Tarea_1.R

Usuario

2025-09-18

```
# 31/08/2025
# 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
```

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

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

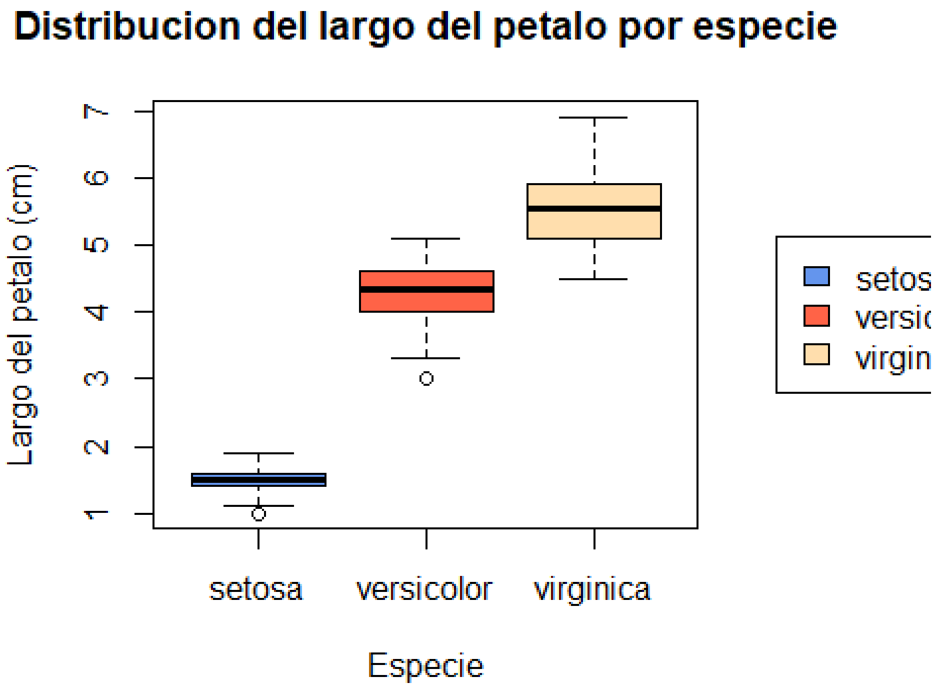
```
library("gt")
library("gtExtras")

# 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)
```

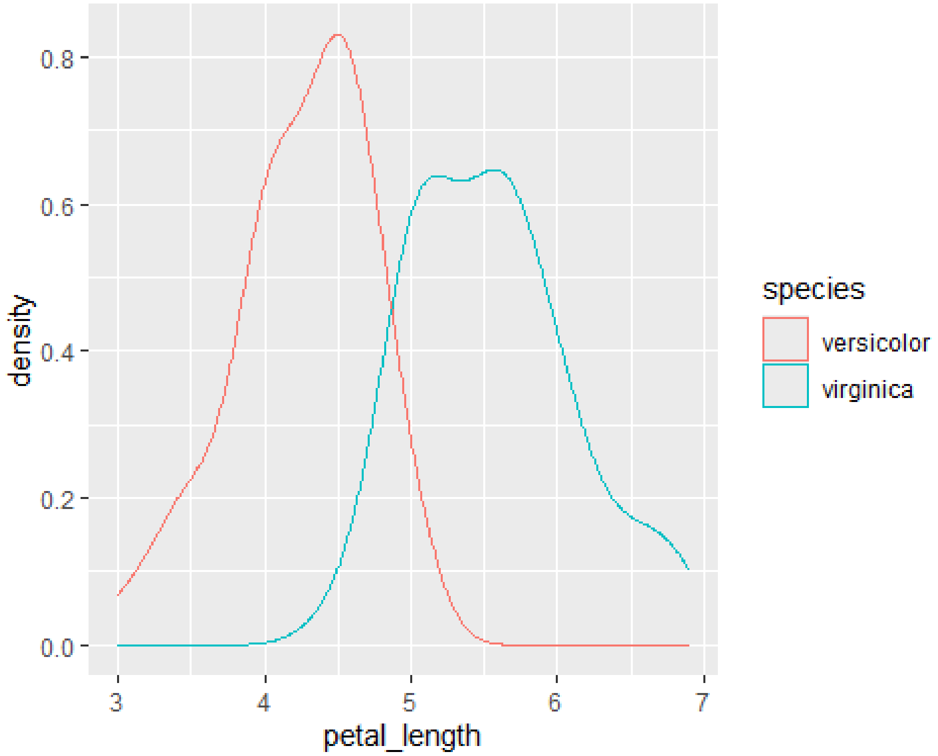


```
# Estadistica descriptiva -------------------------------------------
--

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

```
##      species petal_length
## 1 versicolor          4.7
## 2 versicolor          4.5
## 3 versicolor          4.9
## 4 versicolor          4.0
## 5 versicolor          4.6
## 6 versicolor          4.5
```

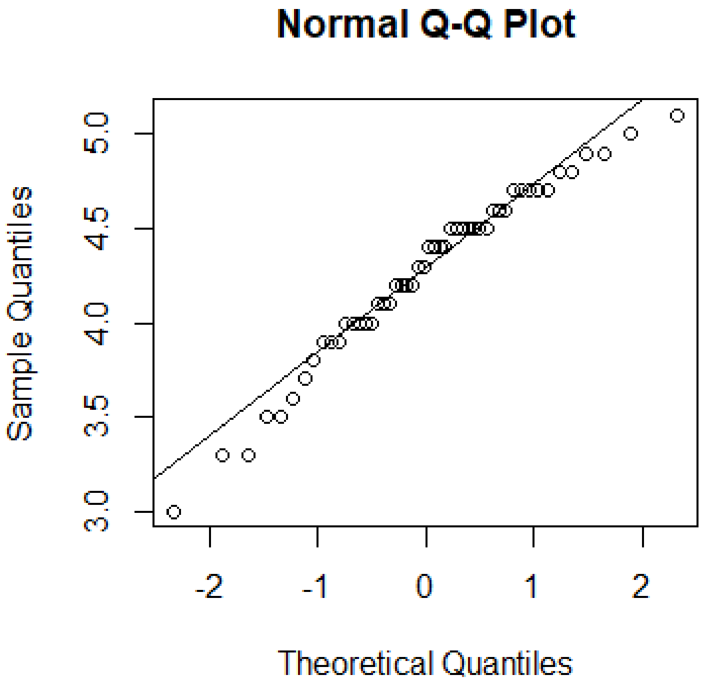
```
summary(iris_sp)
```

```
##        species    petal_length
##  setosa    : 0   Min.   :3.000
##  versicolor:50   1st Qu.:4.375
##  virginica :50   Median :4.900
##                  Mean   :4.906
##                   3rd Qu.:5.525
##                   Max.   :6.900

# Media, desv.est y varianza

tapply(iris_sp$petal_length, iris_sp$species, mean)

##     setosa versicolor  virginica
##         NA      4.260      5.552

tapply(iris_sp$petal_length, iris_sp$species, sd)

##     setosa versicolor  virginica
##         NA  0.4699110  0.5518947

tapply(iris_sp$petal_length, iris_sp$species, var)

##     setosa versicolor  virginica
##         NA  0.2208163  0.3045878

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

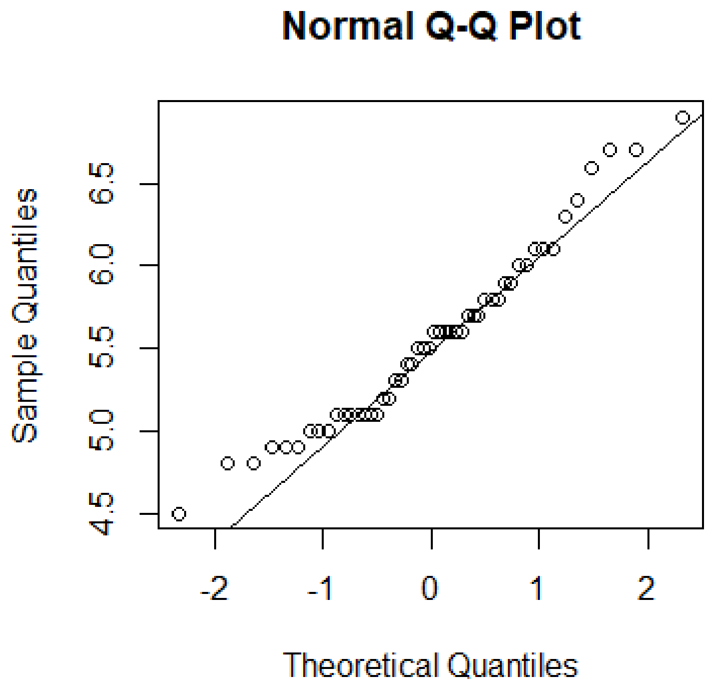


```
df_versicolor <- subset(iris_sp, species  == "versicolor")
df_virginica <- subset(iris_sp, species != "versicolor")

# Hipotesis -------------------------------------------------------------
```

```
--

# ¿Existe una diferencia significante entre el largo del petalo de ambas especies?
# H0 = no hay diferencia
# H1 = si hay diferencia


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



```
qqnorm(df_virginica$petal_length); qqline(df_virginica$petal_length)
```



```
# Ambos tienen datos normales

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

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

shapiro.test(df_virginica$petal_length)

##
##  Shapiro-Wilk normality test
##
## data:  df_virginica$petal_length
## W = 0.96219, p-value = 0.1098

# Mayor a 0.05 (p-value = 0.1585), por lo que existe normalidad en
variables

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

##
##  F test to compare two variances
##
```

```
## data:  df_versicolor$petal_length and df_virginica$petal_length
## F = 0.72497, num df = 49, denom df = 49, p-value = 0.2637
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.411402 1.277530
## sample estimates:
## ratio of variances
##          0.7249678
```

```
# p-value = 0.2637, varianzas relativamente similares, se puede utilizar
prueba de t

# Prueba de t
t.test(df_versicolor$petal_length, df_virginica$petal_length,
       alternative = "two.sided",
       var.equal = T)
```

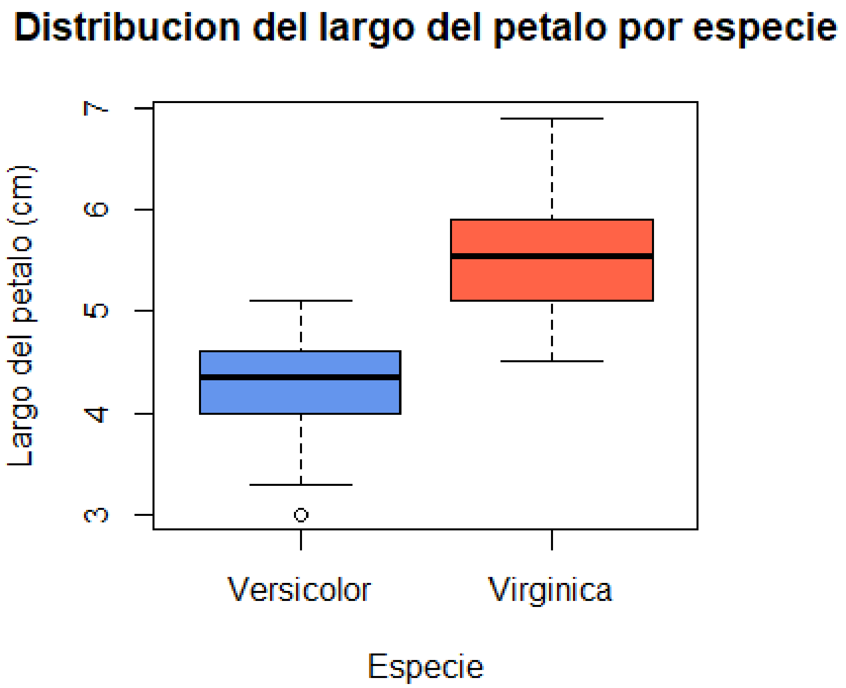
```
##
##  Two Sample t-test
##
## data:  df_versicolor$petal_length and df_virginica$petal_length
## t = -12.604, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.495426 -1.088574
## sample estimates:
## mean of x mean of y
##     4.260     5.552
```

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

# 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))
  (mean(x) - mean(y)) / sp
}

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

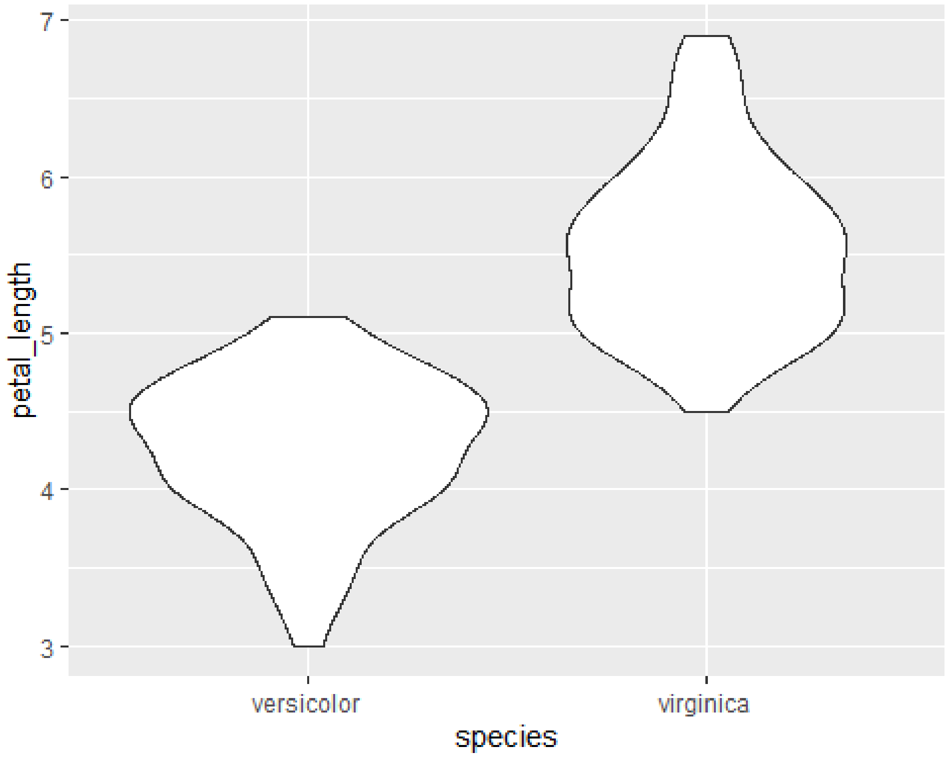
```
## [1] -2.520756
```

```
# Valor que representa la diferencia entre las medias de ambas variables
# Mientras mas grande, mayor diferencia habra entre sus medias y datos

# Grafico boxplot simple de ambas especies
```

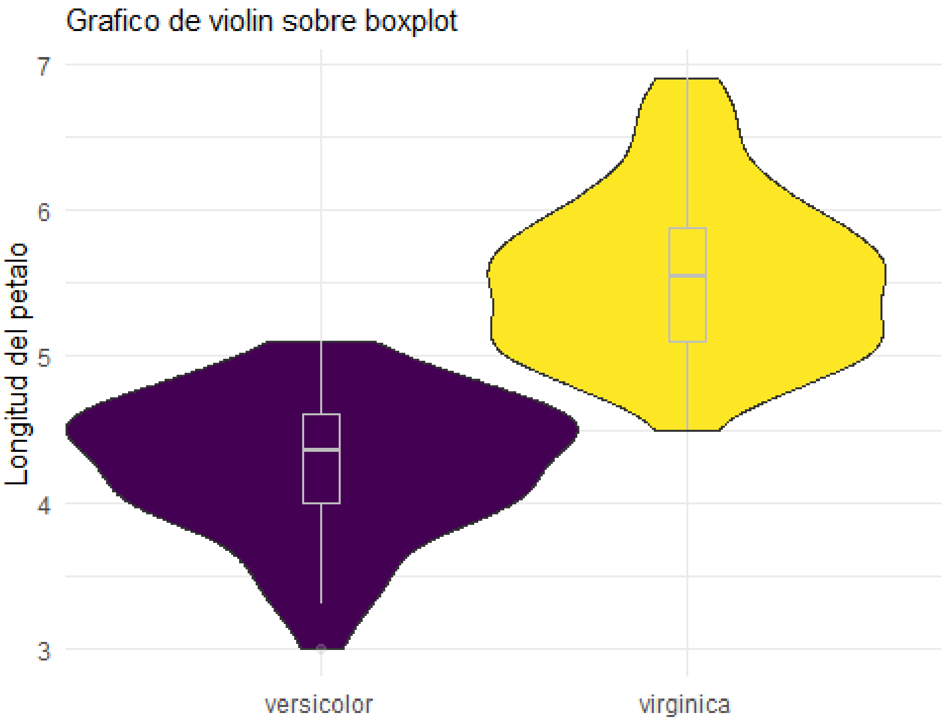
```
boxplot(df_versicolor$petal_length, df_virginica$petal_length,
        names = c("Versicolor","Virginica"),
        col = color,
        main = "Distribucion del largo del petalo por especie",
        xlab = "Especie",
        ylab = "Largo del petalo (cm)")
```



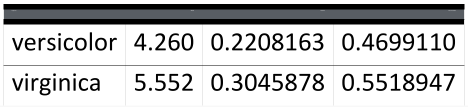
```
# Grafico de violin
ggplot(iris_sp, aes(x = species, y = petal_length))+
  geom_violin()
```



```
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("versicolor", "virginica"),
  Media = c(4.260, 5.552),
  Varianza = c(0.2208163, 0.3045878),
  Desv.Estandar = c(0.4699110, 0.5518947)
)
table_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```



versicolor	4.260	0.2208163	0.4699110
virginica	5.552	0.3045878	0.5518947

```r
mean_iris_sp <- data.frame(                # Media de ambas especies
  Especies = c("versicolor", "virginica"),
  Media = c(4.260, 5.552)
)
mean_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```

versicolor	4.260
virginica	5.552

```r
var_iris_sp <- data.frame(                 # Varianza de ambas especies
  Especies = c("versicolor", "virginica"),
  Varianza = c(0.2208163, 0.3045878)
)
var_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```

versicolor	0.2208163
virginica	0.3045878

```r
sd_iris_sp <- data.frame(                  # Desv.Est de ambas especies
  Especies = c("versicolor", "virginica"),
  Desv.Estandar = c(0.4699110, 0.5518947)
)
sd_iris_sp %>%
  gt() %>%
  gt_theme_pff()
```

versicolor	0.4699110
virginica	0.5518947