Ejercicio de Evaluación. Descriptiva e Inferencia.

Jonás Jiménez Gil

ÍNDICE

EJERCICIO 1

La tabla de datos de anchuras de cráneos y épocas (Predinástica temprana y predinástica tardía) ha sido importado a el script de R como un dataframe mediante código.

```
Libro1 <- read_excel("C:/Users/jjimenez/Desktop/Estadistica R/Tarea/Libro1.xlsx")
View(Libro1)
```

A partir de este dataframe inicial ("Libro1") se crearon dos dataframes diferentes, uno para la época Predinástica Temprana ("P1") y otro para Predinástica Tardía ("P2"). Estos dataframes no se convirtieron en vectores sino que se operó con ellos siendo dataframes.

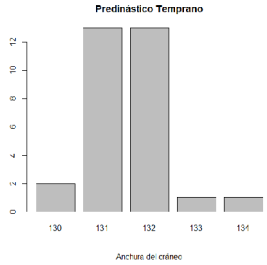
Una vez separados en dos dataframes diferentes se procedió a obtener sus medidas de centralidad, dispersión, asimetría y curtosis.

MEDIDAS DE CENTRALIZACIÓN

PREDINÁSTICO TEMPRANO

```
P1.Centralizacion <- P1%>%
                     summarise(Mean=mean(P1$'Anchura del cráneo'), Max=max(P1$'Anchura del cráneo'),
                     Min=min(P1$'Anchura del cráneo'), Median=median(P1$'Anchura del cráneo'),
                     Mode=mlv(P1$'Anchura del cráneo', method = "mfv"))
```

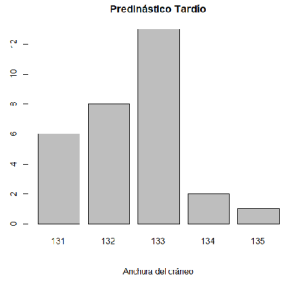
	Mean	Max	Min	Median	Mode
1	131.5333	134	130	131.5	131
2	131.5333	134	130	131.5	132



PREDINÁSTICO TARDÍO

```
P2.Centralizacion <- P2%>%
                     summarise(Mean=mean(P2$'Anchura del cráneo'), Max=max(P2$'Anchura del cráneo'),
                     Min=min(P2$'Anchura del cráneo'), Median=median(P2$'Anchura del cráneo'),
                     Mode=mlv(P2$'Anchura del cráneo', method = "mfv"))
```

	Mean	Max	Min	Median	Mode
1	132.4667	135	131	133	133



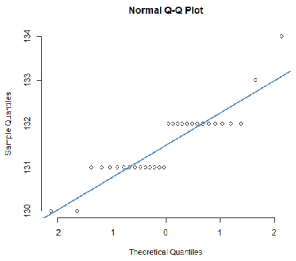
Análisis:
Observamos que la media muestra un valor menor para el Predinástico Temprano (1) que para Predinástico Tardío (2) y el mismo resultado obtenemos para la mediana, luego ha aumentado ligeramente de un período a otro.
En el caso de (1) la media y la mediana tienen un valor muy similar luego tenemos una distribución simétrica en sus valores, mientras que en el caso de (2), la mediana es ligeramente superior a la media, por lo tanto tenemos una distribución con un ligero sesgo negativo.
En cuanto a la moda en (1) obtenemos dos valores que se repiten con la misma asiduidad (131 y 132), mientras que en (2) se repite 133, observamos que estos valores coinciden con la zona central de la curva.

CUARTILES

PREDINÁSTICO TEMPRANO

```
P1.Cuartiles <- quantile(P1$'Anchura del cráneo')
```

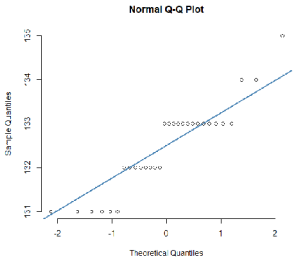
```
  0%  25%  50%  75% 100%
130.0 131.0 131.5 132.0 134.0
```



PREDINÁSTICO TARDÍO

```
P2.Cuartiles <- quantile(P2$'Anchura del cráneo')
```

```
  0%  25%  50%  75% 100%
 131  132  133  133  135
```



Análisis cuartiles: Observamos que dentro de la similitud del rango existe un comportamiento bastante parecido dentro de los cuartiles.
En el caso 1 podemos comparar que los valores de la moda coinciden con los cuartiles Q1 y Q3, agrupando la mayor cantidad de datos entre los dos
En el caso 2 se observa que siendo la moda 133, coincide con el valor del Q3, lo cual nos indica que al menos un 75% de los datos tienen un valor igual o menor a 133, siendo solamente un 25% de ellos mayores.
Con estas graficas podemos observar como ambas se separan de la normalidad en las colas y en el caso de Predinástico Tardío el pico esta ligeramente desplazada hacia arriba con respecto a la distribución normal que lo generaría. Esta observación es preliminar y necesitamos realizar algún test de hipótesis que nos ayude a comprobar su normalidad.

MEDIDAS DE DISPERSIÓN

PREDINÁSTICO TEMPRANO

```
range(P1$'Anchura del cráneo')
P1.Dispersión <-  P1%>%
                  summarise(range=max(P1$'Anchura del cráneo')-min(P1$'Anchura del cráneo'),
                  var=var(P1$'Anchura del cráneo'), sd=sd(P1$'Anchura del cráneo'))
```

	range	var	sd
1	4	0.6712644	0.8193072

PREDINÁSTICO TARDÍO

```
range(P2$'Anchura del cráneo')
P2.Dispersión <-  P2%>%
                  summarise(range=max(P2$'Anchura del cráneo')-min(P2$'Anchura del cráneo'),
                  var=var(P2$'Anchura del cráneo'), sd=sd(P2$'Anchura del cráneo'))
```

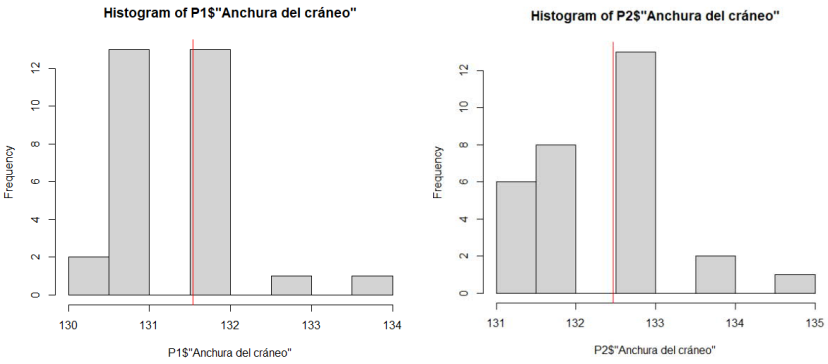
	range	var	sd
1	4	1.016092	1.008014

Análisis medidas de dispersión: Podemos observar que la desviación estándar es mucho mayor en el 2 (sd=1.008), lo que nos muestra una mayor dispersión en los datos en comparación con 1(sd=0.819).
Lo mismo nos sucede con la varianza, ya que esta medida que también mide la dispersión es la SD al cuadrado.
En el caso de los rangos, ambos tienen el mismo ancho, sin embargo el rango 1 está ligeramente más escorado hacia la parte izquierda de los valores.

ASIMETRÍA

En primer lugar calculamos la asimetría de ambos periodos para poder analizar sus resultados.

```
tapply(Libro1$`Anchura del cráneo`, Libro1$`Época histórica`, 'skewness')
> tapply(Libro1$`Anchura del cráneo`, Libro1$`Época histórica`, 'skewness')
        1         2
0.6244092 0.1854320
```



A su vez también calculamos el coeficiente de asimetría de Fisher.

```
skewness(Libro1$`Anchura del cráneo`, na.rm = TRUE, type = 3)
```

Coef.Fisher 1 = 0.4641
Coef.Fisher 2 = 0.1854

Análisis: El resultado de ambas curvas nos muestra una asimetría positiva (al estar este valor por encima de 0), lo cual significa que contiene una mayor cantidad de valores a la izquierda de lamedia, siendo ligeramente más pronunciado este caso en el 1. Observamos que con el coeficiente de asimetría de Fisher obtenemos los mismos valores.

CURTOSIS

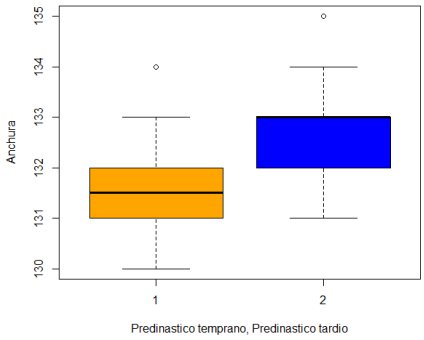
Calculamos la curtosis mediante la función tapply y 'kurtosis'

Curtosis <- tapply(Libro1$`Anchura del cráneo`, Libro1$`Época histórica`, 'kurtosis')

Curtosis 1 = 1.0221
Curtosis 2 = -0.3069

Análisis: La curtosis de 1 muestra ser leptocúrtica ya que concentra una gran cantidad de valores en su tramo central y ser mayor a 0, a su vez es más empinada que la curva de 2, la cual es platicúrtica, al tener un valor inferior a 0 y ser, por tanto, mucho más aplanada

DIAGRAMA DE CAJAS Y BIGOTES



El diagrama muestra que el (1) tiene una media muy inferior en comparación con el (2). También se observa que la media se encuentra bastante centrada entre los datos mientras que para el (2), la media está muy pegada a los valores máximos, presentando un sesgo en la curva gaussiana hacia la izquierda con una asimetría negativa. En los dos diagramas podemos observar un outlier (valor atípico) en los valores más altos.

PRUEBA DE NORMALIDAD DE KOLMOGOROV-SMIRNOV

Teniendo dos muestras continuas lo lógico sería realizar el test de Kolmogorov-Smirnov pero tenemos algunos problemas y es que aparte de tener una muestra de datos muy pequeña, tenemos valores en la distribución que están repetidos, lo cual nos genera un warning en R, por tanto el test pierde fiabilidad. Para solventar este problema se desarrolló una modificación del KS Test llamada Lilliefors Test, es la alternativa al test de Shapiro-Wilk cuando el número de observaciones es mayor de 50 (nuestro caso). Este es el que aplicaremos nosotros.

tapply(Libro1$`Anchura del cráneo`, Libro1$`Época histórica`, lillie.test)

```
$`1`

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.24246, p-value = 9.677e-05


$`2`

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  X[[i]]
D = 0.23496, p-value = 0.0001938
```

A su vez se aplicó el test de Shapiro-Wilk para comprobar y contrastar resultados.

tapply(Libro1$`Anchura del cráneo`, Libro1$`Época histórica`, shapiro.test)

```
$`1`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.83781, p-value = 0.0003481


$`2`

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.8832, p-value = 0.003341
```

Podemos observar que en ambos tests tanto en el Kolmogorov - Smirnov (más apropiado para tests con n>50) como la de Shapiro-Wilk (más apropiado para tests con n<50), el **p-value** es menor a 0.05 siendo menor al nivel de significancia, por lo que se puede concluir que ambas **NO siguen** una distribución normal.

EJERCICIO 2

Condiciones de un t-test para muestras independientes:

• **Independencia** (Verdadero): Las observaciones tienen que ser independientes unas de las otras. Para ello el muestreo debe ser aleatorio y el tamaño de la muestra inferior al 10% de la población.
- Se puede asumir que ambas muestras son independientes.

• **Normalidad** (Falso): Las poblaciones que se comparan tienen que distribuirse de forma normal. A pesar de que la condición de normalidad recae sobre las poblaciones, normalmente no se dispone de información sobre ellas por lo que las muestras (dado que son reflejo de la población) tiene que distribuirse de forma aproximadamente normal. En caso de cierta asimetría los t-test son considerablemente robustos cuando el tamaño de las muestras es mayor o igual a 30.
- Se ha demostrado en el apartado anterior que la muestras de anchuras de cráneo para ambas épocas **NO** siguen una distribución normal.

• **Igualdad de varianza** (Verdadero): la varianza de ambas poblaciones comparadas debe de ser igual. Tal como ocurre con la condición de normalidad, si no se dispone de información de las poblaciones, esta condición se ha de asumir a partir de las muestras.
- Levene Test no ha podido rechazar la hipótesis de igualdad de varianzas por lo que esta se puede asumir.

En primer caso realizamos el test de Levene para la anchura de los Cráneos para comprobar la condición de igualdad de varianza

leveneTest(Libro1$`Anchura del cráneo`, Libro1$`Época histórica`)

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.6195 0.4344
      58
```

En este caso el p-value es de 0.4344 siendo mayor a 0.05 y por lo tanto NO hay suficientes indicios para rechazar H0. Por esto mismo, no se puede rechazar la hipótesis de que las varianzas de ambas muestras son iguales.

Aunque no se cumplen todas las condiciones para emplear el T-Test, se nos indica en el enunciado que lo implementemos, y por tanto realizamos el t-test para cada uno de los intervalos de confianza que nos interesan.

#90%

```
t.test(Libro1$`Anchura del cráneo`~ Libro1$`Época histórica`, var.eq = TRUE, conf.int = TRUE, conf.level = 0.90)
```

```
        Two Sample t-test

data:  Libro1$`Anchura del cráneo` by Libro1$`Época histórica`
t = -3.9354, df = 58, p-value = 0.0002248
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
90 percent confidence interval:
 -1.3297600 -0.5369067
sample estimates:
mean in group 1 mean in group 2
       131.5333        132.4667
```

#95%

```
t.test(Libro1$`Anchura del cráneo`~ Libro1$`Época histórica`, var.eq = TRUE, conf.int = TRUE)
```

```
        Two Sample t-test

data:  Libro1$`Anchura del cráneo` by Libro1$`Época histórica`
t = -3.9354, df = 58, p-value = 0.0002248
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
95 percent confidence interval:
 -1.4080621 -0.4586046
sample estimates:
mean in group 1 mean in group 2
       131.5333        132.4667
```

#99%

```
t.test(Libro1$`Anchura del cráneo`~ Libro1$`Época histórica`, var.eq = TRUE, conf.int = TRUE, conf.level = 0.99)
```

```
        Two Sample t-test

data:  Libro1$`Anchura del cráneo` by Libro1$`Época histórica`
t = -3.9354, df = 58, p-value = 0.0002248
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
99 percent confidence interval:
 -1.5649604 -0.3017063
sample estimates:
mean in group 1 mean in group 2
       131.5333        132.4667
```

El p-value es de 0.00022248, es menor a 0.05 por lo tanto podemos rechazar la hipótesis nula y concluimos que las medias de anchura de cráneo son estadísticamente diferentes. El intervalo de la diferencia no tiene a 0 dentro de intervalo, por lo que hay una nula probabilidad de que en la población las medias de ambas épocas sean iguales.

Según los tests y pruebas realizadas durante todo el ejercicio podemos determinar que la cabeza era más ancha en el período Predinástico Tardío.