Índices ecológicos

Santos G

Tabla de contenidos

1 Contexto del proyecto **1**

2 Limpieza y revisión inicial de los datos **1**

3 Construcción de la matriz de abundancia **7**

4 Calcular índices de diversidad y curvas de rarefacción **8**

5 Conclusiones generales **10**

1 Contexto del proyecto

El presente documento guía describe el dataset palmerpenguins (mediciones morfométricas y metadatos de tres especies de pingüinos: *Adelie*, *Chinstrap* y *Gentoo*). Antes de calcular índices ecológicos o métricas de biodiversidad, es fundamental evaluar la calidad y consistencia de los datos. Para ello se realizará la limpieza y revisión inicial de los datos (nombres, estructura, NA, duplicados, categorías).

2 Limpieza y revisión inicial de los datos

```
#|label: prep
#Librerías
library(tidyverse)
library(janitor)
library(skimr)
library(palmerpenguins)

df_raw <- penguins %>% as_tibble() # guardo raw para auditoría
```

```
#|label: clean_names
df <- df_raw %>% clean_names()
names(df) # comprobar
```

```
[1] "species"           "island"            "bill_length_mm"
[4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
[7] "sex"               "year"
```

```
#|label: skim
skim(df) #Resumen rápido (estructura + NA)
```

Tabla 1: Data summary

Name	df
Number of rows	344
Number of columns	8
_____	
Column type frequency:	
factor	3
numeric	5
_____	
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	0	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	11	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bill_length_mm	2	0.99	43.92	5.46	32.1	39.23	44.45	48.5	59.6	□□□□□
bill_depth_mm	2	0.99	17.15	1.97	13.1	15.60	17.30	18.7	21.5	□□□□□
flipper_length_mm	2	0.99	200.92	14.06	172.0	190.00	197.00	213.0	231.0	□□□□□
body_mass_g	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.0	6300.0	□□□□□
year	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.0	2009.0	□□□□□

De acuerdo con la **Tabla 1**, la base de datos `penguins` contiene 344 registros y 8 variables, de las cuales 3 son categóricas (`species`, `island`, `sex`) y 5 numéricas (`bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`, `year`).

En las variables categóricas:

- `species`: tres categorías (*Adelie* = 152, *Gentoo* = 124, *Chinstrap* = 68), sin valores faltantes.
- `island`: tres categorías (*Biscoe* = 168, *Dream* = 124, *Torgersen* = 52), sin valores faltantes.
- `sex`: dos categorías (*male* = 168, *female* = 165) con 11 valores faltantes (3%).

En las variables numéricas:

- `bill_length_mm`: media ≈ 43.9 mm, rango 32.1–59.6 mm, con 2 valores faltantes.
- `bill_depth_mm`: media ≈ 17.2 mm, rango 13.1–21.5 mm, con 2 faltantes.
- `flipper_length_mm`: media ≈ 200.9 mm, rango 172–231 mm, con 2 faltantes.

- `body_mass_g`: media ≈ 4201 g, rango 2700–6300 g, con 2 faltantes.
- `year`: muestreos entre 2007–2009, sin valores faltantes.

Interpretación general:

- Los nombres de variables ya están estandarizados.
- El dataset presenta un bajo porcentaje de NA (<1% en mediciones y 3% en `sex`). Estos casos podrán eliminarse o imputarse según el análisis.
- No se observan inconsistencias de escritura en categorías ni rangos numéricos irreales.

```
#|label: duplicates
n_dup <- sum(duplicated(df))
n_dup # Número de filas duplicadas
```

```
[1] 0
```

```
# si quieres, ver filas duplicadas (opcional)
df %>% filter(duplicated(.)) %>% head()
```

```
# A tibble: 0 x 8
# i 8 variables: species <fct>, island <fct>, bill_length_mm <dbl>,
#   bill_depth_mm <dbl>, flipper_length_mm <int>, body_mass_g <int>, sex <fct>,
#   year <int>
```

```
# eliminar duplicados exactos (opcional)
df <- df %>% distinct()
```

En este caso no se encontraron filas duplicadas (`n_dup = 0`).

Interpretación general:

- Cuando no hay duplicados, no se requieren cambios.
- Si en futuros proyectos aparecen duplicados, se recomienda, verificar primero si son errores de registro o réplicas biológicas válidas, ya que si son errores (mismo individuo registrado más de una vez), deben eliminarse, caso contrario deben mantenerse o promediarse según el objetivo del estudio.

```
#Variables categóricas
df %>% summarise(across(where(is.character), n_distinct))
```

```
# A tibble: 1 x 0
```

```
df %>% summarise(across(where(is.factor), n_distinct))
```

```
# A tibble: 1 x 3
  species island   sex
    <int>  <int> <int>
1       3      3     3
```

```r
df %>% count(species)
```

```
# A tibble: 3 x 2
  species       n
  <fct>     <int>
1 Adelie      152
2 Chinstrap    68
3 Gentoo      124
```

```r
df %>% count(island)
```

```
# A tibble: 3 x 2
  island        n
  <fct>     <int>
1 Biscoe      168
2 Dream       124
3 Torgersen    52
```

```r
df %>% count(sex)
```

```
# A tibble: 3 x 2
  sex        n
  <fct>  <int>
1 female   165
2 male     168
3 <NA>      11
```

```r
#Tipos de variables (convertir a factor si hace falta)
df <- df %>%
  mutate(species = as.factor(species),
         island  = as.factor(island),
         sex     = as.factor(sex))

#Revisar posibles inconsistencias de texto
unique(df$species)
```

```
[1] Adelie    Gentoo    Chinstrap
Levels: Adelie Chinstrap Gentoo
```

```r
unique(df$island)
```

```
[1] Torgersen Biscoe    Dream
Levels: Biscoe Dream Torgersen
```

```r
unique(df$sex)
```

```
[1] male   female <NA>
Levels: female male
```

```r
df1 <- df %>% select(-year) #Eliminar columnas innecesarias
```

Las variables categóricas (`species`, `island`, `sex`) presentan 3, 3 y 2 categorías respectivamente, sin inconsistencias de escritura. Se ajustaron los tipos de variable a `factor` y se eliminó la columna `year`, ya que no será utilizada en los análisis posteriores.

```r
# Boxplots rápidos para ver rangos y "outliers" visuales
key_vars <- c("bill_length_mm","bill_depth_mm","flipper_length_mm",
              "body_mass_g")

df %>%
  select(all_of(key_vars)) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = variable, y = value)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "Boxplots rápidos: rangos y valores extremos",
       x = "", y = "Valor") +
  theme_minimal(base_size = 13) +
  theme(
   plot.title = element_text(hjust = 0.5, face = "bold"),
   legend.position = "none",
   strip.text = element_text(face = "bold")
   )
```

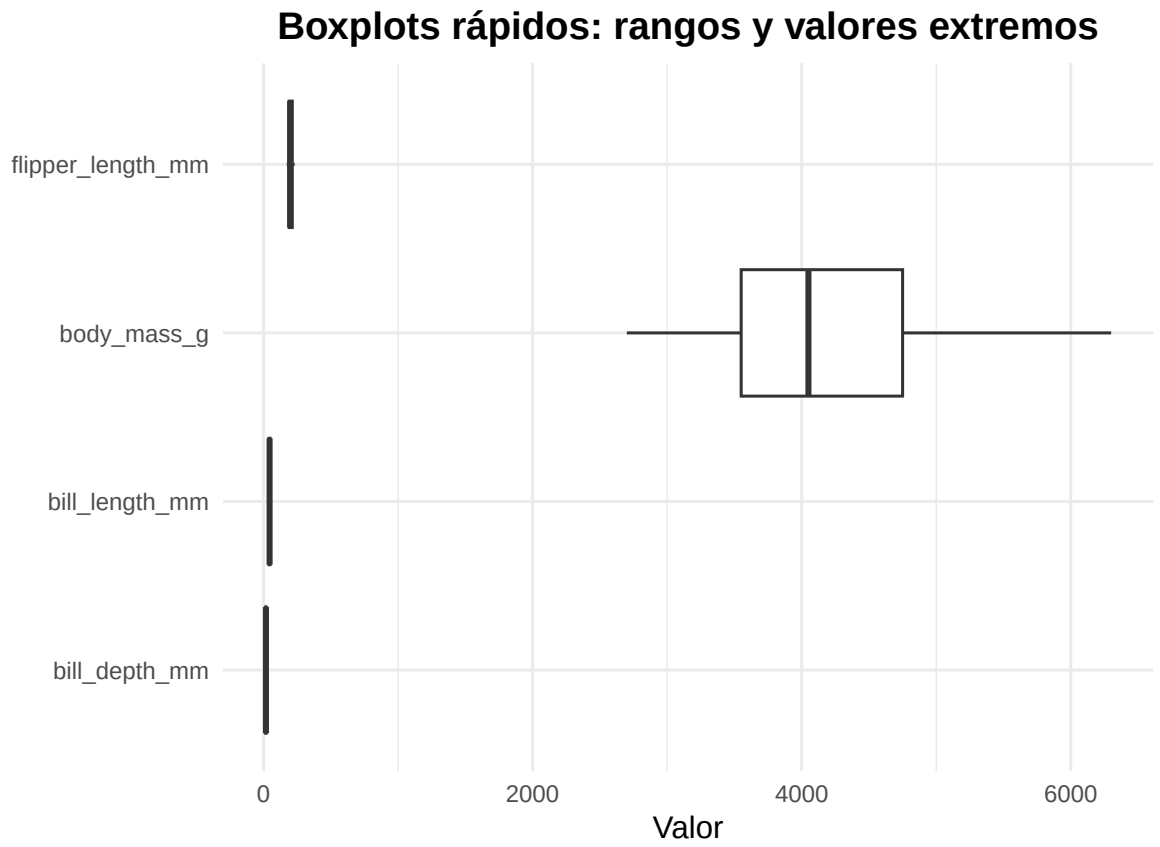

Figura 1: Rangos y valores extremos de variables morfométricas.

En la **Figura 1**, se observa que los rangos numéricos se encuentran dentro de lo esperado para las especies registradas, y los gráficos de caja confirman que los valores extremos corresponden a variabilidad natural más que a errores de registro.

```
#|label: Manejo de los NA

# Manejo de los NA
df <- df %>%
  mutate(
    n_na_numeric = rowSums(is.na(select(., bill_length_mm, bill_depth_mm,
                                        flipper_length_mm, body_mass_g)))
  )

df <- df %>%
  mutate(
    sex = case_when(
      is.na(sex) & n_na_numeric <= 1 ~ "Unknown",  # casi toda la info presente
      TRUE ~ as.character(sex)                     # dejar como está
    )
  )

df <- df %>%
  filter(!(is.na(bill_length_mm) &
           is.na(bill_depth_mm) &
```

```
            is.na(flipper_length_mm) &
            is.na(body_mass_g)))


df <- df %>% select(-n_na_numeric) #Eliminar columnas innecesarias
```

Se contabilizó cuántas mediciones morfométricas tiene cada registro (bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g).

- Registros con ≥ 3 mediciones y sex = NA fueron etiquetados como sex = "Unknown": son registros con información morfométrica suficiente como para conservarlos en análisis de biodiversidad/morfometría, pero sin identificación sexual.
- Registros con 0 mediciones (probablemente avistamientos sin mediciones) fueron eliminados, ya que no aportan datos morfométricos para los análisis previstos.

3 Construcción de la matriz de abundancia

Aunque `palmerpenguins` no es un dataset clásico de abundancias de especies (es más bien morfométrico), podemos adaptarlo, para ello tomamos las especies como categorías biológicas (3 especies: *Adelie*, *Gentoo*, *Chinstrap*) y contamos número de individuos por especie y por isla, para tener una matriz de abundancias que nos permita calcular índices de biodiversidad.

```
#Librerías
library(knitr)
library(kableExtra)
```

```
# Matriz especie x isla
abund <- df %>%
  count(island, species) %>%
  pivot_wider(names_from = species, values_from = n, values_fill = 0)

# Crear tabla
abund %>%
  kable(caption = "Abundancia de pingüinos por especie e isla
        en el archipiélago Palmer.",
        align = "lccc",
        col.names = c("Isla", "Adelie", "Gentoo", "Chinstrap")) %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped",
      "hover", "condensed"))
```

Tabla 4: Abundancia de pingüinos por especie e isla en el archipiélago Palmer.

Isla	Adelie	Gentoo	Chinstrap
Biscoe	44	123	0
Dream	56	0	68
Torgersen	51	0	0

En la **Tabla 4** se muestra la distribución de individuos por especie e isla:

- **Biscoe**: domina la especie *Gentoo* con 123 individuos, seguida por *Adelie* con 44. No se registran individuos en *Chinstrap*.
- **Dream**: comunidad más equilibrada, con 56 *Adelie* y 68 *Chinstrap*. No se registran individuos en *Gentoo*.
- **Torgersen**: exclusiva de *Adelie*, con 51 individuos; no se observan individuos de *Gentoo* ni *Chinstrap*.

Interpretación general: Cada isla presenta una composición particular. Mientras Biscoe está claramente dominada por *Gentoo*, Dream muestra coexistencia entre *Adelie* y *Chinstrap*, y Torgersen resulta la más restringida, con presencia exclusiva de *Adelie*. Esta variación espacial será clave al analizar los índices de diversidad y equidad.

4 Calcular índices de diversidad y curvas de rarefacción

- **Riqueza (S):** número de especies observadas. Valores enteros (1, 2,…).
- **Shannon (H')**: mide diversidad considerando riqueza y abundancia relativa. Valores cercanos a 0 indican baja diversidad (monoespecífico). Valores mayores indican mayor complejidad.
- **Simpson (D)**: en la salida de `vegan::diversity(..., "simpson")` se obtiene D (probabilidad de que dos individuos sean distintos). Valores mayores → más diversidad efectiva.
- **Equidad de Pielou (J')**: H' / ln(S). Rango 0–1. J' ≈ 1 indica que las especies están en proporciones muy similares (alta equidad). J' bajo indica dominancia de una especie. Para S = 1 definimos J' = 0 (no aplicable).
- **Curvas de rarefacción**: Las curvas de rarefacción muestran cómo aumenta la riqueza observada (n° especies) con el número de individuos muestreados. Permiten comparar riqueza entre lugares con distinto esfuerzo muestral, estandarizando por número de individuos. Si una curva se estabiliza (nivelación) significa que el muestreo fue suficiente para capturar la riqueza local. Si una curva sigue subiendo a mayor número de individuos, indica que con más muestreo probablemente se observarían más especies.

```r
#Librerías
library(vegan)
```

```r
# Preparar matriz numérica (filas = islas, columnas = especies)
mat <- abund %>% select(-island) %>% as.matrix()
rownames(mat) <- abund$island

# Cálculo de índices
shannon  <- diversity(mat, index = "shannon")          # H'
simpson  <- diversity(mat, index = "simpson")          # D (Simpson)
richness <- specnumber(mat)                            # S
# Equidad (Pielou) -> si S <= 1, definimos equidad = 0 para evitar NaN
evenness_raw <- ifelse(richness > 1, shannon / log(richness), 0)

indices <- tibble(
  Isla    = rownames(mat),
  Riqueza = richness,
  Shannon = round(shannon, 2),
```

```
  Simpson = round(simpson, 2),
  Equidad = round(evenness_raw, 2)
)

# Imprimir tabla
indices %>%
  kable(align = "lcccc") %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped",
      "hover","condensed"))
```

Tabla 5: Índices de diversidad por isla (Riqueza, Shannon, Simpson, Equidad de Pielou).

Isla	Riqueza	Shannon	Simpson	Equidad
Biscoe	2	0.58	0.39	0.83
Dream	2	0.69	0.50	0.99
Torgersen	1	0.00	0.00	0.00

En la **Tabla 5** se presentan los resultados obtenidos en los análisis de los índices de diversidad:

- **Biscoe (Riqueza = 2; H' = 0.58; S= 0.39; J' = 0.83):** diversidad moderada pero desequilibrada, ya que aunque tiene 2 especies , la abundancia está claramente dominada por *Gentoo* (123 individuos frente a 44 de *Adelie*).

- **Dream (Riqueza = 2; H' = 0.69; S= 0.50; J' = 0.99):** mayor diversidad efectiva y casi máxima equidad. Biológicamente, sugiere condiciones que permiten coexistencia equilibrada. Presenta dos especies con abundancias similares (*Adelie* 56, *Chinstrap* 68).

- **Torgersen (Riqueza = 1; H' = 0; S= 0; J' = 0):** comunidad monoespecífica y baja complejidad de la comunidad local. Solo *Adelie* registrada (51 individuos).

```
# Dibujar rarefacción
cols <- c("darkblue","darkgreen","darkred")
rarecurve(mat, step = 10, sample = min(rowSums(mat)), col = cols, label = TRUE,
          lwd = 2, ylab = "Riqueza de especies")
legend("bottomright", legend = rownames(mat), col = cols, lwd = 2, bty = "n")
```

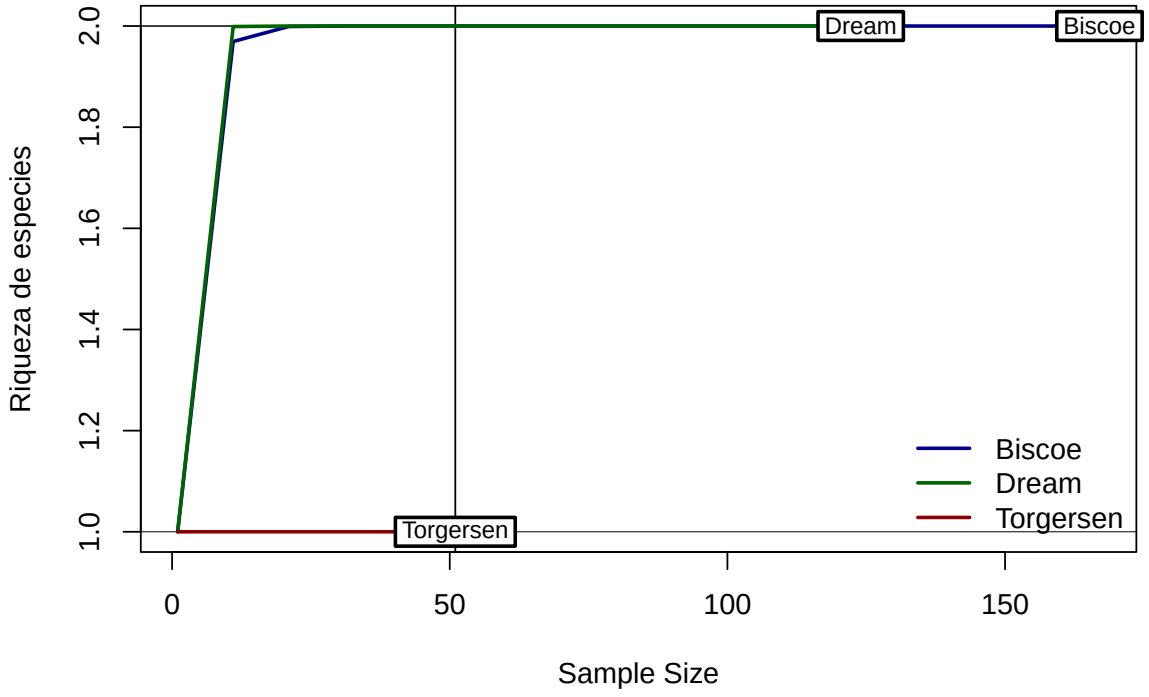

Figura 2: Curvas de rarefacción de riqueza de especies por isla.

En la **Figura 2** se presentan los resultados obtenidos en los análisis de las curvas de rarefacción de riqueza:

- Dream y Biscoe alcanzan un máximo de 2 especies en la curva; sus curvas se estabilizan, sugiriendo que el muestreo capturó la riqueza efectiva.
- Torgersen se estabiliza rápidamente en 1 especie, confirmando su carácter monoespecífico.
- Aun cuando Biscoe tiene más individuos totales , su riqueza no supera a Dream; la rarefacción muestra que la mayor abundancia no implica mayor riqueza.

5 Conclusiones generales

- Riqueza por isla es baja (1–2 especies), pero la distribución de abundancias (equidad) diferencia fuertemente la estructura comunitaria.
- Dream es la isla con mayor diversidad efectiva (H' = 0.69; J' = 0.99), por la coexistencia equilibrada de *Adelie* y *Chinstrap*.
- Biscoe tiene mayor número de individuos pero menor equidad por la dominancia de *Gentoo*, resultando en H' más bajo.
- Torgersen es monoespecífica (solo *Adelie*) y muestra diversidad mínima.
- Las curvas de rarefacción confirman que el muestreo capturó la riqueza observable en cada isla (curvas estabilizadas) y permiten comparar riqueza independientemente del tamaño muestral.