Taller 1. Análisis Estadístico de Datos.

2023-02-03

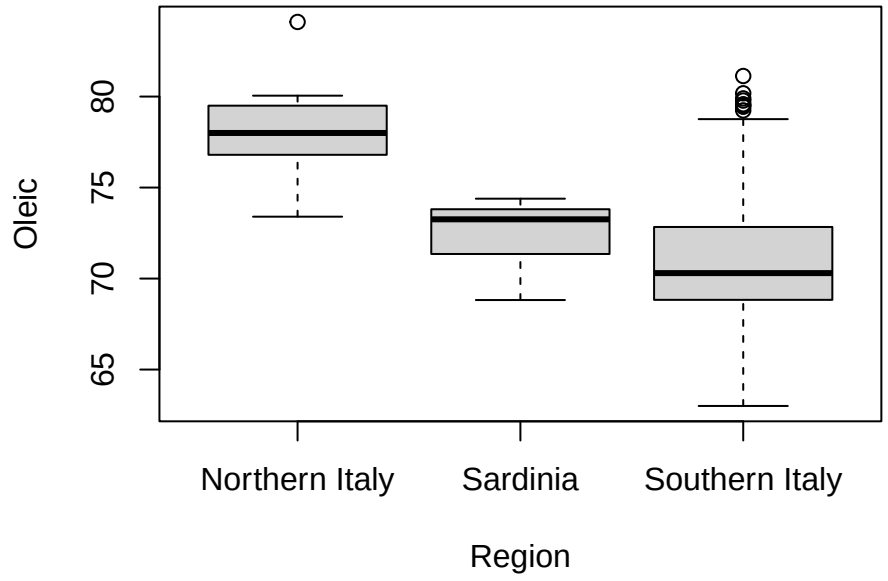
Integrantes

- Diryon Yonith Mora Romero
- Fabio Andrés Rizo Montoya
- Laura Valentina González Rodríguez

Primer punto

«Graficar los boxplots de la variable oleic vs region del dataset olive.»

```
boxplot(olive$oleic ~ olive$region, xlab = "Region", ylab = "Oleic")
```



Segundo punto

Definición variables

```r
a = mtcars;
a_num = subset(a, select = c(-cyl, -vs, -am, -gear, -carb));
a_cat = subset(a, select = c(cyl, vs, am, gear, carb));
```

Primer parte

«Determine la media, la mediana, la moda y la desviación estándar de cada una de las variables. Se puede calcular a todas la variables? a cuales no? Justifique su respuesta.»

Variables numéricas En estas variables se pueden hallar todos los datos solicitados dado que son variables discretas.

```r
sapply(a_num, function(x) {
  data.frame(
    mean = round(mean(x), digits = 3),
    median = median(x),
    mode = names(which.max(table(x))),
    sd = round(sd(x), digits = 3)
  )
}) %>%  kable() %>% kable_styling(latex_options = "hold_position")
```

	mpg	disp	hp	drat	wt	qsec
mean	20.091	230.722	146.688	3.597	3.217	17.849
median	19.2	196.3	123	3.695	3.325	17.71
mode	10.4	275.8	110	3.07	3.44	17.02
sd	6.027	123.939	68.563	0.535	0.978	1.787

Variables categóricas Estas variables se están agrupando en valores ya predefinidos, por lo cual no tiene sentido aplicar medidas de tendencia central y desviación estándar. Sin embargo, si se puede hallar el elemento más repetido; **la moda**.

```r
sapply(a_cat, function(x) {
  names(which.max(table(x)))
}) %>% t() %>% kable() %>% kable_styling(latex_options = "hold_position")
```

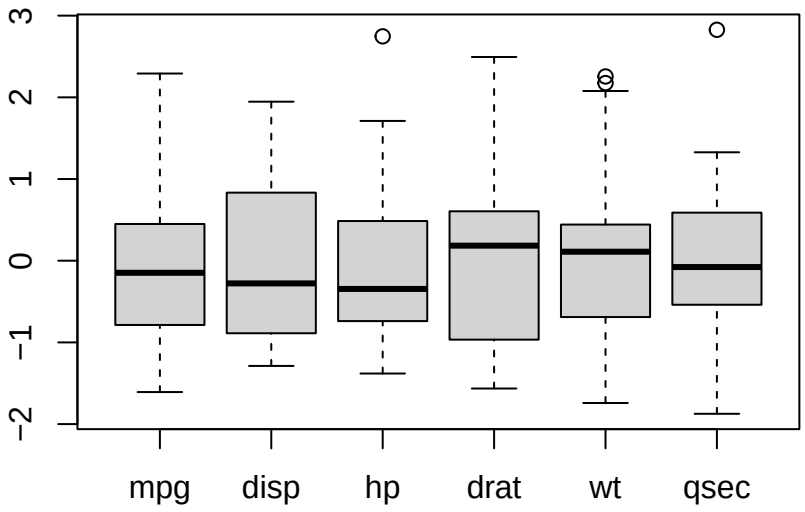
cyl	vs	am	gear	carb
8	0	0	3	2

Segunda parte

«Determinar qué variable presenta valores atípicos, ¿cómo los ha encontrado?»

Con el gráfico de cajas y bigotes normalizado podemos comprobar los valores atípicos; aquellas observaciones no presentes en ningún cuadril. Así, sabemos que `hp`, `wt` y `qsec` poseen datos atípicos.

```r
a_num_norm = as.data.frame(scale(a_num));
boxplot(a_num_norm);
```

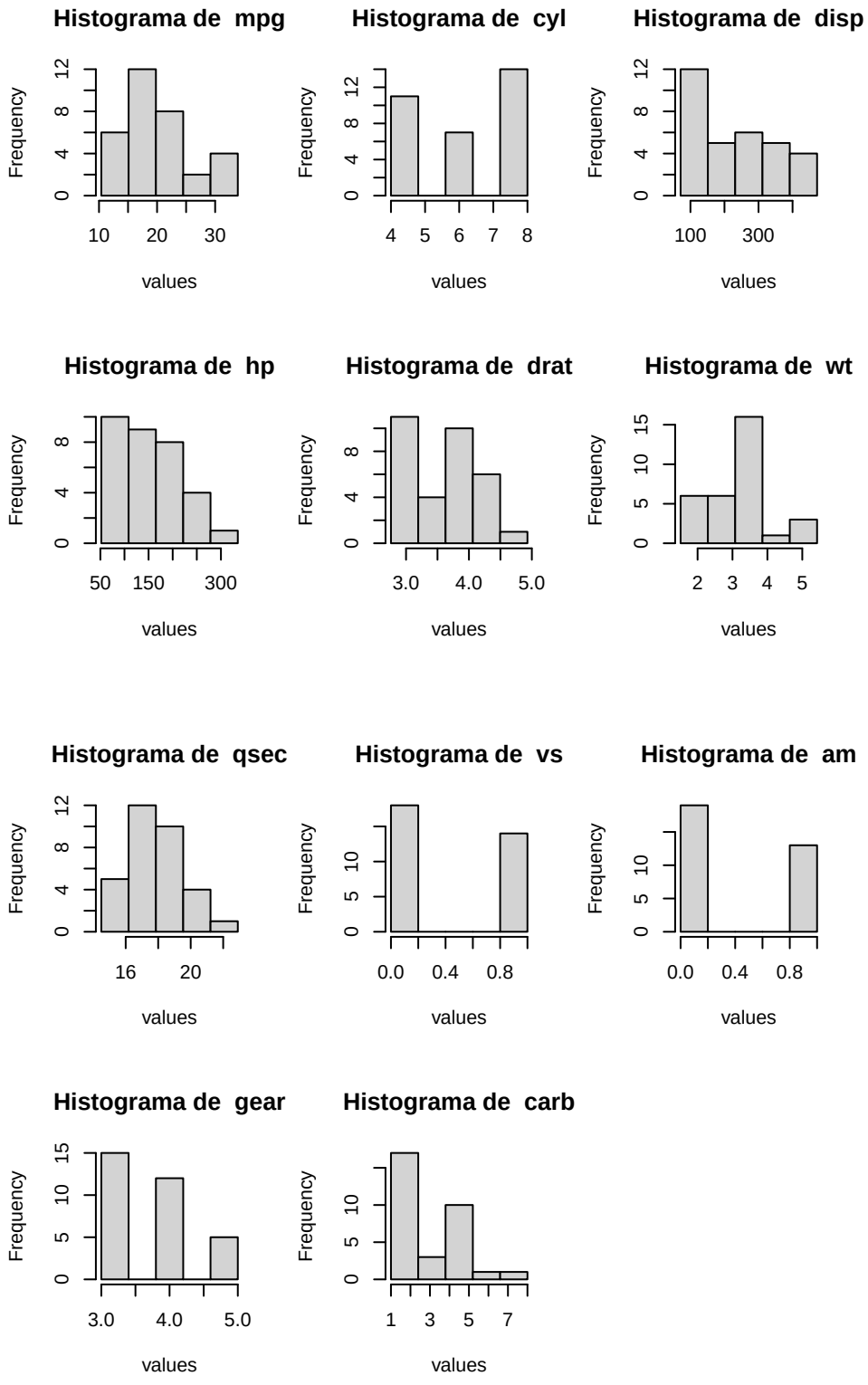


Tercera parte

«Hacer el histograma para cada una de las variable usando 5 intervalos. De nuevo, está gráfica es útil para todas las variables? justifique su respuesta.»

Si bien es posible aplicarlo en variables categóricas, simplemente es redundante a nivel interpretatipo. El eje x del histograma comprende intervalos en cuales se pueden agrupar información y así poder, justamente, categorizar la información discreta / continua.

```
par(mfrow=c(2, 3));
sapply(colnames(a), function(columna) {
  values = unlist(a[columna]);
  breaks = seq(min(values), max(values), length.out = 6);
  hist(
    x = values,
    breaks = breaks,
    main = paste("Histograma de ", columna)
  );
})
```

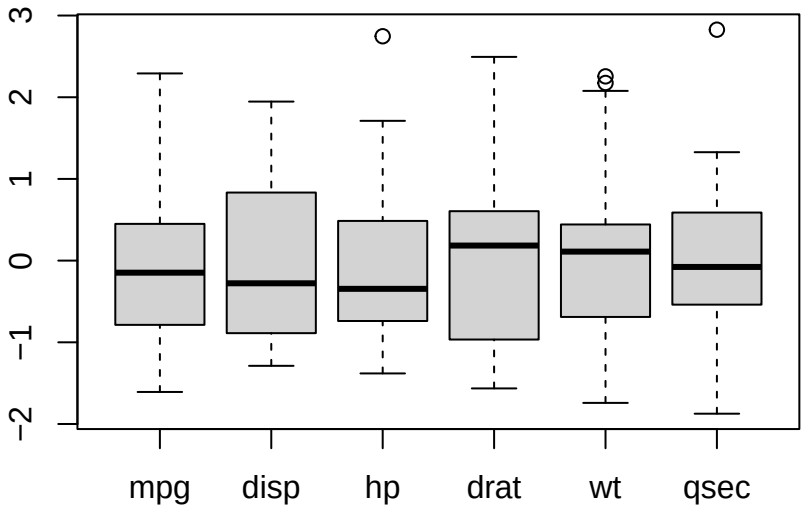

Histograma de mpg
Frequency
values
Histograma de cyl
Frequency
values
Histograma de disp
Frequency
values
Histograma de hp
Frequency
values
Histograma de drat
Frequency
values
Histograma de wt
Frequency
values
Histograma de qsec
Frequency
values
Histograma de vs
Frequency
values
Histograma de am
Frequency
values
Histograma de gear
Frequency
values
Histograma de carb
Frequency
values

Cuarta parte

«Realice una gráfica que incluya el diagrama de cajas de todas las variables de tal manera de que se puedan comparar.»

Básicamente es lo mismo que se realizó en la segunda parte.

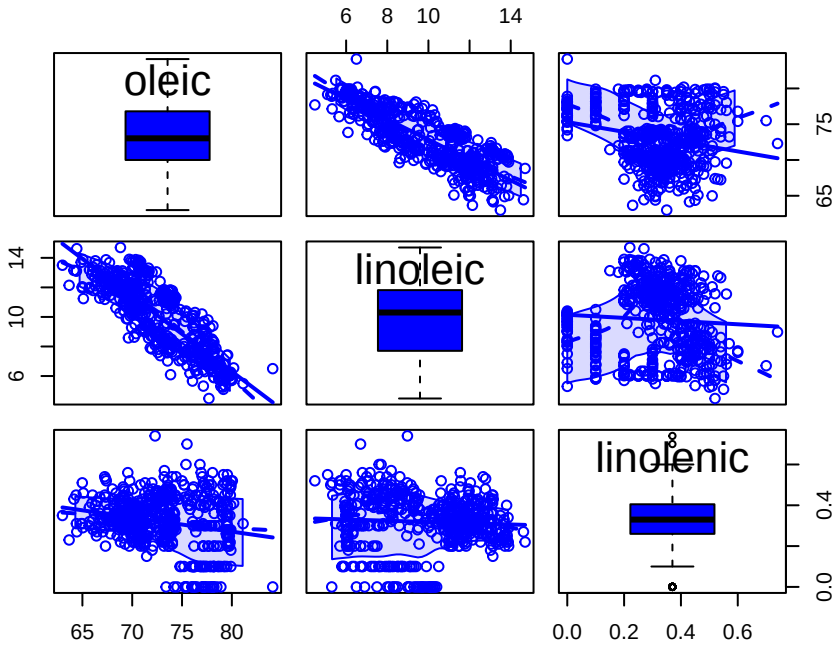
```
boxplot(a_num_norm);
```



Tercer punto

«Graficar una matrix de dispersion de tres variables del dataset olive, con la diagonal mostrando boxplots de las variables.»

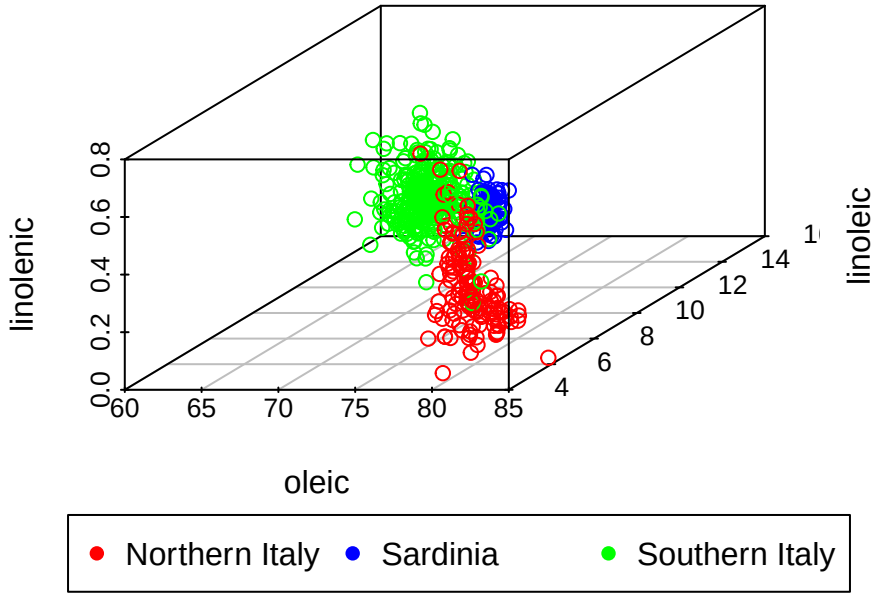
```
olive_subset = subset(olive, select = c(oleic, linoleic, linolenic))
scatterplotMatrix(~ +., data = olive_subset, diagonal = list(method = "boxplot"))
```



Cuarto punto

«Graficar un diagrama de dispersión en 3D, de tres variables numéricas del dataset olive graficando en colores diferentes las regiones.»

```
colors = c("red", "blue", "green");
region_levels = levels(olive$region);
names(colors) = region_levels;
scatterplot3d(
  olive_subset,
  color = colors[factor(olive$region)]
)
legend("bottom", legend = region_levels,
       col = colors, pch = 16,
       inset = -0.45, xpd = TRUE, horiz = TRUE)
```



Quinto punto

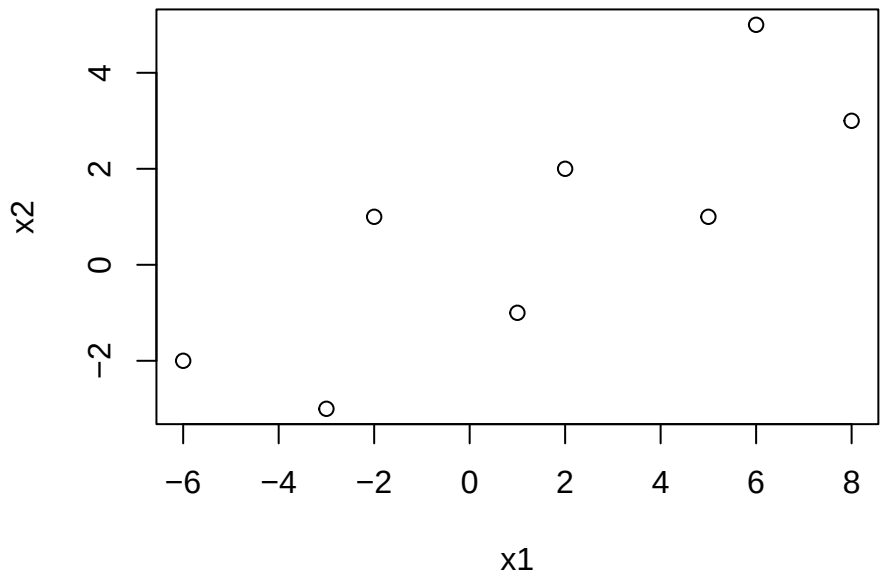
«Dados los siguientes pares de medidas sobre y dos variables x1 y x2:

x1	-6	-3	-2	1	2	5	6	8
x2	-2	-3	1	-1	2	1	5	3

Grafique los datos como un diagrama de dispersión y calcule s_{11}, s_{22} y s_{12}*»*

```
x = data.frame(x1 = c(-6, -3, -2, 1, 2, 5, 6, 8),
               x2 = c(-2, -3, 1, -1, 2, 1, 5, 3))

plot(x)
```



Finalmente, con el siguiente código:

```
varianza_muestral = function(df, i, k) {
  cols = colnames(df)
  col1 = unlist(df[cols[i]])
  col2 = unlist(df[cols[k]])
  n = length(col1)
  x1 = mean(col1)
  x2 = mean(col2)
  return ((1 / n) * sum((col1 - x1) * (col2 - x2)))
}
s11 = varianza_muestral(x, 1, 1)
s22 = varianza_muestral(x, 2, 2)
s12 = varianza_muestral(x, 1, 2)
```

Obtuvimos que $s_{11} = 20.484375$, $s_{22} = 6.1875$ y $s_{12} = 9.09375$.