LAB.-5—BASE-DE-DATOS-IRIS-04-09-25.R

angel

2025-11-26

```
################################################################################
# LAB. 5 - BASE DE DATOS IRIS - TAREA HW_02
# FLOR ANGELI CRUZ ROSALES
# DR. MARCO A. GONZALEZ TAGLE
# 04/09/25
################################################################################

# La base de datos iris es uno de los conjuntos de datos más utilizados en
# estadística y aprendizaje automático

# El conjunto contiene 150 observaciones correspondientes a tres especies de iris
# (setosa, versicolor y irginica), con 50 muestras por especie.

# Para cada flor se registran cuatro variables cuantitativas:
# Sepal.Length: longitud del sépalo (cm)
# Sepal.Width: ancho del sépalo (cm)
# Petal.Length: longitud del pétalo (cm)
# Petal.Width: ancho del pétalo (cm)


# Explorar la base de datos iris usando funciones como head(), summary()
data("iris")
head(iris)
```

```
##   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
```

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
summary(iris)
```

```
##   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
# Identificar las variables Petal.Length y determina las estadísticas
# descriptivas para las dos especies

data_sub <- subset(iris, Species %in% c("versicolor", "virginica"))
head(data_sub)
```

```
##    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 51          7.0         3.2          4.7         1.4 versicolor
## 52          6.4         3.2          4.5         1.5 versicolor
## 53          6.9         3.1          4.9         1.5 versicolor
## 54          5.5         2.3          4.0         1.3 versicolor
## 55          6.5         2.8          4.6         1.5 versicolor
## 56          5.7         2.8          4.5         1.3 versicolor
```

```r
table(data_sub$Species)
```

```
##
##     setosa versicolor  virginica
##          0         50         50
```

```r
# Función tapply para obtener promedio
tapply(data_sub$Petal.Length, data_sub$Species, mean)
```

```
##     setosa versicolor  virginica
##         NA      4.260      5.552
```

```r
# Setosa versicolor  virginica
#  NA      4.260      5.552
# Los pétalos de virginica son más largos  1.3cm más

# Función tapply para obtener desviación estándar
tapply(data_sub$Petal.Length, data_sub$Species, sd)
```

```
##     setosa versicolor  virginica
##         NA  0.4699110  0.5518947
```

```
# Setosa versicolor  virginica
#    NA 0.4699110  0.5518947


# Función tapply para obtener varianza
tapply(data_sub$Petal.Length, data_sub$Species, var)
```

```
##     setosa versicolor  virginica
##         NA  0.2208163  0.3045878
```

```
# Setosa versicolor  virginica
#   NA  0.2208163  0.3045878
# Mayor variabilidad en la especie virginica

summary(data_sub$Petal.Length)
```

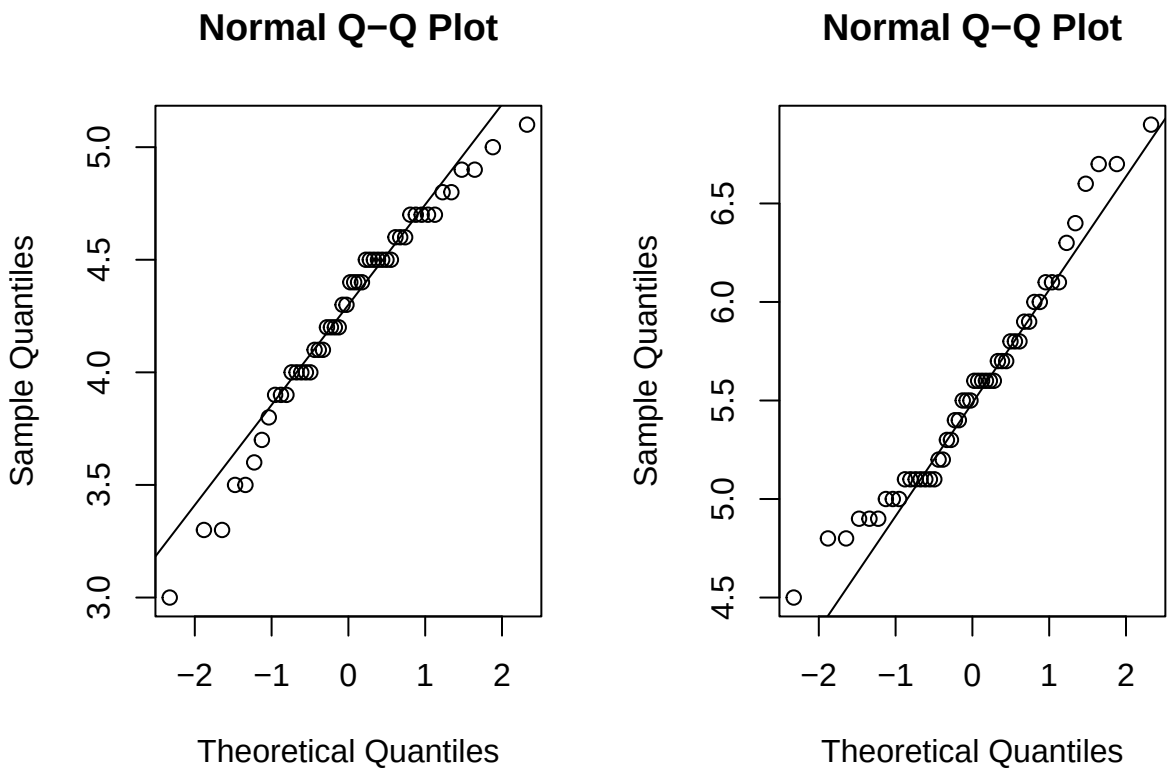
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.375   4.900   4.906   5.525   6.900
```

```
tapply(data_sub$Petal.Length, data_sub$Species, summary)
```

```
## $setosa
## NULL
##
## $versicolor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    4.00    4.35    4.26    4.60    5.10
##
## $virginica
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.500   5.100   5.550   5.552   5.875   6.900
```

```
# Defina una pregunta de investigación sobre la variable Petal.Length
# Pregunta: ¿Hay una diferencia en la longitud de los pétalos entre las especies versicolor y virginica

# Si hay una diferencia en la longitud de los pétalos, para obtener información
# más confiable si realizamos una prueba de t student


# Separar los datos por especie usando subset

df_versicolor <- subset(iris, Species == "versicolor")
df_virginica<- subset(iris, Species == "virginica")


# qqnorm  es un gráfico que nos ayudará a revisar la normalidad de los datos

# Función para que la ventana de gráficas permita que nos
# aparezca dos gráficos par(mfrow)
```

```r
# Una fila con dos columnas (1,2)

par(mfrow=c(1,2))
qqnorm(df_versicolor$Petal.Length); qqline(df_versicolor$Petal.Length)
qqnorm(df_virginica$Petal.Length); qqline(df_virginica$Petal.Length)
```




```r
# El gráfico muestra que una normalidad en nuestros datos ya que
# en la mayoría los puntos estan muy cercanos a la línea recta

# Pero podemos verificarlo de manera más precisa realizando una prueba de Shapiro-Wilkins

shapiro.test(df_versicolor$Petal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_versicolor$Petal.Length
## W = 0.966, p-value = 0.1585
```

```r
# 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
```

```
# data:  df_virginica$Petal.Length
# W = 0.96219, p-value = 0.1098

# *Los resultados son mayores a p-value 0.05,por lo que se acepta la normalidad*
# Cumple con uno de los tres criterios para realizar la prueba de t student



# Revisar homogeneidad de varianzas (segundo criterio)
var.test(Petal.Length ~ Species, data= subset(iris, Species %in% c("versicolor", "virginica")))
```

```
##
##  F test to compare two variances
##
## data:  Petal.Length by Species
## 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
```

```
# El resultado es un p-value= 0.2637 lo que nos indica un valor mayor
# a p-value=0.05, por lo que no se rechaza mi hipótesis nula
# Hay homogeneidad de varianzas pues se toman como iguales

# Se cumplen con los 3 criterios necesarios para realizar una prueba de t-student

iris_sub <- subset(iris,Species %in% c("versicolor", "virginica"))
t.test(Petal.Length ~ Species, data= iris_sub,
       var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Petal.Length by Species
## t = -12.604, df = 98, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group versicolor and group virginica is not
## 95 percent confidence interval:
##  -1.495426 -1.088574
## sample estimates:
## mean in group versicolor  mean in group virginica
##                    4.260                    5.552
```

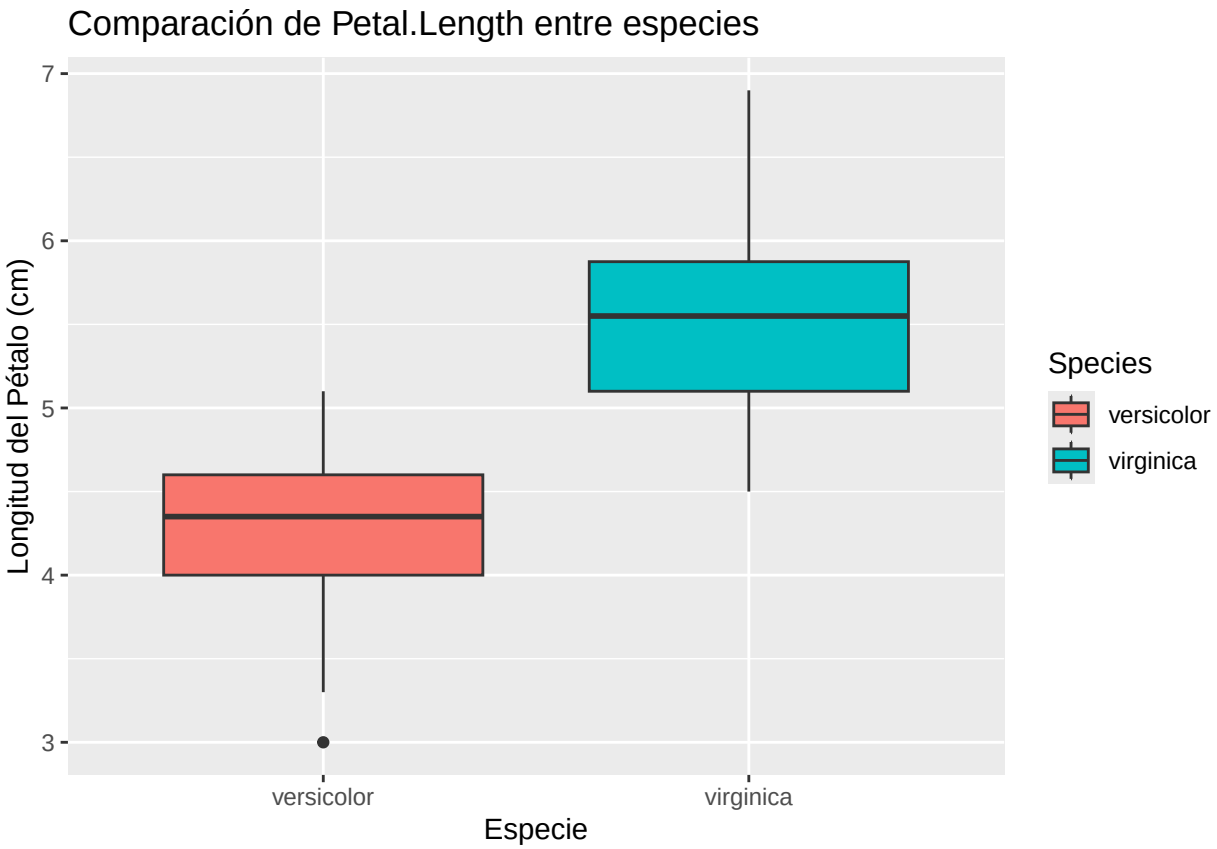
```r
# Resultados= t = -12.604, valor negativo que nos indica que hay una diferencia grande
# entre las medias
# p-value < 2.2e-16 , valor pequeño y menor a p-value= 0.05 por lo que
# se rechaza la hipótesis nula (H0)


# Intervalo de confianza
# -1.495426 -1.088574

# Diferencia de las medias: 4.260 - 5.552
#[1]-1.292

# La diferencia esta fuera del intervalo de confianza, lo que
# nos confirma que es una hipótesis alternativa
# Si no entra en el intervalo de confianza es H1 (alternativa)
# Si esta dentro del intervalo es H0 (nula)


# Se puede afirmar que los pétalos de la especie virginica son más largos
# que los de la especie versicolor
# Se acepta la hipótesis alternativa que nos indica que si existe diferencia en la
# longitud de los pétalos entre las especies versicolor y virginica

# Se determina con el Tamaño de efecto (Cohen´s)

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_virginica$Petal.Length, df_versicolor$Petal.Length)
cohens_efecto(df_virginica$Petal.Length, df_versicolor$Petal.Length)
```

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

```r
# 2.520756 es mayor al rango de 1.3, por lo que se puede deducir que la
# diferencia en la longitud del pétalo es estadisticamente significativa
# y relevante


#Visualización: boxplot que muestra las diferencias entre especies

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```r
ggplot(iris_sub, aes(x = Species, y = Petal.Length, fill = Species)) +
  geom_boxplot() +
  labs(title = "Comparación de Petal.Length entre especies",
       x = "Especie",
       y = "Longitud del Pétalo (cm)")
```


Comparación de Petal.Length entre especies
Longitud del Pétalo (cm)
7
6
5
4
3
versicolor
virginica
Especie
Species
versicolor
virginica