Class Activity: Visualization and Data Understanding

Máster en Ingeniería y Tecnología de Sistemas Software (MITSS)

Student Answer

Full name: Sergi Sanz Carreres

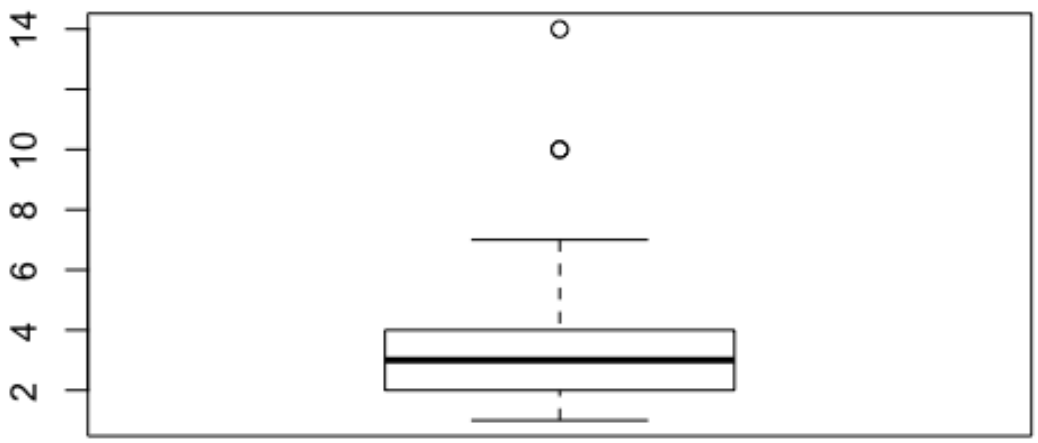
Answer the following questions:

1. How many hours a day on average does each child watch television? What is the range of television hours watched?

La cantidad de horas medias que cada niño ve la televisión es de 4,03333, para calcular este valor se ha utilizado la función: **mean(datos$tvhours),** mientras que, en lo concerniente al rango de horas de televisión observadas, utilizando la función **range(datos$tvhours)** , obtenemos un resultado de 1-14 que corresponde al rango de las horas en las que se visualiza la televisión por parte de los niños.

2. In computing the mean hours watched, were there any apparent outliers? What effect might this have on the mean hours watched?
   NOTE: Visualize graphically (using a boxplot) the hours watched in order to easily answer the question (read the , R manual for boxplots if you need some help).

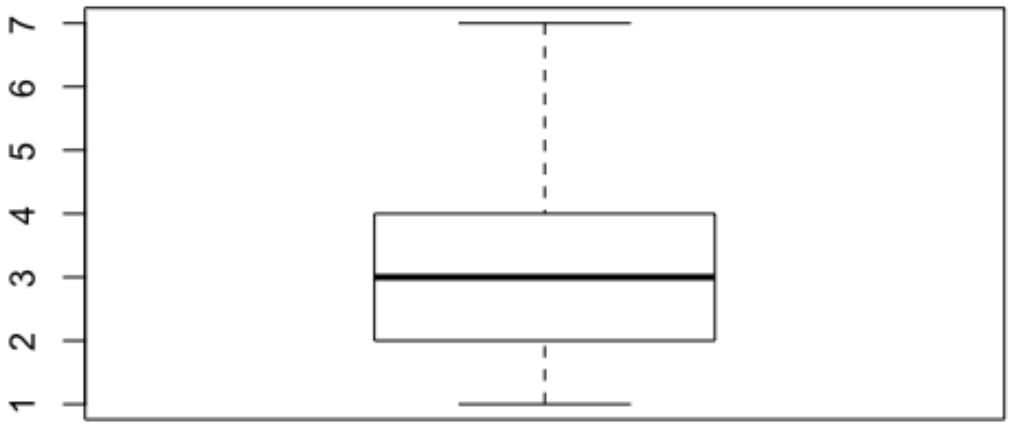
Al calcular el boxplot (utilizando el comando **v = boxplot(datos$tvhours)**) se obtiene el siguiente resultado:



Donde se puede apreciar que disponemos de dos datos atípicos con valores 10 y 14, esto afecta significativamente para el calculo de la media debido a que distorsiona los valores obtenidos en el calculo final.

3. Recompute the mean without the outlier(s) and remove them from the original dataframe.
   NOTE: You can first grab the outliers (this can be done using the boxplot). Take a look to the arguments of the R function boxplot().

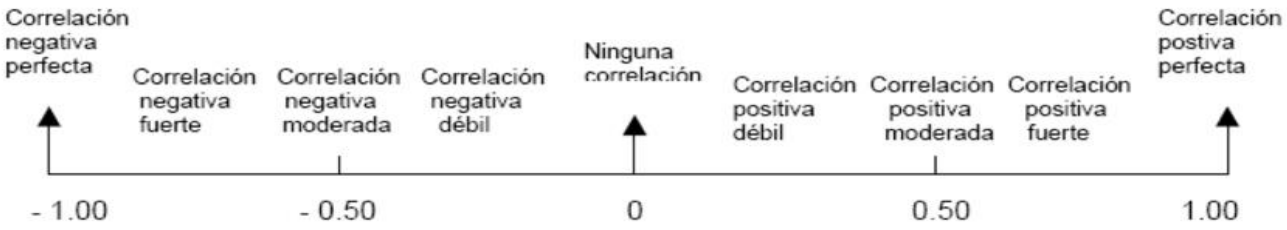
Una manera de eliminar los outliers del dataframe original, es utilizar esta extensión de boxplot, de manera que usando **boxplot(tvhours, outline = FALSE)** Automáticamente se eliminarán todos los outliers, devolviendo un boxplot como el siguiente:



Donde, podemos apreciar que la media desciende de 4,033 del calculo original a **3**. Pero el problema que presenta este comando es que no almacena los datos sin los outliers, por tanto para la correcta realización del ejercicio utilizaremos
tv = subset(datos, !(tvhours %in% v$out)) se obtendrá el mismo resultado que en la imagen anterior.

4. What is the overall correlation between the numbers of hours watched and the obedience? What is the correlation between TV hours watched and attitude?

La correlación es la relación entre dos variables, para calcular la correlación entre tvhours y obedience, se ha utilizado la siguiente formula: **cor(tv$tvhours, tv$obedience)** dando como resultado **-0.1788204**, si observamos detenidamente la siguiente imagen se aprecia de que los resultados obtenidos se acercan a una correlación negativa débil.



Mientras que si aplicamos la formula anterior adaptada para los requisitos correspondientes al calculo de la correlación entre tvhours y attitude, con una formula como la siguiente: **cor(tv$tvhours, tv$attitude),** se obtiene como resultado**: -0.08427498**, lo que indica en la imagen anterior que se acerca a una correlación negativa débil.

5. Describe the relationships indicated in question 4. Are either of these correlations statistically significant?
NOTE: Use the extension of the cor() function to provide the significance testing required.

Para determinar si la correlación entre las variables es significativa, es necesario comparar el valor de p con su nivel de significancia. Por lo general, un nivel de significancia (denotado como α o alfa) de 0.05 es adecuado, ya que un α de 0.05 indica que el riesgo de concluir que existe una correlación, cuando en realidad no es así, es 5%. El valor p indica si el coeficiente de correlación es significativamente diferente de 0. (Un coeficiente de 0 indica que no existe una relación lineal).

Para saber si una correlación es estadísticamente significativa, se debe cumplir $p \leq \alpha$. Mientras que si $p > \alpha$ significa que la correlación no es estadísticamente significativa

Por lo tanto, para la realización de este ejercicio, primero se ha utilizado la función cor.test(…) , siendo en el primer caso **cor.test(tv$tvhours, tv$obedience)** , dando como resultado **t = -0.90875, df = 25, p-value = 0.3722** , por tanto como **p> α** , lo cual significa que no será estadísticamente significativa. Mientras que para el otro ejemplo, al aplicar **cor.test(tv$tvhours, tv$attitude)** dando como result **t = -0.42288, df = 25, p-value = 0.676,** de manera que al igual que en el ejemplo anterior, **p> α** , lo cual significa que no será estadísticamente significativa.

6. Do a simple frequency count on attitude. What fundamental problem does this data present for the hypothesis? What sampling changes could be made to better test the hypothesis that "children who watch more TV are more aggressive"?

En este caso, se ha utilizado **table(tv$attitude)** obteniendo como resultados:
1 2
15 12

De manera que como se aprecía solo tenemos actitudes de 1 y 2 cuando el rango es hasta 5 de modo que no es suficientemente representativo.