Informe N3

11.77 Estadística Aplicada I

Comisión C (Especial) - 20222Q

Grupo 5

Ezequiel Garcia Longo - Azul de los Angeles Makk

X= Longitud del Culmen de un pingüino Adelie macho del archipiélago Palmer (Antártida) tomado al azar (mm)

Y= Longitud del Culmen de un pingüino Adelie hembra del archipiélago Palmer (Antártida) tomado al azar (mm)

En primer lugar, identificamos una distribución Normal Estándar en la muestra trabajada. Todas las expresiones se asemejan a la misma de manera tal que:

- El coeficiente de asimetría es cercano a 0 y el coeficiente de curtosis es cercano a 3 para ambas muestras (Para la muestra X, el coeficiente de asimetría es 0.05642379 y el de la muestra Y es 0.02785804.
- La curtosis para la muestra X es 3.26313785075775 y para Y es 2.69051801021496).
- La mediana es similar a la media (la media de X es 40.39, y la mediana es 40.6; y para Y la media es 37.25 y la mediana es 37).
- Histograma de las muestras dan una imagen que se asemeja a una campana de Gauss, calculamos las estimaciones y los errores paramétricos con la distribución normal.

Como los parámetros de la distribución normal son la media y el desvío estándar, su cálculo no varía si se hace por el método de máxima verosimilitud o si se lo realiza por el método de momentos. A continuación, se incluyen los parámetros de la misma, así como sus expresiones y resultados numéricos.

Expresión	Resultado numérico
$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	Para X = 40,3904
	Para Y = 37,25753
$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$	Para X = 2,26148
	Para Y = 2,014939

Elegimos las 4 cantidades(q)

- q1= Estimar la diferencia entre las medias de las dos muestras poblacionales (m1-m2)

Forma paramétrica y no paramétrica
Utilizamos la expresión

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

para ambas muestras, y luego calculamos su diferencia $\hat{\mu}_1 - \hat{\mu}_2$

- q2= Estimar el cociente entre las dos varianzas poblacionales (v1/v2)

Forma paramétrica y no paramétrica
$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$ para ambas muestras, y luego calculamos el cociente V1/V2

Para q3 y q4, estandarizamos la función normal para luego calcular la probabilidad y el cuantil solicitado.

- q3= Estimar el cuantil 0.5 de ambas muestras poblacionales (mediana)

Forma paramétrica	**Forma no paramétrica**
$P(Z > r) = 1 - \phi(r) = 0.5$ z=0 x=40.39041 y=37.25753	$X_{0.5} = F^{-1}(0.5) = min\, \widehat{X_i}\, tq\, \{X_i : \hat{F}(Xi) \geq 0.5\}$

- q4= Estimar la P(X>41) y P(Y>37.5).

Forma paramétrica	**Forma no paramétrica**
x=41 → Zx=0.2695531 y=37.5 →Zy=0.1203341 $P(Z > 0.2695531) = 1 - \phi(0.2695531) = 0.393752$ $P(Z > 0.1203341) = 1 - \phi(0.1203341) = 0.4521093$	$\hat{F}(x) = \frac{\sum 1\{Xi = x\}}{n}$

Notas a tener en cuenta:

- q1 y q2 fueron propuestas por el profesor.
- q3 fue elegida ya que, al ser la mediana, permite compararla con la media y sacar conclusiones respecto a la distribución de la muestra.
- q4 es una probabilidad, que nos permite saber, en nuestro caso, qué probabilidad existe de que se supere cierto tamaño, lo que nos permite obtener más información sobre la muestra.

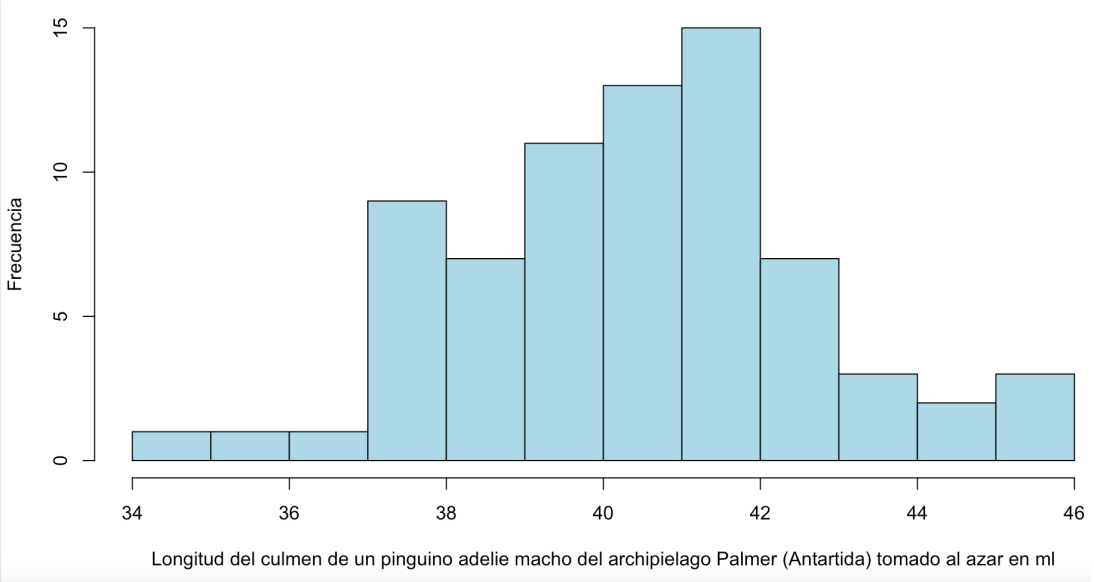

Figura I: Gráfica de variable X (Longitud del culmen de un pingüino adelie macho del archipiélago palmer (Antártida) tomado al azar en ml)

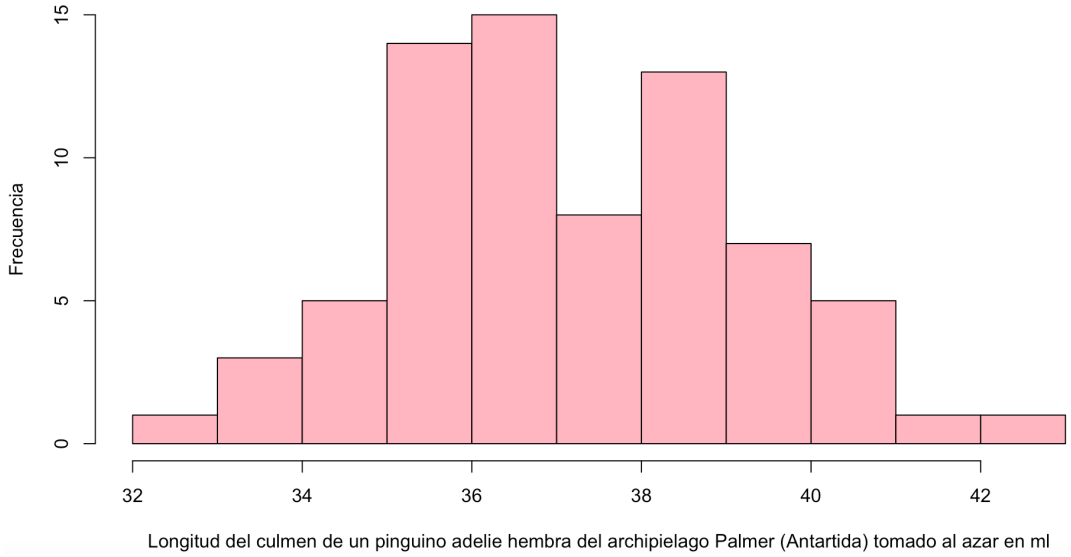

Figura II: Gráfica de variable Y (Longitud del culmen de un pingüino adelie hembra del archipiélago palmer (Antártida) tomado al azar en ml)

Cálculo de las estimaciones y sus respectivos errores

Q	Método no paramétrico	Método paramétrico
1	estimacion: 3.132877	
	error: 0.3573	
2	estimacion: 1.259685	

Q		Método no paramétrico	Método paramétrico
		error: 0.3026	
3	X	estimación: 40.6	estimación: 40.39041
		error: 0.3023286	error: 0.2612089
	Y	estimación: 37	estimacion: 37.25753
		error: 0.415578	error: 0.2330535
4	X	estimacion: 0.9041096	estimacion: 0.393752
		error: 0.0340644	error: 0.04656957
	Y	estimacion: 0.4520548	estimacion: 0.4521093
		error: 0.0593	error: 0.04651747

Cabe aclarar que q1 y q2 fueron calculados de igual manera para el método paramétrico como para el no paramétrico ya que, al ser los parámetros de la distribución normal la media y la varianza, se calculan de igual manera para ambos casos.

Comparativamente, analizando los resultados obtenidos en el caso de q3 podemos observar que este es mayor para la variable X tanto para su forma paramétrica como no paramétrica. De forma simultánea, esta situación se replica aún más significativamente para q4 no paramétrica, mientras que q4 para X es menor para q4 de la variable Y.

A su vez, el cálculo del error fue igual en ambos casos ya que si bien existen expresiones para el cálculo del error de la varianza y de la media para una distribución normal, no es así para la diferencia de medias o el cociente de varianzas. Por esa razón, se debe calcular por el método bootstrap, utilizando el mismo procedimiento que para el método no paramétrico.

No se observa una gran diferencia entre los resultados paramétricos y los no paramétricos. Esto nos indica que hay una gran correlación entre la muestra no parametrizada y su asociación a una distribución normal.

Los estimadores paramétricos son mejores que los no paramétricos, ya que su error es menor. Esto se debe a que, asumiendo una distribución, los cálculos probabilísticos son muy exactos, y no hay muchos valores atípicos, o que se alejen demasiado de la distribución propuesta, por lo que los cálculos realizados asumiendo esta distribución no se ven afectados por estos valores.

Respecto a los outliers, utilizamos el método de caja y bigotes para la determinación de la existencia de estos para ambas muestras. Mediante el mismo

creamos las variables xmod e ymod, las cuales contienen los datos que se encuentran entre el extremo inferior y superior. Comparando el tamaño de las muestras modificadas y el de la original, creamos la variable booleana que señala que en caso de que la variable creada sea de menor tamaño, implica que la misma posee outliers. Esto se probó con la muestra X, la cual posee un total de 5 outliers. De esta forma, al poseer outliers ello implica un impacto en el cálculo de los estimadores no robustos. Por otro lado, ellos se atribuyen como representativos al formar parte de la población,y no por afectarlos. Todos los cálculos están especificados en el código de R.