Diferencia entre medias

Pia Trnovec

2023-01-29

El data set presenta diferentes variables sobre las ventas de una cadena en diferentes tiendas de sillitas de bebé. El data set ya esta arreglado para poder analizar.

- Cargad los datos mediante la instrucción data <- read.csv("ChildCarSeats_clean.csv", stringsAsFactors = TRUE)
- Echad un vistazo a los datos (instrucción str(data) o summary(data))
- Estudiad si hay diferencias en las medias de las ventas (variable Sales) para las tiendas de USA y de fuera de USA (variable US)
- Estudiad si hay diferencias en las medias de las ventas para las tiendas de zona rural y de zona urbana (variable Urban)
- Estudiad si hay diferencias en las medias de las ventas para los diferentes tipos de calidad en la ubicación en los estantes de las tiendas (variable ShelveLoc)

```
# Cargar los paquetes necesarios
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(ggplot2)

install.packages("knitr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(knitr)

# Cargar el conjunto de datos
data <- read.csv("ChildCarSeats_clean.csv", stringsAsFactors = TRUE)

# Ver la estructura de los datos
str(data)
```

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : int  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : int  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: int  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : int  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : int  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
```

```
##  $ Age         : int  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education   : int  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban       : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US          : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```
# Ver el resumen de los datos
summary(data)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.435   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.410   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.160   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

```
# Realice la prueba t para comparar las ventas medias de las tiendas en los EE. UU. y las tiendas fuera
ttest_US <- t.test(Sales ~ US, data=data)
ttest_US
```

```
##
##  Welch Two Sample t-test
##
## data:  Sales by US
## t = -4.9705, df = 354.64, p-value = 1.042e-06
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -1.7963824 -0.7778386
## sample estimates:
##  mean in group No mean in group Yes
##          6.579789          7.866899
```

```
# Realice la prueba t para comparar las ventas medias de tiendas en áreas rurales y urbanas
ttest_Urban <- t.test(Sales ~ Urban, data=data)
ttest_Urban
```

```
##
##  Welch Two Sample t-test
##
## data:  Sales by Urban
## t = 0.47068, df = 220.63, p-value = 0.6383
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.4487328  0.7303280
## sample estimates:
##  mean in group No mean in group Yes
##          7.509237           7.368440
```

```
# Realice la prueba ANOVA para comparar las ventas medias para diferentes tipos de calidad en la ubicac
aov_ShelveLoc <- aov(Sales ~ ShelveLoc, data=data)
aov_ShelveLoc
```

```
## Call:
##    aov(formula = Sales ~ ShelveLoc, data = data)
##
## Terms:
##                 ShelveLoc Residuals
## Sum of Squares   832.8471 2146.4830
## Deg. of Freedom         2       397
##
## Residual standard error: 2.325244
## Estimated effects may be unbalanced
```

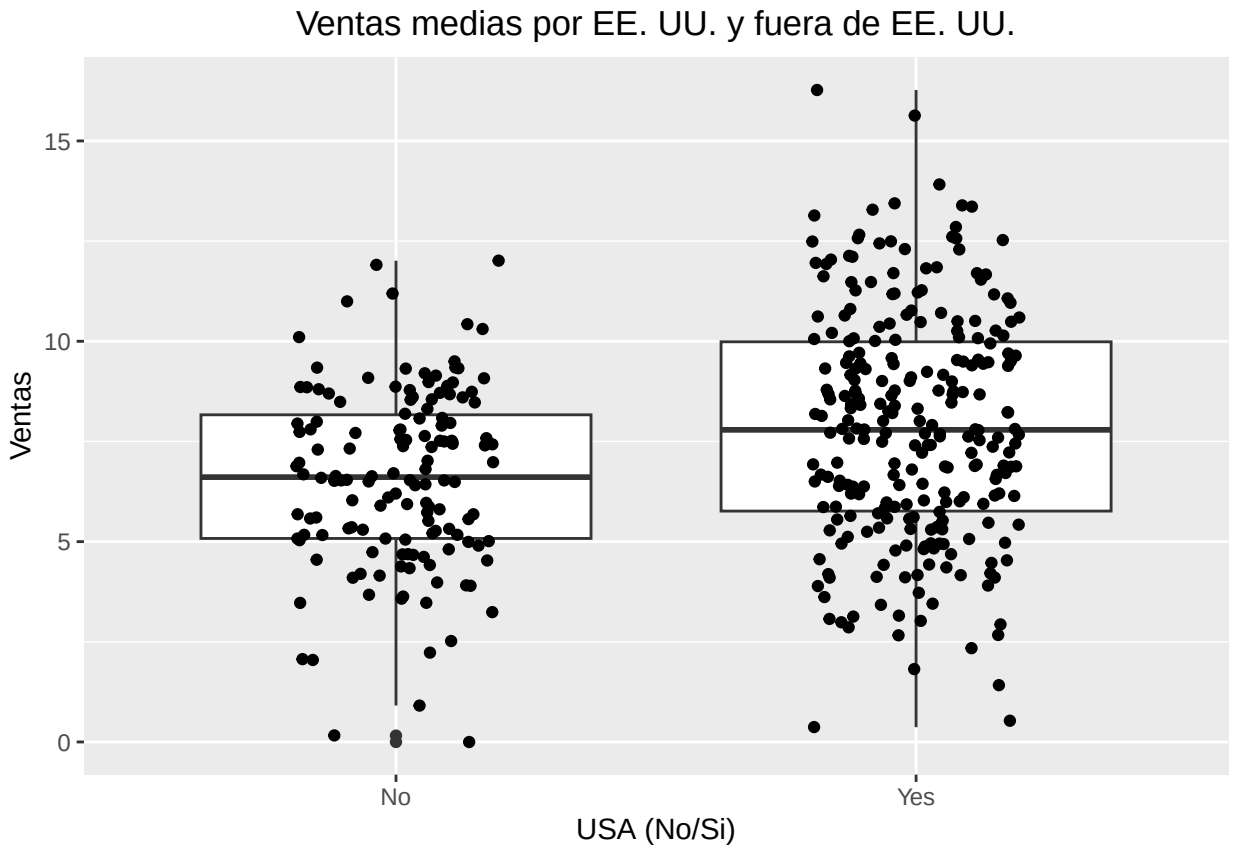
```
# Realice una prueba post-hoc como la prueba HSD de Tukey para identificar qué niveles específicos de S
TukeyHSD_ShelveLoc <- TukeyHSD(aov_ShelveLoc)
TukeyHSD_ShelveLoc
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sales ~ ShelveLoc, data = data)
##
## $ShelveLoc
##                  diff       lwr       upr p adj
## Good-Bad     4.284730  3.470028  5.099433     0
## Medium-Bad   1.783659  1.114079  2.453239     0
## Medium-Good -2.501072 -3.200125 -1.802019     0
```

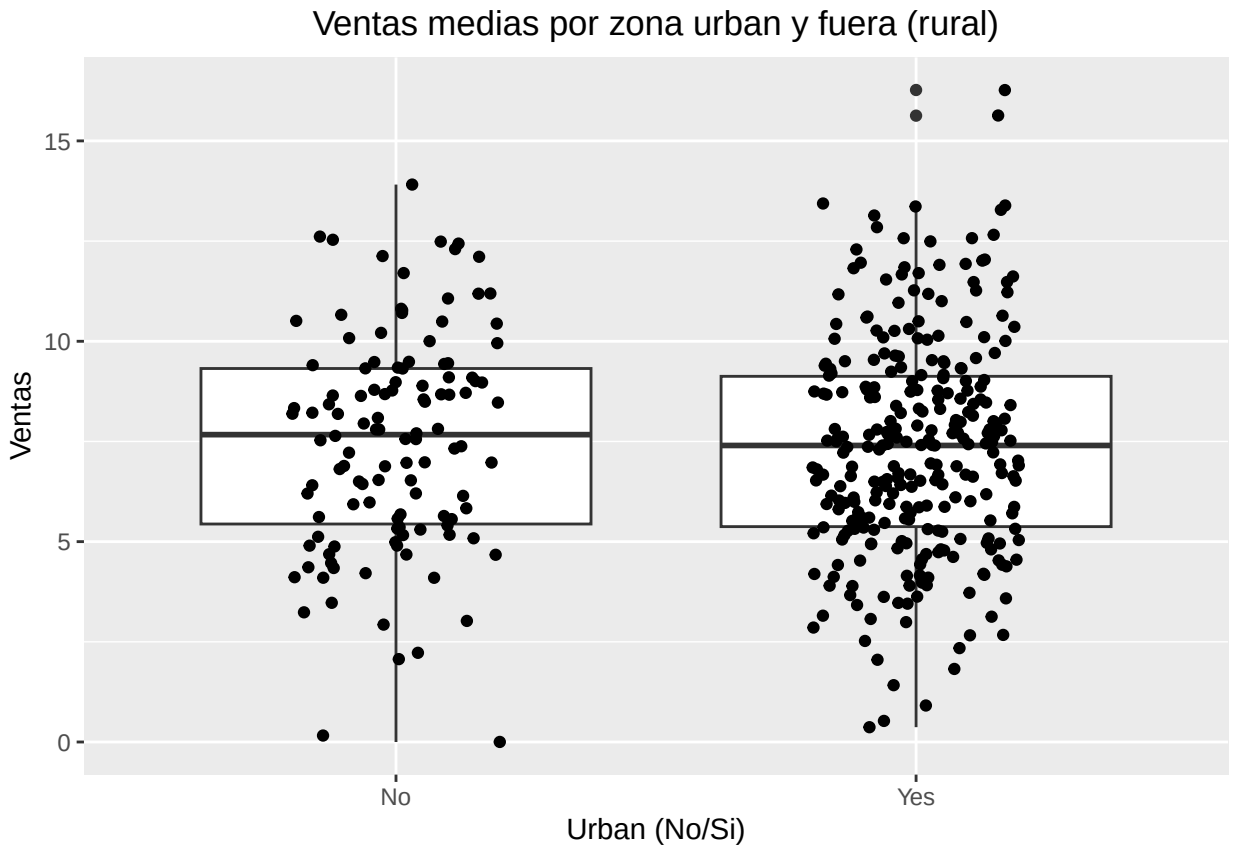
```
# Realice la prueba de Kruskal-Wallis para comparar las ventas medias para diferentes tipos de calidad
kruskal_ShelveLoc <- kruskal.test(Sales ~ ShelveLoc, data=data)
kruskal_ShelveLoc
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Sales by ShelveLoc
## Kruskal-Wallis chi-squared = 106.51, df = 2, p-value < 2.2e-16
```

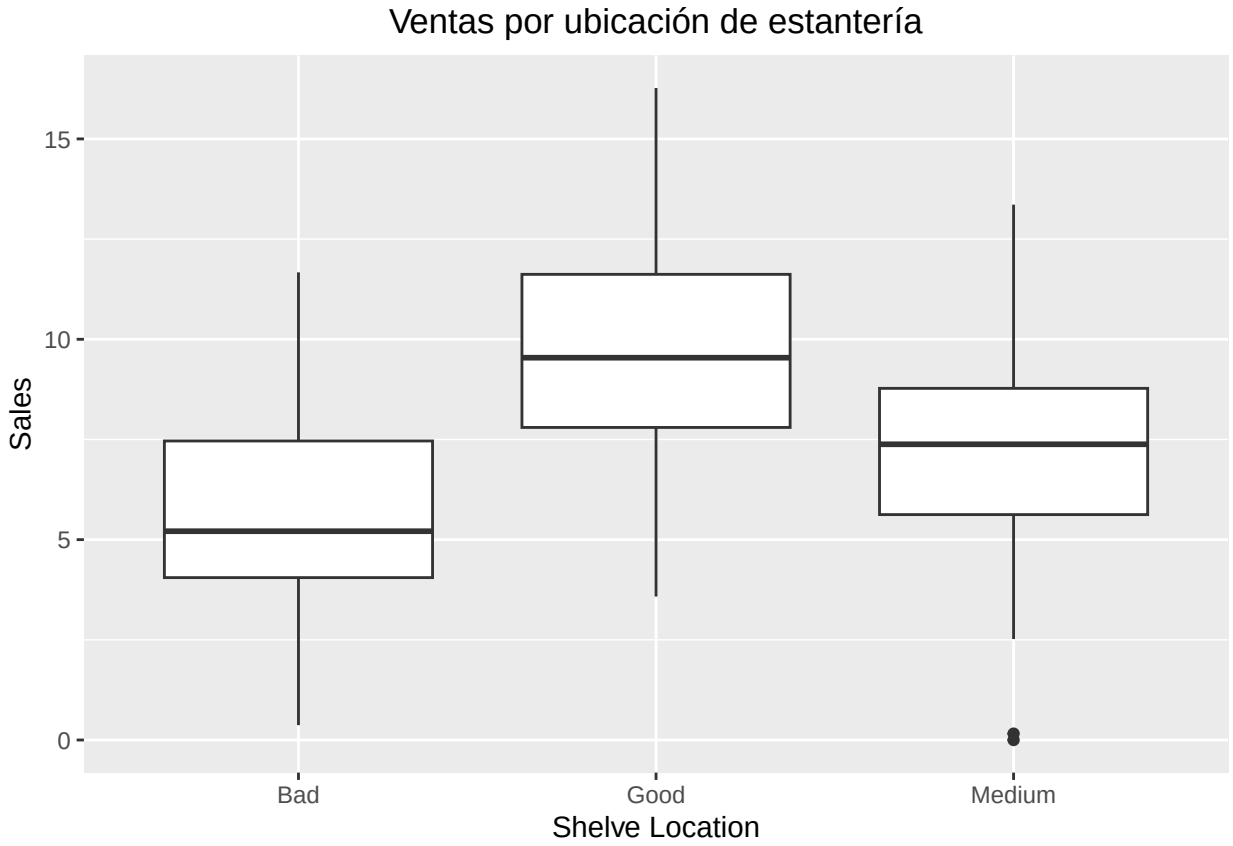
```
#Ventas medias para tiendas en EE. UU. y tiendas fuera de EE. UU.
ggplot(data, aes(x = US, y = Sales)) +
  geom_boxplot() +
  geom_jitter(width = 0.2) +
  labs(x = "USA (No/Si)", y = "Ventas") +
  ggtitle("Ventas medias por EE. UU. y fuera de EE. UU.") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Ventas medias para las tiendas de zona rural y de zona urbana
library(ggplot2)
ggplot(data, aes(x = Urban, y = Sales)) +
  geom_boxplot() +
  geom_jitter(width = 0.2) +
  labs(x = "Urban (No/Si)", y = "Ventas") +
  ggtitle("Ventas medias por zona urban y fuera (rural)") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Crear diagrama de caja para mostrar las ventas por ShelveLoc
ggplot(data, aes(x = ShelveLoc, y = Sales)) +
  geom_boxplot() +
  labs(title = "Ventas por ubicación de estantería",
       x = "Shelve Location", y = "Sales") +
  theme(plot.title = element_text(hjust = 0.5))
```



En la comparación de las ventas promedio de las tiendas en EE. UU. y las tiendas fuera de EE. UU., el t-test Welch de dos muestras muestra un valor t de -5 y un valor p de 1e-06, lo que indica que hay una diferencia significativa en los promedios de los dos grupos (intervalo de confianza del 95%: -1.80 a -0.78). Las ventas promedio de las tiendas en EE. UU. son más altas (7.9) que las tiendas fuera de EE. UU. (6.6).

En la comparación de las ventas promedio de las tiendas en zonas rurales y urbanas, el t-test Welch de dos muestras muestra un valor t de 0.5 y un valor p de 0.6, lo que indica que no hay una diferencia significativa en los promedios de los dos grupos (intervalo de confianza del 95%: -0.45 a 0.73).

En la comparación de las ventas promedio de diferentes tipos de ubicación de estanterías en las tiendas, el test ANOVA muestra que hay una diferencia significativa entre los promedios ($p<0.05$). El test Tukey's HSD, un test post-hoc, identifica que la ubicación de la estantería de buena tienda tiene la venta promedio más alta (11.7) y la ubicación de la estantería de mala tienda tiene la venta promedio más baja (7.4).