Miguel Ángel Castaño Ibáñez

Entregable estadística

Ejercicio 1.

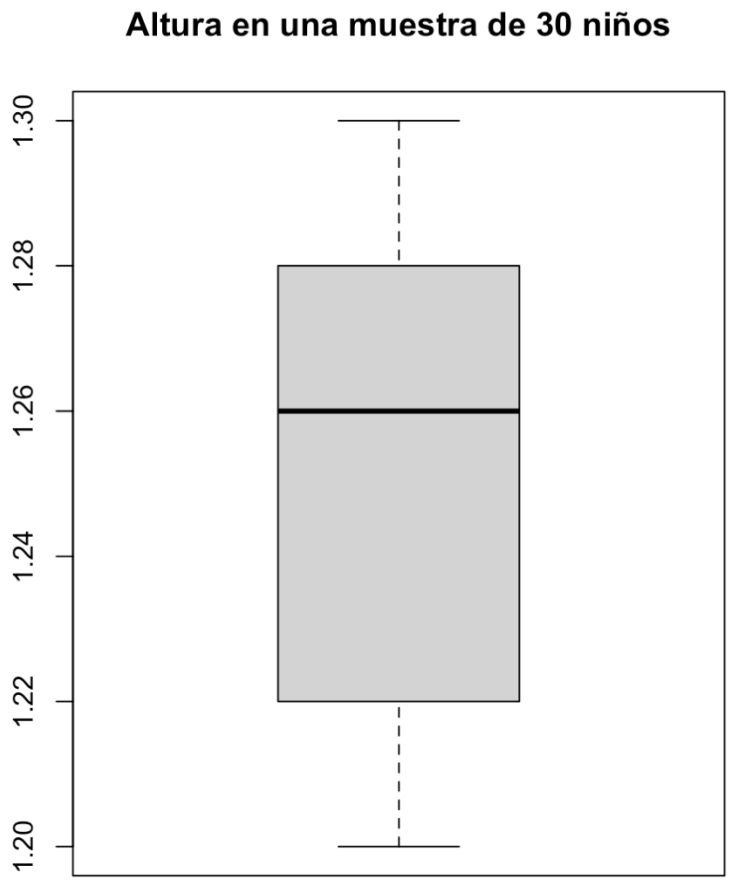
a) Dado el siguiente conjunto de datos, obtener con R las diferentes medidas de centralización y dispersión estudiadas. Así mismo obtener el diagrama de caja y bigotes.

El código empleado para la realización de este ejercicio se puede contemplar en el script adjunto a esta práctica, en el apartado 1

Todas las medidas están representadas en metros (m)

- Medidas de centralización:
  - Media (μ): 1.253
  - Mediana (Me): 1.26
  - Moda (Mo): 1.21, 1.22, 1.28
- Medidas de dispersión
  - Rango (R): (1.2 – 1.3) => 0.1
  - Varianza (S^2): 0.001050575
  - Desviación típica (s): 0.03241257
  - Coeficiente variación (CV): 0.02586109
- Medidas de posición
  - Mínimo (min): 1.200
  - Primer cuartil (1QU): 1.220
  - Segundo cuartil o mediana (Me): 1.260
  - Tercer cuartil (3QU): 1.280
  - Máximo (max): 1.300

Estos valores los podemos ver representados a continuación en el diagrama de cajas y bigotes (boxplot)



b) Dado el siguiente conjunto de datos, obtener la tabla de correspondencias, con R, agrupando cada variable en cuatro clases o intervalos. Estos deberán ser elegidos por el alumno.

En esta tabla se muestra la tabla de correspondencia para todos los valores:

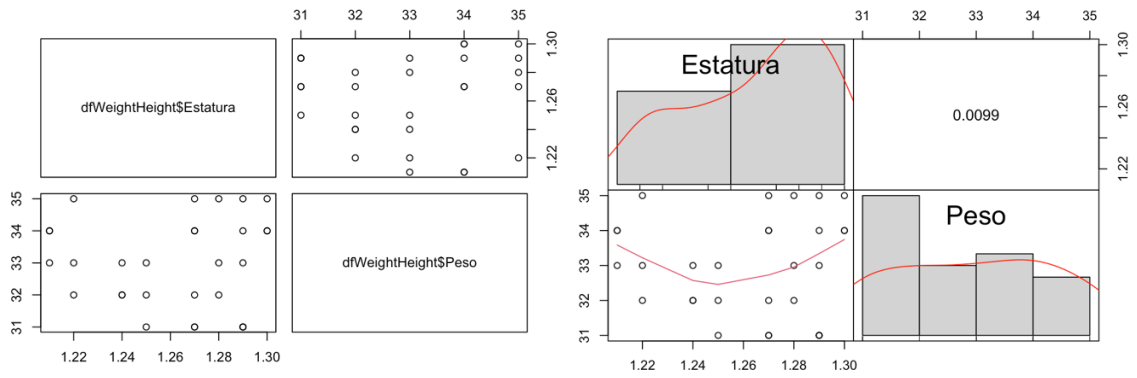
	1.21	1.22	1.24	1.25	1.27	1.28	1.29	1.3
31	0	0	0	1	2	0	3	0
32	0	1	2	1	1	1	0	0
33	1	1	1	1	0	1	1	0
34	2	0	0	0	2	0	1	2
35	0	1	0	0	1	1	1	1

Mientras que en estas podemos observar los valores ya agrupados en 4 clases:

	1.21-1.225	1.226-1.25	1.251-1.275	1.276-1.3
31	0	1	2	3
32	1	3	1	1
33	2	2	0	2
34	2	0	2	3
35	1	0	1	3

Tras obtener la tabla de correspondencias, he decidido pintar en un grafico estos resultados para comprobar si esta correlación estatura-peso podría presentar un coeficiente de correlación lineal o exponencial. Finalmente podemos observar que esta relación es irregular, no solo en la s gráficas mostradas, si no debido a que el coeficiente de correlación es muy próximo a 0 (r = 0.0099). Esto demuestra que efectivamente, aunque la relación es irregular, es positiva, es decir generalmente a mas estatura, más peso.

Por otro lado, creo que aun que vemos en la segunda gráfica una relación con una pequeña forma parabólica, el tamaño de la muestra debería ser mas grande, para obtener un resultado mas fiable, e incluso podríamos llegar a apreciar una relación lineal.



Ejercicio 2.

Considerando, de nuevo, los datos de la primera pregunta del ejercicio anterior, se pide obtener un intervalo de confianza para la diferencia de medias teóricas entre las observaciones de los primeros 15 casos y de los segundos 15 casos.

Así mismo, se pide contrastar la hipótesis nula de que ambas submuestras tienen la misma media, es decir, proceden de la misma población. Detallar las hipótesis necesarias para hacer tal contraste, aunque no es preciso comprobarlas. El análisis debe realizarse con R.

El código empleado para la realización de este ejercicio se puede contemplar en el script adjunto a esta práctica, en el apartado 2

En este ejercicio hemos aplicado la función *t.test()* para ambos dataframes, que nos ha permitido de manera inmediata contestar a todas las cuestiones propuestas en el enunciado.

Como resultado por la consola obtendremos:

```
        Welch Two Sample t-test

data:  dfHeightFirst$Estatura and dfHeightSecond$Estatura
t = 1.132, df = 27.958, p-value = 0.2672
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01079522  0.03746189
sample estimates:
mean of x mean of y
 1.260000  1.246667
```

Podemos observar un intervalo de confianza entre (-0.011, 0.037), un error a priori despreciable.

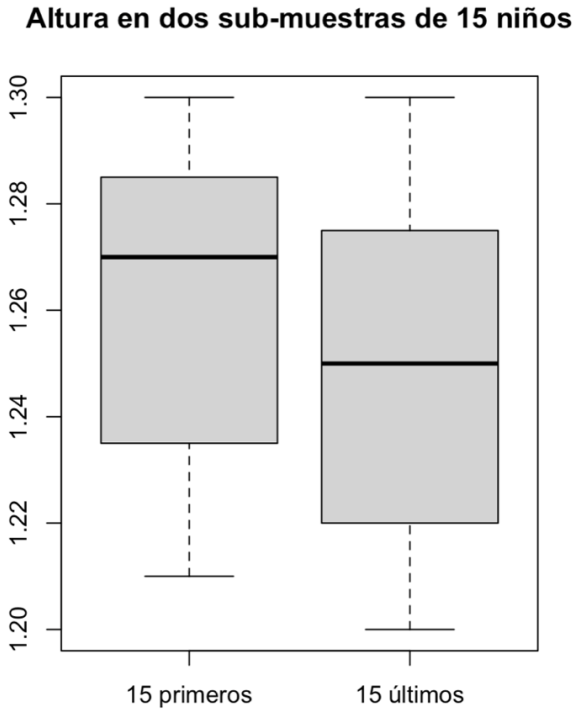
Partiendo de la hipótesis nula (*H0*), de que ambas medias son iguales, es decir, proceden de la misma población. Como hipótesis alternativa (*H1*) tenemos que estas muestras no procederán de la misma población, por lo que sus medias no son iguales.

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Tras aplicar la función *t.test()* obtenemos que el cero esta contemplado dentro del intervalo de confianza, además p-value (0.2672) > α (0.05), por lo tanto, llegamos a la conclusión de que *H0* es cierta. Donde podemos afirmar con un 95% de probabilidad de que la operación (μ_1 - μ_2) tendrá un resultado comprendido entre los valores (-0.011, 0.037), por lo tanto, ambas muestras sí proceden de la misma población y la media poblacional será la misma.

Para apreciarlo de una forma más visual y sencilla he querido representar estas muestras en un diagrama de cajas y bigotes donde ambos diagramas son muy parecidos, donde los cuartiles y el limite inferior y superior, quedan bastante igualados.



Conclusiones

Esta práctica me ha resultado de gran ayuda para recordar todos los conocimientos aprendidos en mis estudios de grado, pero me hubiera gustado manejar muestras mas grandes, para aprovechar toda la potencia de R o incluso algunos casos mas reales con más variables.

Para finalizar, considero que la estadística es una parte fundamental en el análisis de datos, y todo lo aprendido en este modulo será de gran ayuda para módulos posteriores.