Análisis exploratorios

Santos G

Tabla de contenidos


```
# Librerías
library(tidyverse)   # Manipulación de datos: dplyr, tidyr, readr
library(janitor)     # Limpieza: clean_names(), tabyl()
library(ggplot2)     # Gráficos profesionales
library(skimr)       # EDA rápido y completo (skim())
library(GGally)      # Matriz de gráficos para variables múltiples
library(car)         # Homocedasticidad (Levene)
library(MVN)         # Normalidad multivariada
library(robustbase)  # Detección de outliers
library(knitr)       # Tablas en Quarto
library(kableExtra)  # Tablas formateadas para informes
```

1 Contexto general del proyecto

El presente proyecto corresponde a un análisis exploratorio de datos morfométricos de tres especies del género *Iris* (*Iris setosa*, *Iris versicolor* e *Iris virginica*). Este conjunto de datos, ampliamente utilizado en estadística y aprendizaje automático, contiene mediciones de longitud y anchura de sépalos y pétalos en un total de 150 individuos (50 por especie).

El objetivo principal del análisis es describir y comparar la variación morfológica entre especies, evaluando qué rasgos vegetativos (sépalos) y reproductivos (pétalos) permiten una mejor discriminación taxonómica.

El trabajo se estructura en tres ejes:

1. **Análisis descriptivo univariado y multivariado**: resumen numérico, visualizaciones (diagramas de dispersión, histogramas, gráficos de caja) y medidas de correlación.

2. **Evaluación de supuestos estadísticos**: normalidad univariada y multivariada, homocedasticidad, detección de valores atípicos. Esto permite fundamentar el uso de correlaciones paramétricas (Pearson) o no paramétricas (Spearman).

3. **Interpretación ecológica de resultados**: se discute el valor discriminante de cada variable morfométrica, así como el significado biológico de los patrones encontrados.

2 Carga y verificación inicial de datos

```
#|label: data-load

# Carga de datos (ejemplo iris) y limpieza mínima
data("iris")
df <- as_tibble(iris) %>%
  janitor::clean_names()   # convierte a snake_case: sepal_length, etc.

# Información básica
n_rows <- nrow(df); n_cols <- ncol(df)
tbl1<-skim(df)
tbl1 # Resumen compacto por variable
```

Tabla 1: Data summary

Name	df
Number of rows	150
Number of columns	5
_____	
Column type frequency:	
factor	1
numeric	4
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1	FALSE	3	set: 50, ver: 50, vir: 50

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
sepal_length	0	1	5.84	0.83	4.3	5.1	5.80	6.4	7.9	□□□□□
sepal_width	0	1	3.06	0.44	2.0	2.8	3.00	3.3	4.4	□□□□□
petal_length	0	1	3.76	1.77	1.0	1.6	4.35	5.1	6.9	□□□□□
petal_width	0	1	1.20	0.76	0.1	0.3	1.30	1.8	2.5	□□□□□

El dataset contiene **N = 150 observaciones** y **5 variables**. Cuatro son cuantitativas continuas en centímetros (*Sepal.Length, Sepal.Width, Petal.Length, Petal.Width*), y una categórica (*Species*), que clasifica en tres grupos balanceados (n = 50 por especie). No se detectaron valores faltantes ni duplicados tras la inspección inicial. Esta estructura balanceada y sin NA permite aplicar análisis univariados, comparativos y multivariados con mínimo preprocesamiento.

La **Tabla 1** de estadísticos descriptivos muestra lo siguiente:

- **Sepal.Length:** media ≈ 5.84 cm, SD ≈ 0.83, rango 4.3–7.9. Variación moderada, con solapamiento esperado entre especies.
- **Sepal.Width:** media ≈ 3.06 cm, SD ≈ 0.44, rango 2.0–4.4. Es la variable más estable, aunque con ligera asimetría negativa.
- **Petal.Length:** media ≈ 3.76 cm, SD ≈ 1.77, rango 1.0–6.9. Mayor dispersión relativa, con clara separación de *setosa*.
- **Petal.Width:** media ≈ 1.20 cm, SD ≈ 0.76, rango 0.1–2.5. Alta variabilidad, con potencial de discriminación entre las tres especies.

Aspectos destacados del dataset:

- **Escala homogénea de medidas:** todas las variables en centímetros → comparaciones y análisis multivariados sin necesidad de reescalado inmediato.
- **Colinealidad esperada:** Petal.Length y Petal.Width muestran alta correlación, lo que debe considerarse en regresiones o PCA.
- **Grupos biológicos claros y balanceados:** un escenario ideal para aprendizaje, aunque poco frecuente en estudios ecológicos reales.
- **Potencial de discriminación:** las variables de pétalos concentran el mayor poder de separación, coherente con su relevancia funcional en la biología reproductiva de las plantas.

3 Matriz de correlaciones y distribuciones entre variables numéricas

Las **Figuras 1 y 2** combina tres tipos de información: distribuciones univariadas, relaciones bivariadas y correlacciones númericas.

```
# Correlación paramétrica (Pearson en ggpairs)
num_df <- df %>% select(where(is.numeric))
Fig1<- GGally::ggpairs(
  df,
  columns = 1:4, # solo variables numéricas
  mapping = aes(color = species), # color por especie
  upper = list(continuous = wrap("cor", method = "pearson", size = 3)),
  diag = list(continuous = wrap("densityDiag", alpha = 0.6))
)
Fig1
```

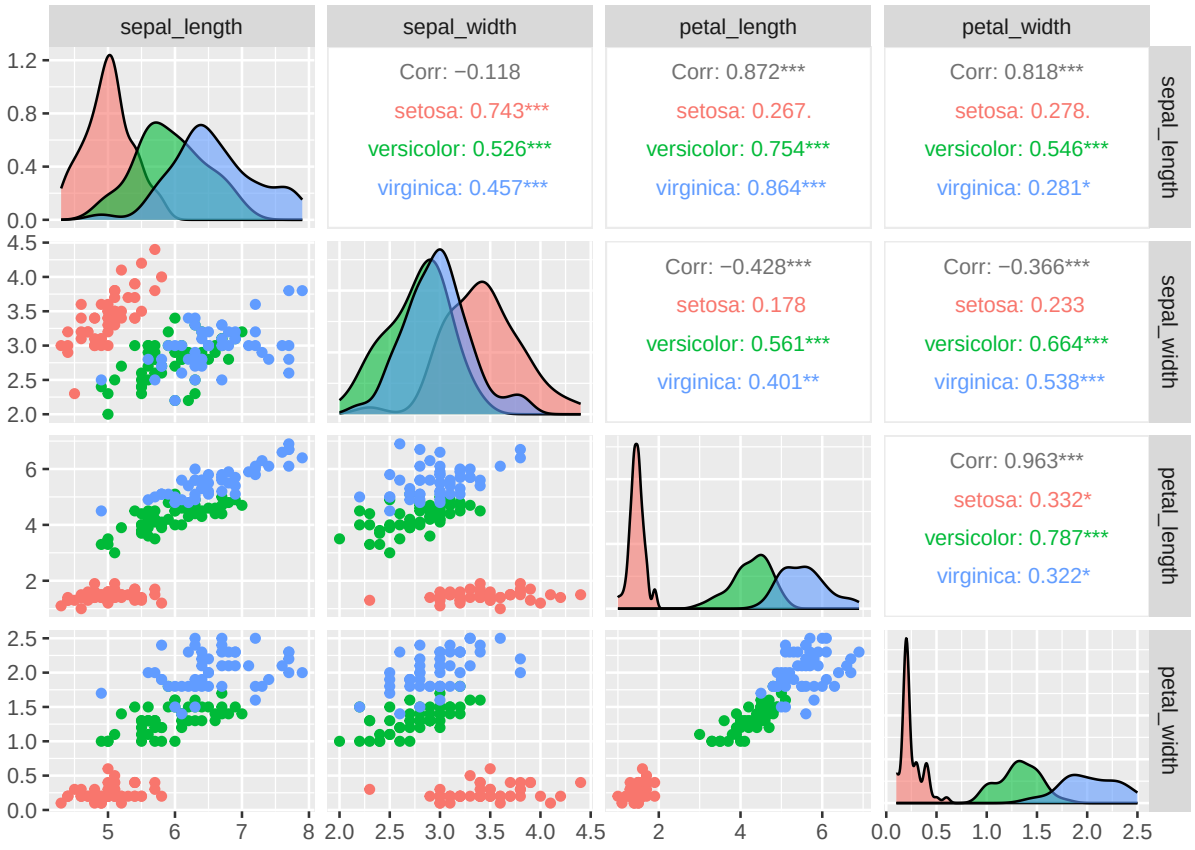

Figura 1: Matriz de dispersión y correlación de las variables cuantitativas (Pearson).

Dado que no se cumple normalidad, también se empleó el método de Spearman para estimar las correlaciones.

```
# Correlación no paramétrica (Spearman en ggpairs)
Fig2 <- ggpairs(
  df,
  columns = 1:4,
```

```
  mapping = aes(color = species),
  upper = list(continuous = wrap("cor", method = "spearman", size = 3)),
  diag = list(continuous = wrap("densityDiag", alpha = 0.6))
)
Fig2
```

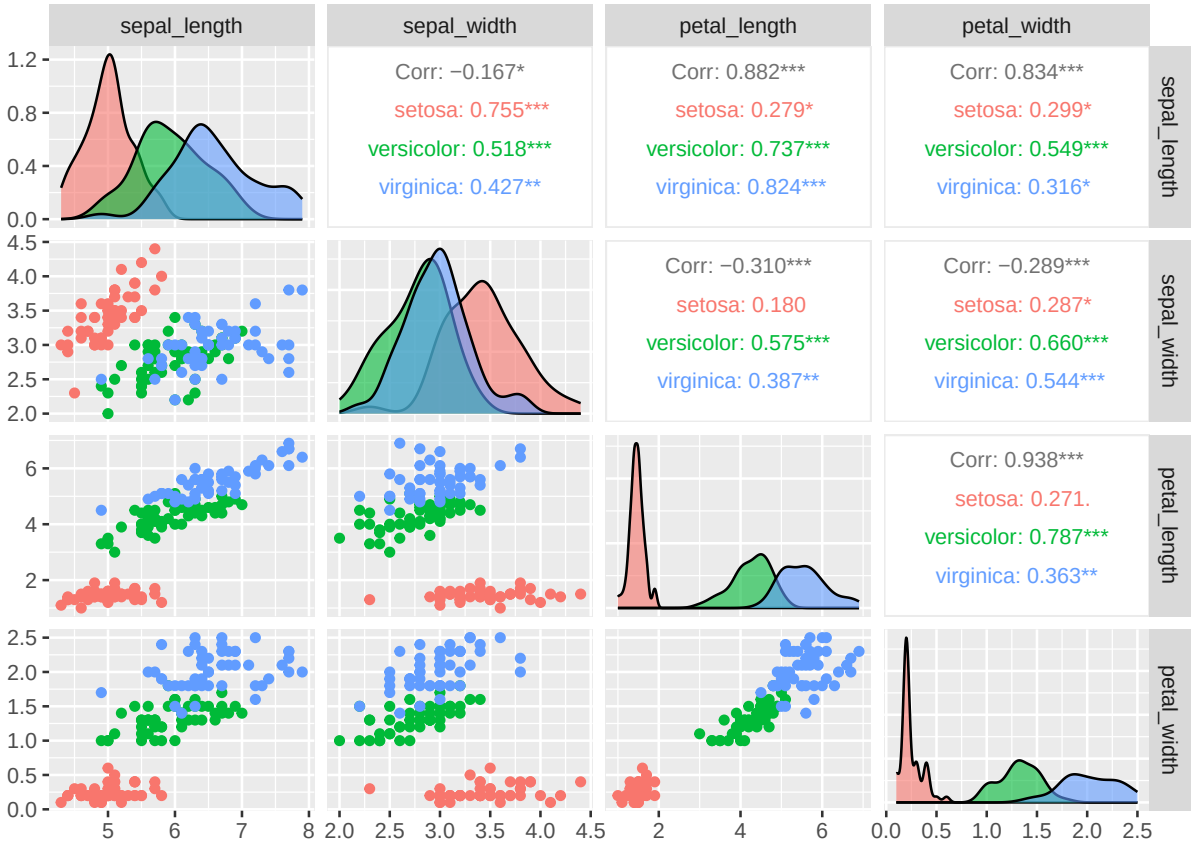

Figura 2: Matriz de dispersión y correlación de las variables cuantitativas (Spearman).

3.1 Distribuciones univariadas

- **Sepal.Length:**
  - *Setosa* concentrada en valores bajos (4.3–5.8 cm), muy homogénea.
  - *Versicolor* rango intermedio (≈ 4.9–7.0 cm).
  - *Virginica* valores altos (≈ 4.9–7.9 cm), con ligera superposición con Versicolor.

  Interpretación: útil para separar Setosa, pero Versicolor y Virginica se solapan.
- **Sepal.Width:**
  - Distribución amplia en todas las especies.
  - *Setosa* tiende a valores promedio mayores, pero con solapamiento considerable.

  Interpretación: bajo poder discriminante, refleja variabilidad natural.

- **Petal.Length:**
  - *Setosa* muy bajos (≈ 1.0–1.9 cm), sin solapamiento con otras especies.
  - *Versicolor* rango medio (≈ 3.0–5.1 cm).
  - *Virginica* altos (≈ 4.5–6.9 cm).

  Interpretación: variable clave, separa Setosa y discrimina relativamente bien Versicolor vs Virginica.
- **Petal.Width:**
  - *Setosa* muy bajos (≈ 0.1–0.6 cm).
  - *Versicolor* rango medio (≈ 1.0–1.8 cm).
  - *Virginica* altos (≈ 1.4–2.5 cm).

  Interpretación: la más robusta para separar las tres especies, casi sin solapamiento.

3.2 Relaciones bivariadas

- **Sepal.Length vs Sepal.Width:** nubes muy mezcladas → baja capacidad de discriminación.
- **Sepal.Length vs Petal.Length:** relación positiva moderada; Setosa queda aislada por pétalos cortos.
- **Sepal.Length vs Petal.Width:** tendencia positiva clara; ayuda a diferenciar más que Sepal.Length solo, aunque con solapamientos.
- **Sepal.Width vs Petal.Length / Petal.Width:** relaciones débiles, confirman poco valor discriminante del sépalo.
- **Petal.Length vs Petal.Width:** relación lineal muy fuerte, tres grupos claramente separados → la mejor combinación para clasificar especies.

3.3 Correlaciones numéricas

Dado que no se cumple normalidad, se utilizaron correlaciones de Spearman. Los resultados fueron muy consistentes con Pearson, mostrando que las conclusiones son robustas:

- **Petal.Length vs Petal.Width:** $\rho \approx 0.93$ (Spearman) / $r \approx 0.96$ (Pearson). Correlación extremadamente alta; variables casi redundantes, pero en conjunto definen un espacio morfológico clave.
- **Sepal.Length vs Petal.Length:** $\rho \approx 0.88$ / $r \approx 0.87$. Relación fuerte, ambas reflejan gradiente de tamaño.
- **Sepal.Length vs Petal.Width:** $\rho \approx 0.83$ / $r \approx 0.82$. Correlación alta, consistente con patrón de tamaño floral.

- **Sepal.Width con el resto:** ρ entre -0.1 y -0.3 (similares a Pearson). Confirma su escaso poder discriminante.

3.4 Interpretación general

- Los **pétalos** son rasgos reproductivos clave: su longitud y anchura diferencian a las especies porque están ligados a la atracción de polinizadores.
- Los **sépalos**, en cambio, son estructuras más plásticas y menos específicas, lo que explica su bajo poder discriminante.
- La fuerte correlación entre variables de pétalo refleja que ambas describen el mismo fenómeno biológico (tamaño floral), pero su combinación refuerza la clasificación.

La comparación entre Spearman y Pearson muestra que, aunque el supuesto de normalidad no se cumple, las correlaciones se mantienen prácticamente idénticas. Esto da confianza en la robustez del patrón biológico: la diferenciación entre especies de *Iris* depende más de rasgos reproductivos (pétalos) que de rasgos de soporte (sépalos).

4 Distribución de variables morfométricas entre especies

La **Figura 3** presenta diagramas de caja y bigotes que resumen la variación de las cuatro variables morfométricas en las tres especies de *Iris*. Este análisis complementa al resumen numérico y a la matriz de relaciones bivariadas, ya que enfatiza tendencias centrales, dispersión y presencia de datos atípicos.

```r
# Pasar el dataset a formato largo
iris_long <- df %>%
  pivot_longer(cols = -species,
               names_to = "Variable",
               values_to = "Valor")

# Gráfico unificado
Fig3 <- ggplot(iris_long, aes(x = species, y = Valor, fill = species)) +
  geom_boxplot(outlier.shape = 21, alpha = 0.7) +
  facet_wrap(~ Variable, scales = "free_y") +
  labs(
    title = "Comparación de variables morfométricas en especies de Iris",
    x = "Especies",
    y = "Valor (cm)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "none",
    strip.text = element_text(face = "bold")
  )
Fig3
```

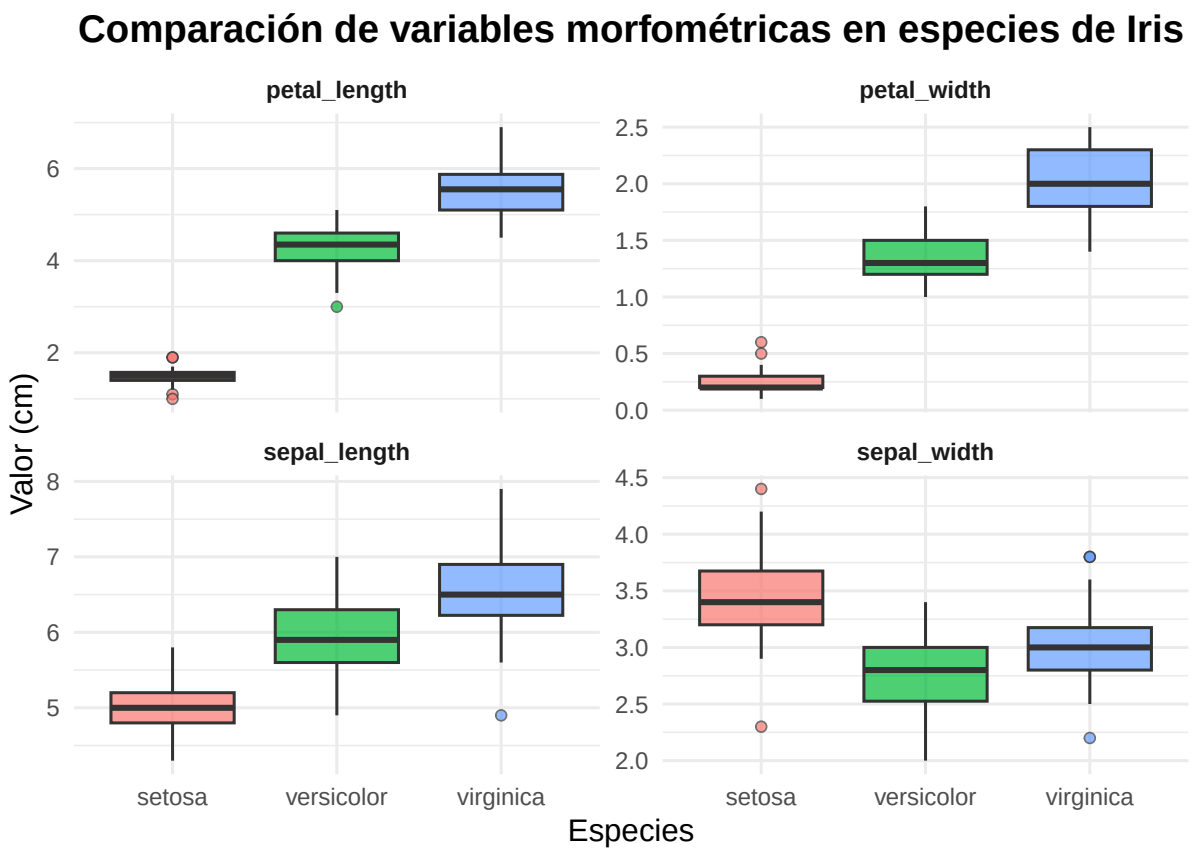

Figura 3: Distribución de variables morfométricas en tres especies de Iris.

4.1 Sepal.Length (longitud del sépalo)

- *I. setosa* muestra valores concentrados entre **4.3 y 5.8 cm**, con una mediana cercana a **5.0 cm**. La caja es compacta, indicando baja variabilidad.
- *I. versicolor* se distribuye entre **4.9 y 7.0 cm**, con mediana ≈ **5.9 cm**, rango intermedio.
- *I. virginica* alcanza los valores más altos (**4.9–7.9 cm**), con mediana ≈ **6.5 cm**.

Interpretación: hay cierto solapamiento entre *versicolor* y *virginica*. Útil para distinguir a *setosa*, pero no para discriminar con precisión entre *versicolor* y *virginica*.

4.2 Sepal.Width (anchura del sépalo)

- *I. setosa* tiene la mediana más alta (≈ **3.4 cm**), con valores entre **2.3 y 4.4 cm**.
- *I. versicolor* oscila entre **2.0 y 3.4 cm**, mediana ≈ **2.8 cm**.
- *I. virginica* se ubica entre **2.2 y 3.8 cm**, mediana ≈ **3.0 cm**.

Se observan varios outliers tanto en *setosa* como *virginica,* individuos con sépalos inusualmente estrechos. Estos valores atípicos podrían reflejar variación intraespecífica natural o condiciones ambientales particulares. La fuerte dispersión y el solapamiento reducen el valor discriminante de esta variable.

Interpretación: aunque setosa tiende a mayor anchura, la amplia dispersión y solapamiento hacen que esta variable tenga bajo poder discriminante.

4.3 Petal.Length (longitud del pétalo)

- *I. setosa* presenta valores muy bajos (**1.0–1.9 cm**) con mediana ≈ **1.5 cm**.
- *I. versicolor* ocupa un rango intermedio (**3.0–5.1 cm**), con mediana ≈ **4.3 cm**.
- *I. virginica* concentra los valores más altos (**4.5–6.9 cm**), mediana ≈ **5.5 cm**.

Interpretación: el solapamiento entre *versicolor* y *virginica* es reducido y se da en los límites superiores/inferiores de sus cajas. Esta variable es altamente informativa; separa completamente a *setosa* y discrimina en gran medida a *versicolor* y *virginica*.

4.4 Petal.Width (anchura del pétalo)

- *I. setosa* tiene los valores más bajos (**0.1–0.6 cm**), con mediana ≈ **0.2 cm**, sin solapamiento con las otras especies.
- *I. versicolor* se concentra entre **1.0 y 1.8 cm**, mediana ≈ **1.3 cm**.
- *I. virginica* presenta los valores más altos (**1.4–2.5 cm**), mediana ≈ **2.0 cm**.

Interpretación: junto con *Petal.Length*, constituye la variable más robusta para separar especies; su poder discriminante es muy alto y con solapamiento mínimo.

4.5 Interpretación general

Las variables de pétalo muestran diferencias netas entre especies, con cajas bien separadas. Por lo que se considera un rasgo clave para clasificación, debido a su alta capacidad de separar grupos. A diferencias de las variables de sépalo que presentan mayor dispersión y solapamiento, lo que limita su valor clasificatorio, debido a su menor capacidad diagnóstica, más influenciados por plasticidad ambiental.

5 Análisis de supuestos

Se realizó una evaluación de los supuestos básicos, como normalidad, homocedasticidad y detección de datos atípicos.

5.1 Normalidad (univariada y multivariada)

```r
# Normalidad (univariada y multivariada)
mvn_norm <- mvn(df[,1:4], mvn_test = "hz")

# Resultados
mvn_norm$univariate_normality
```

```
             Test     Variable Statistic p.value    Normality
1 Anderson-Darling sepal_length     0.889   0.023   Not normal
2 Anderson-Darling  sepal_width     0.908    0.02   Not normal
3 Anderson-Darling petal_length     7.679  <0.001   Not normal
4 Anderson-Darling  petal_width     5.106  <0.001   Not normal
```

```r
mvn_norm$multivariate_normality
```

```
           Test Statistic p.value     Method          MVN
1 Henze-Zirkler     2.336  <0.001 asymptotic   Not normal
```

La prueba de Henze–Zirkler indicó que las variables en conjunto no siguen una distribución normal multivariada (**HZ = 2.336, p < 0.001**). A nivel univariado, el test de Anderson–Darling mostró que ninguna de las cuatro variables es normal:

- *Sepal.Length*: **p = 0.023**
- *Sepal.Width*: **p = 0.020**
- *Petal.Length*: **p < 0.001**
- *Petal.Width*: **p < 0.001**

Interpretación: El conjunto de datos no cumple con la suposición de normalidad, por lo que deben preferirse pruebas no paramétricas (ej. Kruskal-Wallis en lugar de ANOVA) y coeficientes de correlación de Spearman en lugar de Pearson.

5.2 Homocedasticidad

```r
# Homocedasticidad (igualdad de varianzas por especie)
levene_sepal_length  <- leveneTest(sepal_length ~ species, data = df)
levene_sepal_width   <- leveneTest(sepal_width ~ species, data = df)
levene_petal_length  <- leveneTest(petal_length ~ species, data = df)
levene_petal_width   <- leveneTest(petal_width ~ species, data = df)

# Resultados
levene_sepal_length
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   2  6.3527 0.002259 **
      147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
levene_sepal_width
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.5902 0.5555
      147
```

```
levene_petal_length
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2   19.48 3.129e-08 ***
      147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
levene_petal_width
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   2  19.892 2.261e-08 ***
      147
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El test de Levene mostró resultados mixtos:

- *Sepal.Length*: **p = 0.002** → varianzas heterogéneas.
- *Sepal.Width*: **p = 0.556** → varianzas homogéneas.
- *Petal.Length*: **p < 0.001** → varianzas heterogéneas.
- *Petal.Width*: **p < 0.001** → varianzas heterogéneas.

Interpretación: Solo Sepal.Width cumple homogeneidad de varianzas. Las demás variables presentan heterocedasticidad, lo cual refuerza la necesidad de usar pruebas no paramétricas para comparar grupos.

5.3 Outliers multivariados

```r
# Outliers multivariados (Mahalanobis global y robusto)
X <- df[,1:4]

center_global <- colMeans(X)
cov_global <- cov(X)
d2_global <- mahalanobis(X, center_global, cov_global)
threshold_global <- qchisq(0.975, df = ncol(X))

mcd <- covMcd(X) # robusto
d2_robust <- mahalanobis(X, center = mcd$center, cov = mcd$cov)
threshold_robust <- qchisq(0.975, df = ncol(X))

```

```r
df_out <- df %>%
  mutate(
    d2_global = d2_global,
    is_out_global = d2_global > threshold_global,
    d2_robust = d2_robust,
    is_out_robust = d2_robust > threshold_robust
  )

table(df_out$is_out_global)
```

```
FALSE  TRUE
  144     6
```

```r
table(df_out$is_out_robust)
```

```
FALSE  TRUE
   95    55
```

Se evaluaron outliers multivariados con la distancia de Mahalanobis:

- **Mahalanobis global (media/covarianza clásica):** detectó **6 observaciones atípicas** de 150 (4%).
- **Mahalanobis robusto (MCD):** detectó **55 observaciones atípicas** (37%).

Interpretación: El método robusto es más estricto y revela que una parte considerable de las observaciones no se ajusta al patrón multivariado esperado. Esto indica que el dataset, aunque muy usado como ejemplo didáctico, presenta estructuras internas de alta variabilidad (particularmente en Versicolor y Virginica), lo cual podría influir en análisis sensibles a outliers (ej. PCA, discriminante).
En un reporte real, se recomienda analizar los outliers para verificar si corresponden a errores de medición, variabilidad biológica genuina o presencia de subgrupos dentro de las especies.

5.4 Interpretación general

Los supuestos paramétricos clásicos (normalidad y homocedasticidad) no se cumplen plenamente en este conjunto de datos. Por ello, se optó por:

- Usar correlaciones de Spearman en lugar de Pearson.
- Considerar técnicas robustas o no paramétricas en análisis comparativos posteriores.
- Los outliers no deben eliminarse automáticamente, sino interpretarse en su contexto biológico/ecológico (p. ej. plasticidad intraespecífica en Iris).