PROYECTO ESTADÍSTICA

Uxía Taboada Nieto

Máster Big Data & Business Analytics

EJERCICIO 1

- **APARTADO A**

Obtener con R 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.

*Para leer el script, tener en cuenta que los subgrupos en los que se ha dividido el data set original son.

df1 → Predinóstico temprano y df2 → Predinóstico tardío.

Medidas predinástico temprano

Medidas de posición central			
Media artimética	**Media geométrica**	**Mediana**	**Moda**
134.4	**134.3959**	**134**	**134**

Medidas de posición no centrales			
	Cuartiles		**Deciles**
0%	**132**	**0%**	**132**
25%	**134**	**10%**	**133**
50%	**134**	**20%**	**134**
75%	**135**	**30%**	**134**
100%	**137**	**40%**	**134**
		50%	**134**
		60%	**135**
		70%	**135**
		80%	**135**
		90%	**135.1**
		100%	**137**

* Los percentiles se aprecian mejor en el script.

Medidas de dispersión				
Rango	**Varianza**	**Desviación típica**	**Coef. Pearson**	**Coef. Pearson (%)**
5	**1.144828**	**1.069966**	**0.007961058**	**0.7961058**

Medidas de forma	
Asimetría	Curtosis
-0.1567441	0.4226792

Medidas predinástico tardío

Medidas de posición central			
Media artimética	Media geométrica	Mediana	Moda
132.9	132.8961	133	133

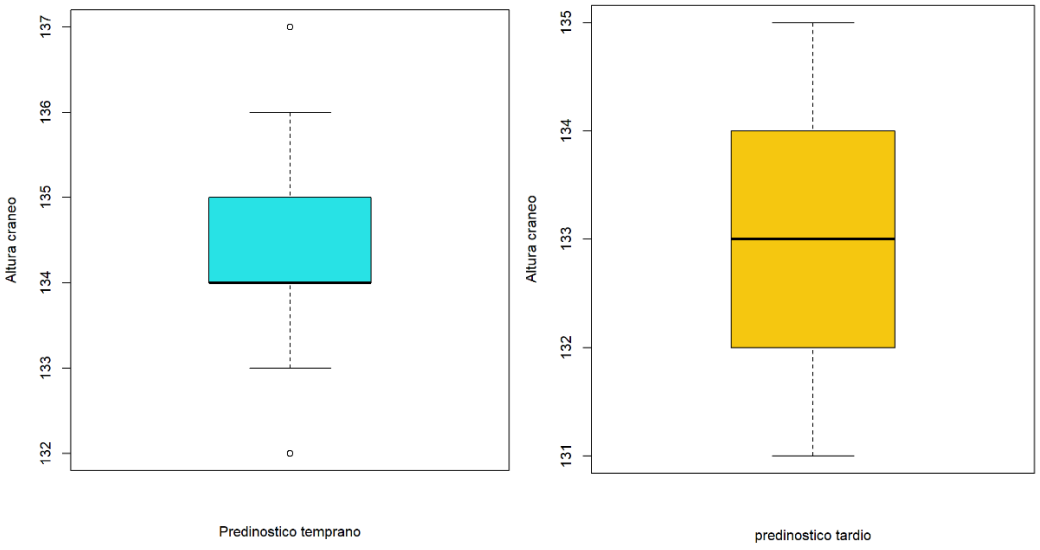
Medidas de posición no centrales			
	Cuartiles		Deciles
0%	131	0%	131
25%	132	10%	131.9
50%	133	20%	132
75%	134	30%	132
100%	135	40%	133
		50%	133
		60%	133
		70%	133
		80%	134
		90%	134
		100%	135

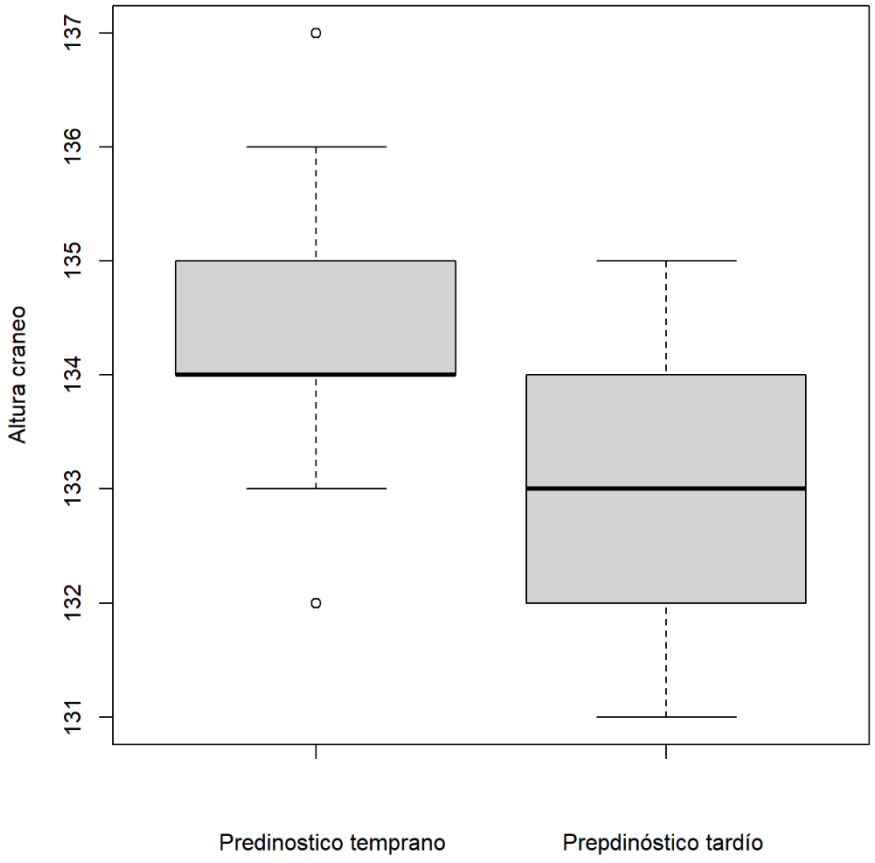
* Los percentiles se aprecian mejor en el script.

Medidas de dispersión				
Rango	Varianza	Desviación típica	Coef. Pearson	Coef. Pearson (%)
4	1.058621	1.028893	0.007741858	0.7741858

Medidas de forma	
Asimetría	Curtosis
-0.176275	-0.7736389

DIAGRAMA DE CAJA DE BIGOTES PARA AMBOS PERÍODOS





- **APARTADO B**

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

En este caso, no es demasiado fiable utilizar el test KS ya que está pensado para un volumen más grande de datos. Otro problema es que tampoco conocemos ni la media ni la varianza que son parámetros que el KS test necesita. Por último, nos lanza el siguiente warning cuando intentamos utilizarlo, este se debe a la presencia de valores repetidos en las muestras, que visualmente vemos en las regresiones 1 y 2 representadas más adelante en el documento al realizar el normal QQ-PLOT.

```
Warning message:
In ks.test.default(df1$Altura, pnorm, mean = media1, sd = desv1) :
  ties should not be present for the Kolmogorov-Smirnov test
```

Como alternativa más fiable, procedemos a utilizar el **test de Lilliefors.**

TEST DE LILLIEFORS

Este test asume media y varianza desconocidas. Además es más adecuado porque en cada subgrupo tenemos menos de 50 muestras, en particular se ha dividido el dataset original en 2 subgrupos de 30 medidas cada uno.

Hipótesis:

$$H_0: \textit{ Las muestras siguen una distribución normal}$$

$$H_1: \textit{Las muestras NO siguen una distribución normal}$$

Test para el **predinóstico temprano**:

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  df1$Altura
D = 0.22093, p-value = 0.0006631
```

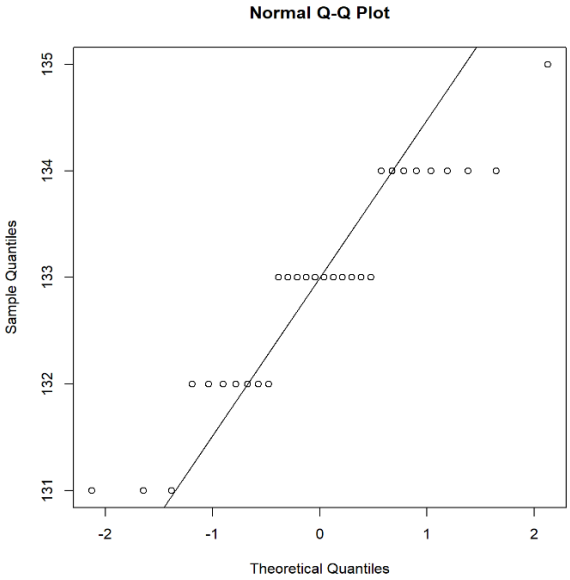
$$\boldsymbol{p - valor = 0.0006631 < \alpha = 0.05}$$

Test para el **predinóstico tardío**:

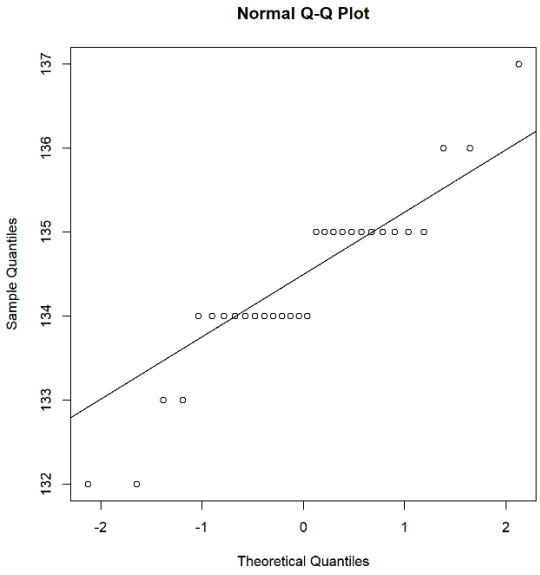
```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  df2$Altura
D = 0.20538, p-value = 0.002331
```

$$\boldsymbol{p - valor = 0.002331 < \alpha = 0.05}$$

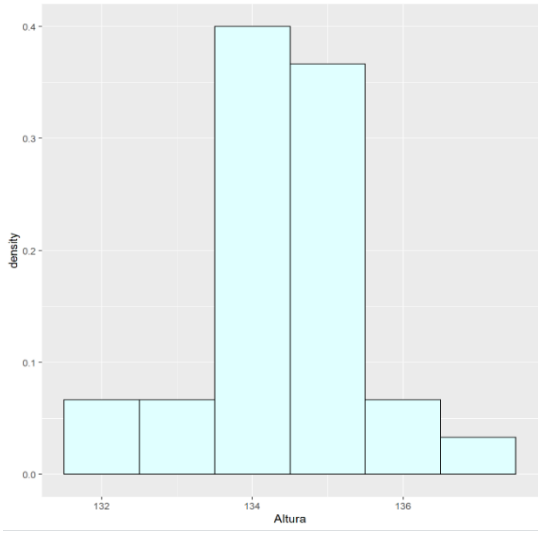
El test concluye, **para ambos períodos**, que **hay suficientes indicios para rechazar la hipótesis nula**, y por lo tanto NO podemos afirmar que las distribuciones sean normales.



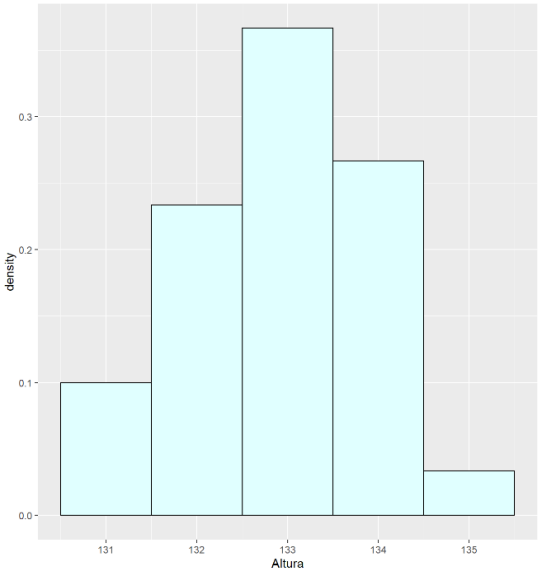
Regresión 1: Predinóstico temprano.



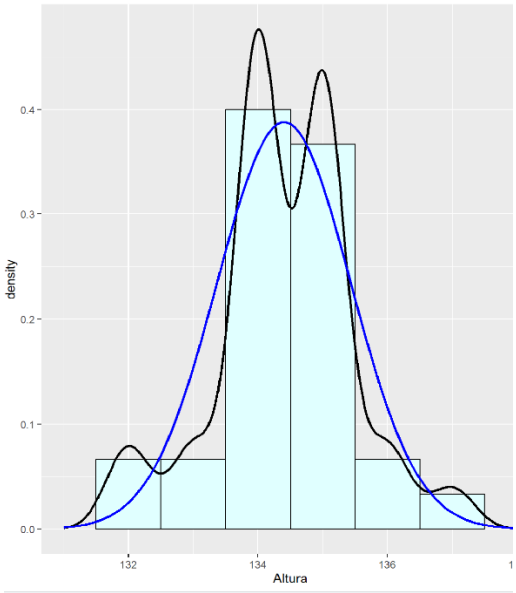
Regresión 2: Predinósotico tardío.



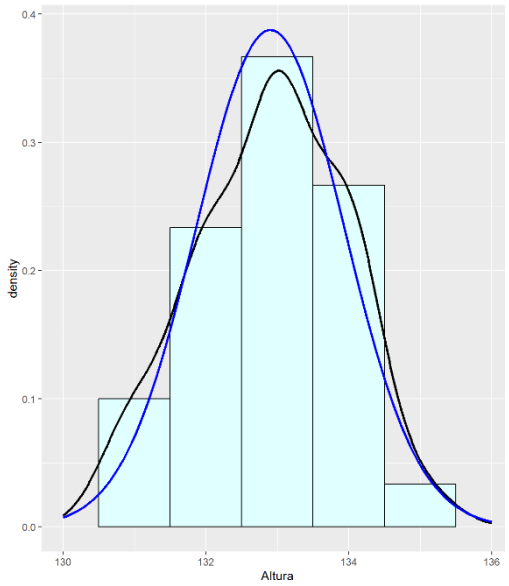
Histograma 1: Predinóstico temprano.



Histograma 2: Predinósotico tardío.



Curva 1: Predinóstico temprano.



Curva 2: Predinósotico tardío.

En estos últimos gráficos, se ha representado de color negro, la curva de densidad real de las medidas, y en azul la curva de la distribución normal. Vemos que en para el cuadro *Curva 1*, la distribución presenta unos picos en su máximo y se aleja de la normal en los extremos, su cuerpo es como el de una normal muy estrecha. En el cuadro *Curva 2*, vemos que se asemeja bastante a una distribución normal, pero bastante deformada sobre todo en la parte central. Estos gráficos junto con el test anterior, nos muestran que **no podemos considerar ninguna de las distribuciones como normales**.

EJERCICIO 2

- **APARTADO 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 altura 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?**

Hipótesis:

$$H_0: \textit{Las muestras tienen medias iguales}$$

$$H_1: \textit{Las muestras NO tiene medias iguales}$$

Para contrastar estas hipótesis, se han realizado 3 t test con los niveles de confianza y significación siguientes:

- **Nivel de confianza 0.9 → α = 0.1**

```
        Welch Two Sample t-test

data:  df1$Altura and df2$Altura
t = 5.5348, df = 57.911, p-value = 7.887e-07
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 1.046976 1.953024
sample estimates:
mean of x mean of y
    134.4     132.9
```

- **Nivel de confianza 0.95 → α = 0.05**

```
        Welch Two Sample t-test

data:  df1$Altura and df2$Altura
t = 5.5348, df = 57.911, p-value = 7.887e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.95749 2.04251
sample estimates:
mean of x mean of y
    134.4     132.9
```

- **<u>Nivel de confianza 0.99 → α = 0.01</u>**

```
        Welch Two Sample t-test

data:  df1$Altura and df2$Altura
t = 5.5348, df = 57.911, p-value = 7.887e-07
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 0.7781761 2.2218239
sample estimates:
mean of x mean of y
    134.4     132.9
```

En los tres casos, el **p-valor=$7.887 \cdot 10^{-7}$, es menor que el nivel de significación α** en cada caso, por lo tanto **hay suficientes indicios para rechazar la hipótesis nula H_0**. Por lo tanto, podemos concluir afirmando que **las medias** de las alturas de los cráneos en el predinóstico temprano y en el tardío, **NO son iguales.**
Analizando los resultados, se puede observar que **en el predinóstico temprano (df1) los cráneos eran más alargados que en el predinóstico tardío (df2)**.

- **APARTADO 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.

<u>Para poder realizar el t test, se han de cumplir **3 condiciones:**</u>

1. **Que las distribuciones sean normales.**

Esta condición no se cumple y se ha demostrado en el ejercicio anterior. De todas maneras haremos el t-test.

2. **Que las muestras sean independientes.**

Se asume como verdadera.

3. **Que la varianza de ambas muestras sea igual.**

Este último punto se puede determinar mediante el test de Levene o el de Bartlett, nos decantamos por el **test de Levene** porque este es menos sensible a la falta de normalidad de las muestras y por eso se adapta mejor a nuestro caso particular.

Hipótesis:

$$H_0: \ Las\ muestras\ presentan\ varianzas\ iguales$$

$$\mu_1 - \mu_2 = 0$$

$$H_1: Las\ muestras\ presentan\ varianzas\ distintas$$

$$\mu_1 - \mu_2 \neq 0$$

Antes de hacer el test, vamos a ver de forma explícita las varianzas de las dos épocas y su diferencia.

$$var(df1) = 1.144828$$

$$var(df2) = 1.058621$$

$$\Delta(var) = var(df1) - var(df2) = 0.086240$$

TEST DE LEVENE

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  4  2.7165 0.05257 .
      25
```

$$\boldsymbol{p - valor = 0.05257 > \ \alpha = 0.05}$$

De forma que **NO hay suficientes indicios para rechazar la hipótesis H_0** y por lo tanto podemos considerar que **las varianzas de las dos épocas son iguales** y por ende podemos realizar el t test con normalidad.

TEST T

Hipótesis:

$$H_0: \ Las\ muestras\ presentan\ medias\ iguales$$

$$H_1: Las\ muestras\ presentan\ medias\ distintas$$

```
        Paired t-test

data:  df1$Altura and df2$Altura
t = 5.7358, df = 29, p-value = 3.296e-06
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.965139 2.034861
sample estimates:
mean difference
            1.5
```

$$\boldsymbol{p - valor = 3.296 \cdot 10^{-6} < \alpha = 0.05}$$

Por lo tanto, **hay suficientes indicios para rechazar H_0** y por lo tanto, podemos concluir que las muestras presentan diferencias significativas en sus medias. Analizando los resultados, podemos afirmar que hay evidencias de que **el cráneo ha ido perdiendo altura con respecto al predinóstico temprano** y por ello las alturas se han ido haciendo más cortas, como hemos detallado en el apartado A.