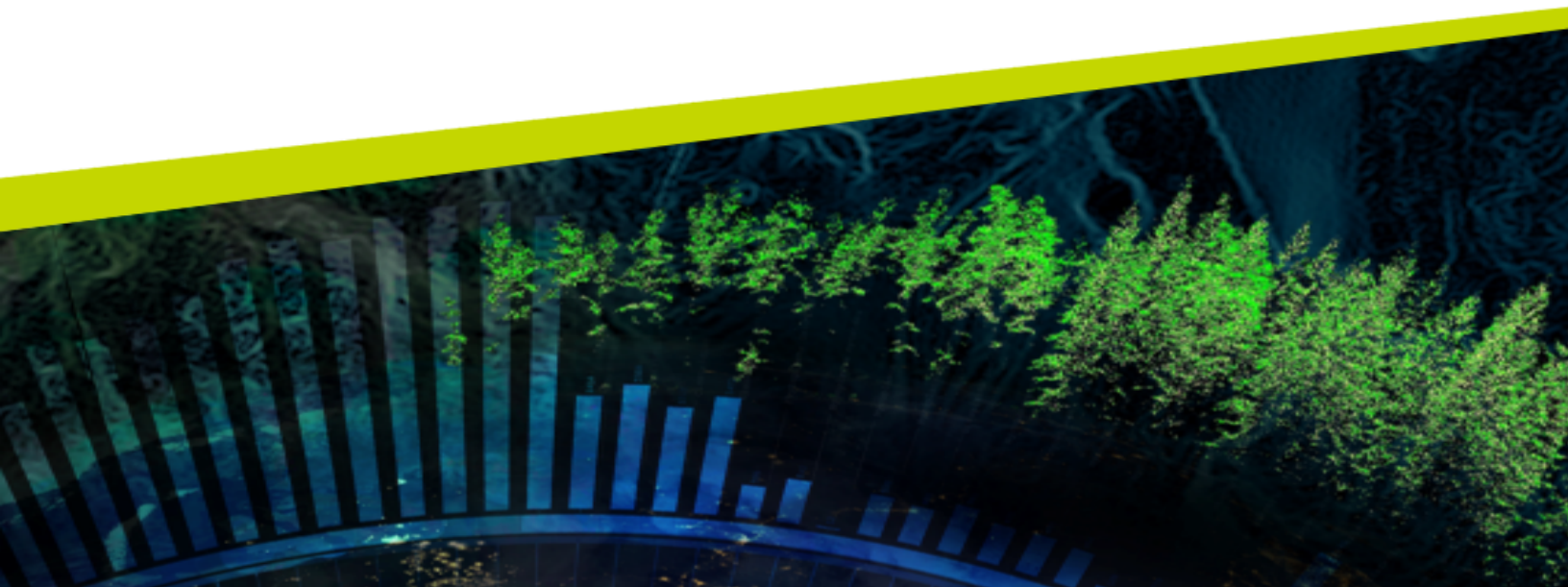
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 10. MÉTODOS CLÁSICOS PARA ENFRENTAR DATOS PROBLEMÁTICOS

La mayoría de los procedimientos estadísticos que hemos estudiado hasta ahora requieren que los datos cumplan con ciertas propiedades o condiciones. Sabemos que esto no siempre ocurre. Estudiamos que para el caso de inferencias con frecuencias (variables categóricas) podemos usar, con cierta pérdida de información, métodos no paramétricos como alternativa (capítulo 7). En este capítulo abordaremos estrategias "clásicas" que podemos utilizar cuando necesitamos hacer inferencias con variables numéricas a partir de "datos problemáticos", que no cumplen plenamente con las condiciones requeridas por los métodos que hemos aprendido hasta ahora. En este contexto, "clásicos" se refiere a que fueron desarrolladas hace bastante tiempo y que se encuentran en la mayoría de los textos introductorios de inferencia. 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).

10.1 TRANSFORMACIÓN DE DATOS

La primera alternativa que se tiene para enfrentar datos problemáticos es transformarlos a una escala diferente donde estos sí cumplan las condiciones necesarias para aplicar la prueba deseada. Estas transformaciones, como aprendimos en los cursos de matemáticas, no son otra cosa que una función que al aplicarla sobre la variable aleatoria original X, obtenemos como resultado una nueva variable aleatoria Y.

Como explica Lane (s.f., pp. 1, 16), existen diversas familias de funciones que usualmente se usan 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.

10.1.1 Transformación lineal

Las **transformaciones lineales** son las más sencillas de las transformaciones, y para hacerlas nos basta con aplicar una función lineal, es decir, de la forma presentada en la ecuación 10.1, donde m y n son constantes.

$$y_i = m \cdot x_i + n \tag{10.1}$$

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:

$$F = 1,8 \cdot C + 32$$

En R, podemos hacer este tipo de transformaciones de forma muy sencilla mediante operaciones aritméticas que se aplican vectorialmente, como muestra el script 10.1. En él se transforma un vector con 4 temperaturas en grados Celsius a grados Fahrenheit. El resultado se muestra en la figura 10.1.

Script 10.1: transformación lineal para convertir grados Celcius a grados Fahrenheit.

```
# Crear un vector con cuatro observaciones en grados Celsius.
Celsius <- c(-8, 0, 29.8, 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)
```

```
Temperaturas en grados Celsius
[1]  -8.0   0.0  29.8 100.0

Temperaturas en grados Fahrenheit
[1]  17.60  32.00  85.64 212.00
```

Figura 10.1: resultado de la transformación lineal del script 10.1.

10.1.2 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. 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 con posibles valores iguales a 0. El script 10.2 aplica esta transformación al peso corporal y al peso cerebral de los animales (líneas 22–23). La figura 10.2 (generada en las líneas 28–43 del script 10.2) muestra gráficamente el resultado de esta transformación para el peso cerebral de los animales.

Muchas veces la transformación logarítmica hace que nos sea más fácil interpretar los datos, evidenciando patrones más claros para la relación entre variables. En la figura 10.3 (creada en las líneas 47–60 del script 10.2), a la derecha, se evidencia una fuerte relación entre el peso corporal y el peso del cerebro tras transformar ambas variables, relación que no podemos percibir con los datos originales (izquierda).

Script 10.2: transformación logarítmica.

```
library(ggpubr)

# Cargar datos
animal <- c("Mountain beaver", "Cow", "Grey wolf", "Goat", "Guinea pig",
            "Dipliodocus", "Asian elephant", "Donkey", "Horse",
            "Potar monkey", "Cat", "Giraffe", "Gorilla", "Human",
            "African elephant", "Triceratops", "Rhesus monkey", "Kangaroo",
            "Golden hamster", "Mouse", "Rabbit", "Sheep", "Jaguar",
            "Chimpanzee", "Brachiosaurus", "Mole", "Pig")

body_weight <- 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,
```

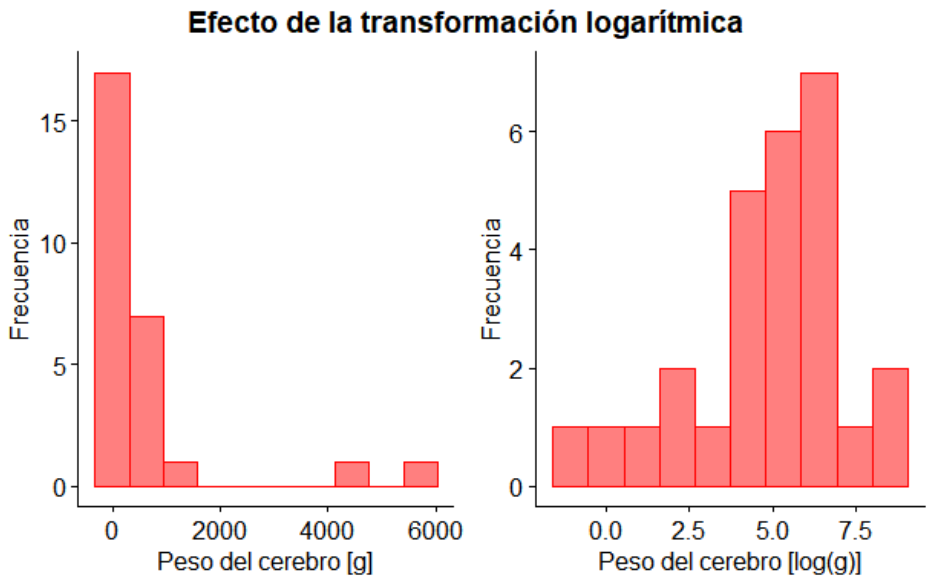


Figura 10.2: histogramas del peso cerebral antes y después de la transformación logarítmica.

```
                  55.5, 100, 52.16, 87000, 0.122, 192)

brain_weight <- 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, body_weight, brain_weight)

# Aplicar transformación logarítmica
log_cuerpo <- log(body_weight)
log_cerebro <- log(brain_weight)
datos <- data.frame(datos, log_cuerpo, log_cerebro)

# Histogramas para el peso cerebral antes y después de la transformación
# logarítmica.
g3 <- gghistogram(datos, x = "brain_weight", bins = 10,
                  xlab = "Peso del cerebro [g]", ylab = "Frecuencia",
                  color = "red", fill = "red")

g4 <- gghistogram(datos, x = "log_cerebro", bins = 10,
                  xlab = "Peso del cerebro [log(g)]", ylab = "Frecuencia",
                  color = "red", fill = "red")

# Crear una única figura con ambos histogramas.
histograma  <- ggarrange(g3, g4, ncol = 2, nrow = 1)

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

histograma <- annotate_figure(histograma, top = titulo)
print(histograma)

# 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 = "body_weight", y = "brain_weight",
```

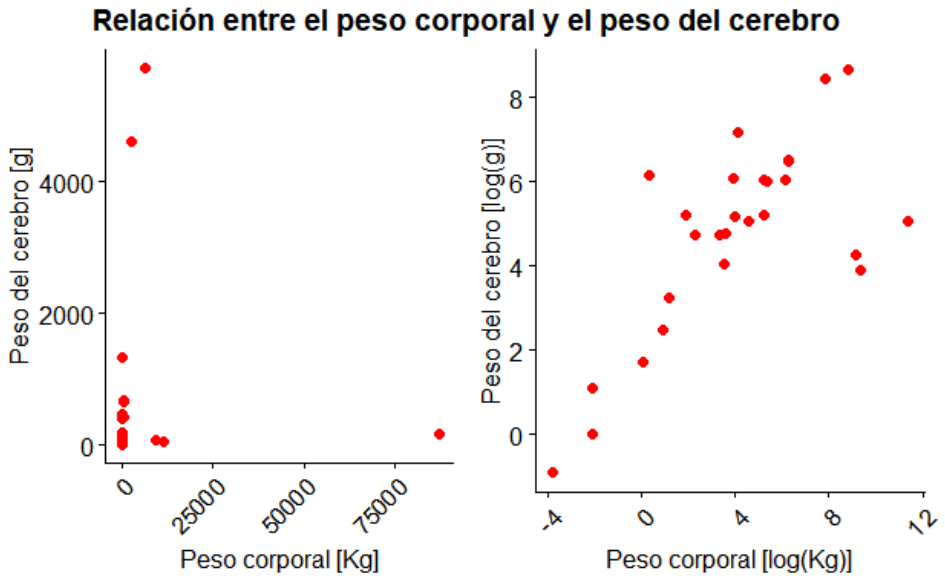


Figura 10.3: gráficos de dispersión para el peso corporal y el peso del cerebro antes y después de las transformaciones logarítmicas.

```
                color = "red", xlab = "Peso corporal [Kg]",
                ylab = "Peso del cerebro [g]") + rotate_x_text(45)

g2 <- ggscatter(datos, x = "log_cuerpo", y = "log_cerebro",
                color = "red", xlab = "Peso corporal [log(Kg)]",
                ylab = "Peso del cerebro [log(g)]") + rotate_x_text(45)

# Crear una única figura con los gráficos de dispersión.
dispersion  <- ggarrange(g1, g2, ncol = 2, nrow = 1)
texto <- "Relación entre el peso corporal y el peso del cerebro"
titulo  <- text_grob(texto, face = "bold", size = 14)
dispersion <- annotate_figure(dispersion, top = titulo)
print(dispersion)
```

Eso sí, tenemos que ser cuidadosos al interpretar los resultados que obtengamos al emplear esta transformación, porque cuando comparamos medias de datos tras una transformación logarítmica, en realidad estamos comparando ¡**medias geométricas**! Recordemos que la media geométrica se calcula de acuerdo a la ecuación 10.2 y suele ocuparse para representar tasas de crecimiento o de interés (Glen, 2021a).

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \tag{10.2}$$

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:

$$\frac{0+1+2}{3} = 1$$

Al revertir la transformación para la media, tenemos que:

$$10^1 = 10$$

Lo que es distinto a la media aritmética de los datos originales:

$$\frac{1 + 10 + 100}{3} = 37$$

A su vez, la media geométrica del vector original es:

$$\sqrt[3]{1 \cdot 10 \cdot 100} = 10$$

Así, si dos variables a las que se ha aplicado la transformación logarítmica tienen igual media, entonces **las medias geométricas de las variables originales son iguales**.

10.1.3 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 10.3, donde λ puede tomar cualquier valor real y se escoge de modo que la distribución de los datos transformados sea lo más cercana a la normal posible. 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.

$$y = x^\lambda \tag{10.3}$$

Formalmente, la transformación de Tukey se define según la ecuación 10.4, aunque (por la falta de computadores) suelen usarse únicamente aquellas que se muestran en la tabla 10.1. 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.

$$\tilde{x}_\lambda = \begin{cases} x^\lambda & \lambda > 0 \\ \log(x) & \lambda = 0 \\ -(x^\lambda) & \lambda < 0 \end{cases} \tag{10.4}$$

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

Tabla 10.1: 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. La figura 10.4 (que resulta de líneas 18–30 del script 10.3) muestra, a la izquierda, el histograma para la población (en millones de habitantes) y, a la derecha, un gráfico de dispersión para la población por año. Podemos ver claramente que la distribución de la población presenta una fuerte asimetría hacia la izquierda y que la población parece aumentar de manera exponencial durante ese periodo.

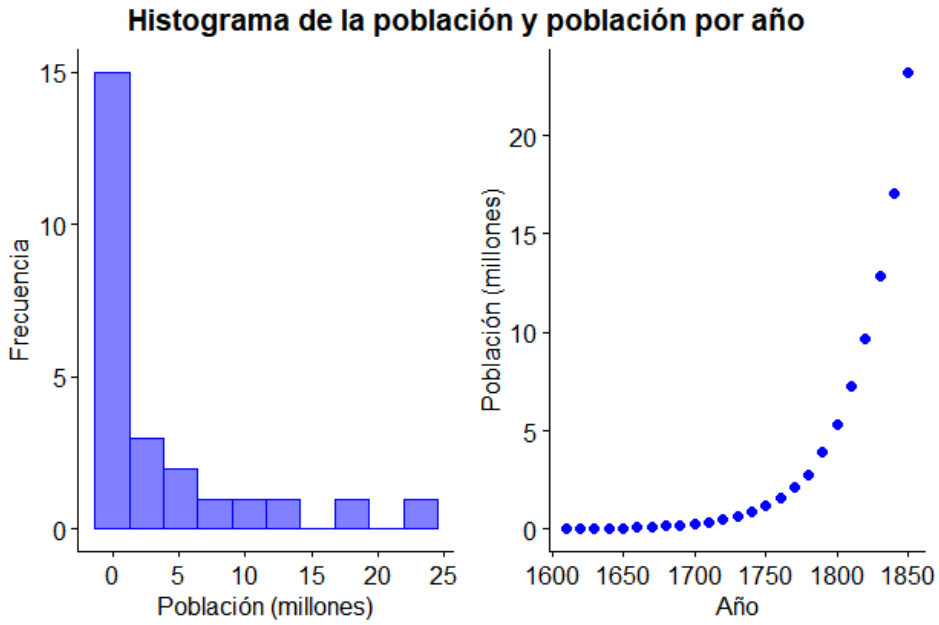


Figura 10.4: histograma de la población histórica de Estados Unidos y gráfico de dispersión de la población por año.

La figura 10.5 (generada en las líneas 46–75 del script 10.3) muestra los gráficos de dispersión para la población por año tras aplicar la transformación de Tukey con diferentes valores de λ a la población. En ella podemos observar que la curva cambia gradualmente de convexa a cóncava a medida que aumenta el valor de λ, lo que deja entrever que, para obtener un resultado que sea lo más cercano posible a una línea recta, tenemos que resolver un problema de optimización que minimice los valores residuales tras ajustar una recta a los puntos transformados. De las transformaciones presentadas en la figura 10.5, la más cercana a una recta se obtiene para $\lambda = 0$.

En párrafos anteriores mencionamos que la transformación de Tukey también permite reducir la asimetría en la distribución de los datos, como muestra la figura 10.6 (construida en las líneas 79–108 del script 10.3). En ella podemos notar que, a medida que λ aumenta, se reduce la asimetría negativa.

Como podemos suponer, reducir la asimetría nos ayuda a cumplir el requisito de normalidad que imponen muchas pruebas estadísticas, permitiéndonos así lograr resultados teóricamente más precisos. Sin embargo, una vez más tenemos que ser cuidadosos y 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!

En R, el paquete `rcompanion` incluye la función `transformTukey(x, start, end, int, plotit, verbose, quiet, statistic, returnLambda)`, donde:

- `x`: vector de valores a transformar.
- `start`: valor inicial de λ a evaluar.
- `end`: valor final de λ a evaluar.
- `int`: intervalo entre los valores de λ a evaluar.
- `plotit`: si toma valor `TRUE`, 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.

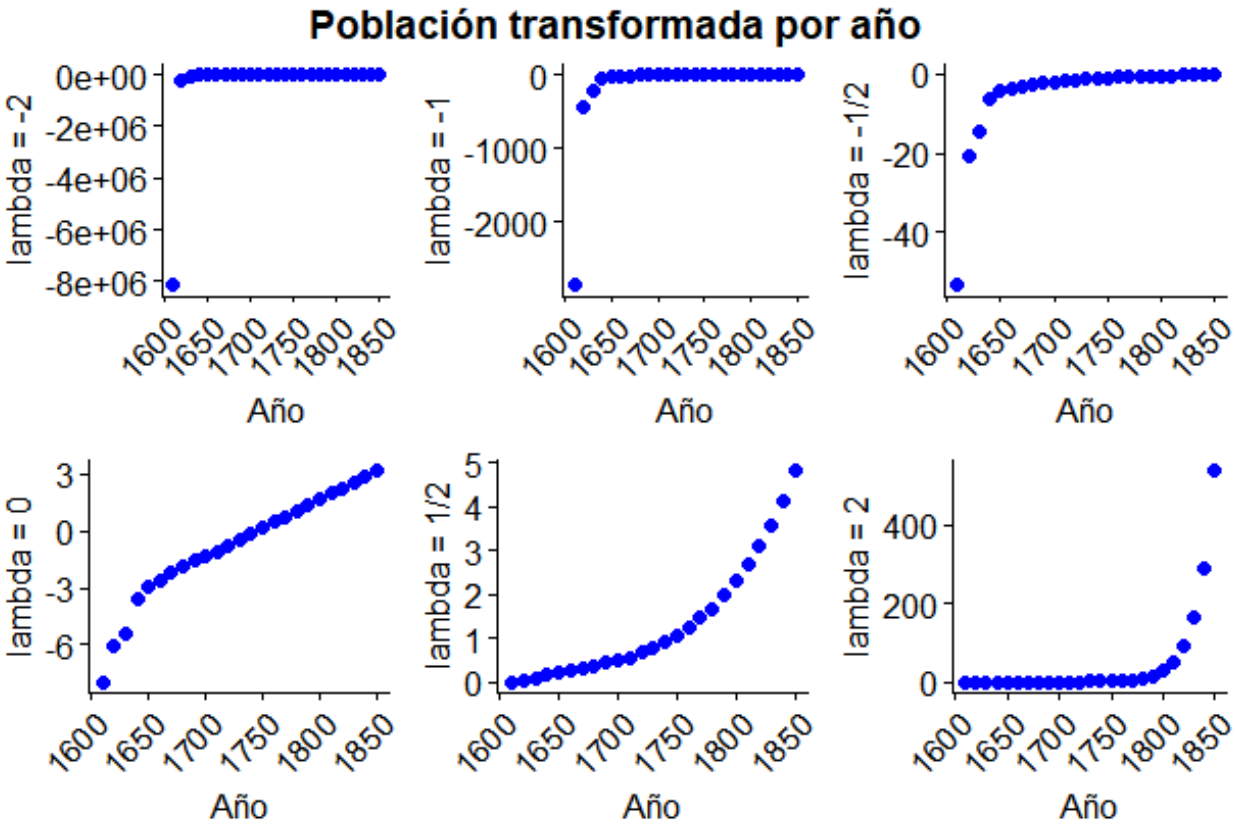


Figura 10.5: población de Estados Unidos por año tras aplicar la transformación de Tukey con distintos valores de λ.

- `quiet`: si toma valor `TRUE`, no muestra información alguna por pantalla.
- `statistic`: si toma valor 1, usa la prueba de normalidad de Shapiro-Wilk. Con valor 2, usa la prueba de Anderson-Darling.
- `returnLambda`: si toma valor `TRUE`, devuelve el valor de λ. Si toma valor `FALSE`, devuelve los datos transformados.

Tras llamar a la función `transformTukey()` con los datos de la población de Estados Unidos (script 10.3, líneas 111–112), obtenemos que el valor óptimo de λ es 0,12 y los gráficos de la figura 10.7. El gráfico 10.7a nos muestra que el valor óptimo de λ es aquel que maximiza el estadístico entregado por la prueba de normalidad. A su vez, en la figura 10.7b vemos que la distribución obtenida con $\lambda = 0,12$ se asemeja mucho más a la normal que la de los datos originales o que cualquiera de las presentadas en la figura 10.6, lo que se ve confirmado por el gráfico Q-Q de la figura 10.7c.

Script 10.3: transformación de Tukey para la población total de Estados Unidos.

```
1  library(ggpubr)
2  library(rcompanion)
3
4  # Cargar datos
5  Year <- c(1610, 1620, 1630, 1640, 1650, 1660, 1670, 1680, 1690, 1700, 1710,
6            1720, 1730, 1740, 1750, 1760, 1770, 1780, 1790, 1800, 1810, 1820,
7            1830, 1840, 1850)
8
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, 1.17076, 1.593625, 2.148076, 2.780369,
```

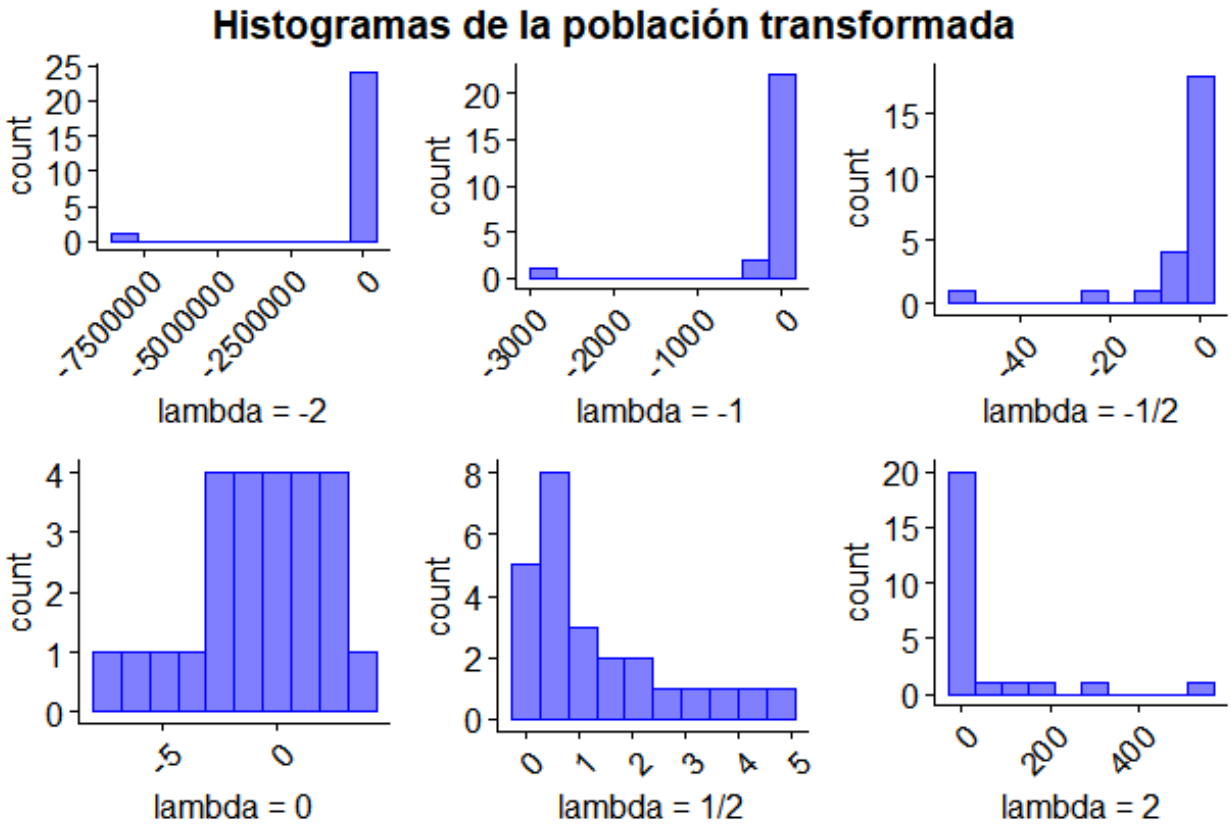
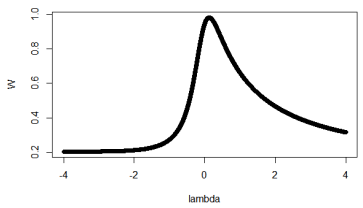
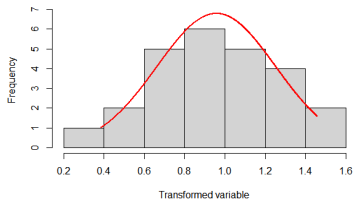


Figura 10.6: histograma de la población de Estados Unidos tras aplicar la transformación de Tukey con distintos valores de λ.

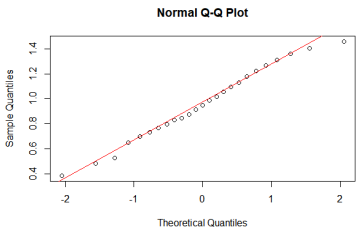
```
                3.929214, 5.308483, 7.239881, 9.638453, 12.86602, 17.069453,
                23.191876)

datos <- data.frame(Year, Population)

# Gráfico de dispersión e histograma.
g1 <- gghistogram(datos, x = "Population", bins = 10,
                  xlab = "Población (millones)", ylab = "Frecuencia",
                  color = "blue", fill = "blue")

g2 <- ggscatter(datos, x = "Year", y = "Population", color = "blue",
                xlab = "Año", ylab = "Población (millones)")

# Histograma de la población y población por año
original  <- ggarrange(g1, g2, ncol = 2, nrow = 1)
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)

# Transformaciones de la población
lambda_menos_dos <- -1 / (datos$Population ** 2)
lambda_menos_uno <- -1 / datos$Population
lambda_menos_un_medio <- -1 / sqrt(datos$Population)
lambda_cero <- log(datos$Population)
```



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



(b) histograma de la población transformada con el valor óptimo de λ.



(c) gráfico Q-Q de la población transformada con el valor óptimo de λ.

Figura 10.7: gráficos entregados por `transformTukey()`.

```
lambda_un_medio <- sqrt(datos$Population)
lambda_dos <- datos$Population ** 2

transformaciones <- data.frame(datos, 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 = "blue", xlab = "Año",
                 ylab = "lambda = -2") + rotate_x_text(45)

gt2 <- ggscatter(transformaciones, x = "Year", y = "lambda_menos_uno",
                 color = "blue", xlab = "Año",
                 ylab = "lambda = -1") + rotate_x_text(45)

gt3 <- ggscatter(transformaciones, x = "Year", y = "lambda_menos_un_medio",
                 color = "blue", xlab = "Año",
                 ylab = "lambda = -1/2") + rotate_x_text(45)

gt4 <- ggscatter(transformaciones, x = "Year", y = "lambda_cero",
                 color = "blue", xlab = "Año",
                 ylab = "lambda = 0") + rotate_x_text(45)

gt5 <- ggscatter(transformaciones, x = "Year", y = "lambda_un_medio",
                 color = "blue", xlab = "Año",
                 ylab = "lambda = 1/2") + rotate_x_text(45)

gt6 <- ggscatter(transformaciones, x = "Year", y = "lambda_dos",
                 color = "blue", xlab = "Año",
                 ylab = "lambda = 2") + rotate_x_text(45)

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

# Histogramas para la transformación de Tukey de la población y el año,
# usando distintos valores de lambda.
h1 <- gghistogram(transformaciones, bins = 10, x = "lambda_menos_dos",
```

```
                    color = "blue", fill = "blue",
                    xlab = "lambda = -2") + rotate_x_text(45)

h2 <- gghistogram(transformaciones, bins = 10, x = "lambda_menos_uno",
                    color = "blue", fill = "blue",
                    xlab = "lambda = -1") + rotate_x_text(45)

h3 <- gghistogram(transformaciones, bins = 10, x = "lambda_menos_un_medio",
                    color = "blue", fill = "blue",
                    xlab = "lambda = -1/2") + rotate_x_text(45)

h4 <- gghistogram(transformaciones, bins = 10, x = "lambda_cero",
                    color = "blue", fill = "blue",
                    xlab = "lambda = 0") + rotate_x_text(45)

h5 <- gghistogram(transformaciones, bins = 10, x = "lambda_un_medio",
                    color = "blue", fill = "blue",
                    xlab = "lambda = 1/2") + rotate_x_text(45)

h6 <- gghistogram(transformaciones, bins = 10, x = "lambda_dos",
                    color = "blue", fill = "blue",
                    xlab = "lambda = 2") + rotate_x_text(45)

# Crear una única figura con todos los gráficos de dispersión.
histograma  <- ggarrange(h1, h2, h3, h4, h5, h6, ncol = 3, nrow = 2)
texto <- "Histogramas de la población transformada"
titulo  <- text_grob(texto, face = "bold", size = 14)
histograma <- annotate_figure(histograma, top = titulo)
print(histograma)

# Buscar la mejor transformación de Tukey usando una función de R.
transformacion <- transformTukey(datos$Population, start = -4, end = 4,
                                 int = 0.001, returnLambda = TRUE)
```

10.1.4 Transformaciones Box-Cox

La **transformación Box-Cox** es una versión escalada de la transformación de Tukey, dada por la ecuación 10.5:

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda} \tag{10.5}$$

Si bien a primera vista parece muy diferente a la ecuación 10.4, podemos reescribirla 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:

$$\lim_{\lambda \to 0} \frac{\left(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \ldots\right) - 1}{\lambda} = \frac{0}{0}$$

Tras aplicar la regla de l'Hôpital, obtenemos finalmente que:

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

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

La figura 10.8 (creada con el script 10.4, líneas 39–64) muestra los 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 al que obtuvimos con la transformación de Tukey (figura 10.5).

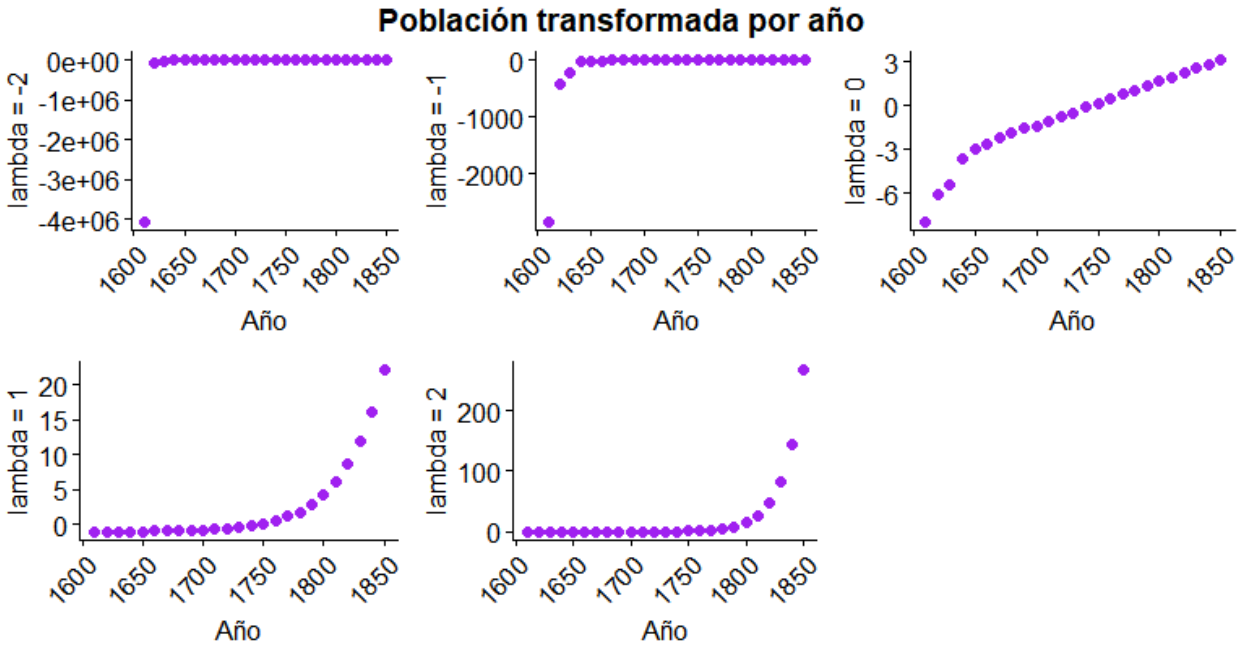


Figura 10.8: población de Estados Unidos por año tras aplicar la transformación de Box-Cox con distintos valores de λ.

Una característica interesante de esta transformación es que, para cualquier valor de λ, $x'_\lambda = 0$ cuando $x = 1$. Podemos observar esto claramente en la figura 10.9, que compara transformaciones Box-Cox con distintos valores de λ con $\log(x)$.

El paquete `DescTools` de R incluye varias funciones que permiten efectuar la transformación Box-Cox (Carchedi y col., 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`.

Donde:

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

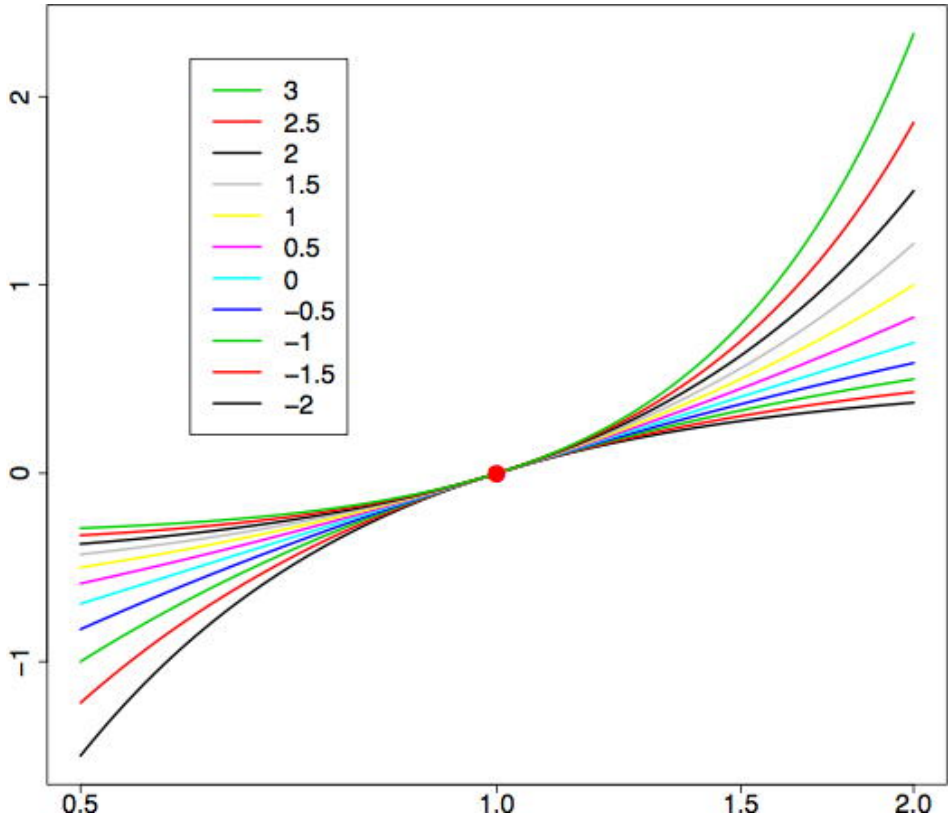
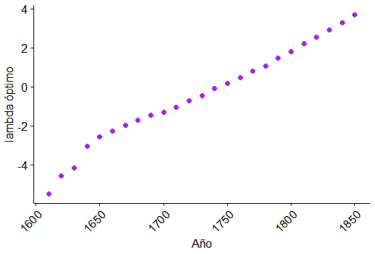


Figura 10.9: ejemplos de la transformación Box-Cox versus $\log(x)$. Fuente: (Lane, s.f., p. 16).

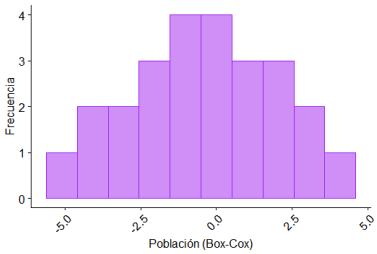
En las líneas 67–70 del script 10.4 se determina el valor óptimo del parámetro λ, obteniéndose como resultado $\lambda = 0,09942656$, para luego efectuar la transformación Box-Cox correspondiente de la población de Estados Unidos. La figura 10.10 (creada en el script 10.4, líneas 84–87) muestra gráficamente el resultado de la transformación. En el gráfico 10.10a podemos ver que la relación entre la población transformada y el año se asemeja a una recta, mientras que el histograma de la figura 10.10b se parece bastante a una distribución normal, hecho que vemos confirmado por el gráfico Q-Q de la figura 10.10c.

Script 10.4: transformación de Box-Cox para la población total de Estados Unidos.

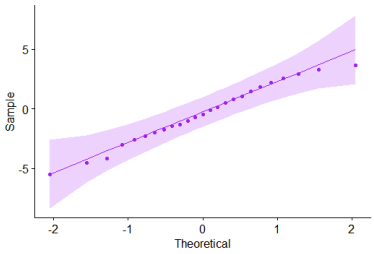
```
1  library(ggpubr)
2  library(DescTools)
3
4  # Cargar datos
5  Year <- c(1610, 1620, 1630, 1640, 1650, 1660, 1670, 1680, 1690, 1700, 1710,
6            1720, 1730, 1740, 1750, 1760, 1770, 1780, 1790, 1800, 1810, 1820,
7            1830, 1840, 1850)
8
9  Population <- c(0.00035, 0.002302, 0.004646, 0.026634, 0.050368, 0.075058,
10                 0.111935, 0.151507, 0.210372, 0.250888, 0.331711, 0.466185,
11                 0.629445, 0.905563, 1.17076, 1.593625, 2.148076, 2.780369,
12                 3.929214, 5.308483, 7.239881, 9.638453, 12.86602, 17.069453,
13                 23.191876)
14
15 datos <- data.frame(Year, Population)
16
17 # Transformación de Box-cox
18 box_cox <- function(x, lambda) {
19   if(lambda == 0) {
```



(a) población de Estados Unidos por año tras aplicar la transformación de Box-Cox con λ óptimo.



(b) histograma de la población transformada con el valor óptimo de λ.



(c) gráfico Q-Q de la población transformada con el valor óptimo de λ.

Figura 10.10: gráficos de población de Estados Unidos por año usando la transformación de Box-Cox.

```
    return(log(x))
  }

  resultado <- (x ** lambda -1) / lambda
  return(resultado)
}

# Transformaciones de la población
lambda_menos_dos <- box_cox(datos$Population, -2)
lambda_menos_uno <- box_cox(datos$Population, -1)
lambda_cero <- box_cox(datos$Population, 0)
lambda_uno <- box_cox(datos$Population, 1)
lambda_dos <- box_cox(datos$Population, 2)

transformaciones <- data.frame(datos, lambda_menos_dos, lambda_menos_uno,
                               lambda_cero, lambda_uno, lambda_dos)

# Gráficos de dispersión para la transformación de Box-Cox de la población y
# el año, usando distintos valores de lambda.
gt1 <- ggscatter(transformaciones, x = "Year", y = "lambda_menos_dos",
                 color = "purple", xlab = "Año",
                 ylab = "lambda = -2") + rotate_x_text(45)

gt2 <- ggscatter(transformaciones, x = "Year", y = "lambda_menos_uno",
                 color = "purple", xlab = "Año",
                 ylab = "lambda = -1") + rotate_x_text(45)

gt3 <- ggscatter(transformaciones, x = "Year", y = "lambda_cero",
                 color = "purple", xlab = "Año",
                 ylab = "lambda = 0") + rotate_x_text(45)

gt4 <- ggscatter(transformaciones, x = "Year", y = "lambda_uno",
                 color = "purple", xlab = "Año",
                 ylab = "lambda = 1") + rotate_x_text(45)

gt5 <- ggscatter(transformaciones, x = "Year", y = "lambda_dos",
                 color = "purple", xlab = "Año",
                 ylab = "lambda = 2") + rotate_x_text(45)

# Crear una única figura con todos los gráficos de dispersión.
dispersion  <- ggarrange(gt1, gt2, gt3, gt4, gt5, ncol = 3, nrow = 2)
texto <- "Población transformada por año"
```

```r
titulo  <- text_grob(texto, face = "bold", size = 14)
dispersion <- annotate_figure(dispersion, top = titulo)
print(dispersion)

# Buscar la mejor transformación Box-Cox usando funciones de R.
lambda <- BoxCoxLambda(datos$Population, lower = -4, upper = 4)
cat("Lambda óptimo:", lambda)
transformacion <- BoxCox(datos$Population, lambda)
datos <- data.frame(datos, transformacion)

# Graficar los datos transformados.
g1 <- ggqqplot(transformacion, color = "purple")
print(g1)

g2 <- gghistogram(datos, bins = 10, x = "transformacion", color = "purple",
                  fill = "purple", xlab = "Población (Box-Cox)",
                  ylab = "Frecuencia") + rotate_x_text(45)

print(g2)

# Gráfico de dispersión para la transformación de Box-Cox de la población y
# el año, usando lambda óptimo.
g3 <- ggscatter(datos, x = "Year", y = "transformacion", color = "purple",
                xlab = "Año", ylab = "lambda óptimo") + rotate_x_text(45)

print(g3)
```

10.2 INFERENCIA NO PARAMÉTRICA CON VARIABLES NUMÉRICAS

En el capítulo 7 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, las pruebas de Wald o de Wilson. Mencionamos también que este problema también puede ocurrir para el caso de inferir con medias, por lo que en este capítulo conoceremos alternativas no paramétricas para las pruebas t de Student (para una y dos medias) y ANOVA (para más de dos medias). Para ello nos basaremos principalmente en Lowry (1999, caps. 11a, 12a, 14a, 15a), Glen (2021b) y Lærd Statistics (2020).

10.3 PRUEBAS PARA UNA O DOS MUESTRAS

En el capítulo 5 aprendimos que la prueba t de Student es adecuada para inferir acerca de una o dos medias muestrales, siempre y cuando se verifiquen algunas condiciones. En el caso de la prueba t de una muestra (o de la diferencia de dos muestras pareadas):

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

En el caso de dos muestras independientes:

1. Cada muestra cumple las condiciones para usar la distribución t.

2. Las muestras son independientes entre sí.

Es importante mencionar también que la distribución normal es continua, de donde se desprende que la escala de medición empleada para la medición de las muestras debe ser de intervalos iguales.

Como ya vimos en el capítulo 7, si usamos la prueba t en un escenario en que no se cumple alguna de estas condiciones, el resultado no sería válido pues carecería de sentido y, en consecuencia, también lo harían las conclusiones que se obtengan a partir de él.

10.3.1 Prueba de suma de rangos de Wilcoxon

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 con muestras independientes. Pese a ser no paramétrica, 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").

Consideremos el siguiente contexto para estudiar 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 de software. Con este fin, la empresa ha seleccionado al azar a 23 voluntarias y voluntarios, 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 corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. La tabla 10.2 muestra las evaluaciones realizadas por cada participante.

	Interfaz A	**Interfaz B**
	2,7	5,0
	6,6	1,4
	1,6	5,6
	5,1	4,6
	3,7	6,7
	6,1	2,7
	5,0	1,3
	1,4	6,3
	1,8	3,7
	1,5	1,3
	3,0	6,8
	5,3	
Media	3,65	4,13

Tabla 10.2: evaluación de las interfaces de usuario 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 existe la misma diferencia de usabilidad (2 puntos)? Pocas escalas Likert tienen estudios de reproducibilidad que aseguren esta consistencia, por lo que no podríamos asumir que la escala es de intervalos iguales en este

ejemplo. En segundo lugar, al revisar los histogramas para las muestras (figura 10.11) podemos observar que las distribuciones no se asemejan a una normal.

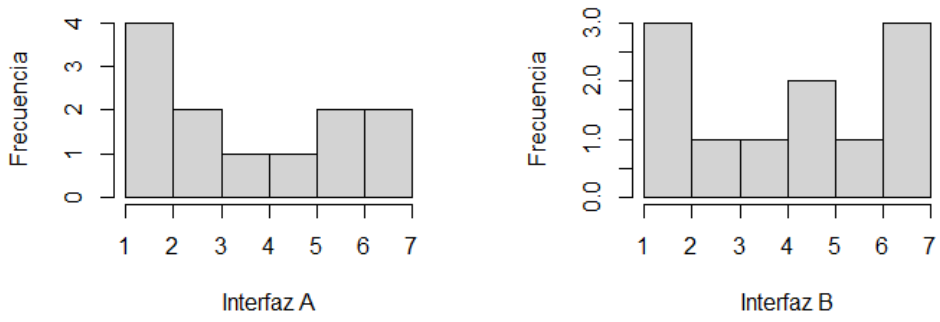


Figura 10.11: histogramas de las muestras.

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 (se distribuyen de igual forma).
H_A: sí hay diferencia en la usabilidad de ambas interfaces (distribuciones distintas).

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 las poblaciones de puntuaciones de usabilidad, es decir, nos entregan **menos información** que la prueba paramétrica equivalente.

El primer paso de la prueba 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, *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 10.3 muestra el resultado de este proceso. Podemos notar que hay dos observaciones con 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 $5,0$.

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

$$T_A = 3,5 + 5,0 + 6,0 + 7,0 + 8,5 + 10,0 + 11,5 + 14,5 + 16,0 + 17,0 + 19,0 + 21,0 = 139$$

De manera análoga, para la muestra B se tiene:

$$T_B = 1,5 + 1,5 + 3,5 + 8,5 + 11,5 + 13,0 + 14,5 + 18,0 + 20,0 + 22,0 + 23,0 = 137$$

La suma de rangos para la muestra combinada siempre está dada por la ecuación 10.6.

$$T_T = \frac{n_T \cdot (n_T + 1)}{2} \tag{10.6}$$

Para el ejemplo:

$$T_T = \frac{23 \cdot (23 + 1)}{2} = 276$$

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

Tabla 10.3: muestras combinadas con rango.

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 se determina siempre mediante la ecuación 10.6 y la media de rangos de la muestra combinada es siempre como muestra la ecuación 10.7.

$$\mu = \frac{n_T \cdot (n_T + 1)}{2} \cdot \frac{1}{n_T} = \frac{n_T + 1}{2} \tag{10.7}$$

Para el ejemplo:

$$\mu = \frac{23 + 1}{2} = 12$$

En consecuencia, la hipótesis nula en el dominio de los rangos es que las medias de los rangos de las dos muestras son iguales. Si la hipótesis nula fuera cierta, las observaciones en ambas muestras serían similares, por lo que, al ordenar la muestra combinada, ambas muestras se mezclarían de manera homogénea. En consecuencia, deberíamos esperar que los promedios de rangos para cada muestra se aproximen al rango promedio de la muestra combinada, es decir, que T_A y T_B se aproximen a los siguientes valores:

$$\mu_A = n_A \cdot \frac{(n_T) + 1}{2} = 12 \cdot \frac{(23 + 1)}{2} = 144$$
$$\mu_B = n_B \cdot \frac{(n_T) + 1}{2} = 11 \cdot \frac{(23 + 1)}{2} = 132$$

La prueba de Wilcoxon-Mann-Whitney tiene dos variantes, una para muestras grandes y otra para muestras pequeñas, que se diferencian a partir de este punto.

10.3.1.1 Prueba de suma de rangos de Wilcoxon para muestras grandes

Hasta ahora, hemos determinado que:

- El valor observado $T_A = 139$ pertenece a una distribución muestral con media $\mu_A = 144$.
- El valor observado $T_B = 137$ pertenece a una distribución muestral con media $\mu_B = 132$.

Bajo el supuesto de que la hipótesis nula sea verdadera, podríamos demostrar que las distribuciones muestrales de T_A y T_B tienen la misma desviación estándar, dada por la ecuación 10.8.

$$\sigma_T = \sqrt{\frac{n_A \cdot n_B \cdot (n_T + 1)}{12}} \tag{10.8}$$

Con lo que:

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

Cuando **ambas muestras tienen tamaño mayor o igual a 5**, siguiendo un procedimiento similar al descrito en la primera sección del capítulo 4, podemos demostrar que las distribuciones muestrales de T_A y T_B tienden a aproximarse a la 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 T_A o T_B, dado por la ecuación 10.9, donde:

- T_{obs} es cualquiera de los valores observados, T_A o T_B.
- μ_{obs} es la media de la distribución muestral de T_{obs}.
- σ_T es la desviación estándar de la distribución muestral de T_{obs} (es decir, el error estándar).

$$z = \frac{(T_{obs} - \mu_{obs}) \pm 0,5}{\sigma_T} \tag{10.9}$$

Puesto que las distribuciones muestrales de T son intrínsecamente discretas (solo pueden asumir valores con decimales cuando existen rangos empatados), debemos emplear un factor de corrección de continuidad:

- $-0,5$ si $T_{obs} > \mu_{obs}$.
- $0,5$ si $T_{obs} < \mu_{obs}$.

Volviendo al ejemplo, tenemos:

$$z_A = \frac{(139 - 144) + 0,5}{16,248} = -0,277$$
$$z_B = \frac{(137 - 132) - 0,5}{\sigma_T} = 0,277$$

Los valores z obtenidos a partir de T_A y T_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 z: si para el ejemplo tuviésemos como hipótesis alternativa que la interfaz

A es mejor que la interfaz B, entonces esperaríamos que las observaciones de mayor rango estuvieran en el grupo A, por lo que z_A tendría que ser positivo.

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, cuya hipótesis alternativa es bilateral, podemos calcular el valor p en R mediante la llamada `2 * pnorm(-0.277, mean = 0, sd = 1, lower.tail = TRUE)`, obteniéndose como resultado $p = 0,782$.

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 tienen niveles de usabilidad similares.

10.3.1.2 Prueba de suma de rangos de Wilcoxon para muestras pequeñas

Cuando las muestras son pequeñas (menos de 5 observaciones), 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 equivalentes a los ya obtenidos.

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 como indica la ecuación 10.10. Fijémonos en que el valor máximo para la suma de rangos de una muestra se produce cuando esta contiene los n_x rangos mayores de la muestra combinada.

$$T_{x[max]} = n_x \cdot n_y + \frac{n_x \cdot (n_x + 1)}{2} \tag{10.10}$$

Así, para el ejemplo:

$$T_{A[max]} = 12 \cdot 11 + \frac{12 \cdot (12 + 1)}{2} = 210$$
$$T_{B[max]} = 11 \cdot 12 + \frac{11 \cdot (11 + 1)}{2} = 198$$

Con esto podemos definir un nuevo estadístico de prueba U, como muestra la ecuación 10.11.

$$U_x = T_{x[max]} - T_x \tag{10.11}$$

Por lo que:

$$U_A = 210 - 139 = 71$$
$$U_B = 198 - 137 = 61$$

El valor del estadístico de prueba es el mínimo entre U_A y U_B, por lo que $U = 61$.

Debemos notar que siempre se cumple la identidad presentada en la ecuación 10.12, por lo que podemos escoger cualquiera de los valores U obtenidos para realizar el resto del procedimiento.

$$U_A + U_B = n_A \cdot n_B \tag{10.12}$$

Si la hipótesis nula fuese cierta:

$$U_A = T_{A[max]} - \mu_A = 210 - 144 = 66$$
$$U_B = T_{B[max]} - \mu_B = 198 - 132 = 66$$

Formalmente, entonces, si la hipótesis nula fuera verdadera, esperaríamos que:

$$U_A = U_B = \frac{n_A \cdot n_B}{2}$$

En consecuencia, la pregunta asociada a la prueba de hipótesis es: si la hipótesis nula es verdadera (no hay diferencias significativas en la usabilidad de ambas interfaces), ¿qué tan probable es obtener un valor de U al menos tan pequeño como el observado ($U = 61$)? Para responder a esta pregunta, seguimos un procedimiento similar al que ya conocimos para la prueba exacta de Fisher (capítulo 7): 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 ¡existen 676.039 combinaciones posibles!

Aunque R no ofrece herramientas para calcular el valor p a partir del estadístico U (pues utiliza el estadístico W, propuesto por Frank Wilcoxon en 1945, que lleva a los mismos resultados), afortunadamente existen tablas que permiten conocer el máximo valor de U para el cual se rechaza la hipótesis nula para un nivel de significación dado sin tener que revisar todas las combinaciones. Considerando $\alpha = 0,05$ para una prueba bilateral, el valor crítico es $U = 33$ (Real Statistics Using Excel, s.f.). Puesto que $61 > 33$, 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.

10.3.1.3 Prueba de suma de rangos de Wilcoxon en R

Como ya dijimos, 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 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. Para aplicar la prueba con una única muestra, `y` debe ser nulo (por defecto, lo es).
- `paired`: booleano con valor falso para indicar que las muestras son independientes (se asume por defecto).
- `alternative`: señala el tipo de hipótesis alternativa: bilateral ("two.sided") o unilateral ("less" o "greater").
- `mu`: valor nulo (ver nota más abajo).
- `conf.level`: nivel de confianza.

El script 10.5 muestra la aplicación de esta prueba para el ejemplo, obteniéndose los resultados que se presentan en la figura 10.12.

Script 10.5: prueba de Mann-Whitney para el ejemplo.

```
# Ingresar los datos.
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)
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.
```

```
Wilcoxon rank sum test with continuity correction

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

Figura 10.12: resultado de la prueba de Mann-Whitney (en rigor, de la prueba para el ejemplo.

```r
alfa <- 0.05

# Hacer la prueba de Mann-Whitney.
prueba <- wilcox.test(a, b, alternative = "two.sided", conf.level = 1 - alfa)
print(prueba)
```

Nota. Hay un poco de confusión (especialmente en Internet) respecto a las hipótesis que son contrastadas por estas pruebas. El argumento `mu` de la función `wilcox.test()` define el valor nulo de la prueba. Cuando se trabaja con una muestra (no ejemplificado en este capítulo) o la diferencia de dos muestras apareadas (como se discute en la siguiente sección), se prueba la hipótesis nula que la distribución de origen es simétrica en torno al valor `mu`. Esto equivale a decir que `mu` es el valor nulo para **la mediana** de la distribución de origen, en el primer caso, o de la distribución de las diferencias de las variables de origen, en el segundo. Cuando se comparan dos grupos independientes, se prueba la hipótesis que los parámetros de localización de las distribuciones de `x` e `y` difieren en `mu`. Solo cuando estas distribuciones de origen tienen **la misma forma** (igual simetría y varianza), esto es equivalente a verificar que poblaciones tienen **las mismas medianas**. Cuando la prueba es unilateral, solo se revisa si el parámetro de localización de la distribución de `x` está a la izquierda (`alternative = "less"`) o a la derecha (`alternative = "greather"`) del de la distribución de `y`.

10.3.2 Prueba de rangos con signo de Wilcoxon

La **prueba de rangos con signo de Wilcoxon** es, conceptualmente, parecida a la prueba de suma de rangos de Wilcoxon presentada en la sección anterior. Sin embargo, en este caso es la alternativa no paramétrica a la prueba t de Student con muestras pareadas. 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 las observaciones es intrínsecamente continua.
3. La escala de medición empleada para ambas muestras debe ser a lo menos ordinal.

Consideremos ahora un nuevo contexto para la aplicación de esta prueba. Una empresa de desarrollo desea evaluar la usabilidad de dos interfaces alternativas, A y B, para un nuevo producto de software, 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 un participante da a la interfaz evaluada corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. La tabla 10.4 muestra las evaluaciones realizadas por cada participante a cada una de las interfaces.

Formalmente, las hipótesis son:

H_0: las mismas personas no perciben diferencia en la usabilidad de ambas interfaces.
H_A: las mismas personas consideran que la interfaz A tiene mejor usabilidad que la interfaz B.

Participante	Interfaz A	Interfaz B
1	2,9	6,0
2	6,1	2,8
3	6,7	1,3
4	4,7	4,7
5	6,4	3,1
6	5,7	1,8
7	2,7	2,9
8	6,9	4,0
9	1,7	2,3
10	6,4	1,6

Tabla 10.4: evaluación de las interfaces de usuario A y 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, 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. La tabla 10.5 ilustra el proceso descrito.

Participante	Interfaz A	Interfaz B	A−B	|A−B|	Rango absoluto	Rango con signo
4	4,7	4,7	0,0	0	-	-
7	2,7	2,9	-0,2	0,2	1	-1
9	1,7	2,3	-0,6	0,6	2	-2
8	6,9	4,0	2,9	2,9	3	+3
1	2,9	6,0	-3,1	3,1	4	-4
2	6,1	2,8	3,3	3,3	5,5	+5,5
5	6,4	3,1	3,3	3,3	5,5	+5,5
6	5,7	1,8	3,9	3,9	7	+7
10	6,4	1,6	4,8	4,8	8	+8
3	6,7	1,3	5,4	5,4	9	+9

Tabla 10.5: asignación de rangos con signo.

Una vez realizado este proceso, se calcula el estadístico de prueba W, correspondiente a la suma de los rangos con signo. Debemos notar que, tras eliminar aquellas observaciones con diferencia igual a 0, el tamaño de las muestras para el ejemplo es $n = 9$. Así:

$$W = -1 + -2 + 3 + -4 + 5,5 + 5,5 + 7 + 8 + 9 = 31$$

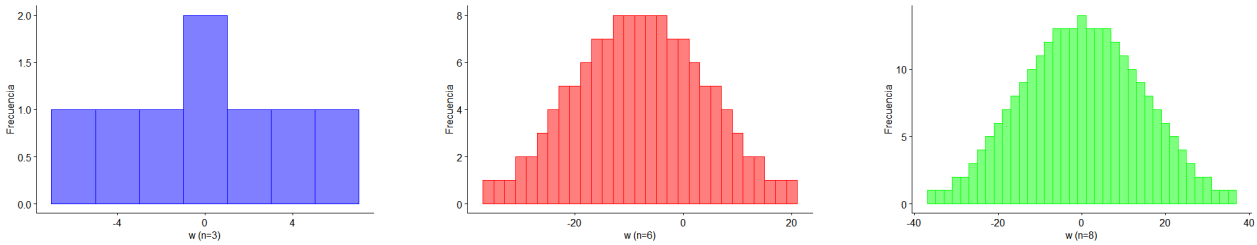
Desde luego, el máximo valor posible para W (W_{max}) corresponde a la suma de los n rangos cuando estos son todos positivos (ecuación 10.7), y el valor mínimo cuando todos son son negativos; así, $W_{min} = -W_{max}$.

Para entender mejor la distribución de W, una muestra de tamaño n genera n rangos no empatados sin signo (columna "Rango absoluto" de la tabla 10.5). A su vez, cada uno de dichos rangos podría tomar valores positivos o negativos, por lo que para W se tienen 2^n combinaciones para los signos de los rangos. La tabla 10.6 muestra todas las posibles combinaciones para $n = 3$. Si la hipótesis nula fuese cierta, los rangos positivos y negativos se distribuirían de manera homogénea, por lo que esperaríamos que el valor de W fuese cercano a 0 (hipótesis nula en el dominio de los rangos).

La figura 10.13 muestra la distribución de W para distintos valores de n. En ella podemos apreciar que, a medida que n aumenta, la distribución de W se aproxima cada vez más a una distribución normal centrada en $\mu_W = 0$.

Rango			
1	**2**	**3**	W
+	+	+	6
+	+	-	0
+	-	+	2
+	-	-	-4
-	+	+	4
-	+	-	-2
-	-	+	0
-	-	-	-6

Tabla 10.6: valores que puede adoptar el estadístico W para $n = 3$.



(a) $n = 3$ (b) $n = 6$ (c) $n = 8$

Figura 10.13: distribución de W.

La desviación estándar de la distribución muestral de W está dada por la ecuación 10.13.

$$\sigma_W = \sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{6}} \tag{10.13}$$

Para el ejemplo:

$$\sigma_W = \sqrt{\frac{9 \cdot (9+1) \cdot (2 \cdot 9+1)}{6}} = 16,882$$

Puesto que estamos trabajando bajo el supuesto de normalidad, calculamos el estadístico de prueba z, dado por la ecuación 10.14.

$$z = \frac{W - 0,5}{\sigma_W} \tag{10.14}$$

Así, para el ejemplo tenemos que:

$$z = \frac{31 - 0,5}{16,882} = 1,807$$

Una vez conocido el estadístico de prueba, podemos obtener el valor p mediante la llamada `pnorm(1.807, mean = 0, sd = 1, lower.tail = FALSE)` (no multiplicamos por 2 en este ejemplo, pues es una prueba unilateral), obteniendo como resultado $p = 0,035$. 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.

Siempre debemos tener en cuenta que el supuesto de normalidad es válido únicamente para $n > 10$, por lo que en caso de que las muestras sean más pequeñas, tenemos que consultar la tabla de valores críticos para la distribución W.

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 la llamada es `wilcox.test(x, y, paired = TRUE, alternative, conf.level)`, donde:

- `x, y`: vectores numéricos con las observaciones.
- `paired`: booleano con valor verdadero para indicar que las muestras son pareadas.
- `alternative`: señala el tipo de hipótesis alternativa: bilateral ("two.sided") o unilateral ("less" o "greater"). `mu`: valor nulo de la prueba.
- `conf.level`: nivel de confianza.

Así, el valor por defecto para el parámetro `paired` es `FALSE`, en cuyo caso se efectúa la prueba de suma de rangos de Wilcoxon; mientras que si explícitamente indicamos `paired = TRUE`, se aplica la prueba de rangos con signo de Wilcoxon.

El script 10.6 muestra la aplicación de la prueba de rangos con signo de Wilcoxon para el ejemplo, obteniéndose los resultados que se presentan en la figura 10.14. Es importante tener en cuenta en cuenta que R usa una variante ligeramente diferente. En lugar del estadístico de prueba W, calcula el estadístico V, correspondiente a la suma de los rangos con signo positivo.

```
Wilcoxon signed rank test with continuity correction

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

Figura 10.14: resultado de la prueba de rangos con signo de Wilcoxon para el ejemplo.

Script 10.6: prueba de rangos con signo de Wilcoxon para el ejemplo.

```
# Ingresar los datos.
a <- c(2.9, 6.1, 6.7, 4.7, 6.4, 5.7, 2.7, 6.9, 1.7, 6.4)
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 la prueba de rangos con signo de Wilcoxon.
prueba <- wilcox.test(a, b, alternative = "greater", paired = TRUE,
                      conf.level = 1 - alfa)

print(prueba)
```

10.4 PRUEBAS PARA MÁS DE DOS MUESTRAS

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

10.4.1 Prueba de Kruskal-Wallis

En el capítulo 8 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 k muestras son obtenidas de manera aleatoria e independiente desde la(s) población(es) de origen.
3. Se puede suponer razonablemente que la(s) población(es) de origen sigue(n) una distribución normal.
4. Las k muestras tienen varianzas aproximadamente iguales.

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

Para ilustrar esta prueba, tomemos el ejemplo de un ingeniero que cuenta con cuatro algoritmos (A, B, C y D) para resolver un determinado problema (en iguales condiciones y para instancias de tamaño fijo) y desea comparar su eficiencia. Para cada algoritmo, selecciona una muestra aleatoria independiente de instancias y registra el tiempo de ejecución (en milisegundos) del algoritmo en cuestión para cada una de las instancias de la muestra correspondiente, obteniendo las siguientes mediciones:

- Algoritmo A: 21, 22, 22, 23, 23, 23, 23, 24, 24, 24, 25, 26
- Algoritmo B: 15, 17, 18, 18, 19, 19, 20, 20, 21
- Algoritmo C: 9, 10, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 14, 15
- Algoritmo D: 15, 15, 16, 16, 16, 18, 18, 18

Las hipótesis a contrastar son, entonces:

H_0: todos los algoritmos son igual de eficientes (o, de manera similar, ningún algoritmo es menos ni más eficiente que los demás).
H_A: al menos uno de los algoritmos presenta una eficiencia diferente a al menos algún otro algoritmo.

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 10.7.

A continuación se calcula la suma (T_x) y la media (M_x) de los rangos en cada grupo y en la muestra combinada. La tabla 10.8 presenta los valores obtenidos para el ejemplo, incluyendo además el tamaño muestral (n_x).

De manera similar a ANOVA, se requiere determinar la diferencia entre las medias grupales. Para ello se calculan las desviaciones cuadradas de las medias grupales de los rangos con respecto a la media total de los rangos. Así, la variabilidad entre grupos está dada por la ecuación 10.15.

$$SS_{bg(R)} = \sum_{i=1}^{k} n_i \cdot (M_i - M_T)^2 \tag{10.15}$$

Para el ejemplo, entonces:

$$SS_{bg(R)} = n_A \cdot (M_A - M_T)^2 + n_B \cdot (M_B - M_T)^2 + n_C \cdot (M_C - M_T)^2 + n_D \cdot (M_D - M_T)^2 = \\ 12 \cdot (37,46 - 22)^2 + 9 \cdot (25,56 - 22)^2 + 14 \cdot (7,61 - 22)^2 + 8 \cdot (20 - 22)^2 = 5.913,21$$

Observaciones				Ranking de obs.			
A	**B**	**C**	**D**	**A**	**B**	**C**	**D**
21	15	9	15	31,5	15,5	1,0	15,5
22	17	10	15	33,5	21,0	3,5	15,5
22	18	10	16	33,5	24,0	3,5	19,0
23	18	10	16	36,5	24,0	3,5	19,0
23	19	10	16	36,5	27,5	3,5	19,0
23	19	11	18	36,5	27,5	8,5	24,0
23	20	11	18	36,5	29,5	8,5	24,0
24	20	12	18	40,0	29,5	8,5	24,0
24	21	12		40,0	31,5	8,5	
24		12		40,0		8,5	
25		12		42,0		8,5	
26		13		43,0		12,0	
		14				13,0	
		15				15,5	

Tabla 10.7: asignación de rangos a la muestra combinada.

	A	**B**	**C**	**D**	**Combinada**
n	12	9	14	8	43
T	449,50	230,00	106,50	160,00	946,00
M	37,46	25,56	7,61	20,00	22,00

Tabla 10.8: resumen de los rangos.

La hipótesis nula, llevada al dominio de los rangos, es que los rangos medios de los distintos grupos no serán muy diferentes entre sí. Podría esperarse que el valor nulo para $SS_{bg(R)}$ fuera 0, no obstante, no es así. Supongamos por un momento que tenemos 3 muestras con dos observaciones cada una, con lo que tendríamos un total de 6 rangos. Dichos rangos pueden combinarse de 90 maneras distintas para formar tres grupos con dos elementos. La distribución muestral de $SS_{bg(R)}$ estaría dada, entonces, por los valores de $SS_{bg(R)}$ obtenidos para cada una de las 90 combinaciones, de los cuales únicamente 6 son iguales a 0 y todos los restantes, mayores que 0 (recuerde que es matemáticamente imposible obtener desviaciones cuadradas con valor negativo). La media de la distribución muestral para $SS_{bg(R)}$ está dada por la ecuación 10.16.

$$\mu_{SS} = (k-1)\frac{n_T \cdot (n_T+1)}{12} \tag{10.16}$$

Para el ejemplo, entonces, tenemos que el valor nulo es:

$$\mu_{SS} = (4-1)\frac{43 \cdot (43+1)}{12} = 473$$

Llegado este punto, se define el estadístico de prueba H, el cual se construye en torno al valor obtenido para $SS_{bg(R)}$ y parte de la fórmula empleada para calcular el valor nulo, como muestra la ecuación 10.17.

$$H = \frac{SS_{bg(R)}}{\frac{n_T \cdot (n_T+1)}{12}} = \frac{12 \cdot SS_{bg(R)}}{n_T \cdot (n_T+1)} \tag{10.17}$$

En consecuencia, el valor del estadístico de prueba para el ejemplo es:

$$H = \frac{12 \cdot 5.913,21}{43 \cdot (43+1)} = 37.5$$

Cuando cada uno de los k grupos tiene a lo menos 5 observaciones, el estadístico de prueba H sigue una distribución χ^2 con $\nu = k - 1$ grados de libertad. Así, podemos calcular el valor p para el ejemplo (en R) mediante la llamada `pchisq(37.5, 3, lower.tail = FALSE)`, obteniéndose como resultado $p = 3.606 \cdot 10^{-8}$. Este valor indica que la evidencia es suficientemente fuerte como para rechazar la hipótesis nula en favor de la hipótesis alternativa, incluso para un nivel de significación $\alpha = 0,01$. En consecuencia, podemos concluir con 99 % de confianza que existen diferencias significativas entre los tiempos promedio de ejecución de los algoritmos A, B, C y D.

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 8, 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 métodos de corrección que ya conocimos en el capítulo 7 (por ejemplo, Holm y Bonferroni) (Amat Rodrigo, 2016b).

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 (factor)>.
- `data`: matriz de datos en formato largo.

Para los procedimientos post-hoc, las pruebas de Bonferroni y Holm pueden realizarse mediante la función `pairwise.wilcox.test(x, g, p.adjust.method, paired = FALSE)`, donde:

- `x`: vector con la variable dependiente.
- `g`: factor o agrupamiento.
- `p.adjust.method`: puede ser "holm" o "bonferroni", entre otras alternativas.
- `paired`: valor booleano que indica si la prueba es pareada (verdadero) o no. Para la prueba de Kruskal-Wallis debe ser `FALSE`.

El script 10.7 muestra la realización de la prueba de Kruskal-Wallis para el ejemplo e incorpora el procedimiento post-hoc de Holm. Los resultados se presentan en la figura 10.15. Podemos ver que el valor p difiere ligeramente al obtenido anteriormente, debido a errores de redondeo. A partir de los resultados del procedimiento post-hoc, considerando un nivel de significación $\alpha = 0,01$, podemos concluir con 99 % de confianza que existen diferencias significativas entre los tiempos promedio de ejecución de todos los pares de algoritmos con excepción de los algoritmos B y D.

En resumen, la elección entre Bonferroni y Holm como procedimiento post-hoc después de una prueba de Kruskal-Wallis depende de tus objetivos de control de errores y del contexto específico de tu estudio. Bonferroni es más conservador y rígido, mientras que Holm es más flexible y puede ser preferible en situaciones en las que se desee un equilibrio entre control de errores y poder estadístico.

```
Kruskal-Wallis rank sum test

data:  Tiempo by Algoritmo
Kruskal-Wallis chi-squared = 37.714, df = 3,
p-value = 3.249e-08


Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  datos$Tiempo and datos$Algoritmo

  A       B       C
B 0.00060 -       -
C 9.3e-05 0.00042 -
D 0.00060 0.02738 0.00060

P value adjustment method: holm
```

Figura 10.15: resultado de la prueba de Kruskal-Wallis y el procedimiento post-hoc de Holm para el ejemplo.

Script 10.7: prueba de Kruskal-Wallis y el procedimiento post-hoc de Holm para el ejemplo.

```
# Construir la mariz de datos.
A <- c(24, 23, 26, 21, 24, 24, 25, 22, 23, 22, 23, 23)
B <- c(17, 15, 18, 20, 19, 21, 20, 18, 19)
C <- c(10, 11, 14, 11, 15, 12, 12, 10, 9, 13, 12, 12, 10, 10)
D <- c(18, 16, 18, 15, 16, 15, 18, 16)
Tiempo <- c(A, B, C, D)

Algoritmo <- c(rep("A", length(A)),
               rep("B", length(B)),
               rep("C", length(C)),
               rep("D", length(D)))

Algoritmo <- factor(Algoritmo)

datos <- data.frame(Tiempo, Algoritmo)

# Establecer nivel de significación
alfa <- 0.01

# Hacer la prueba de Kruskal-Wallis.
prueba <- kruskal.test(Tiempo ~ Algoritmo, data = datos)
print(prueba)

# Efectuar procedimiento post-hoc de Holm si se encuentran diferencias
# significativas.
if(prueba$p.value < alfa) {
  post_hoc <- pairwise.wilcox.test(datos$Tiempo,
                                   datos$Algoritmo,
                                   p.adjust.method = "holm",
                                   paired = FALSE)

  print(post_hoc)
}
```

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

10.4.2 Prueba de Friedman

Es frecuente encontrar documentos que consideran a la **prueba de Friedman** como una alternativa no paramétrica al procedimiento ANOVA de una vía con muestras correlacionadas descrito en el capítulo 9. Sin embargo, debemos saber que no es exactamente una extensión de esta prueba, puesto que no considera las diferencias relativas entre sujetos (como lo hace ANOVA y la prueba de rangos con signo de Wilcoxon), y en consecuencia, como señala Baguley (2012), el poder estadístico es bastante menor.

Recordemos las condiciones que se deben verificar para poder aplicar la prueba ANOVA para muestras correlacionadas:

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 la(s) población(es) de origen sigue(n) una distribución normal.
4. La matriz de varianzas-covarianzas es esférica. Como explica Horn (2008, p. 1), esta condición establece que las varianzas entre los diferentes niveles de las medidas repetidas deben ser iguales.

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 las siguientes:

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

Como ejemplo para esta prueba, supongamos ahora que un equipo de desarrolladores desea establecer qué interfaz gráfica (A, B o C) resulta más atractiva para las usuarias y los usuarios de un nuevo sistema, por lo que han seleccionado una muestra aleatoria representativa de los distintos tipos de usuarios/as 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 tabla 10.9 muestra las puntuaciones totales asignadas por cada participante a las diferentes interfaces. En consecuencia, las hipótesis a contrastar son:

H_0: las interfaces tienen preferencias similares.
H_A: al menos una interfaz obtiene una preferencia distinta a las demás.

Usuario/a	**A**	**B**	**C**
1	21	6	13
2	10	21	25
3	7	18	18
4	21	7	20
5	24	24	24
6	27	13	8
7	17	13	29

Tabla 10.9: evaluación realizada por usuarios/as a cada una de las distintas interfaces.

El primer paso del proceso consiste en asignar rangos a las observaciones de cada sujeto. La interfaz con puntuación más baja recibe un rango de 1 y la más alta, un rango de 3 (generalizando, si se tienen k observaciones pareadas, se asignan rangos con valores de 1 a k). En caso de empate, se asigna el promedio de los rangos correspondientes. La tabla 10.10 muestra el resultado de este proceso.

	Originales			**Rangos**		
Usuario/a	**A**	**B**	**C**	**A**	**B**	**C**
1	21	6	13	3	1	2
2	10	21	25	1	2	3
3	7	18	18	1	2,5	2,5
4	21	7	20	3	1	2
5	24	24	24	2	2	2
6	27	13	8	3	2	1
7	17	13	29	2	1	3

Tabla 10.10: ranking de las interfaces por usuario/a.

La hipótesis nula para la prueba de Friedman es que los rangos promedio de cada interfaz son muy similares. Si denotamos el rango promedio de un grupo (interfaz) por M_x, para cada grupo esperamos, entonces, que se cumpla la igualdad de la ecuación 10.18, donde k es la cantidad de grupos.

$$M_x = \frac{k+1}{2} \tag{10.18}$$

A continuación se calculan algunas estadísticas de resumen, donde n corresponde al tamaño de cada muestra y M, a la media de los rangos (tabla 10.11).

	A	**B**	**C**	**Combinada**
n	7	7	7	21
M	2,14	1,64	2,21	2

Tabla 10.11: resumen de los rangos.

Con estos valores, podemos definir una medida para la variabilidad de los grupos agregados, dada por la ecuación 10.19.

$$SS_{bg(R)} = \sum_{i=1}^{k} n_i \cdot (M_i - M_T)^2 \tag{10.19}$$

Haciendo el cálculo para el ejemplo, tenemos:

$$SS_{bg(R)} = 7 \cdot [(2,14-2)^2 + (1,64-2)^2 + (2,21-2)^2] = 1,357$$

Con el resultado anterior, podemos ahora calcular el estadístico de prueba (10.20), que sigue una distribución χ^2 con $k-1$ grados de libertad.

$$\chi^2 = \frac{\frac{SS_{bg(R)}}{k \cdot (k+1)}}{12} = \frac{12 \cdot SS_{bg(R)}}{k \cdot (k+1)} \tag{10.20}$$

Para el ejemplo:

$$\chi^2 = \frac{12 \cdot 1,357}{3 \cdot (3+1)} = 1,357$$

Una vez más, calculamos el valor p mediante la llamada `pchisq(1.357, 2, lower.tail = FALSE)`, obteniéndose $p = 0,507$. Considerando un nivel de significación $\alpha = 0,05$, se falla al rechazar la hipótesis nula. En consecuencia, concluimos con 95 % de confianza que no hay diferencias significativas de preferencia entre las distintas interfaces.

En este caso no es necesario realizar un procedimiento post-hoc, pues la prueba ómnibus no encontró diferencias estadísticamente significativas. No obstante, si fuese necesario, podemos efectuar una prueba de rangos con signo de Wilcoxon por cada par de grupos y aplicar algún factor de corrección.

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

- `formula`: tiene la forma <variable dependiente>∼ <variable independiente>| <identificador de sujeto o bloque>.
- `data`: matriz de datos en formato largo.

Para los procedimientos post-hoc, podemos aplicar los ajustes de Bonferroni y 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 10.8 muestra la realización de la prueba de Friedman para el ejemplo, cuyo resultado se presenta en la figura 10.16, e incorpora el procedimiento post-hoc de Holm por fines académicos, ya que solo debería realizase si la prueba ómnibus encuentra diferencias significativas.

Script 10.8: prueba de Friedman y el procedimiento post-hoc de Holm para el ejemplo.

```
# Construir la mariz de datos.
A <- c(21, 10, 7, 21, 24, 27, 17)
B <- c(6, 21, 18, 7, 24, 13, 13)
C <- c(13, 25, 18, 20, 24, 8, 29)

Puntuacion <- c(A, B, C)

Interfaz <- c(rep("A", length(A)),
              rep("B", length(B)),
              rep("C", length(C)))

Sujeto <- rep(1:7, 3)

Interfaz <- factor(Interfaz)

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

# Establecer nivel de significación
alfa <- 0.05

# Hacer la prueba de Friedman.
prueba <- friedman.test(Puntuacion ~ Interfaz | Sujeto, data = datos)
print(prueba)

# Efectuar 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)

  print(post_hoc)
}
```

```
        Friedman rank sum test

data:  Puntuacion and Interfaz and Sujeto
Friedman chi-squared = 1.6522, df = 2, p-value = 0.4378
```

Figura 10.16: valores p obtenidos en las pruebas t para cada par de grupos mediante los métodos de Bonferroni y Holm.

Por último, debemos saber que va tomando fuerza la idea de no 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 incluso para muestras independientes y el análisis de dos muestras.

10.5 EJERCICIOS PROPUESTOS

1. ¿Para qué se usa la transformación logarítmica?
2. Explica por qué comparar medias aritméticas de datos en escala logarítmica compara las medias geométricas en la escala normal de la variable.
3. El paquete `rcompanion` proporciona una función para aplicar la escala de potencias de Tukey. Experimenta con su uso.
4. Explica la relación entre la escala de potencias de Tukey y la transformación Box-Cox.
5. El paquete `DescTools` proporciona funciones para aplicar la transformación Box-Cox. Experimenta con su uso.
6. Menciona tres situaciones que complican las pruebas de hipótesis paramétricas tradicionales.
7. ¿Qué alternativas se mencionan para tratar datos problemáticos? En particular, ¿qué recomienda el capítulo para muestras pequeñas que muestran desviaciones de normalidad?
8. ¿Qué riesgos se corren si se aplica la prueba t de Student con dos muestras que no cumplen con las suposiciones que hace esta prueba?
9. La prueba de Wilcoxon-Mann-Whitney es una alternativa no paramétrica ¿para qué versión de la prueba t de Student?
10. ¿Qué suposiciones hace la prueba de Wilcoxon-Mann-Whitney?
11. Explica cómo la prueba de Wilcoxon-Mann-Whitney construye el ranking de los datos.
12. ¿Qué estadístico usa la prueba de Wilcoxon-Mann-Whitney y cómo se calcula?
13. ¿Por qué a la prueba de Wilcoxon-Mann-Whitney también se le conoce como U-test?
14. La prueba de los rangos con signo de Wilcoxon es una alternativa no paramétrica ¿para qué versión de la prueba t de Student?
15. ¿Qué suposiciones hace la prueba de los rangos con signo de Wilcoxon?
16. Explica cómo la prueba de los rangos con signo de Wilcoxon construye el ranking de los datos.
17. ¿Qué estadístico usa la prueba de los rangos con signo de Wilcoxon y cómo se calcula?
18. ¿Cuándo es más relevante preocuparse de las violaciones de las condiciones del procedimiento ANOVA para muestras independientes?
19. Explica cómo la prueba de Kruskal-Wallis construye el ranking de los datos.
20. ¿Qué estadístico usa la prueba de Kruskal-Wallis y cómo se calcula? ¿Qué distribución sigue dicho estadístico?
21. ¿Cuál es la hipótesis nula de la prueba de Kruskal-Wallis?
22. ¿Qué suposiciones hace la prueba de Kruskal-Wallis?
23. Explique cómo la prueba de Friedman construye el ranking de los datos.
24. ¿Qué estadístico usa la prueba de Friedman y cómo se calcula? ¿Qué distribución sigue dicho estadístico?
25. ¿Cuál es la hipótesis nula de la prueba de Friedman?
26. ¿Qué suposiciones hace la prueba de Friedman?

REFERENCIAS

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/

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/

Horn, R. A. (2008). *Sphericity in repeated measures analysis*. Consultado el 11 de mayo de 2021, desde http://oak.ucc.nau.edu/rh232/courses/EPS625/Handouts/RM-ANOVA/Sphericity.pdf

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/

Real Statistics Using Excel. (s.f.). *Mann-Whitney Table*. Consultado el 28 de mayo de 2021, desde https://www.real-statistics.com/statistics-tables/mann-whitney-table/

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.