TAREA ESTADÍSTICA

Carlos Matías Sáez

Máster big data, data science & inteligencia artificial

2023-2024

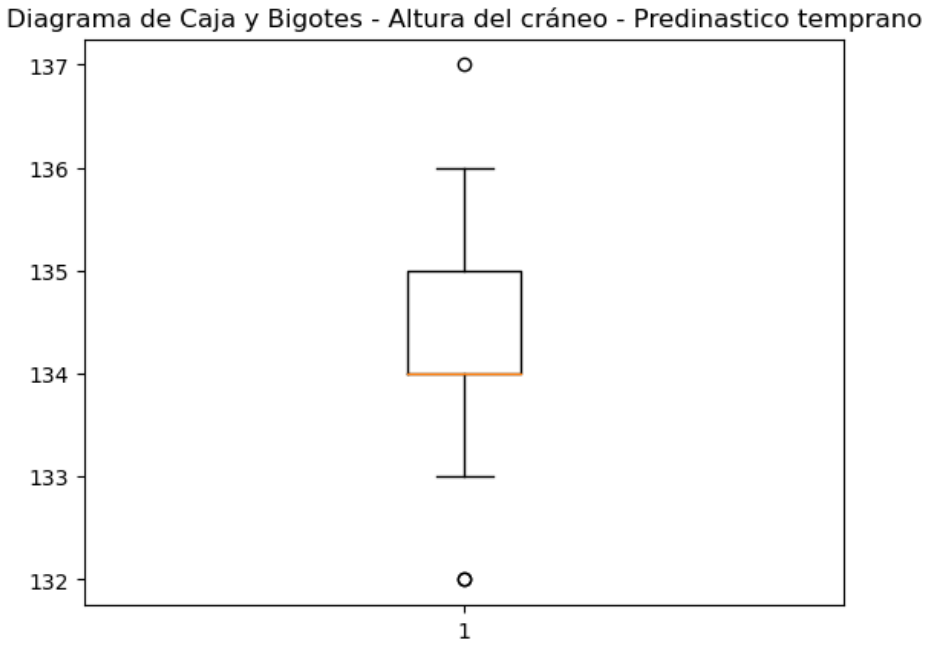
Universidad Complutense de Madrid

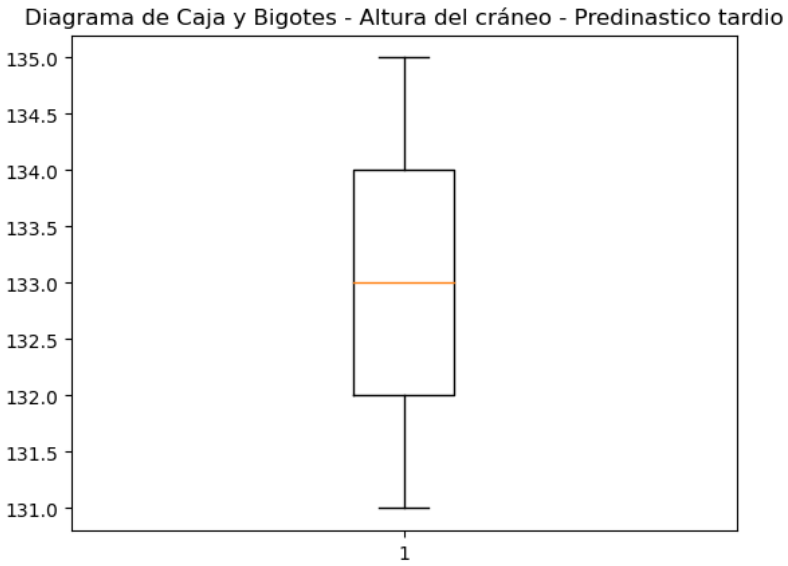
Ejercicio 1.

Resultados

Apartado a)

```
La media de los craneos tempranos es:  134.4
La media de los craneos tardios es:  132.9
-----------------------------------------------------
La mediana de los craneos tempranos es:  134.0
La mediana de los craneos tardios es:  133.0
-----------------------------------------------------
La moda de los craneos tempranos es:  134
La moda de los craneos tardios es:  133
-----------------------------------------------------
La varianza de los craneos tempranos es:  1.144827586206896
La varianza de los craneos tardios es:  1.0586206896551718
-----------------------------------------------------
La desviacion tipica de los craneos tempranos es:  1.069966161243848
La desviacion tipica de los craneos tardios es:  1.0288929437289245
-----------------------------------------------------
El rango de los craneos tempranos es:  5
El rango de los craneos tardios es:  4
-----------------------------------------------------
La curtosis de los craneos tempranos es:  1.0183245892347403
La curtosis de los craneos tardios es:  -0.5040834924441051
-----------------------------------------------------
La asimetría de los craneos tempranos es:  -0.1737311042975376
La asimetría de los craneos tardios es:  -0.19537866713306773
```

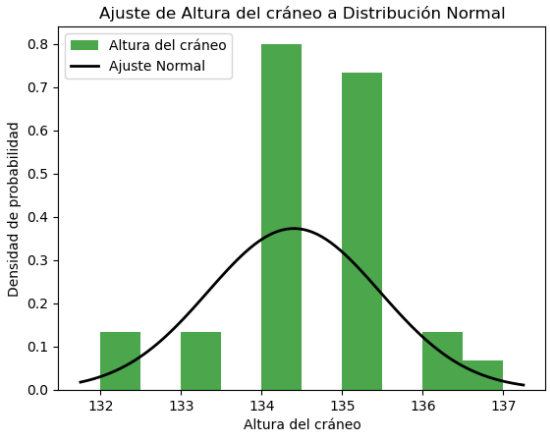




Apartado b)

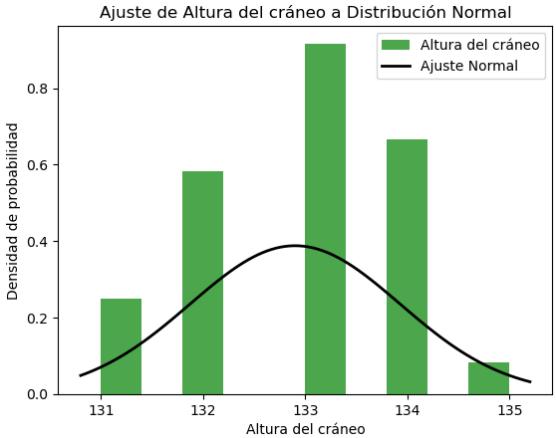
- Para cráneos tempranos:

Estadístico de Kolmogorov-Smirnov: 0.2209
P-valor: 0.0912



- Para cráneos tardíos:

Estadístico de Kolmogorov-Smirnov: 0.2054
P-valor: 0.1379



Comentarios:

Apartado a)

La diferencia observada entre las dos medias es de 1.5 unidades, la diferencia de 1.5 unidades representa una proporción relativamente pequeña de la variabilidad total en las alturas del cráneo.

Respecto a las medianas se puede observar cómo en ambas se cumple que son cercanas a las medias de forma que podemos sugerir que nos hay ningún valor desorbitado como podemos observar viendo el Excel.

La moda es prácticamente la media, esto sugiere que la distribución de los datos es aproximadamente simétrica o centrada alrededor de un solo valor central.

A través de la varianza y de la desviación típica podemos ver como son valores bajos, de forma que el conjunto de ambos datos está más concentrado alrededor de la media y hay poca variabilidad entre los valores individuales y la media. De otra forma, a partir del rango también se puede interpretar lo mismo.

Respecto a la curtosis vemos como en los cráneos tempranos es positiva, por tanto, la concentración de los datos alrededor de la media es mayor, lo que indica una mayor probabilidad de observar valores extremos. La curtosis positiva sugiere picos más altos y colas más pesadas en comparación con una distribución normal. En cambio, en los cráneos tardíos, la curtosis es negativa, por tanto, las colas de la distribución son más ligeras (menos concentradas) que las colas de una distribución normal, la concentración de los datos alrededor de la media es menor, lo que indica una menor probabilidad de observar valores extremos. La curtosis negativa sugiere picos más bajos y colas más ligeras en comparación con una distribución normal.

Respecto a la asimetría se observa como ambas son negativas, esto implica que la distribución está sesgada a la izquierda (tiene una cola más larga a la izquierda).

A partir de los diagramas se puede ver como los datos obtenidos son correctos, tales como la mediana (línea de color). El diagrama de los cráneos tempranos es menos largo que el de los cráneos tardíos, esto representa que la distribución de los datos de los cráneos tardíos es mayor que la otra.

Apartado b)

En cuanto a determinar si cada una de las dos sub-muestras sigue una distribución normal utilizando el test de Kolmogorov-Smirnov, hemos obtenido

distintos valores para ambas sub-muestras. Estos valores se pueden interpretar de forma que utilizando un umbral de 0.05, si el valor *p* es mayor a ese umbral, tal como ha sucedido en nuestro caso para ambos casos, podremos decir que las muestras siguen una distribución normal.

Ejercicio 2.

Resultados

Apartado a)

```
La diferencia de medias es:  1.5
El error estándar de la diferencia de medias es:  0.2710134237672904
El intervalo de confianza para el nivel de confianza 0.9 es:
(1.0469867629210619, 1.953013237078938)
El intervalo de confianza para el nivel de confianza 0.95 es:
(0.9575076915140499, 2.04249230848595)
El intervalo de confianza para el nivel de confianza 0.99 es:
(0.7782134841732306, 2.2217865158267696)
```

Apartado b)

```
Resultados de verificar normalidad para la epoca 1:
Estadística: 0.8997325897216797
Valor p: 0.008273127488791943
-----------------------------------------------------
Resultados de verificar normalidad para la epoca 2:
Estadística: 0.9101247787475586
Valor p: 0.014988579787313938
-----------------------------------------------------
Resultados Verificar homogeneidad de varianzas:
Estadística: 0.03005181347150258
Valor p: 0.8629763102722959
-----------------------------------------------------
Resultados de realizar el test t:
Estadística: 5.53478118961368
Valor p: 7.857638259879935e-07
```

Comentarios

Apartado a)

En primer lugar, se puede observar como para distintos valores de confianza se cumple que el valor de la diferencia entre las medias se encuentra en los intervalos de confianza obtenidos. En el anterior ejercicio habíamos comprobado como las distribuciones siguen una distribución normal, algo que nos ha permitido realizar el intervalo de confianza. Respondiendo a la pregunta que cabeza es más alta podemos decir a través de los resultados obtenidos que la cabeza más alargada es la cabeza de egipcia.

Apartado b)

Después de utilizar el test t para contrastar la hipótesis de que ambas medias son iguales, mostramos las condiciones que se deben cumplir para poder aplicar ese contraste. Estas son las siguientes:

1. Independencia de las muestras:

Las observaciones en una muestra no deben influir en las observaciones de la otra muestra. Esto suele asumirse cuando las muestras se seleccionan aleatoriamente o cuando se aplica un diseño experimental adecuado.

2. Normalidad de las muestras o tamaño de muestra suficientemente grande:

La normalidad de las muestras es una condición importante para tamaños de muestra pequeños. Sin embargo, si el tamaño de la muestra es lo suficientemente grande, el Teorema del Límite Central permite que las medias muestrales sigan una distribución normal, lo que facilita la aplicación del test t incluso si los datos originales no son normalmente distribuidos.

3. Homogeneidad de varianzas (para el test t de muestras independientes):

Las varianzas en ambas muestras deben ser aproximadamente iguales. Puedes verificar esto mediante pruebas estadísticas como la prueba de Levene o visualmente mediante gráficos de dispersión. Algunas versiones del test t son más robustas ante violaciones de esta condición, pero es importante tenerla en cuenta.

La condición 1 ya lo suponemos que se cumple. La condición de normalidad vemos como se cumple en ambas distribuciones a partir del resultado obtenido para los p-valores ya que ambos son menores de 0,05 y por tanto se podría decir que se cumple. La condición de homogeneidad en cambio no se cumple ya que el p-valor es menor de 0,05.