Práctica Diferencia Medias

Mireia Gómez Diaz

07/01/2022

Carga de los datos

El primer paso es cargar los datos ya arreglados, listos para poder ser analizados:

```
data <- read.csv("ChildCarSeats_clean.csv", stringsAsFactors = TRUE)
```

Observamos que con esta instrucción cargará los datos agrupando los campos de texto en factores (diferentes opciones para una misma variable).

El data set presenta diferentes variables sobre las ventas de una cadena en diferentes tiendas de sillitas de bebé. Vamos ahora a visualizar la cabecera, los tipos de datos que encontramos y un resumen con las principales características de cada variable observada:

```
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 ...
```

```
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
##
##
##
##
```

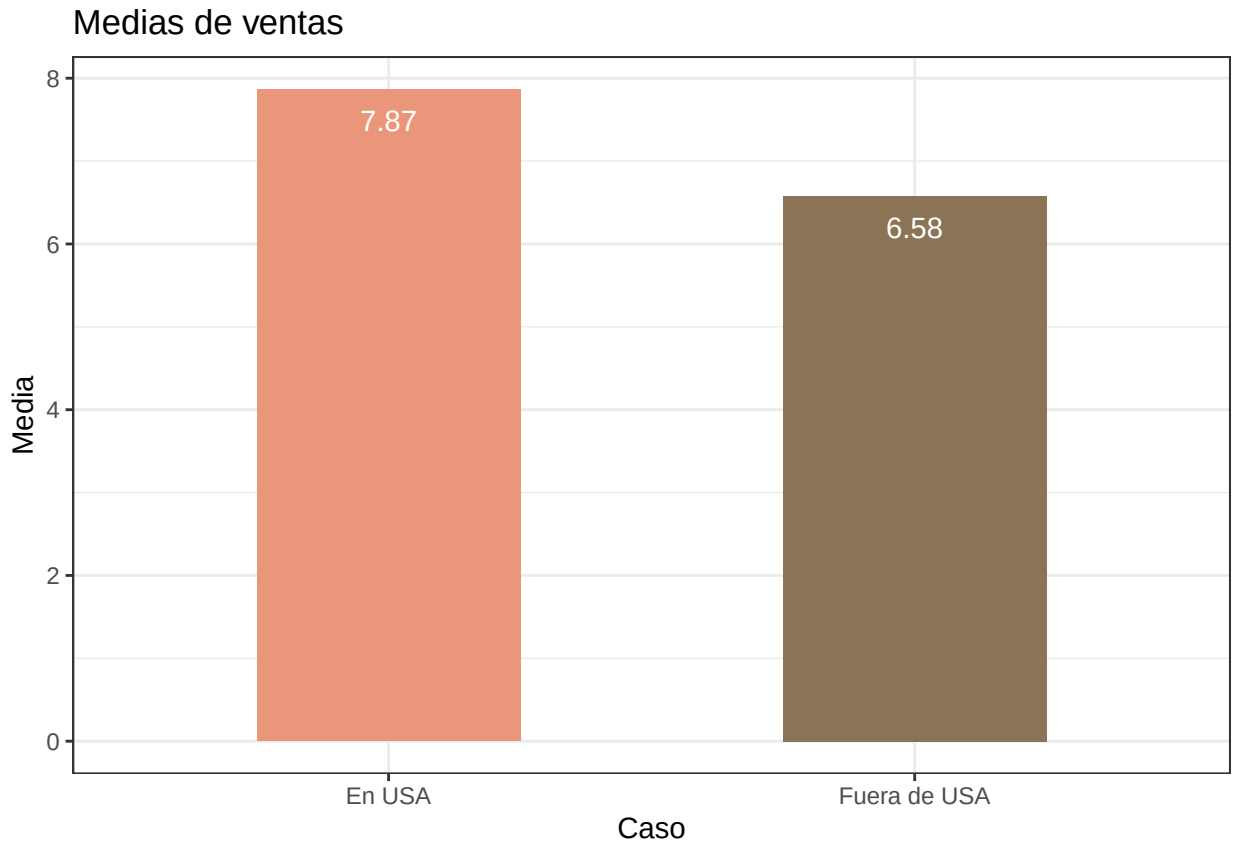
Tiendas en USA y fuera de USA

Queremos estudiar si hay diferencias en las medias de las ventas (variable Sales) para las tiendas de USA y de fuera de USA (variable US). Se trata de muestras independientes, ya que incluso tienen tamaños distintos (como se aprecia en el resumen anterior).

```
sales_in_US <- data %>% filter(US == "Yes") %>% pull(Sales)
sales_out_US <- data %>% filter(US == "No") %>% pull(Sales)
```

Graficamos las medias:

```
sales_mean <- data.frame(Mean = c(mean(sales_in_US), mean(sales_out_US)))
sales_mean$Case <- c("En USA", "Fuera de USA")
ggplot(data = sales_mean, aes(x = Case, y = Mean, label = round(Mean, 2))) +
  geom_bar(stat = "identity", width = 0.5, fill = c("darksalmon", "burlywood4")) +
  geom_text(vjust = 2, colour = "white") +
  labs(title = "Medias de ventas", x = "Caso", y = "Media") +
  theme_bw()
```

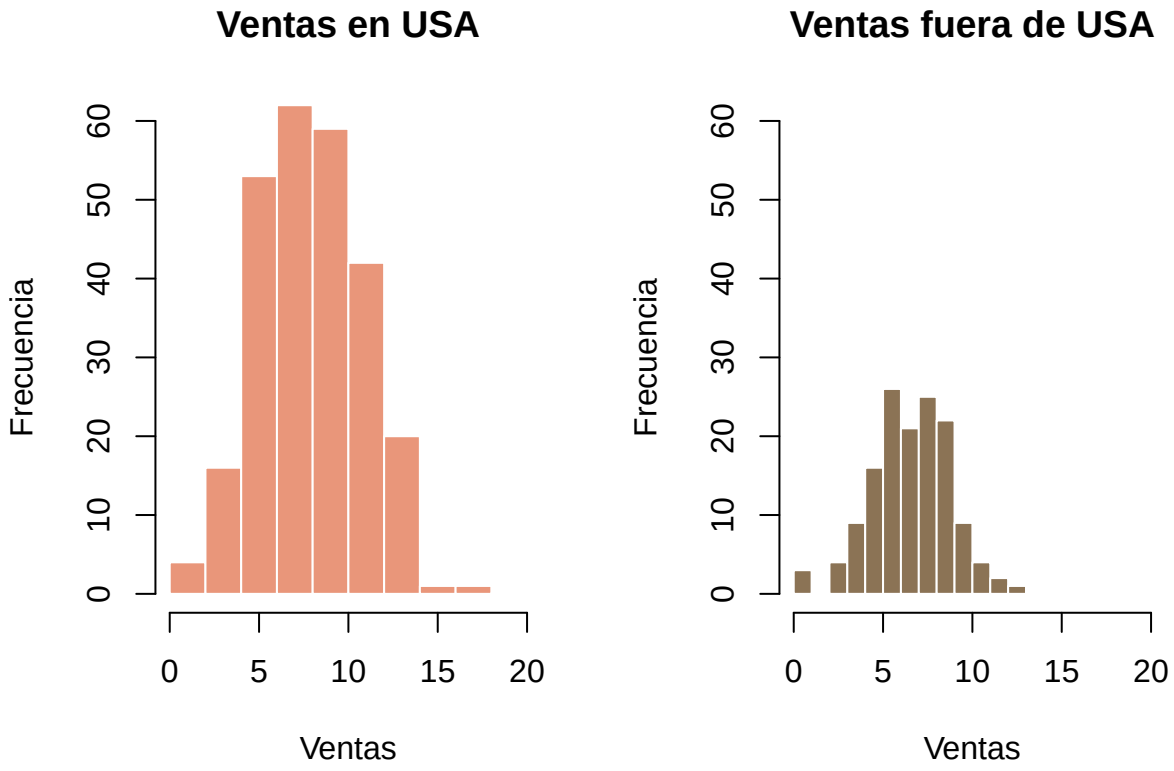


La hipótesis nula es que no hay diferencia entre las ventas. Vamos a comprobar una serie de características para escoger el test adecuado que nos permita intentar rechazar esa hipótesis.

Primero veamos un histograma de las ventas:

```
par(mfrow = c(1,2))
hist(sales_in_US, main = "Ventas en USA", col = "darksalmon",
```

```
     xlab = "Ventas", ylab = "Frecuencia", xlim = c(0,20), ylim = c(0, 60),
     border = "white")
hist(sales_out_US, main = "Ventas fuera de USA", col = "burlywood4",
     xlab = "Ventas", ylab = "Frecuencia", xlim = c(0,20), ylim = c(0, 60),
     border = "white")
```



Visualmente parecen normales. Realizamos el test de Shapiro para comprobar si debemos rechazar la normalidad:

```
shapiro.test(sales_in_US)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_in_US
## W = 0.99545, p-value = 0.6499
```

```
shapiro.test(sales_out_US)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_out_US
## W = 0.98729, p-value = 0.2181
```

Los p-valores son altos, por lo que no podemos rechazar la hipótesis que son normales.

Realizamos ahora el test de diferencia de varianzas para ver si debemos rechazar la hipotésis de que son iguales:

```
var.test(sales_in_US, sales_out_US)
```

```
##
##  F test to compare two variances
##
## data:  sales_in_US and sales_out_US
## F = 1.667, num df = 257, denom df = 141, p-value = 0.0008679
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.237737 2.217092
## sample estimates:
## ratio of variances
##           1.666969
```

El p-valor es muy pequeño, por lo que rechazamos la hipotésis de que tienen la misma varianza. Ya podemos aplicar el t-test para saber si hay una diferencia de medias:

```
t.test(sales_in_US, sales_out_US, paired=FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  sales_in_US and sales_out_US
## t = 4.9705, df = 354.64, p-value = 1.042e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7778386 1.7963824
## sample estimates:
## mean of x mean of y
##  7.866899  6.579789
```

Conclusión: El p-valor es muy pequeño, por lo que podemos rechazar la hipótesis nula. Es decir, la media de ventas de este producto es distinta en USA que fuera de USA.

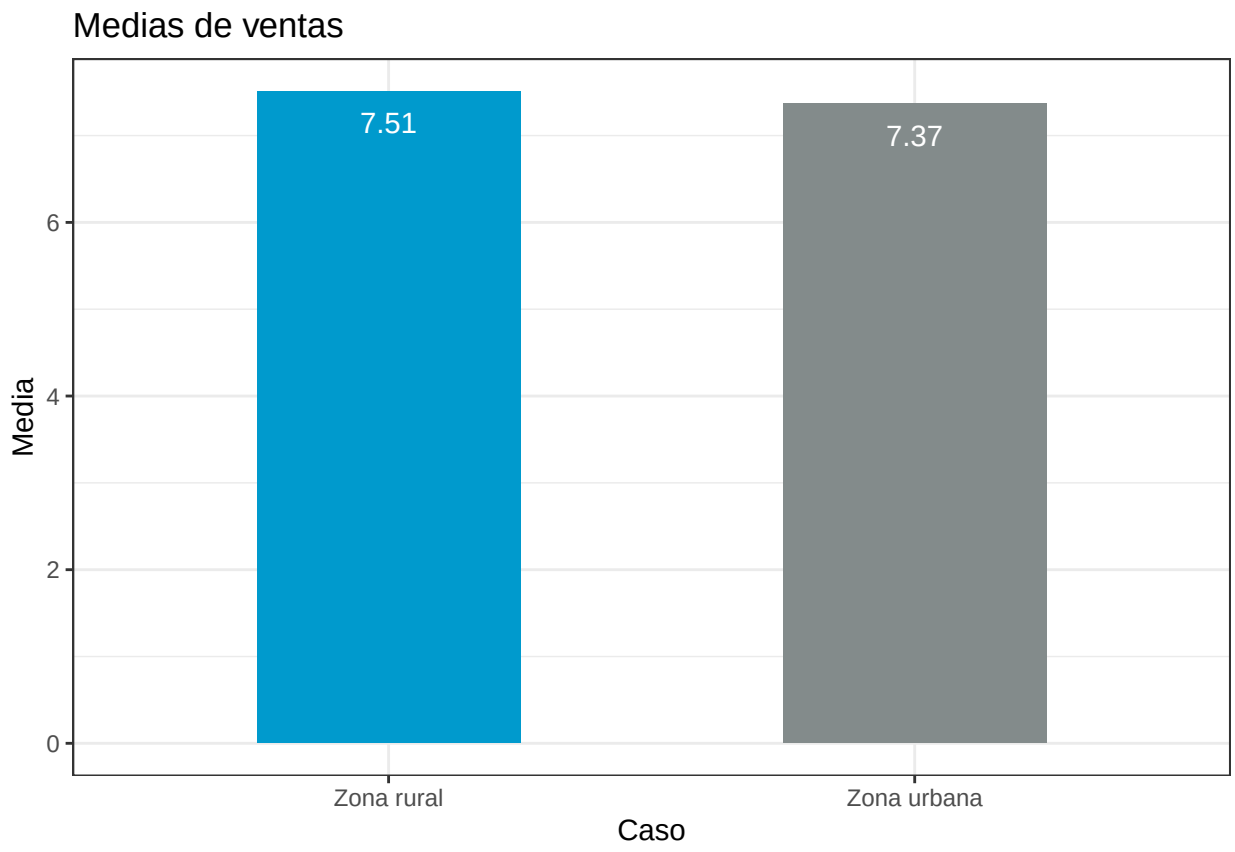
Zonas rurales y zonas urbanas

Queremos estudiar si hay diferencias en las medias de las ventas para las tiendas de zona rural y de zona urbana (variable Urban). Se trata de muestras independientes.

```
sales_rural <- data %>% filter(Urban == "No") %>% pull(Sales)
sales_urban <- data %>% filter(Urban == "Yes") %>% pull(Sales)
```

Graficamos las medias:

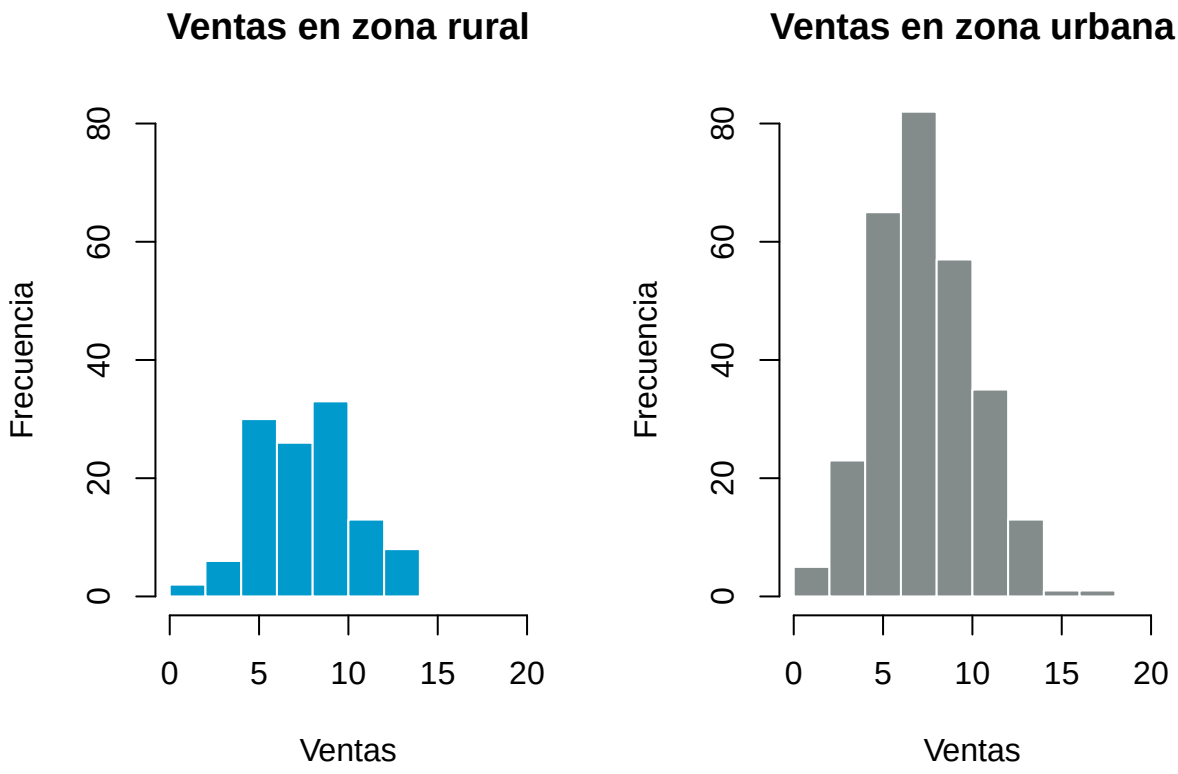
```
sales_mean <- data.frame(Mean = c(mean(sales_urban), mean(sales_rural)))
sales_mean$Case <- c("Zona urbana", "Zona rural")
ggplot(data = sales_mean, aes(x = Case, y = Mean, label = round(Mean, 2))) +
  geom_bar(stat = "identity", width = 0.5, fill = c("azure4", "deepskyblue3")) +
  geom_text(vjust = 2, colour = "white") +
  labs(title = "Medias de ventas", x = "Caso", y = "Media") +
  theme_bw()
```



Primero veamos un histograma de las ventas:

```
par(mfrow = c(1,2))
hist(sales_rural, main = "Ventas en zona rural", col = "deepskyblue3",
     xlab = "Ventas", ylab = "Frecuencia", xlim = c(0,20), ylim = c(0, 80),
     border = "white")
hist(sales_urban, main = "Ventas en zona urbana", col = "azure4",
```

```
    xlab = "Ventas", ylab = "Frecuencia", xlim = c(0,20), ylim = c(0, 80),
    border = "white")
```



Visualmente parecen normales. Realizamos el test de Shapiro para comprobar si debemos rechazar la normalidad:

```
shapiro.test(sales_rural)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_rural
## W = 0.99092, p-value = 0.6306
```

```
shapiro.test(sales_urban)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_urban
## W = 0.99379, p-value = 0.2993
```

Los p-valores son altos, por lo que no podemos rechazar la hipótesis que son normales.

Realizamos ahora el test de diferencia de varianzas para ver si debemos rechazar la hipotésis de que son iguales:

```r
var.test(sales_rural, sales_urban)
```

```
##
##  F test to compare two variances
##
## data:  sales_rural and sales_urban
## F = 0.98783, num df = 117, denom df = 281, p-value = 0.9545
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7344536 1.3548403
## sample estimates:
## ratio of variances
##          0.9878318
```

El p-valor es alto, por lo que no rechazamos la hipotésis de que tienen la misma varianza. Ya podemos aplicar el t-test para saber si hay una diferencia de medias:

```r
t.test(sales_urban, sales_rural, paired=FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  sales_urban and sales_rural
## t = -0.4695, df = 398, p-value = 0.639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7303640  0.4487689
## sample estimates:
## mean of x mean of y
##  7.368440  7.509237
```

Conclusión: El p-valor es alto, por lo que no podemos rechazar la hipótesis nula. Es decir, la media de ventas de este producto en zonas urbanas es la misma que en zonas rurales.

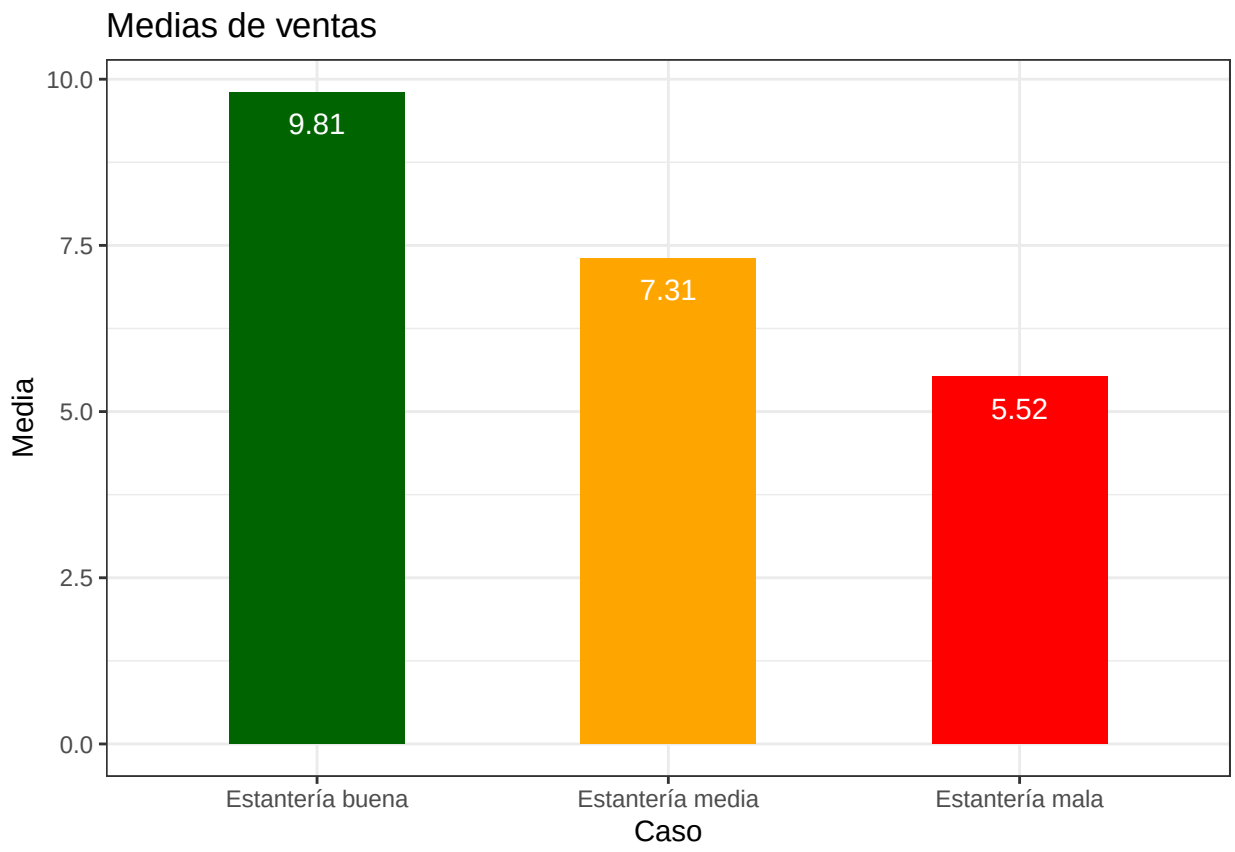
Estanterías buenas, medias y malas

Queremos estudiar 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). Se trata de muestras independientes.

```
sales_good_shelve <- data %>% filter(ShelveLoc == "Good") %>% pull(Sales)
sales_medium_shelve <- data %>% filter(ShelveLoc == "Medium") %>% pull(Sales)
sales_bad_shelve <- data %>% filter(ShelveLoc == "Bad") %>% pull(Sales)
```

Graficamos las medias:

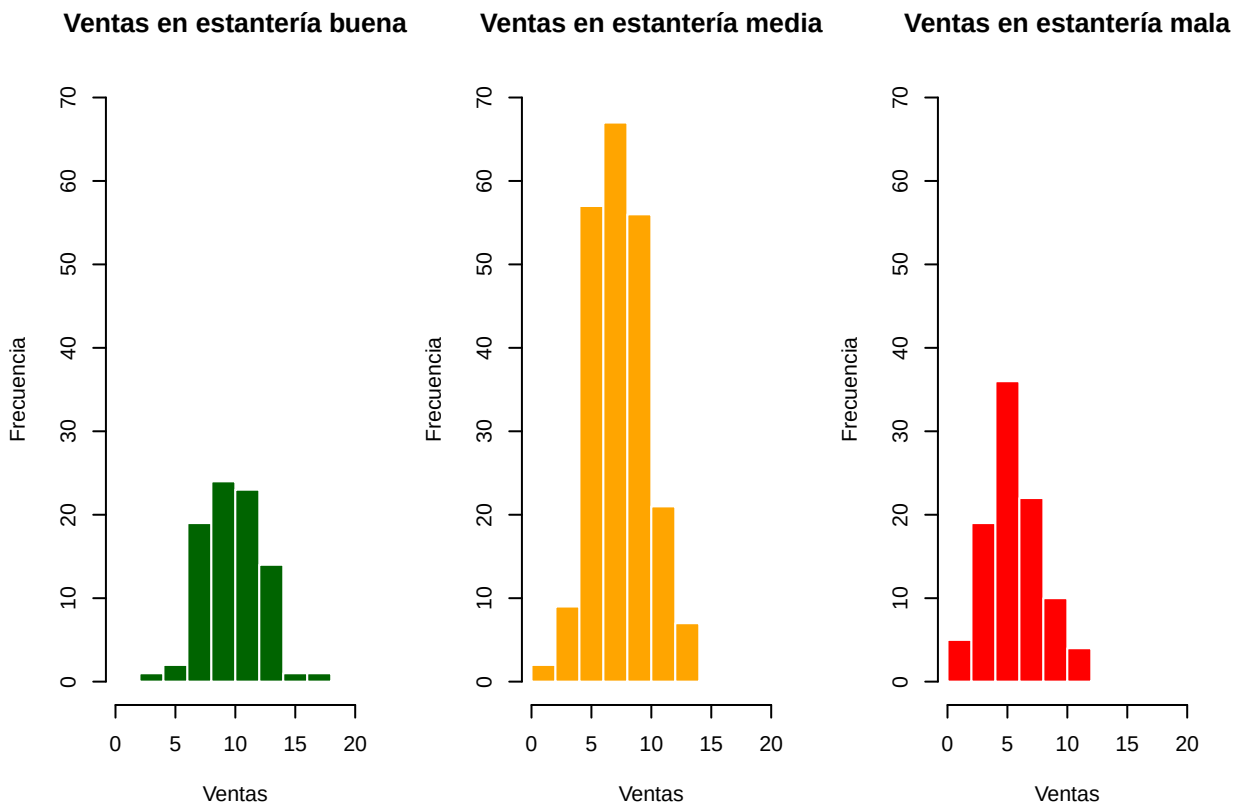
```
sales_mean <- data.frame(Mean = c(mean(sales_good_shelve), mean(sales_medium_shelve),
                                  mean(sales_bad_shelve)))
sales_mean$Case <- c("Estantería buena", "Estantería media", "Estantería mala")
ggplot(data = sales_mean, aes(x = reorder(Case, -Mean), y = Mean,
                              label = round(Mean, 2))) +
  geom_bar(stat = "identity", width = 0.5, fill = c("darkgreen", "orange", "red")) +
  geom_text(vjust = 2, colour = "white") +
  labs(title = "Medias de ventas", x = "Caso", y = "Media") +
  theme_bw()
```



Primero veamos un histograma de las ventas:

```
par(mfrow = c(1,3))
hist(sales_good_shelve, main = "Ventas en estantería buena", col = "darkgreen",
```

```
     xlab = "Ventas", ylab = "Frecuencia", xlim = c(0,20), ylim = c(0, 70),
     border = "white")
hist(sales_medium_shelve, main = "Ventas en estantería media", col = "orange",
     xlab = "Ventas", ylab = "Frecuencia", xlim = c(0,20), ylim = c(0, 70),
     border = "white")
hist(sales_bad_shelve, main = "Ventas en estantería mala", col = "red",
     xlab = "Ventas", ylab = "Frecuencia", xlim = c(0,20), ylim = c(0, 70),
     border = "white")
```



Visualmente parecen normales. Realizamos el test de Shapiro para comprobar si debemos rechazar la normalidad:

```
shapiro.test(sales_good_shelve)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_good_shelve
## W = 0.9912, p-value = 0.8411
```

```
shapiro.test(sales_medium_shelve)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  sales_medium_shelve
## W = 0.99236, p-value = 0.3139
```

```
shapiro.test(sales_bad_shelve)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sales_bad_shelve
## W = 0.98891, p-value = 0.6066
```

Los p-valores son altos, por lo que no podemos rechazar la hipótesis que son normales.

Realizamos ahora el test Bartlett, parecido al test de diferencia de varianzas, pero para más de dos variables:

```
bartlett.test(Sales ~ ShelveLoc, data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Sales by ShelveLoc
## Bartlett's K-squared = 0.7014, df = 2, p-value = 0.7042
```

El p-valor es alto, por lo que no rechazamos la hipótesis de que no tienen varianzas iguales.

Como se cumple la normalidad y la homocedasticidad (varianzas iguales) y se trata de muestras independientes (no relacionadas), podemos aplicar el test ANOVA:

```
result <- aov(Sales ~ ShelveLoc, data)
summary(result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ShelveLoc     2  832.8   416.4   77.02 <2e-16 ***
## Residuals   397 2146.5     5.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusión: El p valor es prácticamente nulo y por tanto podemos decir casi sin riesgo que las medias de ventas son diferentes en función de la estantería donde se coloque este producto.