ESTADÍSTICA PARA INGENIERÍA Y CIENCIAS

PRÁCTICA 5: Estadística no paramétrica

Ivan Svetlich

```
#Librerias
library(IRdisplay)
library(formattable)
library(ggplot2)
library(cowplot)
library(dplyr)
library(stringr)
```

Ejercicio 1

Los siguientes datos representan el tiempo, en minutos, que un paciente tiene que esperar durante 12 visitas al consultorio de una doctora antes de ser atendido por ésta:

Tiempo [min]	17	15	20	20	32	28	12	26	25	25	35	24

Utilice la prueba del signo al nivel de significancia de 0.05 para probar la afirmación de la doctora, de que la media del tiempo de espera para sus pacientes no es mayor que 20 minutos antes de entrar al consultorio.

```
data_1 <- read.csv("./TP5_tables/data1.csv") # Leo los datos desde archivo .csv
```

Las hipótesis son:

$$H_0: \mu = 20 \text{ y } H_1: \mu > 20$$

```
# Plot
options(repr.plot.width=10, repr.plot.height=4)
ggplot(data_1, aes(x=wait_time)) +
    labs(
        title="Diagrama de caja del tiempo de espera",
        x="Tiempo de espera [min]") +
    geom_boxplot(alpha=0.5, fill="#baddf7") +
    theme_light() +
    theme(text=element_text(size=20),
         plot.title = element_text(size=24, hjust = 0.5),
         axis.text.y = element_blank())
```


```
mu0 <- 20
wait_time <- data_1$wait_time
wait_time.filtered <- wait_time[which(data_1$wait_time != mu0)] # filtro los valores iguales as la media
```

```
n <- length(wait_time.filtered)
```

```r
magnitud_normalizada <- wait_time.filtered - mu0
x <- sum(magnitud_normalizada > 0)
p <- 0.5
X_binom <- dbinom(x, n, p)
```

```r
display_markdown(sprintf('Para $n = %d$ y $p = %.2f$, la probabilidad de $X \\geq %d$ es %.4f',
                         n, p, x, X_binom))
```

Para $n = 10$ y $p = 0.50$, la probabilidad de $X \geq 7$ es 0.1172

Como 0.1172 es mayor que el nivel de significancia $\alpha = 0.05$, la hipótesis nula no es rechazada y se concluye que la media del tiempo de espera no es mayor a 20 minutos.

Ejercicio 2

Analice los datos del ejercicio anterior usando la prueba de rango con signo.

```r
df1 <- data.frame(data_1)
mu0 <- 20
df1 <- df1 %>% filter(wait_time != mu0) %>% arrange(abs(wait_time - mu0))
```

```r
magnitud_normalizada <- df1$wait_time - mu0
magnitud_absoluta <- sort(abs(df1$wait_time - mu0))
rango <- rank(abs(magnitud_normalizada))
rango_con_signo <- rango*ifelse(magnitud_normalizada < 0, -1, 1)
```

```r
display_markdown('#### **Rangos con signo**')
display_markdown('\n')
rangos_con_signo <- data.frame(magnitud_normalizada, magnitud_absoluta, rango, rango_con_signo)
colnames(rangos_con_signo) <- c('Valor normalizado', 'Magnitud absoluta', 'Rango', 'Rango con signo')
table <- formattable(rangos_con_signo, align='c')
as.htmlwidget(table, width="70%", height=NULL)
```

Rangos con signo

Valor normalizado	Magnitud absoluta	Rango	Rango con signo
-3	3	1.0	-1.0
4	4	2.0	2.0
-5	5	4.0	-4.0
5	5	4.0	4.0
5	5	4.0	4.0
6	6	6.0	6.0
8	8	7.5	7.5
-8	8	7.5	-7.5
12	12	9.0	9.0
15	15	10.0	10.0

```r
s_plus <- sum(rango_con_signo[which(rango_con_signo > 0)]) # Sumo los rangos positivos
display_markdown(sprintf('Valor de estadístico de prueba: $$s_{+} = %.2f$$', s_plus))
```

Valor de estadístico de prueba:

$$s_+ = 42.50$$

```r
alpha <- 0.05
c1 <- qsignrank(alpha, length(data_1$wait_time), lower=FALSE) # calculo la región de rechazo para alpha=0.05
display_markdown(sprintf('Región de rechazo para prueba de nivel $\\alpha = %.2f$: $$s_{+} \\geq c_{1} = %d$$', alpha, c
1))
```

Región de rechazo para prueba de nivel $\alpha = 0.05$:

$$s_+ \geq c_1 = 60$$

Como el valor del estadístico de prueba se encuentra fuera de la zona de rechazo, no es posible rechazar la hipótesis nula con un nivel de significancia 0.05 y se concluye que la media del tiempo de espera no es mayor a 20 minutos. Este resultado coincide con el obtenido en el inciso anterior, utilizando la prueba del signo.

Utilizando la funcion ***wilcox.test***:

```
wilcox.test(data_1$wait_time,
            alternative = "greater",
            exact = FALSE,
            mu = mu0)
```

```
        Wilcoxon signed rank test with continuity correction

data:  data_1$wait_time
V = 42.5, p-value = 0.06906
alternative hypothesis: true location is greater than 20
```

Como el p-valor de la prueba es mayor a 0.05, no es posible rechazar la hipótesis nula.

Ejercicio 3

Los pesos de 5 personas antes de que dejen de fumar y cinco semanas después de dejar de fumar, en kg, son los siguientes:

	Individuo				
	1	2	3	4	5
Antes	66	80	69	52	75
Después	71	82	68	56	73

Utilice la prueba de rango con signo para observaciones pareadas para probar la hipótesis, en el nivel de significancia de 0.05, de que dejar de fumar no tiene efecto en el peso de una persona, contra la alternativa de que el peso aumenta si se deja de fumar.

```
data_3 <- read.csv("./TP5_tables/data3.csv") # Leo los datos desde archivo .csv
```

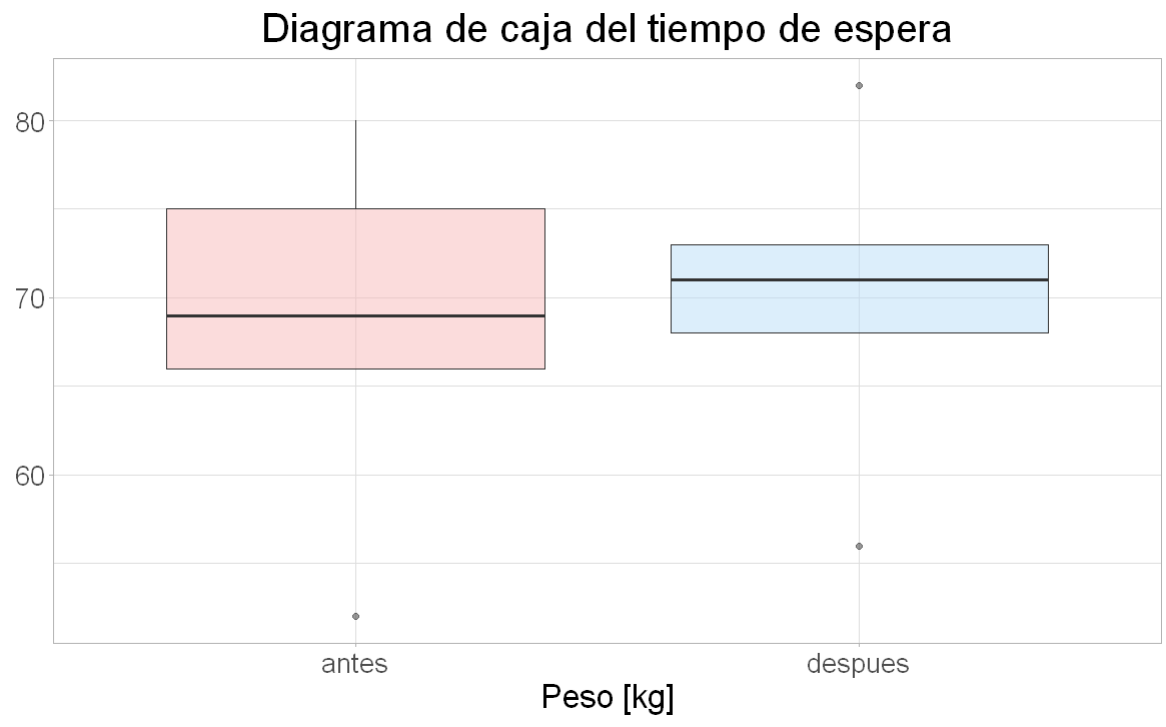
Las hipótesis son:

$$H_0\colon \mu_D = 0 \text{ y } H_1\colon \mu_D > 0$$

siendo μ_D la media de la diferencia de peso:

$$D = X_{\text{después}} - X_{\text{antes}}$$

```
# Plot
options(repr.plot.width=10, repr.plot.height=6)
data_3 %>% melt(id.var=NULL,
  variable.name = "antes_despues",
  value.name = "peso") %>%
ggplot(aes(y=peso, x=antes_despues)) +
    labs(
        title="Diagrama de caja del tiempo de espera",
        x="Peso [kg]",
        y="") +
    geom_boxplot(alpha=0.5, fill=c("#f7baba","#baddf7")) +
    theme_light() +
    theme(text=element_text(size=20),
         plot.title = element_text(size=24, hjust = 0.5))
```




```
df3 <- data.frame(1:nrow(data_3), data_3$antes, data_3$despues, data_3$despues - data_3$antes)
colnames(df3) <- c("individuo", "antes", "despues", "diferencia")
df3 <- df3 %>% arrange(abs(diferencia))
rango <- rank(abs(df3$diferencia))
rango_con_signo <- rango*ifelse(df3$diferencia < 0, -1, 1)
df3 <- data.frame(df3, rango_con_signo)
```

```
display_markdown('#### **Rangos con signo**')
display_markdown('\n')
table <- formattable(df3, align='c')
as.htmlwidget(table, width="50%", height=NULL)
```

Rangos con signo

individuo	antes	despues	diferencia	rango_con_signo
3	69	68	-1	-1.0
2	80	82	2	2.5
5	75	73	-2	-2.5
4	52	56	4	4.0
1	66	71	5	5.0

```
s_plus <- sum(rango_con_signo[which(rango_con_signo > 0)]) # Sumo los rangos positivos
display_markdown(sprintf('Valor de estadístico de prueba: $$s_{+} = %.2f$$', s_plus))
```

Valor de estadístico de prueba:

$$s_+ = 11.50$$

```
alpha <- 0.05
c1 <- qsignrank(alpha, length(df3$diferencia), lower=FALSE) # calculo la región de rechazo para alpha=0.05
display_markdown(sprintf('Región de rechazo para prueba de nivel $\\alpha = %.2f$: $$s_{+} \\geq c_{1} = %d$$', alpha, c
1))
```

Región de rechazo para prueba de nivel $\alpha = 0.05$:

$$s_+ \geq c_1 = 14$$

Como el valor del estadístico de prueba se encuentra fuera de la zona de rechazo, no es posible rechazar la hipótesis nula con un nivel de significancia 0.05 y se concluye que dejar de fumar no tiene efecto en el peso de la persona.

Utilizando la funcion ***wilcox.test***:

```
wilcox.test(x = data_3$antes, y = data_3$despues,
            alternative = "greater")
```

```
        Wilcoxon rank sum exact test

data:  data_3$antes and data_3$despues
W = 11, p-value = 0.6548
alternative hypothesis: true location shift is greater than 0
```

Como el p-valor de la prueba es mayor a 0.05, no es posible rechazar la hipótesis nula.

Ejercicio 4

Los siguientes son los números de recetas surtidas por dos farmacias A y B en un período de 20 días:

día	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
A	19	21	15	17	24	12	19	14	20	18	23	21	17	12	16	15	20	18	14	22
B	17	15	12	12	16	15	11	13	14	21	19	15	11	10	20	12	13	17	16	18

Utilice la prueba de rango con signo al nivel de significancia de 0.01 para determinar si las dos farmacias, en promedio, surten el mismo número de recetas, contra la alternativa de que la farmacia A surte más recetas que la farmacia B.

```
data_4 <- read.csv("./TP5_tables/data4.csv") # Leo los datos desde archivo .csv
```

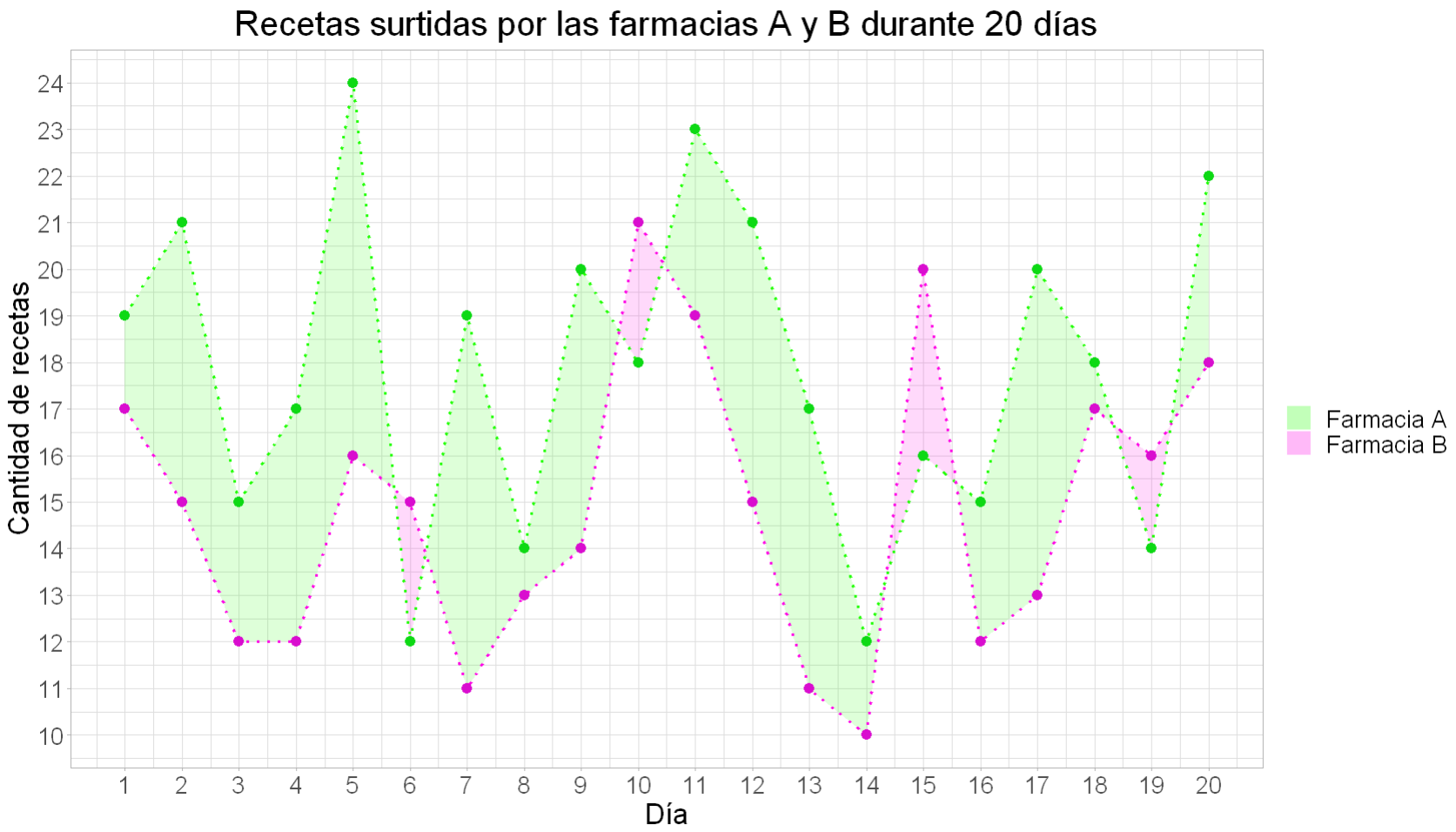
Las hipótesis son:

$$H_0: \mu_D = 0 \text{ y } H_1: \mu_D > 0$$

siendo μ_D la media de la diferencia entre el número de recetas surtidas por la farmacia A y la farmacia B:

$$D = X_A - X_B$$

```
A_approx <- approx(data_4$A, n = 500)
B_approx <- approx(data_4$B, n = 500)
options(repr.plot.width=14, repr.plot.height=8)
data_4.interp <- data_frame(dia=c(A_approx$x, data_4$dia), A=c(A_approx$y, data_4$A), B=c(B_approx$y, data_4$B))
data_4.interp %>%
ggplot(aes(x=dia)) +
    labs(
        title="Recetas surtidas por las farmacias A y B durante 20 días",
        x="Día",
        y="Cantidad de recetas",
        fill="",
        size="",
        alpha="",
        color="") +
    geom_ribbon(aes(ymin=B, ymax = pmin(B, A), fill = "Farmacia B"), alpha=0.15) +
    geom_ribbon(aes(ymin=A, ymax = pmin(A, B), fill = "Farmacia A"), alpha=0.15) +
    geom_line(alpha=1, aes(y=A), col="#20ff03", linetype = "dotted", size=1) +
    geom_line(alpha=1, aes(y=B), col="#ff03e6", linetype = "dotted", size=1) +
    geom_point(aes(y=A, size=ifelse(A%in%data_4$A & dia%in%data_4$dia, "2","0"), alpha=ifelse(A%in%data_4$A & dia%in%dat
a_4$dia, "2","0")), col="#0dd914", show.legend = F) +
    geom_point(aes(y=B, size=ifelse(B%in%data_4$B & dia%in%data_4$dia, "2","0"), alpha=ifelse(B%in%data_4$B & dia%in%dat
a_4$dia, "2","0")), col="#d90dcf", show.legend = F) +
    scale_fill_manual(values = c("#20ff03", "#ff03e6")) +
    scale_alpha_manual(values =c(0,1)) +
    scale_size_manual(values =c(0,3)) +
    scale_x_continuous(breaks = 1:20) +
    scale_y_continuous(breaks = min(min(round(A_approx$y)), min(round(B_approx$y))) : max(max(round(A_approx$y)), max(ro
und(B_approx$y)))) +
    theme_light() +
    theme(text=element_text(size=20),
         plot.title = element_text(size=24, hjust = 0.5))
```




```
df4 <- data.frame(data_4, diferencia = data_4$A - data_4$B)
colnames(df4)[2:3] <- c("farmacia_A", "farmacia_B")
df4 <- df4 %>% arrange(abs(diferencia))
rango <- rank(abs(df4$diferencia))
rango_con_signo <- rango*ifelse(df4$diferencia < 0, -1, 1)
df4 <- data.frame(df4, rango_con_signo)
```

```
display_markdown('#### **Rangos con signo para muestras pareadas**')
display_markdown('\n')
table <- formattable(df4, align='c')
as.htmlwidget(table, width="60%", height=NULL)
```

Rangos con signo para muestras pareadas

dia	farmacia_A	farmacia_B	diferencia	rango_con_signo
8	14	13	1	1.5
18	18	17	1	1.5
1	19	17	2	4.0
14	12	10	2	4.0
19	14	16	-2	-4.0
3	15	12	3	7.5
6	12	15	-3	-7.5
10	18	21	-3	-7.5
16	15	12	3	7.5
11	23	19	4	11.0
15	16	20	-4	-11.0
20	22	18	4	11.0
4	17	12	5	13.0
2	21	15	6	15.5
9	20	14	6	15.5
12	21	15	6	15.5
13	17	11	6	15.5
17	20	13	7	18.0
5	24	16	8	19.5
7	19	11	8	19.5

```
s_plus <- sum(rango_con_signo[which(rango_con_signo > 0)]) # Sumo los rangos positivos
display_markdown(sprintf('Valor de estadístico de prueba: $$s_{+} = %.2f$$', s_plus))
```

Valor de estadístico de prueba:

$$s_+ = 180.00$$

```
alpha <- 0.01
```

```r
c1 <- qsignrank(alpha, length(df4$diferencia), lower=FALSE) # calculo la región de rechazo para alpha=0.05
display_markdown(sprintf('Región de rechazo para prueba de nivel $\\alpha = %.2f$: $$s_{+} \\geq c_{1} = %d$$', alpha, c
1))
```

Región de rechazo para prueba de nivel $\alpha = 0.01$:

$$s_+ \geq c_1 = 166$$

Como el valor del estadístico de prueba se encuentra dentro de la zona de rechazo, se rechaza la hipótesis nula con un nivel de significancia 0.01 y se concluye que la farmacia A surte más recetas que la farmacia B.

Utilizando la funcion ***wilcox.test***:

```r
diferencia <- data_4$A - data_4$B
wilcox.test(diferencia, alternative = "greater", exact = FALSE)
```

```
        Wilcoxon signed rank test with continuity correction

data:  diferencia
V = 180, p-value = 0.002647
alternative hypothesis: true location is greater than 0
```

Como el p-valor de la prueba es menor a 0.01, se rechaza la hipótesis nula.

Ejercicio 5

Un fabricante de cigarrillos afirma que el contenido de alquitrán de la marca de cigarrillos B es menor que la de la marca A. Para probar esta afirmación, se registran las siguientes determinaciones de contenido de alquitrán, en miligramos:

	Concentración de alquitrán (mg)					
Marca A	1	12	9	13	11	14
Marca B	8	10	7	-	-	-

Utilice la prueba de suma de rangos con $\alpha = 0.05$ para probar si tal afirmación es válida.

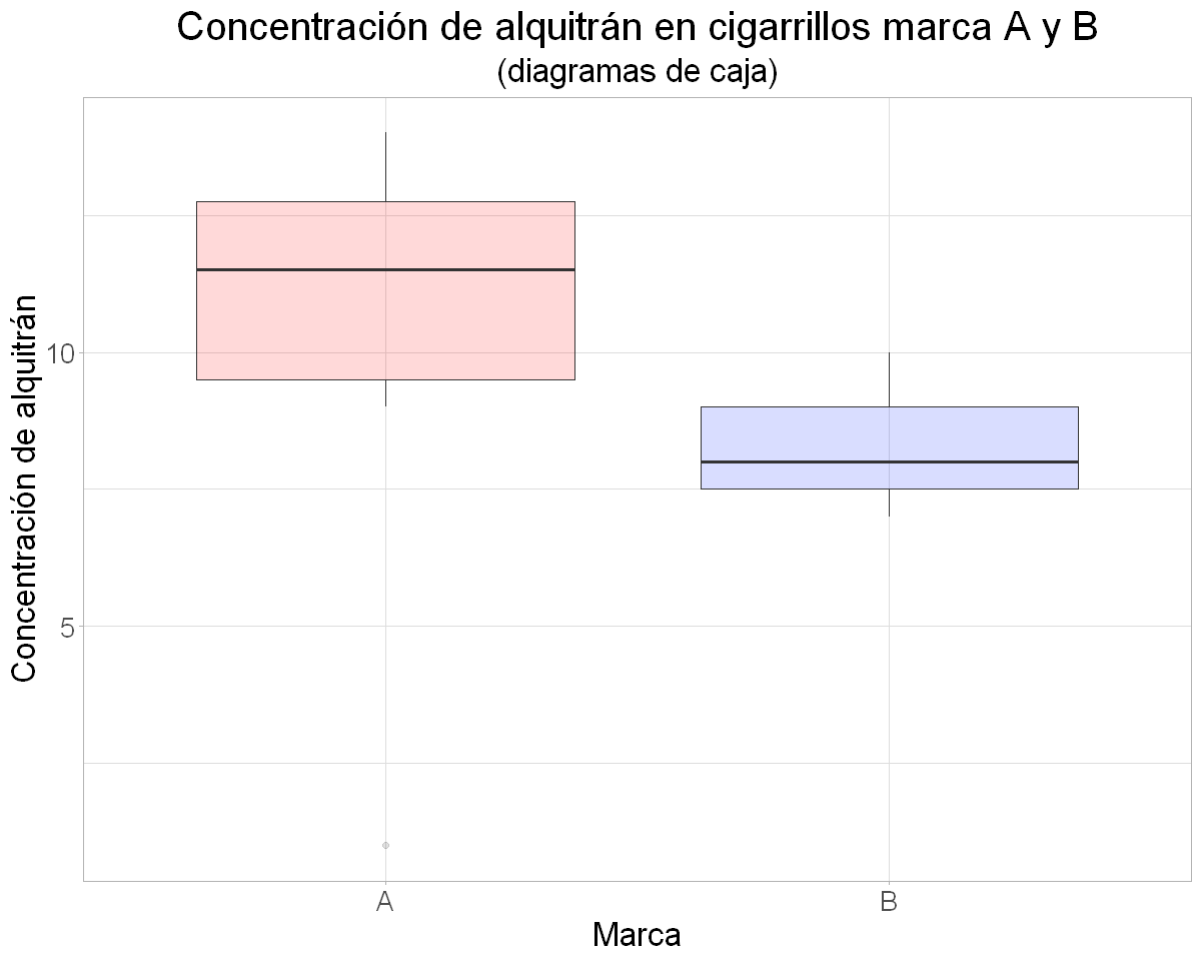
```r
data_5 <- read.csv("./TP5_tables/data5.csv") # Leo los datos desde archivo .csv
data_5$marca <- factor(data_5$marca)
```

Las hipótesis son:

$$H_0\colon \mu_B - \mu_A = 0 \text{ y } H_1\colon \mu_B - \mu_A < 0$$

donde μ_A es la media verdadera de alquitrán en la marca A y μ_B es la media verdadera de alquitrán en la marca B.

```r
# Plot
options(repr.plot.width=10, repr.plot.height=8)
data_5 %>%
ggplot(aes(y=concentracion_alquitran, x=marca)) +
    labs(
        title="Concentración de alquitrán en cigarrillos marca A y B",
        subtitle="(diagramas de caja)",
        x="Marca",
        y="Concentración de alquitrán") +
    geom_boxplot(alpha=0.15, fill=c("#ff0303", "#031cff")) +
    theme_light() +
    theme(text=element_text(size=20),
         plot.title = element_text(size=24, hjust = 0.5),
         plot.subtitle = element_text(size=20, hjust = 0.5))
```




```
concentracion.A <- data_5[data_5$marca == 'A',]$concentracion_alquitran
concentracion.B <- data_5[data_5$marca == 'B',]$concentracion_alquitran

wilcox.test(concentracion.A,concentracion.B,alternative="greater",paired=FALSE)
```

```
	Wilcoxon rank sum exact test

data:  concentracion.A and concentracion.B
W = 14, p-value = 0.131
alternative hypothesis: true location shift is greater than 0
```

Como el p-valor de la prueba es mayor a 0.05, no hay evidencia suficiente para afirmar que la concentración de alquitrán es mayor en los cigarrillos de la marca A que en los de la marca B.

Ejercicio 6

Los siguientes datos representan el número de horas que operan dos diferentes tipos de calculadoras científicas de bolsillo, antes de que necesiten recargarse:

	Duración de la bateria (hs)								
Calculadora A	5.5	5.6	6.3	4.6	5.3	5.0	6.2	5.8	5.1
Calculadora B	3.8	4.8	4.3	4.2	4.0	4.9	4.5	5.2	4.5

Utilice la prueba de la suma de rangos con $\alpha = 0.01$ para determinar si la calculadora A opera más tiempo que la calculadora B con una carga completa de la batería.

```
data_6 <- read.csv("./TP5_tables/data6.csv") # Leo los datos desde archivo .csv
data_6$calculadora <- factor(data_6$calculadora)
```

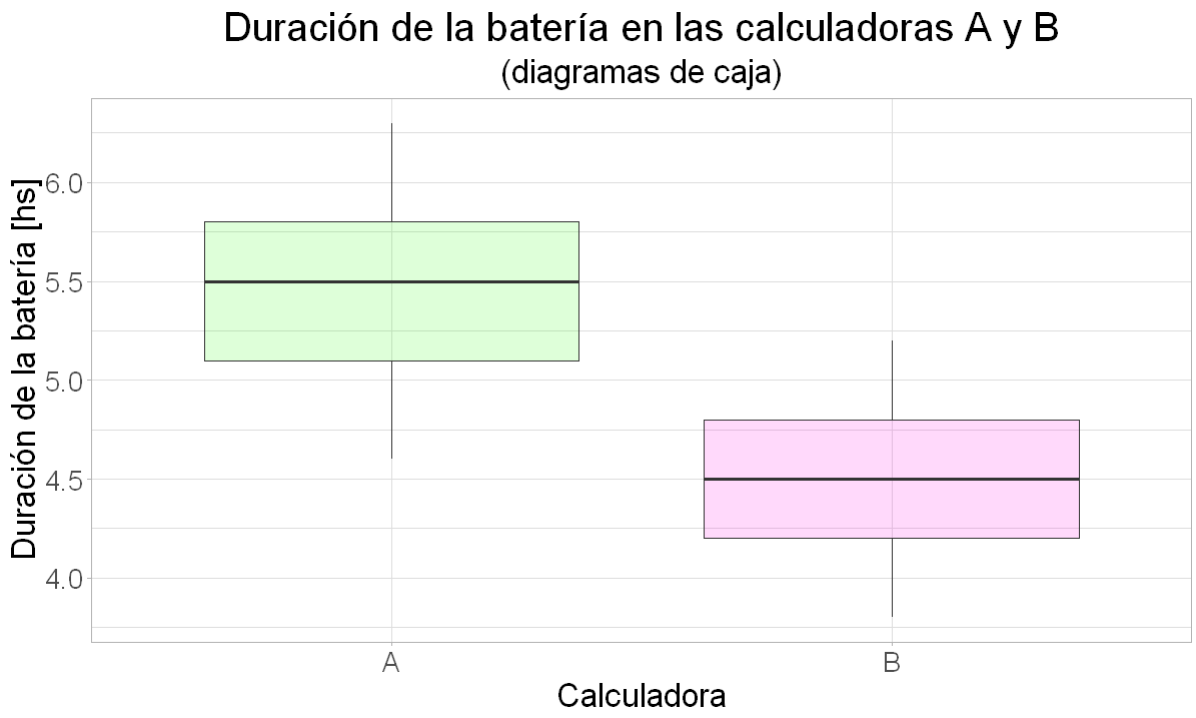
Las hipótesis son:

$$H_0\colon \mu_B - \mu_A = 0 \text{ y } H_1\colon \mu_B - \mu_A < 0$$

donde μ_A y μ_B son las medias verdaderas de duración de la batería en la calculadora A es la media verdadera de duración de la batería en la calculadora B.

```
# Plot
options(repr.plot.width=10, repr.plot.height=6)
data_6 %>%
ggplot(aes(y=duracion_bateria, x=calculadora)) +
    labs(
```

```
        title="Duración de la batería en las calculadoras A y B",
        subtitle="(diagramas de caja)",
        x="Calculadora",
        y="Duración de la batería [hs]") +
    geom_boxplot(alpha=0.15, fill=c("#20ff03", "#ff03e6")) +
    theme_light() +
    theme(text=element_text(size=20),
         plot.title = element_text(size=24, hjust = 0.5),
         plot.subtitle = element_text(size=20, hjust = 0.5))
```




```
calculadora_A <- data_6[data_6$calculadora == 'A',]$duracion_bateria
calculadora_B <- data_6[data_6$calculadora == 'B',]$duracion_bateria

n1<-length(calculadora_A)
n2<-length(calculadora_B)

wilcox.test(calculadora_A,calculadora_B,alternative="greater",paired=FALSE, exact=FALSE)
```

```
        Wilcoxon rank sum test with continuity correction

data:  calculadora_A and calculadora_B
W = 76, p-value = 0.0009935
alternative hypothesis: true location shift is greater than 0
```

Como el valor p-valor de la prueba es menor a 0.01, se rechaza la hipótesis nula y se concluye que la duración de la bateria de la calculadora A es mayor a la de la calculadora B.

Ejercicio 7

La información siguiente se refiere a la concentración del isótopo radiactivo estroncio-90, obtenida en muestras de leche de cinco lecherías seleccionadas al azar en cuatro regiones diferentes.

Región	Concentración del isótopo				
	1	2	3	4	5
1	6.4	5.8	6.5	7.7	6.1
2	7.1	9.9	11.2	10.5	8.8
3	5.7	5.9	8.2	6.6	5.1
4	9.5	12.1	10.3	12.4	11.7

Pruebe al nivel 0.10 para ver si el promedio verdadero de la concentración de estroncio-90 difiere en al menos dos de las regiones.

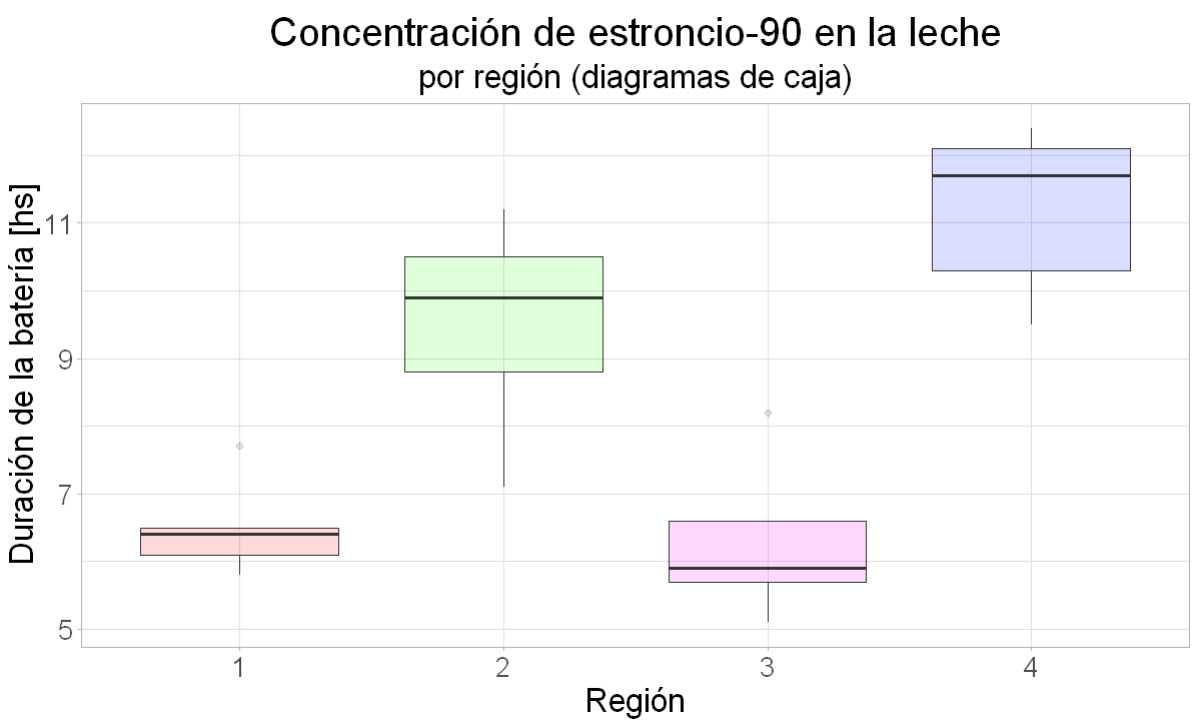
```
data_7 <- read.csv("./TP5_tables/data7.csv") # Leo los datos desde archivo .csv
data_7$region <- factor(data_7$region)
```

Las hipótesis son:

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ y } H_1\text{: al menos un par de } \mu_i \text{ difiere entre sí}$$

donde μ_i es la media verdadera de concentración del isótopo en la región i. $i = 1, 2, 3, 4$

```
# Plot
options(repr.plot.width=10, repr.plot.height=6)
data_7 %>%
ggplot(aes(y=concentracion_isotopo, x=region)) +
    labs(
        title="Concentración de estroncio-90 en la leche",
        subtitle="por región (diagramas de caja)",
        x="Región",
        y="Duración de la batería [hs]") +
    geom_boxplot(alpha=0.15, fill=c("#ff0303", "#20ff03", "#ff03e6", "#031cff")) +
    theme_light() +
    theme(text=element_text(size=20),
         plot.title = element_text(size=24, hjust = 0.5),
         plot.subtitle = element_text(size=20, hjust = 0.5))
```




```
kruskal.test(concentracion_isotopo ~ region, data_7)
```

```
        Kruskal-Wallis rank sum test

data:  concentracion_isotopo by region
Kruskal-Wallis chi-squared = 14.063, df = 3, p-value = 0.002821
```

Como el valor p-valor de la prueba es menor a 0.01, se rechaza la hipótesis nula y se concluye que la concentración del isótopo estroncio-90 difiere en al menos dos de las regiones estudiadas.

Ejercicio 8

Se presentan los resultados de un experimento para comparar cuatro técnicas de mezclado diferentes sobre la resistencia a la tensión del cemento portland. ¿Existe algún indicador de que las técnicas de mezclado afectan la resistencia?. Utilice $\alpha = 0.05$.

Método	Resistencia (lb/in^2)			
1	3129	3000	2865	2890
2	3200	3000	2975	3150
3	2800	2900	2985	3050
4	2600	2700	2600	2765

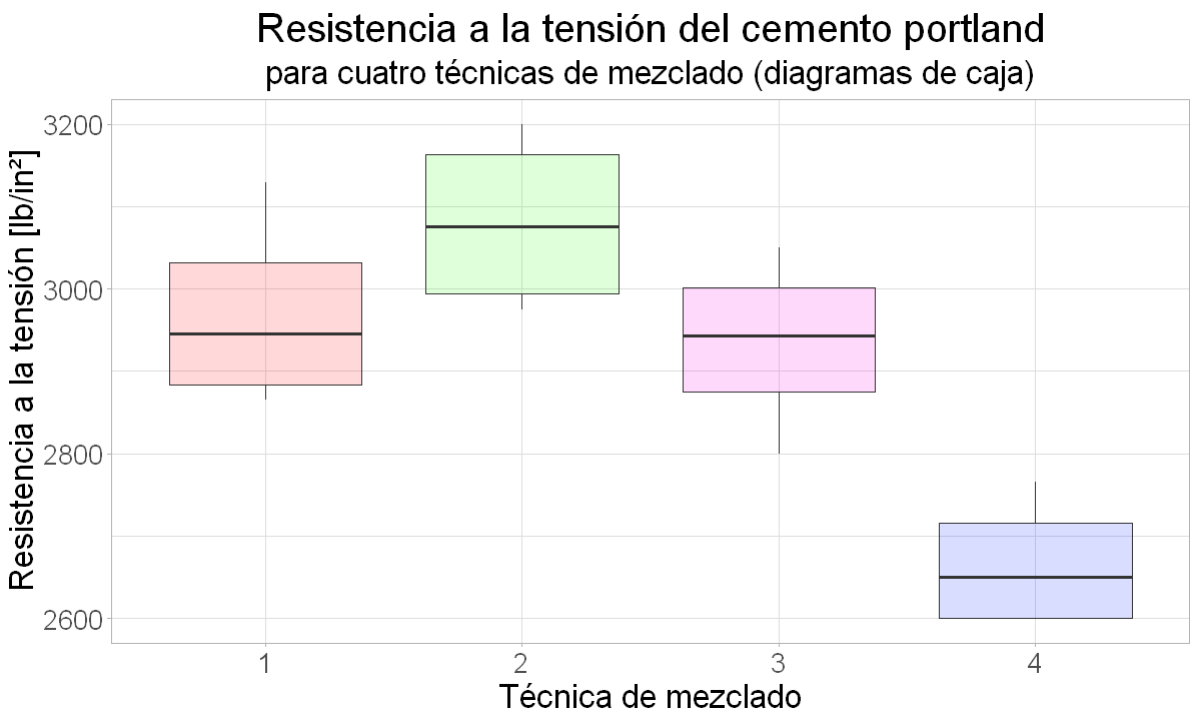
```
data_8 <- read.csv("./TP5_tables/data8.csv") # Leo los datos desde archivo .csv
data_8$metodo <- factor(data_8$metodo)
```

Las hipótesis son:

$$H_0\text{: } \mu_1 = \mu_2 = \mu_3 = \mu_4 \text{ y } H_1\text{: al menos un par de } \mu_i \text{ difiere entre sí}$$

siendo μ_i la media verdadera de resistencia a la tensión obtenida con el método i. $i = 1, 2, 3, 4$

```r
# Plot
options(repr.plot.width=10, repr.plot.height=6)
data_8 %>%
ggplot(aes(y=resistencia, x=metodo)) +
    labs(
        title="Resistencia a la tensión del cemento portland",
        subtitle="para cuatro técnicas de mezclado (diagramas de caja)",
        x="Técnica de mezclado",
        y="Resistencia a la tensión [lb/in²]") +
    geom_boxplot(alpha=0.15, fill=c("#ff0303", "#20ff03", "#ff03e6", "#031cff")) +
    theme_light() +
    theme(text=element_text(size=20),
         plot.title = element_text(size=24, hjust = 0.5),
         plot.subtitle = element_text(size=20, hjust = 0.5))
```




```r
kruskal.test(resistencia ~ metodo, data_8)
```

```
        Kruskal-Wallis rank sum test

data:  resistencia by metodo
Kruskal-Wallis chi-squared = 10.028, df = 3, p-value = 0.01833
```

Como el valor p-valor de la prueba es menor a 0.05, se rechaza la hipótesis nula y se concluye que las técnicas de mezclado afectan la resistencia.
