Módulo 6: Estadística

Alumno: Pablo Pérez Calvo

Ejercicio 1 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

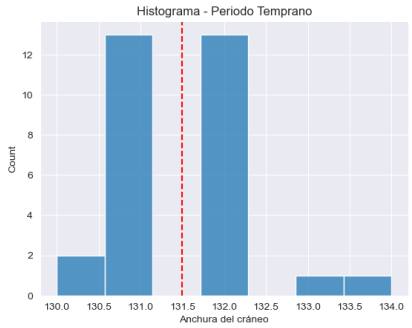
	Medida	Anchura del cráneo
0	count	30.0
1	mean	131.533333
2	std	0.819307
3	min	130.0
4	25%	131.0
5	50%	131.5
6	75%	132.0
7	max	134.0
8	moda	[131, 132]
9	rango	4
10	varianza	0.671264
11	CoeficientePerson	0.006229
12	CoeficienteFisher	0.645941
13	CoeficienteCurtosis	1.160893

(a) Predinástico temprano

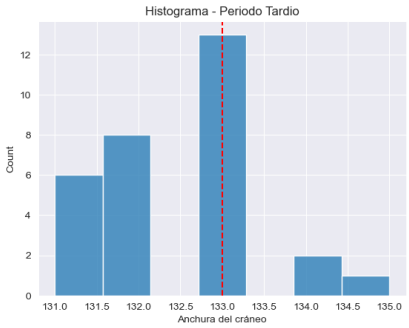
	Medida	Anchura del cráneo
0	count	30.000000
1	mean	132.466667
2	std	1.008014
3	min	131.000000
4	25%	132.000000
5	50%	133.000000
6	75%	133.000000
7	max	135.000000
8	moda	133.000000
9	rango	4.000000
10	varianza	1.016092
11	CoeficientePerson	0.007610
12	CoeficienteFisher	0.191826
13	CoeficienteCurtosis	-0.280029

(b) Predinástico tardío

Figura 1: Medidas calculadas

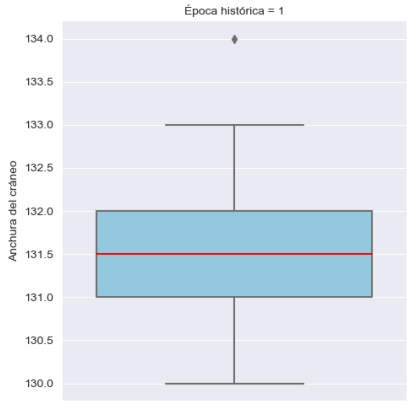


(a) Predinástico temprano

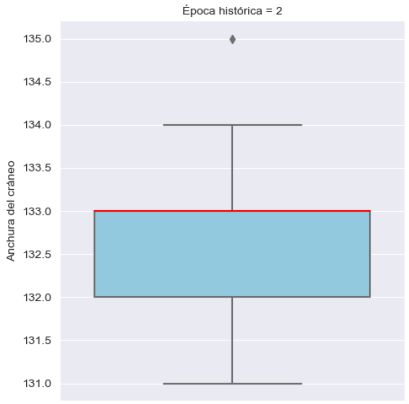


(b) Predinástico tardío

Figura 2: Histogramas



(a) Predinástico temprano



(b) Predinástico tardío

Figura 3: Diagrama de cajas y bigotes

Una vez obtenidas las tablas con las medidas correspondientes, y a su vez, los gráficos de histograma y cajas y bigotes de la muestra, podemos destacar las siguientes características:

Período temprano

- Presenta bimodalidad, destacan en términos de frecuencia dos valores: 131 y 132.
- Al ser **el coeficiente de Fisher** $0{,}6459 > 0$ indica una asimetría positiva, la distribución posee mayor concentración de valores a la derecha de la mediana, como podemos observar en el histograma 2a.
- El **coeficiente de curtosis** al ser $1{,}16089 > 0$ nos indica una distribucion

leptocúrtica, valores muy concentrados alrededor de la mediana, como se puede observar claramente en el histograma 2a.

- El valor de la mediana es muy cercano al valor de la media y en el diagrama de cajas y bigotes 3a podemos observar que la mediana se encuentra centrada en la caja, esto indica que la distribución es apróximadamente simétrica. Aunque con una ligera concentración a la derecha como se ha mencionado anteriormente.
- Esta distribución presenta un valor atípico, 134, que está alejado de la mediana y se identifica en el diagrama de cajas y bigotes 3a.

Periodo tardío

- La mediana (Quartil 50) es igual al Quartil 75, además en el diagrama 3b se puede ver que coinciden en la misma línea, el valor 133 se repite en el $75\,\%$ de los datos.
- En ese diagrama también podemos identificar un valor atípico, el 135.
- El **coeficiente de Fisher** es $0{,}1918 > 0$, la distribución posee mayor concentración de valores a la derecha de la mediana. En el histograma 2b parece que hay más valores a la izquierda pero hay que recordar que el 133 es moda, mediana y Q75.
- El **coeficiente de curtosis** es $-0{,}2800 < 0$ nos indica una distribución platicúrtica, es decir, valores menos concentrados alrededor de la mediana a comparación de la distribución normal y más distribuidos en las colas.

Comparativa de diagramas de cajas de ambas muestras

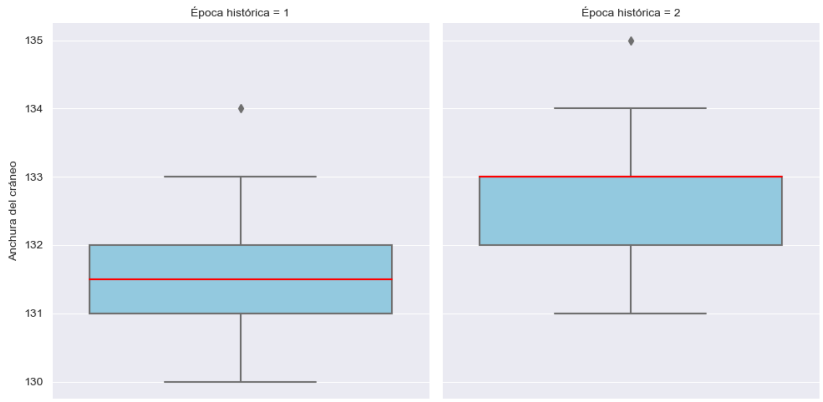


Figura 4: Comparación de ambos diagramas de cajas y bigotes

Al comparar ambos diagramas, podemos deducir rápidamente que el ancho de los cráneos en el periodo tardío es mayor al de los cráneos en el periodo temprano. El

valor de la mediana del tardío es mayor a la del temprano. En los próximos ejercicios seguiremos analizando esta cuestión para llegar a una conclusión más reforzada.

Ejercicio 1 b)

Determinar si cada una de las dos sub-muestras sigue una distribución normal utilizando el test de KolmogorovSmirnov

Para determinar si las dos sub-muestras siguen una distribución normal utilizando el test de Kolmogorov-Smirnov, planteamos las siguiente hipótesis.

- H_0: Los datos siguen una distribución normal.
- H_1: Los datos no siguen una distribución normal.

Luego normalizamos los valores

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Para terminar, con el código de Python obtenemos los siguientes resultados.

Resultados de la submuestra 1*:*
Estadístico: $0{,}2460415331404474$*, Valor* p*:* $0{,}04379464338101191$
La submuestra 1 *no sigue una distribución normal.*

Resultados de la submuestra 2*:*
Estadístico: $0{,}23809252465886277$*, Valor* p*:* $0{,}05572704984817678$
La submuestra 2 *sigue una distribución normal.*

Primera muestra - La muestra posee un estadístico $D_{observado} = 0,2460$ y el valor crítico según la tabla de K-SMR para una muestra de tamaño $n = 30$ y un nivel de significancia $\alpha = 0{,}05$ es $D_{tabla} = 0{,}2417$. Como $Dobservado > Dtabla$ se rechaza la hipotesis nula (H_0).

Por lo tanto,con un nivel de confianza de $95\,\%$ existe suficiente evidencia para concluir que la submuetra 1 no sigue una distribución normal. Adicionalmente podemos determinar que $pvalor = 0{,}04 < 0{,}05$, lo que refuerza el rechazo de la H_0.

Segunda muestra - La muestra posee un estadístico de $D_{observado} = 0{,}2381$ y el valor crítico según la tabla de K-SMR para una muestra de tamaño $n = 30$ y un nivel de significancia $\alpha = 0{,}05$ es $D_{tabla} = 0{,}2417$. Como $Dobservado < Dtabla$ no se rechaza la hipotesis nula (H_0).

Por lo tanto, con un nivel de confianza de $95\,\%$ no existe suficiente evidencia para rechazar la H_0 luego la submuestra 2 sigue una distribución normal. Además, podemos determinar que el $pvalor$ es $0{,}06 > 0{,}05$ lo que confirma que H_0 no puede ser rechazada.

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 alta?

Para abordar este problema se requiere:

1. Saber si las muestras son independientes - sin embargo el ejericico nos comenta que asumamos esta condición.
2. Sabemos que las varianzas poblacionales son desconocidas.
3. Demostrar si las varianzas de las poblaciones son iguales o diferentes.

Nos falta por demostrar si las varianzas poblacionales son iguales o diferentes, plantearemos las siguientes hipótesis.

- H_0: Las muestras poseen varianzas poblacionales iguales $S1 = S2$.
- H_1: Las muestras poseen varianzas poblacionales diferentes $S1 \neq S2$.

Tomaremos un intervalo de confianza del $90\,\%$, por lo que $\alpha = 0{,}1$. Como esta prueba es de 2 colas, ya que probamos igualdad o diferencia, se tomará $\frac{\alpha}{2} = 0{,}05$ para obtener los valores críticos.

Aunque por el ejercicio anterior concluimos que con un $95\,\%$ de confianza la submuestra 1 no seguía una distribución normal, asumiremos normalidad para realizar este ejercicio. Por este motivo, los resultados que se obtengan deberán ser interpretados con precaución.

Dado que las muestras son de tamaño $n = 30$, los grados de libertad para ambas serán de 29. Realizando los cálculos necesarios en *Python*, obtenemos los siguientes resultados.

La zona de aprobación de la H_0 es: $0{,}5373999648406917 < F < 1{,}8608114354760754$

Además, calculando el coeficiente de Fisher $F = \frac{s_1^2}{s_2^2}$, siendo s_1 la variaza de la primera muestra y s_2 la de la segunda, obtenemos.

$F = 0{,}6606334841628961$

Figura 5: Distribución F

Como podemos observar el valor obtenido se encuentra en la zona de aceptación de la H_0. Podemos concluir a partir del gráfico y los valores obtenidos que a un $90\,\%$ de confianza, no existe evidencia estadística suficiente como para rechazar la H_0. Por lo tanto, las desviaciones estándar de ambas muestras son iguales.

Ahora sabiendo que las muestras son independientes, las varianzas desconocidas y además estas son iguales, podemos calcular el intervalo de confianza para la diferencia de medias.

Para la condición de muestras independientes, muestras normales y varianzas poblacionales desconocidas pero iguales tenemos.

$$\text{ERROR ESTÁNDAR ESTIMADO} = \frac{\sqrt{\left(n_1 S_X^2 + n_2 S_y^2\right)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}{\sqrt{n_1 + n_2 - 2}}$$

Con esto nos falta calcular el valor crítico T con grados de libertad $(n1 + n2 - 2)$ y un alfa dependiendo de que intervalo de confianza adoptemos. Con ayuda de *Python* tendremos el siguiente resultado:

Los intervalos de confianza para la diferencia de medias con un $90\,\%$ de confianza es: $(-1{,}3365370178425788, -0{,}5301296488240955)$.

Los intervalos de confianza para la diferencia de medias con un $95\,\%$ *de confianza es:* $(-1{,}4161777177709258, -0{,}4504889488957485)$.

Los intervalos de confianza para la diferencia de medias con un $99\,\%$ *de confianza es:* $(-1{,}5757582313054241, -0{,}2909084353612502)$.

Dado que en los intervalos para los 3 niveles de confianza determinados $(90 - 95 - 99)$ no está incluido el 0, hay evidencia suficiente para concluir que las medias son diferentes. Además, como todos los intervalos de confianza de diferencia de medias $(X_1 - X_2)$ incluyen valores negativos, podemos concluir que la submuestra 2 tiene una media de anchura mayor a la submuestra 1. En otras palabras, los cráneos del periodo tardío son más anchos que los del periodo temprano.

Desde el punto de vista histórico, el aumento de la anchura de los cráneos a lo largo del tiempo puede verse debido a unas mayores condiciones de vida y mayor acceso a alimentos esenciales para el desarrollo óseo, así como posibles interacciones genéticas con otros grupos humanos en el periodo tardío que haya podido favorecer el aumento de la anchura.

Hay que recordar que los resultados obtenidos e interpretaciones se han de tomar con precaución, ya que para este ejercicio hemos supuesto que la submuestra 1 seguía una distribución normal. Cuando en la anterior actividad, demostramos que no seguía una distribución normal con un $95\,\%$ de confianza.

Ejercicio 2 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. Observación: Quiero insistir en que debéis hacer el test t para la diferencia de medias aunque las condiciones no se cumplan. En ese caso discutir la validez de los resultados obtenidos

Para poder probar la diferencia de medias de dos poblaciones. debemos considerar que se deben cumplir 3 condiciones:

1. Normalidad de los datos (con la prueba de kolmogrov- smirnoy se probó que una de las muestras sigue una Distribución Normal, para realizar este ejercicio supondremos que la muestra 1 también sigue una distribución normal).
2. Homogeneidad de la varianza (se demostró en el ejercicio anterior que las varianzas son homogéneas con prueba F)
3. Independencia de las observaciones (el ejercicio nos exige que esta condición ya se cumple).

Como se satisfacen las 3 condiciones, planteamos las siguientes hipótesis.

- H_0: Las medias de ambas muestras son iguales $M_0 = M_1$.
- H_1: Las medias de las muestras son diferentes $M_0 \neq M_1$.

Como se trata de una prueba de hipótesis para comprobar si son iguales o distintas, se trata de un prueba de dos colas. Tomaremos un intervalo de confianza del $95\,\%$ por lo que $\alpha = 0{,}025$ a cada lado de la curva. Además, tomaremos los grados de libertad correspondientes a $N_1 + N_2 - 2$.

Junto al apoyo de *Python* calcularemos los valores críticos correspondientes a la distribución T.

la zona de aceptación para la prueba es : $-2{,}0017174830120927$, $2{,}0017174830120923$

Ahora calculamos el valor t de la siguiente forma:

$$t = \frac{(\overline{M}_1 - \overline{M}_2) - (\mu_1 - \mu_2)}{\text{Error Estándar Estimado}} = \frac{(\overline{M}_1 - \overline{M}_2)}{\text{Error Estándar Estimado}}$$

Se tiene que $(\mu_1 - \mu_2) = 0$ debido a que estamos probando que las medias de ambas muestras son iguales (H_0).

De nuevo con ayuda de *Python* y recordando que el error estándar estimado ya fue calculado en el ejercicio $2a)$, obtenemos el siguiente resultado.

El valor critico de T es : -3.869299739267822

Figura 6: Distribución t

Como podemos observar con el valor obtenido y el gráfico de la distribución, el valor crítico no se encuentra dentro del intervalo para aceptar la H_0. Concluimos que con un $95\,\%$ de confianza se rechaza H_0, por lo que las medias poblacionales son diferentes.

Como vemos en el gráfico, el valor observado se encuentra muy a la izquierda de los valores críticos, esto quiere decir que la anchura media de los cráneos del periodo tardío es mayor a la de los cráneos en el periodo temprano, reforzando la interpretación dada en los anteriores ejercicios.

Por último, quería recalcar que se ha realizado el test t para la diferencia de medias no cumpliéndose la normalidad para una de las muestras. Aunque el uso del test t es una aproximación útil hay que tomar los resultados con precaución, se recomienda el aumento del tamaño de las muestras para mejorar la validez en futuros estudios o utilizar tests no parámetricos.