Trabajo Práctico 2 - Estadística Descriptiva

María Eugenia Fontecha

EL archivo dieta.csv contiene los datos correspondientes a la ingesta de alcohol, grasas y calorías de dos tipos de dietas distintas.

Variables:

- **Grasas:** cantidad de grasas en gramos consumidas por semana
- **Alcohol:** cantidad de alcohol en gramos consumidos por semana
- **Calorias:** cantidad de calorías consumidas en Kcal por semana.
- **Dieta:** tipo de dieta dividida por meta de consumo semanal de calorías. (0: 2000 Kal/día y 1: 1500Kcal/dia)

1. **Configurar el Working-Directory.**

```
setwd('C:/Users/meuge/Documents/R/TP2')
```

2. **Cree el objeto diabetes con la base de datos del archivo *dieta1.csv* ¿Qué tipo de objeto es? Analice la estructura y las características de las variables.**

Creo el objeto diabetes importando el archivo dieta1.csv:

```
diabetes <- read.csv2('dieta1.csv')
summary(diabetes)
```

```
##     Grasas             Alcohol             calorias        Dieta
##  Length:108         Length:108         Min.   : 910   Min.   :0.0000
##  Class :character   Class :character   1st Qu.:1440   1st Qu.:1.0000
##  Mode  :character   Mode  :character   Median :1645   Median :1.0000
##                                        Mean   :1694   Mean   :0.8333
##                                        3rd Qu.:1952   3rd Qu.:1.0000
##                                        Max.   :2760   Max.   :1.0000
```

La estructura es un **dataframe**. Tiene **4 variables** y **108 registros**. Las variables son:

- Grasa: character
- Alcohol: character
- calorias: numeric entre 800-2800
- Dieta: numeric entre 0 y 1.

3. **Debido a que el n resulta escaso se decide agregar a la base de datos actual, la base de datos contenida en el archivo *dieta2.csv*. Analice la estructura de nuestra nueva base de datos. ¿Coinciden las clases con los tipos de variables establecidos? Si no es así, cámbielos según corresponda.**

Importo una nueva base de datos de dieta2.csv:

```
diabetes2 <- read.csv2('dieta2.csv')
summary(diabetes2)
```

```
##     Grasas             Alcohol             calorias        Dieta
##  Length:65          Length:65          Min.   :1263   Min.   :0
##  Class :character   Class :character   1st Qu.:1861   1st Qu.:0
##  Mode  :character   Mode  :character   Median :2179   Median :0
##                                        Mean   :2171   Mean   :0
##                                        3rd Qu.:2479   3rd Qu.:0
##                                        Max.   :3125   Max.   :0
```

También es un **dataframe** con las mismas **4 variables** y tiene **65 registros**.

Para unir las dos bases de datos uso la librería dplyr.

```
bind_rows(diabetes, diabetes2, id=NULL) -> diabetes
```

Las variables Grasas y Alcohol están cargadas como character, por lo que las transformo en numeric:

```
diabetes$Grasas <- as.numeric(diabetes$Grasas)
diabetes$Alcohol <- as.numeric(diabetes$Alcohol)
```

Por otro lado, la variable Dieta es categórica dicotómica porque solo puede tomar los valores 1 o 0, por lo que tiene que ser de tipo factor:

```
diabetes$Dieta <- as.factor(diabetes$Dieta)
```

4. **Identifique si hay datos faltantes**

```
sum(is.na(diabetes))
```

```
## [1] 0
```

No hay datos faltantes.

5. **Guarde la nueva tabla en un archivo CSV.**

```
write.csv(diabetes, 'diabetes.csv')
```

6. **Cree una función que devuelva las pruebas de normalidad que se deberían realizar para comprobar la distribución de una variable continua, tanto con Test estadísticos (Shapiro-Wilk y Kolgomorov-Smirnov) como con métodos gráficos (Histograma, Q-Q plot). Puede utilizar las funciones propias de R como *lillie.test* o *shapiro.test* o las corresondientes a los gráficos.**

```r
normal_prueba <- function(x, nombre='')
{
  #pruebas de normalidad con Test estadísticos
  print(nombre)
  print(lillie.test(x))
  print(shapiro.test(x))

  #pruebas de normalidad con métodos gráficos
  qqnorm(x, main = paste("Q-Q Plot: ",nombre))
  qqline(x)

  hist(x, col = 'blue', border = 'black', main= paste("Hist: ", nombre), freq =FALSE)
  curve( dnorm(x,mean(x),sd(x)),  min(x), max(x), add= T, col = 'red')
}
```

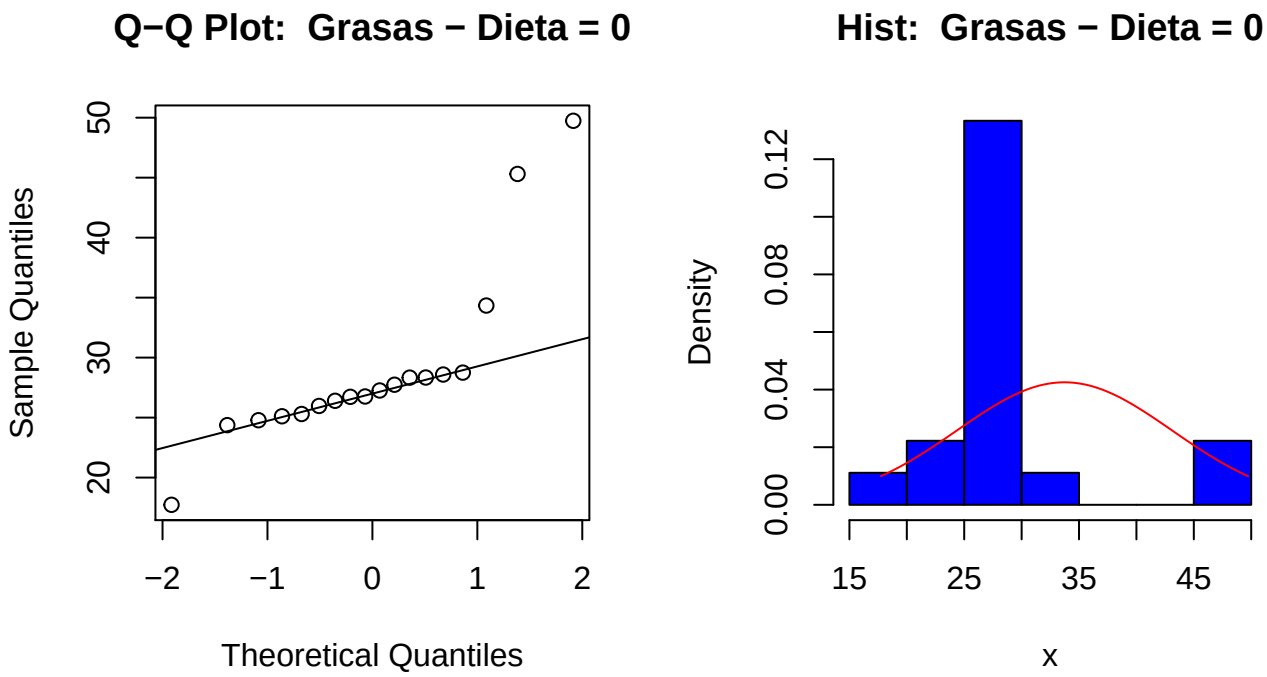
Esta función toma como parámetros la variable a la cual se le quiere realizar el test de normalidad y el nombre de dicha variable. Realiza el test lillie, el shapiro y los métodos gráficos, Q-Q plot e histograma.

7. **Para cada variable indique la distribución. En caso de ser cuantitativa realice las pruebas de normalidad (cuando se comparan dos grupos de tratamiento lo correcto es realizar las pruebas de normalidad por grupo). Puede utilizar la función previamente creada.**

Analizo normalidad para de la variable Grasas para el grupo Dieta = 0 y para el grupo Dieta = 1:

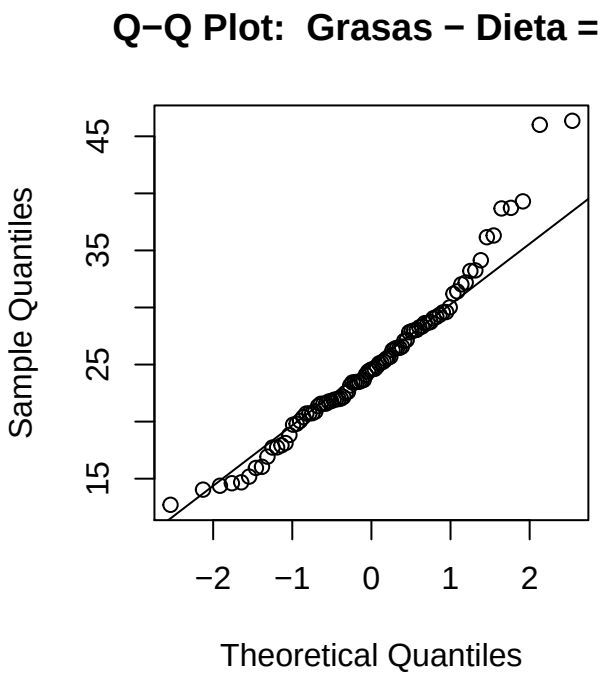
```r
normal_prueba(diabetes$Grasas[diabetes$Dieta == 0], 'Grasas - Dieta = 0')
```

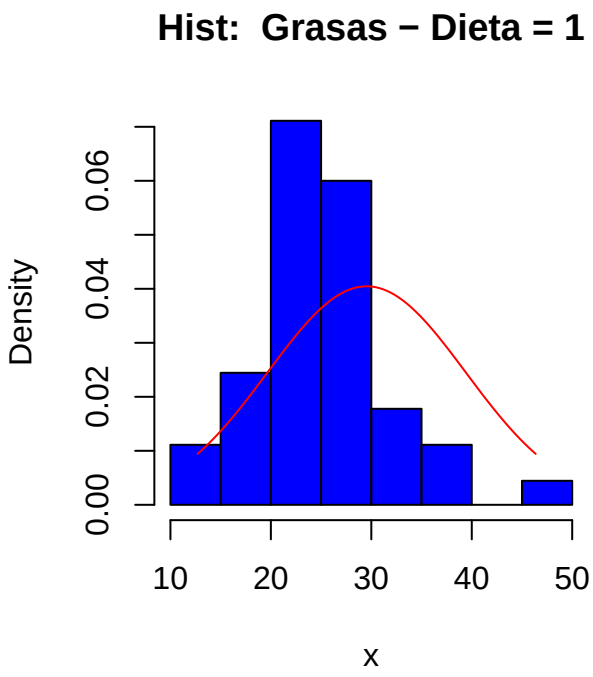
```
## [1] "Grasas - Dieta = 0"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.34533, p-value = 3.876e-06
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.7468, p-value = 0.0002952
```




```
normal_prueba(diabetes$Grasas[diabetes$Dieta == 1],'Grasas - Dieta = 1')
```

```
## [1] "Grasas - Dieta = 1"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.08214, p-value = 0.1413
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95786, p-value = 0.005287
```





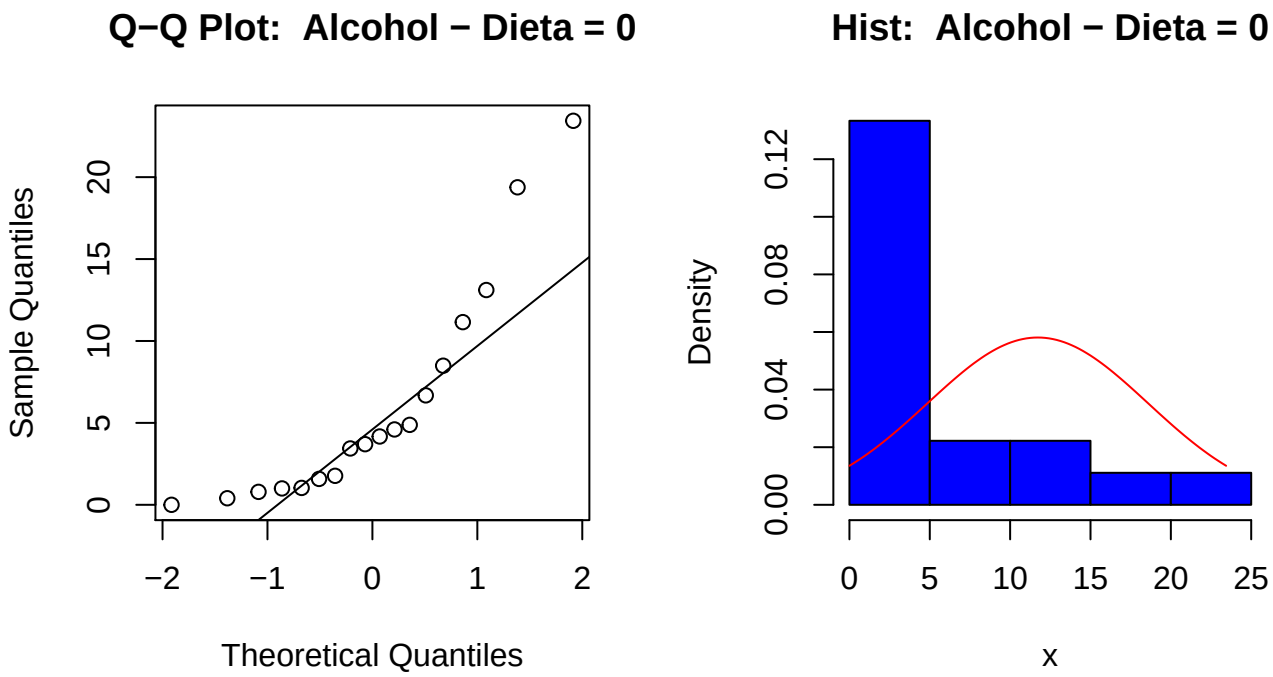
A partir de los tests de Lilliefors y de Shapiro-Wilk se puede concluir que la variable Grasas para el grupo Dieta = 0 no tiene distribución normal ya que con los dos tests se obtuvo un P-Valor menor al nivel de significancia (0.05). Esto implica que la hipótesis nula, que dice que la variable tiene distribución normal, sea rechazada. Por orto lado, en el caso de Dieta = 1, se obtuvo un P-Valor mayor al nivel de significancia en el test de Lilliefors y un P-Valor menor al nivel de significancia con el test de Shapiro-Wilk, por lo que vemos los métodos gráficos.

En el caso del Q-Q Plot puede verse que los puntos no siguen a la recta, especialmente en los extremos inferior y superior en el caso del grupo Dieta = 0, y tampoco en el centro en el caso de Dieta = 1. Además, los histogramas tampoco presentan una distribución normal en ninguno de los casos.Por lo que concluyo que la variable Grasas para el grupo Dieta = 1 tampoco tiene distribución normal.

Analizo normalidad para de la variable Alcohol para el grupo Dieta = 0 y para el grupo Dieta = 1:

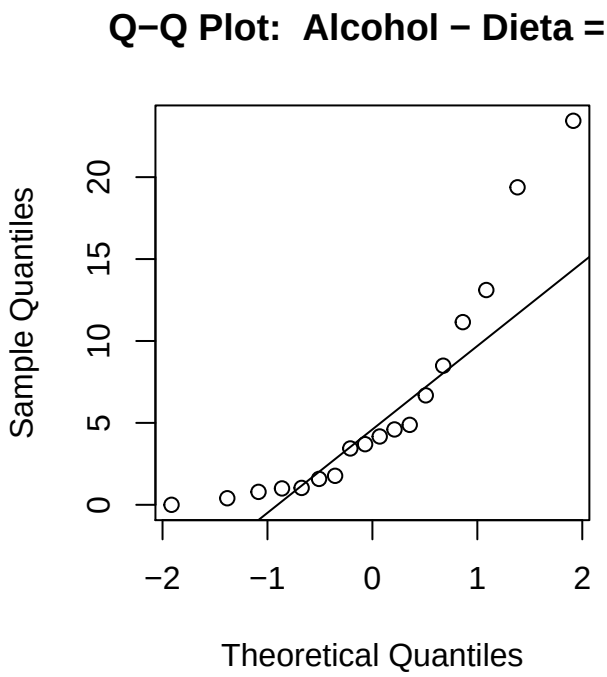
```
normal_prueba(diabetes$Alcohol[diabetes$Dieta == 0],'Alcohol - Dieta = 0')
```

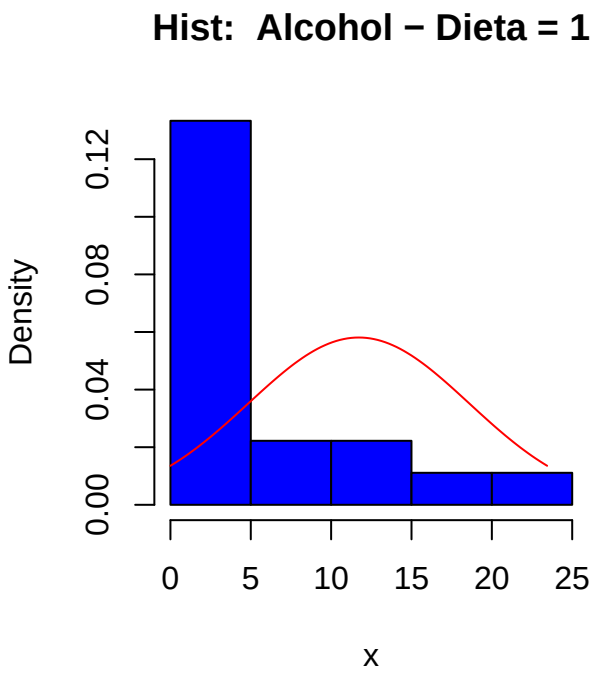
```
## [1] "Alcohol - Dieta = 0"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.23812, p-value = 0.008149
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.81177, p-value = 0.002238
```




```
normal_prueba(diabetes$Alcohol[diabetes$Dieta == 0], 'Alcohol - Dieta = 1')
```

```
## [1] "Alcohol - Dieta = 1"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.23812, p-value = 0.008149
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.81177, p-value = 0.002238
```

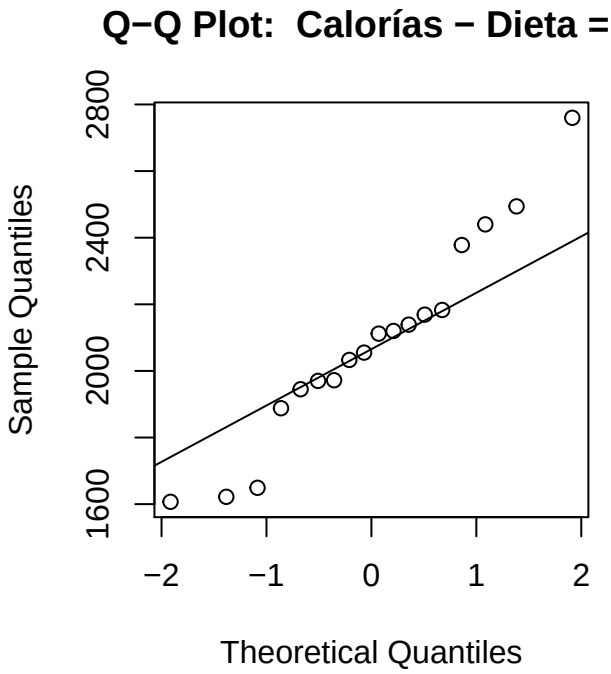


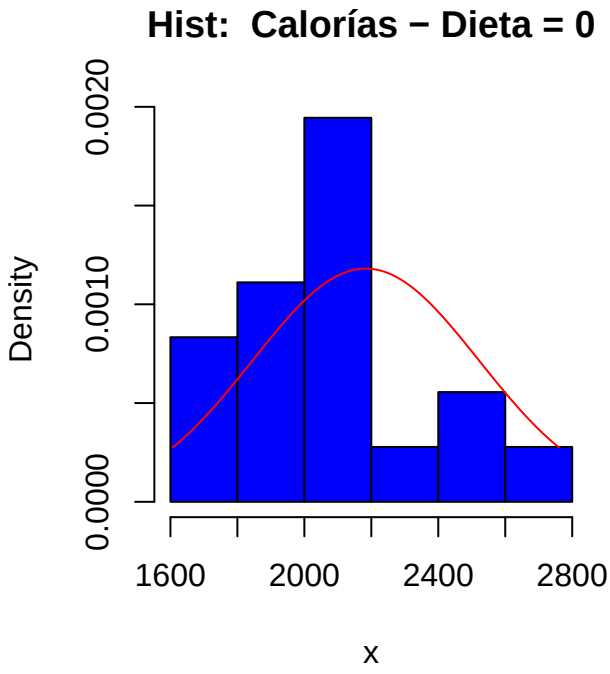


Para la variable Alcohol se puede concluir también que no presenta normalidad, tanto con los test estadísticos como con los gráficos. Por un lado, para ambos grupos el P-Valor de los dos tests estadísticos fue menor a 0.05. Por el otro se puede ver en los Q-Q Plots que los datos no tienen distribución normal con bastante claridad, porque los datos están muy alejados de la recta. Por último, en los histogramas también se aprecia que no existe una distribución normal de los datos en niguno de los dos grupos.

```
normal_prueba(diabetes$calorias[diabetes$Dieta == 0],'Calorías - Dieta = 0')
```

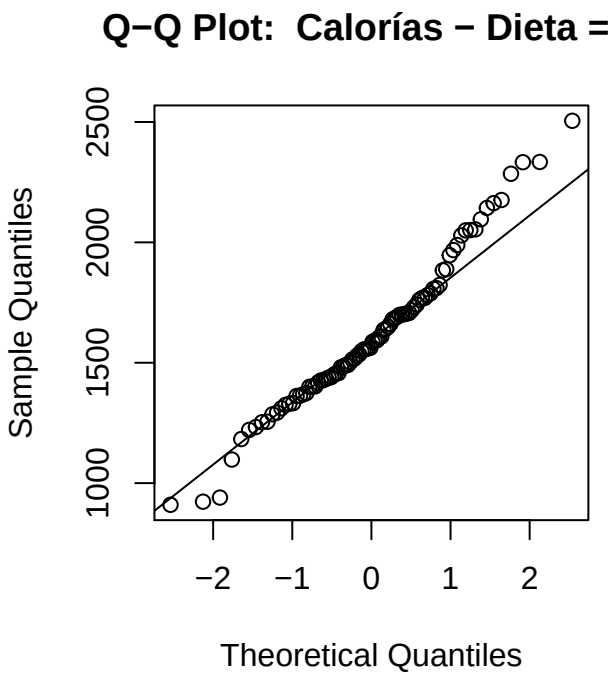
```
## [1] "Calorías - Dieta = 0"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.15202, p-value = 0.3313
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.95745, p-value = 0.5532
```

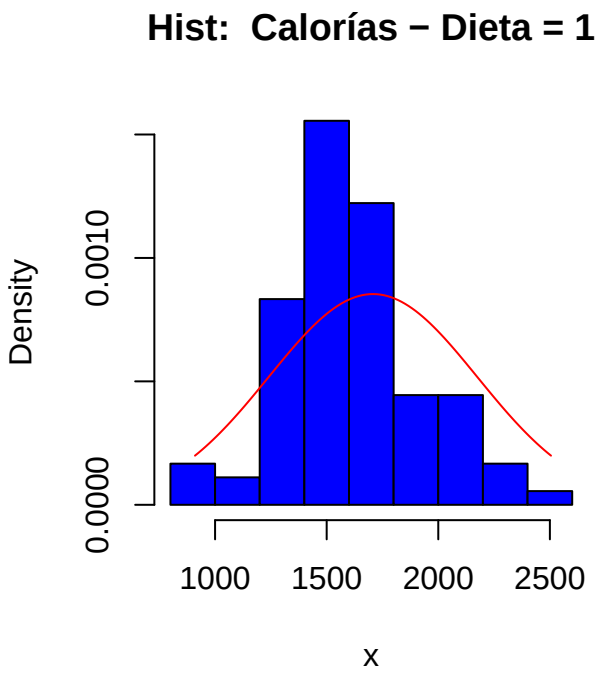





```
normal_prueba(diabetes$calorias[diabetes$Dieta == 1], 'Calorías - Dieta = 1')
```

```
## [1] "Calorías - Dieta = 1"
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.074122, p-value = 0.2584
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.97666, p-value = 0.105
```





Los resultados de los tests estadísticos, para ambos grupos de estudio, muestran que el P-Valor es mayor a 0.05, por lo que en ambos casos se acepta la hipótesis nula, que dice que la variable Calorías tiene distribución normal.Esto se puede ver también en los histogramas de ambos grupos, aunque no logré poner en escala la curva normal. Por otro lado, el Q-Q Plot muestra algunas desviaciones en los extremos superior e inferior, pero por lo mencionado anteriormente conlcuyo que esta variable tiene distribución normal en ambos grupos.

8. **Realice una tabla que resuma las pruebas de normalidad por variable. ¿Qué conclusiones sacó?**

```
filas = c("Grasas", "Alcohol", "Calorías")
df <- data.frame(Dieta_0= c("no normal", "no normal", "normal"),
                 Dieta_1 = c("no normal", "no normal", "normal"),  row.names = filas)

kable(df)
```

	Dieta_0	Dieta_1
Grasas	no normal	no normal
Alcohol	no normal	no normal
Calorías	normal	normal

9. **Para las variables numéricas calcule las medidas de tendencia central que mejor las representen y las medidas de dispersión. Realice una tabla que resuma los resultados.**

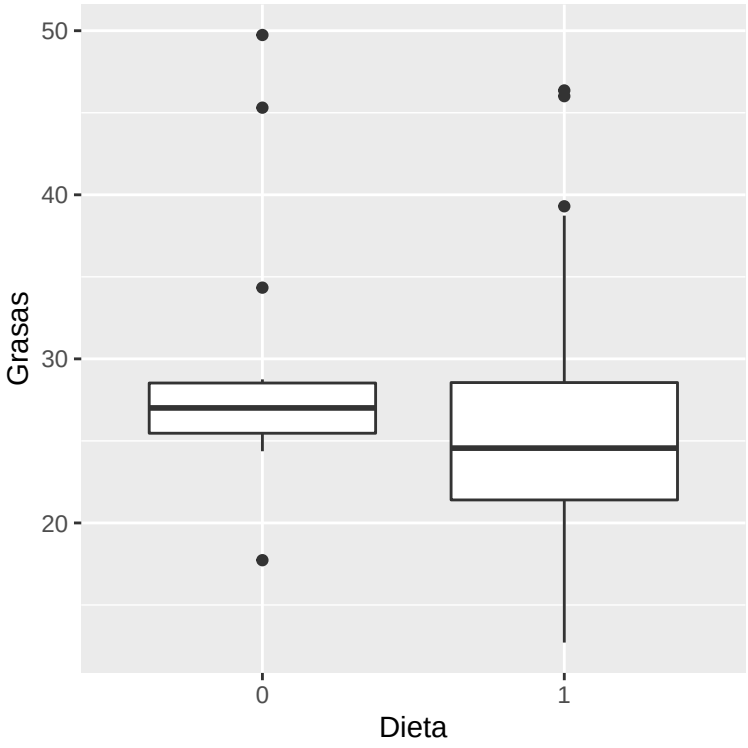
Para cada una de las tres variables numéricas calculo dos medidas de tendencia central, media y mediana, y una de dispersión, el desvío estándar.

```
variables = list( diabetes$Grasas, diabetes$Alcohol, diabetes$calorias)

medidas = double(length = 9)

i = 1
for (var in variables){
  k = i+2
  medidas[i:k] = c(round(mean(var), digits=4), round(median(var),digits=4), round(sd(var), digits=4))
  i = i +3
}
filas = c("Media", "Mediana", "Desvío Estándar")

tabla <- data.frame(Grasas = medidas[1:3],
                    Alcohol = medidas[4:6],
                    Calorias = medidas[7:9], row.names = filas)
kable(tabla)
```

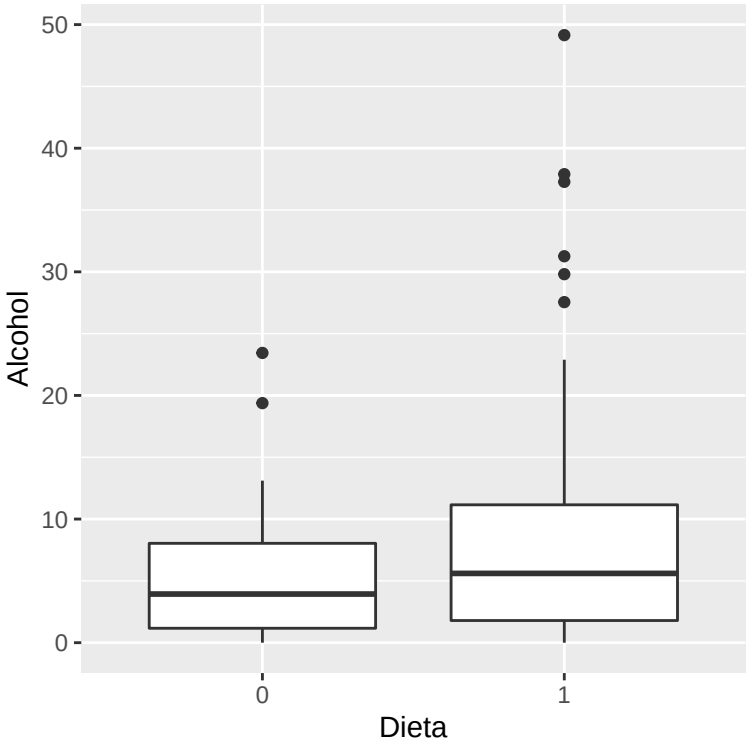
	Grasas	Alcohol	Calorias
Media	25.7805	8.3920	1693.6019
Mediana	25.2250	5.1850	1645.0000
Desvío Estándar	6.8480	9.4193	357.7166

10. **Para cada variable realice un Box-plot que compare ambos grupos de tratamiento. Analice los resultados obtenidos.**

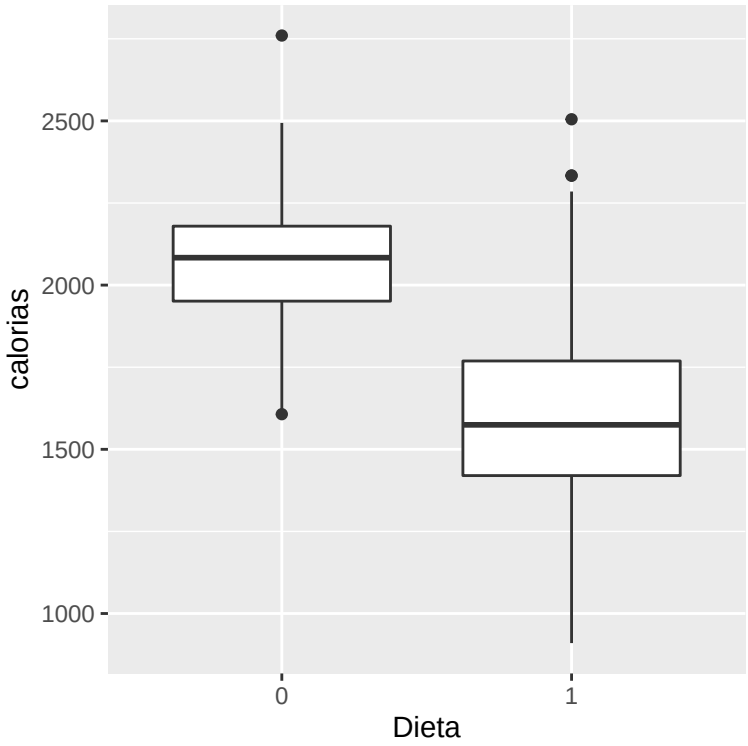
```
plt <- ggplot(diabetes, aes(x=Dieta ,y=Grasas)) +
geom_boxplot()
plt
```




```
plt2 <- ggplot(diabetes, aes(x=Dieta ,y=Alcohol)) +
geom_boxplot()
plt2
```



```
plt2 <- ggplot(diabetes, aes(x=Dieta ,y=calorias)) +
geom_boxplot()
plt2
```



Se puede observar que en las tres variables, el grupo de Dieta = 1 presenta una mayor simetría que el grupo de Dieta = 0, ya que la mediana se encuentra en el centro de la caja. En el caso particular de la varible Alcohol, existe una gran diferencia de tamaño entre los bigotes superiores e inferiores en ambos grupos, lo cual se pudo ver en los histogramas que presentaban un corrimiento hacia la izquierda. Por otro lado, comparando las variables entre sí se puede ver que Calorías es la más simétrica de las tres ya que ambos grupos presentan alta simetría y tiene que ver con que la distrubución es normal.