

# 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

## 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

```
1 #|label: prep
2 #Librerías
3 library(tidyverse)
4 library(janitor)
5 library(skimr)
6 library(palmerpenguins)
7
8 df_raw <- penguins %>% as_tibble() # guardo raw para auditoría
```

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

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

```

1 #|label: skim
2 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  $\approx 43.9$  mm, rango 32.1–59.6 mm, con 2 valores faltantes.
- *bill\_depth\_mm*: media  $\approx 17.2$  mm, rango 13.1–21.5 mm, con 2 faltantes.
- *flipper\_length\_mm*: media  $\approx 200.9$  mm, rango 172–231 mm, con 2 faltantes.

- `body_mass_g`: media  $\approx 4201$  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.

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

```
[1] 0
```

```
1 # si quieres, ver filas duplicadas (opcional)
2 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>
```

```
1 # eliminar duplicados exactos (opcional)
2 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.

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

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

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

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

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

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

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

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

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

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

```
1 #Tipos de variables (convertir a factor si hace falta)
2 df <- df %>%
3   mutate(species = as.factor(species),
4           island  = as.factor(island),
5           sex     = as.factor(sex))
6
7 #Revisar posibles inconsistencias de texto
8 unique(df$species)
```

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

```
1 unique(df$island)
```

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

```
1 unique(df$sex)
```

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

```
1 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.

```
1 # Boxplots rápidos para ver rangos y "outliers" visuales
2 key_vars <- c("bill_length_mm", "bill_depth_mm", "flipper_length_mm",
3             "body_mass_g")
4
5 df %>%
6   select(all_of(key_vars)) %>%
7   pivot_longer(everything(), names_to = "variable", values_to = "value") %>%
8   ggplot(aes(x = variable, y = value)) +
9   geom_boxplot() +
10  coord_flip() +
11  labs(title = "Boxplots rápidos: rangos y valores extremos",
12       x = "", y = "Valor") +
13  theme_minimal(base_size = 13) +
14  theme(
15    plot.title = element_text(hjust = 0.5, face = "bold"),
16    legend.position = "none",
17    strip.text = element_text(face = "bold")
18  )
```

## Boxplots rápidos: rangos y valores extremos

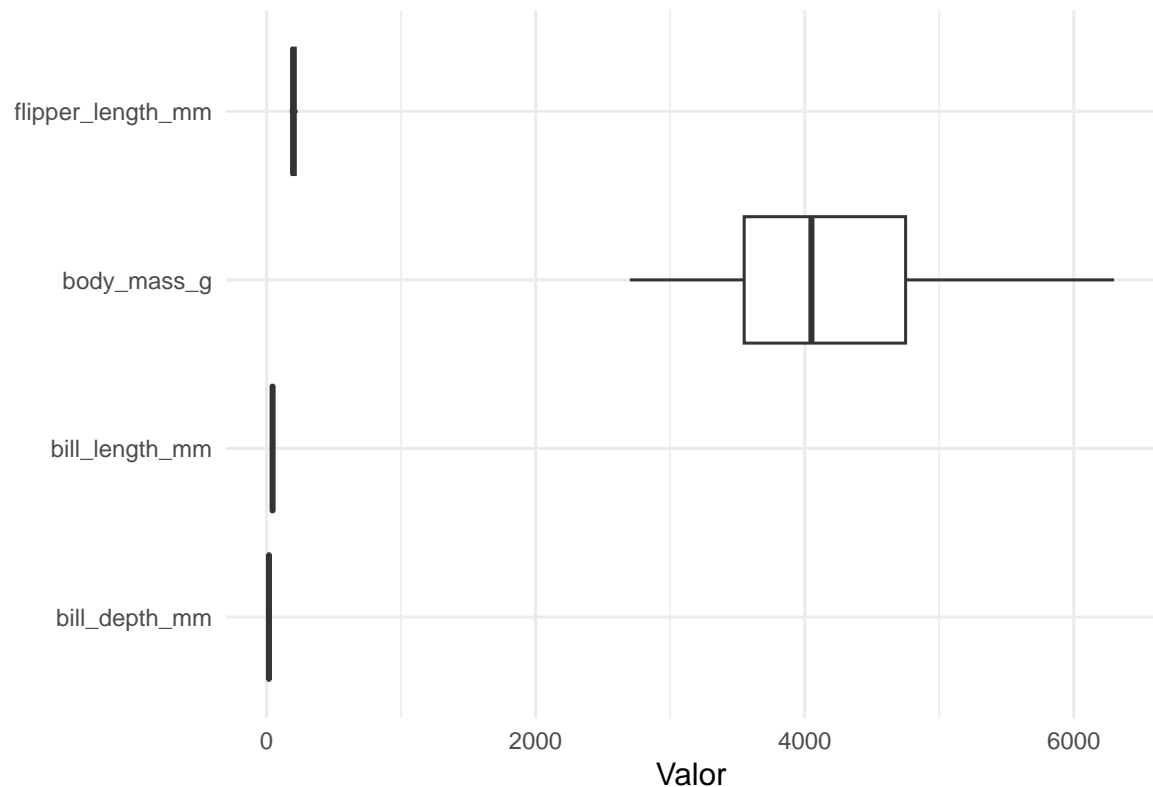


Figura 1: Rangos y valores extremos de variables morfométricas.

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.

```
1 #|label: Manejo de los NA
2
3 # Manejo de los NA
4 df <- df %>%
5   mutate(
6     n_na_numeric = rowSums(is.na(select(., bill_length_mm, bill_depth_mm,
7     flipper_length_mm, body_mass_g)))
8   )
9
10 df <- df %>%
11   mutate(
12     sex = case_when(
13       is.na(sex) & n_na_numeric <= 1 ~ "Unknown", # casi toda la info presente
14       TRUE ~ as.character(sex)                  # dejar como está
15     )
16   )
17
18 df <- df %>%
19   filter(!(is.na(bill_length_mm) &
20     is.na(bill_depth_mm) &
```

```

21     is.na(flipper_length_mm) &
22     is.na(body_mass_g)))
23
24
25 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  $\geq 3$  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

```

1 #Librerías
2 library(knitr)
3 library(kableExtra)

1 # Matriz especie x isla
2 abund <- df %>%
3   count(island, species) %>%
4   pivot_wider(names_from = species, values_from = n, values_fill = 0)
5
6 # Crear tabla
7 abund %>%
8   kable(caption = "Abundancia de pingüinos por especie e isla
9         en el archipiélago Palmer.",
10        align = "lccc",
11        col.names = c("Isla", "Adelie", "Gentoo", "Chinstrap")) %>%
12   kable_styling(full_width = FALSE, bootstrap_options = c("striped",
13   "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

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.

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

```

1 #Librerías
2 library(vegan)

1 # Matriz especie x isla
2 mat <- abund %>% select(-island) %>% as.matrix()
3 rownames(mat) <- abund$island
4
5 # índices
6 shannon <- diversity(mat, index = "shannon")
7 simpson <- diversity(mat, index = "simpson")
8 richness <- specnumber(mat)
9 evenness <- ifelse(richness > 1, shannon / log(richness), 0) # corregido
10
11 indices <- data.frame(
12   Isla      = rownames(mat),
13   Riqueza   = richness,
14   Shannon   = round(shannon, 2),
15   Simpson   = round(simpson, 2),
16   Equidad   = round(evenness, 2)
17 )
18
19 knitr::kable(indices, caption = NULL)

```

Tabla 5: Índices de diversidad de especies de pingüinos por isla.

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



```

1 # Curvas de rarefacción
2 rarecurve(mat, step = 10, sample = min(rowSums(mat)), col = 1:3,
3           lwd = 2, ylab = "Riqueza de especies")
4 legend("bottomright", legend = rownames(mat), col = 1:3,
5       lwd = 2, bty = "n")

```

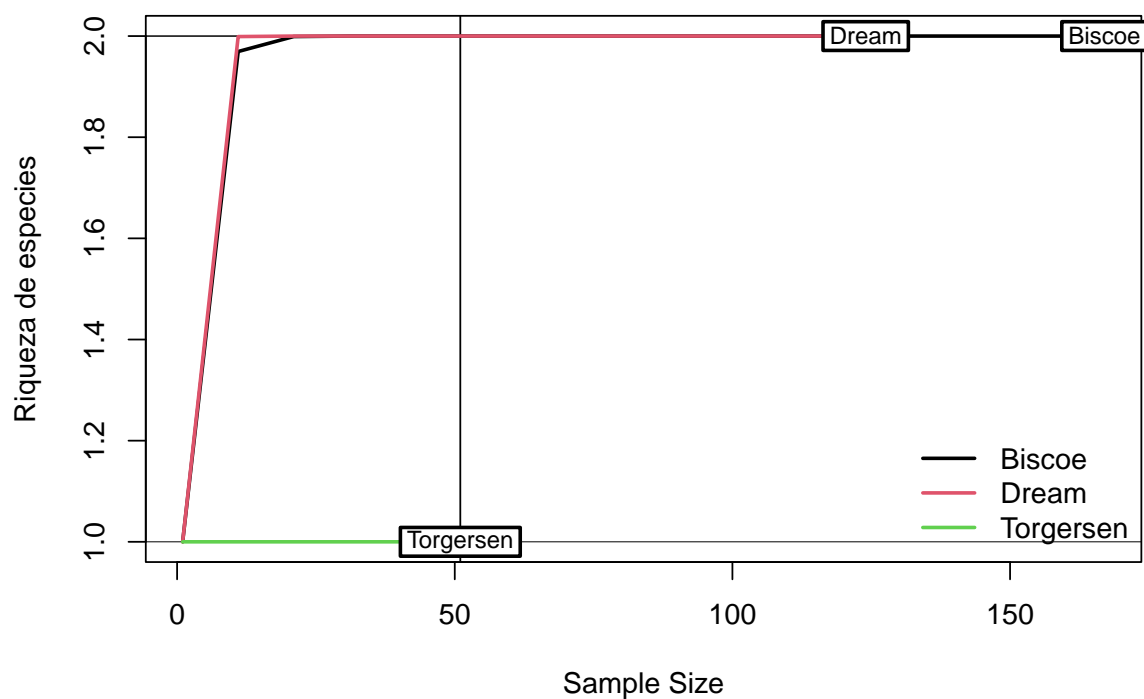


Figura 2: Curvas de rarefacción de riqueza de especies de pingüinos por isla.