Tarea_2.R

Usuario

2025-09-18

```r
# 04/09/2025
# JEGR
# Base de datos Iris

library("ggplot2")
library("dplyr")
```

```
##
## Adjuntando el paquete: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library("hrbrthemes")
library("viridis")
```

```
## Cargando paquete requerido: viridisLite
```

```r
library("gt")
library("gtExtras")

# Base de datos
data("iris")
View(iris)

# Renombrar base de datos, para hacerlo mas facil de seguir
iris_df <- rename(iris,
                  petal_length = Petal.Length,
                  petal_width = Petal.Width,
                  sepal_length = Sepal.Length,
                  sepal_width = Sepal.Width,
                  species = Species)

# Datos estadisticos descriptivos simples
summary(iris_df)
```

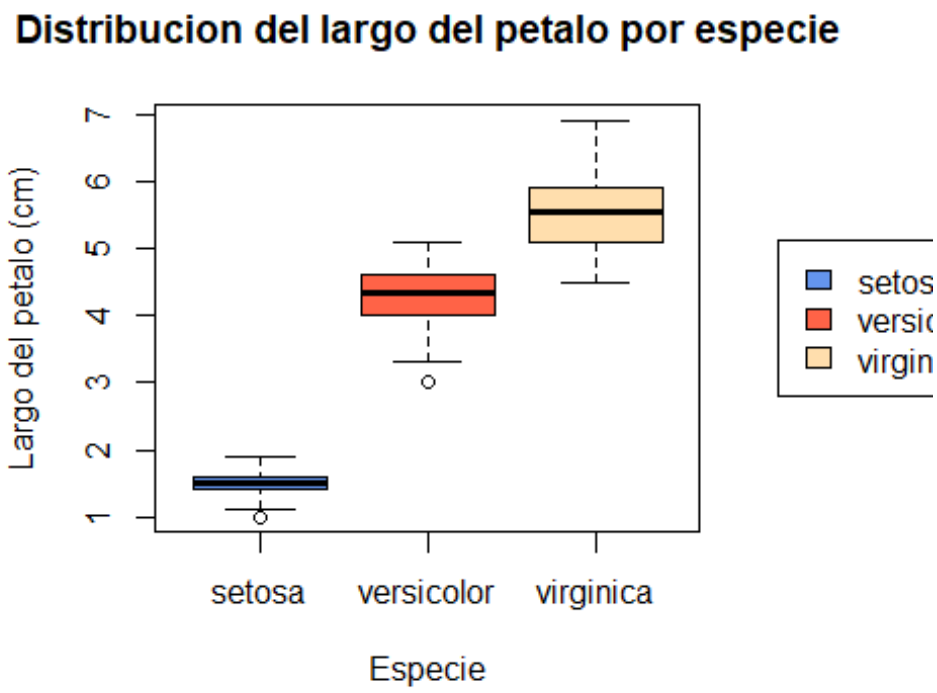
```
##   sepal_length    sepal_width     petal_length    petal_width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
```

```
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```r
head(iris_df)
```

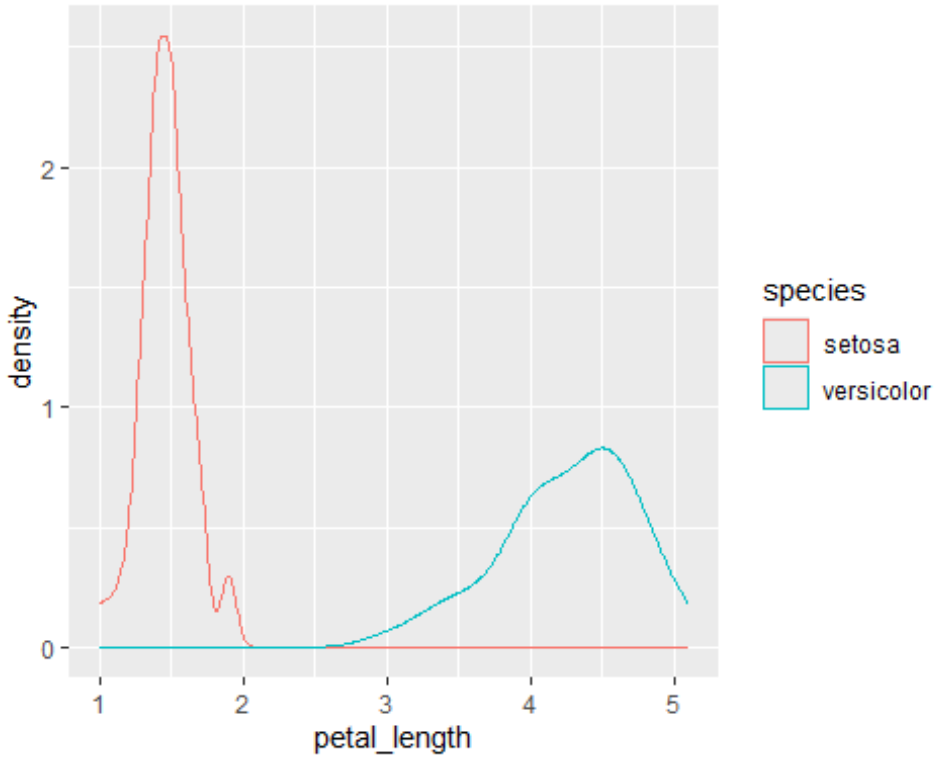
```
##   sepal_length sepal_width petal_length petal_width species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```r
# Grafico boxplot simple
color <- c("cornflowerblue", "tomato", "navajowhite")

par(mar = c(5, 5, 4, 6)) # Cambiar margenes de la grafica

boxplot(iris_df$petal_length ~ iris_df$species,
        col = color,
        main = "Distribucion del largo del petalo por especie",
        xlab = "Especie",
        ylab = "Largo del petalo (cm)")
legend("right",
       legend = c("setosa", "versicolor", "virginica"),
       inset = c(-0.57, 0),
       fill = color,
       col = color,
       xpd = T)
```

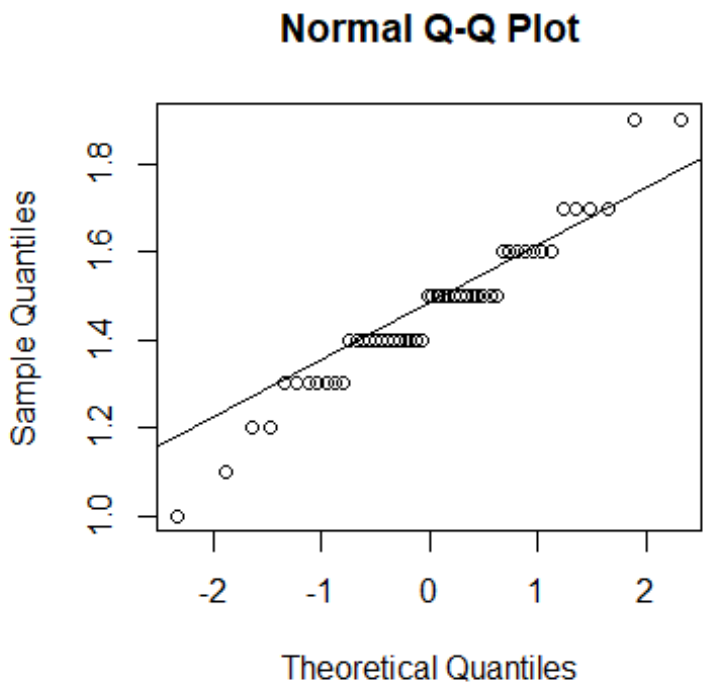



```
# Estadistica descriptiva ---------------------------------------------
--

data_sub <- subset(iris_df, species %in% c("setosa", "versicolor"))
iris_sp <- data.frame(species = data_sub$species,
                      petal_length = data_sub$petal_length) # Dataframe
con solo
# species (en orden, setosa y versicolor) y petal_length
View(iris_sp)
head(iris_sp)

##   species petal_length
## 1  setosa          1.4
## 2  setosa          1.4
## 3  setosa          1.3
## 4  setosa          1.5
## 5  setosa          1.4
## 6  setosa          1.7

summary(iris_sp)

##        species    petal_length
##  setosa    :50   Min.   :1.000
##  versicolor:50   1st Qu.:1.500
##  virginica : 0   Median :2.450
##                  Mean   :2.861
```

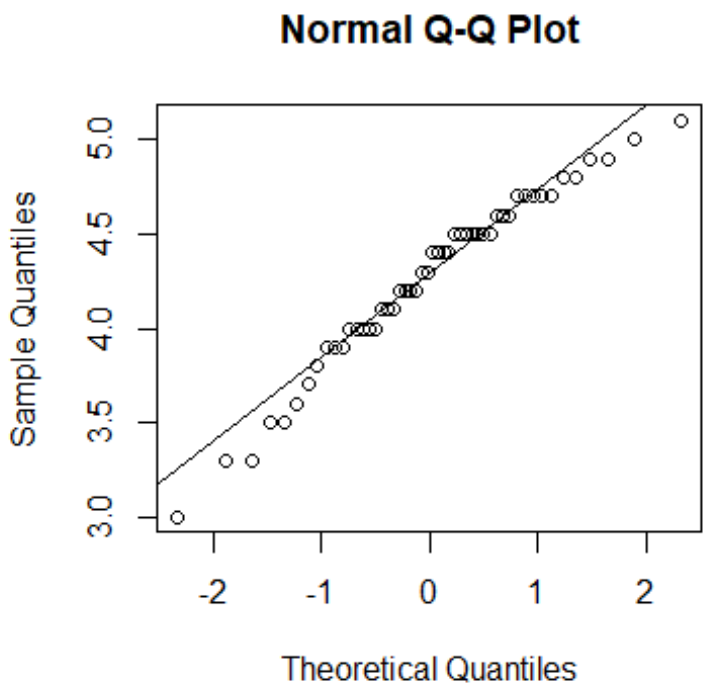
```
##                   3rd Qu.:4.325
##                   Max.   :5.100

# Media, desv.est y varianza

tapply(iris_sp$petal_length, iris_sp$species, mean) # Media de vers. 3
veces mayor

##     setosa versicolor  virginica
##      1.462      4.260         NA

tapply(iris_sp$petal_length, iris_sp$species, sd)   # Var. de vers. 3
veces mayor

##     setosa versicolor  virginica
##   0.173664   0.469911         NA

tapply(iris_sp$petal_length, iris_sp$species, var)  # Sd. de vers. mucho
mayor

##     setosa versicolor  virginica
## 0.03015918 0.22081633         NA

# Grafica de densidad
ggplot(iris_sp, aes(x = petal_length, color = species,))+
  geom_density()
```



```r
df_setosa <- subset(iris_sp, species  == "setosa")
df_versicolor <- subset(iris_sp, species != "setosa")

# Hipotesis ------------------------------------------------------------
--

# ¿En comparacion?
# H0 = no hay una diferencia significativa entre las medias de ambas
especies
# H1 = hay una gran diferencia entre las medias de setosa con respecto a
versicolor


# Grafico de normalidad para ambas especies
qqnorm(df_setosa$petal_length); qqline(df_setosa$petal_length)
```




```r
qqnorm(df_versicolor$petal_length); qqline(df_versicolor$petal_length)
```




```
# Setosa muestra datos una distribucion no normal, al contrario de
versicolor,
# que cuenta con una distribucion normal en su grafica

# Prueba de normalidad
shapiro.test(df_setosa$petal_length)

##
##  Shapiro-Wilk normality test
##
## data:  df_setosa$petal_length
## W = 0.95498, p-value = 0.05481

shapiro.test(df_versicolor$petal_length)

##
##  Shapiro-Wilk normality test
##
## data:  df_versicolor$petal_length
## W = 0.966, p-value = 0.1585

# df_setosa presenta datos no normales (p-value = 0.05481), mientras que
df_versicolor
# si presenta datos normales (p-value = 0.1585)

# Homogeneidad de varianzas
var.test(df_setosa$petal_length, df_versicolor$petal_length)
```

```
##
##  F test to compare two variances
##
## data:  df_setosa$petal_length and df_versicolor$petal_length
## F = 0.13658, num df = 49, denom df = 49, p-value = 1.026e-10
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.07750613 0.24068043
## sample estimates:
## ratio of variances
##          0.1365804
```

```
# IC(95%) = [0.07750613, 0.24068043], p -value =  1.026e-10, varianzas
muy diferentes
# en ambos grupos con un p-value significativo >.005.

##############################################################################
#
##############################################################################
#
# Prueba de t ()
t.test(df_setosa$petal_length, df_versicolor$petal_length,
       alternative = "less",
       var.equal = F)
```

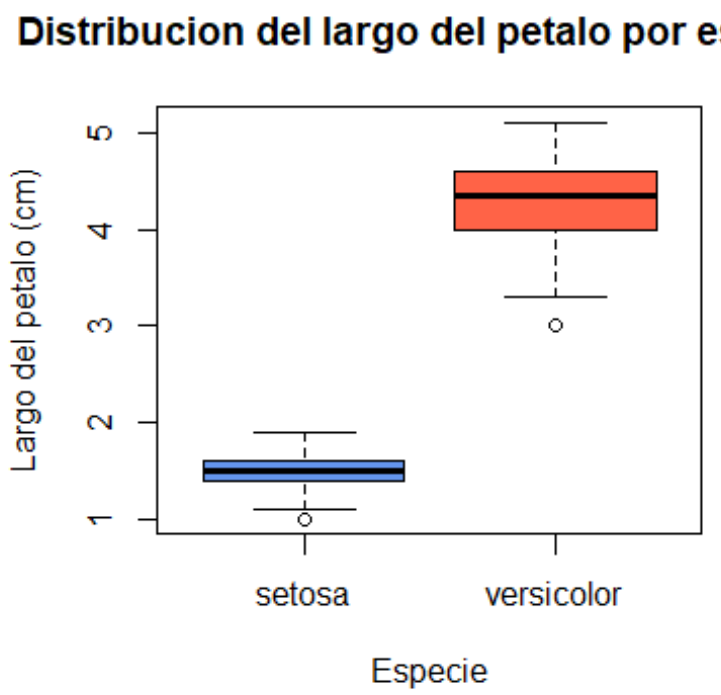
```
##
##  Welch Two Sample t-test
##
## data:  df_setosa$petal_length and df_versicolor$petal_length
## t = -39.493, df = 62.14, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2.679701
## sample estimates:
## mean of x mean of y
##     1.462     4.260
```

```
# p-value < 2.2e-16, menor a 0.05, por lo que se rechaza H0, hay una gran
diferencia
# en tamaño de petalos comparando setosa y versicolor
##############################################################################
#
##############################################################################
#

# Prueba de cohen´s d
cohens_efecto <- function(x,y) {
  n1 <- length(x); n2 <- length(y)
  s1 <- sd(x); s2 <- sd(y)
  sp <- sqrt(((n1-1) * s1^2 + (n2 - 1) * s2^2) / (n1 + n2 -2))
```

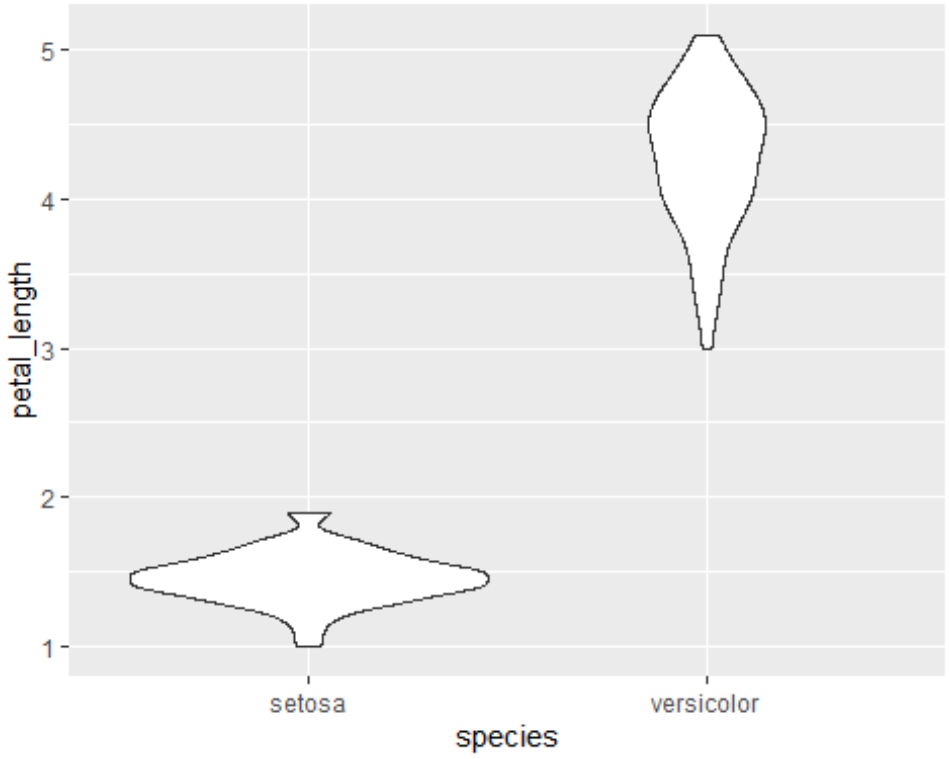
```r
  (mean(x) - mean(y)) / sp
}

d_cal <- (cohens_efecto(df_setosa$petal_length,
df_versicolor$petal_length))
d_cal
```

```
## [1] -7.898544
```

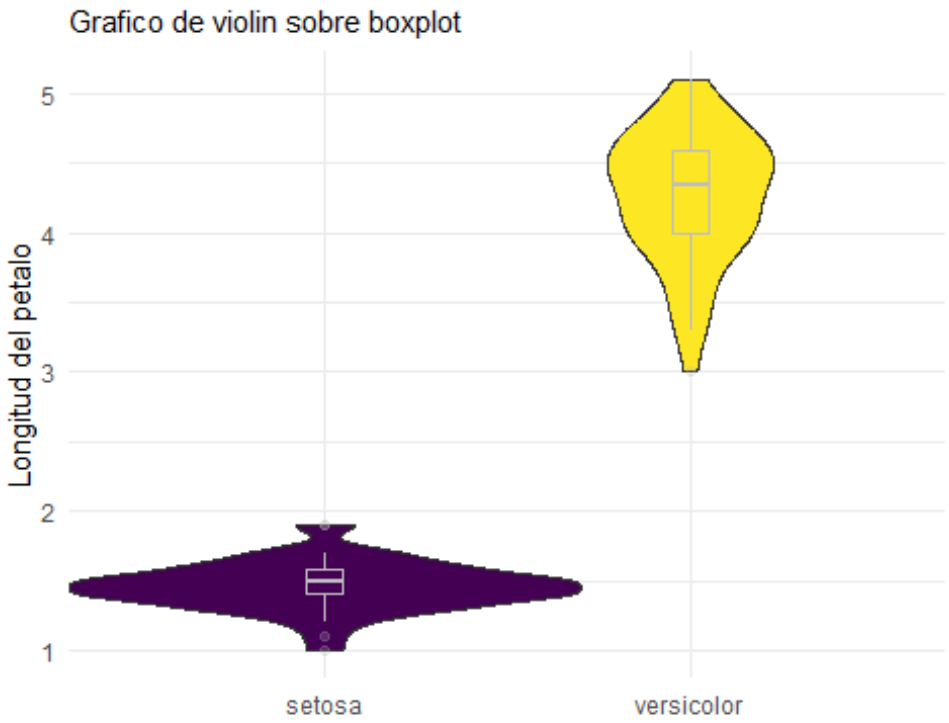
```r
# Valor que representa la diferencia entre las medias de ambas variables
# Mientras mas grande, mayor diferencia habra entre sus medias y datos
# Valor muy significativo, cohen´s D = -7.898544

# Grafico boxplot simple de ambas especies
boxplot(df_setosa$petal_length, df_versicolor$petal_length,
        names = c("setosa","versicolor"),
        col = color,
        main = "Distribucion del largo del petalo por especie",
        xlab = "Especie",
        ylab = "Largo del petalo (cm)")
```




```r
# Grafico de violin
ggplot(iris_sp, aes(x = species, y = petal_length))+
  geom_violin()
```




```
sample_size = iris_sp %>%group_by(species) %>%summarize(num=n())

iris_sp %>%
  left_join(sample_size) %>%
  mutate(myaxis = paste0(species, "", "")) %>%
  ggplot( aes(x=myaxis, y= petal_length, fill=species))+
  ylab("Longitud del petalo")+
  geom_violin(width=1.4)+
  geom_boxplot(width=0.1, color="grey", alpha=0.2)+
  scale_fill_viridis(discrete= T)+
  scale_fill_viridis(discrete = T)+
  theme_minimal()+
  theme(
    legend.position = "none",
    plot.title = element_text(size=11)
  )+
  ggtitle("Grafico de violin sobre boxplot")+
  xlab("")

## Joining with `by = join_by(species)`

## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.

## Warning: `position_dodge()` requires non-overlapping x intervals.
```




```r
# Tablas

table_iris_sp <- data.frame(
  Especies = c("setosa", "versicolor"),
  Media = c(1.462, 4.260),
  Varianza = c(0.03015918, 0.22081633),
  Desv.Estandar = c(0.173664, 0.469911)
)
table_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```

setosa	1.462	0.03015918	0.173664
versicolor	4.260	0.22081633	0.469911

```r
mean_iris_sp <- data.frame(                    # Media de ambas especies
  Especies = c("setosa", "versicolor"),
  Media = c(1.462, 4.260)
)
mean_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```

setosa	1.462
versicolor	4.260

```
var_iris_sp <- data.frame(                    # Varianza de ambas especies
  Especies = c("setosa", "versicolor"),
  Varianza = c(0.03015918, 0.22081633)
)
var_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```

setosa	0.03015918
versicolor	0.22081633

```
sd_iris_sp <- data.frame(                     # Desv.Est de ambas especies
  Especies = c("setosa", "versicolor"),
  Desv.Estandar = c(0.173664, 0.469911)
)
sd_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```

setosa	0.173664
versicolor	0.469911