Estudio de datos biométricos sobre el ciervo volante. Inferencia estadística.

Olivia Marcos Diaz de Otazu Irene Canales Giménez
Samuel Melián Benito

2025-12-28

Índice

Objetivo del estudio

Continuaremos el estudio de datos biométricos del **ciervo volante** (*Lucanus cervus*) estudiando los valores de anchura de cabeza (KB) y longitud de los élitros (EL), y sus diferencias dependiendo de su sexo y procedencia, esta vez usando inferencia estadística.

El objetivo del estudio es responder a las siguentes preguntas:

1. ¿Entre qué valores fluctúa el promedio de la anchura de la cabeza y de la longitud de los élitros del Lucanus Cervus?
2. ¿Existe diferencia estadísticamente significativa entre el promedio de la anchura de la cabeza y la longitud de los élitros según sexo?
3. ¿Existe diferencia estadísticamente significativa entre el promedio de la anchura de la cabeza y la longitud de los élitros según si habitan en Cantabria o en Asturias?
4. Suponiendo que la muestra es representativa de la población, ¿los hábitats son igual de numerosos en Cantabria y Asturias?, ¿y en términos de sexo?

Y concluiremos comparando las respuestas obtenidas con las conclusiones de la primera práctica de estadística descriptiva. Comprobaremos, utilizando técnicas de inferencia si dichas conclusiones son estadísticamente significativas o sólo responden al azar.

Metodología

Para esta práctica de inferencia y contrastes de hipótesis, utilizaremos en todo momento el conjunto de datos al completo, es decir, los 250 individuos. Excluiremos aquellos cuyo origen es desconocido. Esto supone un cambio en relación con la práctica anterior, en la que se empleaba una muestra aleatoria.

En primer lugar, para estimar el valor medio poblacional de ambas variables biométricas, se han calculado intervalos de confianza al 95% para la media, empleando la distribución t de Student, dado que la desviación típica poblacional es desconocida.

Para analizar la existencia de diferencias entre sexos, se han realizado contrastes de hipótesis para la igualdad de medias mediante la prueba t de Student para muestras independientes, utilizando la versión de Welch al no asumir igualdad de varianzas. Además, se ha calculado el tamaño del efecto mediante el estadístico d de Cohen.

El efecto de la procedencia geográfica sobre las variables biométricas se ha evaluado mediante un análisis de la varianza (ANOVA) de un factor, considerando las provincias como niveles del factor. Cuando ha sido necesario, se ha aplicado la prueba post hoc de Tukey HSD para comparar las medias por pares.

Finalmente, para estudiar la representatividad de los hábitats según provincia y sexo, se han realizado contrastes de proporciones mediante el test chi-cuadrado, acompañados de sus correspondientes intervalos de confianza al 95%.

Se incluirán tablas a modo de resumen con los resultados de los análisis ya realizados previamente sobre las variables.

Análisis de resultados

Pregunta 1. ¿Entre qué valores fluctúa el promedio de la anchura de la cabeza y de la longitud de los élitros del *Lucanus Cervus*?

El intervalo de confianza al 95% para la media de la **anchura de la cabeza (KB)** es:

$$[\,11.56,\ 12.52\,]\ \text{mm}$$

Este primer cálculo del intervalo de confianza para el promedio de la anchura de la cabeza lo hemos realizado mediante la fórmula de intervalo de confianza para la media, siendo desconocida la desviación estándar poblacional. Recordemos que esta fórmula es

$$IC = \overline{X} \pm t_{\frac{\alpha}{2},n-1} \cdot \left(\frac{s}{\sqrt{n}}\right)$$

Interpretación: Con un nivel de confianza del 95%, el promedio de la anchura de la cabeza de la **población** se encuentra en este intervalo obtenido.

El intervalo de confianza al 95% para la media de la **longitud de los élitros (EL)** es:

$$[\,20.60,\ 21.22\,]\ \text{mm}$$

Interpretación: Con un nivel de confianza del 95%, el promedio de la longitud de los élitros de la **población** se encuentra en este intervalo obtenido. Es decir, si repitiéramos el procedimiento con muchas muestras distintas, en el 95% de las ocasiones aproximadamente la media poblacional estaría en este intervalo de confianza.

Pregunta 2.1 ¿Existe diferencia estadísticamente significativa entre el promedio de la anchura de la cabeza según sexo?

Primero presentamos un resumen estadístico de la anchura de la cabeza para cada grupo (hembras y machos) con los datos de la muestra de 250 individuos.

Table 1: Resumen estadístico de la anchura de la cabeza (KB) según sexo

SEXO	N_validos	Pct_validos	Media	Mediana	DE	Min	Max
hembra	118	100	8.894	8.9	0.882	7.2	11.2
macho	132	100	14.846	14.6	3.293	9.0	24.1

Para poder responder esta pregunta tenemos que realizar un contraste de hipótesis para la igualdad de medias.

Hipótesis nula (H_0): La media de KB es igual en ambos sexos.

Hipótesis alternativa (H_a): La media de KB es diferente entre los dos sexos.

Para ello hacemos un test t de Student sobre la muestra de los machos y la muestra de las hembras. Concretamente haremos una prueba t de Welch, que se usa con dos muestras independientes cuando las varianzas no son iguales. Además, conviene su uso al ser las dos muestras de diferentes tamaños.

Table 2: Test t de medias para KB por sexos (Parte 1)

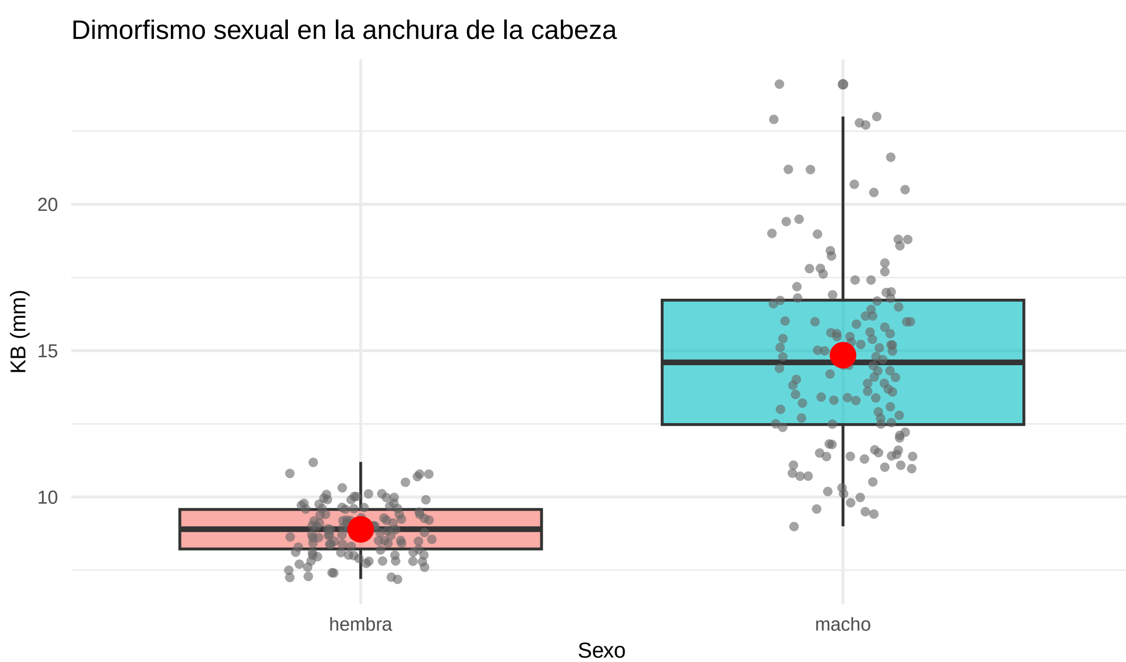
Estadístico t	Grados de libertad	p-Valor
-19.97606	151.7822	0

Table 3: Test t de medias para KB por sexos (Parte 2)

Tamaño del efecto (d de Cohen)	IC 95% (inf)	IC 95% (sup)
2.410409	-6.539955	-5.362729

Interpretación: Existe una diferencia significativa entre las medias de la anchura de la cabeza entre machos y hembras, el p-Valor es <0.001. El tamaño del efecto es muy grande, lo que indica dimorfismo sexual muy pronunciado. La media para machos es de 14.85mm y la de hembras 8.89mm.

En el siguiente gráfico se puede observar claramente la diferencia entre sexos:



Pregunta 2.2 ¿Existe diferencia estadísticamente significativa entre el promedio de la longitud de los élitros según sexo?

Primero presentamos un resumen estadístico de la longitud de los élitros para cada grupo (hembras y machos) con los datos de la muestra de 250 individuos.

Table 4: Resumen estadístico de la longitud de los élitros (EL) según sexo

SEXO	N_validos	Pct_validos	Media	Mediana	SD	Min	Max
hembra	118	100	19.65	19.7	1.85	15.75	23.5
macho	132	100	22.05	22.2	2.41	15.30	28.3

Para resolver esta pregunta seguimos exactamente el mismo procedimiento, haciendo el t test esta vez para la variable EL.

Table 5: Test t de medias para EL por sexos (Parte 1)

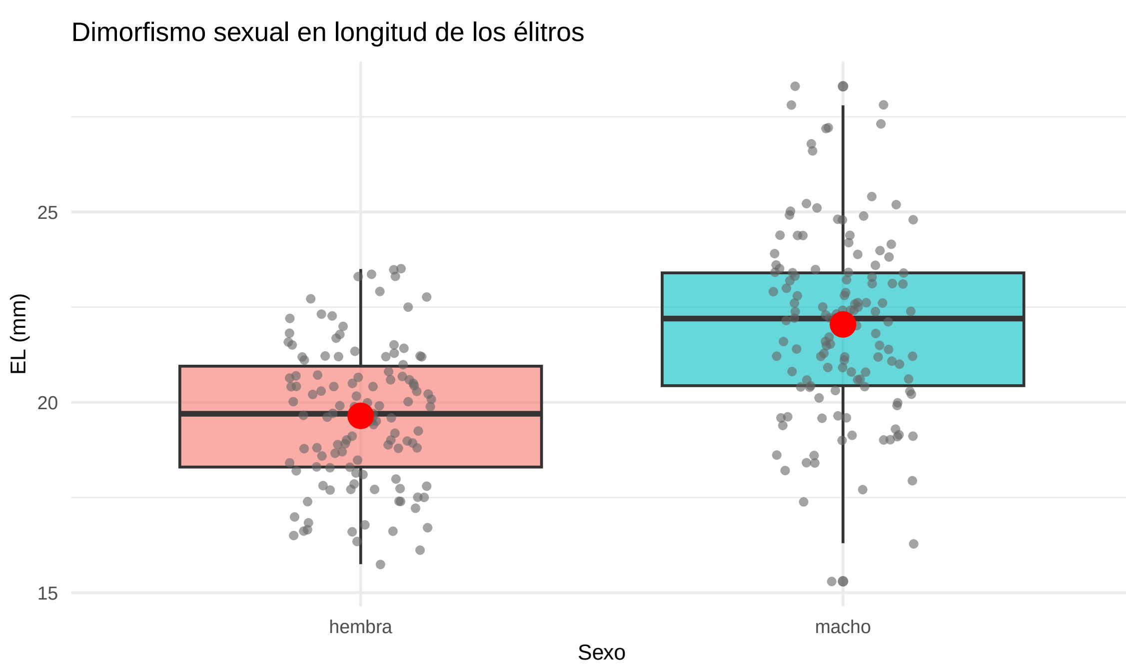
Estadístico t	Grados de libertad	p-Valor
-8.887837	242.6702	0

Table 6: Test t de medias para EL por sexos (Parte 2)

Tamaño del efecto (d de Cohen)	IC 95% (inf)	IC 95% (sup)
1.109816	-2.932047	-1.868184

Interpretación: Existe una diferencia significativa entre las medias de la longitud de los élitros entre machos y hembras, el p-Valor es <0.001. El tamaño del efecto no es tan grande como para la anchura de la cabeza, con lo cual no existe tanta diferencia como en la anterior pero sí que hay diferencia.

En el siguiente gráfico se puede observar la diferencia entre sexos:



Pregunta 3.1 ¿Existe diferencia estadísticamente significativa entre el promedio de la anchura de la cabeza según si habitan en Cantabria o en Asturias?

Primero presentamos un resumen estadístico de la anchura de la cabeza para cada procedencia geográfica con los datos de la muestra de 250 individuos.

Table 7: Resumen estadístico de la anchura de la cabeza (KB) según provincia

PROVINCIA	N_validos	Pct_validos	Media	Mediana	SD	Min	Max
Asturias	180	100	11.88	10.80	3.72	7.2	22.7
Cantabria	28	100	13.16	13.35	3.67	8.0	21.2
Otras	42	100	11.94	10.02	4.53	7.4	24.1

Al igual que en el apartado anterior, para responder esta pregunta podríamos realizar un contraste de hipótesis para la igualdad de medias teniendo en cuenta solo el subconjunto de Asturias y de Cantabria.

Hipótesis nula (H_0): La media de KB es igual en ambas provincias.

Hipótesis alternativa (H_a): La media de KB es diferente entre las dos provincias.

Para ello haríamos un test t de de Student, concretamente haríamos una prueba t de Welch. No obstante, para ir algo más allá y ofrecer algún resultado adicional, podemos realizar un test ANOVA para entre otras cosas, poder obtener conclusiones sobre la pregunta planteada. Ahora el contraste es:

H_0: La media de la anchura de la cabeza los 3 grupos es igual.

H_1: Al menos una media es distinta.

Table 8: ANOVA de un factor para la anchura de la cabeza (KB) según procedencia geográfica

Estadístico F	Grados de libertad entre grupos	Grados de libertad dentro de grupos (residuales)	p-Valor
1.332528	2	247	0.265699

Una vez hecho el test ANOVA, podemos hacer la prueba Tukey HSD (post-hoc) para identificar los pares específicos cuyas medias difieren.

Table 9: Comparaciones múltiples (Tukey HSD) para KB por provincia

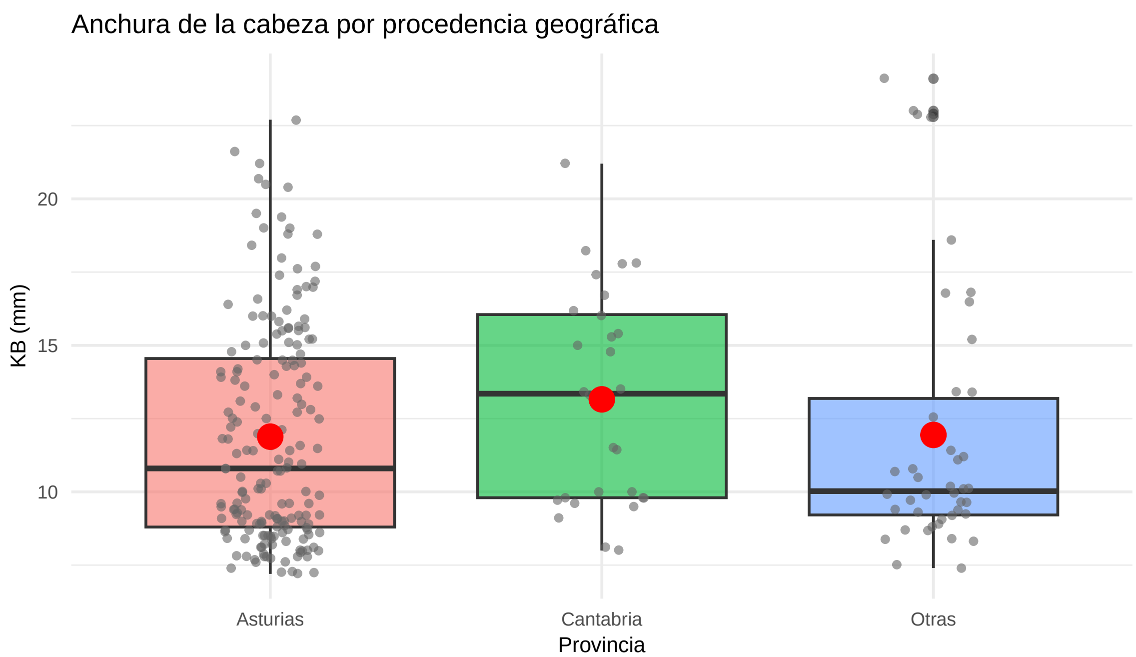
Comparación	Diferencia medias	IC 95% (inf)	IC 95% (sup)	p-Valor ajustado
Cantabria-Asturias	1.272698	-0.575950	3.121347	0.237723
Otras-Asturias	0.058413	-1.500973	1.617798	0.995708
Otras-Cantabria	-1.214286	-3.434440	1.005869	0.402350

El p-Valor que hemos obtenido en el análisis de varianza de un factor (ANOVA) es elevado ($p = 0.265699$), por lo que no se rechaza la hipótesis nula de igualdad de medias.

En consecuencia, no se detectan diferencias estadísticamente significativas en la anchura de la cabeza entre las tres procedencias geográficas (Asturias, Cantabria y Otras).

Además, los resultados del análisis de comparaciones múltiples nos permiten ver que particularmente tampoco hay diferencias de medias estadísticamente significativas entre **Asturias y Cantabria**, lo que refuerza la conclusión anterior.

El siguiente gráfico nos permite ver la variación de la anchura de la cabeza por provincias:



Pregunta 3.2 ¿Existe diferencia estadísticamente significativa entre la longitud de los élitros según si habitan en Cantabria o en Asturias?

Primero presentamos un resumen estadístico de la longitud de los élitros para cada procedencia geográfica con los datos de la muestra de 250 individuos.

Table 10: Resumen estadístico de la longitud de los élitros (EL) según provincia

PROVINCIA	N_validos	Pct_validos	Media	Mediana	SD	Min	Max
Asturias	180	100	20.76	20.65	2.46	15.3	27.3
Cantabria	28	100	21.53	21.50	2.11	16.7	27.2
Otras	42	100	21.14	20.55	2.68	16.6	28.3

De nuevo, hacemos un test ANOVA y un análisis post hoc de Tukey para confirmar el resultado, esta vez para la variable EL.

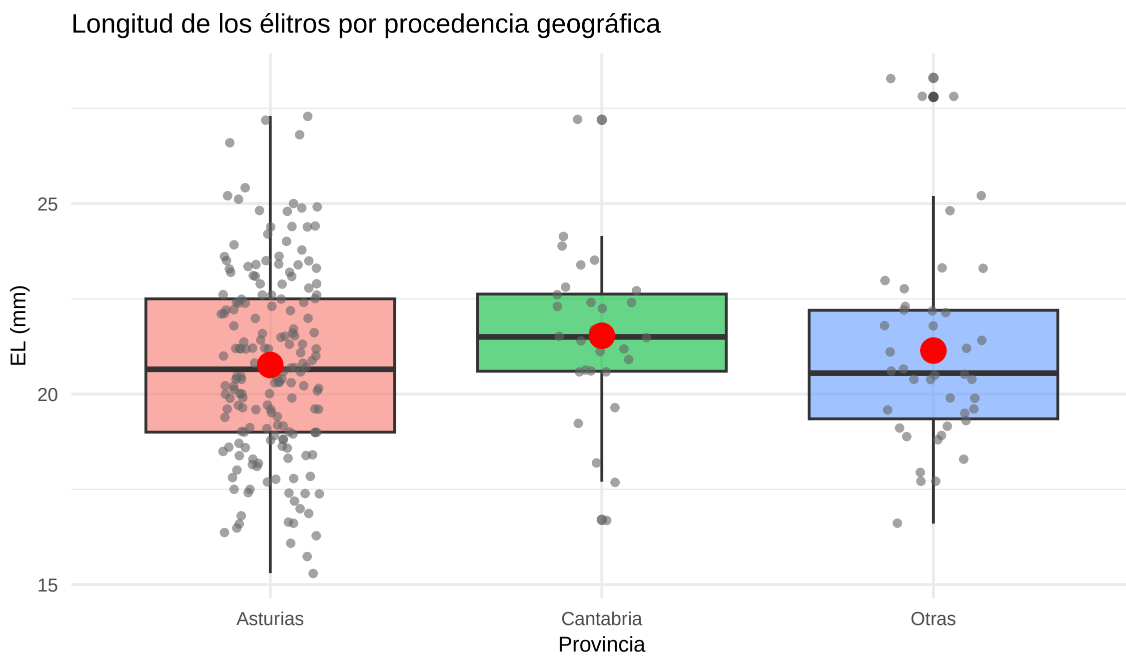
Table 11: ANOVA de un factor para la longitud de los élitros (EL) según procedencia geográfica

Estadístico F	Grados de libertad entre grupos	Grados de libertad dentro de grupos (residuales)	p-Valor
1.390908	2	247	0.250791

Table 12: Comparaciones múltiples (Tukey HSD) para EL por provincia

Comparación	Diferencia medias	IC 95% (inf)	IC 95% (sup)	p-Valor ajustado
Cantabria-Asturias	0.767302	-0.413877	1.948480	0.277856
Otras-Asturias	0.378611	-0.617745	1.374968	0.643339
Otras-Cantabria	-0.388690	-1.807240	1.029859	0.794754

De nuevo, el análisis de varianza de un factor no muestra diferencias estadísticamente significativas en la longitud de los élitros entre las procedencias geográficas ($p = 0.250791$). El análisis post hoc de Tukey confirma el resultado, ya que ninguna de las comparaciones por pares presenta un p-valor ajustado inferior a 0.05, incluyendo la comparación entre **Asturias y Cantabria**.



Pregunta 4. Suponiendo que la muestra es representativa de la población, ¿los hábitats son igual de numerosos en Cantabria y Asturias?, ¿y en términos de sexo?

Para responder a esta pregunta vamos a hacer un test de contraste de proporciones que nos permita evaluar si las dos provincias están igualmente representadas en la población. Las hipótesis son

Hipótesis nula (H_0): Asturias y Cantabria están igualmente representadas en la población (es decir, $p_{Asturias} = p_{Cantabria}$)

Hipótesis alternativa (H_a): Asturias y Cantabria no están igualmente representadas en la población (es decir, $p_{Asturias} \neq p_{Cantabria}$)

Table 13: Contraste de diferencia de proporciones para la representación de los hábitats entre Asturias y Cantabria (IC 95%)

Estadístico chi-cuadrado	p-valor	IC 95%
222.154	0	[0.665,0.796]

Interpretación: El p-Valor es del orden de 10^{-16}, lo que indica que podemos rechazar la hipótesis nula de igualdad de proporciones. Además, el intervalo de confianza no contiene al 0, lo que refuerza este hecho. Con esto, podemos afirmar que los hábitats no son igual de numerosos en Asturias y Cantabria.

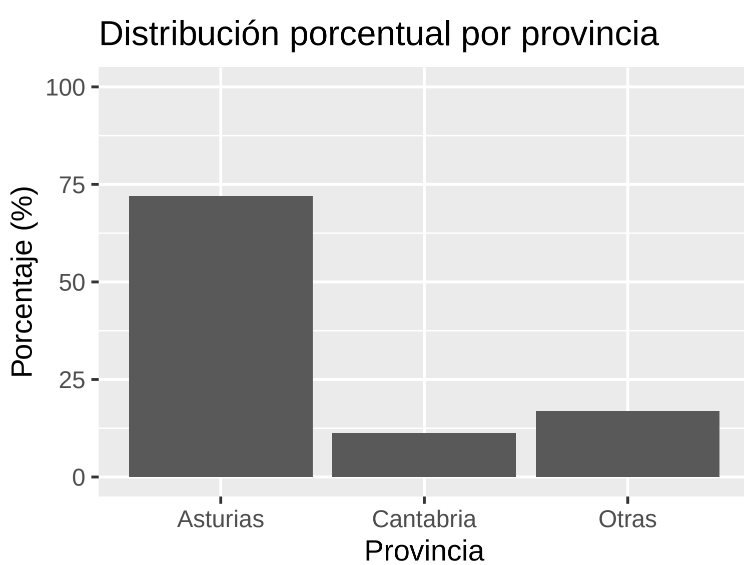
Para responder a la segunda pregunta y ver si son igual de numerosos en términos de sexo, podemos de nuevo realizar un test de contraste de proporciones, esta vez para el sexo.

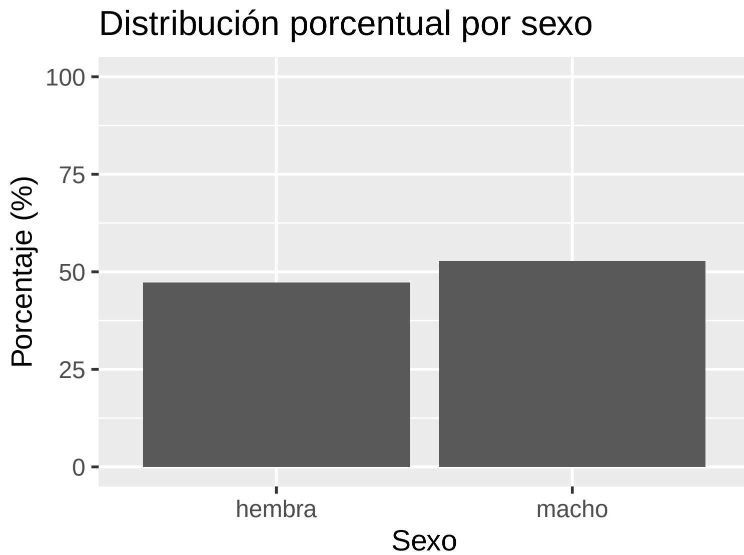
Table 14: Contraste de diferencia de proporciones para la representación de los hábitats entre sexos (IC 95%)

Estadístico chi-cuadrado	p-valor	IC 95%
1.568	0.21	[-0.032,0.144]

Interpretación: El p-Valor obtenido es 0.21, bastante superior al nivel de significación habitual, por lo que no se rechaza la hipótesis nula de igualdad de proporciones. Este resultado se ve reforzado por el intervalo de confianza al 95%, que contiene al valor 0, lo que indica que no se detectan diferencias estadísticamente significativas entre machos y hembras en lo que a la representatividad de los hábitats se refiere.

Presentamos a continuación los dos gráficos ilustrativos que representan el porcentaje de individuos de cada provincia y de cada sexo:





Distribución porcentual por sexo
100
75
50
25
0
Porcentaje (%)
hembra
macho
Sexo

Resumen de resultados y conclusiones

Table 15: Resumen de los análisis estadísticos realizados

Analisis	Test	Estadistico	p.valor	Resultado
KB (por sexo)	t de Student	t = -19.976	0.000	Concluyente
KB (por provincia)	ANOVA	F = 1.333	0.266	Alto, no concluyente
EL (por sexo)	t de Student	t = -8.888	0.000	Concluyente
EL (por provincia)	ANOVA	F = 1.391	0.251	Alto, no concluyente
Contraste (provincia)	Test de proporciones	Chi-cuadrado = 222.154	0.000	Concluyente
Contraste (sexo)	Test de proporciones	Chi-cuadrado = 1.568	0.210	Alto, no concluyente

En la tabla anterior podemos ver un resumen de los test que hemos llevado a cabo, así como una rápida interpretación de sus resultados.

Conclusiones

Tras este análisis empleando inferencia estadística sobre los datos biométricos del ciervo volante, hemos obtenido las siguientes conclusiones. Cabe destacar que gran parte del interés de esta práctica radica en la comparación de este estudio con el realizado previamente a través de la estadística descriptiva, en la práctica 1.

Para comenzar, hemos analizado la variable Anchura de la cabeza (KB) y los resultados confirman lo que ya se veía en la práctica anterior. En ella se tenía que la estimación de la media de la muestra escogida estaba en un 12.14mm. Este valor se encuentra dentro del intervalo de confianza que se ha calculado para la media poblacional en la pregunta 1.

A continuación, gracias a los resultados de los análisis de la pregunta 2, podemos confirmar un dimorfismo sexual muy acentuado. Para ambas variables se observa un p-Valor tal que $p < 0.001$, empleando el análisis de t de Student. Esto evidencia la existencia de diferencias reales entre machos y hembras. Esto también es algo que se había observado en la práctica anterior, y que queda de nuevo confirmado.

Tomando ahora el sesgo geográfico, recordemos que no obtuvimos conclusiones claras en el análisis descriptivo, pues las diferencias no eran significativas. En este nuevo análisis los p-Valores de entre 0.20 y 0.27 tampoco indican una evidencia clara de la influencia del factor geográfico en las diferencias morfológicas que presentan los ejemplares. El análisis de la varianza y las comparaciones realizadas indican que las diferencias entre provincias podrían ser debidas al azar hasta en un 27% de los casos, lo cual no permite una conclusión.

En la última parte de esta práctica, la pregunta 4, se buscaba confirmar si existían diferencias en términos del número de ejemplares según el sexo y la provincia. El test de proporciones nos ha servido para comparar las proporciones entre grupos. Primero, entre provincias, los resultados nos indican que sí hay una diferencia significativa entre ellas, y que no se distribuyen equitativamente. Por otro lado, según su sexo, se obtiene un p-Valor que sí es significativo, por lo que no podemos descartar la hipótesis nula y no se puede afirmar estadísticamente que las diferencias se deban a este sesgo. Cabe destacar que estas conclusiones son una reiteración de lo observado en la práctica anterior, si tomamos la muestra dada como una buena representación de la población total.

En conclusión, los resultados obtenidos empleando inferencia estadística confirman aquellos obtenidos en el análisis descriptivo. Las variaciones morfológicas observadas en la especie de estudio se pueden explicar debido al factor sexual, y sin embargo no podemos dar una conclusión acerca de la infuencia del factor geográfico, que no parece tener un efecto significativo. Esto demuestra una validez en el estudio realizado entre ambas prácticas, y confirma que se han obtenido resultados robustos en torno a la población de estudio.

Anexo: Código R

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
library(tidyverse)
library(summarytools)
library(GGally)
library(gt)
library(flextable)
library(knitr)
library(corrplot)
library(ggplot2)
library(dplyr)
library(kableExtra)
library(effsize)
library(ggstatsplot)
datos <- read.table("./datos/CIERVO.txt", fill=TRUE, dec = ",", header=TRUE );

datos$SEXO <- factor(datos$SEXO, levels = c("hembra", "macho"))
datos$PROVINCIA <- factor(datos$PROVINCIA, levels = c("Asturias", "Cantabria",
 ↪  "Otras"))

datos <- datos |> dplyr::filter(!is.na(PROVINCIA))
calcular_ic <- function(datos, alpha) {
  n <- length(datos) #Tamaño del dataset
  media <- mean(datos) #Media de los datos
  s <- sd(datos) #Desviación estándar de los datos
  t_critico <- qt(1 - alpha/2, df = n - 1) #valor crítico de la t-student (alpha
 ↪  es el nivel de confianza t n los grados de libertad de la t-student)
  error_estandar <- s / sqrt(n)
  ME <- t_critico * error_estandar
  c(media - ME, media + ME) #Intervalo de confianza
}
IC_KB <- calcular_ic(datos$KB, 0.05)
IC_EL <- calcular_ic(datos$EL, 0.05)
datos %>%
  group_by(SEXO) %>%
  summarise(
    N_validos = sum(!is.na(KB)),
    Pct_validos = N_validos / n() * 100,
    Media = mean(KB, na.rm = TRUE),
    Mediana = median(KB, na.rm = TRUE),
    DE = sd(KB, na.rm = TRUE),
    Min = min(KB, na.rm = TRUE),
    Max = max(KB, na.rm = TRUE)
  ) %>%
  kable(
    caption = "Resumen estadístico de la anchura de la cabeza (KB) según sexo",
```

```
    digits = 3,
    booktabs = TRUE
  ) %>%
  kable_styling(
    latex_options = c("hold_position", "striped")
  )
test_KB_sexo = t.test(KB ~ SEXO, data = datos, var.equal = FALSE)

cohen_d <- cohen.d(KB ~ SEXO, data = datos, pooled = TRUE, hedges.correction =
 ↪  FALSE)

tabla_kb <- data.frame(
  t = unname(test_KB_sexo$statistic),
  gl = unname(test_KB_sexo$parameter),
  p_valor = test_KB_sexo$p.value,
  d_cohen = abs(cohen_d$estimate),
  IC_inf = test_KB_sexo$conf.int[1],
  IC_sup = test_KB_sexo$conf.int[2]
)
tabla_kb[, 1:3] |>
  kable(
    col.names = c("Estadístico t", "Grados de libertad", "p-Valor"),
    caption = "Test t de medias para KB por sexos (Parte 1)",
    digits = 6,
    booktabs = TRUE,
    align = "c"
  ) |>
  kable_styling(latex_options = c("hold_position","striped"), font_size = 9)

tabla_kb[, 4:6] |>
  kable(
    col.names = c("Tamaño del efecto (d de Cohen)", "IC 95% (inf)", "IC 95%
     ↪  (sup)"),
    caption = "Test t de medias para KB por sexos (Parte 2)",
    digits = 6,
    booktabs = TRUE,
    align = "c"
  ) |>
  kable_styling(latex_options = c("hold_position","striped"), font_size = 9)
ggplot(datos, aes(x = SEXO, y = KB, fill = SEXO)) +
  geom_boxplot(
    alpha = 0.6,
    outlier.shape = 16) +
  geom_jitter(
    width = 0.15,
    alpha = 0.6,
    size = 1,
    color = "grey40") +
```

```r
  stat_summary(
    fun = mean,
    geom = "point",
    size = 4,
    color = "red"
  ) +
  labs(
    title = "Dimorfismo sexual en la anchura de la cabeza",
    x = "Sexo",
    y = "KB (mm)"
  ) +
  theme_minimal() +
  theme(legend.position = "none") +
  theme(
    plot.title = element_text(size = 10),
    axis.title = element_text(size = 8),
    axis.text  = element_text(size = 7)
  )
datos %>%
  group_by(SEXO) %>%
  summarise(
    N_validos = sum(!is.na(EL)),
    Pct_validos = N_validos / n() * 100,
    Media = mean(EL, na.rm = TRUE),
    Mediana = median(EL, na.rm = TRUE),
    SD = sd(EL, na.rm = TRUE),
    Min = min(EL, na.rm = TRUE),
    Max = max(EL, na.rm = TRUE)
  ) %>%
  kable(
    caption = "Resumen estadístico de la longitud de los élitros (EL) según
     ↪  sexo",
    digits = 2,
    booktabs = TRUE
  ) %>%
  kable_styling(
    latex_options = c("hold_position", "striped")
  )
test_EL_sexo = t.test(EL ~ SEXO, data = datos, var.equal = FALSE)

cohen_d <- cohen.d(EL ~ SEXO, data = datos, pooled = TRUE, hedges.correction =
 ↪  FALSE)

tabla_el <- data.frame(
  t = unname(test_EL_sexo$statistic),
  gl = unname(test_EL_sexo$parameter),
  p_valor = test_EL_sexo$p.value,
  d_cohen = abs(cohen_d$estimate),
```

```
  IC_inf = test_EL_sexo$conf.int[1],
  IC_sup = test_EL_sexo$conf.int[2]
)
# Primera mitad
tabla_el[, 1:3] |>
  kable(
    col.names = c("Estadístico t", "Grados de libertad", "p-Valor"),
    caption = "Test t de medias para EL por sexos (Parte 1)",
    digits = 6,
    booktabs = TRUE,
    align = "c"
  ) |>
  kable_styling(latex_options = c("hold_position","striped"), font_size = 9)

# Segunda mitad
tabla_el[, 4:6] |>
  kable(
    col.names = c("Tamaño del efecto (d de Cohen)", "IC 95% (inf)", "IC 95%
     ↪ (sup)"),
    caption = "Test t de medias para EL por sexos (Parte 2)",
    digits = 6,
    booktabs = TRUE,
    align = "c"
  ) |>
  kable_styling(latex_options = c("hold_position","striped"), font_size = 9)
ggplot(datos, aes(x = SEXO, y = EL, fill = SEXO)) +
  geom_boxplot(
    alpha = 0.6,
    outlier.shape = 16) +
  geom_jitter(
    width = 0.15,
    alpha = 0.6,
    size = 1,
    color = "grey40") +
  stat_summary(
    fun = mean,
    geom = "point",
    size = 4,
    color = "red"
  ) +
  labs(
    title = "Dimorfismo sexual en longitud de los élitros",
    x = "Sexo",
    y = "EL (mm)"
  ) +
  theme_minimal() +
  theme(legend.position = "none") +
  theme(
```

```r
    plot.title = element_text(size = 10),
    axis.title = element_text(size = 8),
    axis.text  = element_text(size = 7)
  )
datos %>%
  group_by(PROVINCIA) %>%
  summarise(
    N_validos = sum(!is.na(KB)),
    Pct_validos = N_validos / n() * 100,
    Media = mean(KB, na.rm = TRUE),
    Mediana = median(KB, na.rm = TRUE),
    SD = sd(KB, na.rm = TRUE),
    Min = min(KB, na.rm = TRUE),
    Max = max(KB, na.rm = TRUE)
  ) %>%
  kable(
    caption = "Resumen estadístico de la anchura de la cabeza (KB) según
     ↪  provincia",
    digits = 2,
    booktabs = TRUE
  ) %>%
  kable_styling(
    latex_options = c("hold_position", "striped")
  )
anova_KB <- aov(KB ~ PROVINCIA, data = datos)

res_anova <- summary(anova_KB)[[1]]

tabla_anova_KB <- data.frame(
  est_f = res_anova["PROVINCIA", "F value"],
  gl_1 = res_anova["PROVINCIA", "Df"],
  gl_2 = res_anova["Residuals", "Df"],
  p = res_anova["PROVINCIA", "Pr(>F)"]
)

tabla_anova_KB |>
  kable(
    col.names = c(
      "Estadístico F",
      "Grados de libertad entre grupos",
      "Grados de libertad dentro de grupos (residuales)",
      "p-Valor"
    ),
    caption = "ANOVA de un factor para la anchura de la cabeza (KB) según
     ↪  procedencia geográfica",
    digits = 6,
    booktabs = TRUE,
    align = "c"
```

```r
  ) |>
  kable_styling(
    latex_options = c("hold_position","striped"),
    font_size = 9
  )

tuk <- TukeyHSD(anova_KB, which = "PROVINCIA")
tuk_tab <- as.data.frame(tuk$PROVINCIA)
tuk_tab$Comparación <- rownames(tuk_tab)
rownames(tuk_tab) <- NULL

tuk_tab <- tuk_tab |>
  dplyr::select(Comparación, diff, lwr, upr, `p adj`)

tuk_tab |>
  knitr::kable(
    col.names = c("Comparación", "Diferencia medias", "IC 95% (inf)", "IC 95%
     ↪  (sup)", "p-Valor ajustado"),
    digits = 6,
    booktabs = TRUE,
    align = "c",
    caption = "Comparaciones múltiples (Tukey HSD) para KB por provincia"
  ) |>
  kableExtra::kable_styling(
    latex_options = c("hold_position", "striped"),
    font_size = 9
  )

ggplot(datos, aes(x = PROVINCIA, y = KB, fill = PROVINCIA)) +
  geom_boxplot(
    alpha = 0.6,
    outlier.shape = 16) +
  geom_jitter(
    width = 0.15,
    alpha = 0.6,
    size = 1,
    color = "grey40") +
  stat_summary(
    fun = mean,
    geom = "point",
    size = 4,
    color = "red"
  ) +
  labs(
    title = "Anchura de la cabeza por procedencia geográfica",
    x = "Provincia",
    y = "KB (mm)"
  ) +
```

```r
  theme_minimal() +
  theme(legend.position = "none") +
  theme(
    plot.title = element_text(size = 10),
    axis.title = element_text(size = 8),
    axis.text  = element_text(size = 7)
  )
datos %>%
  group_by(PROVINCIA) %>%
  summarise(
    N_validos = sum(!is.na(EL)),
    Pct_validos = N_validos / n() * 100,
    Media = mean(EL, na.rm = TRUE),
    Mediana = median(EL, na.rm = TRUE),
    SD = sd(EL, na.rm = TRUE),
    Min = min(EL, na.rm = TRUE),
    Max = max(EL, na.rm = TRUE)
  ) %>%
  kable(
    caption = "Resumen estadístico de la longitud de los élitros (EL) según
     ↪  provincia",
    digits = 2,
    booktabs = TRUE
  ) %>%
  kable_styling(
    latex_options = c("hold_position", "striped")
  )
anova_EL <- aov(EL ~ PROVINCIA, data = datos)

res_anova <- summary(anova_EL)[[1]]

tabla_anova_EL <- data.frame(
  est_f = res_anova["PROVINCIA", "F value"],
  gl_1 = res_anova["PROVINCIA", "Df"],
  gl_2 = res_anova["Residuals", "Df"],
  p = res_anova["PROVINCIA", "Pr(>F)"]
)

tabla_anova_EL |>
  kable(
    col.names = c(
      "Estadístico F",
      "Grados de libertad entre grupos",
      "Grados de libertad dentro de grupos (residuales)",
      "p-Valor"
    ),
    caption = "ANOVA de un factor para la longitud de los élitros (EL) según
     ↪  procedencia geográfica",
```

```r
    digits = 6,
    booktabs = TRUE,
    align = "c"
  ) |>
  kable_styling(
    latex_options = c("hold_position","striped"),
    font_size = 9
  )

tuk <- TukeyHSD(anova_EL, which = "PROVINCIA")
tuk_tab <- as.data.frame(tuk$PROVINCIA)
tuk_tab$Comparación <- rownames(tuk_tab)
rownames(tuk_tab) <- NULL

tuk_tab <- tuk_tab |>
  dplyr::select(Comparación, diff, lwr, upr, `p adj`)

tuk_tab |>
  knitr::kable(
    col.names = c("Comparación", "Diferencia medias", "IC 95% (inf)", "IC 95%
     ↪  (sup)", "p-Valor ajustado"),
    digits = 6,
    booktabs = TRUE,
    align = "c",
    caption = "Comparaciones múltiples (Tukey HSD) para EL por provincia"
  ) |>
  kableExtra::kable_styling(
    latex_options = c("hold_position", "striped"),
    font_size = 9
  )

ggplot(datos, aes(x = PROVINCIA, y = EL, fill = PROVINCIA)) +
  geom_boxplot(
    alpha = 0.6,
    outlier.shape = 16) +
  geom_jitter(
    width = 0.15,
    alpha = 0.6,
    size = 1,
    color = "grey40") +
  stat_summary(
    fun = mean,
    geom = "point",
    size = 4,
    color = "red"
  ) +
  labs(
    title = "Longitud de los élitros por procedencia geográfica",
```

```r
    x = "Provincia",
    y = "EL (mm)"
  ) +
  theme_minimal() +
  theme(legend.position = "none") +
  theme(
    plot.title = element_text(size = 10),
    axis.title = element_text(size = 8),
    axis.text  = element_text(size = 7)
  )
datos_AC <- datos |>
  dplyr::filter(PROVINCIA %in% c("Asturias", "Cantabria"))

tabla_habitat <- table(datos_AC$PROVINCIA)

ic_habitat <- prop.test(
  x = c(tabla_habitat["Asturias"], tabla_habitat["Cantabria"]),
  n = c(sum(tabla_habitat), sum(tabla_habitat)),
  correct = FALSE
)

tabla_prop <- data.frame(
  Estadístico = paste0( round(ic_habitat$statistic, 3)),
  "p-valor" = signif(ic_habitat$p.value, 3),
  `IC 95` = paste0("[",
    round(ic_habitat$conf.int[1], 3), ",", round(ic_habitat$conf.int[2], 3),"]"
  )
)

tabla_prop |>
  knitr::kable(
    col.names = c("Estadístico chi-cuadrado", "p-valor", "IC 95%"),
    digits = 3,
    booktabs = TRUE,
    align = "c",
    caption = "Contraste de diferencia de proporciones para la representación de
     ↪ los hábitats entre Asturias y Cantabria (IC 95\\%)"
  ) |>
  kableExtra::kable_styling(
    latex_options = c("hold_position", "striped"),
    font_size = 9
  )

test_provincia <- ic_habitat
tabla_habitat <- table(datos$SEXO)

ic_habitat <- prop.test(
  x = c(tabla_habitat["macho"], tabla_habitat["hembra"]),
```

```
  n = c(sum(tabla_habitat), sum(tabla_habitat)),
  correct = FALSE
)

tabla_prop <- data.frame(
  Estadístico = paste0( round(ic_habitat$statistic, 3)),
  "p-valor" = signif(ic_habitat$p.value, 3),
  `IC 95` = paste0("[",
    round(ic_habitat$conf.int[1], 3), ",", round(ic_habitat$conf.int[2], 3),"]"
  )
)

tabla_prop |>
  knitr::kable(
    col.names = c("Estadístico chi-cuadrado", "p-valor", "IC 95%"),
    digits = 3,
    booktabs = TRUE,
    align = "c",
    caption = "Contraste de diferencia de proporciones para la representación de
     ↪  los hábitats entre sexos (IC 95\\%)"
  ) |>
  kableExtra::kable_styling(
    latex_options = c("hold_position", "striped"),
    font_size = 9
  )

test_sexo <- ic_habitat
datos |>
  count(PROVINCIA) |>
   mutate(porcentaje = n / sum(n) * 100) |>
  ggplot(aes(x = PROVINCIA, y = porcentaje)) +
  geom_col()+
  labs(
    title = "Distribución porcentual por provincia",
    x = "Provincia",
    y = "Porcentaje (%)"
  ) + scale_y_continuous(limits = c(0, 100))

datos |>
  count(SEXO) |>
   mutate(porcentaje = n / sum(n) * 100) |>
  ggplot(aes(x = SEXO, y = porcentaje)) +
  geom_col() +
  labs(
    title = "Distribución porcentual por sexo",
    x = "Sexo",
    y = "Porcentaje (%)"
  )+ scale_y_continuous(limits = c(0, 100))
```

```r
tabla_resumen <- data.frame(
  Analisis = c(
    "KB (por sexo)",
    "KB (por provincia)",
    "EL (por sexo)",
    "EL (por provincia)",
    "Contraste (provincia)",
    "Contraste (sexo)"
  ),

  Test = c(
    "t de Student",
    "ANOVA",
    "t de Student",
    "ANOVA",
    "Test de proporciones",
    "Test de proporciones"
  ),

  Estadistico = c(
    paste0("t = ", round(unname(test_KB_sexo$statistic), 3)),
    paste0("F = ", round(summary(anova_KB)[[1]]$`F value`[1], 3)),
    paste0("t = ", round(unname(test_EL_sexo$statistic), 3)),
    paste0("F = ", round(summary(anova_EL)[[1]]$`F value`[1], 3)),
    paste0("Chi-cuadrado = ", round(test_provincia$statistic, 3)),
    paste0("Chi-cuadrado = ", round(test_sexo$statistic, 3))
  ),

  `p-valor` = c(
    signif(test_KB_sexo$p.value, 8),
    signif(summary(anova_KB)[[1]]$`Pr(>F)`[1], 3),
    signif(test_EL_sexo$p.value, 8),
    signif(summary(anova_EL)[[1]]$`Pr(>F)`[1], 3),
    signif(test_provincia$p.value, 3),
    signif(test_sexo$p.value, 3)
  ),

  Resultado = c(
    "Concluyente",
    "Alto, no concluyente",
    "Concluyente",
    "Alto, no concluyente",
    "Concluyente",
    "Alto, no concluyente"
  )
)
```

```r
# Mostrar tabla
tabla_resumen |>
  kable(
    caption = "Resumen de los análisis estadísticos realizados",
    booktabs = TRUE,
    align = "l"
  ) |>
  kable_styling(
    latex_options = c("hold_position", "striped"),
    font_size = 9
  )
```