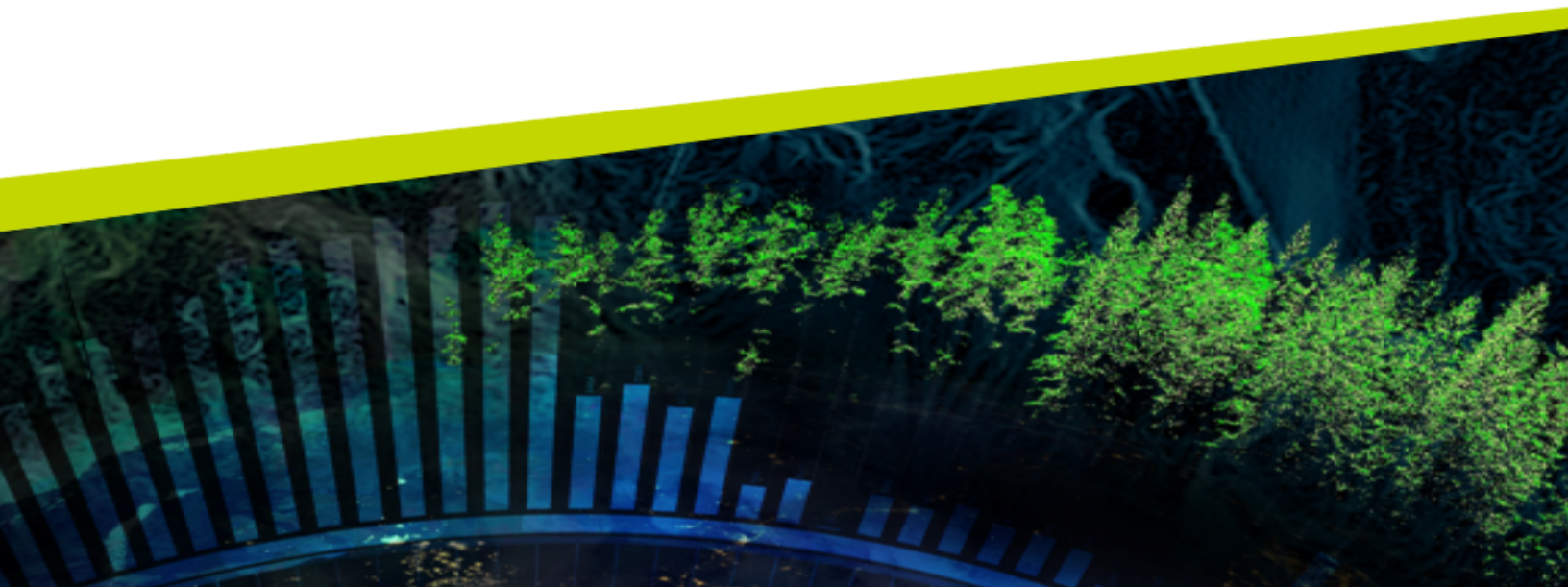
INFERENCIA Y MODELOS ESTADÍSTICOS

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. En el capítulo 8 nos enfrentamos a un escenario similar para cuando trabajamos con proporciones, para lo cual conocimos la prueba Q de Cochran. Aprendimos que esta última prueba es de tipo ómnibus: es decir, comprueba la igualdad de todos los grupos, pero que si encuentra diferencias no nos indica dónde están. En consecuencia, conocimos los procedimientos post-hoc para identificar entre qué grupos existen estas diferencias.

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).

Intuitivamente, podríamos abordar este problema efectuando pruebas t independientes para cada pareja de grupos con un nivel de significación α. Por ejemplo, para tres muestras A, B y C con medias $\overline{x}_A$, $\overline{x}_B$ y $\overline{x}_C$, respectivamente, se tendrían tres pruebas t de Student para diferencia de medias:

1. $\overline{x}_A - \overline{x}_B$
2. $\overline{x}_A - \overline{x}_C$
3. $\overline{x}_B - \overline{x}_C$

Sin embargo, este enfoque presenta un grave inconveniente: por cada una de las pruebas t anteriores, se tiene una probabilidad α de cometer un error tipo I. Al efectuar cada una de las pruebas anteriores, la probabilidad total de que en alguna de ellas se cometa un error tipo I se acerca a $3 \cdot \alpha$, bastante superior al nivel de significación nominal establecido para la prueba[1]. El método de **análisis de varianza**, comúnmente conocido como **ANOVA** o AoV (del inglés Analysis of Variance), surge, en esencia, como un método para combatir este problema al comparar simultáneamente tres o más medias muestrales.

De manera similar, también existe el procedimiento ANOVA para muestras correlacionadas (que se aborda en el capítulo siguiente), semejante a la prueba t de Student con muestras pareadas. Los procedimientos ANOVA para muestras independientes y muestras correlacionadas 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 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: un ingeniero cuenta con tres algoritmos (A, B y C) para resolver un determinado problema (en iguales condiciones y para instancias de tamaño fijo, digamos con E elementos) y desea comparar su eficiencia. Para cada algoritmo, selecciona una muestra aleatoria independiente de instancias y registra el tiempo de ejecución (en milisegundos) para cada una de las instancias de la muestra correspondiente, obteniendo las siguientes observaciones:

- Algoritmo A: 23, 19, 25, 23, 20

[1] Aunque el cálculo exacto de esta probabilidad disjunta escapa a los alcances de este curso, es intuitivo ver que la probabilidad de no cometer un error de tipo I en cada prueba es $(1 - alpha)$. Así, la probabilidad de no cometer un error de tipo I en las tres comparaciones es $(1 - alpha)^3$. Luego, la probabilidad de cometer un error de tipo I para la hipótesis global (no hay diferencias entre los grupos) es $(1 - (1 - alpha)^3)$. Si, por ejemplo, nominalmente $\alpha = 0,05$, el nivel de significación para la hipótesis global sería aproximadamente 0.143

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

La pregunta detrás de ANOVA para este ejemplo es: ¿se diferencian los tiempos medios que requieren los algoritmos para resolver todas las posibles instancias del problema de tamaño E? De donde se desprende que:

H_0 : el tiempo de ejecución promedio para instancias de tamaño E es igual para los tres algoritmos.
H_A : el tiempo de ejecución promedio para instancias de tamaño E es diferente para al menos un algoritmo.

Notemos que, como en el caso de la prueba Q de Cochran, la hipótesis nula no es específica, sino que comprueba la igualdad de todas las medias, por lo que ANOVA es una prueba ómnibus.

9.1 CONDICIONES PARA USAR ANOVA DE UNA VÍA PARA MUESTRAS INDEPENDIENTES

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. Las k muestras tienen varianzas aproximadamente iguales.

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 el ingeniero 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 observan algunos valores que podrían ser atípicos y las muestras son pequeñas, es mejor que procedamos con cautela y usemos un nivel de significación $\alpha = 0,025$.

Una regla sencilla para comprobar la cuarta condición, llamada también **homogeneidad de las varianzas** u **homocedasticidad**, es comprobar que la razón entre la máxima y la mínima varianza muestral de los grupos no sea superior a 1,5.

$$\overline{x}_B = \frac{26+24+28+23+29}{5} = 26,0$$
$$s_A^2 = \frac{(23-22)^2+(19-22)^2+(25-22)^2+(23-22)^2+(20-22)^2}{4} = 6,0$$
$$\overline{x}_B = \frac{26+24+28+23+29}{5} = 26,0$$
$$s_B^2 = \frac{(26-26)^2+(24-26)^2+(28-26)^2+(23-26)^2+(29-26)^2}{4} = 6,5$$
$$\overline{x}_C = \frac{19+24+20+21+17}{5} = 20,2$$
$$s_C^2 = \frac{(19-20,2)^2+(24-20,2)^2+(20-20,2)^2+(21-20,2)^2+(17-20,2)^2}{4} = 6,7$$

En el caso del ejemplo, la muestra obtenida para el algoritmo A tiene la menor varianza, mientras que la muestra del algoritmo C tiene la mayor:

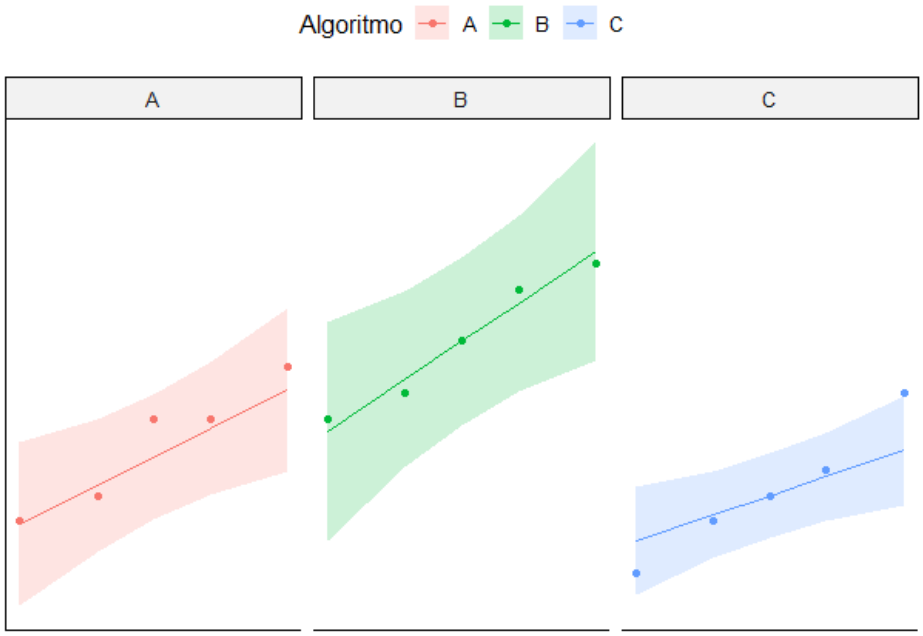


Figura 9.1: gráfico para comprobar el supuesto de normalidad en las tres muestras del ejemplo.

$$\frac{s_C^2}{s_A^2} = \frac{6,7}{6,0} = 1,117 \tag{9.1}$$

En consecuencia, la condición de homocedasticidad se verifica para el ejemplo.

Se ha encontrado que ANOVA es una prueba **robusta**, que resiste razonablemente bien a 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.

9.2 PROCEDIMIENTO ANOVA DE UNA VÍA PARA MUESTRAS INDEPENDIENTES

Como su nombre indica, ANOVA se centra en la **variabilidad** de las muestras, una generalización de la varianza, que se calcula en base a la suma de los cuadrados de las desviaciones, como muestra la ecuación 9.2, donde n corresponde al tamaño de la muestra, y $\overline{x}$ a la media muestral.

$$SS = \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{9.2}$$

9.2.1 Variabilidad total

La variabilidad total, SS_T, se calcula mediante la ecuación 9.2 considerando la totalidad de observaciones, vale decir, combinando las muestras correspondientes a los diferentes grupos.

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

$$\begin{aligned}SS_T = &(23-22,733)^2+(19-22,733)^2+(25-22,733)^2+(23-22,733)^2+(20-22,733)^2+ \\ &(26-22,733)^2+(24-22,733)^2+(28-22,733)^2+(23-22,733)^2+(29-22,733)^2+ \\ &(19-22,733)^2+(24-22,733)^2+(20-22,733)^2+(21-22,733)^2+(12-22,733)^2 \\ = &\ 164,933\end{aligned}$$

La variabilidad total puede descomponerse en dos partes: una de ellas corresponde a la variabilidad existente al interior de cada uno de los grupos (o variabilidad intra-grupos), *within groups* en inglés, denotada por SS_{wg}; la otra corresponde a la variabilidad entre los diferentes grupos, *between groups* en inglés, denotada como SS_{bg}. La ecuación 9.3 muestra una **identidad importante** que relaciona ambas componentes.

$$SS_T = SS_{bg} + SS_{wg} \tag{9.3}$$

9.2.2 Variabilidad entre grupos

La **variabilidad entre grupos** nos permite medir de manera agregada la magnitud de las diferencias entre las distintas medias muestrales. Se calcula como muestra la ecuación 9.4, donde:

- k: cantidad de grupos.
- n_i: cantidad de observaciones en el i-ésimo grupo.
- $\overline{x}_i$: media del i-ésimo grupo.
- $\overline{x}_T$: media de la muestra combinada.

$$SS_{bg} = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x}_T)^2 \tag{9.4}$$

Esta medida corresponde a la suma de las desviaciones cuadradas de la media de cada grupo con respecto a la media combinada, donde la diferencia de cada grupo se pondera por la cantidad de observaciones que este contiene, a fin de mantener la representatividad de cada grupo. Mide el grado en que los grupos difieren unos de otros. Para el ejemplo tenemos:

$$SS_{bg} = 5(22,0-22,733)^2+5(26,0-22,733)^2+5(20,2-22,733)^2 = 88,133$$

9.2.3 Variabilidad al interior de cada grupo

La **variabilidad intra-grupos**, a su vez, corresponde a la suma total de las desviaciones cuadradas al interior de cada grupo, por lo que representa la variabilidad aleatoria de cada uno de los diferentes grupos. Esta medida se calcula de acuerdo a la ecuación 9.5, donde SS_i corresponde a la variabilidad del i-ésimo grupo, calculada mediante la ecuación 9.2.

$$SS_{wg} = \sum_{i=1}^{k} SS_i \tag{9.5}$$

Para el ejemplo:

$$\begin{aligned}
SS_A &= (23-22)^2 + (19-22)^2 + (25-22)^2 + (23-22)^2 + (20-22)^2 = 24,0 \\
SS_B &= (26-26)^2 + (24-26)^2 + (28-26)^2 + (23-26)^2 + (29-26)^2 = 26,0 \\
SS_C &= (19-20,2)^2 + (24-20,2)^2 + (20-20,2)^2 + (21-20,2)^2 + (17-20,2)^2 = 26,8 \\
SS_{wg} &= SS_A + SS_B + SS_C = 24,0 + 26,0 + 26,8 = 76,8
\end{aligned}$$

9.2.4 El estadístico de prueba F

En el capítulo 2 vimos que la varianza se calcula como:

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

Podemos generalizar esta ecuación como muestra la ecuación 9.6, donde ν corresponde a los grados de libertad.

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{9.6}$$

En el contexto de análisis de varianza, llameremos MS, del inglés *mean square*, a la media de las desviaciones cuadradas. 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 (para el ejemplo con los tres algortimos, $\nu_{bg} = 3 - 1 = 2$). Con ello, la media cuadrada entre grupos queda dada por la ecuación 9.7.

$$MS_{bg} = \frac{SS_{bg}}{\nu_{bg}} \tag{9.7}$$

Entonces, en nuestro ejemplo:

$$MS_{bg} = \frac{88,133}{2} = 44,067$$

De manera similar, los grados de libertad para la componente de la variabilidad al interior de los grupos está dada por la suma de los grados de libertad en cada grupo, como se ve en la ecuación 9.8 (siendo k la cantidad

de grupos), y su media de las desviaciones cuadradas se normaliza con estos grados de libertad (ecuación 9.9).

$$\nu_{wg} = \sum_{i=1}^{k}(n_k - 1) \tag{9.8}$$

$$MS_{wg} = \frac{SS_{wg}}{\nu_{wg}} \tag{9.9}$$

Así, para el ejemplo tenemos:

$$\begin{aligned} \nu_{wg} &= (5-1) + (5-1) + (5-1) = 12 \\ MS_{wg} &= \frac{76,8}{12} = 6,4 \end{aligned}$$

En ocasiones resulta útil conocer también la cantidad total de grados de libertad, que podemos obtener mediante la ecuación 9.10, donde n_T es el tamaño de la muestra combinada.

$$\nu_T = n_T - 1 = \nu_{bg} + \nu_{wg} \tag{9.10}$$

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

- Si la hipótesis nula es verdadera, 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 (el estadístico de prueba para ANOVA), que se calcula como muestra la ecuación 9.11, donde MS_{efecto} corresponde a una estimación de la varianza del efecto que se desea medir y MS_{error}, a la variabilidad aleatoria pura asociada a la situación. En este punto puede ser útil revisar nuevamente lo que ya hemos aprendido de la distribución F (capítulo 3).

$$F = \frac{MS_{\text{efecto}}}{MS_{\text{error}}} \tag{9.11}$$

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:

$$F = \frac{44,067}{6,4} = 6,885$$

De manera similar a lo que hemos visto en otras pruebas, el p-valor corresponde al área bajo la cola superior de la distribución F (en este caso con 2 y 12 grados de libertad) mayor o igual al estadístico obtenido, que en R puede calcularse mediante la llamada `pf(6,885, 2, 12, lower.tail = FALSE)`, obteniéndose $p = 0,010$.

9.2.5 Resultado del procedimiento ANOVA

El resultado del procedimiento ANOVA suele representarse en forma tabular, 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, podemos 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.

9.2.6 Resumen del procedimiento ANOVA de una vía para muestras independientes

El procedimiento ANOVA de una vía para variables independientes puede resumirse en los siguientes pasos:

1. Calcular la suma de los cuadrados de las desviaciones para la muestra combinada (SS_T).
2. Para cada grupo g, calcular la suma de los cuadrados de las desviaciones dentro de dicho grupo (SS_g).
3. Calcular la variabilidad entre grupos (SS_{bg}).
4. Calcular la variabilidad al interior de los grupos (SS_{wg}).
5. Calcular los grados de libertad (ν_T, ν_{bg} y ν_{wg}).
6. Calcular las medias de las desviaciones cuadradas (MS_{bg} y MS_{wg}).
7. Calcular el estadístico de prueba (F).
8. Obtener el valor p.

9.3 ANOVA DE UNA VÍA PARA MUESTRAS INDEPENDIENTES EN R

Desde luego, R nos ofrece funciones para realizar diferentes pruebas ANOVA, incluyendo la de una vía para muestras independientes.

La primera alternativa que conoceremos es la función `aov(formula, data)`, donde:

- `formula`: se escribe de la forma `variable_dependiente ~ variable_independiente`.
- `data`: data frame que contiene las variables especificadas en la fórmula.

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 (numérica con escala de igual intervalo).
- `wid`: variable (factor) con el identificador de cada instancia.
- `between`: variable independiente (factor).

- `return_aov`: si es verdadero, devuelve un objeto de tipo aov para uso posterior.

Un parámetro adicional que no hemos mencionado es `type`, el cual no estudiaremos en detalle porque escapa a los alcances de este libro. Sin embargo, se debe tener en cuenta que este argumento permite incorporar algunas modificaciones y resguardos en caso que las muestras tengan diferentes tamaños o se tengan datos incompletos. Al trabajar con R, para la mayoría de los casos se recomienda mantener el valor por defecto (`type = 2`).

Una ventaja de `ezANOVA()` por sobre `aov()` es que, además de ejecutar la prueba ANOVA, realiza también la **prueba de homocedasticidad de Levene** (NIST/SEMATECH, 2013). Si bien no estudiaremos esta prueba en detalle, es pertinente mencionar que sirve para comprobar si k muestras tienen igual varianza, por lo que su resultado nos ayuda a verificar las condiciones requeridas 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 muestras son iguales.
H_A: al menos una de las muestras tiene varianza diferente a alguna de las demás.

El paquete `ez` contiene también la función `ezPlot(data, dv, wid, between, x)`, la cual nos permite ver gráficamente el tamaño del efecto medido. 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.

El script 9.1 muestra el procedimiento ANOVA de una vía para muestras independientes usando ambas funciones, con igual resultado al obtenido de manera manual. Además, genera el gráfico del tamaño del efecto medido, presentado en la figura 9.2.

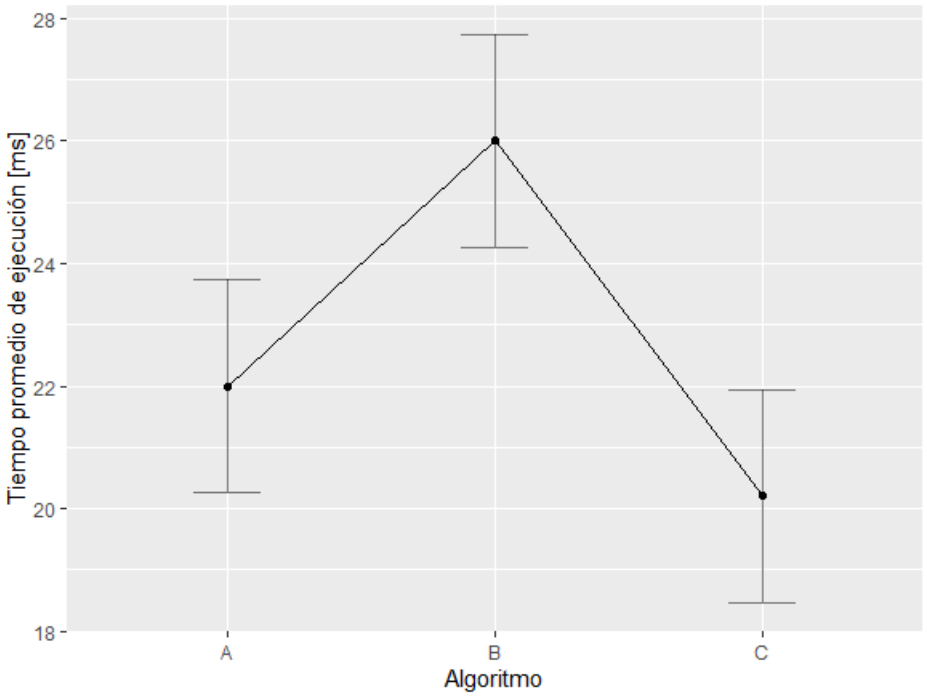


Figura 9.2: tamaño del efecto medido.

Script 9.1: procedimiento ANOVA de una vía para muestras independientes.

```
1 library(tidyverse)
2 library(ggpubr)
3 library(ez)
4
5 # Crear el data frame en formato ancho.
6 A <- c(23, 19, 25, 23, 20)
```

```r
B <- c(26, 24, 28, 23, 29)
C <- c(19, 24, 20, 21, 17)
datos <- data.frame(A, B, C)

# Llevar data frame a formato largo.
datos <- datos %>% pivot_longer(c("A", "B", "C"),
                                names_to = "algoritmo",
                                values_to = "tiempo")

datos[["algoritmo"]] <- factor(datos[["algoritmo"]])
datos[["instancia"]] <- factor(1:nrow(datos))

# Comprobción de normalidad.
g <- ggqqplot(datos,
              x = "tiempo",
              y = "algoritmo",
              color = "algoritmo")

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)

# Procedimiento ANOVA con aov().
cat("Procedimiento ANOVA usando aov\n\n")
prueba <- aov(tiempo ~ algoritmo, data = datos)
print(summary(prueba))

# Procedimiento ANOVA con ezANOVA().
cat("\n\nProcedimiento ANOVA usando ezANOVA\n\n")
prueba2 <- ezANOVA(
  data = datos,
  dv = tiempo,
  between = algoritmo,
  wid = instancia,
  return_aov = TRUE)

print(prueba2)

# Gráfico del tamaño del efecto.
g2 <- ezPlot(
  data = datos,
  dv = tiempo,
  wid = instancia,
  between = algoritmo,
  y_lab = "Tiempo promedio de ejecución [ms]",
  x = algoritmo
)

print(g2)
```

9.4 ANÁLISIS POST-HOC

Al aplicar el procedimiento ANOVA de una vía para muestras independientes a nuestro ejemplo pudimos concluir que **existe al menos un algoritmo cuyo tiempo promedio de ejecución es diferente al de los demás**. Ahora bien, si los algoritmos del ejemplo 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 algún método que permita determinar si los tiempos de ejecución de los algoritmos A y B difieren significativamente, o los de B y C, o bien los de A y C. En el contexto general, si tenemos k grupos, la cantidad de comparaciones (N) que deberíamos efectuar está dada por la ecuación 9.12.

$$N = \binom{k}{2} = \frac{k(k-1)}{2} \tag{9.12}$$

Al igual que aprendimos en el capítulo anterior para la prueba Q de Cochran, existen diversas pruebas *post-hoc* que podemos usar para este fin, algunas de las cuales exploraremos a continuación.

9.4.1 Correcciones de Bonferroni y Holm

Como ya estudiamos en el capítulo anterior, los factores de corrección de Bonferroni y Holm distribuyen el nivel de significación cuando se realizan múltiples comparaciones entre pares de grupos. Las fórmulas para calcularlos son las mismas que ya conocimos, pero ahora se realizan pruebas t para muestras independientes para cada par. 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 9.2 muestra la realización de pruebas t para cada par de grupos usando tanto la corrección de Bonferroni como la de Holm, obteniéndose los resultados que se muestran en la figura 9.3. Debemos recordar que el nivel de significación que se entrega como argumento es el mismo que usamos en el procedimiento ANOVA.

Script 9.2: procedimientos *post-hoc* de Bonferroni y Holm en R.

```
library(tidyverse)

# Crear el data frame 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 <- data.frame(A, B, C)

# Llevar data frame a formato largo.
datos <- datos %>% pivot_longer(c("A", "B", "C"),
```

```
                                names_to = "algoritmo",
                                values_to = "tiempo")

datos[["algoritmo"]] <- factor(datos[["algoritmo"]])
datos[["instancia"]] <- factor(1:nrow(datos))

# Establecer nivel de significación (el mismo usado en ANOVA).
alfa <- 0.025

# Procedimiento post-hoc de Bonferroni.
cat("Procedimiento post-hoc de Bonferroni\n\n")

bonferroni <- pairwise.t.test(datos[["tiempo"]],
                              datos[["algoritmo"]],
                              p.adj = "bonferroni",
                              pool.sd = TRUE,
                              paired = FALSE,
                              conf.level = 1 - alfa)

print(bonferroni)

# Procedimiento post-hoc de Holm.
cat("\n\nProcedimiento post-hoc de Holm\n\n")

holm <- pairwise.t.test(datos[["tiempo"]],
                        datos[["algoritmo"]],
                        p.adj = "holm",
                        pool.sd = TRUE,
                        paired = FALSE,
                        conf.level = 1 - alfa)

print(holm)
```

Los valores p obtenidos con ambos métodos son diferentes (un lector atento recordará que la corrección de Bonferroni es considerada muy conservadora). Sin embargo, en ambos casos podemos ver que únicamente los algoritmos B y C presentan una diferencia significativa al comparar el valor p ajustado que entrega R con el nivel de significación ($\alpha = 0,025$). Si miramos el gráfico del tamaño del efecto obtenido para el procedimiento ANOVA (figura 9.2), podemos concluir entonces con 97,5 % de confianza que el algoritmo C es más rápido que el algoritmo B.

9.4.2 Prueba HSD de Tukey

La prueba **HSD de Tukey** es más poderosa que los factores de corrección de Bonferroni y Holm. Se asemeja a estas últimas en que también busca diferencias significativas (de hecho, el nombre HSD se debe a las siglas inglesas para "diferencia honestamente significativa") entre los diferentes pares de medias, aunque usa un enfoque muy diferente: para ello emplea el estadístico Q, el cual sigue una distribución de rango estudiantizado[2], que para cualquier par de medias en los k grupos se calcula según la ecuación 9.13, donde:

- $\overline{x}_g$ es la mayor de las dos medias comparadas.
- $\overline{x}_p$ es la menor de las dos medias comparadas.
- MS_{wg} corresponde la media cuadrada intra-grupos (entregada por el procedimiento ANOVA).

[2]El detalle de la distribución de rango estudiantizado escapa a los alcances de este texto.

```
Procedimiento post-hoc de Bonferroni


Pairwise comparisons using t tests with pooled SD

data:  datos[["tiempo"]] and datos[["algoritmo"]]

  A     B
B 0.084 -
C 0.848 0.010

P value adjustment method: bonferroni


Procedimiento post-hoc de Holm


Pairwise comparisons using t tests with pooled SD

data:  datos[["tiempo"]] and datos[["algoritmo"]]

  A     B
B 0.056 -
C 0.283 0.010

P value adjustment method: holm
```

Figura 9.3: valores p obtenidos en las pruebas t para cada par de grupos mediante los métodos de Bonferroni y Holm.

- n_m es la cantidad de observaciones por cada muestra. Si las k muestras tienen tamaños diferentes, se calcula mediante la fórmula presentada en la ecuación 9.14.

$$Q = \frac{\overline{x}_g - \overline{x}_p}{\sqrt{\frac{MS_{wg}}{n_m}}} \tag{9.13}$$

$$n_m = \frac{k}{\sum_{i=1}^{k} \frac{1}{n_i}} \tag{9.14}$$

En la práctica, sin embargo, no necesitamos 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 por Q_α), el cual depende de la cantidad de grupos (k) y de los grados de libertad del error, ν_{wg} en el caso de ANOVA de una vía para muestras independientes. La llamada `qtukey(`α`,` n_m`,` ν_{wg}`, lower.tail = FALSE)` en R entrega el valor de Q_α.

El valor crítico Q_α nos permite determinar cuán grande debe ser la diferencia entre las medias de dos grupos para ser considerada significativa, lo cual se logra mediante la ecuación 9.15.

$$HSD_\alpha = Q_\alpha \cdot \sqrt{\frac{MS_{wg}}{n_m}} \tag{9.15}$$

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

Para el ejemplo, se tenemos que $Q_\alpha = 4,324$, de donde:

$$HSD_{0,025} = 4,324 \cdot \sqrt{\frac{6,4}{5}} = 4,892$$

Recordando del procedimiento ANOVA que $\overline{x}_A = 22$, $\overline{x}_B = 26$ y $\overline{x}_C = 20,2$; tenemos:

$$\overline{x}_B - \overline{x}_A = 26 - 22 = 4$$

$$\overline{x}_A - \overline{x}_C = 22 - 20,2 = 1,8$$

$$\overline{x}_B - \overline{x}_C = 26 - 20,2 = 5,8$$

Así, la tercera diferencia es la única que supera $HSD_{0,025}$, con lo que solo existe diferencia significativa entre los tiempos promedio de ejecución de los algoritmos B y C, y se puede concluir que el algoritmo C es más rápido que el algoritmo B, lo que se condice con los resultados presentados en la figura 9.2.

R también permite realizar la prueba HSD de Tukey, como muestra el script 9.3. 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 9.4 muestra el resultado obtenido para la prueba HSD de Tukey mediante el script 9.3.

Script 9.3: procedimiento *post-hoc* de Tukey.

```
library(tidyverse)

# Crear el data frame 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 <- data.frame(A, B, C)

# Llevar data frame a formato largo.
datos <- datos %>% pivot_longer(c("A", "B", "C"),
                                names_to = "algoritmo",
                                values_to = "tiempo")

datos[["algoritmo"]] <- factor(datos[["algoritmo"]])
datos[["instancia"]] <- factor(1:nrow(datos))

# Establecer nivel de significación (el mismo usado en ANOVA).
alfa <- 0.025

# Procedimiento ANOVA.
anova <- aov(tiempo ~ algoritmo, data = datos)

# Prueba HSD de Tukey.
post_hoc <- TukeyHSD(anova,
                     "algoritmo",
                     ordered = TRUE,
```

```
27                       conf.level = 1 - alfa)
28
29 print(post_hoc)
```

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

Fit: aov(formula = tiempo ~ algoritmo, data = datos)

$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 9.4: resultado del procedimiento *post-hoc* HSD de Tukey.

En la figura 9.4 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. 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.

9.4.3 Prueba de comparación de Scheffé

Otra alternativa para hacer un análisis *post-hoc* 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 la ventaja de que permite hacer comparaciones adicionales, además de todos los pares de grupos: por ejemplo, podríamos preguntar si un grupo es mejor que todos los demás. El ingeniero 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 A distinto al tiempo de ejecución promedio de los algoritmos B y C?

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: $\mu_A - \mu_B = 0$
H_A: $\mu_A - \mu_B \neq 0$

La hipótesis nula puede expresarse, entonces, como una combinación lineal de la forma:

$$c_A \cdot \mu_A + c_B \cdot \mu_B + c_C \cdot \mu_C = 0$$

Que puede, a su vez, representarse vectorialmente como:

$$[c_A, c_B, c_C]$$

En este caso, resulta evidente que la combinación lineal es:

$$1 \cdot \mu_A - 1 \cdot \mu_B + 0 \cdot \mu_C = 0$$

Que corresponde al vector:

$$[1, -1, 0]$$

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

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

Vectorialmente dada por:

$$\left[1, -\frac{1}{2}, -\frac{1}{2}\right]$$

Ahora que hemos establecido qué es un contraste, podemos comenzar a explicar el el procedimiento *post-hoc* de Scheffé, el cual ocupa el mismo nivel de significación empleado para el procedimiento ANOVA. Recordemos que, para el ejemplo, $\alpha = 0,025$, $\overline{x}_A = 22$, $\overline{x}_B = 26$ y $\overline{x}_C = 20,2$.

El primer paso consiste en determinar los contrastes a realizar. Supongamos que el ingeniero desea hacer todas las comparaciones entre pares y, además, comparar cada algoritmo contra los dos restantes. Podemos representar esto en forma matricial, donde cada fila de la matriz corresponde a un contraste:

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

Luego calculamos los estimadores para cada contraste C_i:

$$\begin{aligned} C_1 &= |\overline{x}_A - \overline{x}_B| = 4,0 \\ C_2 &= |\overline{x}_A - \overline{x}_C| = 1,8 \\ C_3 &= |\overline{x}_B - \overline{x}_C| = 5,8 \\ C_4 &= |\overline{x}_A - \frac{\overline{x}_B}{2} - \frac{\overline{x}_C}{2}| = 1,1 \\ C_5 &= |-\frac{\overline{x}_A}{2} + \overline{x}_B - \frac{\overline{x}_C}{2}| = 4,9 \\ C_6 &= |-\frac{\overline{x}_A}{2} - \frac{\overline{x}_B}{2} + \overline{x}_C| = 3,8 \end{aligned}$$

El tercer paso consiste en calcular los valores críticos para la prueba de comparación de Scheffé, dados por la ecuación 9.16, donde:

- i es el número de fila del contraste.
- ν_{efecto}, ν_{error} y MS_{error} se obtienen desde la tabla ANOVA (tabla 9.1).
- F^* corresponde al percentil $1-\alpha$ de la distribución $F(\nu_{\text{efecto}}, \nu_{\text{error}})$.
- c_j es el peso del grupo j en la comparación i.
- n_j es el tamaño de la muestra para el grupo j.

$$VC_i = \sqrt{\nu_{\text{efecto}} \cdot F^* \cdot MS_{\text{error}} \cdot \sum_{j=1}^{k} \frac{c_j^2}{n_j}} \tag{9.16}$$

Así, para el ejemplo tenemos que $\nu_{\text{efecto}} = 2$, $\nu_{\text{error}} = 12$ y $MS_{\text{error}} = 6,4$. Podemos calcular F^* en R con la llamada `qf(1 - 0.025, 2, 12, lower.tail = TRUE)`, obteniéndose $F^* = 5,0959$. Debemos notar que en la ecuación 9.16, $\nu_{\text{efecto}} \cdot F^* \cdot MS_{\text{error}} = 2 \cdot 5,0959 \cdot 6,4 = 65,2275$ es constante para todos los contrastes. Así:

$$VC_1 = \sqrt{65,2275 \cdot \left(\frac{1^2}{5} + \frac{(-1)^2}{5} + \frac{0^2}{5}\right)} = 4,9891$$

$$VC_2 = \sqrt{65,2275 \cdot \left(\frac{1^2}{5} + \frac{0^2}{5} + \frac{(-1)^2}{5}\right)} = 4,9891$$

$$VC_3 = \sqrt{65,2275 \cdot \left(\frac{0^2}{5} + \frac{1^2}{5} + \frac{(-1)^2}{5}\right)} = 4,9891$$

$$VC_4 = \sqrt{65,2275 \cdot \left(\frac{1^2}{5} + \frac{(-0,5)^2}{5} + \frac{(-0,5)^2}{5}\right)} = 2,4945$$

$$VC_5 = \sqrt{65,2275 \cdot \left(\frac{(-0,5)^2}{5} + \frac{1^2}{5} + \frac{(-0,5)^2}{5}\right)} = 2,4945$$

$$VC_6 = \sqrt{65,2275 \cdot \left(\frac{(-0,5)^2}{5} + \frac{(-0,5)^2}{5} + \frac{1^2}{5}\right)} = 2,4945$$

Tabulemos los resultados obtenidos hasta ahora, como muestra la tabla 9.2.

i	C_i	VC_i
1	4,0	4,9891
2	1,8	4,9891
3	5,8	4,9891
4	1,1	2,4945
5	4,9	2,4945
6	3,8	2,4945

Tabla 9.2: estimadores y valores críticos para los contrastes de la prueba de comparación de Scheffé.

Finalmente evaluamos cada contraste, comparando el estimador C_i con el valor crítico correspondiente, VC_i. Si $C_i > VC_i$, la comparación es estadísticamente significativa. Podemos ver, entonces, que las comparaciones 3, 5 y 6 son significativas. Esto quiere decir que:

- Existe una diferencia significativa entre las eficiencias de los algoritmos B y C.
- El tiempo promedio de ejecución del algoritmo B es distinto del tiempo promedio de ejecución (combinado) de los algoritmos A y C.
- El tiempo promedio de ejecución del algoritmo C es distinto del tiempo promedio de ejecución (combinado) de los algoritmos A y B.

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 9.4 muestra el ejemplo en R, cuyo resultado se presenta en la figura 9.5. 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. Debemos tener en cuenta que el resultado es ligeramente diferente debido a errores de redondeo. 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.

```
  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 9.5: valores p e intervalos de confianza para las diferencias de las medias obtenidos mediante la prueba de comparación de Scheffé.

Script 9.4: prueba de comparación de Scheffé.

```
library(tidyverse)
library(DescTools)

# Crear el data frame 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 <- data.frame(A, B, C)

# Llevar data frame a formato largo.
datos <- datos %>% pivot_longer(c("A", "B", "C"),
                                names_to = "algoritmo",
                                values_to = "tiempo")

datos[["algoritmo"]] <- factor(datos[["algoritmo"]])
datos[["instancia"]] <- factor(1:nrow(datos))

# Establecer nivel de significación (el mismo usado en ANOVA).
alfa <- 0.025

# Procedimiento ANOVA.
anova <- aov(tiempo ~ algoritmo, data = datos)
```

```r

# Crear matriz de contrastes.
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),
                     nrow=6,
                     byrow = TRUE
)

# Trasponer matriz de contrastes (para que cada contraste sea una columna).
contrastes <- t(contrastes)

# Hacer prueba de Scheffé.
scheffe <- ScheffeTest(x = anova,
                       which = "algoritmo",
                       contrasts = contrastes,
                       conf.level = 1 - alfa
)

print(scheffe)
```

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.

9.5 EJERCICIOS PROPUESTOS

1. Si se tienen datos de tres grupos (A, B y C), ¿por qué no se pueden hacer tres comparaciones con pares de grupos con la prueba t de Student (A-B, A-C, B-C) vista en el capítulo 5?
2. Si SS es la suma de desviaciones cuadradas, ¿qué son SS entre grupos, SS al interior de los grupos y SS total?
3. Define la razón F. ¿Por qué se espera que sea cercana a 1 si es que las poblaciones tienen medias similares?
4. ¿Qué significa que un procedimiento ANOVA sea de una vía?
5. ¿Cuándo un procedimiento ANOVA de una vía es equivalente a una prueba T de Student?
6. ¿Qué sería ANOVA de dos vías?
7. ¿Cuándo aplica el procedimiento ANOVA de una vía para muestras independientes?
8. ¿Cuáles son las hipótesis contrastadas en el procedimiento ANOVA de una vía para muestras independientes?
9. Enumera las condiciones (o supuestos) del procedimiento ANOVA de una vía para muestras independientes para tener confiabilidad.
10. Investiga con más detalle por qué se dice que un procedimiento ANOVA es una prueba omnibus.
11. Investiga en qué consiste, para qué sirve y cómo se aplica en R la prueba de Levene.
12. Investiga algún procedimiento *post-hoc* no abordado en este capítulo, junto con la forma de aplicarlo en R.
13. 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) Independientemente del 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.

REFERENCIAS

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

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

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/

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