Trabajo Práctico Final

Axel Cesar Wood Niella, Lucía Bernochi, Priscilla Vanesa Tenas Vai

Fecha de entrega: 26/04/2021

1. Objetivos

El objetivo del taller y del trabajo es la presentación del diseño de una investigación basado en una base de datos de interés que en este caso se trato de biomarcadores urinarios del cáncer de páncreas y posterior análisis estadístico de la misma.

2. Introducción

El cáncer de páncreas se trata de un tipo de cáncer extremadamente mortal. Una vez diagnosticada la enfermedad, la tasa de supervivencia a cinco años es inferior al 10%. De todas formas, si el cáncer de páncreas es detectado en estadios tempranos las probabilidades de sobrevivir son mucho más altas. Sin embargo, este es difícil de detectar con anticipación ya que no causa síntomas de inmediato. Una vez que los síntomas aparecen suelen ser imperceptibles. Algunos de los síntomas incluyen: coloración amarillenta de la piel y los ojos, dolor en el abdomen, dolor en la espalda, pérdida de peso y fatiga. También, se presenta una dificultad extra en el diagnóstico debido a que el páncreas está oculto detrás de otros órganos por lo que el profesional de la salud no puede palpar los tumores en los exámenes de rutina. Para el diagnóstico efectivo, el profesional de la salud incluye: examen físico, pruebas de sangre, exámenes de imágenes y biopsia. Posteriormente al diagnóstico positivo los posibles tratamientos pueden incluir: cirugía, radiación y quimioterapia y terapia dirigida. Está última usa medicamentos u otras sustancias para combatir células cancerosas específicas y así causar el menor daño posible sobre las células normales y sanas.

En la actualidad no existen biomarcadores específicos con la capacidad de generar una detección temprana sobre el cáncer de páncreas. En la práctica clínica se utiliza como principal biomarcador el CA19-9 que no es específico al cáncer de páncreas, pero sirve para controlar la respuesta al tratamiento. Tradicionalmente, la principal fuente de biomarcadores es la sangre (como lo es para el caso de CA19-9) pero a partir de la orina surge una nueva fuente potencial para la búsqueda de biomarcadores ya que permite un muestreo completamente no invasivo, gran cantidades de recolección en volumen y facilidad en la repetición de mediciones.

Finalmente, el problema principal en la detección está en que el cáncer de páncreas no muestra síntomas hasta que este no ha sido diseminado por todo el cuerpo. En consecuencia, el objetivo está en encontrar una prueba de diagnóstica capaz de identificar a las personas con posibilidades de desarrollar la enfermedad.

En el caso puntual de la investigación elegida y con el objetivo mencionado anteriormente se confecciona un panel con diferentes biomarcadores urinarios (creatinina, LYVE1, REG1B y TFF1) para obtener un PancRISK que permite la estratificación de los pacientes en aquellos con riesgo “normal” o “elevado” de desarrollar cáncer pancreático (adenocarcinoma ductal pancreático).

A continuación, se define la pregunta PICO de la investigación.

- **Población (P):** Hombres y mujeres con un rango etario entre 26 y 89 años con histeorias clínicas diferentes: pacientes saludables, pacientes con afecciones pancreáticas no cancerosas, como pancreatitis crónica y pacientes con adenocarcinoma ductal pancreático.

- **Intervención (I):** Medición de cuatro biomarcadores urinarios: creatinina, LYVE1, REG1B y TFF1.
- **Comparación (C):** Pacientes con adenocarcinoma ductal pancreático vs. Pacientes sin adenocarcinoma ductual pancreático.
- **Outcomes (O):** Presencia o no de cáncer pancreático (adenocarcinoma ductal pancreático).

3. Materiales y métodos

Diseño de estudio, settings y participantes

La enfermedad tiene baja prevalencia (60 mil pacientes al año) por lo que es un estudio de casos y controles. En este estudios de casos y controles se tomo una base de datos de muestras tomadas en el Hospital Boston EEUU. Estas muestras fueron tomadas previo a la realización de una intervención quirúrgica o de tratamiento quimioterapéutico para el tratamiento de cancer de pancreas durante 6 meses. Se utilizaron muestras de orina y plasma y se emparearon por sexo y edad siempre que fue posible. Estas fueron preservadas y guardadas según procedimientos operativos estándar. Las muestras fueron tomadas tanto de hombres como mujeres en un rango etario de 45 a 80 años por ser la edad de mayor prevalencia de la enfermedad. Se tomaron muestras de pacientes sin afecciones pancreáticas o neoplasias malignas conocidas ni antecedentes de enfermedades renales en el momento de la recogida, otras procedían de pacientes con enfermedades hepatobiliares benignas (grupo benigno) y finalmente muestras que eran de pacientes con adenocarcinoma ductal pancreático. Las muestras del grupo benigno incluyen pancreatitis crónica, enfermedades de la vesícula biliar, lesiones quísticas del páncreas y casos con dolor abdominal y síntomas gastrointestinales sugestivos de origen pancreático.

Variables

Se determina como variable respuesta al diagnostico de cancer de pancreas con CA19_9 en plasma y creatinina, LYVE1, REG1B, TFF1, REG1A en orina como variables explicativas. Se debe tener en cuenta en el análisis que conforme la edad aumenta el riesgo de que las funciones renales no sean las adecuadas y que el biomarcador en orina no sea el esperado.

Data sources/ measurement

Las muestras de orina y plamsa fueron procesadas en grupos de 40 para llenar un kit de Elisa. Todas las mediciones fueron realizadas en duplicado para confirmar los resultados.

- REG1B: fue medida utilizando TMB Substrate Set y Stop Solution de BioLegend para su cuantificacion.
- Creatinina: se midió en el Laboratorio de Bioquímica Clínica del Universidad de Westminster usando un analizador ILab Aries del Laboratorio de Instrumentación según el protocolo del fabricante. Unidad: mmol/l

El resto de las variables fueron medidas utilizando el FLUOstar Omega Microplate Reader utilizando los kit de Elisa y sus limites de deteccion son los siguiente:

- CA19_9:0.3 U/ml
- TFF1: 3.91 pg/ml
- REG1B: 8 pg/ml
- LYVE1: 56 pg/ml

Sesgos

El posible sesgo que identificamos a la hora de realizar el estudio es el sesgo de seleccion. Consideramos este puede afectar el estudio ya que se tomaron muestras de paciente en un rango etario de 45 a 80 años siendo estas las edades con mayor prevalencia de la enfermedad. No obstante, las funciones renales de pacientes mayores de edad pueden funcionar de una manera distinta a la esperada y afectar la presencia del biomarcador.

Tamaño del Estudio

Variables Cuantitativa

Las variables cuantitativas fueron separadas en grupos dependiendo de provenian de una muestra de plasma o orina. Los pacientes no fueron agrupados en distintos grupos segun la edad cuestion que podria tener un efecto en los resultados encontrados.

4. Resultados

Lo primero que se realizo fue la carga del archivo datacancer.cvs y se observo el tipo de datos que contenía.

```
# Creo objeto DMO
data <- read.csv2(file.choose(), sep = ";")
# Muestro el tipo de datos en tabla
tab = data.frame(Variable = names(data), Tipo = sapply(data, class),
row.names = NULL)
```

En primer lugar, vemos que las variables sex y diagnosis se importaron como tipo character por lo que hay que cambiarlas a tipo factor ya que se tratan de variables dicotómicas y politómicas respectivamente.

En segundo lugar, vemos que las variables creatinine, LYVE1, REG1B y TFF1 se importaron como tipo character por lo que hay que cambiarlas a tipo numeric ya que se tratan de variables numéricas.

Mediante el siguiente código se cambia el tipo de variables.

```
# Cambio tipo de variables
data$sex <- as.factor(data$sex)
data$diagnosis <- as.factor(data$diagnosis)
data$creatinine <- as.numeric(data$creatinine)
```

```
## Warning: NAs introducidos por coerción
```

```
data$LYVE1 <- as.numeric(data$LYVE1)
```

```
## Warning: NAs introducidos por coerción
```

```
data$REG1B <- as.numeric(data$REG1B)
```

```
## Warning: NAs introducidos por coerción
```

```r
data$TFF1 <- as.numeric(data$TFF1)
```

```
## Warning: NAs introducidos por coerción
```

```r
# Muestro el tipo de datos en tabla
tab = data.frame(Variable = names(data), Tipo = sapply(data, class), row.names = NULL)
```

Se procede a trabajar únicamente con los grupos 1 y 3 correspondientes a sanos vs. pacientes con cáncer de páncreas. Además, dado que hay datos faltantes procedemos a realizar el análisis para observar si los mismos corresponden a un porcentaje menor al 20%.

```r
# En primer lugar nos quedamos con el grupo 1 y 3 (sanos vs cáncer de pancreas)
data1 <- data[data$diagnosis !=2,]

data7 <- data1[!(is.na(data1$LYVE1) & is.na(data1$REG1B) & is.na(data1$TFF1)),]
data8 <- data7[!(is.na(data7$LYVE1) & is.na(data7$REG1B)),]
data8 <- data8[!(is.na(data8$TFF1) & is.na(data8$REG1B)),]
data8 <- data8[!(is.na(data8$TFF1) & is.na(data8$LYVE1)),]
data9 <- rownames(data8[is.na(data8$LYVE1),])

data8 <- data8[sample(1:nrow(data8), 0.984*nrow(data8)),] #50% of data

sum(is.na(data8$REG1B))
```

```
## [1] 21
```

```r
sum(is.na(data8$TFF1))
```

```
## [1] 13
```

```r
sum(is.na(data8$LYVE1))
```

```
## [1] 27
```

Una vez verificado que los datos obtenidos tienen una faltante menor al 20% se procede a reemplazar los datos faltantes por la media correspondiente para cada una de las variables. También, se cambian los grupos 1 y 3 por grupos 0 y 1 respectivamente, es decir, control vs. cáncer de páncreas. Además, se realizan los cambios de tipo de variables correspondientes.

```r
# Cargamos el data set final con el que vamos a trabajar
datafinal <- read.csv2(file.choose(), sep = ",")

# Cambio tipo de variables a numeric
datafinal$creatinine <- as.numeric(datafinal$creatinine)
datafinal$LYVE1 <- as.numeric(datafinal$LYVE1)
datafinal$REG1B <- as.numeric(datafinal$REG1B)
datafinal$TFF1 <- as.numeric(datafinal$TFF1)

# Cambiamos el grupo 1 y 3 por grupo 0 y 1 respectivamente (Control vs Cáncer)
datafinal$diagnosis[datafinal$diagnosis==1]<-0
```

```r
datafinal$diagnosis[datafinal$diagnosis==3]<-1

# Cambio tipo de variables a factor
datafinal$sex <- as.factor(datafinal$sex)
datafinal$diagnosis <- as.factor(datafinal$diagnosis)

# Ponemos la media en los NA
datafinal$LYVE1[is.na(datafinal$LYVE1)]<- mean(datafinal$LYVE1, na.rm = TRUE)
datafinal$REG1B[is.na(datafinal$REG1B)]<- mean(datafinal$REG1B, na.rm = TRUE)
datafinal$TFF1[is.na(datafinal$TFF1)]<- mean(datafinal$TFF1, na.rm = TRUE)

# Muestro el tipo de datos en tabla
tab = data.frame(Variable = names(datafinal), Tipo = sapply(datafinal, class),
                 row.names = NULL)
```

Identificación de variables, su forma de medirlas y su distribución

En primer lugar, se separa la base de datos por grupo correspondientes a Control vs. Cáncer de páncreas.

```r
G0 <- subset(datafinal, datafinal$diagnosis==0)
G1 <- subset(datafinal, datafinal$diagnosis==1)
```

A continuación, se realiza el análisis para cada variable.

Variable: sex En primer lugar, se genera la tabla de frecuencias.

```r
# Tabla de frecuencias
tf <- table(datafinal$sex, datafinal$diagnosis, dnn = c('Genero', 'Cáncer de páncreas'))
colnames(tf) <- c('No', 'Si')
rownames(tf) <- c('Fem', 'Masc')
kable(tf)
```

	No	Si
Fem	68	12
Masc	29	16

La variable sex se trata de una variable del tipo cualitativa nominal categórica y en consecuencia posee una distribución binominal.

Dado que se trata de variables dicotómicas para proceder a realizar el análisis de la independencia entre las mismas se debería realizar test de Chi cuadrado. Además, dado que la tabla de contingencia es de 2x2 se debe realizar el test utilizando la corrección de continuidad. La hipótesis nula y alternativa en este caso serán:

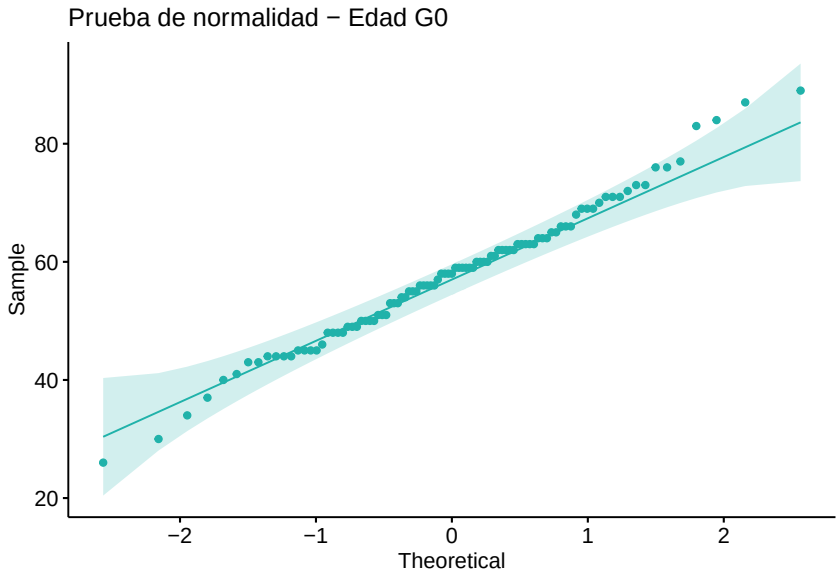
- H_0: Se trata de variables independientes.
- H_1: No se trata de variables independientes.

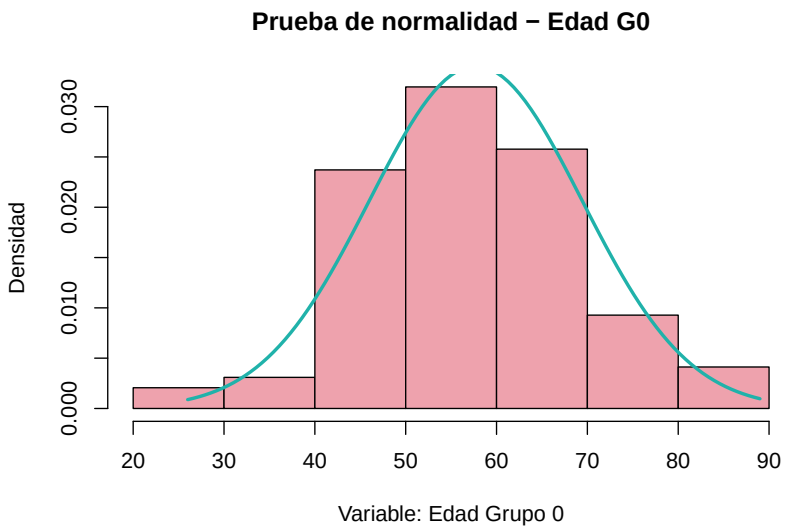
```r
prueba_normalidad <- function(x, xlab='datos', ylab='Densidad', main='Titulo'){
  print(shapiro.test(x))
  print(lillie.test(x))
  print(ggqqplot(x, main= main, col = "lightseagreen"))
  min <- min(x)
  max <- max(x)
  media <- mean(x)
  des <- sd(x)
  hist(x, freq = FALSE,
       main = main, xlab = xlab, ylab = ylab, border = 'black',
       col = "lightpink2")
  curve(dnorm(x, media, des), min, max, add = TRUE,
        col="lightseagreen",lwd=2.5)
}
```

Variable: age Finalmente, una vez armada la función anterior se comprueba mediante la misma si la variable edad para el grupo Control tiene una distribución normal o no.

```r
edad0 <- G0$age
prueba_normalidad(edad0, 'Variable: Edad Grupo 0', main= 'Prueba de normalidad - Edad G0')
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.9901, p-value = 0.6925
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.058762, p-value = 0.5633
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es igual a 0.6925 y mayor a 0.05, entonces rechazo la hipótesis alternativa y acepto la nula. Para el caso del test de Lilliefors el valor de p es igual a 0.5633 y mayor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se rechaza la hipótesis alternativa y se acepta la nula.

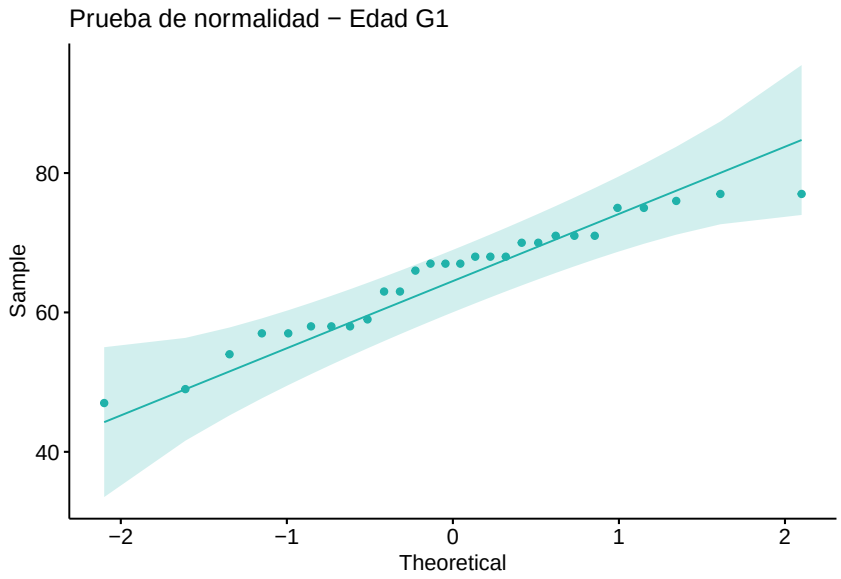
En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante uniforme respecto de la recta de distribución ideal. Por otra parte, el histograma presenta una forma acampanada y simétrica.

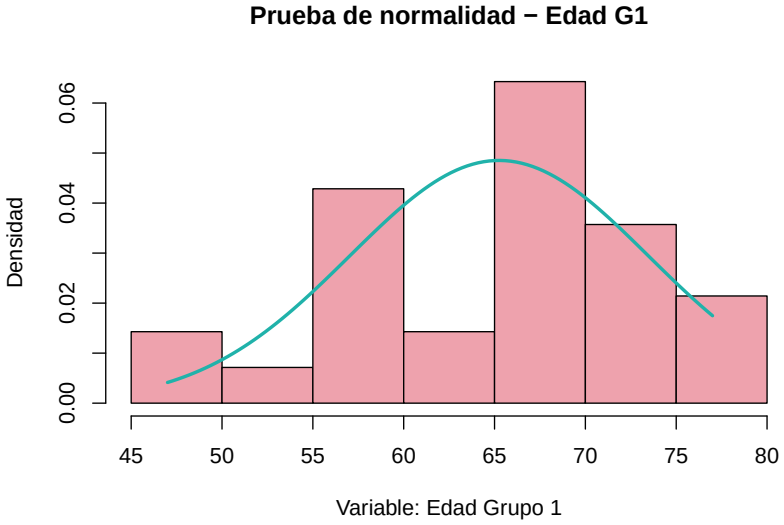
Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable edad para el grupo 0 tiene una distribución normal.

A continuación, se realiza el mismo análisis para el grupo 1.

```
edad1 <- G1$age
prueba_normalidad(edad1, 'Variable: Edad Grupo 1', main= 'Prueba de normalidad - Edad G1')
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.9459, p-value = 0.1561
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.1557, p-value = 0.08022
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es igual a 0.1561 y mayor a 0.05, entonces rechazo la hipótesis alternativa y acepto la nula. Para el caso del test de Lilliefors el valor de p es igual a 0.08022 y mayor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se rechaza la hipótesis alternativa y se acepta la nula.

En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada y simétrica.

Por último, mediante el resultado de ambos tests y pese a el análisis de los gráficos se puede concluir que la variable edad para el grupo 1 tiene una distribución normal.

Finalmente, la variable age/edad se trata de una variable explicatoria del tipo cuantitativa continua que se mide en unidades de años. Como se comprobo anteriormente se trata de una variable que posee distribución normal para ambos grupos y por lo anterior para realizar la comparación entre los mismos se debe analizar la homocedasticidad para poder posteriormente realizar el test Z. Para esto previamente se debe realizar el

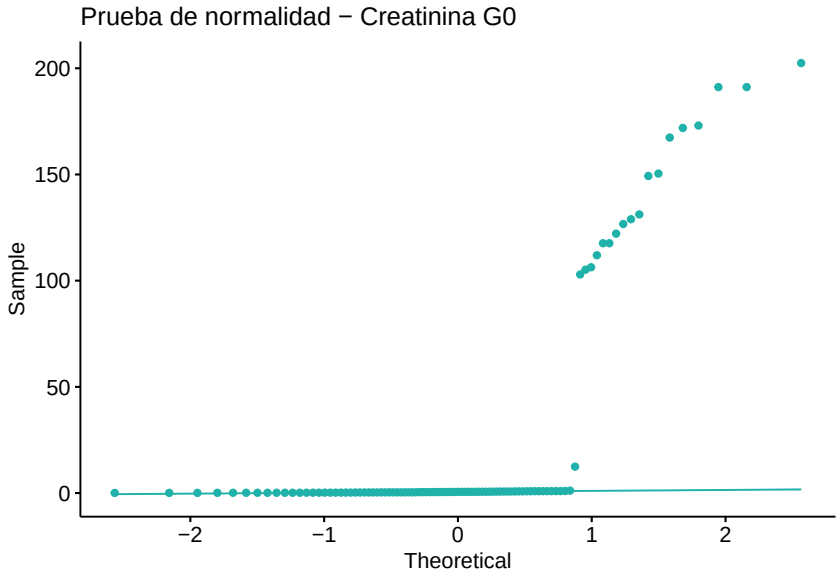
test de Levene. La hipótesis nula para este caso será que las varianzas de ambos grupos son homogéneas, mientras que la hipótesis alternativa será que las varianzas de ambos grupos son heterogéneas. Si se cumplen las condiciones de homocedasticidad y la muestra es mayor a 30 como lo es en este caso, se puede realizar el test Z. Las hipótesis del test serán:

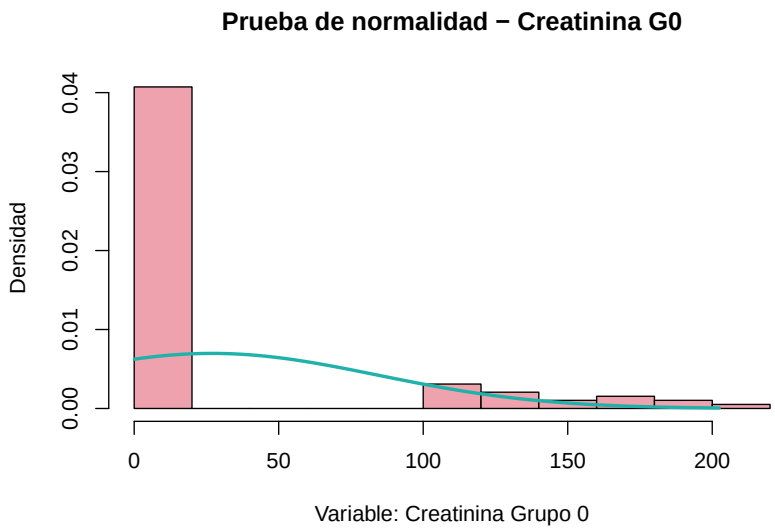
- $H_0 : \mu_{Grupo0} - \mu_{Grupo1} = 0$
- $H_1 : \mu_{Grupo0} - \mu_{Grupo1} \neq 0$

Variable: Creatinine En primer lugar, se realiza el análisis de normalidad de la variable para el grupo Control mediante la función que se encuentra a continuación. La función devuelve 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).

```
creatinina0 <- G0$creatinine
prueba_normalidad(creatinina0, 'Variable: Creatinina Grupo 0', main= 'Prueba de normalidad - Creatinina
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.51607, p-value = 2.512e-16
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.47837, p-value < 2.2e-16
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es menor a 2.512e-16 y menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es menor a 2.2e-16 y por lo tanto menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

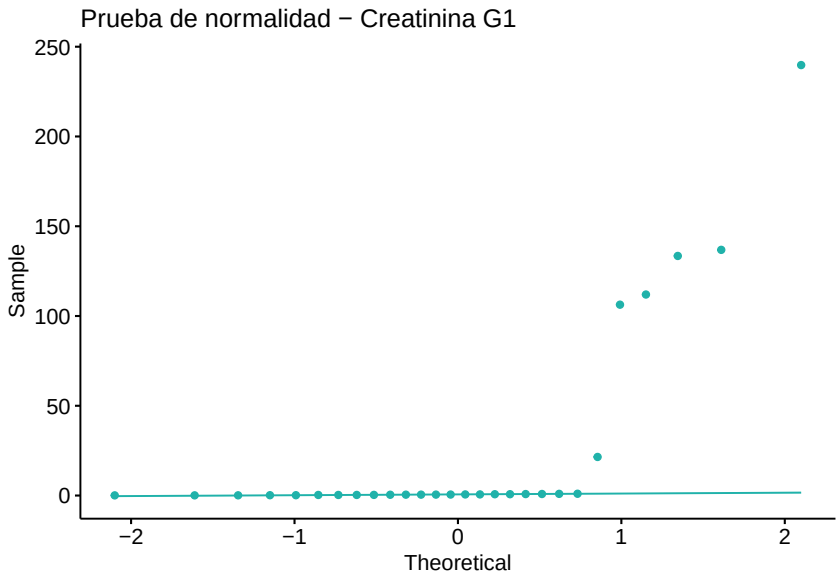
En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

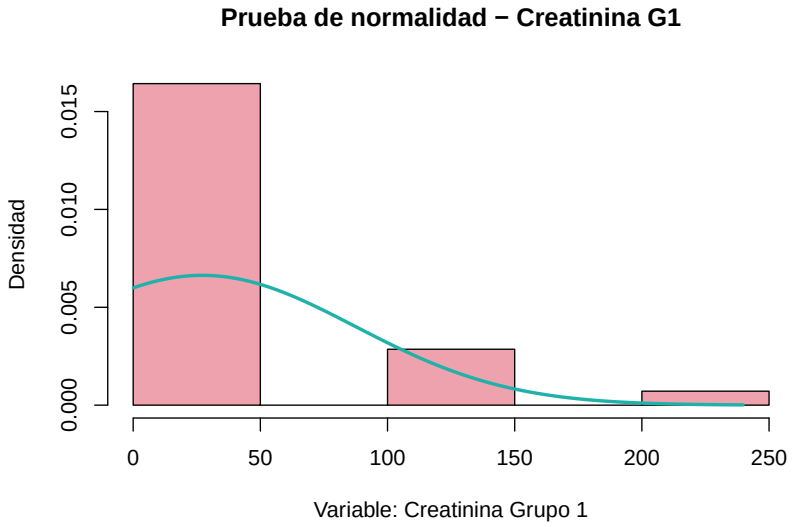
Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable creatinine para el grupo 0 no tiene una distribución normal.

A continuación, se realiza el mismo análisis para el grupo 1.

```
creatinina1 <- G1$creatinine
prueba_normalidad(creatinina1, 'Variable: Creatinina Grupo 1', main= 'Prueba de normalidad - Creatinina
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.51872, p-value = 1.79e-08
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.45397, p-value < 2.2e-16
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es menor a 1.79e-08 y menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es menor a 2.2e-16 y por lo tanto menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable creatinine para el grupo 1 no tiene una distribución normal.

Finalmente, la variable creatinine se trata de una variable explicatoria del tipo cuantitativa continua que se mide en unidades de mmol/l. Como se comprobó anteriormente se trata de una variable que posee distribución no normal para ambos grupos y por lo anterior para realizar la comparación entre los mismos se debería utilizar el test de Wilcoxon. Las hipótesis del test son:

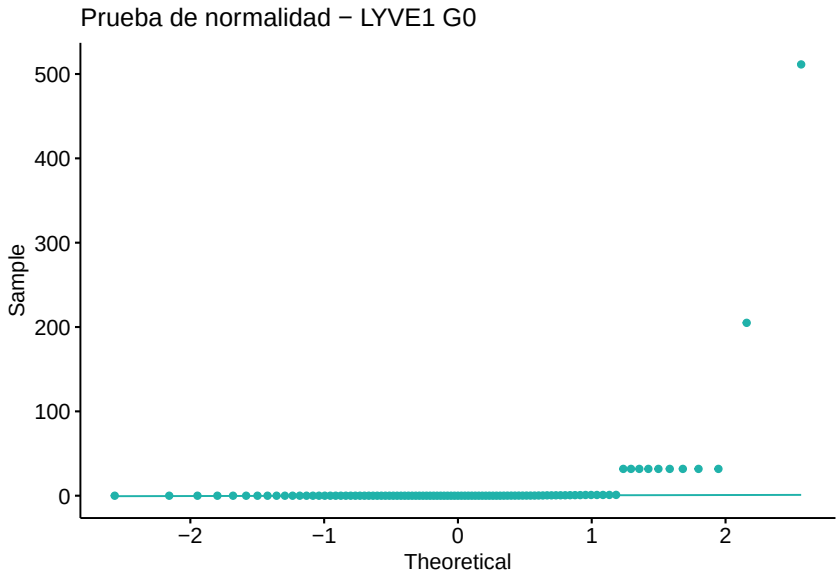
- $H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- $H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

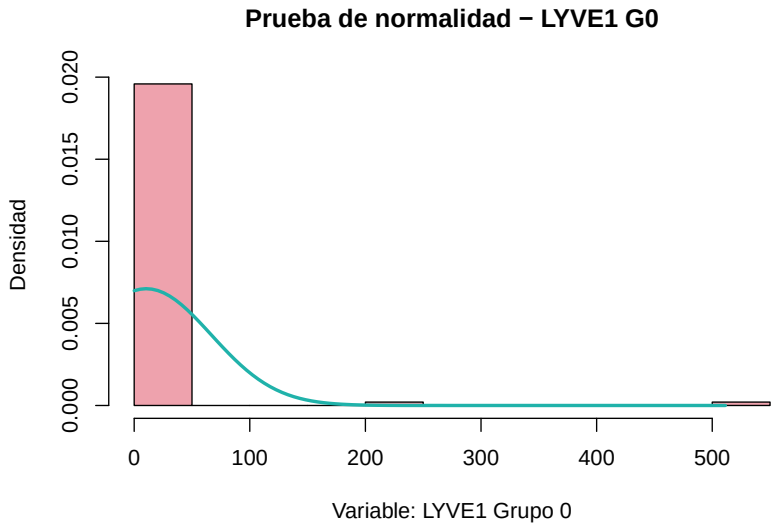
Donde θ corresponde a la mediana de la variable.

Variable: LYVE1 En primer lugar, se realiza el análisis de normalidad de la variable para el grupo Control mediante la función que se encuentra a continuación. La función devuelve 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).

```
LYVE10 <- G0$LYVE1
prueba_normalidad(LYVE10, 'Variable: LYVE1 Grupo 0', main= 'Prueba de normalidad - LYVE1 G0')
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.17755, p-value < 2.2e-16
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.45377, p-value < 2.2e-16
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es menor a 2.2e-16 y por lo tanto menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es menor a 2.2e-16 y por lo tanto menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

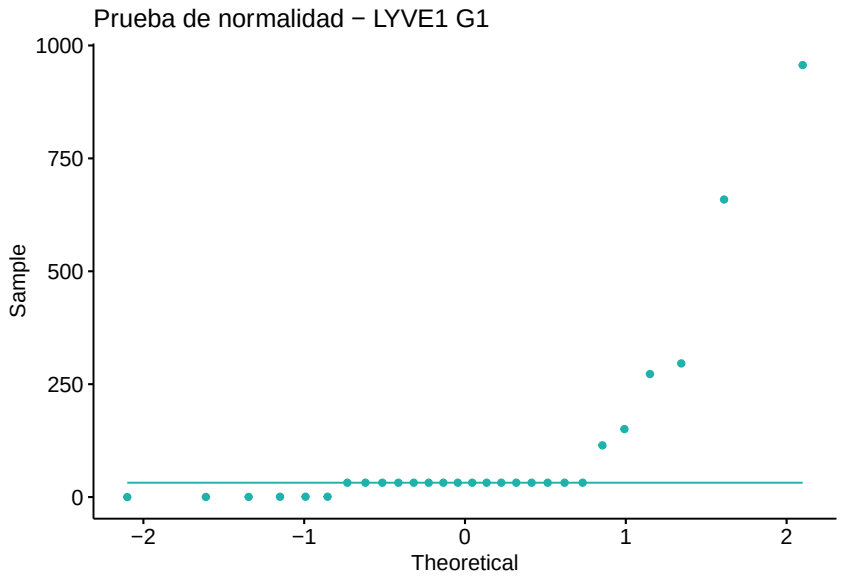
En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

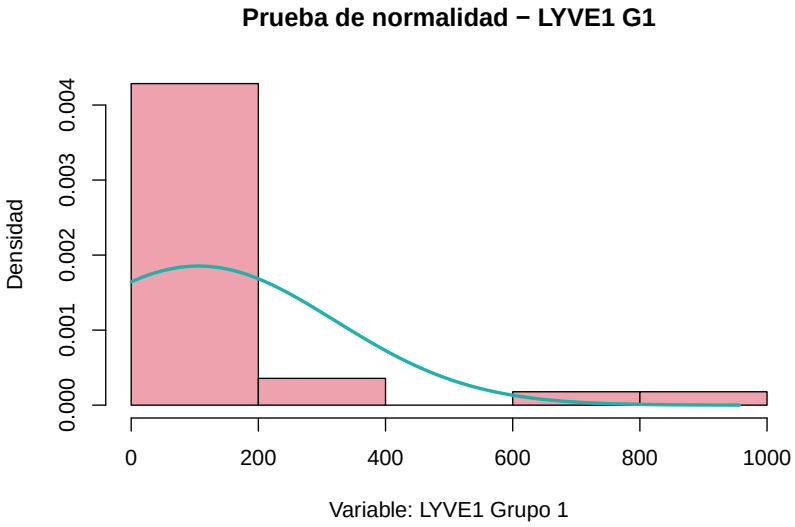
Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable LYVE1 para el grupo 0 no tiene una distribución normal.

A continuación, se realiza el mismo análisis para el grupo 1.

```
LYVE11 <- G1$LYVE1
prueba_normalidad(LYVE11, 'Variable: LYVE1 Grupo 1', main= 'Prueba de normalidad - LYVE1 G1')
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.49652, p-value = 1.052e-08
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.42015, p-value = 3.452e-14
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es igual a 1.052e-08 y menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es igual a 3.452e-14 y menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable LYVE1 para el grupo 1 no tiene una distribución normal.

Finalmente, la variable LYVE1 se trata de una variable explicatoria del tipo cuantitativa continua que se mide en unidades de pg/ml. Como se comprobo anteriormente se trata de una variable que posee distribución no normal para ambos grupos y por lo anterior para realizar la comparación entre los mismos se debería utilizar el test de Wilcoxon. Las hipótesis del test son:

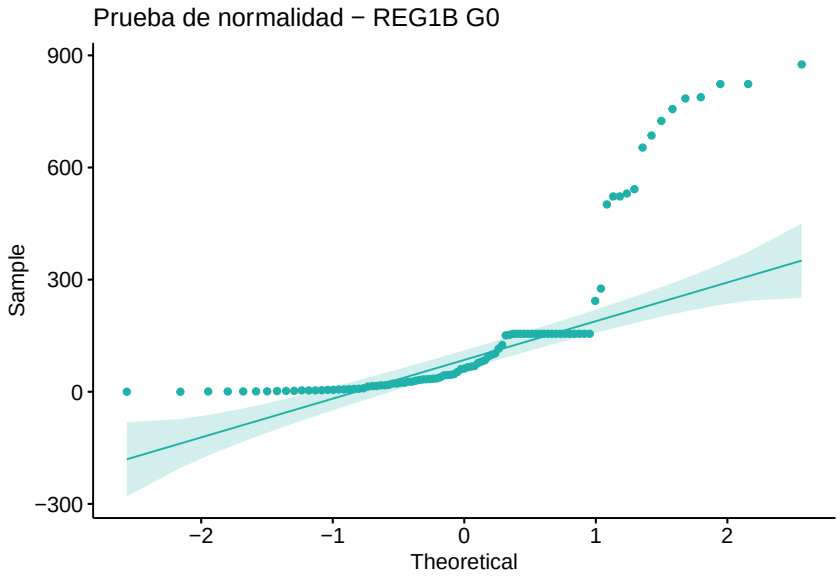
- $H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- $H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

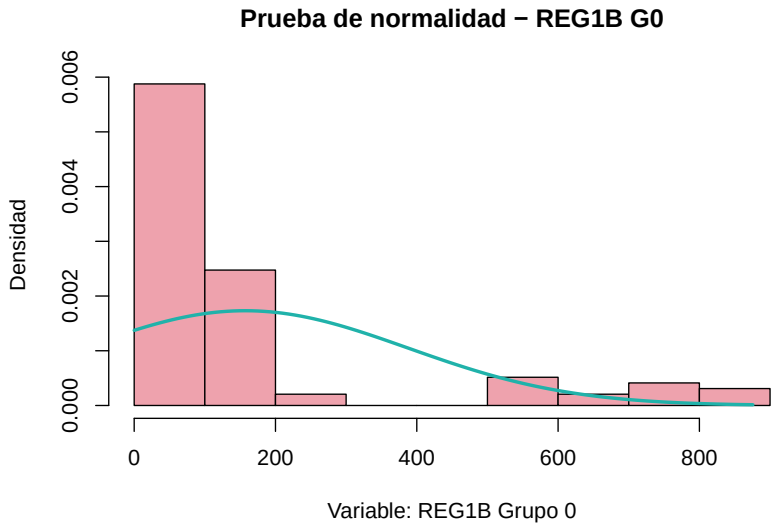
Donde θ corresponde a la mediana de la variable.

Variable: REG1B En primer lugar, se realiza el análisis de normalidad de la variable para el grupo Control mediante la función que se encuentra a continuación. La función devuelve 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).

```
REG1B0 <- G0$REG1B
prueba_normalidad(REG1B0, 'Variable: REG1B Grupo 0', main= 'Prueba de normalidad - REG1B G0')
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.66248, p-value = 1.307e-13
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.33843, p-value < 2.2e-16
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es igual 1.307e-13 y menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es menor a 2.2e-16 y por lo tanto menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

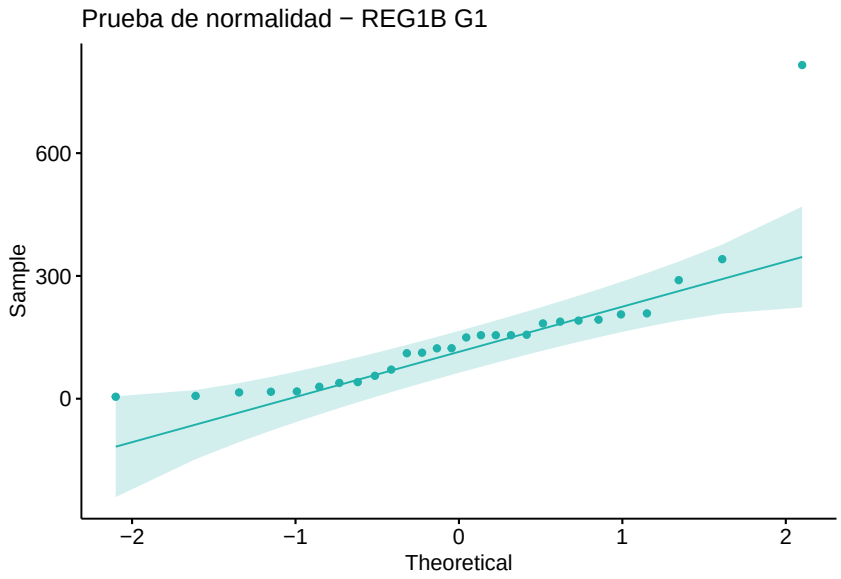
En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

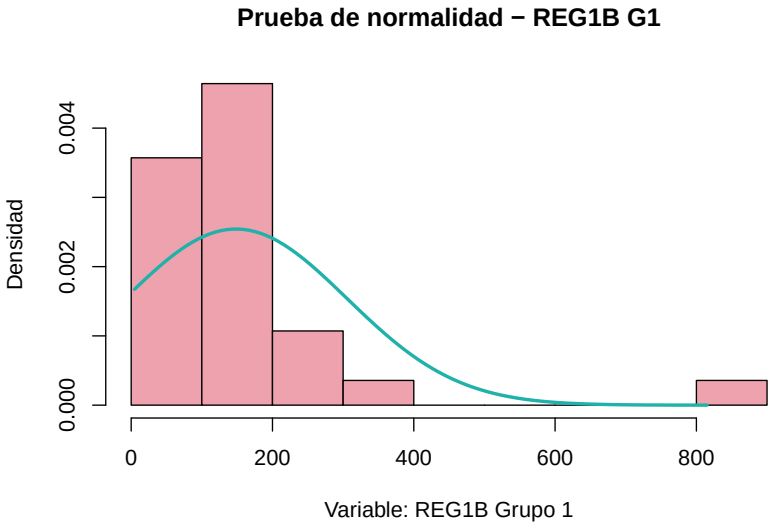
Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable REG1B para el grupo 0 no tiene una distribución normal.

A continuación, se realiza el mismo análisis para el grupo 1.

```
REG1B1 <- G1$REG1B
prueba_normalidad(REG1B1, 'Variable: REG1B Grupo 1', main= 'Prueba de normalidad - REG1B G1')
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.70028, p-value = 2.875e-06
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.24385, p-value = 0.0001734
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es igual 2.875e-06 y menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es menor a 0.0001734 y menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

En el gráfico de Q-Q plot podemos observar que la distribución de puntos es dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable REG1B para el grupo 1 no tiene una distribución normal.

Finalmente, la variable REG1B se trata de una variable explicatoria del tipo cuantitativa continua que se mide en unidades de pg/ml. Como se comprobo anteriormente se trata de una variable que posee distribución no normal para ambos grupos y por lo anterior para realizar la comparación entre los mismos se debería utilizar el test de Wilcoxon. Las hipótesis del test son:

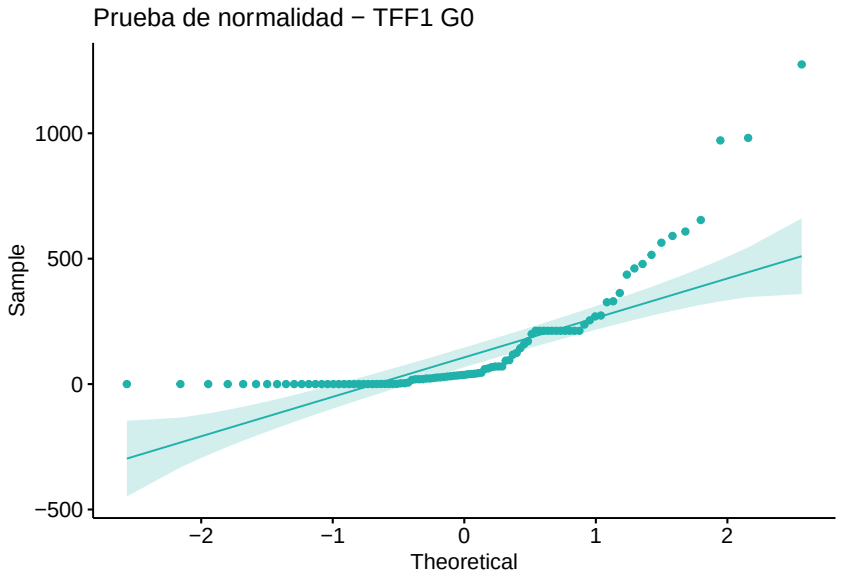
- $H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- $H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

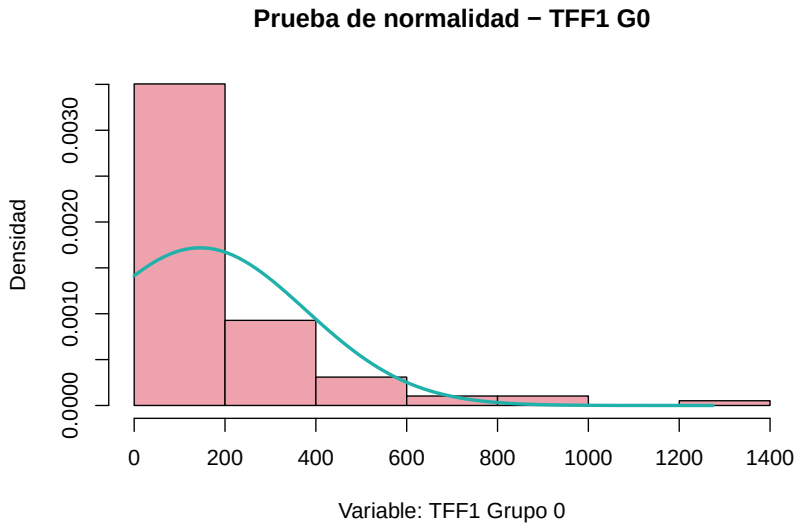
Donde θ corresponde a la mediana de la variable.

Variable: TFF1 En primer lugar, se realiza el análisis de normalidad de la variable para el grupo Control mediante la función que se encuentra a continuación. La función devuelve 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).

```
TFF10 <- G0$TFF1
prueba_normalidad(TFF10, 'Variable: TFF1 Grupo 0', main= 'Prueba de normalidad - TFF1 G0')
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.66383, p-value = 1.396e-13
##
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.26578, p-value < 2.2e-16
```





Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es igual 1.396e-13 y menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es menor a 2.2e-16 y por lo tanto menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

En el gráfico de Q-Q plot podemos observar que la distribución de puntos es bastante dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable TFF1 para el grupo 0 no tiene una distribución normal.

A continuación, se realiza el mismo análisis para el grupo 1.

```
TFF11 <- G1$TFF1
#prueba_normalidad(TFF11, 'Variable: TFF1 Grupo 1', main= 'Prueba de normalidad - TFF1 G1')
```

Analizando los tests estadísticos y sabiendo que la hipótesis nula en este caso sería que la distribución de la variable es normal, podemos observar que para el test de Shapiro-Wilk el valor de p es igual 2.466e-05 y menor a 0.05, entonces acepto la hipótesis alternativa y rechazo la nula. Para el caso del test de Lilliefors el valor de p es igual a 0.001016 y menor a 0.05, por lo que, al igual que en el test de Shapiro-Wilk se acepta la hipótesis alternativa y se rechaza la nula.

En el gráfico de Q-Q plot podemos observar que la distribución de puntos es dispersa respecto de la recta de distribución ideal. Por otra parte, el histograma no presenta una forma acampanada ni simétrica.

Por último, mediante el resultado de ambos tests y el análisis de los gráficos se puede concluir que la variable TFF1 para el grupo 1 no tiene una distribución normal.

Finalmente, la variable TFF1 se trata de una variable explicatoria del tipo cuantitativa continua que se mide en unidades de pg/ml. Como se comprobo anteriormente se trata de una variable que posee distribución no normal para ambos grupos y por lo anterior para realizar la comparación entre los mismos se debería utilizar el test de Wilcoxon. Las hipótesis del test son:

- $H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- $H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

Donde θ corresponde a la mediana de la variable.

Variable: diagnosis

Analisis univariado

Variable: Sex La variable dicotomica Sex se debe analizar utilizando el Test Chi Cuadrado con corrección de continuidad como se menciono anteriormente. La tabla ce contingencia de esta variable se definió como *tf*. Las hipótesis nula y alternativa en este caso serán:

- H_0: Se trata de variables independientes.
- H_1: No se trata de variables independientes.

```
chisq.test(tf, correct = TRUE)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tf
## X-squared = 5.8681, df = 1, p-value = 0.01542
```

Dado que el p-value es menor que 0.05 se acepta la hipótesis alternativa, rechazando la hipótesis nula. Es decir, el genero y el cáncer de páncreas no son variables independientes.

Variable: Age Si se cumplen las condiciones de homocedasticidad y la muestra es mayor a 30 como lo es en este caso, se puede realizar el test Z. Las hipótesis del test serán:

- $H_0 : \mu_{Grupo0} - \mu_{Grupo1} = 0$
- $H_1 : \mu_{Grupo0} - \mu_{Grupo1} \neq 0$

Previamente se realiza un Levene Test para analizar la homocedasticidad, donde la hipótesis nula es que las varianzas de ambas variables son homogéneas y la hipótesis alternativa establece que son distintas.

```
leveneTest(y= datafinal$age, group= datafinal$diagnosis)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  2.9808 0.08677 .
##       123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos observar que el p-valor es mayor a 0.05 por lo que no se rechaza la hipótesis nula, por lo tanto cumple condición para realizar el Test Z.

```
## No se que sigma poner en cada uno
sigma.x <- sd(edad0)
sigma.y <- sd(edad1)
z.test(edad0,edad1, mu = 0, sigma.x=sigma.x, sigma.y =sigma.y, conf.level = 0.95)
```

```
##
##  Two-sample z-Test
##
## data:  edad0 and edad1
## z = -3.8511, p-value = 0.0001176
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.375321  -3.701999
## sample estimates:
## mean of x mean of y
##  57.71134  65.25000
```

Como se puede observar el p-valor es menor que 0.05 por lo que se rechaza la hipotesis nula, y se cumple que la diferencia entre las medias de edad del grupo 1 y del grupo 2 es distinta de cero. Se puede ver que las medias estimadas son 57.71134 y 65.25000 para los grupos 1 y 2 respectivamente.

Variable: Creatinine Entonces, se aplica el test Wilcoxon. Las hipotesis del mismo son las siguientes:

- La hipótesis nula establece que la mediana del grado de dolor del grupo 1 no difiere de la del grupo 2. $\rightarrow$ $H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- La hipótesis alternativa afirma que las medianas son distintas $\rightarrow$ $H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

Donde θ corresponde a la mediana de la variable.

Recordamos que anteriormente se definieron *creatinina0* y *creatinina1* como la creatinina para los grupos 0 y 1 respectivamente. Los datos no están apareados por lo que se determina $paired = F$ en el test.

```
wilcox.test(creatinina0, creatinina1, paired = F, conf.int =0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  creatinina0 and creatinina1
## W = 1285.5, p-value = 0.6698
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.2375452  0.1470490
## sample estimates:
## difference in location
##            -0.03394606
```

Como se puede observar, el p-valor es mayor a 0.05 por lo que se rechaza la hipótesis nula.

Variable: LYVE1 Recordamos que las hipotesis para el test Wilcoxon son las siguientes:

- $H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- $H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

Donde θ corresponde a la mediana de la variable.

Como vimos anteriormente, se definieron *LYVE10* y *LYVE11* como la LYVE1 para los grupos 0 y 1 respectivamente. Los datos no están apareados por lo que se determina *paired* = F en el test.

```
wilcox.test(LYVE10, LYVE11, paired = F, conf.int =0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  LYVE10 and LYVE11
## W = 285.5, p-value = 1.845e-10
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -31.79223 -31.61921
## sample estimates:
## difference in location
##              -31.78906
```

Como se puede observar, el p-valor es menor a 0.05 por lo que se rechaza la hipótesis nula. Además, la mediana de la diferencia entre las muestras del grupo 0 y del grupo 1 para la variable *LYVE1* es de -31.78906 pg/ml, con un intervalo de confianza del 95 % entre -31.79223 pg/ml y -31.61921 pg/ml.

Variable: REG1B Entonces, se aplica el test Wilcoxon. Las hipotesis del mismo son las siguientes:

- La hipótesis nula establece que la mediana del grado de dolor del grupo 1 no difiere de la del grupo 2. $\rightarrow H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- La hipótesis alternativa afirma que las medianas son distintas $\rightarrow H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

Donde θ corresponde a la mediana de la variable.

Recordamos que anteriormente se definieron *REG1B0* y *REG1B1* como la REG1B para los grupos 0 y 1 respectivamente. Los datos no están apareados por lo que se determina *paired* = F en el test.

```
wilcox.test(REG1B0, REG1B1, paired = F, conf.int =0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  REG1B0 and REG1B1
## W = 1056.5, p-value = 0.07389
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -78.73506100   0.04301019
## sample estimates:
## difference in location
##              -29.97464
```

Como se puede observar, el p-valor es mayor a 0.05 por lo que se acepta la hipótesis nula.

	Variables Utilizadas	AUC
Modelo 1	LYVE1	0.5968
Modelo 2	TFF1	0.6572
Modelo 3	LYVE1 + TFF1	0.7204

Variable: TFF1 Recordamos que las hipotesis para el test Wilcoxon son las siguientes:

- $H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$
- $H_1 : \theta_{Grupo0} - \theta_{Grupo1} \neq 0$

Donde θ corresponde a la mediana de la variable.

Recordamos que anteriormente se definieron *TFF10* y *TFF11* como la TFF1 para los grupos 0 y 1 respectivamente. Los datos no están apareados por lo que se determina *paired* = *F* en el test.

```
wilcox.test(TFF10, TFF11, paired = F, conf.int =0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  TFF10 and TFF11
## W = 677, p-value = 5.53e-05
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -255.37294  -79.62802
## sample estimates:
## difference in location
##              -175.7217
```

Como se puede observar, el p-valor es menor a 0.05 por lo que se rechaza la hipótesis nula. Además, la mediana de la diferencia entre las muestras del grupo 0 y del grupo 1 para la variable *TFF1* es de -175.7217, con un intervalo de confianza del 95 % entre -255.37294 pg/ml y -79.62802 pg/ml.

En resumen:

Variable	Test	Hipotesis	p-valor	Inferencia
Sex	Chi Cuadrado	H_0: Se trata de variables independientes	0.01542	Se rechaza H_0
Age	Test Z	$H_0 : \mu_{Grupo0} - \mu_{Grupo1} = 0$	0.0001176	Se rechaza H_0
Creatinina	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	0.6698	Se acepta H_0
LYVE1	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	1.845e-10	Se rechaza H_0
REG1B	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	0.07389	Se acepta H_0
TFF1	Wilcoxon	$H_0 : \theta_{Grupo0} - \theta_{Grupo1} = 0$	5.53e-05	Se rechaza H_0

```
min( datafinal$age)
```

```
## [1] 26
```

LYVE1D	
0	93
1	28

TFF1D	
0	117
1	8