

## Ejercicio de Evaluación. Descriptiva e Inferencia.

*Ejercicio 1. La siguiente tabla contiene, en un editable Excel, dos variables: la primera es dicotómica con valores 1 (predinástico temprano) y 2 (predinástico tardío) y la segunda contiene la anchura de cráneos (mm.) encontrados en un yacimiento arqueológico. La idea es analizar si existen diferencias en la longitud de la anchura de los cráneos egipcios a medida que pasa el tiempo. Creo que mayoritariamente tenemos una idea de que las cabezas egipcias son más alargadas y cuando ya llegamos a los romanos son más redondeadas. El cine se ha encargado de hacer muy gráfico todo esto.*

1	132
1	132
1	132
1	131
1	130
1	132
1	132
1	131
1	132
1	130

*Se pide:*

- a) *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 sub-muestra de los cráneos del predinástico temprano y para la sub-muestra de los del predinástico tardío. Comentar los resultados obtenidos. Estos comentarios son obligatorios.*

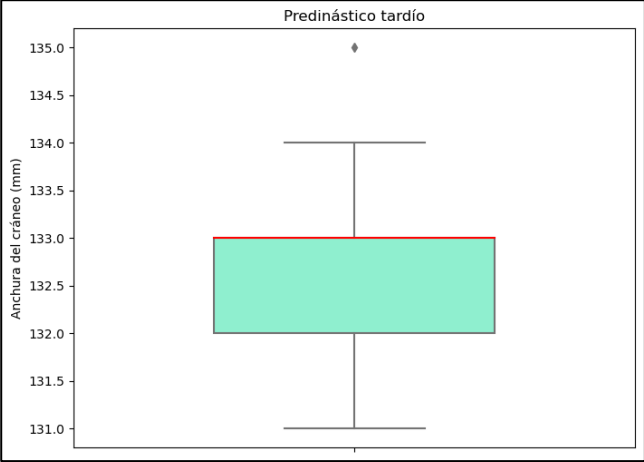
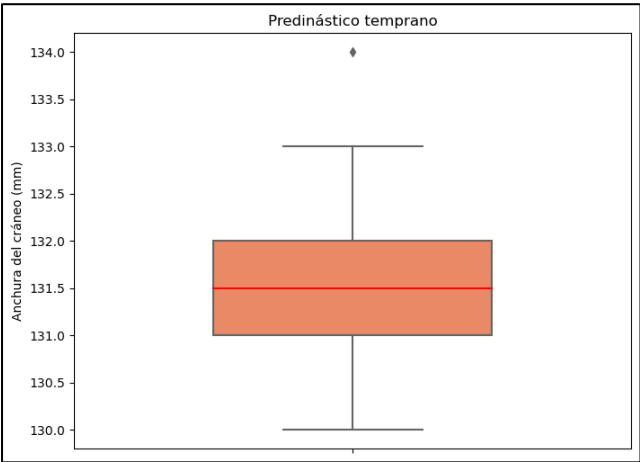
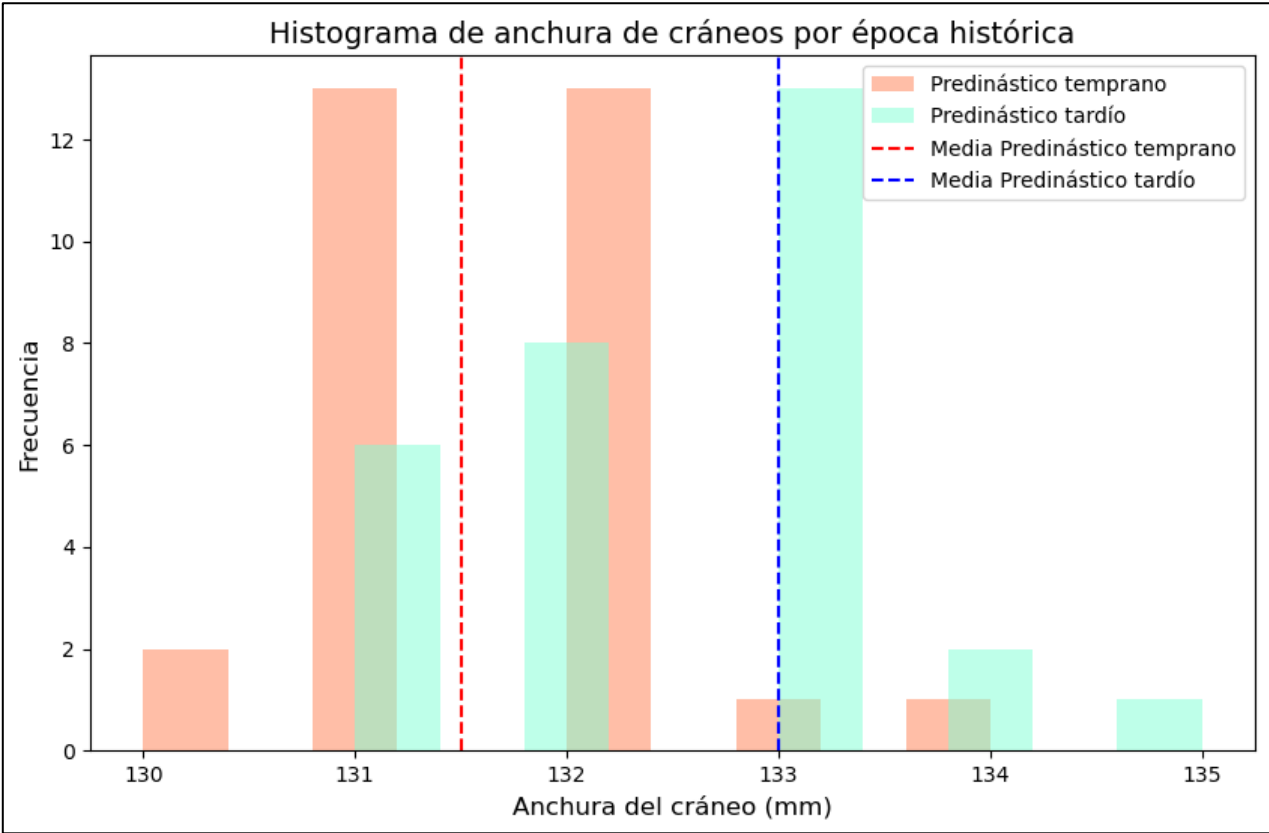
Calculamos las medidas de centralización y dispersión, asimetría y curtosis para las dos sub-muestras y representamos también el histograma para la anchura de los cráneos.

**Submuestra predinástico temprano**

**Submuestra predinástico tardío**

Medida			Medida		
0	Count	30.0	0	Count	30.0
1	Mean	131.53333333333333	1	Mean	132.46666666666667
2	Std	0.8193072487266863	2	Std	1.0080138659874618
3	Min	130.0	3	Min	131.0
4	25%	131.0	4	25%	132.0
5	50%	131.5	5	50%	133.0
6	75%	132.0	6	75%	133.0
7	Max	134.0	7	Max	135.0
8	Moda	131	8	Moda	133

9	Rango	4	9	Rango	4
10	Varianza	0.6488888888888891	10	Varianza	0.9822222222222224
11	Coeficiente Pearson	0.0061241996825111885	11	Coeficiente Pearson	0.007481665197442631
12	Coeficiente Fisher	0.6569830307466812	12	Coeficiente Fisher	0.19510554794191115
13	Coeficiente Curtosis	1.304372302495782	13	Coeficiente Curtosis	-0.18623697303495312



Una vez obtenidos todas las medidas, los histogramas y los diagramas de cajas, pasamos a analizarlos y comentar los resultados obtenidos.

#### **Predinástico temprano:**

- Podemos ver que la mediana en este grupo está muy céntrica y tiene un valor de 131.5.
- Encontramos un valor atípico observando el diagrama de cajas, el cual tiene un valor de 134.
- El coeficiente de asimetría tiene un valor de 0.6569. Es un valor positivo por lo que indica una ligera asimetría hacia la derecha en la distribución de los datos.
- El coeficiente de curtosis con un valor de 1.3043 (valor positivo) nos indica una mayor concentración de los datos alrededor de la media y una distribución leptocúrtica.

#### **Predinástico tardío:**

- En este grupo podemos observar tanto en la tabla de los estadísticos descriptivos como en el diagrama de cajas que su mediana (Q50) coincide con el cuartil 75. Esto quiere decir que el valor 133 se repite en el 75% de los datos. Además, este valor coincide con la moda (valor más frecuente en los datos).
- Encontramos otro caso de atípico con un valor de 135.
- El coeficiente de asimetría en este caso tiene un valor de 0.1951, menor que en el predinástico temprano, lo que sugiere una asimetría menos pronunciada hacia la derecha.
- En este el coeficiente de curtosis tiene un valor de -0.1862, valor negativo, por lo que a diferencia del predinástico temprano, nos indica una distribución menos concentrada alrededor de la media y una distribución platicúrtica.

Respecto al coeficiente de Pearson, en ambos casos es cercano a 0, por lo que indica que ambas distribuciones son relativamente simétricas. Además, observando la media de ambos grupos y los valores en los gráficos, podemos ver que, en el caso del predinástico tardío, la anchura de los cráneos es algo más mayor que en el predinástico temprano.

b) *Determinar si cada una de las dos sub-muestras sigue una distribución normal utilizando el test de Kolmogorov-Smirnov.*

Utilizaremos el contraste de bondad de ajuste de Kolmogorov-Smirnov para ver si los datos se distribuyen normalmente.

Tenemos que:

*$H_0$ : las distribuciones siguen una distribución normal*

*$H_i$ : las distribuciones no siguen una distribución normal*

Para ver si mis datos siguen una distribución  $N(0,1)$  primero debemos normalizar los datos. Esto sirve para que los datos estén centrados en 0 y tengan una dispersión comparable. Lo haremos de la siguiente manera: **datos normalizados** =  $\frac{\text{datos} - \text{media}}{\text{desv típica}}$

Seguidamente utilizaremos la función **kstest()** de Python para ver los resultados del test de Kolmogorov-Smirnov:

Resultados predinóstico temprano:  
estadístico: 0.2460415331404474  
pvalor: 0.04379464338101191  
El grupo temprano no sigue una distribución normal  
Resultados predinóstico tardío:  
estadístico: 0.23809252465886277  
pvalor: 0.05572704984817678  
El grupo tardío sigue una distribución normal

n	Test de Kolmogorov-Smirnov							
	Nivel de significación $\alpha$							
	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1	0.90000	0.95000	0.97500	0.99000	0.99500	0.99750	0.99900	0.99950
2	0.68337	0.77639	0.84189	0.90000	0.92929	0.95000	0.96838	0.97764
3	0.56481	0.63604	0.70760	0.78456	0.82900	0.86428	0.90000	0.92065
4	0.49265	0.56522	0.62394	0.68887	0.73424	0.77639	0.82217	0.85047
26	0.20399	0.23320	0.25908	0.28962	0.30963	0.33022	0.35431	0.37139
27	0.20030	0.22898	0.25438	0.28438	0.30502	0.32425	0.34794	0.36473
28	0.19680	0.22497	0.24993	0.27942	0.29971	0.31862	0.34190	0.35842
29	0.19348	0.22117	0.24571	0.27471	0.29466	0.31327	0.33617	0.35242
30	0.19032	0.21756	0.24170	0.27023	0.28986	0.30818	0.33072	0.34672

Como podemos observar, para el grupo del predinástico temprano tenemos un **estadístico** de **0.246** y un **p-valor** de **0.0437**. Como tenemos 30 observaciones, el valor crítico para  $\alpha=0.05$  es de 0.2417. Por lo que para un nivel de confianza del 95%, tenemos suficiente evidencia estadística para rechazar  $H_0$ , por lo que concluimos que los datos no siguen una distribución normal.

Por el contrario, para el grupo del predinástico tardío, con un **estadístico** de **0.238** y un **p-valor** de **0.0557** y con el mismo nivel de confianza, no hay suficiente evidencia para concluir que los datos no siguen una distribución normal.

Teniendo en cuenta los resultados y el análisis previo, podríamos considerar la eliminación de los dos valores atípicos encontrados en ambas sub-muestras. Esto podría ayudar a que estos valores no afecten negativamente a la normalidad de los datos. Debemos tener en cuenta que la eliminación de estos valores puede cambiar las medidas de tendencia central calculadas y la dispersión de los datos.

Realizamos antes un análisis de sensibilidad para comprobar cómo afectaría la eliminación de estos valores a los resultados y conclusiones. Para ello, volvemos a calcular el test de Kolmogorov-Smirnov sin esos valores y nos quedarían los siguientes resultados:

Resultados predinóstico temprano:  
estadístico: 0.27626019502650634  
pvalor: 0.018984169632381365  
El grupo temprano no sigue una distribución normal

Obtenemos un p-valor de 0.018, por lo que concluimos que los datos no siguen una distribución normal y además ha empeorado.

Podríamos quedarnos con los valores atípicos y considerar métodos estadísticos más robustos que sean menos sensibles a estas observaciones, pero en este caso, puesto que el p-valor del predinástico temprano se acerca a 0.05 (0.0437), consideramos que no es un gran problema y que podríamos considerar que esta distribución se distribuye normalmente.

## Ejercicio 2.

- a) Con los mismos datos del ejercicio anterior, obtener un intervalo de confianza (de nivel 0.9, 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?

Los intervalos de confianza nos dirán que probabilidad hay de que un parámetro desconocido se encuentre entre dos valores. Existen dos métodos para calcular los intervalos de confianza: Método de la función pivote y el método de Neyman.

Para ello primero debemos comprobar si las muestras son independientes (en el apartado b nos dice que asumamos independencia), y además demostrar si las varianzas poblacionales son iguales o diferentes.

Tenemos el siguiente contraste:

\*Siendo  $S_1$  = varianza poblacional sub-muestra predinástico temprano

$S_2$  = varianza poblacional sub-muestra predinástico tardío

$$H_0: S_1 = S_2$$

$$H_1: S_1 \neq S_2$$

Si  $H_0$  es cierto entonces:

$$\frac{\frac{\hat{S}_x^2}{\sigma_1^2}}{\frac{\hat{S}_y^2}{\sigma_2^2}} = \frac{\hat{S}_x^2}{\hat{S}_y^2} \rightarrow F_{n-1, m-1}$$

La región de aceptación es:

$$F_{c1} \leq \frac{\hat{S}_x^2}{\hat{S}_y^2} \leq F_{c2}$$

Tomando un intervalo del 90%, calculamos con Python estos valores y nos queda:

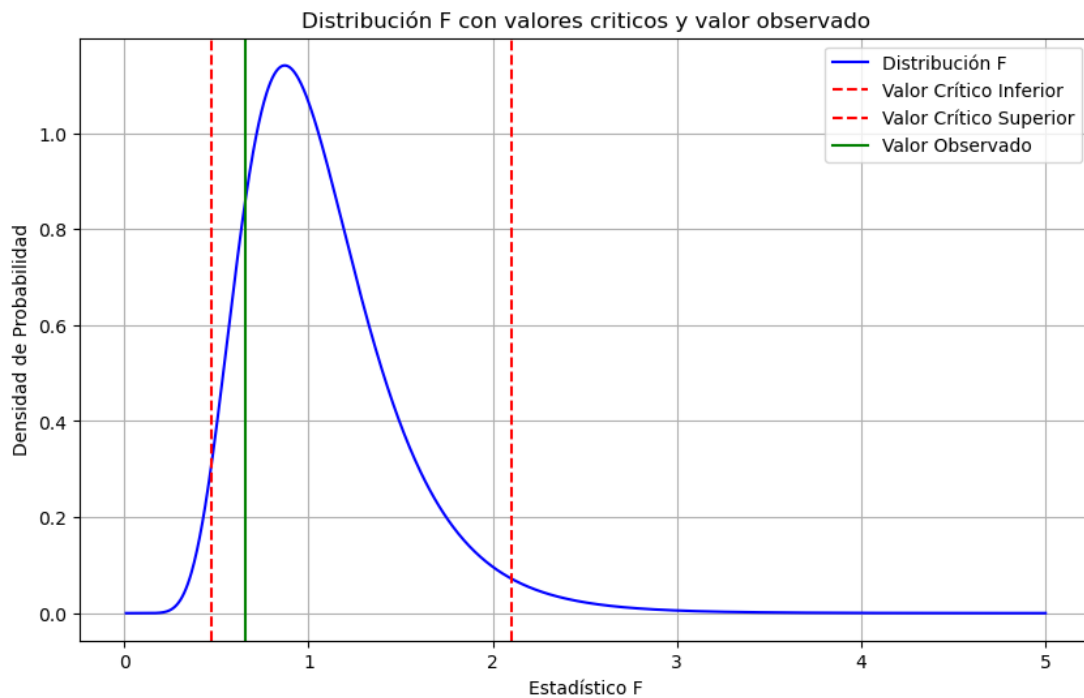
Valores Críticos:

Valor Crítico Inferior: 0.47596477431003137

Valor Crítico Superior: 2.100995817284212

Valor Observado (Estadístico F): 0.660633484162896

Valor p: 0.1350663442179744



El valor observado se encuentra dentro de la zona de aceptación de  $H_0$  por lo que podemos concluir que aun 90% de confianza, no existe evidencia estadística suficiente para rechazar  $H_0$ , por lo que las desviaciones estándar de ambas muestras son iguales.

Ahora podemos proceder con el cálculo de los intervalos de confianza. Para el caso de muestras independientes, distribuciones normales y varianzas poblacionales desconocidas pero iguales tenemos que:

$$IC_{\alpha}(\mu_1 - \mu_2) = \left( \bar{X} - \bar{Y} - t_{n_1+n_2-2, \alpha/2} \frac{\sqrt{(n_1 S_X^2 + n_2 S_Y^2)[(1/n_1) + (1/n_2)]}}{\sqrt{n_1 + n_2 - 2}}, \bar{X} - \bar{Y} + t_{n_1+n_2-2, \alpha/2} \frac{\sqrt{(n_1 S_X^2 + n_2 S_Y^2)[(1/n_1) + (1/n_2)]}}{\sqrt{n_1 + n_2 - 2}} \right)$$

- Para un nivel de confianza del 90%:

Intervalo de confianza del 90%: (-1.3365370178425786, -0.5301296488240956)

- Para un nivel de confianza del 95%:

Intervalo de confianza del 95%: (-1.4161777177709256, -0.45048894889574853)

- Para un nivel de confianza del 99%:

Intervalo de confianza del 99%: (-1.5757582313054241, -0.2909084353612502)

Estos intervalos nos proporcionan un rango de valores donde es probable que se encuentre la verdadera diferencia entre las medias poblacionales con un cierto nivel de confianza. Al observar los intervalos, vemos que todos son negativos y no contienen el valor 0, lo que indica que la diferencia entre las medias poblacionales es significativamente diferente de 0.

Interpretando estos resultados, concluimos que las medias poblacionales son diferentes entre los periodos temprano y tardío. Específicamente, los cráneos en el periodo tardío tienden a ser más anchos que en el periodo temprano, como lo indican las diferencias negativas en los intervalos de confianza.

*b) 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 aplicar el test t para contrastar la hipótesis de que ambas medias sean iguales, tenemos que tener en cuenta 3 cosas:

- Las muestras deben provenir de poblaciones normales o que la distribución muestral de las medias sea aproximadamente normal. Condición que ya verificamos en el ejercicio 1.
- Homogeneidad de la varianza, que se demostró en el anterior apartado.
- Independencia en los datos. El ejercicio nos exige asumir esta condición.

Entonces tenemos que:

$$H_0: \mu_1 = \mu_2$$

$$H_i: \mu_1 \neq \mu_2$$

$$\text{Si } \frac{|\bar{X} - \bar{Y}|}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} > t_{m+n-2, \alpha/2} \text{ se rechaza } H_0$$

ó

Equivalentemente, si  $P_{\text{valor}} = P \left( |t_{m+n-2}| > \frac{|\bar{X} - \bar{Y}|}{\sqrt{S_c \left( \frac{1}{m} + \frac{1}{n} \right)}} \right)$  es tal que  $P_{\text{valor}} < \alpha$  entonces, se rechaza  $H_0$

Con un nivel de confianza de 95%, procedemos a calcular los valores críticos con Python y nos queda:

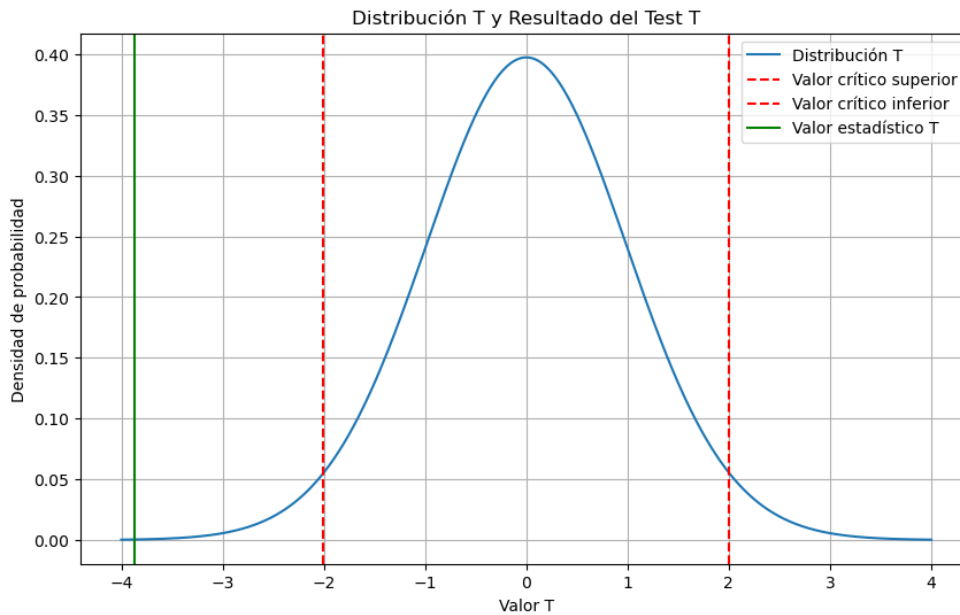
Valor crítico inferior: -2.0017174830120927

Valor crítico superior: 2.0017174830120923

Ahora calculamos el valor T de la forma:

$T = (\mu_1 - \mu_2) - (X_1 - X_2) / \text{errorStandar Estimado}$  sabiendo que  $(X_1 - X_2) = 0$  al ser las medias poblacionales iguales.

Con Python nos quedaría: Valor estadístico T: -3.8692997392678223



Al calcular el valor del estadístico T para la diferencia de medias entre los dos grupos, obtenemos un valor de -3.8693. Este valor está fuera de los límites establecidos para un nivel de confianza del 95%, lo que significa que rechazamos la hipótesis nula de que las medias de ambos grupos son iguales. En otras palabras, hay evidencia estadística para concluir que las medias poblacionales son diferentes.

En el contexto de nuestro análisis, esta diferencia en las medias implica que la anchura de los cráneos en el periodo tardío es mayor que en el periodo temprano. Este resultado puede ser importante para comprender las características físicas y antropológicas de las poblaciones en cada periodo histórico y puede dar información valiosa sobre la evolución y adaptación a lo largo del tiempo.