UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE ESTUDIOS ESTADÍSTICOS

TAREA EVALUABLE

MÓDULO: ESTADÍSTICA

Jorge González Perea 51553561G

Máster en Big Data, Data Science & Inteligencia Artificial

Curso académico 2024-2025

Índice

1. Introducción.

Se tiene un conjunto de 30 datos de anchuras de cráneos humanos para dos períodos prehistóricos (60 en total). Para el cálculo de las variables estadísticas de interés, como las varianzas, desviaciones estándar, medias, etc. se ha creado la librería *funciones_estadistica.py*, que contiene todas las funciones para llevar a cabo los análisis pedidos (medias, desviaciones, pruebas estadísticas, intervalos de confianza, etc.). El código de dicha librería se puede encontrar en un archivo con el mismo nombre.

2. Ejercicio 1.

2.1. Obtener con Python las diferentes medidas de centralización y dispersión, asimetría y curtosis estudiadas. Así mismo, obtener el diagrama de caja y bigotes. Se debe hacer por separado para la submuestra de los cráneos del predinástico temprano y para la submuestra de los del predinástico tardío. Comentar los resultados obtenidos. Estos comentarios son obligatorios.

Los valores de todas las variables estadísticas relevantes se han calculado con las fórmulas proporcionadas por la documentación y se pueden encontrar en la siguiente tabla:

Medida	Predinástico temprano	Predinástico tardío
Cantidad de datos	30	30
Media aritmética	131,53	132,47
Media geométrica	131,53	132,46
Mediana	131,50	133,00
Moda	131 y 132	133
Mínimo	130	131
25%	131,00	132,00
50%	131,50	133,00
75%	132,00	133,00
Máximo	134	135
Rango	4	4
Cuasivarianza	0,671	1,016
Cuasidesviación típica	0,819	1,008
Varianza	0,649	0,982
Desviación típica	0,806	0,991
Coeficiente de variación de Pearson (%)	0,612	0,748
Coeficiente de asimetría de Fisher (%)	0,657	0,195
Coeficiente de curtosis (%)	1,304	-0,186

Figura 1: Variables estadísticas para ambos períodos (en mm y potencias).

Además, el diagrama de caja y bigotes de cada muestra también proporciona información importante de forma intuitiva y más visual que ayuda a la comparación entre las dos muestras de datos.

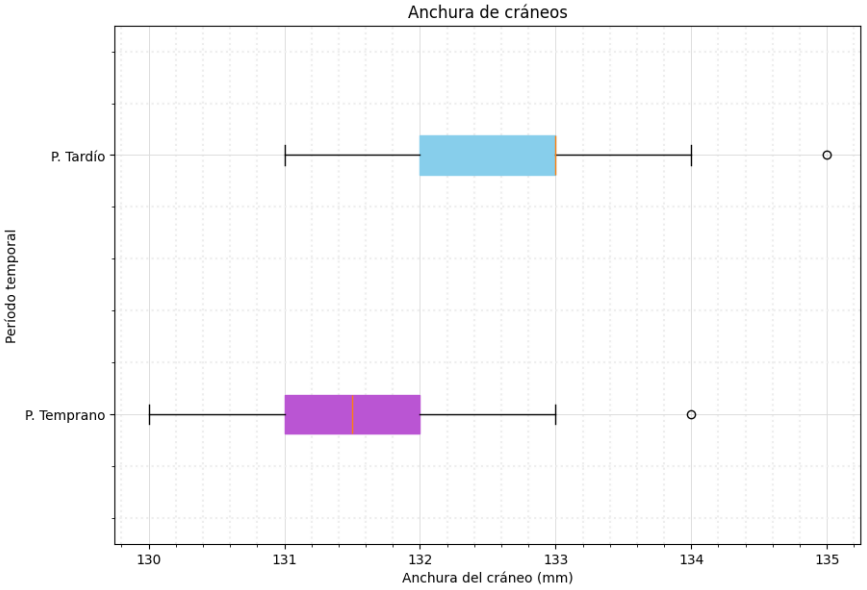

Figura 2: Diagrama de caja y bigotes de cada muestra.

Por otro lado los histogramas permiten visualizar el tipo de distribución de los datos y su simetría de forma rápida:

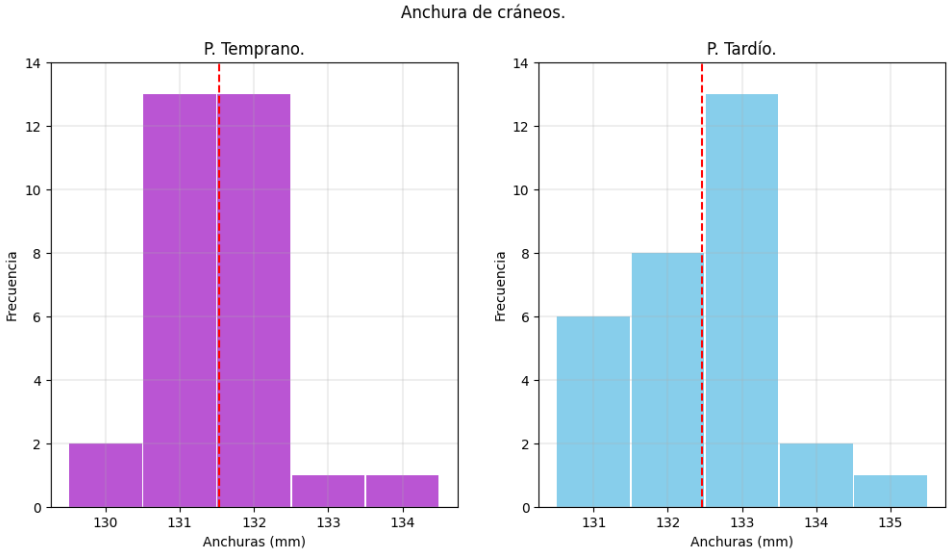

Figura 3: Histograma de cada muestra. Las líneas verticales representan las medias aritméticas.

Con los cálculos hechos y los datos representados es posible llevar a cabo un análisis estadístico de ambas muestras:

Período temprano:

- En el diagrama de caja para este período se puede observar cierta simetría en los cuartiles. Esta simetría está respaldada por el histograma, que es más simétrico para este período que para el predinástico tardío.
- Además, presenta un valor atípico que se aleja mucho de la mediana y sólo aparece una vez tal y como se puede comprobar en el histograma.
- Con respecto al coeficiente de asimetría de Fisher, este tiene un valor de **0.66 %** aproximadamente, lo que implica que se puede considerar como una muestra prácticamente simétrica al ser tan cercano al 0 (de hecho hay el mismo número de valores a la izquierda que a la derecha de la media).
- El coeficiente de curtosis para esta submuestra (**1.30 %**, por lo que se trata de una distribución leptocúrtica. Es decir, hay una mayor concentración de valores cerca de las medidas de centralización, tal y como se puede observar en el primer histograma.

Período tardío:

- A primera vista se observa en el diagrama de caja que el 3^{er} cuartil coincide con la mediana de la submuestra, y por tanto con el 2^o cuartil también, lo que significa que al menos 75 % de los datos está por debajo o son iguales a **133.00 mm**. Esto se debe a su elevada frecuencia relativa, que es de un **43.33 %**, por lo que el agrupamiento significativo en torno a este valor da lugar a estas coincidencias.
- Se observa un valor atípico que sólo se repite una vez, aunque está menos alejado de la mediana que el de la muestra anterior.
- El coeficiente de asimetría tiene un valor aproximado de **0.20 %**, ya que la diferencias con respecto a la media por la izquierda son mayores (más negativas) que en el caso anterior y por la derecha son similares, tal y como se puede observar en el segundo histograma.
- Según el coeficiente de curtosis de **-0.19 %**, los datos siguen una distribución platicúrtica, ya que los datos se concentran de forma más equilibrada en el histograma y no domina una concentración alrededor de la media.

Comparación:

- Se puede observar que en el período tardío los cráneos son en general más anchos, aunque los intervalos de ambas muestras se solapan en casi todo el recorrido tal y como se puede ver en la figura 1.
- Con respecto a las desviaciones respecto a la media, hay mayor dispersión en el período tardío.

2.2. Determinar si cada una de las dos sub-muestras sigue una distribución normal utilizando la prueba de Kolmogorov-Smirnov.

Los resultados obtenidos para la prueba de Kolmogorov-Smirnov para $\alpha = 0{,}05$ se pueden comprobar en la siguiente tabla:

Medida	P. Temprano	P. Tardío
K	0,2460	0,2381
p	0,0438	0,0557

Figura 4: Estadístico K y valor p de las muestras.

El valor crítico de la distribución de Kolmogorov-Smirnov para una muestra con $n = \mathbf{30}$ y $\alpha = 0{,}05$ es de **0.24170**. Por tanto, para la primera muestra se rechaza la hipótesis H_0, pero no hay evidencias para rechazarla en el caso de la segunda distribución. Es decir:

- Para el período temprano, a un nivel de confianza del **95 %** se rechaza la hipótesis nula ya que el estadístico calculado es mayor que el valor crítico y por tanto $p < \alpha$: **la primera muestra NO sigue una distribución normal.**
- En el caso del período tardío, a un nivel de confianza del **95 %** no hay evidencias para rechazar la hipótesis nula: **la segunda muestra sigue una distribución normal.**

3. Ejercicio 2.

3.1. Con los mismos datos del ejercicio anterior, obtener un intervalo de confianza (de nivel 0.90, de nivel 0.95 y de nivel 0.99) para la diferencia entre las medias de la anchura de la cabeza en ambos periodos históricos. Interpretar los resultados obtenidos y discutirlos en función del test de normalidad del ejercicio anterior. La interpretación debe ser rigurosa desde el punto de vista estadístico y también marcada por el story-telling, es decir, comprensible desde el punto de vista de las variables respondiendo a la pregunta ¿en qué época la cabeza era más ancha?

En este caso es necesario asumir que las muestras son independientes y que las varianzas de ambas poblaciones son desconocidas ($S_i^2,\ \ i = 1, 2$), ya que sólo se dispone de las varianzas **muestrales** ($\sigma_i^2,\ \ i = 1, 2$). El problema reside en comprobar si las varianzas **poblacionales** son iguales o no. Para ello se formulan las siguientes hipótesis:

- H_0: las varianzas poblacionales son iguales ($S_1^2 = S_2^2$)
- H_1: las varianzas poblacionales NO son iguales ($S_1^2 \neq S_2^2$)

Para comprobar si es necesario rechazar o no la hipótesis H_0 se puede tomar un intervalo de confianza del 90 %. El valor del estadístico F está comprendido en en siguiente intervalo:

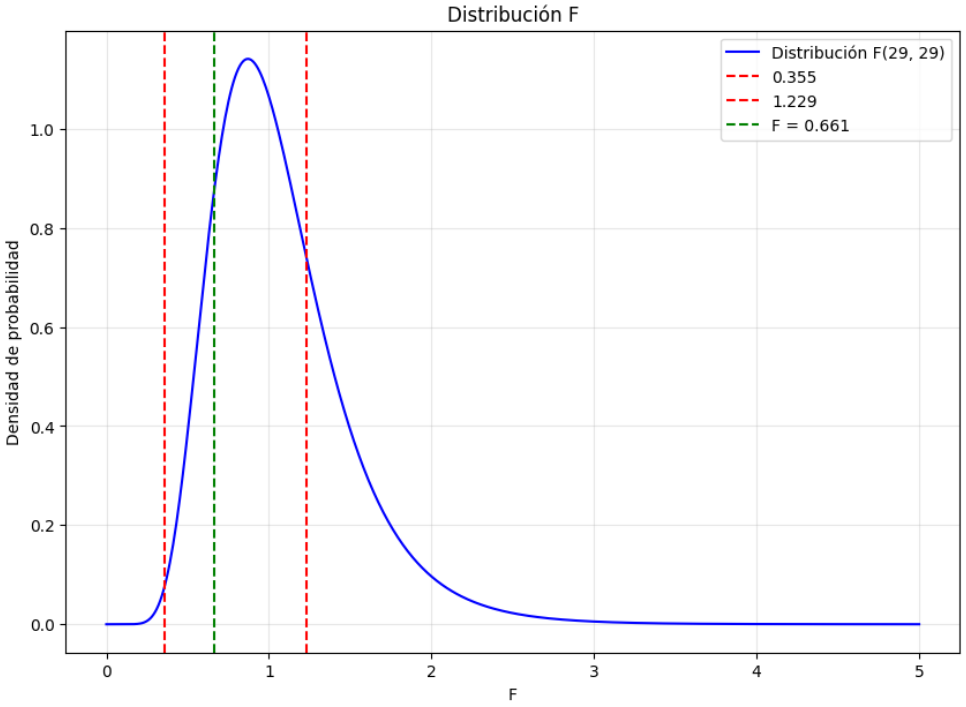

Figura 5: Distribución F para dos muestras con grados de libertad $g_1 = 29$ y $g_2 = 29$.

El valor concreto del estadístico, calculado como el cociente de las desviaciones típicas muestrales es $F = 1{,}081$, y el intervalo de aprobación (o más bien intervalo de no rechazo) de H_0 es, numéricamente aproximado: $(0{,}355,\ 1{,}229)$.

El error que da lugar al intervalo de confianza para la diferencias de medias $(\bar{x}_1 - \bar{x}_2)$ de dos muestras cuyas varianzas son desconocidas se calcula mediante la siguiente expresión:

$$err = t_{n_1+n_2-2,\frac{\alpha}{2}} \frac{\sqrt{(n_1\sigma_1^2 + n_2\sigma_2^2)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}{\sqrt{n_1 + n_2 - 2}} \tag{1}$$

Luego el intervalo de confianza, que depende del nivel escogido y por tanto del valor de la t de Student, será el siguiente:

$$((\bar{x}_1 - \bar{x}_2) - err \ \ , \ \ (\bar{x}_1 - \bar{x}_2) + err) \tag{2}$$

Aplicando esta fórmula a las medias y varianzas muestrales que aparecen en la primera tabla se obtienen los siguientes intervalos para cada nivel de confianza:

Nivel de confianza (%)	Intervalo (mm)
90	(-1.33, -0.537)
95	(-1.408, -0.459)
99	(-1.565, -0.302)

Figura 6: Intervalos de confianza para la diferencia de medias de ambas muestras.

Como se puede observar, los tres intervalos están en la parte negativa de la recta de los números reales, luego no hay evidencias de que las medias de ambas poblaciones (que no muestras) puedan

ser iguales, ya que el 0 no está incluido en ninguno de estos. Además, se puede concluir que la media de la primera población (predinástico temprano) es significativamente menor que la media de la segunda (predinástico tardío). Es decir, **se puede afirmar que las cabezas eran más anchas en la época tardía según los datos**.

Por otro lado, según las pruebas de Kolmogorov-Smirnov llevadas a cabo, la primera muestra no sigue una distribución normal, luego los intervalos de confianza no son tan fiables como podrían llegar a ser de lo contrario. Sin embargo, como las muestras son de gran tamaño ($n_i \geq 30, \;\; i = 1, 2$), se puede asumir que se cumple el teorema central del límite y que la falta de normalidad no afecta tanto a los resultados.

3.2. Utilizar el test t para contrastar la hipótesis de que ambas medias son iguales. Explicar qué condiciones se deben cumplir para poder aplicar ese contraste. Determinar si se cumplen. Admitiremos de forma natural la independencia entre ambas muestras, así que esa condición no hace falta comprobarla.

Para poder llevar a cabo el test t sobre dos muestras es necesario que se cumplan ciertas condiciones:

1. Normalidad. Las pruebas de Kolmogorov-Smirnov llevadas a cabo indican que la primera muestra no sigue una distribución normal y la segunda sí.
2. Igualdad de varianzas. Aunque las varianzas muestrales sean diferentes, las varianzas poblacionales se pueden considerar iguales gracias al cálculo del estadístico F con un nivel de confianza del 90 %.
3. Independencia. Según los enunciados se puede asumir la independencia de las muestras.

La primera condición no se cumple para la muestra del período predinástico temprano. De todas formas, se va a realizar el test t a pesar de que los resultados no sean válidos.

Las hipótesis que se quieren contrastar son las siguientes:

- H_0: las medias poblacionales son iguales ($\mu_1 = \mu_2$).
- H_1: las medias poblacionales son diferentes: ($\mu_1 \neq \mu_2$).

Para llevar a cabo el test t se ha escogido un nivel de confianza del 95 % y $g = n_1 + n_2 - 2$ grados de libertad. Para ello basta con repetir el cálculo del apartado anterior pero, en lugar de con una distribución F, emplear una distribución T:

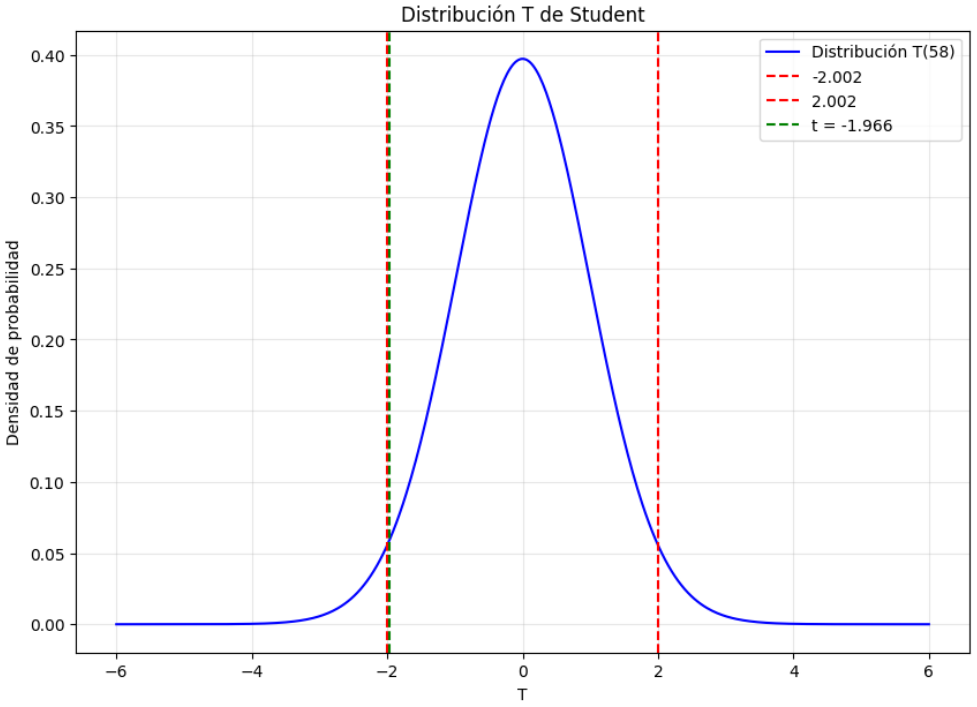

Figura 7: Distribución T para dos muestras con grados de libertad $g_1 = 29$ y $g_2 = 29$.

Donde se ha obtenido un valor $t = -1{,}966$ en un intervalo de confianza del 95 % de $(-2{,}002\ ,\ -2{,}002)$. Es decir, según el test t **no hay evidencias para rechazar la hipótesis** H_0. Por tanto se puede concluir que las medias de ambas poblaciones son iguales para este nivel de confianza y que, en términos del ancho de las cabezas, estas tienen la misma anchura tanto en el período temprano como en el tardío.

Sin embargo, esto contradice a los resultados del apartado anterior, en el que se ha llegado a la conclusión de que la media de la segunda muestra es significativamente mayor que la media de la primera (y por tanto que las cabezas se ensancharon con el tiempo). Esta contradicción se debe a la no normalidad de la muestra del período temprano, que es una condición que invalida la fiabilidad del test t y por tanto de sus resultados.

4. Conclusiones.

El análisis estadístico de estas muestras da a entender que las anchuras de los cráneos en el período predinástico temprano son, en promedio, menores que en el período predinástico tardío. Por otro lado, el test t del ejercicio 2 implica lo contrario, pero este resultado se puede ignorar debido a que la muestra del período temprano no sigue una distribución normal (ver figura 4), ya que esto es una condición necesaria para que la prueba sea válida.

En resumen, **se puede afirmar que los cráneos son significativamente más anchos en el período tardío**.

Nota: para el cálculo de las variables estadísticas de este análisis se han empleado la varianza y la desviación estándar, y no la cuasivarianza o cuasidesviación estándar.