Análisis de Varianza de los pingüinos de Palmer

Mauricio Moreno

2023-11-11

Table of contents

Datos

Los datos en el presente reporte corresponden a la tabla `penguins` del paquete `palmerpenguins`. Esta tabla contiene 8 variables que corresponden a 344 pingüinos de 3 especies (Adelie, Chinstrap y Gentoo).

Hipótesis

El largo de la aleta de los pingüinos de Palmer difiere significativamente dependiendo de su especie.

Estadísticos descriptivos

Las poblaciones de pingüinos se encuentran caracterizadas por los siguientes estadísticos:

```
desc_tabla <- describe(penguins) %>%
  flextable(.) %>%
  colformat_double(., digits = 2)
```

```
autofit(desc_tabla)
```

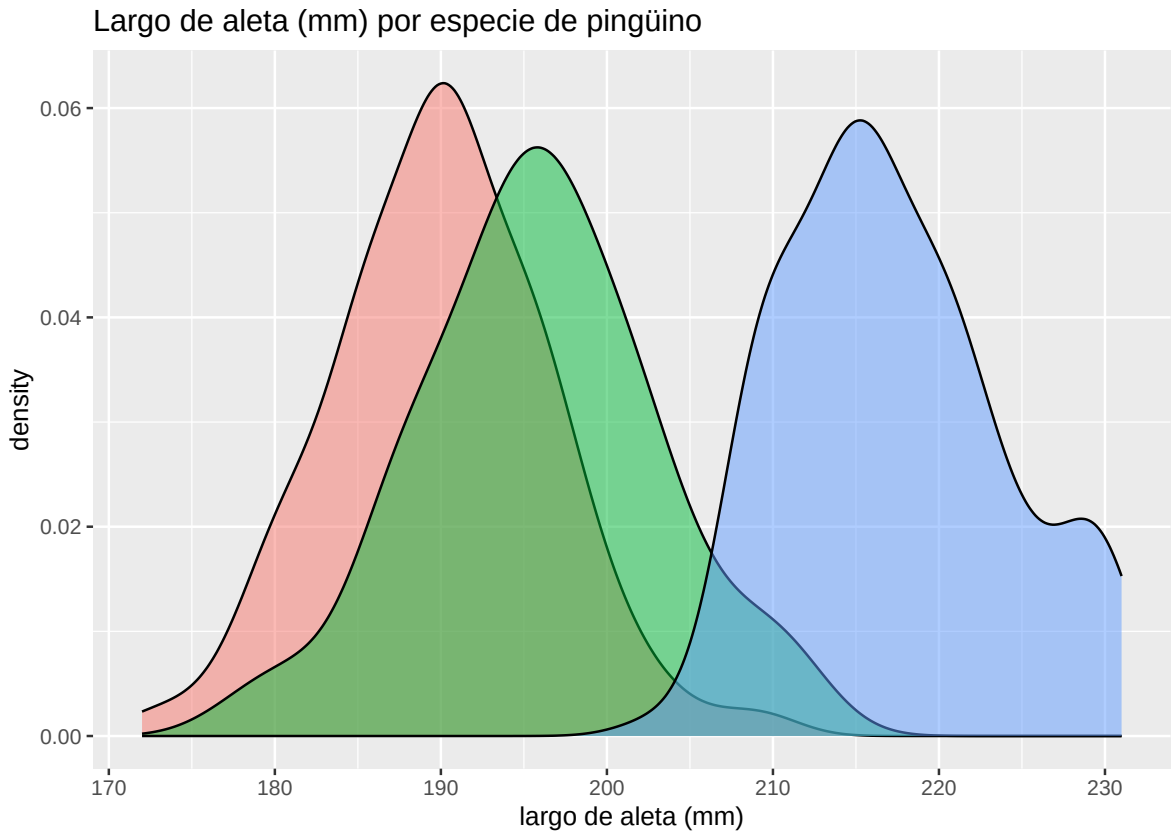
Variable	N	Missing	M	SD	Min	Q25	M
bill_length_mm	342	2	43.92	5.46	32.10	39.23	44
bill_depth_mm	342	2	17.15	1.97	13.10	15.60	17
flipper_length_mm	342	2	200.92	14.06	172.00	190.00	197
body_mass_g	342	2	4,201.75	801.95	2,700.00	3,550.00	4,050
year	344	0	2,008.03	0.82	2,007.00	2,007.00	2,008

Por ejemplo, podemos ver como el largo de la aleta (flipper length) varia de 172 a 231 mm, con una media de 200.9 mm.

Pruebas de los supuestos

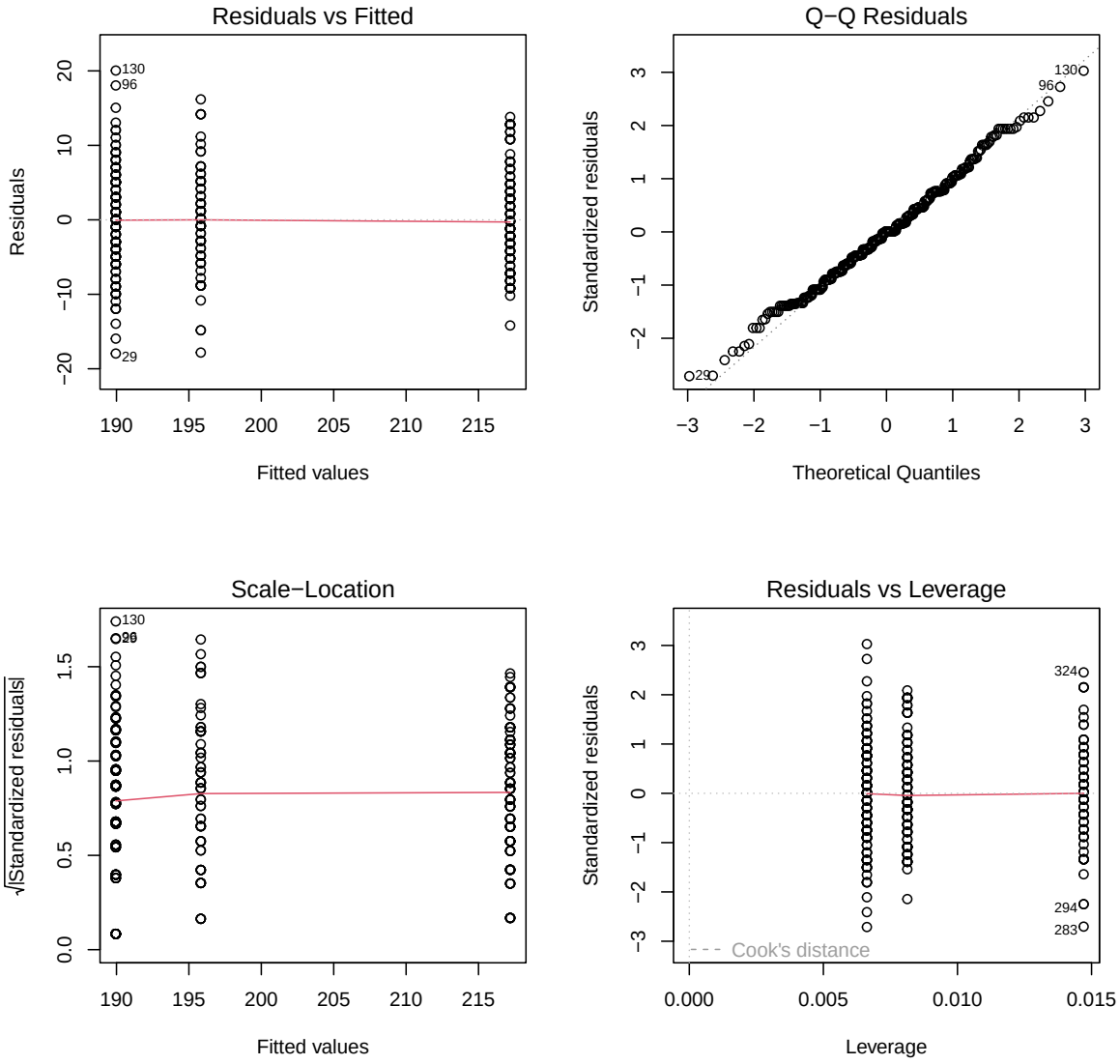
Antes de comenzar con pruebas formales de normalidad y homocedasticidad, podemos dar un vistazo a las distribuciones de las aletas en función de su especie. De acuerdo a la siguiente figura, podemos suponer que los datos están cercanos a cumplir los supuestos.

```
ggplot(penguins, aes(x = flipper_length_mm, fill = species))+
  geom_density(alpha = 0.5)+
  theme(legend.position = "none")+
  labs(x = "largo de aleta (mm)", title = "Largo de aleta (mm) por especie de pingüino")
```



De los gráficos diagnósticos, podemos ver que no existen mayores razones de preocupación de que los supuestos no se cumplan.

```
lm1 <- lm(flipper_length_mm ~ species, data = penguins)
par(mfrow = c(2, 2))
plot(lm1)
```




```
par(mfrow = c(1, 1))
```

Esto se confirma mediante las pruebas formales de Shapiro-Wilk y Levene para la normalidad y la homocedasticidad, respectivamente.

```
residuos <- resid(lm1)
shapiro.test(residuos)
```

```
    Shapiro-Wilk normality test

data:  residuos
W = 0.99452, p-value = 0.2609
```

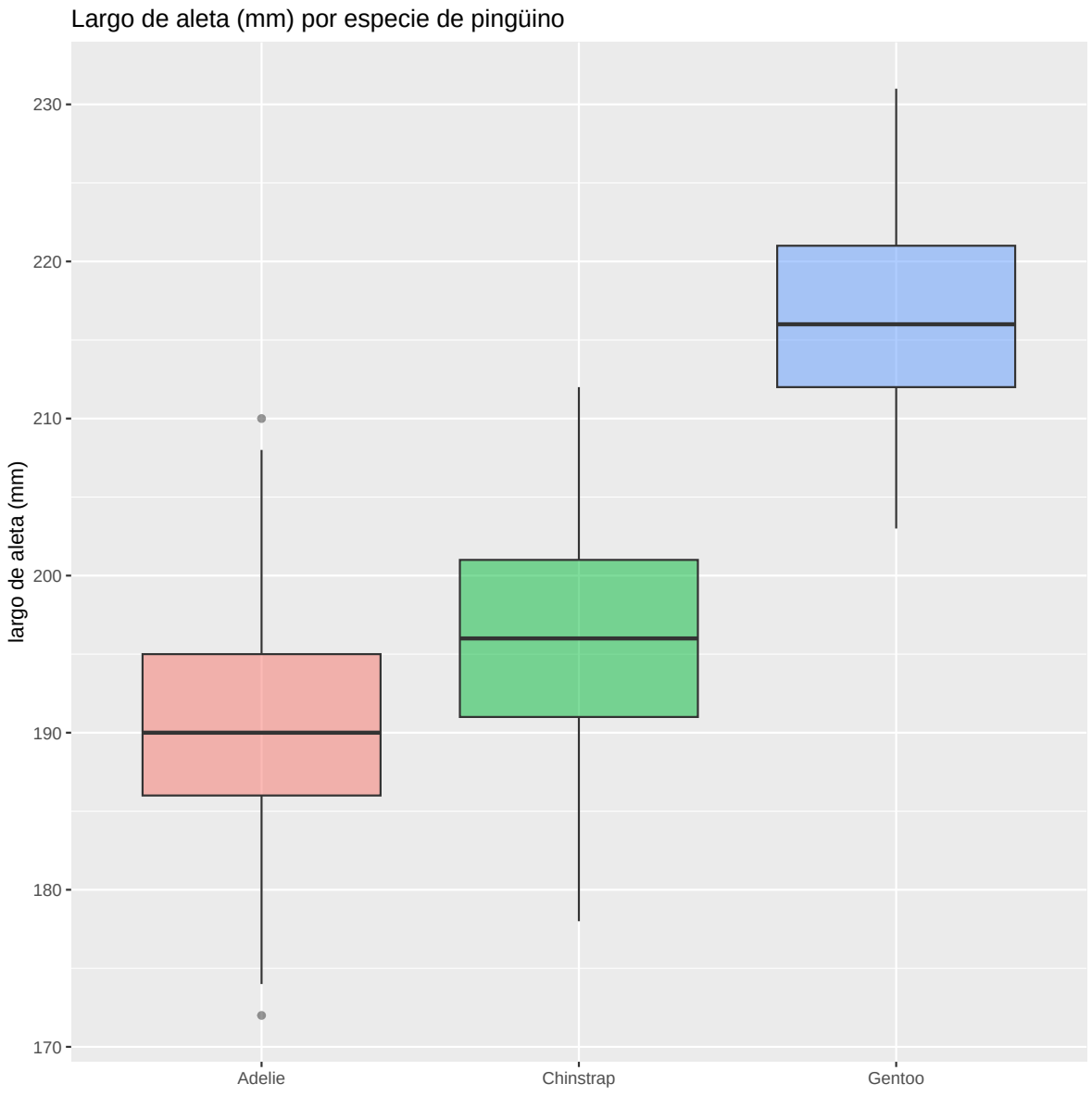
```r
leveneTest(lm1)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.3306 0.7188
      339
```

ANOVA

Una vez que hemos visto que los supuestos se cumplen, podemos llevar a cabo el ANOVA. Sin embargo, antes de ello, es buena práctica explorar un poco más las distribuciones de nuestra variable dependiente.

```r
ggplot(penguins, aes(x = species, y = flipper_length_mm, fill = species))+
  geom_boxplot(alpha = 0.5)+
  theme(legend.position = "none")+
  labs(y = "largo de aleta (mm)", x = NULL, title = "Largo de aleta (mm) por especie de pi
```



Del gráfico podemos intuir dos ideas:

1) Es muy probable que el largo de la aleta entre las especies Adelie y Chinstrap no sean estadísticamente distintas, y

2) Dos observaciones en la especie Adelie, de acuerdo al criterio del rango intercuartílico, podrían ser outliers. Sin embargo, de las pruebas anteriores, podemos estar seguros que estas dos observaciones no apalancan la distribución fuera de la normalidad.

Adicionalmente, podemos dar un vistazo a las medias y desviaciones estándar del largo de aleta por especie. Para ello, podemos hacer uso de la librería `tidycomm`. Pero para este caso, usaremos `dplyr`, con su función `summarise`.

```
penguins %>%
  group_by(species) %>%
  summarise(media = mean(flipper_length_mm, na.rm = T),
            ds = sd(flipper_length_mm, na.rm = T))
```

```
# A tibble: 3 x 3
  species   media    ds
  <fct>     <dbl> <dbl>
1 Adelie     190.  6.54
2 Chinstrap  196.  7.13
3 Gentoo     217.  6.48
```

Ahora, de acuerdo a la tabla del ANOVA, confirmamos lo que ya nos pudo haber dado la idea de los análisis anteriores. Así, podemos concluir, que con un valor p de 0.000, rechazamos la hipótesis nula de la igualdad de las medias del largo de aleta entre las especies, y al menos una es distinta.

```
tabla_anova <- as.data.frame(Anova(lm1))
tabla_anova <- cbind(parametro = row.names(tabla_anova), tabla_anova)
tabla_anova <- add_significance(tabla_anova,
                  p.col = "Pr(>F)",
                  output.col = " ",
                  cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                  symbols = c("***", "**", "*", ".", "ns"))
tabla_anova <- colformat_double(flextable(tabla_anova), digits = 3, j = c(2, 4, 5))
tabla_anova <- add_footer_lines(tabla_anova, "Códigos Signif. 0 '***', 0.001 '**', 0.1 '*'
autofit(tabla_anova)
```

parametro	Sum Sq	Df	F value	Pr(>F)
species	52,473.284	2	594.802	0.000 ***
Residuals	14,953.257	339		

Códigos Signif. 0 '***', 0.001 '**', 0.1 '*', 0.05 '.', 0.1 'ns'

Pruebas post-hoc

Llevamos a cabo la prueba HSD de Tukey, y observamos que todas las pruebas de pares son significativamente distintas. Es decir, contrario a lo que nos dio a pensar el boxplot, existe de hecho diferencias entre el largo de las aletas de las especies Adelie y Chinstrap.

```
medias_marginales <- emmeans(lm1, specs = "species", type = "response")

tukey_comp <- contrast(medias_marginales, specs = "species", method = "tukey")
tukey_comp
```

```
 contrast           estimate    SE  df t.ratio p.value
 Adelie - Chinstrap    -5.87 0.970 339  -6.052  <.0001
 Adelie - Gentoo      -27.23 0.807 339 -33.760  <.0001
 Chinstrap - Gentoo   -21.36 1.004 339 -21.286  <.0001

P value adjustment: tukey method for comparing a family of 3 estimates
```

Ahora, podemos visualizar estos resultados en dos formas, con la tabla de grupos de Tukey, o graficando dichos grupos.

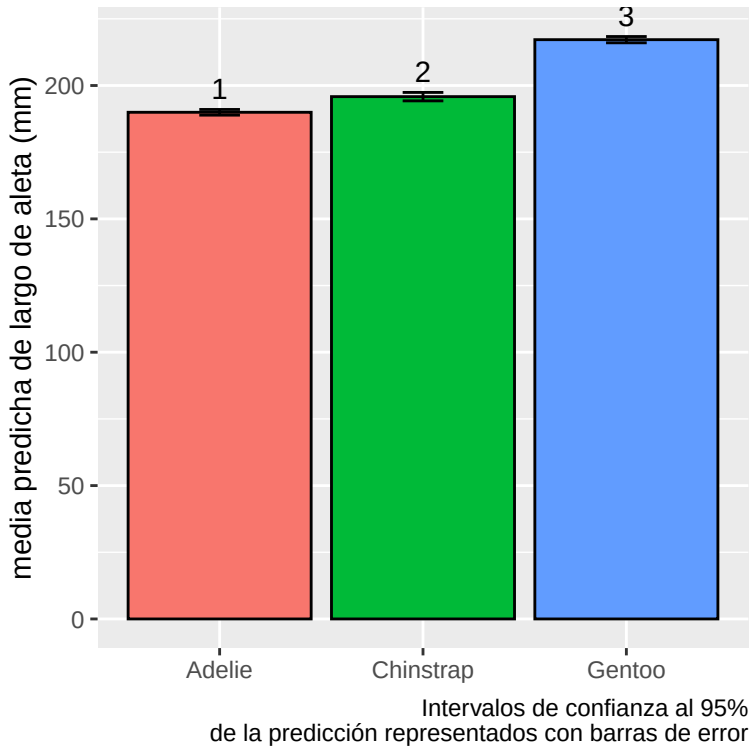
```
grupos_tukey <- cld(medias_marginales)
grupos_tukey
```

```
 species   emmean    SE  df lower.CL upper.CL .group
 Adelie       190 0.540 339      189      191  1
 Chinstrap    196 0.805 339      194      197   2
 Gentoo       217 0.599 339      216      218    3

Confidence level used: 0.95
P value adjustment: tukey method for comparing a family of 3 estimates
significance level used: alpha = 0.05
NOTE: If two or more means share the same grouping symbol,
      then we cannot show them to be different.
      But we also did not show them to be the same.
```

```
gruposvals <- as.data.frame(grupos_tukey)
gruposvals$Pred <- factor(gruposvals$species, levels = c("Adelie", "Chinstrap", "Gentoo"))
ggplot(gruposvals,
       aes(x = species,
```

```
          y = emmean,
          fill = species)) +
geom_bar(stat = "identity",
          show.legend = F,
          color = "black")+
geom_errorbar(aes(ymin = lower.CL,
                  ymax = upper.CL),
              width=0.2)+
geom_text(aes(label=str_trim(.group),
              y = upper.CL, vjust=-0.5))+
labs(caption = "Intervalos de confianza al 95%\nde la predicción representados con barra
     x = NULL, y = "media predicha de largo de aleta (mm)")
```



O también podemos hacer uso de la librería `ggstatsplot`. Esta librería no realiza HSD Tukey *per se*, sino Games-Howell. Games-Howell no asume varianzas iguales, y por tanto realiza corrección de Welch en todos los casos. Cuando los grupos son homocedásticos, sus resultados son los mismos que HSD Tukey.

```r
set.seed(1985) # esta librería realiza sampling aleatorio para calcular los intervalos de
ggbetweenstats(
  data = penguins,
  x = species,
  y = flipper_length_mm,
  pairwise.comparisons = T,
  pairwise.display = "all",
  p.adjust.method = "none"
)
```