como hemos vistos en otros casos, este resultado ómnibus podría no ser suficiente.

Si suponemos que estos algoritmos tienen por objeto resolver un problema crítico, de cuya rápida solución depende aumentar la productividad de una empresa o prevenir una situación de mucho riesgo, desde luego nos interesaría conocer cuál es el mejor (o el peor) de los algoritmos comparados a fin de poder garantizar un menor tiempo de respuesta. En consecuencia, necesitamos contar con un procedimiento post-hoc para medias independientes que permita determinar si los tiempos medios de ejecución de los algoritmos A y B son los que difieren significativamente, o los de A y C, o bien los de B y C.

Una opción es usar **múltiples pruebas t de Student** para la diferencia de dos medias independientes y luego corregir los valores p que se obtengan aplicando los ajustes generales ya estudiados en la sección 8.3. Para esto, R dispone de la función `pairwise.t.test(x, g, p.adjust.method, pool.sd, paired, alternative, ...)`, donde:

- `x`: vector con la variable dependiente.
- `g`: factor o vector de agrupamiento.
- `p.adjust.method`: señala qué método emplear para ajustar los valores p resultantes.
- `pool.sd`: valor booleano que indica si se usa o no varianza combinada.
- `paired`: valor booleano que indica si las pruebas t son pareadas (verdadero) o no.
- `alternative`: indica si la prueba es bilateral (`"two.sided"`) o unilateral (`"greater"` o `"less"`).
- `...`: argumentos adicionales que se pasan a la función `t.test()` que es llamada internamente.

El script 9.5 muestra la realización de pruebas t para cada par de grupos usando tanto la corrección de Holm como la de Holm, obteniéndose los resultados que se muestran en la figura 9.6. Debemos recordar que el nivel de significación que se entrega como argumento es el mismo que usamos en el procedimiento ANOVA.

Script 9.5: (continuación del script 9.4) procedimiento post-hoc para ANOVA de una vía para muestras independientes usando la función `pairwise.t.test()` de R.

```
# Definir el nivel de significación
alfa <- 0.025

# Realizar y mostrar un procedimiento post-hoc de Holm
holm <- pairwise.t.test(datos_largos[["Tiempo"]], datos_largos[["Algoritmo"]],
                        p.adj = "holm", pool.sd = TRUE, paired = FALSE,
                        conf.level = 1 - alfa)
cat("\n\nProcedimiento post-hoc de Holm\n")
cat("-----------------------------\n")
print(holm)

# Realizar y mostrar un procedimiento post-hoc de Benjamini y Hochberg
bh <- pairwise.t.test(datos_largos[["Tiempo"]], datos_largos[["Algoritmo"]],
                      p.adj = "fdr", pool.sd = TRUE, paired = FALSE,
                      conf.level = 1 - alfa)
cat("\nProcedimiento post-hoc de Benjamini y Hochberg\n")
cat("-----------------------------------------------\n")
print(bh)
```

Los valores p obtenidos con ambos métodos son similares y, en ambos casos, podemos ver que únicamente los algoritmos B y C presentan una diferencia significativa al nivel de significación $\alpha = 0{,}025$. Esto confirma lo que intuitivamente observamos en el gráfico del tamaño del efecto de figura 9.4. Podemos concluir entonces con 97,5 % de confianza que el algoritmo C es más rápido que el algoritmo B.

Pero además de esta opción genérica, existen dos procedimientos post-hoc **especialmente concebidos** para la prueba ANOVA, los que discutiremos a continuación.

9.9.1 Prueba HSD de Tukey

Cuando revisamos el gráfico de la figura 9.4, intuitivamente nos pusimos a comparar las medias e intervalos de confianza encontrados para cada grupo. Esta fue precisamente la **primera técnica** de comparación

```
Procedimiento post-hoc de Holm
------------------------------

 Pairwise comparisons using t tests with pooled SD

data:  datos_largos[["Tiempo"]] and datos_largos[["Algoritmo"]]

  A     B
B 0.056 -
C 0.283 0.010

P value adjustment method: holm

Procedimiento post-hoc de Benjamini y Hochberg
----------------------------------------------

 Pairwise comparisons using t tests with pooled SD

data:  datos_largos[["Tiempo"]] and datos_largos[["Algoritmo"]]

  A     B
B 0.042 -
C 0.283 0.010

P value adjustment method: fdr
```

Figura 9.6: resultados del procedimiento post-hoc para la prueba ANOVA del ejemplo usando los ajustes de Holm y Benjamini y Hochberg.

de pares de medias desarrollada por Ronald Fisher en 1935 (Williams & Abdi, 2010), quien propuso una "diferencia significativa mínima" que debía existir entre las medias de dos grupos para considerarlas distintas. Sin embargo, hoy en día su uso no es recomendado puesto que en realidad no hace una corrección para comparaciones múltiples (Midway et al., 2020).

Se entiende entonces que John Wilder Tukey propusiera, en 1949, un procedimiento basado en una **diferencia honestamente significativa**, al que hoy se le conoce como el método **HSD de Tukey** (del inglés *honestly significant difference*), que está diseñado específicamente para comparar todos los pares de grupos luego de una prueba ANOVA de una vía para muestras independientes significativa.

Este procedimiento controla la tasa de error por familia de hipótesis (FWER, como podremos recordar) cuando se compararan todos los pares de grupos (*pairwise comparisons*) y resulta con **mayor potencia estadística** que los ajustes generales para controlar FWER cuando se cumplen sus supuestos: los grupos tienen la misma cantidad de observaciones y se cumplen las condiciones para realizar la prueba ANOVA (vistas en la sección 9.2). Cuando los grupos no están balanceados, se puede recurrir a la versión de Tukey-Kramer que adapta el procedimiento HSD original para este caso. Ante violaciones de la última condición, relacionadas con incumplimiento de las condiciones de normalidad u homocedasticidad, es más seguro usar uno de los métodos genéricos.

El procedimiento HSD de Tukey emplea el estadístico Q que sigue una distribución de rango estudiantizado[2], también llamada "distribución del rango estandarizado de Student", que para cualquier par de medias (de entre los k grupos) se calcula según la ecuación 9.15:

$$Q = \frac{\overline{x}_g - \overline{x}_p}{\sqrt{\frac{MS_{wg}}{\dot{n}}}} \tag{9.15}$$

donde:

[2]Los detalles de esta distribución escapan a los alcances de este texto.

- $\overline{x}_g$ es la mayor de las dos medias muestrales comparadas.
- $\overline{x}_p$ es la menor de las dos medias muestrales comparadas.
- MS_{wg} corresponde al cuadrado medio intra-grupos (entregada por el procedimiento ANOVA).
- $\dot{\overline{n}}$ es, siguiendo la adaptación de Tukey-Kramer, la **media armónica** de la cantidad de observaciones en cada grupo, obtenida con la fórmula de la ecuación 9.16 (que es igual a la media aritmética si ambas muestras tienen el mismo tamaño, resultando en el tamaño único como es requerido por la idea original de Tukey).

$$\dot{\overline{n}} = \frac{k}{\sum_{i=1}^{k} \frac{1}{\overline{n}}} \tag{9.16}$$

En la práctica, sin embargo, no es necesario calcular el estadístico Q para cada par de medias, sino que basta con conocer el valor crítico de este estadístico para el nivel de significación α establecido, denotado Q_α, el cual depende de la cantidad de grupos (k) y de los grados de libertad del error aleatorio, que en el caso de un procedimiento ANOVA de una vía para muestras independientes corresponde a ν_{wg}.

El valor crítico Q_α nos permite determinar **cuán grande** debe ser la diferencia entre las medias de dos grupos para ser considerada significativa para ese nivel de significación, que es más *honesta* para la cantidad de comparaciones múltiples que se están realizando. El valor crítico se calcula mediante la ecuación 9.17:

$$HSD_\alpha = Q_\alpha \sqrt{\frac{MS_{wg}}{\dot{\overline{n}}}} \tag{9.17}$$

Así, una diferencia entre las medias de dos grupos únicamente es significativa si es mayor o igual que HSD_α.

Veamos cómo estas ideas se aplican a nuestro ejemplo de los algoritmos para el problema de la mochila 0-1. Si bien existen tablas o calculadoras en línea para la distribución de rango estudiantizado, para nosotros es más cómodo usar R, por medio de la función `qtukey(p, nmeans, df, lower.tail = FALSE)`. Para el ejemplo `p` $= \alpha =$ `0.025`, `nmeans` $= \dot{\overline{n}} =$ `5` y `df` $= \nu_{wg} =$ `12`, obteniendo que $Q_\alpha \approx 5{,}080$ y, en consecuencia, $HSD_{0{,}025} = 5{,}080 \cdot \sqrt{6{,}4/5} \approx 5{,}747$.

Recordando las medias calculas para cada grupo en la ecuación 9.1, las diferencias encontradas serían:

$$\begin{aligned} \overline{x}_B - \overline{x}_A &= 26 - 22 \quad = \quad 4 \\ \overline{x}_A - \overline{x}_C &= 22 - 20{,}2 = 1{,}8 \\ \overline{x}_B - \overline{x}_C &= 26 - 20{,}2 = 5{,}8 \end{aligned}$$

De esta forma, la tercera diferencia es la única que supera el valor $HSD_{0{,}025}$ para los datos ejemplo, con lo que solo existe diferencia significativa entre los tiempos promedio de ejecución de los algoritmos B y C, y (como la diferencia calculada es positiva) se puede concluir que el algoritmo C es más rápido que el algoritmo B, lo que se condice con los resultados anteriores.

R también permite realizar la prueba HSD de Tukey de manera sencilla, como se muestra en el script 9.6. La función para ello es `TukeyHSD(x, which, ordered, conf.level)`, donde:

- `x`: un modelo ANOVA (objeto de tipo `"aov"`).
- `which`: string con el nombre de la variable para la que se calculan las diferencias.
- `ordered`: valor lógico que, cuando es verdadero, hace que los grupos se ordenen de acuerdo a sus medias a fin de obtener diferencias positivas.
- `conf.level`: nivel de confianza.

La figura 9.7 muestra el resultado obtenido para la prueba HSD de Tukey mediante el script 9.6. En esta figura podemos apreciar que la columna `diff` muestra las diferencias de las medias entre grupos, obteniéndose resultados idénticos a los teóricos, y la columna `p.adj` entrega valores p asociados a cada diferencia, **ajustados** para compararlos con el nivel de significación original.

Script 9.6: (continuación del script 9.5) procedimiento post-hoc de Tukey usando la función `TukeyHSD()` de R.

```r
# Realizar y mostrar un procedimiento post-hoc HSD de Tukey
hsd <- TukeyHSD(prueba, "Algoritmo",
                ordered = TRUE, conf.level = 1 - alfa)
cat("\nProcedimiento HSD de Tukey\n")
cat("--------------------------\n")
print(hsd)
```

```
Procedimiento HSD de Tukey
--------------------------
  Tukey multiple comparisons of means
    97.5% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = Tiempo ~ Algoritmo, data = datos_largos)

$Algoritmo
    diff        lwr       upr     p adj
A-C  1.8 -3.0923417  6.692342 0.5176889
B-C  5.8  0.9076583 10.692342 0.0090297
B-A  4.0 -0.8923417  8.892342 0.0670199
```

Figura 9.7: resultado del procedimiento post-hoc HSD de Tukey.

Cabe destacar que el único valor p menor a este nivel ($\alpha = 0{,}025$) corresponde a la diferencia B-C, siendo esta última la única significativa, lo cual una vez más coincide con el resultado del procedimiento manual. También debemos notar que las columnas `lwr` y `upr` muestran el límite inferior y superior, respectivamente, del intervalo de $(1-\alpha)\cdot 100\,\%$ confianza para la verdadera diferencia entre las medias de los grupos.

9.9.2 Prueba de comparación de Scheffé

Otra alternativa diseñada para hacer un análisis post-hoc luego de una prueba ANOVA significativa es la **prueba de Scheffé**. Al igual que la corrección de Bonferroni, este método también es muy conservador al momento de efectuar comparaciones entre pares. No obstante, tiene como ventaja que permite hacer comparaciones adicionales, además de todos los pares de grupos, como por ejemplo si cierto grupo es mejor que todos los demás.

La ingeniera del ejemplo podría, tras encontrar mediante el procedimiento ANOVA que existen diferencias significativas, plantearse preguntas del siguiente tipo:

1. ¿Existe diferencia entre los tiempos de ejecución de los algoritmos A y B?
2. ¿Es el tiempo promedio de ejecución del algoritmo C distinto al tiempo de ejecución promedio de los algoritmos A y B?

La primera pregunta corresponde a una comparación entre pares, pero la segunda resulta más compleja. En realidad, podemos modelar escenarios para múltiples preguntas, usando para ello **contrastes**, que son **combinaciones lineales** de las medias de cada grupo. Para entender mejor esta idea, veamos la primera pregunta. Matemáticamente, puede formularse como las siguientes hipótesis:

$$H_0\colon\ \mu_A - \mu_B = 0$$
$$H_A\colon\ \mu_A - \mu_B \neq 0$$

La hipótesis nula puede expresarse, entonces, como una combinación lineal de la forma: $c_1\,\mu_A + c_2\,\mu_B + c_3\,\mu_C = 0$, asociada entonces al vector de coeficientes: $[c_1, c_2, c_3]$.

Para la primera pregunta entonces, los coeficientes serían: $[1, -1, 0]$, correspondiente a la combinación lineal: $1\cdot\mu_A - 1\cdot\mu_B + 0\cdot\mu_C = 0$.

La segunda pregunta es algo más compleja, pero las hipótesis asociadas son:

$$H_0\colon\ \mu_C - \frac{\mu_A + \mu_B}{2} = 0$$
$$H_A\colon\ \mu_C - \frac{\mu_A + \mu_B}{2} \neq 0$$

El vector de coeficientes en este caso sería: $[-\frac{1}{2}, -\frac{1}{2}, 1]$, para la combinación lineal $-\frac{1}{2}\cdot\mu_A - \frac{1}{2}\cdot\mu_B + 1\cdot\mu_C = 0$.

Ahora que hemos establecido qué es un contraste, podemos comenzar a explicar el procedimiento post-hoc de Scheffé, con la ayuda del ejemplo usado a lo largo del capítulo.

El primer paso consiste en determinar los contrastes $\{C_i\}$ a realizar. Supongamos que la ingeniera desea hacer todas las comparaciones entre pares y, además, comparar cada algoritmo contra los dos restantes. Podemos representar estos seis contrastes en forma matricial, donde la i-ésima fila de la matriz corresponde al contraste C_i:

$$\begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & -0{,}5 & -0{,}5 \\ -0{,}5 & 1 & -0{,}5 \\ -0{,}5 & -0{,}5 & 1 \end{bmatrix}$$

Luego calculamos los estimaciones para cada contraste considerando las medias calculas para cada grupo (ecuación 9.1):

$$\begin{aligned} C_1 &= \overline{x}_A - \overline{x}_B & &= -4{,}0 \\ C_2 &= \overline{x}_A - \overline{x}_C & &= \ \ 1{,}8 \\ C_3 &= \overline{x}_B - \overline{x}_C & &= \ \ 5{,}8 \\ C_4 &= \overline{x}_A - 0{,}5\cdot\overline{x}_B - 0{,}5\cdot\overline{x}_C & &= -1{,}8 \\ C_5 &= -0{,}5\cdot\overline{x}_A + \overline{x}_B - 0{,}5\cdot\overline{x}_C & &= \ \ 4{,}9 \\ C_6 &= -0{,}5\cdot\overline{x}_A - 0{,}5\cdot\overline{x}_B + \overline{x}_C & &= -3{,}8 \end{aligned}$$

El tercer paso consiste en calcular **valores críticos** para cada contraste, que para la prueba de comparación de Scheffé están dados por la ecuación 9.18:

$$C_i^* = \sqrt{F^*_{1-\alpha}\ \nu_{\text{efecto}}\ MS_{\text{error}}\ \sum_{j=1}^{k}\frac{c_{i,j}^2}{n_j}} \tag{9.18}$$

donde:

- i es el número (de fila) del contraste.
- $F^*_{1-\alpha}$ corresponde al percentil $1-\alpha$ de la distribución F con ν_{efecto} y ν_{error} grados de libertad.
- MS_{error} es la varianza aleatoria.
- $c_{i,j}^2$ es el coeficiente para el grupo (columna) j en el contraste (fila) i.
- n_j es el tamaño de la muestra para el grupo j.

Los valores ν_{efecto}, ν_{error} y MS_{error} se obtienen desde la tabla ANOVA. Así, revisando la tabla 9.1, para el ejemplo tenemos que $\nu_{\text{efecto}} = 2$, $\nu_{\text{error}} = 12$ y $MS_{\text{error}} = 6{,}4$. Podemos calcular $F^*_{1-\alpha}$ en R con la llamada `qf(1 - 0.025, 2, 12, lower.tail = TRUE)`, obteniéndose $F^*_{0{,}975} \approx 5{,}096$. Debemos notar que en la ecuación 9.18, la expresión $F^*_{1-\alpha}\ \nu_{\text{efecto}}\ MS_{\text{error}} = 5{,}096\cdot 2\cdot 6{,}4 \approx 65{,}229$ es constante para todos los contrastes. Así:

$$C_1^* = \sqrt{65{,}229\cdot\left(\frac{1^2}{5} + \frac{(-1)^2}{5} + \frac{0^2}{5}\right)} = 5{,}108$$
$$C_2^* = \sqrt{65{,}229\cdot\left(\frac{1^2}{5} + \frac{0^2}{5} + \frac{(-1)^2}{5}\right)} = 5{,}108$$

$$C_3^* = \sqrt{65{,}229 \cdot \left(\frac{0^2}{5} + \frac{1^2}{5} + \frac{(-1)^2}{5}\right)} = 5{,}108$$

$$C_4^* = \sqrt{65{,}229 \cdot \left(\frac{1^2}{5} + \frac{(-0{,}5)^2}{5} + \frac{(-0{,}5)^2}{5}\right)} = 4{,}424$$

$$C_5^* = \sqrt{65{,}229 \cdot \left(\frac{(-0{,}5)^2}{5} + \frac{1^2}{5} + \frac{(-0{,}5)^2}{5}\right)} = 4{,}424$$

$$C_6^* = \sqrt{65{,}229 \cdot \left(\frac{(-0{,}5)^2}{5} + \frac{(-0{,}5)^2}{5} + \frac{1^2}{5}\right)} = 4{,}424$$

Finalmente evaluamos cada contraste, comparando la estimación C_i con el valor crítico correspondiente, C_i^*. Si $|C_i| > C_i^*$, la comparación es estadísticamente significativa, y el signo del coeficiente nos dice cuál es la dirección de esa diferencia.

Tabulemos los resultados obtenidos como muestra la tabla 9.2.

i	C_i	C_i^*	$\lvert C_i\rvert > C_i^*$
1	−4,0	5,108	No
2	1,8	5,108	No
3	5,8	5,108	Sí
4	−1,8	4,424	No
5	4,9	4,424	Sí
6	−3,8	4,424	No

Tabla 9.2: resultado de aplicar la prueba de comparación de Scheffé a los contrastes definidos para el ejemplo.

Podemos ver, entonces, que las comparaciones 3 ($\overline{x}_B - \overline{x}_C$) y 5 ($\overline{x}_B - (\overline{x}_A + \overline{x}_C)/2$) son significativas, en ambos casos las diferencias son positivas, por lo que es el algoritmo B el que presenta la media más alta. Por lo tanto podemos concluir con 97,5 % confianza que el algoritmo B es menos eficiente (tarda más en promedio) que el algoritmo C, y que el tiempo promedio de ejecución del algoritmo B es significativamente mayor al tiempo promedio combinado de los algoritmos A y C.

Notemos que la conclusión menciona el mismo nivel de confianza que en la prueba ómnibus, pues tuvimos el cuidado de emplear el mismo nivel de significación $\alpha = 0{,}025$ al calcular $F^*_{1-\alpha}$ para las comparaciones de a pares.

En R, este procedimiento puede hacerse mediante la función `ScheffeTest(x, which, contrasts, conf.level)` del paquete `DescTools`, donde:

- `x`: objeto `"aov"` con el resultado de ANOVA.
- `which`: variable independiente en la prueba.
- `contrasts`: matriz con los contrastes (cada contraste es una columna).
- `conf.level`: nivel de confianza.

El script 9.7 muestra el ejemplo en R, cuyo resultado se presenta en la figura 9.8. A diferencia del proceso manual, la función `ScheffeTest()` nos entrega un valor p ajustado para cada contraste e identifica aquellos que son relevantes para diferentes niveles de significación. Aquí, al igual que en el caso de la prueba HSD de Tukey, las columnas `lwr` y `upr` señalan los límites del intervalo de confianza para la verdadera diferencia entre las medias de los grupos. Debemos notar que los resultados son consistentes con los obtenidos con los cálculos manuales, encontrando diferencias significativas en los mismos contrastes.

Un detalle importante a tener en cuenta es que podemos hacer la llamada a `ScheffeTest()` sin entregar los argumentos `which` y `contrasts`, en cuyo caso únicamente se contrastan todos los pares, como en las pruebas post-hoc precedentes.

```
Procedimiento post-hoc de Scheffé
---------------------------------

  Posthoc multiple comparisons of means: Scheffe Test
    97.5% family-wise confidence level

$Algoritmo
      diff     lwr.ci     upr.ci   pval
A-B   -4.0 -9.1079193  1.1079193 0.0808 .
A-C    1.8 -3.3079193  6.9079193 0.5479
B-C    5.8  0.6920807 10.9079193 0.0118 *
A-B,C -1.1 -5.5235879  3.3235879 0.7356
B-A,C  4.9  0.4764121  9.3235879 0.0138 *
C-A,B -3.8 -8.2235879  0.6235879 0.0540 .

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

Figura 9.8: resultado de la prueba de comparación de Scheffé.

Script 9.7: (continuación del script 9.6) procedimiento post-hoc de Scheffé usando la función `ScheffeTest()` de R.

```r
# Crear matriz de contrastes para la prueba de Scheffé
contrastes <- matrix(c(1, -1  ,  0  ,  1  , 0, -1  ,  0  ,  1  , -1,
                       1, -0.5, -0.5, -0.5, 1, -0.5, -0.5, -0.5,  1),
                     ncol = 3, byrow = TRUE)
# Trasponer matriz de contrastes
# (para que cada contraste sea una columna).
contrastes <- t(contrastes)

# Realizar y mostrar un procedimiento post-hoc de Scheffé
scheffe <- ScheffeTest(x = prueba, which = "Algoritmo",
                       contrasts = contrastes, conf.level = 1 - alfa)
cat("\nProcedimiento post-hoc de Scheffé\n")
cat("---------------------------------\n")
print(scheffe)
```

9.10 RESUMEN DEL PROCEDIMIENTO

Si alguien quisiera pensar el procedimiento ANOVA de una vía para variables independientes de forma muy mecánica, cosa que no es recomendable, se podría resumirse en los siguientes pasos:

1. Determinar las hipótesis ómnibus.
2. Verificar el cumplimiento de las condiciones, y no seguir si hay dudas de ello.
3. Definir un nivel de significación.
4. Calcular la suma de cuadrados para la muestra combinada (SS_T).
5. Para cada grupo g, calcular la suma de cuadrados dentro de dicho grupo (SS_g).
6. Calcular la variabilidad entre grupos (SS_{bg}).
7. Calcular la variabilidad intra-grupos (SS_{wg}).
8. Calcular los grados de libertad (ν_T, ν_{bg} y ν_{wg}).
9. Calcular los cuadrados medios (MS_{bg} y MS_{wg}).
10. Calcular el estadístico de prueba (F).
11. Obtener el valor p.
12. Concluir sobre las hipótesis ómnibus.
13. Si la prueba resulta significativa:
    a) Seleccionar una procedimiento post-hoc.

b) Verificar que se cumplen las condiciones de la prueba post-hoc.
c) Aplicar el procedimiento post-hoc.
d) Concluir sobre los resultados post-hoc.

9.11 EJERCICIOS PROPUESTOS

9.1 El conjunto de datos `chickwts`, disponible en R, registra el peso de 71 pollitos a las seis semanas de nacidos y el tipo de alimento que cada pollito recibió. Para este conjunto de datos:

(a) Verifica si se cumplen las condiciones para efectuar un procedimiento ANOVA de una vía para muestras independientes.

(b) Sin considerar el resultado anterior, efectúa el procedmiento ANOVA de una vía para muestras independientes a fin de determinar si existen diferencias en el peso de los pollitos de acuerdo al tipo de alimento recibido.

(c) En caso de identificar que existen diferencias significativas, lleva a cabo los análisis post-hoc y determina qué tipos de alimento presentan dichas diferencias. Compara los resultados obtenidos con los diferentes métodos.

9.2 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba ANOVA independiente. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

9.3 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera utilizar una prueba ANOVA independiente. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

9.4 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera utilizar una prueba ANOVA independiente. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

9.5 Justificando tus suposiciones, inventa muestras de los datos que se podrían encontrar en el estudio propuesto en la pregunta 9.2 que tengan entre 12 y 15 observaciones y que cumplan las condiciones requeridas por la prueba ANOVA independiente.

9.6 Justificando tus suposiciones, inventa muestras de los datos que se podrían encontrar en el estudio propuesto en la pregunta 9.3 que tengan entre 12 y 15 observaciones y que cumplan las condiciones requeridas por la prueba ANOVA independiente.

9.7 Justificando tus suposiciones, inventa muestras de los datos que se podrían encontrar en el estudio propuesto en la pregunta 9.4 que tengan entre 12 y 15 observaciones y que cumplan las condiciones requeridas por la prueba ANOVA independiente.

9.8 Usando la función `ezANOVA()` realiza el análisis de las muestras definidas en la pregunta 9.5. Aplica la prueba HSD de Tukey para el análisis post-hoc (aún cuando este sea innecesario en estricto rigor).

9.9 Usando la función `ezANOVA()` realiza el análisis de las muestras definidas en la pregunta 9.6. Aplica la prueba de Scheffé para el análisis post-hoc (aún cuando este sea innecesario en estricto rigor).

9.10 Usando la función `ezANOVA()` realiza el análisis de las muestras definidas en la pregunta 9.7. Aplica el método de Benjamini y Hochberg para el análisis post-hoc (aún cuando este sea innecesario en estricto rigor).

9.11 Usando la función `pwr.anova.test()` del paquete `pwr` calcule el poder que tiene la prueba ANOVA aplicada en la pregunta 9.8. ¿Cuántas observaciones debería tener cada grupo para obtener una potencia de 0.9 (manteniendo los otros factores)?

9.12 Usando la función `pwr.anova.test()` del paquete `pwr` calcule el poder que tiene la prueba ANOVA aplicada en la pregunta 9.9. ¿Cuántas observaciones debería tener cada grupo para obtener una potencia de 0.9 (manteniendo los otros factores)?

9.13 Usando la función `pwr.anova.test()` del paquete `pwr` calcule el poder que tiene la prueba ANOVA aplicada en la pregunta 9.10. ¿Cuántas observaciones debería tener cada grupo para obtener una potencia de 0.9 (manteniendo los otros factores)?

9.12 BIBLIOGRAFÍA DEL CAPÍTULO

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods*, *37*, 379-384.

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815.

Berman, H. (2000). *Scheffé's Test for Multiple Comparisons.* Consultado el 7 de mayo de 2021, desde https://stattrek.com/anova/follow-up-tests/scheffe.aspx

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159.

Glen, S. (2021). *Post-Hoc Definition and Types of Post Hoc Tests.* Consultado el 7 de mayo de 2021, desde https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/post-hoc/#PHscheffes

IBM. (1989). *ANOVA de un factor: Contrastes post hoc.* Consultado el 30 de abril de 2021, desde https://www.ibm.com/docs/es/spss-statistics/25.0.0?topic=anova-one-way-post-hoc-tests

Kroes, A. D., & Finley, J. R. (2023). Demystifying omega squared: Practical guidance for effect size in common analysis of variance designs. *Psychological Methods.*

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 62627.

Lowry, R. (1999). *Concepts & Applications of Inferential Statistics.* Consultado el 3 de mayo de 2021, desde http://vassarstats.net/textbook/

Meier, L. (2021). *ANOVA: A Short Intro Using R.* Consultado el 7 de mayo de 2021, desde https://stat.ethz.ch/~meier/teaching/anova/

Midway, S., Robertson, M., Flinn, S., & Kaller, M. (2020). Comparing multiple comparisons: practical guidance for choosing the best multiple comparisons test. *PeerJ*, *8*, e10387.

NIST/SEMATECH. (2013). *e-Handbook of Statistical Methods.* Consultado el 29 de abril de 2021, desde http://www.itl.nist.gov/div898/handbook/

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, *8*(4), 434.

Williams, L. J., & Abdi, H. (2010). Fisher's least significant difference (LSD) test. *Encyclopedia of research design*, *218*(4), 840-853.