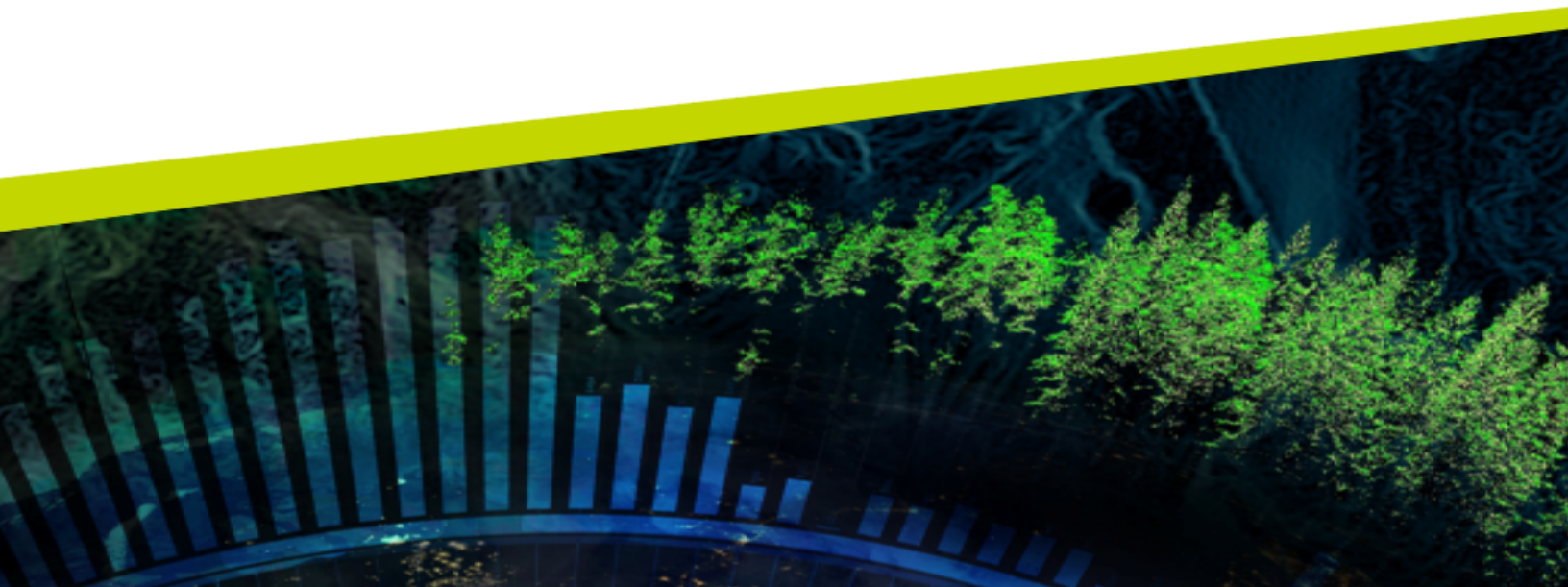
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 11. MÉTODOS CLÁSICOS PARA ANALIZAR VARIABLES NUMÉRICAS PROBLEMÁTICAS

Como ya sabemos, muchos procedimientos estadísticos requieren que los datos cumplan con ciertas propiedades o condiciones, lo que no siempre ocurre. En este capítulo abordaremos algunos métodos para enfrentar estos problemas cuando se intenta inferir sobre las medias poblacionales de variables aleatorias numéricas.

Para ello nos basaremos principalmente en Lane (s.f., pp. 1, 16), Lowry (1999, caps. 11a, 12a, 14a, 15a), Glen (2021b) y Lærd Statistics (2020).

11.1 TRANSFORMACIÓN DE DATOS

Un fenómeno que ocurre a menudo en los estudios, además del incumplimiento de ciertas condiciones, es la necesidad de convertir los datos de una escala a otra diferente. Para hacer tales transformaciones, debemos aplicar una determinada función a una variable aleatoria X, lo que nos entrega como resultado una nueva variable aleatoria Y.

Existen diversos métodos que podemos usar para transformar datos, dependiendo de la forma que tengan los datos originales y la que deseemos obtener como resultado. En esta sección conoceremos, entonces, algunas transformaciones de uso frecuente en estadística.

11.1.1 Transformación lineal

Las **transformaciones lineales** son las más sencillas, y para hacerlas nos basta con aplicar una función lineal de la forma presentada en la ecuación 11.1:, donde m y n son constantes:

$$y_i = m \cdot x_i + n \tag{11.1}$$

La física nos ofrece muchos escenarios en que es necesario aplicar este tipo de transformaciones, pues es el tipo de operación que realizamos cuando convertimos de una unidad a otra. A modo de ejemplo, consideremos la conversión de grados Celsius a grados Fahrenheit: $°F = 1{,}8\ °C + 32$.

En R, podemos hacer este tipo de transformaciones de forma muy sencilla mediante operaciones aritméticas que se aplican vectorialmente, como muestra el script 11.1. En él se transforma un vector con siete temperaturas en grados Celsius a grados Fahrenheit. El resultado se muestra en la figura 11.1.

Script 11.1: transformación lineal para convertir grados Celsius a grados Fahrenheit.

```
# Crear un vector con cuatro observaciones en grados Celsius.
Celsius <- c(-4, 0, 18, 30, 35, 50, 100)

# Aplicar transformación lineal para convertir a grados Fahrenheit.
Fahrenheit <- 1.8 * Celsius + 32

# Mostrar los resultados.
cat("Temperaturas en grados Celsius\n")
print(Celsius)
cat("\nTemperaturas en grados Fahrenheit\n")
print(Fahrenheit)
```

11.1.2 Transformación logarítmica

La transformación logarítmica nos puede servir cuando tenemos distribuciones muy asimétricas, pues ayuda a reducir la desviación y así facilita el cumplimiento de la condición de normalidad requerida por muchas de las pruebas estadísticas que ya conocemos.

```
Temperaturas en grados Celsius
[1]  -4   0  18  30  35  50 100

Temperaturas en grados Fahrenheit
[1]  24.8  32.0  64.4  86.0  95.0 122.0 212.0
```

Figura 11.1: resultado de la transformación lineal del script 11.1.

Para ver este efecto de manera más clara, usaremos un conjunto de datos que registra el peso corporal (en kilogramos) y el peso del cerebro (en gramos) de diversos animales, algunos de ellos extintos (Rousseeuw & Leroy, 1987, p. 57). En R, esta transformación puede hacerse gracias a la función `log(x, base)`, aunque debemos tener cuidado si en los datos pueden haber valores iguales o menores a cero. El script 11.2 aplica esta transformación al peso corporal y al peso cerebral de los animales (líneas 22–23). La figura 11.2.A (líneas 28–32) muestra un histograma del peso cerebral de los animales en su escala original, mientras que la figura 11.2.B (líneas 34–38) muestra gráficamente el resultado de la transformación logarítmica a estos pesos.

Muchas veces la transformación logarítmica hace que nos sea más fácil interpretar un conjunto de datos, evidenciando patrones más claros para la relación entre variables. En la figura 11.2.D (script 11.2, líneas 48–52) se evidencia una clara relación entre el peso corporal y el peso del cerebro después de transformar ambas variables, relación que no podemos percibir con los datos originales (figura 11.2.C, líneas 42–46).

Sin embargo, tenemos que ser cuidadosos al usar esta transformación porque, a diferencia de la trasformación lineal, cuando comparamos medias a través de una prueba estadística paramétrica en datos que han sido sometidos a una transformación logarítmica, en realidad estamos comparando ¡**medias geométricas**! Recordemos que la media geométrica se calcula de acuerdo a la ecuación 11.2 y suele ocuparse para representar tasas de crecimiento o de interés (Glen, 2021a).

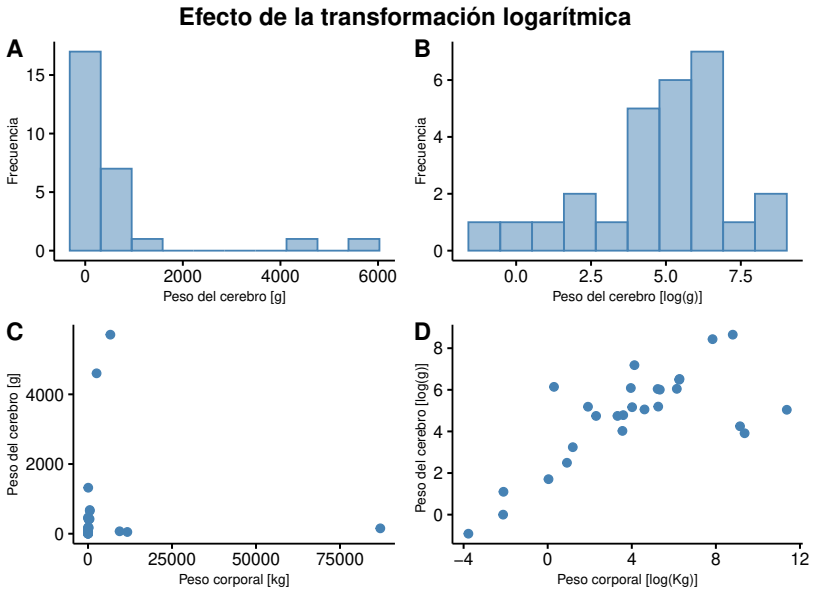


Figura 11.2: histogramas del peso cerebral antes y después de la transformación logarítmica (fila superior). Gráficos de dispersión para el peso corporal y el peso del cerebro antes y después de las transformaciones logarítmicas (fila inferior).

Script 11.2: transformación logarítmica.

```
1  library(ggpubr)
2
3  # Crear estructura con los datos
4  animal <- c("Mountain beaver", "Cow", "Grey wolf", "Goat", "Guinea pig",
5              "Dipliodocus", "Asian elephant", "Donkey", "Horse",
6              "Potar monkey", "Cat", "Giraffe", "Gorilla", "Human",
7              "African elephant", "Triceratops", "Rhesus monkey", "Kangaroo",
8              "Golden hamster", "Mouse", "Rabbit", "Sheep", "Jaguar",
```

```
            "Chimpanzee", "Brachiosaurus", "Mole", "Pig")
peso_cuerpo <- c(1.35, 465, 36.33, 27.66, 1.04, 11700, 2547, 187.1, 521, 10,
                 3.3, 529, 207, 62, 6654, 9400, 6.8, 35, 0.12, 0.023, 2.5,
                 55.5, 100, 52.16, 87000, 0.122, 192)
peso_cerebro <- c(465, 423, 119.5, 115, 5.5, 50, 4603, 419, 655, 115, 25.6,
                  680, 406, 1320, 5712, 70, 179, 56, 1, 0.4, 12.1, 175, 157,
                  440, 154.5, 3, 180)

datos <- data.frame(animal, peso_cuerpo, peso_cerebro)

# Aplicar transformación logarítmica en base e (no hay valores nulos ni negativos)
peso_cuerpo_logaritmo <- log(peso_cuerpo)
peso_cerebro_logaritmo <- log(peso_cerebro)
datos <- data.frame(datos, peso_cuerpo_logaritmo, peso_cerebro_logaritmo)

# Histogramas para el peso cerebral antes y después
# de la transformación logarítmica.
h1 <- gghistogram(datos, x = "peso_cerebro", bins = 10,
                  xlab = "Peso del cerebro [g]", ylab = "Frecuencia",
                  color = "steelblue", fill = "steelblue")
h1 <- h1 + theme(axis.title = element_text(size = rel(0.7)),
                 axis.text = element_text(size = rel(0.8)))

h2 <- gghistogram(datos, x = "peso_cerebro_logaritmo", bins = 10,
                  xlab = "Peso del cerebro [log(g)]", ylab = "Frecuencia",
                  color = "steelblue", fill = "steelblue")
h2 <- h2 + theme(axis.title = element_text(size = rel(0.7)),
                 axis.text = element_text(size = rel(0.8)))

# Gráficos de dispersión para la relación entre peso corporal y peso del
# cerebro, antes y después de aplicar la transformación logarítmica.
g1 <- ggscatter(datos, x = "peso_cuerpo", y = "peso_cerebro",
                color = "steelblue", xlab = "Peso corporal [kg]",
                ylab = "Peso del cerebro [g]")
g1 <- g1 + theme(axis.title = element_text(size = rel(0.7)),
                 axis.text = element_text(size = rel(0.8)))

g2 <- ggscatter(datos, x = "peso_cuerpo_logaritmo", y = "peso_cerebro_logaritmo",
                color = "steelblue", xlab = "Peso corporal [log(Kg)]",
                ylab = "Peso del cerebro [log(g)]  ")
g2 <- g2 + theme(axis.title = element_text(size = rel(0.7)),
                 axis.text = element_text(size = rel(0.8)))

# Crear una única figura con los gráficos y anotarla con un título
grafico  <- ggarrange(h1, h2, g1, g2, nrow = 2, ncol = 2, labels="AUTO")

texto <- "Efecto de la transformación logarítmica"
titulo  <- text_grob(texto, face = "bold", size = 14)
grafico <- annotate_figure(grafico, top = titulo)

print(grafico)
```

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \tag{11.2}$$

Para ilustrar esta idea, supongamos que aplicamos una transformación logarítmica con base 10 al vector $[1, 10, 100]$, obteniendo como resultado $[0, 1, 2]$. La media aritmética del vector transformado es: $(0+1+2)/3 = 1$. Al revertir la transformación para esta media aritmética, tenemos que $10^1 = 10$, lo que es distinto a la media aritmética de los datos originales: $(1 + 10 + 100)/3 = 37$. A su vez, la media geométrica del vector original es: $\sqrt[3]{1 \cdot 10 \cdot 100} = 10$.

Luego, si contrastamos hipótesis paramétricas sobre medias en una o más variables a las que se ha aplicado la transformación logarítmica, entonces las estimaciones obtenidas se refieren a las medias geométricas de las variables originales. Así, podemos reportar, por ejemplo, intervalos de confianza para la diferencia de

medias geométricas, que no son fácilmente transformables a las escalas originales. Sin embargo, se encuentran diferencias significativas con medias geométricas **si y solo si estas diferencias existen** con las medias aritméticas originales. Es necesario, entonces, tener en cuenta estas relaciones al momento de **redactar las conclusiones**.

11.1.3 Escalera de potencias de Tukey

Más general que la transformación logarítmica, la **escalera de potencias de Tukey** nos ayuda a cambiar la forma de una distribución asimétrica para que se asemeje a la normal. Este método consiste en explorar relaciones de la forma que muestra la ecuación 11.3:

$$x' = x^{\lambda} \tag{11.3}$$

donde x es la observación original, x' es la observación transformada, y λ puede tomar cualquier valor real que **se escoge** de modo que la distribución de los datos transformados sea lo más cercana posible a la normal. También es útil al explorar la relación entre dos variables, en cuyo caso se busca obtener un gráfico de dispersión en que los puntos se asemejen a una recta.

Formalmente, la transformación de Tukey se define según la ecuación 11.4:

$$x'_{\lambda} = \begin{cases} x^{\lambda} & \lambda > 0 \\ \log(x) & \lambda = 0 \\ -(x^{\lambda}) & \lambda < 0 \end{cases} \tag{11.4}$$

Sin embargo, por la falta de computadores en la época de su desarrollo, suelen considerarse únicamente las trasformaciones que se presentan en la tabla 11.1. Fijémonos en que si $\lambda = 1$, no se realiza transformación alguna, y que para el caso de $\lambda = 0$, se tiene que $x^0 = 1$, por lo que se reemplaza en este caso por la transformación logarítmica.

λ	-2	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1	2
x'	$-\frac{1}{x^2}$	$-\frac{1}{x}$	$-\frac{1}{\sqrt{x}}$	$\log(x)$	$\sqrt{x}$	x	x^2

Tabla 11.1: escalera de transformaciones de Tukey.

Usemos ahora la población total de Estados Unidos entre los años 1610 y 1850 (United States Census Bureau, 2004, 2021) como ejemplo para entender mejor esta transformación. El script 11.3 define (líneas 6–14) una matriz con estos datos y genera (líneas 18–33) un histograma para la población, en millones de habitantes, mostrado en la figura 11.3.A, y un gráfico de dispersión para la población por año, mostrado en la figura 11.3.B. Podemos ver claramente que la distribución del número de habitantes presenta una fuerte asimetría positiva (o hacia la derecha) y que la población parece aumentar de manera exponencial a partir desde comienzos del siglo XVIII.

Script 11.3: datos de la población total de Estados Unidos.

```
1  library(ggpubr)
2  library(latex2exp)
3  library(rcompanion)
4
5  # Crear estructura con los datos
6  Year <- c(1610, 1620, 1630, 1640, 1650, 1660, 1670, 1680, 1690, 1700, 1710,
7            1720, 1730, 1740, 1750, 1760, 1770, 1780, 1790, 1800, 1810, 1820,
8            1830, 1840, 1850)
9  Population <- c( 0.00035,  0.002302, 0.004646, 0.026634,  0.050368,  0.075058,
10                  0.111935, 0.151507, 0.210372, 0.250888,  0.331711,  0.466185,
11                  0.629445, 0.905563, 1.170760, 1.593625,  2.148076,  2.780369,
12                  3.929214, 5.308483, 7.239881, 9.638453, 12.866020, 17.069453,
13                 23.191876)
14 datos <- data.frame(Year, Population)
15
16 # Obtener un histograma y gráfico de dispersión de la población por año
```

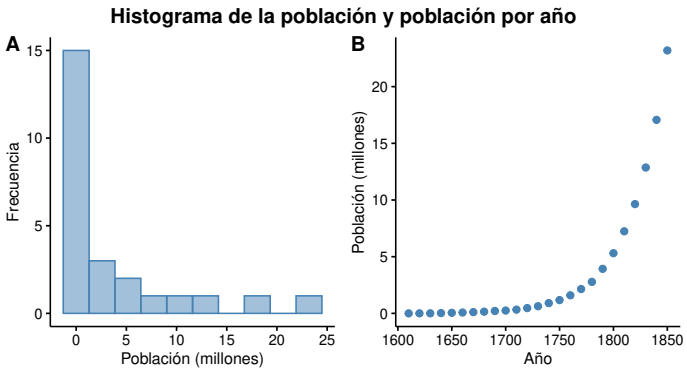


Figura 11.3: histograma de la población histórica de Estados Unidos y gráfico de dispersión de la población por año.

```r
# con los datos originales.
ho <- gghistogram(datos, x = "Population", bins = 10,
                  xlab = "Población (millones)", ylab = "Frecuencia",
                  color = "steelblue", fill = "steelblue") +
      theme(axis.title = element_text(size = rel(0.9)),
            axis.text = element_text(size = rel(0.8)))
go <- ggscatter(datos, x = "Year", y = "Population", color = "steelblue",
                xlab = "Año", ylab = "Población (millones)") +
      theme(axis.title = element_text(size = rel(0.9)),
            axis.text = element_text(size = rel(0.8)))

# Unir y mostrar el histograma y el gráfico de dispersión
original  <- ggarrange(ho, go, ncol = 2, nrow = 1, labels = "AUTO")
texto <- "Histograma de la población y población por año"
titulo  <- text_grob(texto, face = "bold", size = 14)
original <- annotate_figure(original, top = titulo)
print(original)
```

La figura 11.4, producida con el script 11.4, muestra gráficos de dispersión para la población por año tras aplicar la transformación de Tukey a la primera variable con diferentes valores de λ. En ella podemos observar que la curva cambia gradualmente de convexa a cóncava a medida que aumenta el valor de λ. Para obtener un resultado que sea lo más cercano posible a una línea recta, se debe resolver un problema de optimización que minimice los valores residuales tras ajustar una recta a los puntos transformados. De los gráficos presentados en la figura 11.4, el más cercano a una recta se obtuvo para $\lambda = 0$.

Como se dijo anteriormente, al aplicar la transformación de Tukey con este λ óptimo se debería obtener una distribución cercana a la normal, lo que permite cumplir el requisito de normalidad que imponen muchas pruebas estadísticas, permitiéndonos así lograr resultados más confiables con estos datos transformados.

Afortunadamente la búsqueda del valor óptimo de λ ya se encuentra implementada en R en la función `transformTukey(x, start, end, int, plotit, verbose, quiet, statistic, returnLambda)` incluida en el paquete `rcompanion`, donde:

- `x`: vector de valores a transformar.
- `start`: valor inicial de λ para la búsqueda. Si se omite, toma valor -10.
- `end`: valor final de λ para la búsqueda. Si se omite, toma valor 10.
- `int`: intervalo (paso) entre los valores de λ para la búsqueda. Si se omite, toma valor 0,025.
- `plotit`: si toma valor `TRUE`, que es su valor por omisión, entrega los siguientes gráficos:
  - Estadístico de la prueba de normalidad versus λ.
  - Histograma de los valores transformados.
  - Gráfico Q-Q de los valores transformados.
- `verbose`: si toma valor `TRUE`, muestra información adicionar sobre la prueba de normalidad con respecto a λ. Si se omite toma valor `FALSE`.
- `quiet`: si toma valor `TRUE`, no muestra información alguna por pantalla. Toma valor `FALSE` si se omite.
- `statistic`: si toma valor 1 (valor por omisión) usa la prueba de normalidad de Shapiro-Wilk. Con valor

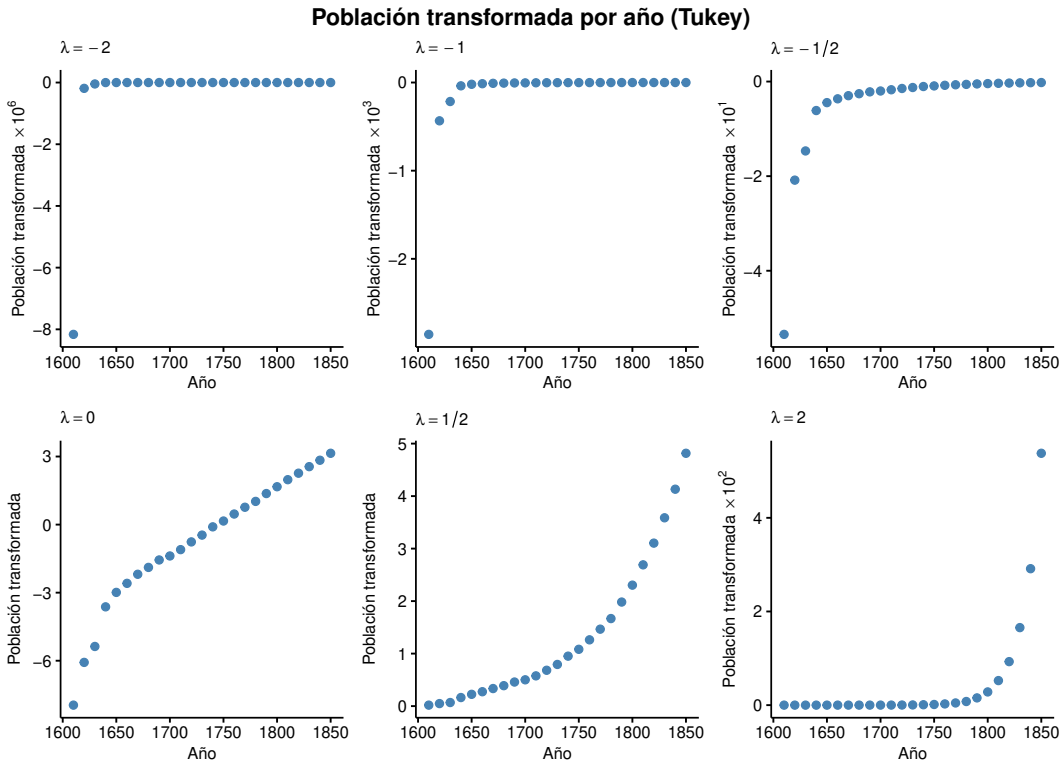


Figura 11.4: gráficos de dispersión de la población de Estados Unidos por año tras aplicar la transformación de Tukey con distintas escalas.

2, usa la prueba de Anderson-Darling.

- `returnLambda`: si toma valor `TRUE`, devuelve el valor de λ. Si toma valor `FALSE`, que es el valor por omisión, devuelve los datos transformados.

Podemos ver que el comportamiento de esta función es bastante flexible gracias a sus argumentos opcionales. En su versión más simple, `transformTukey(x, plotit = FALSE, quiet = TRUE))`, puede utilizarse para obtener los datos luego de aplicar la transformación de Tukey con el valor óptimo de λ que encuentre.

Script 11.4: (continuación del script 11.3) aplicación de la transformacion de Tukey a los datos de la población de Estados Unidos.

```
# Transformaciones de la población para diferentes escalas de Tukey
lambda_menos_dos <- -1 / (datos[["Population"]] ** 2) / 1000000
lambda_menos_uno <- -1 / datos[["Population"]] / 1000
lambda_menos_un_medio <- -1 / sqrt(datos[["Population"]]) / 10
lambda_cero <- log(datos[["Population"]])
lambda_un_medio <- sqrt(datos[["Population"]])
lambda_dos <- datos[["Population"]] ** 2 / 100

transformaciones <- data.frame(Year, lambda_menos_dos, lambda_menos_uno,
                               lambda_menos_un_medio, lambda_cero,
                               lambda_un_medio, lambda_dos)

# Gráficos de dispersión para la transformación de Tukey de la población y el
# año, usando distintos valores de lambda.
gt1 <- ggscatter(transformaciones, x = "Year", y = "lambda_menos_dos",
                 color = "steelblue", xlab = "Año") +
       ylab(TeX(r'(Población transformada $\times 10^{6}$)')) +
       labs(subtitle = TeX(r'($\lambda = -2$)')) +
       theme(title = element_text(size = rel(0.8)),
             axis.text = element_text(size = rel(0.8)))
gt2 <- ggscatter(transformaciones, x = "Year", y = "lambda_menos_uno",
                 color = "steelblue", xlab = "Año") +
```

```
       ylab(TeX(r'(Población transformada $\times 10^{3}$)')) +
       labs(subtitle = TeX(r'($\lambda = -1$)')) +
       theme(title = element_text(size = rel(0.8)),
             axis.text = element_text(size = rel(0.8)))
gt3 <- ggscatter(transformaciones, x = "Year", y = "lambda_menos_un_medio",
                 color = "steelblue", xlab = "Año") +
       ylab(TeX(r'(Población transformada $\times 10^{1}$)')) +
       labs(subtitle = TeX(r'($\lambda = -1/2$)')) +
       theme(title = element_text(size = rel(0.8)),
               axis.text = element_text(size = rel(0.8)))
gt4 <- ggscatter(transformaciones, x = "Year", y = "lambda_cero",
                 color = "steelblue", xlab = "Año",
                 ylab = "Población transformada") +
       labs(subtitle = TeX(r'($\lambda = 0$)')) +
       theme(title = element_text(size = rel(0.8)),
             axis.text = element_text(size = rel(0.8)))
gt5 <- ggscatter(transformaciones, x = "Year", y = "lambda_un_medio",
                 color = "steelblue", xlab = "Año",
                 ylab = "Población transformada") +
       labs(subtitle = TeX(r'($\lambda = 1 / 2$)')) +
       theme(title = element_text(size = rel(0.8)),
             axis.text = element_text(size = rel(0.8)))
gt6 <- ggscatter(transformaciones, x = "Year", y = "lambda_dos",
                 color = "steelblue", xlab = "Año") +
       ylab(TeX(r'(Población transformada $\times 10^{2}$)')) +
       labs(subtitle = TeX(r'($\lambda = 2$)')) +
       theme(title = element_text(size = rel(0.8)),
             axis.text = element_text(size = rel(0.8)))

# Crear y mostrar una única figura con todos los gráficos de dispersión
dispersiones  <- ggarrange(gt1, gt2, gt3, gt4, gt5, gt6,
                           ncol = 3, nrow = 2, align = "hv")
texto <- "Población transformada por año (Tukey)"
titulo  <- text_grob(texto, face = "bold", size = 14)
dispersiones <- annotate_figure(dispersiones, top = titulo)
print(dispersiones)
```

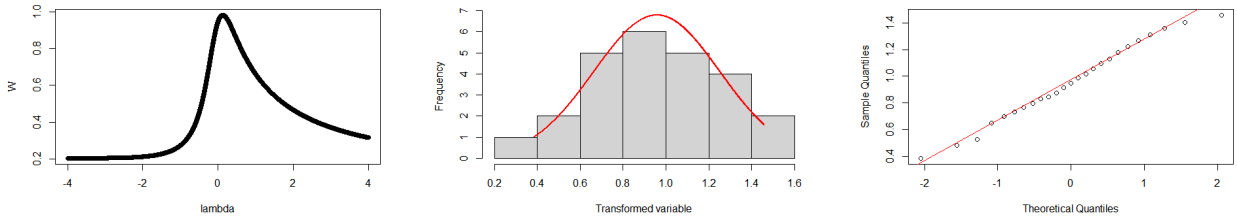
Pero es inusual, ya que normalmente uno debería **reportar** el valor de la λ utilizado en la transformación de los datos. La siguiente sentencia permite aplicar esta búsqueda a los datos del ejemplo de la población total de Estados Unidos (asumiendo que se tiene la matriz de datos creada en las líneas 6–14 del script 11.3): `transformTukey(datos[["Population"]], start = -1, end = 1, int = 0.001)`. Hacemos la búsqueda de la escala óptima en un intervalo más reducido: $\lambda \in [-1; 1]$, ya que sabemos, de la figura 11.4, que el valor óptimo no debería estar alejado del valor cero. Esta sentencia genera (por separado) los gráficos de la figura 11.5 y entrega la siguiente salida en pantalla:

```
     lambda      W Shapiro.p.value
1121   0.12 0.9814          0.9107

if (lambda >  0){TRANS = x ^ lambda}
if (lambda == 0){TRANS = log(x)}
if (lambda <  0){TRANS = -1 * x ^ lambda}

 [1] 0.3848471 0.4824504 0.5248673 0.6472207 0.6986482 0.7329049 0.7689107
 [8] 0.7973554 0.8293895 0.8471057 0.8759739 0.9124877 0.9459647 0.9881668
[15] 1.0190985 1.0575145 1.0960893 1.1305566 1.1784650 1.2217897 1.2681417
[22] 1.3124443 1.3587304 1.4056149 1.4582779
>
```

La figura 11.5a nos muestra gráficamente que el valor óptimo λ^* corresponde al que maximiza el estadístico entregado por la prueba de normalidad. A su vez, en la figura 11.5b vemos que la distribución obtenida aplicando λ^* se asemeja mucho más a una distribución normal, lo que se ve confirmado por el gráfico Q-Q de los mismos datos de la figura 11.5c. La salida por pantalla, por otra parte, nos reporta que el valor óptimo de lambda es $\lambda^* = 0{,}12$, con lo que se obtienen datos transformados que cumplen con el supuesto de normalidad según la prueba de Shapiro y Wilk ($W = 0{,}981$; $p = 0{,}912$).



(a) estadístico W de la prueba de Shapiro-Wilk por cada valor de λ.

(b) histograma de los datos transformados con λ^*.

(c) gráfico Q-Q de los datos transformados con λ^*.

Figura 11.5: gráficos entregados por la función `transformTukey()` para los datos de población de Estados Unidos.

Una vez más debemos tener en cuenta la transformación realizada al momento de interpretar los resultados. Si bien tenemos certeza que si se encuentran diferencias significativas en la variable transformada, estas diferencias **también existen en la variable original**, los estadísticos y los intervalos de confianza **no son los mismos** que arrojarían las pruebas con los datos originales.

11.1.4 Transformaciones Box-Cox

La transformación de Box y Cox, más conocida como la **transformación Box-Cox**, es una versión escalada de la transformación de Tukey, dada por la ecuación 11.5:

$$x''_\lambda = \frac{x^\lambda - 1}{\lambda} \tag{11.5}$$

Notemos que aplicando la expansión de la serie de Taylor, podemos reescribir la ecuación 11.5 como:

$$x''_\lambda = \frac{x^\lambda - 1}{\lambda} \approx \frac{\left(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \ldots\right) - 1}{\lambda}$$

de donde:

$$\lim_{\lambda \to 0} \frac{\left(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \ldots\right) - 1}{\lambda} \quad \text{toma la forma} \quad \frac{0}{0}$$

que tras aplicar la regla de l'Hôpital, finalmente se obtiene:

$$\lim_{\lambda \to 0} x''_\lambda = \log(x)$$

por lo que, al igual que en la escalera de potencias de Tukey, pero de forma más natural, empleamos la transformación logarítmica para $\lambda = 0$.

Si reemplazamos las líneas 35–41 del script 11.4 por las que presenta el script 11.5, puede obtenerse la figura 11.6 que muestra gráficos de dispersión para la población total de Estados Unidos por año tras aplicar la transformación de Box-Cox con diferentes valores de λ. Podemos ver que el resultado se parece mucho al que obtuvimos con la transformación de Tukey (figura 11.4).

Script 11.5: aplicación de transformaciones Box-Cox a los datos de la población de Estados Unidos.

```
35 # Transformaciones Box-Cox de la población para diferentes valores de lambda
36 lambda_menos_dos <- (datos[["Population"]] ** -2) / -2 / 1000000
37 lambda_menos_uno <- (datos[["Population"]] ** -1) / -1 / 1000
38 lambda_menos_un_medio <- (datos[["Population"]] ** (-1/2)) / (-1/2) / 10
39 lambda_cero <- log(datos[["Population"]])
40 lambda_un_medio <- (datos[["Population"]] ** (1/2)) / (1/2)
41 lambda_dos <- (datos[["Population"]] ** 2) / 2 / 100
```

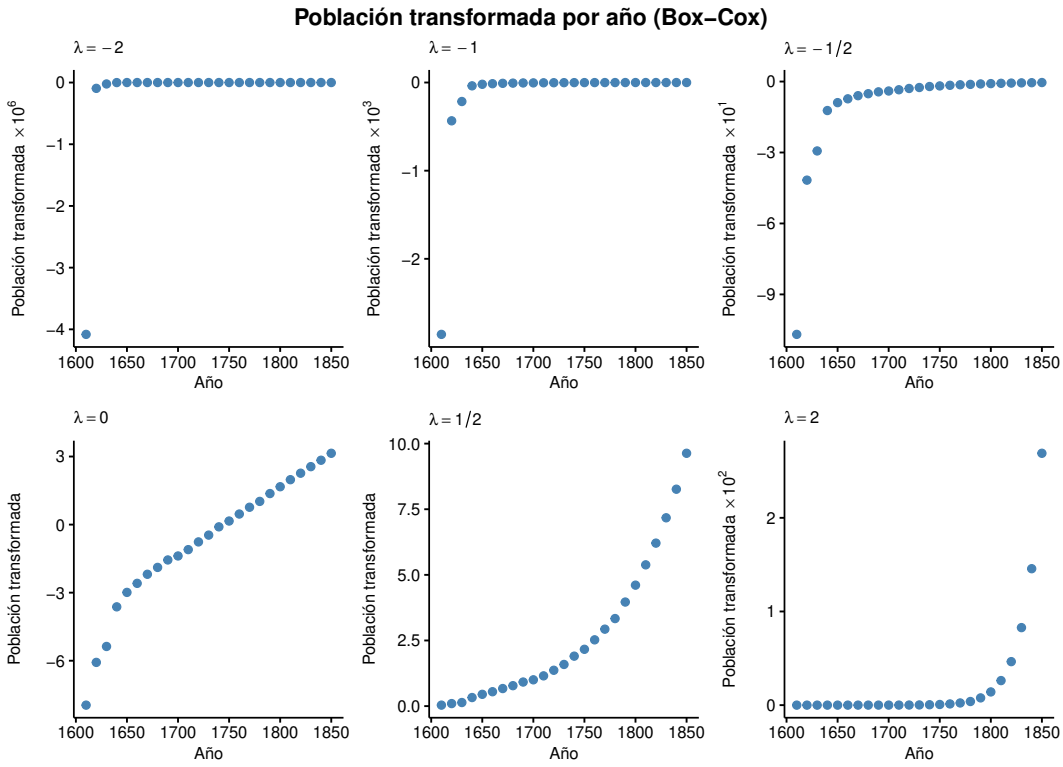


Figura 11.6: población de Estados Unidos por año tras aplicar la transformación de Box-Cox con distintos valores de λ.

Una característica interesante de esta transformación es que $x''_\lambda = 0$ cuando $x = 1$ para cualquier valor de λ. Podemos observar esto claramente en la figura 11.7, que compara transformaciones Box-Cox con distintos valores de λ en el intervalo $[0{,}5; 2{,}0]$.

El paquete `DescTools` de R incluye varias funciones que permiten efectuar la transformación Box-Cox (Carchedi et al., s.f.). Destacan entre ellas:

- `BoxCoxLambda(x, lower, upper)`: devuelve el valor óptimo de λ para la transformación Box-Cox del vector `x`.
- `BoxCox(x, lambda)`: devuelve un vector correspondiente a la transformación Box-Cox de `x` con parámetro `lambda`.
- `BoxCoxInv(x, lambda)`: revierte la transformación Box-Cox del vector `x` con parámetro `lambda`.

Para estas funciones:

- `x`: vector numérico con los valores originales.
- `lower`: límite inferior para los posibles valores de λ.
- `upper`: límite superior para los posibles valores de λ.
- `lambda`: parámetro de la transformación.

El script 11.6 muestra cómo usar estas funciones en los datos del ejemplo de la población de Estados Unidos. En la línea 95 se determina el valor óptimo del parámetro λ, para luego utilizarlo al aplicar la correspondiente transformación Box-Cox en la línea 96. La figura 11.8, generada en las líneas 100–122 del script, muestra gráficamente el resultado de la transformación.

Script 11.6: aplicación de transformaciones Box-Cox a los datos de la población de Estados Unidos.

```
94 # Buscar y aplicar la mejor transformación Box-Cox usando funciones de R
95 lambda <- BoxCoxLambda(datos[["Population"]], lower = -4, upper = 4)
96 transformacion <- BoxCox(datos[["Population"]], lambda)
97 datos <- data.frame(datos, transformacion)
98
```

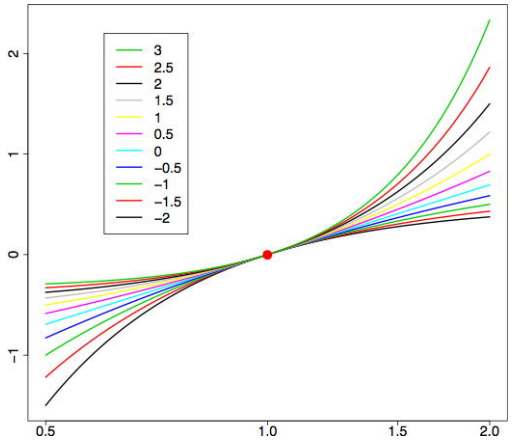


Figura 11.7: ejemplos de la transformación Box-Cox. Fuente: (Lane, s.f., p. 16).

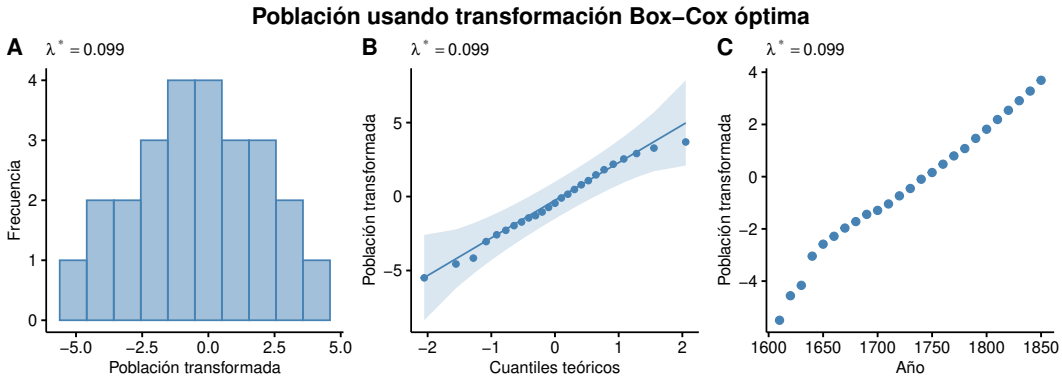


Figura 11.8: población de Estados Unidos por año tras aplicar la transformación de Box-Cox con λ óptimo.

```r
# Crear gráficos de los datos transformados con el lambda óptimo
h1 <- gghistogram(datos, bins = 10, x = "transformacion", color = "steelblue",
                  fill = "steelblue", xlab = "Población transformada", ylab = "
                     Frecuencia") +
      labs(subtitle = TeX(sprintf(r'($\lambda^{*} = %.3f$)', lambda))) +
      theme(title = element_text(size = rel(0.8)),
            axis.text = element_text(size = rel(0.8)))

q1 <- ggqqplot(transformacion, color = "steelblue",
               xlab = "Cuantiles teóricos", ylab = "Población transformada") +
      labs(subtitle = TeX(sprintf(r'($\lambda^{*} = %.3f$)', lambda))) +
      theme(title = element_text(size = rel(0.8)),
            axis.text = element_text(size = rel(0.8)))

g1 <- ggscatter(datos, x = "Year", y = "transformacion", color = "steelblue",
                xlab = "Año", ylab = "Población transformada")  +
      labs(subtitle = TeX(sprintf(r'($\lambda^{*} = %.3f$)', lambda))) +
      theme(title = element_text(size = rel(0.8)),
            axis.text = element_text(size = rel(0.8)))

# Crear y mostrar una única figura con los tres gráficos con datos transformados
dispersion  <- ggarrange(h1, q1, g1, ncol = 3, nrow = 1, labels = "AUTO")
texto <- "Población usando transformación Box-Cox óptima"
titulo  <- text_grob(texto, face = "bold", size = 14)
dispersion <- annotate_figure(dispersion, top = titulo)
print(dispersion)
```

En el gráfico 11.8.A se presenta un histograma de los datos transformados, que es característico de una

distribución normal, lo que es confirmado por el gráfico Q-Q en 11.8.B. De manera congruente, podemos ver en el gráfico 11.8.C que la relación entre la población transformada y el año se asemeja a una recta. Estos gráficos también nos dicen que el valor óptimo del parámetro resultó ser $\lambda^* \approx 0{,}099$.

11.2 PRUEBAS NO PARAMÉTRICAS CON UNA Y DOS MUESTRAS NUMÉRICAS

En el capítulo 8 conocimos algunos métodos no paramétricos que podemos usar para inferir sobre frecuencias cuando nuestro conjunto de datos no cumple con las condiciones para poder usar, por ejemplo, la prueba paramétrica de Wilson. Mencionamos que este problema también puede ocurrir cuando se intenta inferir con medias, por lo que en esta sección conoceremos alternativas no paramétricas para la pruebas t de Student (con una y dos muestras).

11.2.1 Prueba de suma de rangos de Wilcoxon

En el capítulo 5 aprendimos que la prueba t de Student es adecuada para inferir sobre la media de una población a partir de una muestra de observaciones siempre y cuando se verifiquen dos condiciones:

1. Las observaciones elegidas son independientes entre sí.
2. Las observaciones provienen de una población con distribución cercana a la normal.

También vimos que esta prueba se podía extender a inferencias sobre la diferencia de las medias de dos poblaciones a partir de muestras independientes de cada una de ellas cuando se cumple que:

1. Cada muestra cumple las condiciones para usar la distribución t mencionadas arriba.
2. Las muestras son independientes entre sí.

Es importante mencionar también que la distribución normal es continua, por lo que la escala de medición empleada por la variable debe ser, al menos, de intervalos iguales.

Al igual que con las pruebas paramétricas con proporciones, como vimos en el capítulo 8, si usamos la prueba t en un escenario en que no se cumple alguna de estas condiciones, no hay garantías de que el resultado sea válido y, en consecuencia, las conclusiones que se obtengan a partir de ella podrían ser equivocadas.

La **prueba de suma de rangos de Wilcoxon**, también llamada **prueba U de Mann-Whitney** o **prueba de Wilcoxon-Mann-Whitney**, es una alternativa no paramétrica a la prueba t de Student para una o dos muestras independientes. Esta prueba requiere verificar el cumplimiento de las siguientes condiciones:

1. Las observaciones de ambas muestras son independientes.
2. La escala de medición empleada debe ser a lo menos ordinal, de modo que tenga sentido hablar de relaciones de orden ("igual que", "menor que", "mayor o igual que").

En general es más frecuente utilizar esta prueba con dos muestras, que suelen llamarse "muestra A" y "muestra B" en los textos de estudio. Para darle mayor contexto, consideremos el siguiente ejemplo: una empresa de desarrollo de software desea evaluar la usabilidad de dos interfaces alternativas, A y B, para un nuevo producto de software. Con este fin, la empresa ha seleccionado al azar a 23 voluntarias y voluntarios, que no se conocen, quienes son asignados de manera aleatoria a dos grupos, cada uno de los cuales debe probar una de las interfaces ($n_A = 12$, $n_B = 11$). Cada participante debe evaluar 6 aspectos de usabilidad de la interfaz, cada uno de los cuales se mide con una escala Likert de 7 puntos, donde 1 significa "muy malo" y 7, "muy bueno". La valoración que cada participante da a la interfaz evaluada, llamado "índice de usabilidad", corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. La tabla 11.2 muestra los índices de usabilidad otorgados por cada participante.

Interfaz *A*	2,7	6,6	1,6	5,1	3,7	6,1	5,0	1,4	1,8	1,5	3,0	5,3	$\overline{x}_A = 3{,}65$
Interfaz *B*	5,0	1,4	5,6	4,6	6,7	2,7	1,3	6,3	3,7	1,3	6,8		$\overline{x}_B = 4{,}13$

Tabla 11.2: muestras de índices de usabilidad para las interfaces de uso A y B.

En este caso, si bien se cumple la condición de independencia de la prueba t de Student, no podemos usar esta prueba por dos razones: primero, no todas las escalas Likert pueden asegurar que son de igual intervalo. En el ejemplo, si dos participantes califican un aspecto de la interfaz A con notas 3 y 5, mientras que dos

participantes califican esos aspectos con notas 4 y 6 para la interfaz B, ¿se podría asegurar que en ambos casos los participantes consideran que existe la misma diferencia de usabilidad (representada por 2 puntos)? Pocas escalas Likert tienen estudios de reproducibilidad que aseguren esta consistencia, por lo que **no deberíamos asumir** que la escala es de intervalos iguales en este ejemplo. En segundo lugar, al revisar los histogramas para las muestras (figura 11.9) podemos observar que las distribuciones no se asemejan a una normal.

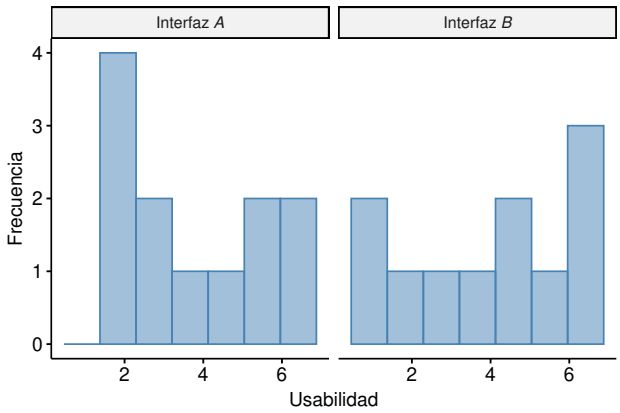


Figura 11.9: histogramas de las muestras descritas en la tabla 11.2.

Como alternativa, podemos usar la prueba no paramétrica de Wilcoxon-Mann-Whitney, cuyas hipótesis para el ejemplo son:

H_0: no hay diferencia en la usabilidad de ambas interfaces (los valores de los índice de usabilidad se distribuyen de igual forma).
H_A: sí hay diferencia en la usabilidad de ambas interfaces (las distribuciones de los índices de usabilidad son distintas).

Notemos que, al igual que en el caso de la prueba χ^2 de Pearson, estas hipótesis no hacen referencia a algún parámetro de una supuesta distribución para los índices de usabilidad, es decir, nos entregan **menos información** que la prueba paramétrica equivalente.

El primer paso del procedimiento consiste en combinar todas las observaciones en un único conjunto de tamaño $n_T = n_A + n_B$ y ordenarlo de menor a mayor. A cada elemento se le asigna un **valor de rango** (*rank* en inglés, posición en el *ranking* en chileno) de 1 a n_T, de acuerdo a la posición que ocupa en el conjunto ordenado. En caso de que un valor aparezca más de una vez, cada repetición toma como valor el rango promedio de todas las ocurrencias del valor. La tabla 11.3 muestra el resultado de este proceso. Podemos notar que hay dos observaciones del valor 1,3 a las que le corresponderían los rangos 1 y 2, por lo que, en consecuencia, ambas reciben el mismo valor de rango, igual al promedio 1,5. Esto también ocurre para las puntuaciones 1,4; 2,7; 3,7 y 5,0.

Muestra	B	B	A	B	A	A	A	A	B	A	A	B	B	A	B	A	A	B	A	B	A	B	B
Observación	1,3	1,3	1,4	1,4	1,5	1,6	1,8	2,7	2,7	3,0	3,7	3,7	4,6	5,0	5,0	5,1	5,3	5,6	6,1	6,3	6,6	6,7	6,8
Rango	1,5	1,5	3,5	3,5	5,0	6,0	7,0	8,5	8,5	10,0	11,5	11,5	13,0	14,5	14,5	16,0	17,0	18,0	19,0	20,0	21,0	22,0	23,0

Tabla 11.3: muestra combinada del ejemplo con rango.

Aquí cabe hacer una aclaración. En la literatura con frecuencia se menciona que otra condición para aplicar la prueba de suma de rangos de Wilcoxon es que la escala de medición sea **intrínsecamente continua**, con un número arbitrario de decimales. Esta condición permite suponer que **no habrá empates** al construir los rangos. Sin embargo, implementaciones actuales de la prueba son capaces de manejar la presencia de un **número acotado** de empates, realizando ciertos ajustes, por lo que también se puede aplicar a datos **ordinales** y **discretos**. Sin embargo, es importante tener en cuenta que cuando hay empates, las correcciones introducidas pueden afectar la precisión y potencia de la prueba.

Volviendo al procedimiento, a continuación se suman los rangos asociados a las observaciones de cada muestra, y para la muestra combinada. Así, para la muestra A del ejemplo obtenemos:

$$S^A = 3{,}5 + 5{,}0 + 6{,}0 + 7{,}0 + 8{,}5 + 10{,}0 + 11{,}5 + 14{,}5 + 16{,}0 + 17{,}0 + 19{,}0 + 21{,}0 = 139$$

y para la muestra B:

$$S^B = 1{,}5 + 1{,}5 + 3{,}5 + 8{,}5 + 11{,}5 + 13{,}0 + 14{,}5 + 18{,}0 + 20{,}0 + 22{,}0 + 23{,}0 = 137$$

La suma de rangos para la muestra combinada está dada por la ecuación 11.6:

$$S^T = S^A + S^B = \frac{n_T\,(n_T + 1)}{2} \tag{11.6}$$

Para el ejemplo:

$$S^T = 139 + 137 = \frac{23 \cdot (23 + 1)}{2} = 276$$

Trabajar con los rangos en lugar de las observaciones nos ofrece **dos ventajas**: la primera es que el foco solo está en las **relaciones de orden** entre las observaciones, sin necesidad de que estas provengan de una escala de intervalos iguales; la segunda es que esta transformación facilita conocer de manera sencilla algunas **propiedades del conjunto de datos**. Por ejemplo, la suma de rangos de la muestra combinada se determina siempre mediante la ecuación 11.6 y el rango promedio es siempre como muestra la ecuación 11.7:

$$\overline{R} = \frac{S^T}{n_T} = \frac{n_T\,(n_T + 1)}{2\,n_T} = \frac{n_T + 1}{2} \tag{11.7}$$

Para el ejemplo entonces:

$$\overline{R} = \frac{276}{23} = \frac{23 + 1}{2} = 12$$

Si la hipótesis nula fuera cierta, las observaciones en ambas muestras serían similares, por lo que ambas muestras se mezclarían de manera más o menos alternada al ordenar la muestra combinada. En consecuencia, deberíamos esperar que los promedios de rangos para cada muestra se aproximen al rango promedio de la muestra combinada. Esto es equivalente a que la suma de los rangos de cada muestra contribuya de igual forma a la suma total, como presenta la ecuación 11.8:

$$\begin{aligned} S^A_{H_0} &= n_A\,\overline{R} = n_A\,\frac{n_T + 1}{2} \\ S^B_{H_0} &= n_B\,\overline{R} = n_B\,\frac{n_T + 1}{2} \end{aligned} \tag{11.8}$$

Para el ejemplo:

$$\begin{aligned} S^A_{H_0} &= 12 \cdot \frac{(23 + 1)}{2} = 144 \\ S^B_{H_0} &= 11 \cdot \frac{(23 + 1)}{2} = 132 \end{aligned}$$

A partir de este punto, la prueba de Wilcoxon-Mann-Whitney tiene dos variantes, una para muestras grandes y otra para muestras pequeñas.

11.2.1.1 Prueba de suma de rangos de Wilcoxon para muestras grandes

Hasta ahora hemos determinado valores para las sumas de los rangos esperada en cada muestra cuando estas provienen de poblaciones con igual distribución (H_0). Así, los valores de las sumas de los rangos observados en las muestras pueden ser consideradas como estimaciones de estas sumas esperadas. En nuestro ejemplo:

$$\begin{aligned} S^A = 139 &\triangleq S^A_{H_0} = 144 \\ S^B = 137 &\triangleq S^B_{H_0} = 132 \end{aligned}$$

Se ha demostrado que, para poblaciones con igual distribución, las sumas de los rangos tienen la misma desviación estándar, dada por la ecuación 11.9:

$$\sigma_S = \sqrt{\frac{n_A\, n_B\, (n_T + 1)}{12}} \tag{11.9}$$

Con lo que para el ejemplo tendríamos:

$$\sigma_S = \sqrt{\frac{12 \cdot 11 \cdot (23 + 1)}{12}} \approx 16{,}248$$

Cuando **ambas muestras tienen tamaño mayor o igual a 5**, siguiendo un procedimiento similar al descrito en la primera sección del capítulo 4, las distribuciones muestrales de S^A y S^B tienden a aproximarse a una distribución normal. En consecuencia, una vez conocidas la media y la desviación estándar de una distribución normal para la muestra, podemos calcular el estadístico z para S^A o S^B, dado por la ecuación 11.10:

$$z^M = \frac{(S^M - S^M_{H_0}) \pm 0{,}5}{\sigma_S} \tag{11.10}$$

donde:

- S^M es cualquiera de los valores observados S^A o S^B.
- $S^M_{H_0}$ es el valor nulo (la media de la distribución muestral de S^M si la hipótesis nula es cierta).
- σ_S es el error estándar de S^M (es decir, la desviación estándar de su distribución muestral, el error estándar).

Notemos que la ecuación 11.10 incluye un factor de corrección de continuidad, puesto que las distribuciones muestrales de las sumas de los rangos son intrínsecamente discretas (solo pueden asumir valores con decimales cuando existen rangos empatados). Este factor es negativo ($-0{,}5$) si $S^M > S^M_{H_0}$ y positivo ($0{,}5$) en caso contrario.

Volviendo al ejemplo, tenemos:

$$\begin{aligned} z^A &= \tfrac{(139-144)+0{,}5}{16{,}248} &&\approx -0{,}277 \\ z^B &= \tfrac{(137-132)-0{,}5}{16{,}248} &&\approx 0{,}277 \end{aligned}$$

Los valores z obtenidos a partir de S^A y S^B siempre tienen igual valor absoluto y signos opuestos, por lo que no importa cuál de ellos usemos para la prueba de significación estadística. No obstante, debemos tener muy claro el significado del signo de este estadístico z: si tuviésemos como hipótesis alternativa que la muestra A contiene **valores mayores** que la muestra B, entonces esperaríamos que las observaciones de mayor rango estuvieran en la primera, por lo que z^A debería ser positivo. Al contrario, si esperamos que la muestra A agrupe **valores menores** a los observados en la muestra B, entonces esperaríamos que las observaciones de mayor rango estuvieran en la segunda, por lo que z^A debería ser negativo.

El valor z obtenido permite calcular el valor p para una hipótesis alternativa unilateral (pues solo delimita la región de rechazo en una de las colas de la distribución normal estándar subyacente).

Así, para el ejemplo, que tiene una hipótesis alternativa bilateral, en R podemos calcular el valor p correspondiente mediante la llamada `2 * pnorm(-0.277, mean = 0, sd = 1, lower.tail = TRUE)`, o alternativamente `2 * pnorm(0.277, mean = 0, sd = 1, lower.tail = FALSE)`, obteniéndose como resultado $p \approx 0{,}782$. Evidentemente, el valor p obtenido es muy alto, por lo que fallamos al rechazar la hipótesis nula. En consecuencia, podemos concluir que no es posible descartar que las dos interfaces, A y B, **tienen niveles de usabilidad similares**.

11.2.1.2 Prueba de suma de rangos de Wilcoxon para muestras pequeñas

Cuando las muestras son pequeñas (menos de 5 observaciones[1]), no podemos usar el supuesto de normalidad del apartado anterior, por lo que necesitamos una vía alternativa. Este método sirve también para muestras más grandes, con resultados similares a los obtenidos con la aproximación normal.

[1] Aunque algunos autores fijan en 10 e incluso ¡30 observaciones! como umbral para usar la aproximación normal.

Aprovechando una vez más las ventajas de considerar los rangos en lugar de las observaciones originales, podemos calcular el máximo valor posible para la suma de rangos de cada muestra, suponiendo alternadamente que cada una recibe los rangos más altos de la muestra combinada, como muestra la ecuación 11.11:

$$\begin{aligned} S^A_{max} &= n_A\, n_B + \frac{n_A\,(n_A+1)}{2} \\ S^B_{max} &= n_B\, n_A + \frac{n_B\,(n_B+1)}{2} \end{aligned} \tag{11.11}$$

Así, para el ejemplo:

$$\begin{aligned} S^A_{max} &= 12 \cdot 11 + \frac{12 \cdot (12+1)}{2} = 210 \\ S^B_{max} &= 11 \cdot 12 + \frac{11 \cdot (11+1)}{2} = 198 \end{aligned}$$

Con esto podemos definir un nuevo estadístico de prueba U, como muestra la ecuación 11.12:

$$\begin{aligned} U^A &= S^A_{max} - S^A \\ U^B &= S^B_{max} - S^B \\ U &= \text{mín}(U^A, U^B) \end{aligned} \tag{11.12}$$

Por lo que en el ejemplo:

$$\begin{aligned} U^A &= 210 - 139 = 71 \\ U^B &= 198 - 137 = 61 \\ U &= \text{mín}(71, 61) = 61 \end{aligned}$$

Si la hipótesis nula fuera verdadera, esperaríamos que:

$$U^A = U^B = \frac{n_A\, n_B}{2}$$

Para el ejemplo, bajo la hipótesis nula esperaríamos que $U^A = U^B = \frac{12 \cdot 11}{2} = 66$.

En consecuencia, la pregunta asociada a la prueba es: si la hipótesis nula fuera verdadera, ¿qué tan probable es obtener un valor de U al menos tan pequeño como el observado?[2] Para el ejemplo esto sería: si no hay diferencias significativas en la usabilidad de ambas interfaces, ¿qué tan probable es obtener un valor $U \leq 61$?

Para responder a esta pregunta, seguimos un procedimiento similar al que ya conocimos para la prueba exacta de Fisher (capítulo 8): se calculan todas las formas en que n_T rangos podrían combinarse en dos grupos de tamaños n_A y n_B, y luego se determina la proporción de las combinaciones que produzcan un valor de U al menos tan pequeño como el encontrado.

Pero, para el ejemplo, ¡existen 676.039 combinaciones posibles! Afortunadamente existen tablas que permiten conocer el máximo valor U^{max} para el cual se rechaza la hipótesis nula para un nivel de significación dado, sin tener que revisar todas estas combinaciones.

Pero R no ofrece herramientas para calcular estos valores críticos, ni el valor p a partir del estadístico U observado, puesto que tiene implementadas funciones para la distribución del estadístico W, propuesto por Frank Wilcoxon en 1945, que lleva **a los mismos resultados**. Así, podemos aproximar el valor U^{max} (en rigor W^{max}), para una prueba bilateral, con la llamada `qwilcox(`$\alpha/2$`, `n_A`, `n_B`, lower.tail = TRUE)`. Sin embargo, debemos recordar que la distribución de U (como la de W) es discreta, por lo que el valor U^{max} devuelto por la llamada anterior está sobreestimado cuando su probabilidad no es exactamente $\alpha/2$. En este caso, debemos corregirlo restándole uno.

[2] Debemos notar que siempre se cumple que $U^A + U^B = n_A\, n_B$, por lo que en realidad podríamos haber escogido cualquiera de los valores U para realizar el procedimiento. Si se usara, en vez del menor, el mayor valor, es decir $U = \text{máx}(U^A, U^B)$, se debería responder qué tan probable es obtener un valor de U al menos tan grande como el observado.

Considerando el ejemplo, y un nivel de significación $\alpha = 0{,}05$, y como estamos realizando una prueba bilateral, obtenemos un primer valor para U^{max} con la llamada `qwilcox(0.05 / 2, 12, 11, lower.tail = TRUE)`, que devuelve 34. Para confirmar su exactitud, ejecutamos la llamada `pwilcox(34, 12, 11, lower.tail = TRUE)` para conocer más exactamente su probabilidad, lo que nos retorna el valor 0,0256094.... Como esta probabilidad no es exactamente 0,25, debemos corregir y quedarnos con el valor $U^{max} = 33$ (que es el encontrado en las tablas de valores críticos para el estadístico U de Mann-Whitney).

Puesto que el valor observado es mayor que el valor crítico, $61 > 33$, **fallamos al rechazar la hipótesis nula**, por lo que concluimos con 95 % de confianza que no se puede descartar que la usabilidad de ambas interfaces sea la misma.

11.2.1.3 Prueba de suma de rangos de Wilcoxon en R

Como podría suponerse, la implementación de esta prueba en R usa el estadístico W (introducido por Wilcoxon) en lugar del estadístico U empleado por Mann y Whitney. Es por ello que esta prueba se realiza mediante la función `wilcox.test(x, y, paired = FALSE, alternative, mu, conf.level)`, donde:

- `x`, `y`: vectores numéricos con las observaciones en cada muestra. Para aplicar la prueba con una única muestra, `y` debe ser nulo (por omisión, lo es).
- `paired`: booleano con valor falso para indicar que las muestras son independientes (que es el valor por omisión).
- `alternative`: señala el tipo de hipótesis alternativa: bilateral (`"two.sided"`) o unilateral (`"less"` o `"greater"`).
- `mu`: valor nulo, igual a cero si se omite.
- `conf.level`: nivel de confianza.

Al usar la función `wilcox.test()` con **una muestra** (no ejemplificado en este apunte), la hipótesis nula corresponde a que la población de origen se **distribuye simétricamente en torno al valor nulo** especificado (`mu` en la función). Cuando se usa con dos muestras independientes, como haremos para el ejemplo seguido, la hipótesis nula es que la **localización** de las tendencias centrales de las poblaciones de origen muestran una **desplazamiento igual a** `mu`.

Cuando la prueba es unilateral, la hipótesis alternativa es que la localización de la distribución poblacional, al trabajar con una muestra, o el verdadero desplazamiento de localización entre las poblaciones, al trabajar con dos muestras independientes, se ubica a la izquierda o a la derecha del valor nulo. Obviamente, en el caso bilateral, la hipótesis alternativa es que la localización o el desplazamiento de localización, respectivamente, es un valor distinto a `mu`.

El script 11.7 muestra la aplicación de esta prueba para el ejemplo, obteniéndose los resultados que se presentan en la figura 11.10.

```
 Wilcoxon rank sum test with continuity correction

data:  Interfaz_A and Interfaz_B
W = 61, p-value = 0.7816
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(Interfaz_A, Interfaz_B, alternative =
"two.sided",  :
  cannot compute exact p-value with ties
```

Figura 11.10: resultado de la prueba de Wilcoxon-Mann-Whitney para el ejemplo de las interfaces.

Podemos notar que la función `wilcox.test()` devuelve el mismo valor p (y el estadístico W con el mismo valor que el estadístico U) que calculamos anteriormente.

Vale la pena mencionar que cuando trabajemos con muestras pequeñas, existen las funciones `wilcox_test()` del paquete `coin` y `wilcox.exact()` del paquete `exactRankTests`, entre otras opciones, para aplicar una prueba de Wilcoxon-Mann-Whitney exacta. Esta última usa, junto a algunos parámetros adicionales, los mismos argumentos que la función `wilcox.test()` mostrada aquí.

Notemos que la salida de la función `wilcox.test()` (figura 11.10) nos advierte que en el procedimiento se han producido empates, por lo que se hicieron ajustes y el valor p que reporta podría ser inexacto. La llamada `wilcox.exact(Interfaz_A, Interfaz_B)` permite conocer el valor p exacto: `0.7741`. Vemos que la diferencia es pequeña y no influye en la decisión tomada.

Script 11.7: prueba de Mann-Whitney para el ejemplo.

```r
# Definir las muestras
Interfaz_A <- c(2.7, 6.6, 1.6, 5.1, 3.7, 6.1, 5.0, 1.4, 1.8, 1.5, 3.0, 5.3)
Interfaz_B <- c(5.0, 1.4, 5.6, 4.6, 6.7, 2.7, 1.3, 6.3, 3.7, 1.3, 6.8)

# Establecer nivel de significación
alfa <- 0.05

# Hacer y mostrar la prueba de suma de rangos de Wilcoxon (Mann-Whitney)
prueba <- wilcox.test(Interfaz_A, Interfaz_B,
                      alternative = "two.sided", conf.level = 1 - alfa)
print(prueba)
```

11.2.2 Prueba de rangos con signo de Wilcoxon

Recordemos que en el capítulo 5 también aprendimos una versión de la prueba t de Student para inferir sobre la media de las diferencias de dos muestras apareadas, siempre y cuando se verifiquen las siguientes dos condiciones:

1. Los pares de observaciones son independientes entre sí.
2. Las diferencias de observaciones apareadas siguen una distribución cercana a la normal.

Si no tenemos certeza de que se esté cumpliendo la segunda condición, recordando que esta no se cumple si la escala de medición no asegura intervalos iguales, podemos recurrir a la **prueba de rangos con signo de Wilcoxon**, que es conceptualmente similar a la prueba de suma de rangos de Wilcoxon presentada en la sección anterior, pero que en este caso corresponde a la alternativa no paramétrica a la prueba t de Student para **muestras apareadas**.

Las condiciones que se deben cumplir para usar esta prueba son:

1. Los pares de observaciones son independientes.
2. La escala de medición empleada para ambas muestras debe ser a lo menos ordinal.

El comentario sobre la condición de que la escala de medición de las observaciones sea "intrínsecamente continua" también aplica a esta prueba, es decir, en la actualidad se puede manejar un número acotado de empates a costa de pérdida de precisión en el valor p que se calcule.

Consideremos ahora un nuevo contexto para la aplicación de esta prueba: una empresa de desarrollo de software desea evaluar la usabilidad de dos interfaces alternativas, A y B, para un nuevo producto, a fin de determinar si, como asegura el departamento de diseño, es mejor la interfaz A. Para ello, la empresa ha seleccionado a 10 participantes al azar, quienes deben evaluar 6 aspectos de usabilidad de cada interfaz, cada uno de los cuales se mide con una escala Likert de 7 puntos, donde 1 significa "muy malo" y 7, "muy bueno". La valoración que cada participante da a la interfaz evaluada, llamada "índice de usabilidad", corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. Designados aleatoriamente, 5 participantes evaluaron primero la interfaz A, mientras que los otros 5 evaluaron primero la interfaz B. La tabla 11.4 muestra los índices de usabilidad otorgados por cada participante a cada una de las interfaces.

Participante	1	2	3	4	5	6	7	8	9	10	
Interfaz *A*	2,9	6,1	6,7	4,7	6,4	5,7	2,7	6,9	1,7	6,4	$\overline{x}_A = 5{,}02$
Interfaz *B*	6,0	2,8	1,3	4,7	3,1	1,8	2,9	4.0	2,3	1,6	$\overline{x}_B = 3{,}05$

Tabla 11.4: muestra de los índices de usabilidad asignados por cada participante.

Formalmente, y notando que en este caso la hipótesis alternativa planteada en el enunciado es unilateral, se necesitaría contrastar las siguientes hipótesis:

H_0: las mismas personas no perciben diferencia en la usabilidad de ambas interfaces (la distribución de los índices de usabilidad tienen la misma localización).

H_A: las mismas personas consideran que la interfaz A tiene mejor usabilidad que la interfaz B (la localización de los índices de usabilidad para la interfaz A está a la derecha de la localización para la interfaz B).

La mecánica inicial para esta prueba consiste en calcular las diferencias entre cada par de observaciones y obtener luego su valor absoluto. Generalmente se descartan aquellas instancias con diferencia igual a cero (los empates), pues no aportan información relevante al procedimiento. A continuación se ordenan las diferencias absolutas en orden creciente y se les asignan rangos de manera correlativa, del mismo modo que en la prueba de Wilcoxon-Mann-Whitney. Una vez asignados los rangos, se les incorpora el signo asociado a la diferencia del par de observaciones. La tabla 11.5 ilustra el proceso descrito.

Participante	4	7	9	8	1	2	5	6	10	3
Interfaz *A*	4,7	2,7	1,7	6,9	2,9	6,1	6,4	5,7	6,4	6,7
Interfaz *B*	4,7	2,9	2,3	4,0	6,0	2,8	3,1	1,8	1,6	1,3
A*−*B	0,0	-0,2	-0,6	2,9	-3,1	3,3	3,3	3,9	4,8	5,4
$\lvert \boldsymbol{A}-\boldsymbol{B} \rvert$	0,0	0,2	0,6	2,9	3,1	3,3	3,3	3,9	4,8	5,4
Rango absoluto	—	1	2	3	4	5,5	5,5	7	8	9
Rango con signo	—	-1	-2	+3	-4	+5,5	+5,5	+7	+8	+9

Tabla 11.5: asignación de rangos con signo para el ejemplo.

En teoría, una muestra de n pares distintos genera n **rangos no empatados** sin signo (fila "Rango absoluto" de la tabla 11.5). A su vez, cada uno de dichos rangos podría tomar valores positivos o negativos, por lo que se tienen ¡2^n posibles combinaciones de rangos con signos!

Por ejemplo, la tabla 11.6 muestra todas las posibles combinaciones de signos para $n = 3$ rangos, junto a las sumas de los rangos positivos (S^+), negativos (S^-) y en general (S^G). Podemos observar que la suma de los rangos positivos varía de 0 a 6, que la suma de los rangos negativos varía de -6 a 0, y que la suma general de los rangos con signo toma algunos valores de -6 a 6. Esto no es un accidente, puesto que para n pares, el rango máximo queda dado por la ecuación 11.13, que para $n = 3$ resulta $(3 \cdot 4)/2 = 6$.

$$R^{max} = \frac{n\,(n+1)}{2} \tag{11.13}$$

Rango			**Sumas**		
1	**2**	**3**	$\mathbf{S}^+$	$\mathbf{S}^-$	$\mathbf{S}^G$
+	+	+	6	0	6
+	+	-	3	-3	0
+	-	+	4	-2	2
+	-	-	1	-5	-4
-	+	+	5	-1	4
-	+	-	2	-4	-2
-	-	+	3	-3	0
-	-	-	0	-6	-6

Tabla 11.6: combinaciones de rangos positivos y negativos para una muestra de $n = 3$ pares.

Si la hipótesis nula fuese cierta, los grupos presentarían valores similares para los rangos positivos y negativos y se distribuirían de manera homogénea, por lo que se esperaría que estas sumas tomaran los valores expresados en la ecuación 11.14, que corresponden a los valores nulos en el dominio de los rangos.

$$\begin{aligned} S^+_{H_0} = \quad S^-_{H_0} = \quad \frac{R^{max}}{2} \quad &= \frac{n\,(n+1)}{4} \\ S^G_{H_0} = \quad S^+_{H_0} + S^-_{H_0} \quad &= 0 \end{aligned} \tag{11.14}$$

La figura 11.11 muestra las distribuciones muestrales de las sumas de los rangos positivos, negativos y en general, para distintos valores de n. En ella podemos apreciar que, a medida que el número de pares observados aumenta, estas distribuciones **rápidamente** se aproximan cada vez más a distribuciones normales con medias en los valores nulos de la ecuación 11.14.

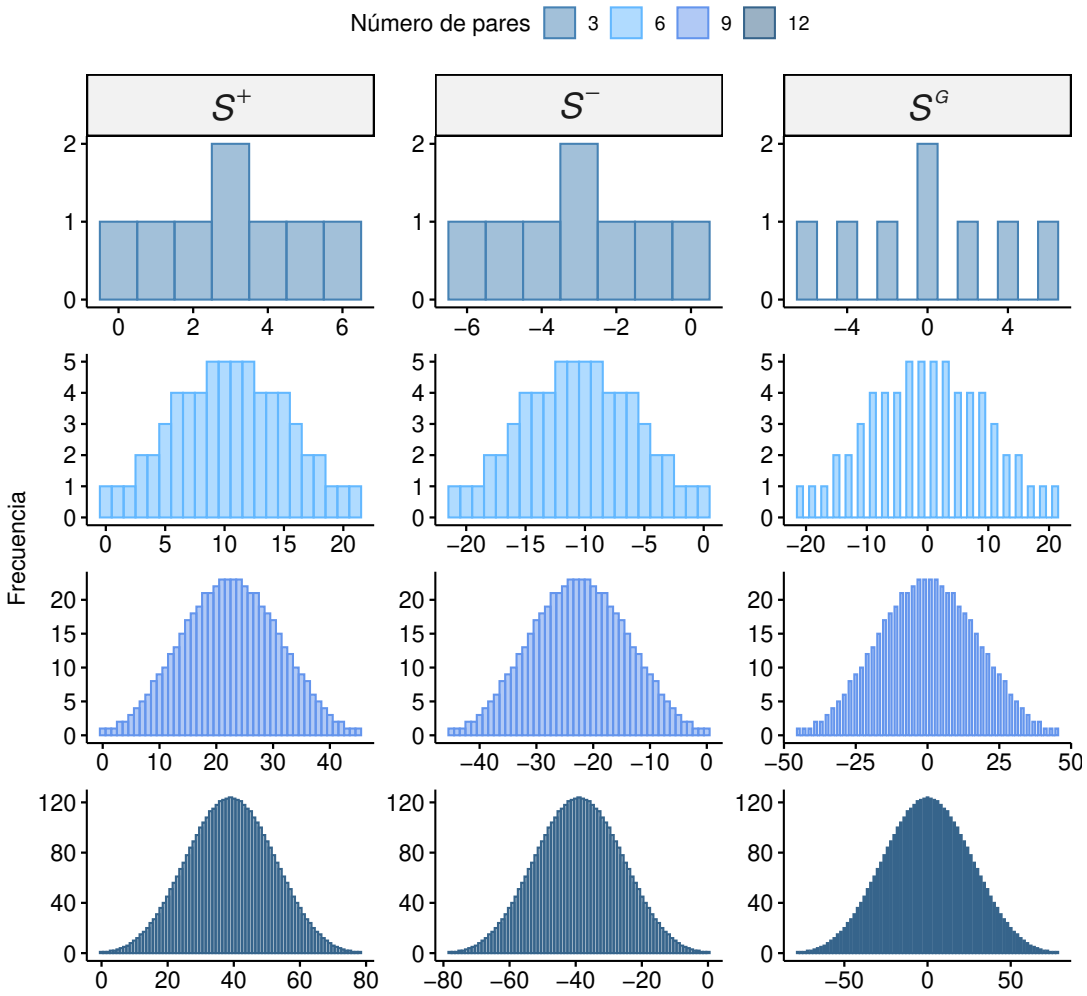


Figura 11.11: ejemplo de distribuciones muestrales de la sumas de los rangos con signo para muestras con 3, 6, 9 y 12 pares de observaciones distintas.

Aquí nos enfrentamos a un dilema, puesto que hay múltiples versiones de prueba de rangos con signo de Wilcoxon que usan estadísticos diferentes pero que llevan a resultados muy similares. Algunas usan como estadístico la suma de los rangos con signo (S^G), mientras que otras usan el menor valor absoluto de las sumas de los rangos positivos (S^+) y negativos (S^-). En este apunte usaremos el **estadístico V** que corresponde a la suma de los rangos con signo positivo, es decir $V = S^+$, sencillamente porque es el estadístico que reporta la función `wilcox.test()` que usaremos luego para aplicar esta prueba usando R. Se sabe que el error estándar para este estadístico está dado por la ecuación 11.15:

$$\sigma_V = \sqrt{\frac{n\,(n+1)\,(2n+1)}{24}} \tag{11.15}$$

Volviendo al ejemplo (más grande) de la comparación de la usabilidad de dos interfaces (tabla 11.5), primero debemos notar que tenemos $n = 9$ pares de observaciones distintas (y un empate). Usando las definiciones anteriores, tendríamos:

$$\begin{aligned} V = &\quad 3 + 5{,}5 + 5{,}5 + 7 + 8 + 9 &&= 38 \\ S^+_{H_0} = V_0 = &\quad \frac{9 \cdot (9+1)}{4} &&= 22{,}5 \\ \sigma_V = &\quad \sqrt{\frac{10 \cdot (10+1) \cdot (2 \cdot 10 + 1)}{24}} &&\approx 8{,}441 \end{aligned}$$

La figura 11.11 presenta la distribución muestral para este caso (tercera fila, primera columna). Podemos notar que esta distribución comienza a parecerse bastante a una distribución normal con media 22,5 y desviación estándar alrededor de 8. Como ha sido tradicional, la prueba de hipótesis necesita responder si el valor observado $V = 38$ está lo suficientemente lejos del valor hipotético $V_0 = 22{,}5$ como para descartar una igualdad en la usabilidad de ambas interfaces.

Cuando la muestra de pares es grande, podemos trabaja bajo el supuesto de normalidad y calcular el estadístico de prueba z, de la forma que ha sido usual, dado por la ecuación 11.16:

$$z = \frac{V - V_0 \pm 0{,}5}{\sigma_V} \tag{11.16}$$

Al igual que para la prueba de suma de rangos de Wilcoxon, el estadístico de prueba incluye un factor de corrección de continuidad que es negativo si $V > V_0$ y positivo en caso contrario.

Para el ejemplo tenemos que:

$$z = \frac{38 - 22{,}5 - 0{,}5}{8{,}441} \approx 1{,}777$$

Como hecho anteriormente, podemos obtener el valor p asociado a este estadístico de prueba mediante la llamada `pnorm(1.777, mean = 0, sd = 1, lower.tail = FALSE)` (no multiplicamos por 2, pues consideramos una prueba unilateral), obteniendo como resultado $p = 0{,}038$[3]. Considerando un nivel de significación $\alpha = 0{,}05$, **rechazamos la hipótesis nula** en favor de la hipótesis alternativa. En consecuencia, concluimos con 95 % de confianza que **la usabilidad de la interfaz *A* es mejor que la de la interfaz *B***.

11.2.2.1 Prueba de rangos con signo de Wilcoxon en R

En R, la prueba de rangos con signo de Wilcoxon está implementada en la misma función que en el caso de muestras independientes, pero ahora debemos asegurarnos de indicar que las muestras están apareadas a través de la llamada `wilcox.test(x, y, paired = TRUE, alternative, conf.level)`. Es decir, el valor por omisión para el parámetro `paired` es `FALSE` y este indica aplicar la prueba de suma de rangos de Wilcoxon a los datos, mientras que si explícitamente indicamos `paired = TRUE`, se aplica la prueba de rangos con signo de Wilcoxon.

El script 11.8 muestra la aplicación de esta última prueba para el ejemplo, obteniéndose los resultados que se presentan en la figura 11.12. Vemos que el valor p entregado por la función `wilcox.test()` es el mismo que obtuvimos de forma manual. También observamos que la función nos advierte que tuvo que hacer correcciones por la presencia de empates (que aquí se les llama *zeroes*).

Por otro lado, debemos tener en cuenta que esta función sigue el supuesto de normalidad, el que es válido para **n > 10 pares distintos**. Para nuestro ejemplo, o con muestras más pequeñas, tenemos que consultar una tabla de valores críticos para V o usar las funciones implementadas para su distribución o alguna implementación de una prueba exacta.

R cuenta con las funciones `psignrank(q, n, lower.tail)` y `qsignrank(p, n, lower.tail)` que nos permiten calcular valores p o umbrales críticos de V, respectivamente, puesto que implementan las funciones de probabilidad acumulada y cuantil para la distribución del estadístico de rango con signo de Wilcoxon obtenido de una muestra de tamaño `n`.

[3]Por supuesto, podríamos usar la llamada `pnorm(38 - 0.5, mean = 22.5, sd = 8.441, lower.tail = FALSE)` para obtener esta probabilidad, evitando la normalización.

```
 Wilcoxon signed rank test with continuity correction

data:  Interfaz_A and Interfaz_B
V = 38, p-value = 0.03778
alternative hypothesis: true location shift is greater than 0

Warning message:
In wilcox.test.default(Interfaz_A, Interfaz_B, paired = TRUE,
alternative = "greater",  :
  cannot compute exact p-value with zeroes
```

Figura 11.12: resultado de la prueba de rangos con signo de Wilcoxon para el ejemplo.

Script 11.8: prueba de rangos con signo de Wilcoxon para el ejemplo.

```r
# Definir las muestras
Interfaz_A <- c(2.9, 6.1, 6.7, 4.7, 6.4, 5.7, 2.7, 6.9, 1.7, 6.4)
Interfaz_B <- c(6.0, 2.8, 1.3, 4.7, 3.1, 1.8, 2.9, 4.0, 2.3, 1.6)

# Establecer nivel de significación
alfa <- 0.05

# Hacer y mostrar la prueba de rangos con signo de Wilcoxon
prueba <- wilcox.test(Interfaz_A, Interfaz_B, paired = TRUE,
                      alternative = "greater", conf.level = 1 - alfa)
print(prueba)
```

En R también existen, entre otras alternativas, las funciones `wilcoxsign_test()` del paquete `coin`, con argumentos `distribution = "exact"` y `zero.method = "Wilcoxon"` para reproducir el procedimiento visto aquí, o la función `wilcox.exact()` del paquete `exactRankTests`, indicando `paired = TRUE`. De hecho, usando la llamada `wilcox.exact(Interfaz_A, Interfaz_B, paired = TRUE, alternative = "greater")` podemos conocer un valor p más confiable para nuestro ejemplo: $p = 0.037$ (que no es muy distinto al valor aproximado usando la suposición de normalidad).

11.2.3 Nota sobre las hipótesis e interpretación de las pruebas basadas en rangos

Es importante hacer una observación respecto a lo que se encuentra en Internet sobre las hipótesis e interpretación de las pruebas con rangos vistas en esta sección, ya que uno se topa con diferentes versiones que pueden llevar a confusión. Esto ocurre porque los estadísticos basados en rangos pueden tener significados ligeramente distintos dependiendo de los supuestos que se hagan sobre las distribuciones de las muestras (Fay & Proschan, 2010).

Una versión muy frecuente es que las pruebas con rangos comparan **medianas**. Cuando se trabaja con una muestra o la diferencia de dos muestras apareadas, la hipótesis nula es que la población de origen está centrada en torno al valor igual, menor o mayor que el valor nulo considerado (especificado con el argumento `mu` de la función `wilcox.test()`). Cuando **además se supone** (o **verifica**) que la población tiene una **distribución simétrica**, este valor nulo efectivamente corresponde al valor hipotético para **la mediana de la población de origen**.

Similarmente, al comparar muestras independientes, la hipótesis nula es que la **diferencia entre las localizaciones** de tendencia central de las poblaciones de origen es igual, menor o mayor que el valor nulo considerado. Cuando **además se supone** que las poblaciones tienen distribuciones con la **misma forma simétrica**, entonces la hipótesis nula se refiere a **la diferencia entre sus medianas**.

Así, estas pruebas pueden resultar significativas por otras razones, como la presencia de asimetrías o diferencias en la dispersión, y solo **descartando estas posibilidades**, se les puede atribuir a diferencias en las medianas. Sin embargo, esto no es fácil de hacer con muestras pequeñas.

Por esta razón, en este apartado se ha usado la forma más robustas y libre de supuestos, considerando hipótesis que apuntan a detectar "diferencias" que no se refieren únicamente a las medianas, sino a si la

población es dominantemente mayor o menor que el valor nulo hipotetizado o, en el caso de dos muestras independientes, que una de las distribuciones es dominantemente mayor o menor que la otra en general.

De aquí que aparece otra interpretación, relacionada con la **probabilidad estocástica** de que una observación aleatoria de la población de origen es mayor que el valor nulo, al trabajar con una muestra, o a otra observación aleatoria de la otra población, cuando se consideran dos muestras independientes. Es decir, se evalúa si la probabilidad $P(A > \mu_0^A)$ o $P(A > B)$, respectivamente, es mayor que 0,5.

Existen otras interpretaciones, pero que son encontradas con menor frecuencia.

11.3 PRUEBAS NO PARAMÉTRICAS CON MÁS DE DOS MUESTRAS NUMÉRICAS

Al igual que existen alternativas no paramétricas para inferir con una o dos muestras de una variable numérica, también las hay para cuando se tienen más de dos muestras. Conoceremos ahora alternativas no paramétricas clásicas para el procedimiento ANOVA de una vía, tanto para muestras independientes como para muestras correlacionadas.

11.3.1 Prueba de Kruskal-Wallis

En el capítulo 9 estudiamos el procedimiento ANOVA de una vía para $k > 2$ muestras independientes, el cual requiere el cumplimiento de los siguientes supuestos:

1. La escala con que se mide la variable dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las muestras son obtenidas de manera aleatoria e independiente desde las poblaciones de origen.
3. Se puede suponer razonablemente que las poblaciones de origen siguen una distribución normal.
4. Se puede suponer razonablemente que las poblaciones de origen tienen varianzas similares.

Si bien ANOVA es usualmente robusto ante desviaciones leves de las condiciones (excepto la segunda) cuando las muestras son de igual tamaño, no ocurre lo mismo cuando los tamaños de las muestras difieren. En este caso, una alternativa es emplear la **prueba de Kruskal-Wallis**, cuyas condiciones son:

1. La escala de la variable dependiente debe ser, a lo menos, ordinal.
2. Las observaciones son independientes entre sí.
3. La variable independiente debe tener a lo menos dos niveles (aunque, para dos niveles, se suele usar la prueba de Wilcoxon-Mann-Whitney).

Para ilustrar esta prueba, tomemos el ejemplo de un estudiante memorista en informática que para su proyecto de título estudia cuatro criterios diferentes (A, B, C y D) para optimizar una secuencia de operaciones *join* en tablas de bases de datos relacionales con millones de registros. Para sus experimentos, el memorista consiguió armar un conjunto de 48 bases de datos públicas de tamaños similares donde se requiere realizar una secuencia de 8 operaciones *join* para obtener un reporte. Luego, asignó cada base de datos a uno de los cuatro criterios de forma aleatoria y ejecutó las consultas en iguales condiciones de hardware y recursos computacionales, registrando el tiempo de ejecución (en segundos) requerido por cada *query*, obteniendo las siguientes mediciones:

Criterio *A*: 95, 36, 58, 11, 56, 77, 49, 9, 11, 29, 28, 13
Criterio *B*: 22, 63, 26, 20, 24, 23, 23, 24, 53
Criterio *C*: 39, 77, 26, 34, 26, 26, 8, 49, 28, 40, 64, 7, 11, 7
Criterio *D*: 14, 8, 15, 10, 20, 6, 10, 13

Podemos suponer que el estudiante memorista necesita conocer si los criterios que estudia tienen eficiencias similares o si existe(n) alguno(s) mejor(es) que otros. La eficiencia entonces, podría medirse como la media del tiempo de ejecución que requieren las secuencia de operaciones *join* optimizadas que genera cada criterio, lo que sugiere que podríamos realizar un análisis estadístico usando un procedimiento ANOVA para muestras independientes. La primera condición para aplicar este procedimiento se cumple, pues el tiempo de ejecución (la variable dependiente) corresponde a una medición física, por lo que tiene escala de razón. Para la segunda condición, debemos considerar que las 48 bases de datos seleccionadas representan mucho menos del 10 % de la población de bases de datos grandes que existen en el mundo, que el experimento asignó cada una de estas a un criterio de forma aleatoria y que el tiempo que tarda un *query* en una base de datos no influye

en el tiempo que requiere otro *query* en la misma o en base de datos. Luego, las cuatro muestras contienen observaciones independientes.

Veamos las otras dos condiciones. El script 11.9 presenta el código que permite revisar los supuestos de normalidad y homogeneidad de las varianzas por medio de pruebas de Shapiro-Wilk y Levene, respectivamente. El resultado se puede apreciar en la figura 11.13, donde vemos que, con 95 % confianza, podemos descartar que una de las poblaciones de origen siga una distribución normal, y que debemos rechazar la idea de que las muestras provienen de distribuciones con varianzas aproximadamente iguales.

Script 11.9: evaluación de las condiciones para aplicar un procedimiento ANOVA al ejemplo.

```r
library(car)

# Construir la mariz de datos
A <- c(95, 36, 58, 11, 56, 77, 49, 9, 11, 29, 28, 13)
B <- c(22, 63, 26, 20, 24, 23, 23, 24, 53)
C <- c(39, 77, 26, 34, 26, 26, 8, 49, 28, 40, 64, 7, 11, 7)
D <- c(14, 8, 15, 10, 20, 6, 10, 13)

Tiempo <- c(A, B, C, D)
Criterio <- c(rep("A", length(A)), rep("B", length(B)),
              rep("C", length(C)), rep("D", length(D)))
Criterio <- factor(Criterio)
datos <- data.frame(Tiempo, Criterio)

# Establecer nivel de significación
alfa <- 0.05

# Revisar normalidad y homocedasticidad
sh_tests <- by(Tiempo, Criterio, shapiro.test)
normalidad <- data.frame(
  W = sapply(sh_tests, function(t) round(t[["statistic"]], 4)),
  p.value = sapply(sh_tests, function(t) round(t[["p.value"]], 4)))
rownames(normalidad) <- levels(Criterio)
homogeneidad_var <- leveneTest(Tiempo ~ Criterio, datos)

cat("Pruebas de normalidad Shapiro-Wilk\n")
cat("----------------------------------\n")
print(normalidad)
cat("\nPrueba de homocedasticidad de Levene\n")
cat("------------------------------------\n")
print(homogeneidad_var)
```

```
Pruebas de normalidad Shapiro-Wilk
----------------------------------
       W p.value
A 0.9109  0.2193
B 0.6629  0.0005
C 0.9174  0.2020
D 0.9675  0.8777

Prueba de homocedasticidad de Levene
------------------------------------
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  3   3.489 0.02456 *
      39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 11.13: normalidad y homocedasticidad de los datos del ejemplo

De esta forma, lo más prudente sería utilizar la prueba no paramétrica de Kruskal-Wallis, para la cual los datos sí cumplen las condiciones, aún cuando perdamos información sobre las diferencias en tiempos de

ejecución.

Para el ejemplo, las hipótesis no paramétricas a contrastar serían:

H_0: todos los criterios llevan a consultas igualmente eficientes (las distribuciones del tiempo de ejecución de cada criterio son las mismas).

H_A: al menos uno de los criterios genera consultas con eficiencia diferente a las generadas por al menos algún otro criterio.

El procedimiento de la prueba de Kruskal-Wallis tiene elementos similares a los descritos en las pruebas no paramétricas para una y dos medias. El primer paso consiste en combinar las muestras y luego asignar el rango a cada elemento, obteniéndose para el ejemplo el resultado de la tabla 11.7.

Observaciones (ordenadas)										**Ranking de observaciones**									
A			***B***		***C***			***D***		***A***			***B***		***C***			***D***	
9	29	77	20	24	7	26	40	6	14	6.0	29.0	41.5	16.5	21.5	2.5	24.5	33.0	1.0	14.0
11	36	95	22	26	7	26	49	8	15	10.0	31.0	43.0	18.0	24.5	2.5	24.5	34.5	4.5	15.0
11	49		23	53	8	28	64	10	20	10.0	34.5		19.5	36.0	4.5	27.5	40.0	7.5	16.5
13	56		23	63	11	34	77	10		12.5	37.0		19.5	39.0	10.0	30.0	41.5	7.5	
28	58		24		26	39		13		27.5	38.0		21.5		24.5	32.0		12.5	

Tabla 11.7: asignación de rangos a la muestra combinada del ejemplo.

A continuación se calcula la suma (S^g) y la media (M^g) de los rangos en cada grupo ($g \in \{A, B, C, D\}$) y en la muestra combinada ($g = T$). La tabla 11.8 presenta los valores obtenidos para el ejemplo, incluyendo además el tamaño muestral de cada caso (n_g).

Grupo:	***A***	***B***	***C***	***D***	***T***
n_g	12	9	14	8	43
S^g	320,00	216,00	331,50	78,50	946,00
M^g	26,67	24,00	23,68	9,81	22,00

Tabla 11.8: resumen de los rangos para el ejemplo.

Como en las pruebas anteriores, usar rangos nos entrega información adicional sobre los datos. Por ejemplo, el promedio de los rangos está siempre dado por la ecuación 11.17.

$$M^T = \frac{n_T + 1}{2} \tag{11.17}$$

La hipótesis nula, llevada al dominio de los rangos, es que los rangos medios de los distintos grupos son iguales. Bajo este supuesto, el promedio y la varianza de las sumas de los rangos de cada grupo están dados, respectivamente, por las ecuaciones 11.18 y 11.19:

$$S^g_{H_0} = \frac{n_g\,(n_T + 1)}{2} \tag{11.18}$$

$$S^g_{H_0} = \sqrt{\frac{n_g\,(n_T + 1)(n_T - n_g)}{12}} \tag{11.19}$$

De manera similar a ANOVA, se requiere determinar la variabilidad que presentan los rangos **entre los grupos**, que denotaremos RSS_{bg} por sus siglas en inglés, que se obtiene con la suma de las desviaciones cuadradas de las medias grupales de los rangos con respecto a la media total de los rangos, como se presenta en la ecuación 11.20:

$$RSS_{bg} = \sum_g n_g\,(M^g - M^T)^2 \tag{11.20}$$

Con un poco de álgebra:

$$\begin{aligned} RSS_{bg} &= \sum_{g=1}^{k} n_g \, (\frac{S^g}{n_g} - \frac{n_T + 1}{2})^2 \\ &= \sum_{g} (S^g - \frac{n_g \, (n_T + 1)}{2})^2 \\ &= \sum_{g} (S^g - S^g_{H_0})^2 \end{aligned}$$

Si consideramos que $\frac{S^g - S^g_{H_0}}{\sigma^g_{H_0}} \sim \mathcal{N}(0, 1)$, entonces $\sum_g \frac{(S^g - S^g_{H_0})^2}{\sigma^g_{H_0}} \sim \chi^2(\nu = k - 1)$, para k grupos.

En base a este análisis, Kruskal y Wallis (1952) definieron el estadístico de prueba H como muestra la ecuación 11.21, que sigue aproximadamente una distribución χ^2 con $k - 1$ grados de libertad cuando **cada grupo tiene a lo menos 5 observaciones**.

$$H = \frac{12}{n_T \, (n_T + 1)} RSS_{bg} \tag{11.21}$$

Volviendo al ejemplo de las secuencias de operaciones *join*, tendríamos:

$$\begin{aligned} RSS_{bg} &= n_A \, (M^A - M^T)^2 + n_B \, (M^B - M^T)^2 + n_C \, (M^C - M^T)^2 + n_D \, (M^D - M^T)^2 \\ &= 12 \cdot (26{,}67 - 22)^2 + 9 \cdot (24{,}00 - 22)^2 + 14 \cdot (23{,}68 - 22)^2 + 8 \cdot (9{,}81 - 22)^2 \\ &\approx 1.525{,}989 \end{aligned}$$

y

$$H = \frac{12}{43 \cdot (43 + 1)} \cdot 1.525{,}989 \approx 9{,}679$$

Como todas las muestras tienen más 5 observaciones, el estadístico se distribuye aproximadamente como χ^2 con $\nu = 4 - 1 = 3$ grados de libertad. Así, podemos calcular el valor p para el ejemplo (en R) mediante la llamada `pchisq(9.679, 3, lower.tail = FALSE)`, obteniéndose como resultado $p = 0{,}022$. Este valor indica que la evidencia es suficientemente fuerte como para **rechazar la hipótesis nula** en favor de la hipótesis alternativa con un nivel de significación $\alpha = 0{,}05$. En consecuencia, podemos concluir con 95 % confianza que existe al menos un criterio de optimización que lleva a consultas con eficiencias distintas a las generadas por al menos algún otro criterio.

Fijémonos en que, al igual que ANOVA, la prueba de Kruskal-Wallis es de **tipo ómnibus**, por lo que no entrega información en relación a cuáles grupos presentan diferencias. En consecuencia, una vez más es necesario efectuar un análisis post-hoc cuando se detectan diferencias significativas. De manera similar a la estudiada en el capítulo 9, podemos hacer comparaciones entre pares de grupos con la prueba de Wilcoxon-Mann-Whitney (equivalentes a las realizadas con la prueba t de Student para ANOVA de una vía para muestras independientes), usando alguno de los factores de corrección que ya conocimos en el capítulo 8, como los métodos de Holm o de Benjamini y Hochberg (Amat Rodrigo, 2016b).

11.3.1.1 Prueba de Kruskal-Wallis en R

En R, podemos ejecutar la prueba de Kruskal-Wallis mediante la función `kruskal.test(formula, data)`, donde:

- `formula`: tiene la forma `<variable_dependiente>~<variable_independiente>`.
- `data`: matriz de datos en formato largo.

La función `pairwise.wilcox.test(x, g, p.adjust.method, paired = FALSE)`, permite realizar pruebas post-hoc para todos los pares de tratamientos, cuando corresponda, donde:

- `x`: vector con la variable dependiente.
- `g`: factor de agrupamiento (vector con la variable independiente).
- `p.adjust.method`: string con el nombre del método para ajustar los valores p de las pruebas múltiples (`"holm"`, `"BH"`, etc.).
- `paired`: valor booleano que indica si la prueba es pareada (verdadero) o no. Debe tener valor `FALSE` para aplicar pruebas de Wilcoxon-Mann-Whitney.

El script 11.10 muestra la realización de la prueba de Kruskal-Wallis para el ejemplo e incorpora el procedimiento post-hoc de Benjamini y Hochberg. Observemos que hemos indicado el argumento `exact = FALSE` en este último, con la intención de evitar la advertencia, para cada par de tratamientos, que no es posible obtener un valor p exacto en presencia de rangos empatados. Los resultados se presentan en la figura 11.14.

Primero, debemos notar que el estadístico H que reporta la función difiere ligeramente al obtenido anteriormente ($9{,}6903 \neq 9{,}679$). Esto se debe al redondeo que aplicamos y a que el método debe introducir ajustes en la ecuación 11.21 por la presencia de empates en los datos analizados. En consecuencia, el valor p que reporta la función también presenta una pequeña diferencia con el obtenido más arriba.

```
Resultados de la prueba ómnibus
--------------------------------
 Kruskal-Wallis rank sum test

data:  Tiempo by Criterio
Kruskal-Wallis chi-squared = 9.6903, df = 3, p-value = 0.02139

Resultados del análisis post-hoc
--------------------------------
 Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  datos[["Tiempo"]] and datos[["Criterio"]]

  A      B      C
B 0.5937 -      -
C 0.5714 0.5937 -
D 0.0615 0.0044 0.0680

P value adjustment method: BH
```

Figura 11.14: resultado de la prueba de Kruskal-Wallis y el procedimiento post-hoc de Benjamini y Hochberg para el ejemplo.

Script 11.10: (continuación del script 11.9) prueba de Kruskal-Wallis y el procedimiento post-hoc de Benjamini y Hochberg para el ejemplo.

```r
# Aplicar la prueba de Kruskal-Wallis y mostrar los resultados
prueba <- kruskal.test(Tiempo ~ Criterio, data = datos)
cat("\nResultados de la prueba ómnibus\n")
cat("--------------------------------")
print(prueba)

# Efectuar un procedimiento post-hoc de Benjamini y Hochberg
# si se encuentran diferencias significativas.
if(prueba[["p.value"]] < alfa) {
  post_hoc <- pairwise.wilcox.test(datos[["Tiempo"]], datos[["Criterio"]],
                                   p.adjust.method = "BH",
                                   paired = FALSE, exact = FALSE)

  cat("Resultados del análisis post-hoc\n")
  cat("--------------------------------")
  print(post_hoc)
}
```

A partir de los resultados del procedimiento post-hoc, podemos concluir con 95 % confianza que existen diferencias significativas entre la eficiencia conseguida por los criterios de optimización B y D. Consultando

la tabla 11.8, vemos que el promedio de los rangos obtenidos por el criterio D es menor al del criterio B, por lo que el primero es significativamente más eficiente que el segundo.

Notemos que `pairwise.wilcox.test()` solo reporta los p valores ajustados. Si queremos conocer el tamaño del efecto de las diferencias detectadas, como se recomienda reportar, debemos realizar las correspondientes pruebas de Wilcoxon-Mann-Whitney para todos los pares de grupos que presenten diferencias significativas.

11.3.2 Prueba de Friedman

Como es natural suponer, podemos considerar la **prueba de Friedman** como una alternativa no paramétrica al procedimiento ANOVA de una vía con muestras correlacionadas descrito en el capítulo 10. Sin embargo, debemos saber que no es exactamente una extensión de esta prueba, puesto que no considera las diferencias relativas entre casos (como lo hace ANOVA y la prueba de rangos con signo de Wilcoxon), y en consecuencia, como señala Baguley (2012), el poder estadístico de esta prueba es bastante menor.

Recordemos las condiciones que se deben verificar para poder aplicar la prueba ANOVA para muestras correlacionadas (vistas en el capítulo 10):

1. La escala con que se mide la variable dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las mediciones son independientes al interior de cada grupo.
3. Se puede suponer razonablemente que las poblaciones de origen siguen una distribución normal.
4. La matriz de varianzas-covarianzas es esférica.

Existen situaciones en las que no podemos comprobar que la escala de medición de la variable dependiente sea de intervalos iguales:

- Cuando las observaciones se miden en una escala logarítmica (por ejemplo, la escala de pH para medir la acidez o la escala de Richter para medir la intensidad de los sismos).
- Cuando las mediciones provienen de una escala ordinal, por ejemplo, un orden de preferencia.
- Cuando las mediciones de base provienen de una escala ordinal. Por ejemplo, cuando se suman o promedian puntajes de diversos elementos evaluados con una escala Likert.

Las condiciones requeridas por la prueba de Friedman son menos exigentes:

1. La escala de la variable dependiente debe ser, a lo menos, ordinal.
2. Las observaciones son una muestra aleatoria e independiente de la población.
3. La variable independiente debe ser categórica y tener a lo menos tres niveles.

Como ejemplo para esta prueba, supongamos ahora que el equipo de desarrollo de una exitosa *app* desea establecer qué interfaz gráfica (A, B o C) resultaría más atractiva para la siguiente versión, por lo que han seleccionado una muestra aleatoria representativa de los distintos tipos de usuarias y usuarios, y les han solicitado evaluar 6 aspectos de cada interfaz con una escala Likert de 5 puntos, donde el valor 1 corresponde a una valoración muy negativa y 5, a una muy positiva. La valoración que cada participante da a la interfaz evaluada, llamado "índice de usabilidad", corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. El orden en que cada participante evaluó cada una de las interfaces fue asignado aleatoriamente, obteniendo las siguientes mediciones:

Interfaz *A*: 3,6; 4,2; 3,5; 3,2; 3,6; 3,5; 3,3; 3,5; 4,1; 3,7; 4,0; 3,5; 3,3; 3,4; 3,6
Interfaz *B*: 4,4; 5,0; 4,3; 3,6; 4,5; 4,2; 3,9; 4,3; 4,8; 4,5; 4,8; 4,2; 3,7; 4,0; 4,5
Interfaz *C*: 4,9; 5,0; 4,7; 3,2; 5,0; 4,6; 3,6; 4,9; 5,0; 5,0; 5,0; 4,4; 3,2; 3,7; 4,9

En esta situación, el equipo de desarrollo necesita determinar si existe alguna interfaz que sea más atractiva para las usuarias y los usuarios del nuevo sistema. Podemos suponer, inicialmente, que "más atractivo" significa "mayor índice de usabilidad promedio", por lo que nuestra primera intuición sería aplicar un procedimiento ANOVA para muestras correlacionadas. Sin embargo, como este índice se obtiene utilizando una escala Likert, **no sería apropiado asumir** que esta medida (la variable dependiente) tiene escala de intervalos iguales, por lo que este análisis paramétrico queda descartado.

De esta forma, en este ejemplo es más prudente utilizar la prueba no paramétrica de Friedman, cuyas condiciones sí parecen cumplirse: las escalas Likert son por esencia ordinales, el enunciado describe una selección aleatoria de un conjunto reducido de potenciales usuarias y usuarios, y se estudian tres posibles interfaces (la variable independiente).

En consecuencia, las hipótesis a contrastar serían:

H_0: las interfaces obtienen índices de usabilidad similares.
H_A: al menos una interfaz obtiene índices de usabilidad distintos que al menos otra interfaz.

El primer paso del proceso consiste en asignar rangos a las k observaciones de **cada participante**. La medición con menor valor recibe el rango 1, la siguiente el rango 2, y así sucesivamente, hasta que la observación más alta obtiene un rango k.

Para el ejemplo, la interfaz con puntuación más baja recibe un rango 1, la que sigue el rango 2 y la que tiene la puntuación más alta, el rango 3. Como en los procedimientos estudiados anteriormente, en caso de empate se asigna el promedio de los rangos correspondientes. La tabla 11.9 muestra el resultado de este proceso.

Participante	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Interfaz *A*	1	1	1	1,5	1	1	1	1	1	1	1	1	2	1	1
Interfaz *B*	2	2,5	2	3	2	2	3	2	2	2	2	2	3	3	2
Interfaz *C*	3	2,5	3	1,5	3	3	2	3	3	3	3	3	1	2	3

Tabla 11.9: rangos de las interfaces por persona usuaria.

Asociada a la hipótesis nula para la prueba de Friedman está la idea que los rangos promedio de cada medición son iguales. Entonces, si denotamos el rango promedio de una medición por M^m, para cada uno de ellos esperamos que se cumpla la igualdad de la ecuación 11.22:

$$M^m = \frac{k+1}{2} \tag{11.22}$$

La tabla 11.10 presenta un resumen de estas estadísticas para el ejemplo de las tres interfaces, donde n_m y M^m corresponden, respectivamente, al tamaño de la muestra y al promedio de los rangos de cada interfaz ($m \in \{A, B, C\}$) y en la muestra combinada ($m = T$).

Grupo:	***A***	***B***	***C***	***T***
n^m	15	15	15	45
M^m	1,1	2,3	2,6	2,0

Tabla 11.10: resumen de los rangos para el ejemplo.

Con estos valores, podemos definir una medida para la variabilidad agregada de las mediciones, dada por la ecuación 11.23:

$$RSS_{bm} = \sum_g n_m \left(M^m - M^T\right)^2 \tag{11.23}$$

Haciendo el cálculo para el ejemplo, tenemos:

$$RSS_{bm} = 15 \cdot (1{,}1 - 2{,}0)^2 + 15 \cdot (2{,}3 - 2{,}0)^2 + 15 \cdot (2{,}6 - 2{,}0)^2 = 18{,}9$$

Con el resultado anterior, podemos ahora calcular el estadístico de prueba que sigue una distribución χ^2 con $k - 1$ grados de libertad, siendo k el número de mediciones repetidas, como muestra la ecuación 11.24:

$$\chi^2 = \frac{12}{k\,(k+1)}\, RSS_{bg} \tag{11.24}$$

Para el ejemplo:

$$\chi^2 = \frac{12}{3 \cdot (3+1)} \cdot 18{,}9 = 18{,}9$$

Una vez más, calculamos el valor p mediante la llamada `pchisq(18.9, 2, lower.tail = FALSE)`, obteniéndose $p < 0{,}002$ ($p \approx 0{,}00007869$). Considerando un nivel de significación $\alpha = 0{,}01$, se **rechazar la hipótesis nula** en favor de la hipótesis alternativa. En consecuencia, concluimos con 99 % confianza que hay evidencia suficiente para descartar que los índices de usabilidad entre las distintas interfaces sean los mismos.

Por supuesto esta conclusión es de tipo ómnibus y es necesario realizar un procedimiento post-hoc para determinar dónde se encuentran las diferencias, que usualmente utiliza múltiples pruebas de rangos con signo de Wilcoxon por cada par de mediciones y aplicando algún factor de corrección.

11.3.2.1 Prueba de Friedman en R

Para aplicar la prueba de Friedman en R, podemos usar la función `friedman.test(formula, data)`, donde:

- `formula`: tiene la forma `<variable_dependiente>~<variable_independiente>| <identificador_caso` [4].
- `data`: matriz de datos en formato largo.

Para los procedimientos post-hoc, podemos utilizar los ajustes ya vistos, como el de Holm, mediante la función `pairwise.wilcox.test()`, del mismo modo descrito para la prueba de Kruskal-Wallis, cuidando en este caso que el argumento `paired` debe tomar forzosamente el valor `TRUE`. Si además queremos conocer el tamaño del efecto detectado para aquellos pares identificados como relevantes, debemos realizar las correspondientes pruebas de rangos con signo de Wilcoxon para todos los pares de grupos que presenten diferencias significativas (Amat Rodrigo, 2016a).

El script 11.11 muestra la realización de la prueba de Friedman para el ejemplo, cuyo resultado se presenta en la figura 11.15, junto al procedimiento post-hoc de Holm. Vemos que, una vez más, la implementación en R introduce correcciones debido a los empates, por lo que el estadístico de prueba y el valor p ómnibus son algo diferentes a los calculados a mano.

Script 11.11: prueba de Friedman y el procedimiento post-hoc de Holm para el ejemplo.

```r
# Construir la mariz de datos
A <- c(3.6, 4.2, 3.5, 3.2, 3.6, 3.5, 3.3, 3.5, 4.1, 3.7, 4.0, 3.5, 3.3, 3.4, 3.6)
B <- c(4.4, 5.0, 4.3, 3.6, 4.5, 4.2, 3.9, 4.3, 4.8, 4.5, 4.8, 4.2, 3.7, 4.0, 4.5)
C <- c(4.9, 5.0, 4.7, 3.2, 5.0, 4.6, 3.6, 4.9, 5.0, 5.0, 5.0, 4.4, 3.2, 3.7, 4.9)

Puntuacion <- c(A, B, C)
Interfaz <- c(rep("A", length(A)), rep("B", length(B)), rep("C", length(C)))
Interfaz <- factor(Interfaz)
Caso <- rep(1:15, 3)

datos <- data.frame(Caso, Puntuacion, Interfaz)

# Establecer nivel de significación
alfa <- 0.01

# Aplicar y mostrar la prueba de Friedman
prueba <- friedman.test(Puntuacion ~ Interfaz | Caso, data = datos)
cat("Resultados de la prueba ómnibus\n")
cat("--------------------------------")
print(prueba)

# Efectuar un procedimiento post-hoc de Holm
# si se encuentran diferencias significativas.
if(prueba[["p.value"]] < alfa) {
  post_hoc <- pairwise.wilcox.test(datos[["Puntuacion"]], datos[["Interfaz"]],
                                   p.adjust.method = "holm",
                                   paired = TRUE, exact = FALSE)

  cat("Resultados del análisis post-hoc\n")
  cat("--------------------------------")
  print(post_hoc)
}
```

11.3.2.2 Nota importante

Por último, debemos saber que va tomando fuerza la idea de que **no se debe usar** la prueba de Friedman. En reemplazo, se está recomendando transformar los datos en rangos y luego aplicar directamente el análisis de varianza sobre los datos *rankeados* (Zimmerman & Zumbo, 1993). Es más, esta idea va ganando adeptos

[4] Este último también puede ser `<identificador_bloque>`

```
Resultados de la prueba ómnibus
-------------------------------
 Friedman rank sum test

data:  Puntuacion and Interfaz and Caso
Friedman chi-squared = 19.552, df = 2, p-value = 5.681e-05

Resultados del análisis post-hoc
--------------------------------
 Pairwise comparisons using Wilcoxon signed rank test with continuity correction

data:  datos[["Puntuacion"]] and datos[["Interfaz"]]

  A     B
B 0.002 -
C 0.002 0.123

P value adjustment method: BH
```

Figura 11.15: salida generada por el script 11.11 considerando fines académicos.

incluso para muestras independientes y el análisis de dos muestras (aplicando pruebas t sobre los datos transformados a rangos).

11.4 EJERCICIOS PROPUESTOS

11.4.1 Transformación de datosd (sección 11.1)

11.1 En la década del 1920 se hicieron los primeros estudios sobre la relación entre la velocidad de un automóvil con la distancia que necesita para detenerse. Los datos de estas pruebas se pueden encontrar el conjunto `cars` del paquete `datasets`. Con ellos, responde la siguiente pregunta: en promedio, ¿se requieren más de 40 pies para detener un vehículo que viaja a más de 10 millas por hora?

11.2 El conjunto `airquality` del paquete `datasets` contiene mediciones diarias de la calidad del aire en la ciudad de New York, EE.UU., registradas de mayo a septiembre de 1973. Verifica si las mediciones de ozono [partes por billón] son, en promedio, iguales en mayo y junio. Usa la escalera de potencias de Tukey si los datos presentan problemas para su análisis.

11.3 Para el conjunto `airquality` del paquete `datasets`, verifica si las mediciones de ozono [partes por billón] son, en promedio, iguales en agosto y septiembre. Usa la transformación Box-Cox si los datos presentan problemas para su análisis.

11.4 El conjunto `WorldPhones` del paquete `datasets` contiene la cantidad de teléfonos (en miles) instalados para diferentes regiones del planeta en algunos años entre 1951 y 1961. Verifica si, en promedio, la cantidad de teléfonos instalados en aquella época eran iguales en Centroamérica, Sudamérica y África. Usa la escalera de potencias de Tukey si los datos presentan problemas para su análisis.

11.5 Para el conjunto `WorldPhones` del paquete `datasets`, verifica si, en promedio, los datos sugieren que la cantidad de teléfonos instalados en Asia, Oceanía y África era similares por aquellos años. Usa la transformación Box-Cox si los datos presentan problemas para su análisis.

11.6 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que necesite aplicar una prueba t de Student para dos muestras independientes, pero que probablemente, por la naturaleza de las variables estudiadas, requiera una transformación de los datos. Identifica bien las variables involucradas y enuncia las hipótesis a docimar. (*Ayuda: variables que presentan distribuciones exponenciales o logarítmicas, entre otras, suelen mostrar mucha asimetría. Ejemplos se encuentran en variables relacionadas con tiempos, precios, ingresos, etc.*)

11.7 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que necesite aplicar una prueba t de Student para dos muestras apareadas, pero que probablemente, por la naturaleza de las variables estudiadas, requiera una transformación de los datos. Identifica bien las variables involucradas y enuncia las hipótesis a docimar. (*Ver la nota de ayuda en* 11.6)

11.8 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera utilizar una prueba ANOVA para muestras independientes, pero que probablemente, por la naturaleza de las variables estudiadas, requiera una transformación de los datos. Identifica bien las variables involucradas y enuncia las hipótesis a docimar. (*Ver la nota de ayuda en* 11.6)

11.9 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera utilizar una prueba ANOVA para medidas repetidas, pero que probablemente, por la naturaleza de las variables estudiadas, requiera una transformación de los datos. Identifica bien las variables involucradas y enuncia las hipótesis a docimar. (*Ver la nota de ayuda en* 11.6)

11.10 Justificando tus suposiciones, inventa muestras plausibles de los datos que se podrían encontrar en el estudio propuesto en la pregunta 11.6 que tengan entre 15 y 18 observaciones cada una.

11.11 Justificando tus suposiciones, inventa muestras plausibles de los datos que se podrían encontrar en el estudio propuesto en la pregunta 11.7 que tengan 22 pares observaciones cada una, y que cumplan las condiciones requeridas por la prueba.

11.12 Justificando tus suposiciones, inventa muestras plausibles de los datos que se podrían encontrar en el estudio propuesto en la pregunta 11.8 que tengan entre 21 y 28 observaciones cada una.

11.13 Justificando tus suposiciones, inventa muestras plausibles de los datos que se podrían encontrar en el estudio propuesto en la pregunta 11.9 con 20 casos.

11.14 Usando las muestras inventadas en 11.10, aplica la escalera de potencias de Tukey y verifica que los datos transformados cumplen las condiciones requeridas por la prueba de hipótesis propuesta.

11.15 Usando las muestras inventados en 11.11, aplica la transformación de Box-Cox y verifica que los datos transformados cumplen las condiciones requeridas por la prueba de hipótesis propuesta.

11.16 Usando las muestras inventadas en 11.12, aplica la escalera de potencias de Tukey y verifica que los datos transformados cumplen las condiciones requeridas por la prueba de hipótesis propuesta.

11.17 Usando las muestras inventados en 11.13, aplica la transformación de Box-Cox y verifica que los datos transformados cumplen las condiciones requeridas por la prueba de hipótesis propuesta.

11.18 Usando los datos transformados en 11.14, realiza la prueba propuesta y entrega una conclusión a la pregunta planteada.

11.19 Usando los datos transformados en 11.15, realiza la prueba propuesta y entrega una conclusión a la pregunta planteada.

11.20 Usando los datos transformados en 11.16, realiza la prueba propuesta, usando la prueba HSD de Tukey para el análisis post-hoc de ser necesario, y entrega una conclusión a la pregunta planteada.

11.21 Usando los datos transformados en 11.17, realiza la prueba propuesta, usando el método de Benjamini y Hochberg para el análisis post-hoc de ser necesario, y entrega una conclusión a la pregunta planteada.

11.4.2 Pruebas no paramétricas con una y dos muestras numéricas (sección 11.2)

11.22 En la década del 1920 se hicieron los primeros estudios sobre la relación entre la velocidad de un automóvil con la distancia que necesita para detenerse. Los datos de estas pruebas se pueden encontrar el conjunto `cars` del paquete `datasets`. Con ellos, responde la siguiente pregunta: la distribución de las distancias necesitadas para detener vehículos antiguos que viajaban a más de 10 millas por hora ¿se centra en un valor menor a 60 pies? No olvides verificar si se cumplen las condiciones de la prueba que estás aplicando.

11.23 El conjunto `airquality` del paquete `datasets` contiene mediciones diarias de la calidad del aire en la ciudad de New York, EE.UU., registradas de mayo a septiembre de 1973. Verifica si la calidad del aire respecto del ozono es la misma los primeros 9 días de agosto que los primeros 9 días de septiembre. No olvides verificar si se cumplen las condiciones de la prueba que estás aplicando.

11.24 El conjunto `ChickWeight` del paquete `datasets` contiene los resultados de un experimento del efecto de 4 tipos de dietas en el crecimiento temprano de pollitos. Verifica si las dietas 1 y 2 producen crecimientos similares. No olvides verificar si se cumplen las condiciones de la prueba que estás aplicando.

11.25 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba de Mann-Whitney para una muestra. Identifica bien las variables involucradas, justifica por qué no usar una prueba paramétrica equivalente, y enuncia las hipótesis a docimar.

11.26 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera utilizar una prueba Mann-Whitney para dos muestras debido a que la escala de la variable dependiente no permite usar una prueba t de Student. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.27 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera utilizar una prueba de sumas con signo de Wilcoxon por problemas con la escala de las mediciones. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.28 Da un ejemplo de una pregunta de investigación sobre el financiamiento de la educación superior en Chile que requiera utilizar una prueba de sumas con signo de Wilcoxon por problemas de normalidad. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.29 Investiga qué alternativas existen para estimar el tamaño del efecto para la prueba de sumas con rangos de Wilcoxon-Mann-Whitney. ¿Están implementadas en R?

11.30 Investiga qué alternativas existen para conocer el poder de una prueba de sumas con rangos de Wilcoxon-Mann-Whitney. ¿Están implementadas en R?

11.31 Investiga qué alternativas existen para estimar el tamaño del efecto para la prueba de sumas de rangos con signo de Wilcoxon. ¿Están implementadas en R?

11.32 Investiga qué alternativas existen para conocer el poder de una prueba de sumas de rangos con signo de Wilcoxon. ¿Están implementadas en R?

11.4.3 Pruebas no paramétricas con más de dos muestras numéricas (sección 11.3)

11.33 Considerando los datos del primer estudio sobre el frenado de automóviles (conjunto `cars` del paquete `datasets`), determina si las distancias requeridas para detenerse ya eran distintas dependiendo si el vehículo viajaba a una velocidad baja, media o alta. En aquella época, década de 1920s, estas categorías estaban definidas, respectivamente, por los siguientes rangos: [8, 13), [13, 20), [21, 25) [millas por hora]. Utiliza una prueba de Kruskal-Wallis para el análisis, y la corrección de Benjamini y Hochberg para comparaciones múltiples. No olvides enunciar las hipótesis y verificar si se cumplen las condiciones de las pruebas que estás aplicando.

11.34 El conjunto `airquality` del paquete `datasets` contiene mediciones de la radiación solar (expresada en Langleys) recibida diariamente (con algunos datos perdidos) en la ciudad de New York, EE.UU., desde mayo a septiembre de 1973. Verifica si la radiación recibida en la cuidad es la misma durante los meses de verano (junio a agosto). No olvides enunciar las hipótesis y verificar si se cumplen las condiciones de las pruebas aplicadas. Utiliza la corrección de Benjamini y Hochberg para comparaciones múltiples de ser necesario.

11.35 El conjunto `ChickWeight` del paquete `datasets` contiene los resultados de un experimento del efecto de 4 tipos de dietas en el crecimiento temprano de pollitos. Verifica si las dietas producen crecimientos disímiles al sexto día de vida de los pollitos, cuando deben trasladarse a los galpones no calefaccionados. Enuncia las hipótesis y verifica si se cumplen las condiciones de las pruebas aplicadas. Utiliza la corrección de Benjamini y Hochberg para comparaciones múltiples de ser necesario.

11.36 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba de Kruskal-Wallis. Identifica bien las variables involucradas, justifica por qué no se podría usar una prueba paramétrica, y enuncia las hipótesis a docimar.

11.37 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera utilizar una prueba Kruskal-Wallis debido a que la escala de la variable dependiente no permite usar una prueba ANOVA. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.38 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera utilizar una prueba de Kruskal-Wallis. Identifica bien las variables involucradas, justifica por qué no se podría usar una prueba paramétrica, y enuncia las hipótesis a docimar.

11.39 Investiga qué alternativas existen para estimar el tamaño del efecto para la prueba Kruskal-Wallis. ¿Están implementadas en R?

11.40 Investiga qué alternativas conocer el poder o estimar el tamaño de las muestras necesitadas para de una prueba de Kruskal-Wallis. ¿Están implementadas en R?

11.41 Considera los resultados que tiene la dieta 4 en el crecimiento temprano de pollitos, que se reportan en el conjunto `ChickWeight` del paquete `datasets`. Verifica si el peso de los pollitos difieren entre los días 4, 5, 8 y 10 de vida. Enuncia las hipótesis y verifica si se cumplen las condiciones de las pruebas aplicadas. Utiliza la corrección de Benjamini y Hochberg para comparaciones múltiples de ser necesario.

11.42 Una farmacéutica comparó tratamientos para la picazón usando como grupo de estudio 10 hombres de entre 20 y 30 años. Cada sujeto recibió un tratamiento diferente cada día, durante siete días consecutivos, en un orden temporal aleatorio. Para cada tratamiento, se les indujo comezón en los antebrazos mediante un estímulo llamado "cowage", y los sujetos anotaron la duración del picor, en segundos. Se puede acceder a estos datos usando la función `read_csv()` del paquete `tidyverse` mediante la siguiente llamada: `read_csv("http://statacumen.com/teach/ADA2/notes/ADA2_notes_Ch05_itch.csv")`. Verifica si existen diferencias en la duración de la comezón entre los fármacos Papaverina (`Papv`), Pentobarbitol (`Pento`) y el placebo (`Placebo`). Enuncia las hipótesis y verifica si se cumplen las condiciones de las pruebas aplicadas. Utiliza la corrección de Benjamini y Hochberg para comparaciones múltiples de ser necesario.

11.43 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba de Friedman. Identifica bien las variables involucradas, justifica por qué no se podría usar una prueba paramétrica, y enuncia las hipótesis a docimar.

11.44 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera utilizar una prueba Friedman debido a que la escala de la variable dependiente no permite usar una prueba ANOVA. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.45 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera utilizar una prueba de Friedman. Identifica bien las variables involucradas, justifica por qué no se podría usar una prueba paramétrica, y enuncia las hipótesis a docimar.

11.46 Investiga qué alternativas existen para estimar el tamaño del efecto para la prueba Friedman. ¿Están implementadas en R?

11.47 Investiga qué alternativas conocer el poder o estimar el tamaño de las muestras necesitadas para de una prueba de Friedman. ¿Están implementadas en R?

11.5 BIBLIOGRAFÍA DEL CAPÍTULO

Amat Rodrigo, J. (2016a). *Test de Friedman*. Consultado el 29 de mayo de 2021, desde https://www.cienciadedatos.net/documentos/21_friedman_test

Amat Rodrigo, J. (2016b). *Test Kruskal-Wallis*. Consultado el 29 de mayo de 2021, desde https://www.cienciadedatos.net/documentos/20_kruskal-wallis_test

Baguley, T. (2012). *Beware the Friedman test!* Consultado el 13 de diciembre de 2021, desde https://seriousstats.wordpress.com/2012/02/14/friedman/

Carchedi, N., De Mesmaeker, D., & Vannoorenberghe, L. (s.f.). RDocumentation. Consultado el 2 de abril de 2021, desde https://www.rdocumentation.org/

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, *4*, 1.

Glen, S. (2021a). *Geometric Mean Definition and Formula*. Consultado el 27 de mayo de 2021, desde https://www.statisticshowto.com/geometric-mean-2/

Glen, S. (2021b). *Kruskal Wallis H Test: Definition, Examples & Assumptions*. Consultado el 5 de junio de 2021, desde https://www.statisticshowto.com/kruskal-wallis/

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583-621.

Lærd Statistics. (2020). *Friedman Test in SPSS Statistics* [Lund Research Ltd.]. Consultado el 5 de junio de 2021, desde https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php

Lane, D. (s.f.). *Online Statistics Education: A Multimedia Course of Study*. Consultado el 4 de mayo de 2021, desde https://onlinestatbook.com/

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

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. John wiley & sons.

United States Census Bureau. (2004). *CT1970p2-13: Colonial and Pre-Federal Statistics*. Consultado el 26 de mayo de 2021, desde https://www2.census.gov/prod2/statcomp/documents/CT1970p2-13.pdf

United States Census Bureau. (2021). Decennial Census of Population and Housing. Consultado el 26 de mayo de 2021, desde https://www.census.gov/programs-surveys/decennial-census/decade.html

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education*, *62*(1), 75-86.