# Mercurio

Franco Quintanilla

2022-09-05

Importamos los datos

```
df = read.csv("/Users/francoquintanilla/Documents/R/mercurio.csv",
row.names=1)
head(df)
```

```
##             X2    X3  X4   X5    X6   X7 X8   X9  X10  X11 X12
## 1    Alligator   5.9 6.1  3.0   0.7 1.23  5 0.85 1.43 1.53   1
## 2        Annie   3.5 5.1  1.9   3.2 1.33  7 0.92 1.90 1.33   0
## 3       Apopka 116.0 9.1 44.1 128.3 0.04  6 0.04 0.06 0.04   0
## 4 Blue Cypress  39.4 6.9 16.4   3.5 0.44 12 0.13 0.84 0.44   0
## 5        Brick   2.5 4.6  2.9   1.8 1.20 12 0.69 1.50 1.33   1
## 6       Bryant  19.6 7.3  4.5  44.1 0.27 14 0.04 0.48 0.25   1
```

En donde el nombre de las variables son las siguientes:

- X1 = número de identificación
- X2 = nombre del lago
- X3 = alcalinidad (mg/l de carbonato de calcio)
- X4 = PH
- X5 = calcio (mg/l)
- X6 = clorofila (mg/l)
- X7 = concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago
- X8 = número de peces estudiados en el lago
- X9 = mínimo de la concentración de mercurio en cada grupo de peces
- X10 = máximo de la concentración de mercurio en cada grupo de peces
- X11 = estimación (mediante regresión) de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)
- X12 = indicador de la edad de los peces (0: jóvenes; 1: maduros)

## Variables Cuantitativas

### Medidas de tendencia central y Medidas de dispersión.

Para las medidas de tendencia de cada una de estas variables, vamos a sacar el promedio, la mediana, y la moda. Para las medidas de dispersión vamos a calcular la varianza y la desviación estándar.

**Para la moda, vamos a definir la siguiente función, ya que R no cuenta con la función de la moda que a nosotros nos interesa.**

```
moda <- function(x)
  {
   return(as.numeric(names(which.max(table(x)))))
  }
```

- Alcalinidad (mg/l de carbonato de calcio)

```
m_X3 = mean(df$X3)
cat("El promedio de la Alcalinidad es de:", m_X3, "\n")

## El promedio de la Alcalinidad es de: 37.53019

# Median
med_X3 = median(df$X3)
cat("La mediana de la Alcalinidad es de:", med_X3, "\n")

## La mediana de la Alcalinidad es de: 19.6

# Mode
moda_X3 = moda(df$X3)
cat("La moda de la Alcalinidad es de:", moda_X3, "\n")

## La moda de la Alcalinidad es de: 17.3

# Variance
v_X3 = var(df$X3)
cat("La varianza de la Alcalinidad es de:", v_X3, "\n")

## La varianza de la Alcalinidad es de: 1459.509

# Standard deviation
sd_X3 = sd(df$X3)
cat("La desviación estándar de la Alcalinidad es de:", sd_X3, "\n")

## La desviación estándar de la Alcalinidad es de: 38.20353
```

- PH

```
m_X4 = mean(df$X4)
cat("El promedio del PH es de:", m_X4, "\n")

## El promedio del PH es de: 6.590566

# Median
med_X4 = median(df$X4)
cat("La mediana del PH es de:", med_X4, "\n")

## La mediana del PH es de: 6.8

# Mode
moda_X4 = moda(df$X4)
cat("La moda del PH es de:", moda_X4, "\n")
```

```
## La moda del PH es de: 5.8

# Variance
v_X4 = var(df$X4)
cat("La varianza del PH es de:", v_X4, "\n")

## La varianza del PH es de: 1.660102

# Standard deviation
sd_X4 = sd(df$X4)
cat("La desviación estándar del PH es de:", sd_X4, "\n")

## La desviación estándar del PH es de: 1.288449
```

- Calcio (mg/l)

```
m_X5 = mean(df$X5)
cat("El promedio del Calcio es de:", m_X5, "\n")

## El promedio del Calcio es de: 22.20189

# Median
med_X5 = median(df$X5)
cat("La mediana del Calcio es de:", med_X5, "\n")

## La mediana del Calcio es de: 12.6

# Mode
moda_X5 = moda(df$X5)
cat("La moda del Calcio es de:", moda_X5, "\n")

## La moda del Calcio es de: 3

# Variance
v_X5 = var(df$X5)
cat("La varianza del Calcio es de:", v_X5, "\n")

## La varianza del Calcio es de: 621.6333

# Standard deviation
sd_X5 = sd(df$X5)
cat("La desviación estándar del Calcio es de:", sd_X5, "\n")

## La desviación estándar del Calcio es de: 24.93257
```

- Clorofila (mg/l)

```
m_X6 = mean(df$X6)
cat("El promedio de la Clorofila es de:", m_X6, "\n")

## El promedio de la Clorofila es de: 23.11698

# Median
med_X6 = median(df$X6)
cat("La mediana de la Clorofila es de:", med_X6, "\n")
```

```
## La mediana de la Clorofila es de: 12.8

# Mode
moda_X6 = moda(df$X6)
cat("La moda de la Clorofila es de:", moda_X6, "\n")

## La moda de la Clorofila es de: 1.6

# Variance
v_X6 = var(df$X6)
cat("La varianza de la Clorofila es de:", v_X6, "\n")

## La varianza de la Clorofila es de: 949.6457

# Standard deviation
sd_X6 = sd(df$X6)
cat("La desviación estándar de la Clorofila es de:", sd_X6, "\n")

## La desviación estándar de la Clorofila es de: 30.81632
```

- Concentración media de mercurio (parte por millón) en el tejido muscular del grupo de peces estudiados en cada lago

```
m_X7 = mean(df$X7)
cat("El promedio de la Concentración de mercurio es de:", m_X7, "\n")

## El promedio de la Concentración de mercurio es de: 0.5271698

# Median
med_X7 = median(df$X7)
cat("La mediana de la Concentración de mercurio es de:", med_X7, "\n")

## La mediana de la Concentración de mercurio es de: 0.48

# Mode
moda_X7 = moda(df$X7)
cat("La moda de la Concentración de mercurio es de:", moda_X7, "\n")

## La moda de la Concentración de mercurio es de: 0.34

# Variance
v_X7 = var(df$X7)
cat("La varianza de la Concentración de mercurio es de:", v_X7, "\n")

## La varianza de la Concentración de mercurio es de: 0.1163053

# Standard deviation
sd_X7 = sd(df$X7)
cat("La desviación estándar de la Concentración de mercurio es de:",
sd_X7, "\n")

## La desviación estándar de la Concentración de mercurio es de:
0.3410356
```

- Número de peces estudiados en el lago

```
m_X8 = mean(df$X8)
cat("El promedio del Número de peces estudiados es de:", m_X8, "\n")

## El promedio del Número de peces estudiados es de: 13.0566

# Median
med_X8 = median(df$X8)
cat("La mediana del Número de peces estudiados es de:", med_X8, "\n")

## La mediana del Número de peces estudiados es de: 12

# Mode
moda_X8 = moda(df$X8)
cat("La moda del Número de peces estudiados es de:", moda_X8, "\n")

## La moda del Número de peces estudiados es de: 12

# Variance
v_X8 = var(df$X8)
cat("La varianza del Número de peces estudiados es de:", v_X8, "\n")

## La varianza del Número de peces estudiados es de: 73.2852

# Standard deviation
sd_X8 = sd(df$X8)
cat("La desviación estándar del Número de peces estudiados es de:",
sd_X8, "\n")

## La desviación estándar del Número de peces estudiados es de: 8.560677
```

- Mínimo de la concentración de mercurio en cada grupo de peces

```
m_X9 = mean(df$X9)
cat("El promedio del Mínimo de la concentración de mercurios es de:",
m_X9, "\n")

## El promedio del Mínimo de la concentración de mercurios es de:
0.2798113

# Median
med_X9 = median(df$X9)
cat("La mediana del Mínimo de la concentración de mercurio es de:",
med_X9, "\n")

## La mediana del Mínimo de la concentración de mercurio es de: 0.25

# Mode
moda_X9 = moda(df$X9)
cat("La moda del Mínimo de la concentración de mercurio es de:", moda_X9,
"\n")

## La moda del Mínimo de la concentración de mercurio es de: 0.04
```

```r
# Variance
v_X9 = var(df$X9)
cat("La varianza del Mínimo de la concentración de mercurio es de:",
v_X9, "\n")
```

```
## La varianza del Mínimo de la concentración de mercurio es de:
0.05125958
```

```r
# Standard deviation
sd_X9 = sd(df$X9)
cat("La desviación estándar del Mínimo de la concentración de mercurio es
de:", sd_X9, "\n")
```

```
## La desviación estándar del Mínimo de la concentración de mercurio es
de: 0.2264058
```

- Máximo de la concentración de mercurio en cada grupo de peces

```r
m_X10 = mean(df$X10)
cat("El promedio del Máximo de la concentración de mercurios es de:",
m_X10, "\n")
```

```
## El promedio del Máximo de la concentración de mercurios es de:
0.8745283
```

```r
# Median
med_X10 = median(df$X10)
cat("La mediana del Máximo de la concentración de mercurio es de:",
med_X10, "\n")
```

```
## La mediana del Máximo de la concentración de mercurio es de: 0.84
```

```r
# Mode
moda_X10 = moda(df$X10)
cat("La moda del Máximo de la concentración de mercurio es de:",
moda_X10, "\n")
```

```
## La moda del Máximo de la concentración de mercurio es de: 0.06
```

```r
# Variance
v_X10 = var(df$X10)
cat("La varianza del Máximo de la concentración de mercurio es de:",
v_X10, "\n")
```

```
## La varianza del Máximo de la concentración de mercurio es de:
0.2725329
```

```r
# Standard deviation
sd_X10 = sd(df$X10)
cat("La desviación estándar del Máximo de la concentración de mercurio es
de:", sd_X10, "\n")
```

```
## La desviación estándar del Máximo de la concentración de mercurio es
de: 0.5220469
```

- Estimación de la concentración de mercurio en el pez de 3 años (o promedio de mercurio cuando la edad no está disponible)

```
m_X11 = mean(df$X11)
cat("El promedio de la Estimación de la concentración de mercurio es
de:", m_X11, "\n")
```

```
## El promedio de la Estimación de la concentración de mercurio es de:
0.5132075
```

```
# Median
med_X11 = median(df$X11)
cat("La mediana de la Estimación de la concentración de mercurio es de:",
med_X11, "\n")
```

```
## La mediana de la Estimación de la concentración de mercurio es de:
0.45
```

```
# Mode
moda_X11 = moda(df$X11)
cat("La moda de la Estimación de la concentración de mercurio es de:",
moda_X11, "\n")
```

```
## La moda de la Estimación de la concentración de mercurio es de: 0.16
```

```
# Variance
v_X11 = var(df$X11)
cat("La varianza de la Estimación de la concentración de mercurio es
de:", v_X11, "\n")
```

```
## La varianza de la Estimación de la concentración de mercurio es de:
0.1147376
```

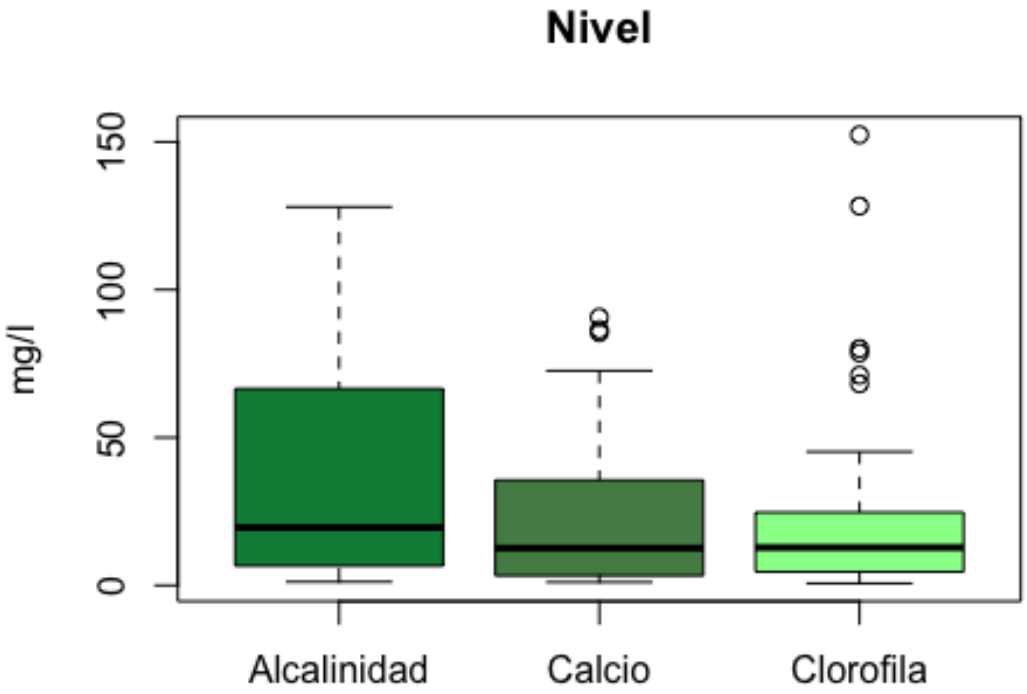
```
# Standard deviation
sd_X11 = sd(df$X11)
cat("La desviación estándar de la Estimación de la concentración de
mercurio es de:", sd_X11, "\n")
```

```
## La desviación estándar de la Estimación de la concentración de
mercurio es de: 0.3387294
```

## Visualización de las medidas de tendencia central y de dispersión

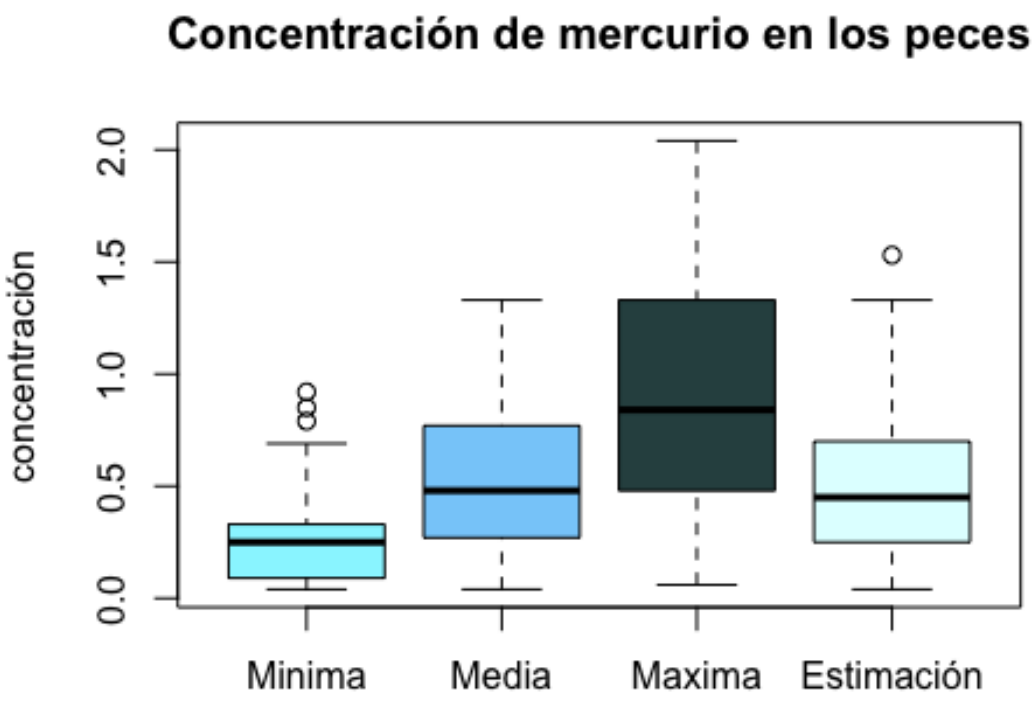
Primero hacemos el boxplot juntando los datos que tienen la misma unidad $\frac{mg}{l}$.

```
boxplot(df$X3, df$X5, df$X6, main="Nivel", ylab="mg/l", col=c("#008B45",
"#548B54", "#98FB98"), names=c("Alcalinidad", "Calcio", "Clorofila"))
```



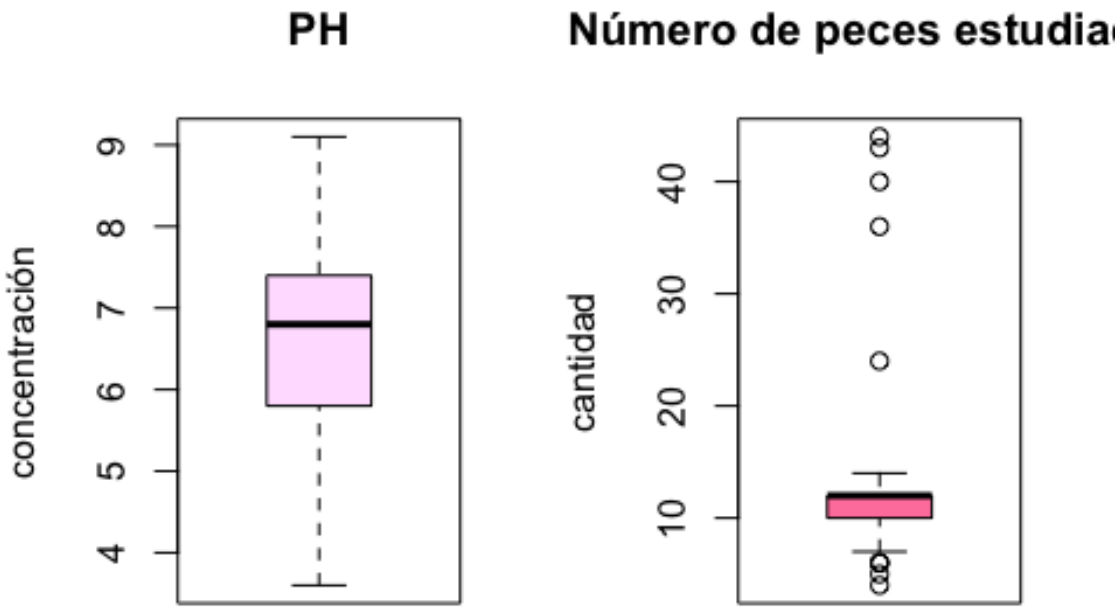

Como podemos ver, tanto el calcio como la clorofila tienen outliers, por lo que vamos a limpiar esos datos más adelante en base a los quantiles.

```
boxplot(df$X9, df$X7, df$X10, df$X11, main="Concentración de mercurio en los peces", ylab="concentración", col=c("#98F5FF", "#87CEFA", "#2F4F4F", "#E0FFFF"), names=c("Minima", "Media", "Maxima", "Estimación"))
```




De igual manera, en la concentración de mercurio vamos a hacer una limpia de los outliers, para tener un modelo más limpio y consiso.

```
par(mfrow=c(1,2))
boxplot(df$X4, main="PH", ylab="concentración", col="thistle1")
boxplot(df$X8, main="Número de peces estudiados", ylab="cantidad",
col="palevioletred1")
```




En el caso del PH vemos que no hay outliers, y todos los datos se encuentran dentro de los quantiles. Por otra parte, aunque en el número de peces estudiados tenemos varios outliers, no los vamos a eliminar, en este caso solo nos sirve para ver en donde está centrada la media de la cantidad de peces estudiados en cada lago.

## Variables cualitativas

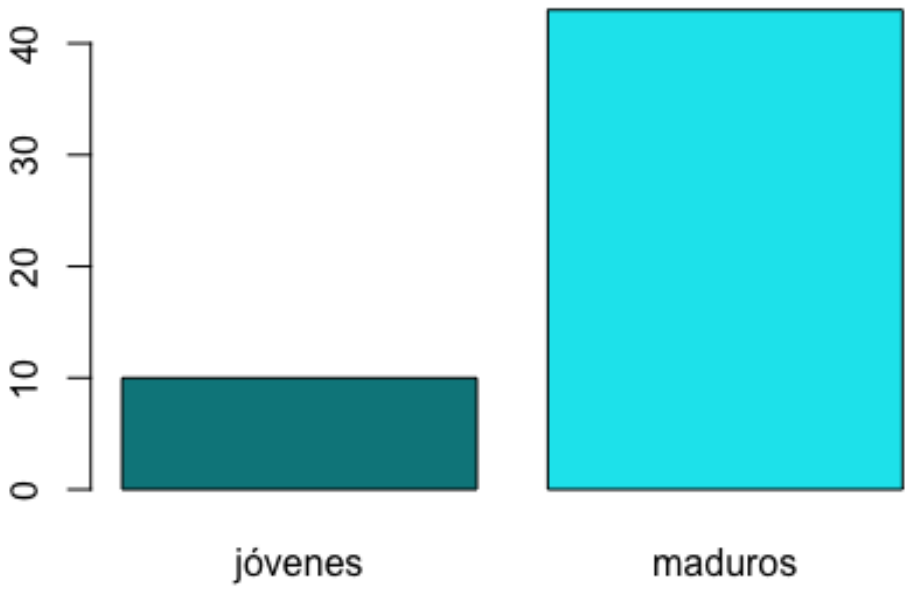
- Indicador de la edad de los peces (0: jóvenes; 1: maduros)

```
m_X12 = mean(df$X12)
cat("El promedio de la edad de los peces es de:", m_X12, "\n")

## El promedio de la edad de los peces es de: 0.8113208

# Median
med_X12 = median(df$X12)
cat("La mediana de la edad de los peces es de:", med_X12, "\n")

## La mediana de la edad de los peces es de: 1

# Mode
moda_X12 = moda(df$X12)
cat("La moda de la edad de los peces es de:", moda_X12, "\n")
```

```
## La moda de la edad de los peces es de: 1

# Variance
v_X12 = var(df$X12)
cat("La varianza de la edad de los peces es de:", v_X12, "\n")

## La varianza de la edad de los peces es de: 0.1560232

# Standard deviation
sd_X12 = sd(df$X12)
cat("La desviación estándar de la edad de los peces es de:", sd_X12,
"\n")

## La desviación estándar de la edad de los peces es de: 0.3949977
```

## Visualización

```
barplot(table(df$X12), col=c("turquoise4", "turquoise2"),
names=c("jóvenes", "maduros"))
```



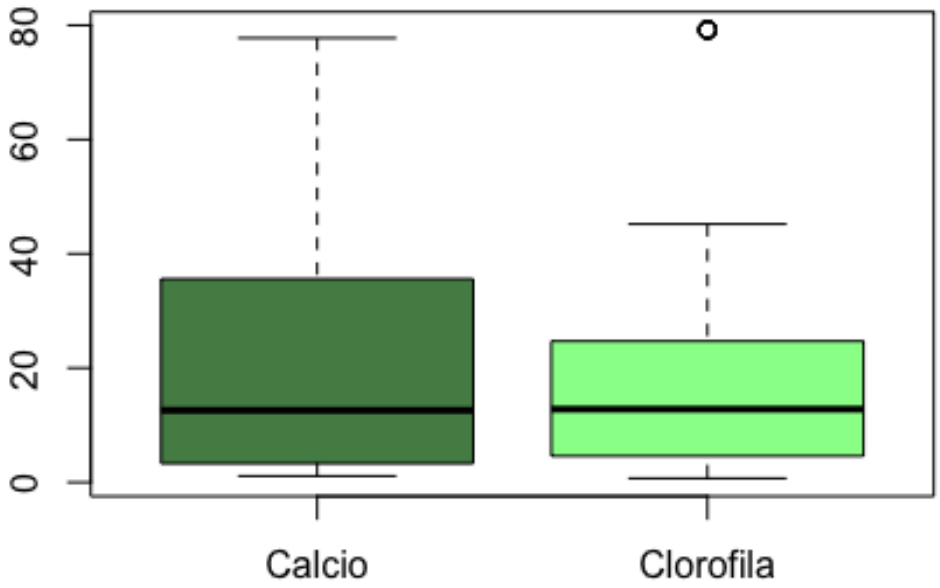
## Limpieza de datos

Para la limpieza de los datos, vamos a hacer una función para la eliminación de los outliers.

```
f_outliers <- function(x, removeNA = TRUE)
  {
  qrts <- quantile(x, probs = c(0.25, 0.75), na.rm = removeNA)
  caps <- quantile(x, probs = c(.05, .95), na.rm = removeNA)
  iqr <- qrts[2] - qrts[1]
  x[x<qrts[1] - 1.5*iqr] <- caps[1]
  x[x>qrts[2] + 1.5*iqr] <- caps[2]
  x
  }
```

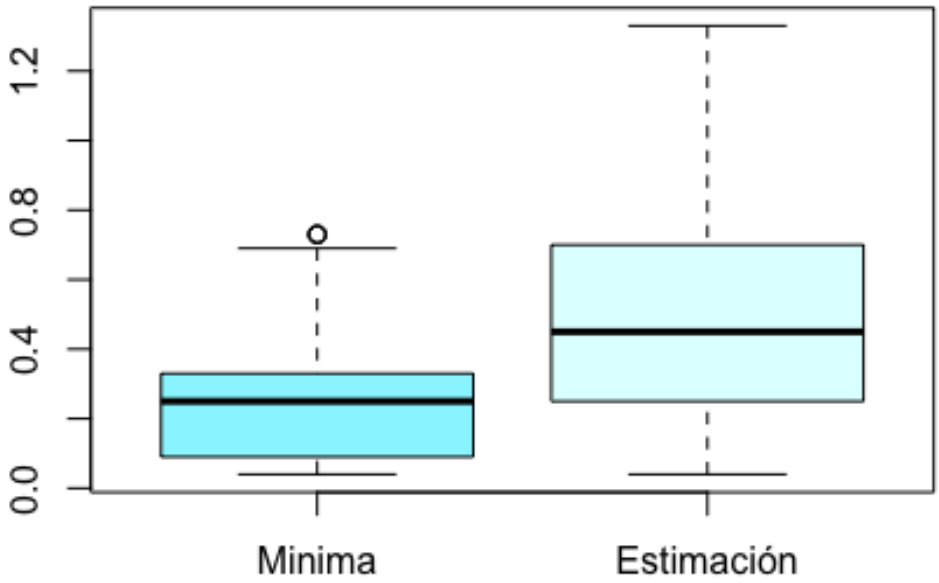
## Calcio y Clorofila

```
x5 <- f_outliers(df$X5)
x6 <- f_outliers(df$X6)
boxplot(x5, x6, col=c("#548B54", "#98FB98"), names=c("Calcio",
"Clorofilla"))
```



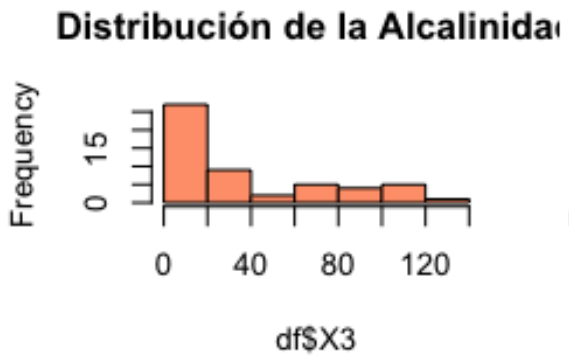
## Mínima y Estimación

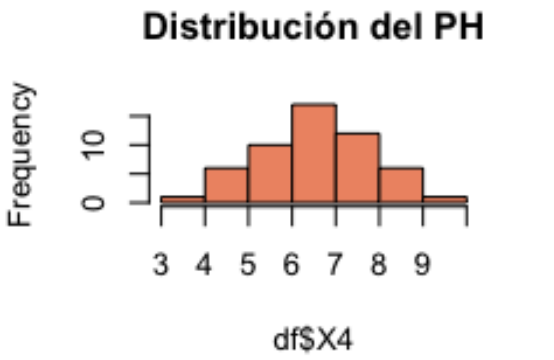
```
x9 <- f_outliers(df$X9)
x11 <- f_outliers(df$X11)
boxplot(x9, x11, col=c("#98F5FF", "#E0FFFF"), names=c("Minima",
"Estimación"))
```

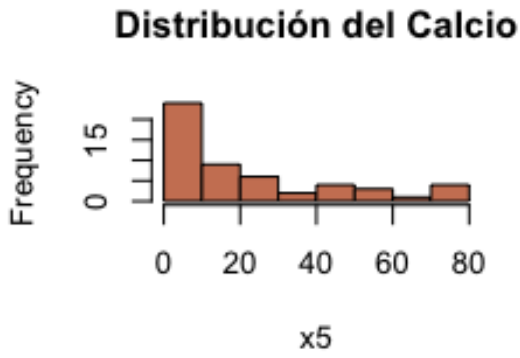


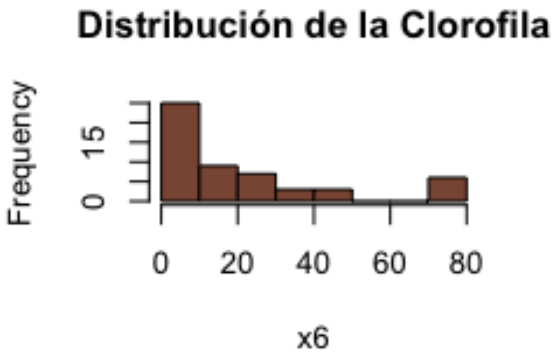

Con los datos limpios, vamos a hacer la visualización de los histogramas

```
par(mfrow=c(2,2))

hist(df$X3, col="#FFA07A", main="Distribución de la Alcalinidad")
hist(df$X4, col="#EE9572", main="Distribución del PH")
hist(x5, col="#CD8162", main="Distribución del Calcio")
hist(x6, col="#8B5742", main="Distribución de la Clorofila")
```

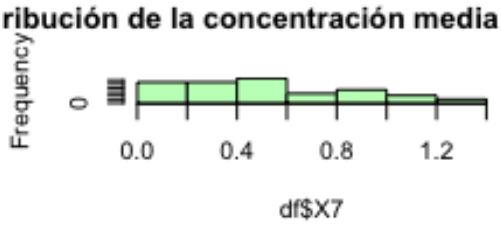


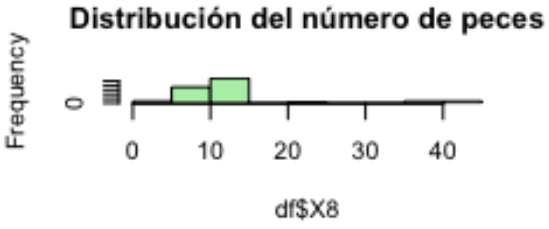



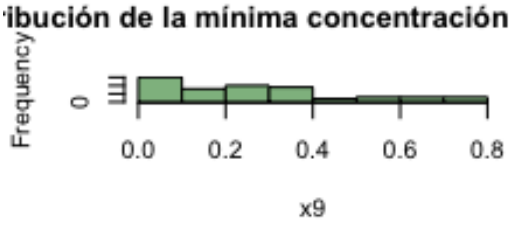



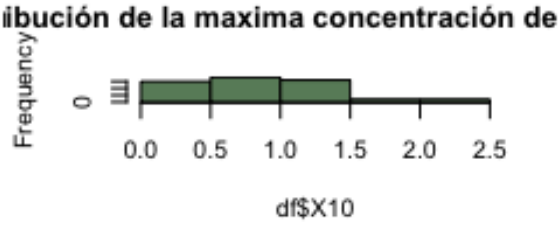



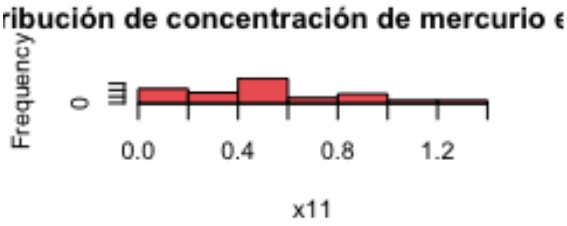

```
par(mfrow=c(3,2))

hist(df$X7, col="#C1FFC1", main="Distribución de la concentración media
de mercurio")
hist(df$X8, col="#B4EEB4", main="Distribución del número de peces")
hist(x9, col="#8FBC8F", main="Distribución de la mínima concentración de
mercurio")
hist(df$X10, col="#698B69", main="Distribución de la maxima concentración
de mercurio")
hist(x11, col="#EE6363", main="Distribución de concentración de mercurio
estimada")
```




Después de esto, vamos a hacer el análisis de los datos mediante:

- Regresión lineal simple
- ANOVA

Escogimos estos 2 análisis de datos , ya que estamos haciendo nuestra investigación en base a la siguiente pregunta:

- ¿Cuáles son los principales factores que influyen en el nivel de contaminación por mercurio en los peces de los lagos de Florida?

Por lo que decidimos hacer los siguientes análisis.

## Regresión lineal simple

Hacemos la matriz de correlación

```
cm_ph = subset(df, select = c(X3, X4, X5, X6, X7, X8, X9, X10, X11, X12))
cor(cm_ph)
```

```
##               X3          X4           X5          X6          X7
X8
## X3   1.00000000  0.71916568  0.832604192  0.47753085 -0.59389671
```

```
0.01029074
## X4   0.71916568  1.00000000  0.577132721  0.60848276 -0.57540012
-0.01860607
## X5   0.83260419  0.57713272  1.000000000  0.40991385 -0.40067958
-0.08937901
## X6   0.47753085  0.60848276  0.409913846  1.00000000 -0.49137481
-0.01182027
## X7  -0.59389671 -0.57540012 -0.400679584 -0.49137481  1.00000000
0.07903426
## X8   0.01029074 -0.01860607 -0.089379013 -0.01182027  0.07903426
1.00000000
## X9  -0.52535654 -0.54196524 -0.332476229 -0.40045856  0.92720506
-0.08165278
## X10 -0.60479558 -0.55181523 -0.407916635 -0.48497215  0.91586397
0.16109174
## X11 -0.62795845 -0.61284905 -0.464409465 -0.50644193  0.95921481
0.02580046
## X12 -0.09493882  0.03800021 -0.002111124 -0.28300234  0.10873896
0.20795617
##              X9         X10         X11          X12
## X3  -0.52535654 -0.60479558 -0.62795845 -0.094938825
## X4  -0.54196524 -0.55181523 -0.61284905  0.038000214
## X5  -0.33247623 -0.40791663 -0.46440947 -0.002111124
## X6  -0.40045856 -0.48497215 -0.50644193 -0.283002338
## X7   0.92720506  0.91586397  0.95921481  0.108738958
## X8  -0.08165278  0.16109174  0.02580046  0.207956171
## X9   1.00000000  0.76535319  0.91908939  0.100661967
## X10  0.76535319  1.00000000  0.85975810  0.093752072
## X11  0.91908939  0.85975810  1.00000000  0.089411267
## X12  0.10066197  0.09375207  0.08941127  1.000000000
```

Con esta matriz, nos vamos a dar cuenta de las variables que más correlación tienen con nuestra variable objetivo.

También, vemos que hay variables de más, así que solo vamos a usar, X3, X4, X5, X6, y X9

```
cm_ph = subset(df, select = c(X3, X4, X5, X6, X9))
cor(cm_ph)
```

```
##            X3         X4         X5         X6         X9
## X3  1.0000000  0.7191657  0.8326042  0.4775308 -0.5253565
## X4  0.7191657  1.0000000  0.5771327  0.6084828 -0.5419652
## X5  0.8326042  0.5771327  1.0000000  0.4099138 -0.3324762
## X6  0.4775308  0.6084828  0.4099138  1.0000000 -0.4004586
## X9 -0.5253565 -0.5419652 -0.3324762 -0.4004586  1.0000000
```

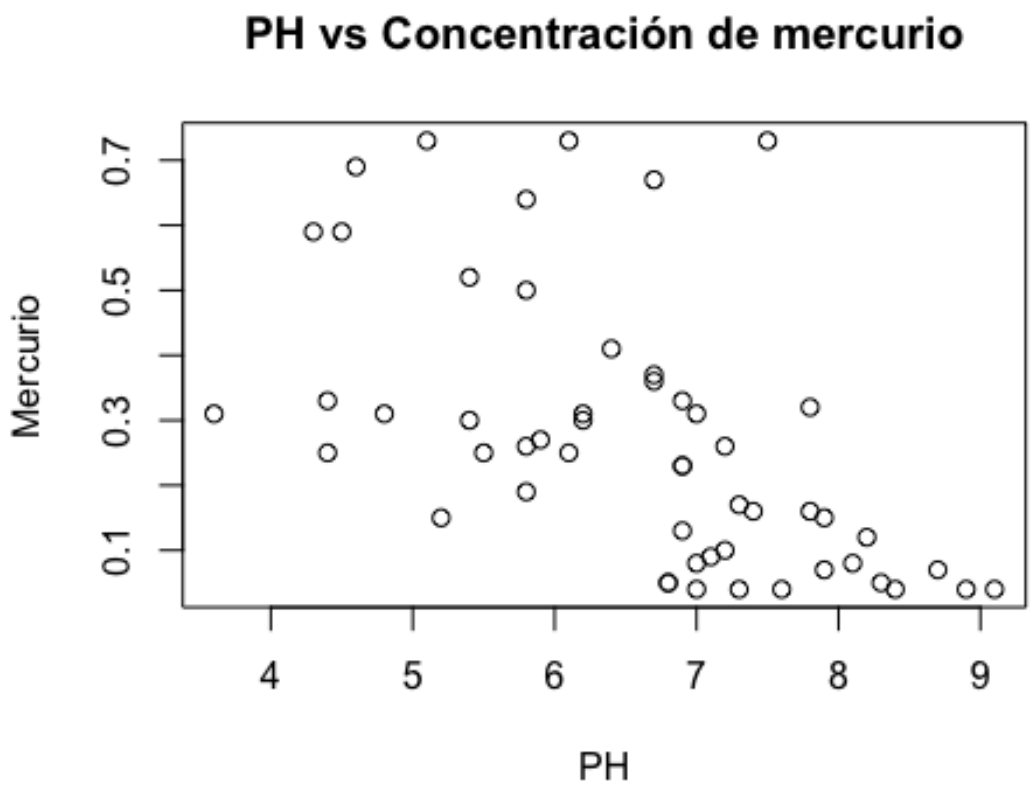
## Modelo

Definimos nuestras variables

```
# PH
x = df$X4

# Mínimo de la concentración de mercurio
y = x9
```

Visualizamos la tendencia

```
plot(x, y, main="PH vs Concentración de mercurio", xlab="PH",
ylab="Mercurio")
```




Hacemos el fit para el modelo lineal en base al PH

```
model = lm(y ~ x)
summary(model)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25063 -0.12306 -0.02059  0.09723  0.54075
##
```

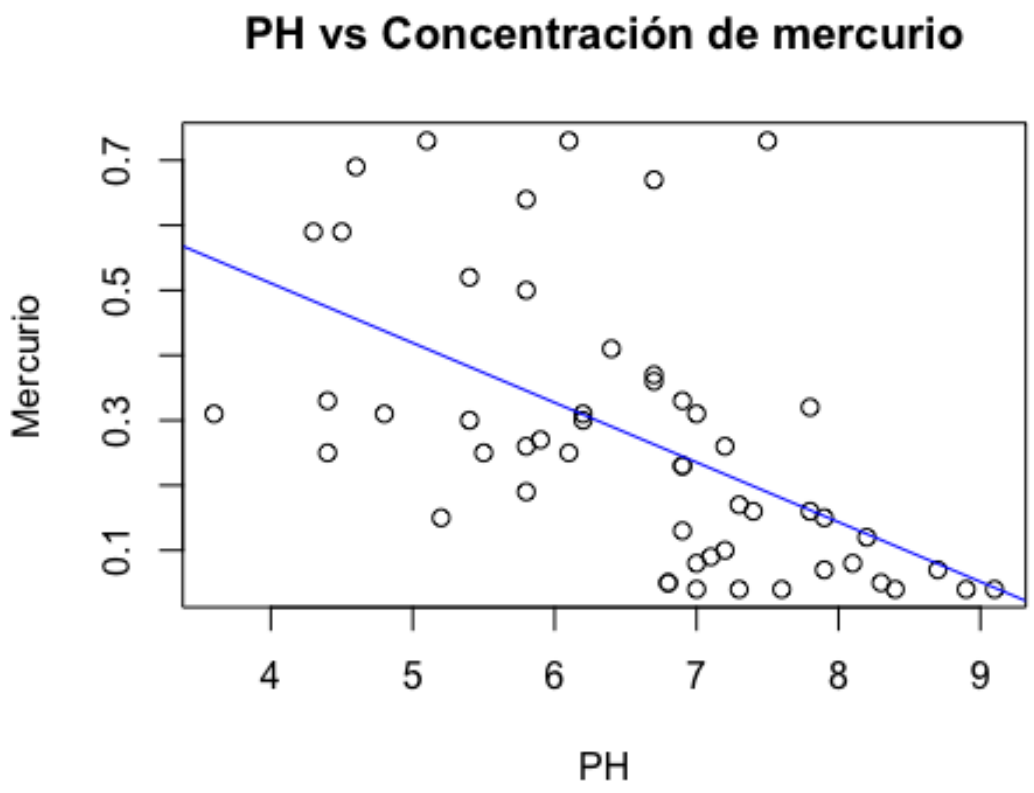
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.87853    0.12582   6.982 5.80e-09 ***
## x           -0.09190    0.01874  -4.903 9.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1741 on 51 degrees of freedom
## Multiple R-squared:  0.3204, Adjusted R-squared:  0.3071
## F-statistic: 24.04 on 1 and 51 DF,  p-value: 9.993e-06
```

Nuestra ecuación que define el valor esperado de Y, nos queda que:

$$E(Y) = 0.878 - 0.0919x,$$

Con los coeficientes del modelo de regresión lineal, lo ploteamos para ver el comportamiento con los datos.

```
par(mfrow=c(1,1))
plot(x, y, main="PH vs Concentración de mercurio", xlab="PH",
ylab="Mercurio")
abline(model, col="blue")
```



También, podemos obtener los intervalos de confianza de nuestro modelo con un nivel de confianza del 95 %.

```
confint(model, level=0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept)  0.6259342  1.13111675
## x           -0.1295306 -0.05427618
```

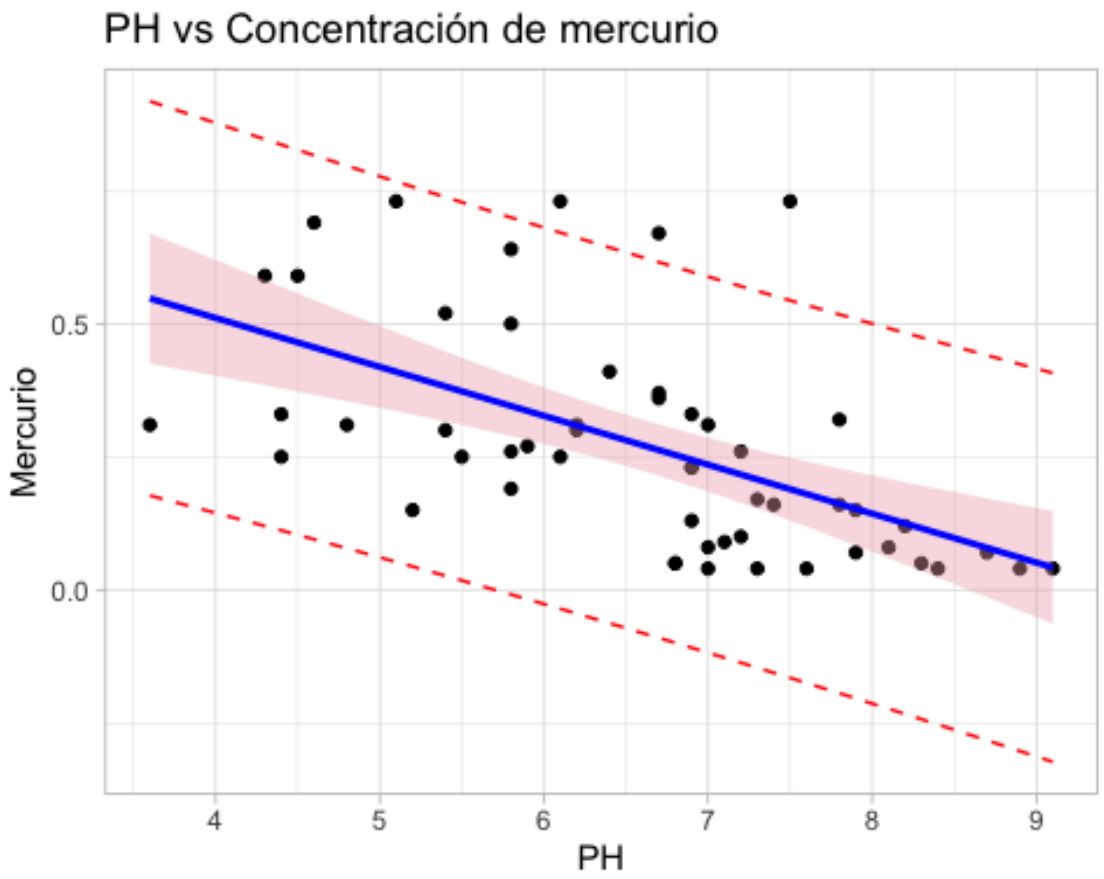
Donde para visualizar mejor los intervalos, los podemos agregar a nuestro gráfico.

```
Yp = predict(object=model, interval="prediction", level=0.95)
```

```
## Warning in predict.lm(object = model, interval = "prediction", level =
0.95): predictions on current data refer to _future_ responses
```

```
datos1 = cbind(df, Yp)
```

```
library(ggplot2)
ggplot(datos1, aes(x=x, y=y)) + geom_point() + geom_line(aes(y=lwr),
color="red", linetype="dashed") +
geom_line(aes(y=upr), color="red", linetype="dashed") +
geom_smooth(method=lm, formula=y~x, se=TRUE, level=0.95, col='blue',
fill='pink2') +
theme_light() + labs(title="PH vs Concentración de mercurio", x="PH",
y="Mercurio")
```




## Análisis de los residuos

### Normalidad de los residuos

Para la prueba de normalidad de residuos, hacemos una prueba de hipótesis con el shapiro test.

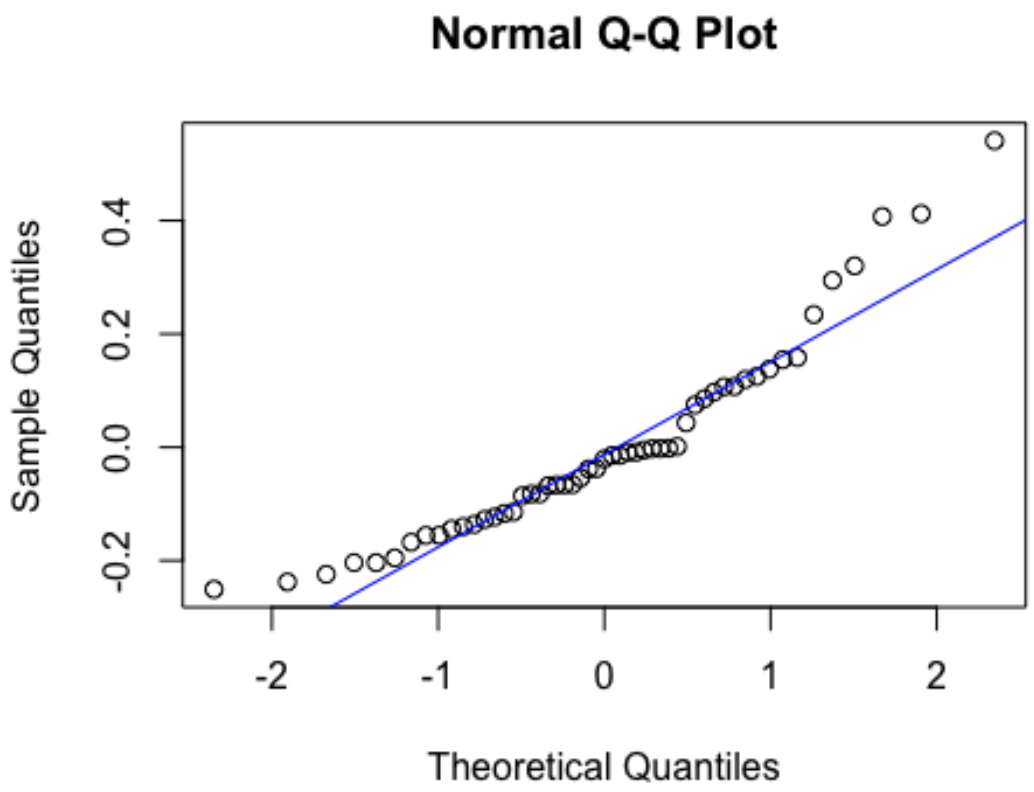
```
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.91803, p-value = 0.001413
```

Pero como es difícil de interpretar lo, agregaremos sus respectivas gráficas.

**QQ-Plot**

```
qqnorm(model$residuals)
qqline(model$residuals, col="blue")
```

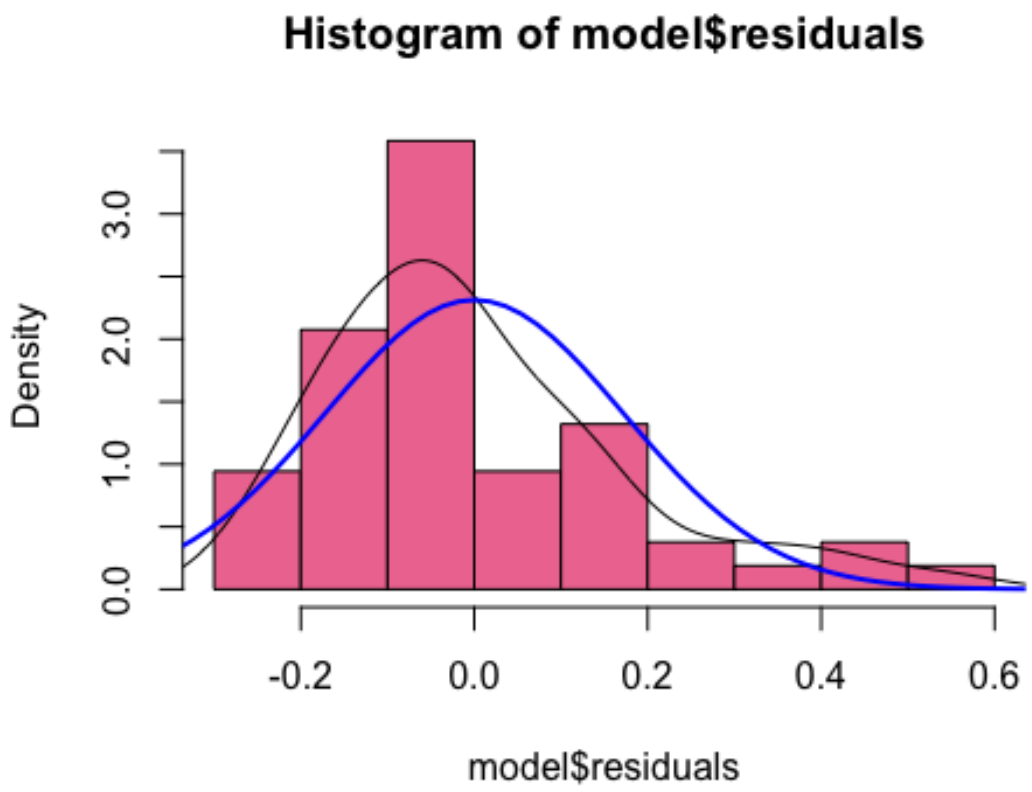



Como podemos ver, el modelo se comporta como una distribución con colas gruesas es decir, que tiene baja curtosis, del tipo platicúrtica, pero que también tiene una asimetría positiva es decir que su sesgo va a la derecha.

Otra manera de visualizar los datos, es con la ayuda de un histograma, lo cual es lo siguiente que vamos a hacer.

**Histograma**

```
# Histograma
hist(model$residuals, freq=FALSE, breaks=10, col="palevioletred2")

# Comportamiento de los residuos
lines(density(model$residuals), col="black")

# Comportamiento de los residuos estandarizados.
curve(dnorm(x, mean=mean(model$residuals), sd=sd(model$residuals)),
from=-0.9,to=0.9, add=TRUE, col="blue",lwd=2)
```




Con el histograma podemos ver el comportamiento que concluimos con la ayuda de las pruebas de normalidad, que nuestro modelo cuenta con una curtosis leptocúrtica, y lo más visual es que tiene una asimetría positiva lo que nos dice que su sesgo va a la derecha.

### Verificación de media cero

```
t.test(model$residuals)
```

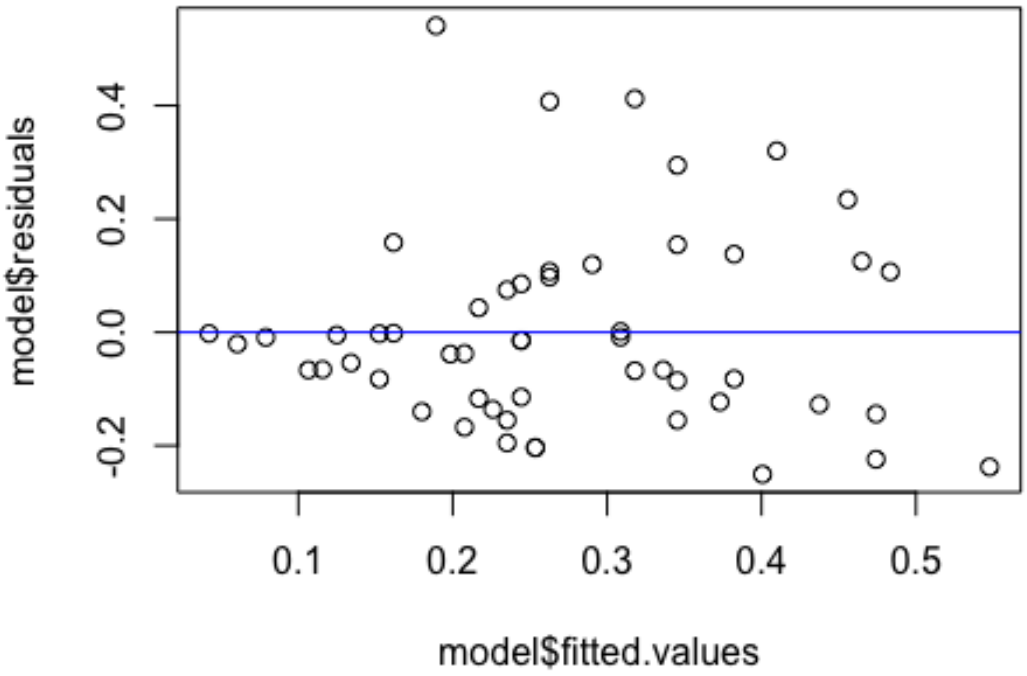
```
##
##  One Sample t-test
##
## data:  model$residuals
## t = 3.6434e-16, df = 52, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.04753501  0.04753501
## sample estimates:
##    mean of x
## 8.630766e-18
```

Desde aquí podemos ver que, por el valor de $t^* = 3.6434e-16$ a comparación de $t_0$, es que $t^*$ es muy cercano a 0, por lo que está dentro de los límites para poder aceptar la hipótesis de que los errores se comportan como una normal.

Por último, nos queda hacer el análisis de la homocedasticidad.

### Homocedasticidad

```
plot(model$fitted.values, model$residuals)
abline(h=0, col="blue")
```




Con este análisis de la homocedasticidad nos damos cuenta de una buena vez por todas que nuestro modelo es el adecuado para definir la comparación entre el PH y la concentración de mercurio, ya que se ve que el modelo cuenta con homocedasticidad, y sus residuos no cuentan con una tendencia específica.

## ANOVA

Definimos nuestras variables para el ANOVA.

```
# La edad de los peces
edad = df$X12
```

```r
# Mínimo de la concentración de mercurio
merc = x9
```

## Hacemos el ANOVA

```r
anova = aov(merc ~ edad)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## edad         1 0.0427 0.04266   0.974  0.328
## Residuals   51 2.2330 0.04378
```

### Análisis de cada tratamiento

```r
m = tapply(merc, edad, mean)
cat("Media de la concentración de mercurio por la edad de los peces:", m,
"\n")
```

```
## Media de la concentración de mercurio por la edad de los peces: 0.214
0.2865116
```

```r
s = tapply(merc, edad, sd)
cat("Desviación estándar de la concentración de mercurio por la edad de
los peces:", s, "\n")
```

```
## Desviación estándar de la concentración de mercurio por la edad de los
peces: 0.2148488 0.208028
```

```r
n = tapply(merc, edad, length)
cat("Tamaño de la muestra de la concentración de mercurio por la edad de
los peces:", n, "\n")
```

```
## Tamaño de la muestra de la concentración de mercurio por la edad de
los peces: 10 43
```

### Intervalos de confianza

```r
sm = s/sqrt(n)
E = abs(qt(0.025,n-1))*sm

In = m - E
cat("Intervalos de confianza inferiores:", In, "\n")
```

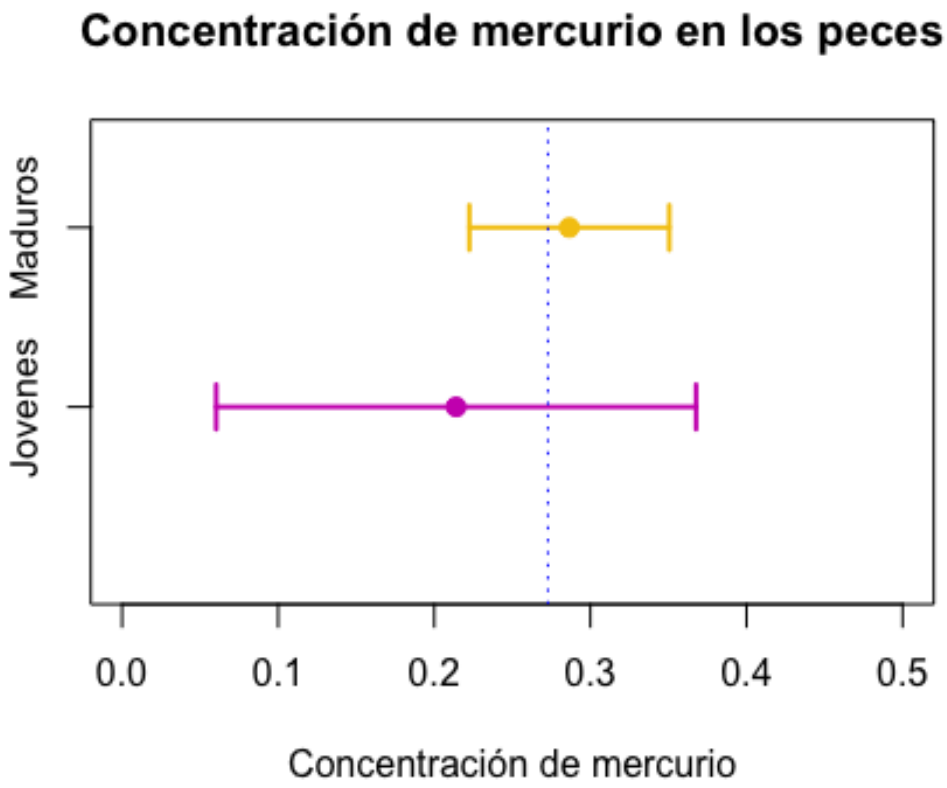
```
## Intervalos de confianza inferiores: 0.06030644 0.2224901
```

```r
Sup = m + E
cat("Intervalos de confianza superiores:", Sup, "\n")
```

```
## Intervalos de confianza superiores: 0.3676936 0.3505332
```

## Gráfico de los intervalos de confianza de cada grupo de edad de peces

```
plot(0, ylim=c(0,2.5), xlim=c(0,0.5), yaxt="n",
ylab="",xlab="Concentración de mercurio", main="Concentración de mercurio
en los peces")
axis(2,at=c(1:2),labels=c("Jovenes","Maduros"))

for(i in 1:2)
  {
arrows(In[i],i,Sup[i],i, angle=90, code=3, length = 0.1, lwd = 2,col=i+5)
points(m[i], i, pch=19, cex=1.1,col=i+5)
}

abline(v=mean(merc),lty=3,col="blue")
```

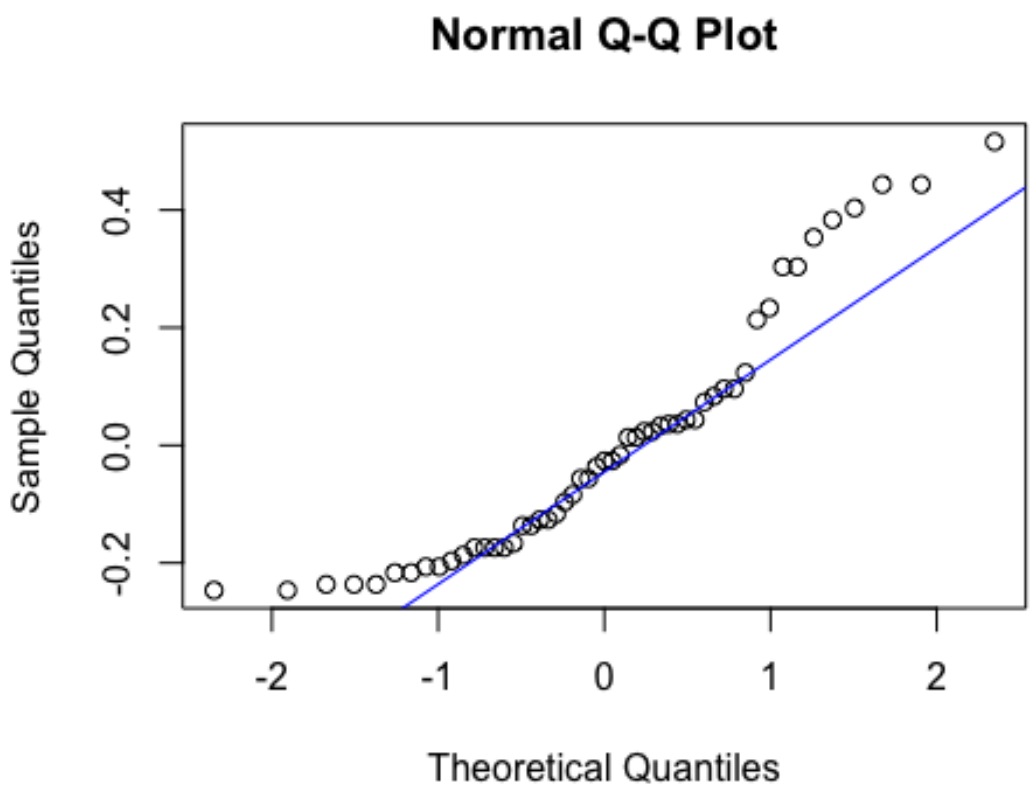



Desde aquí podemos observar que la concentración de mercurio por la edad de los peces no afecta, ya que una contiene a la otra en los intervalos de confianza, por lo que se podría decir que son iguales, donde además, ambas se encuentran dentro de la media poblacional.

## Verificación de supuestos

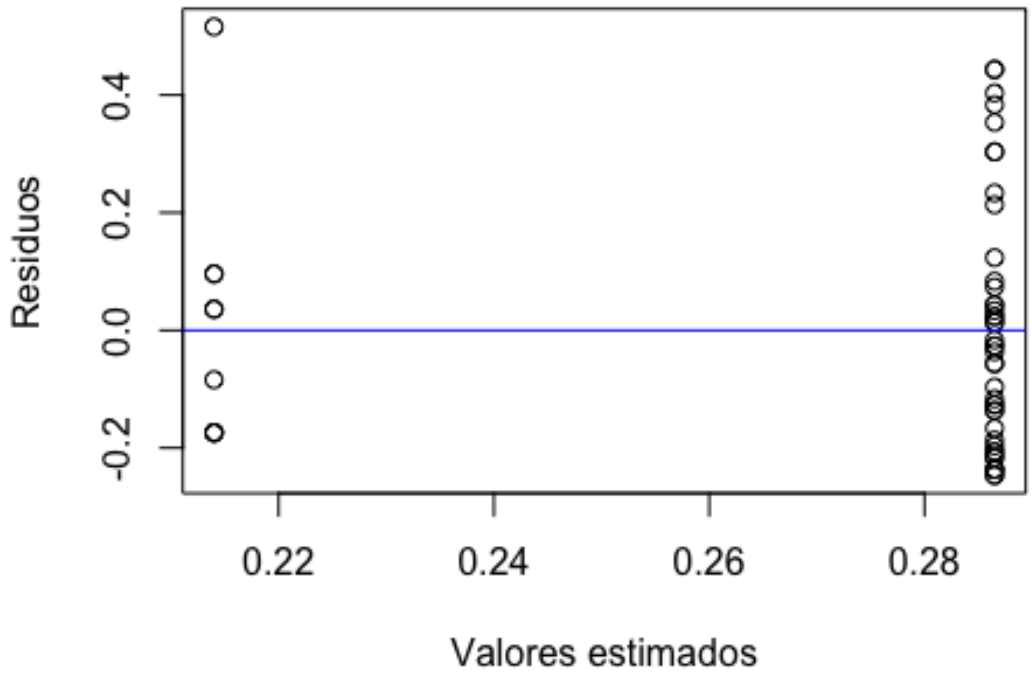
**Prueba de normalidad QQ plot**

```
qqnorm(anova$residuals)
qqline(anova$residuals, col="blue")
```



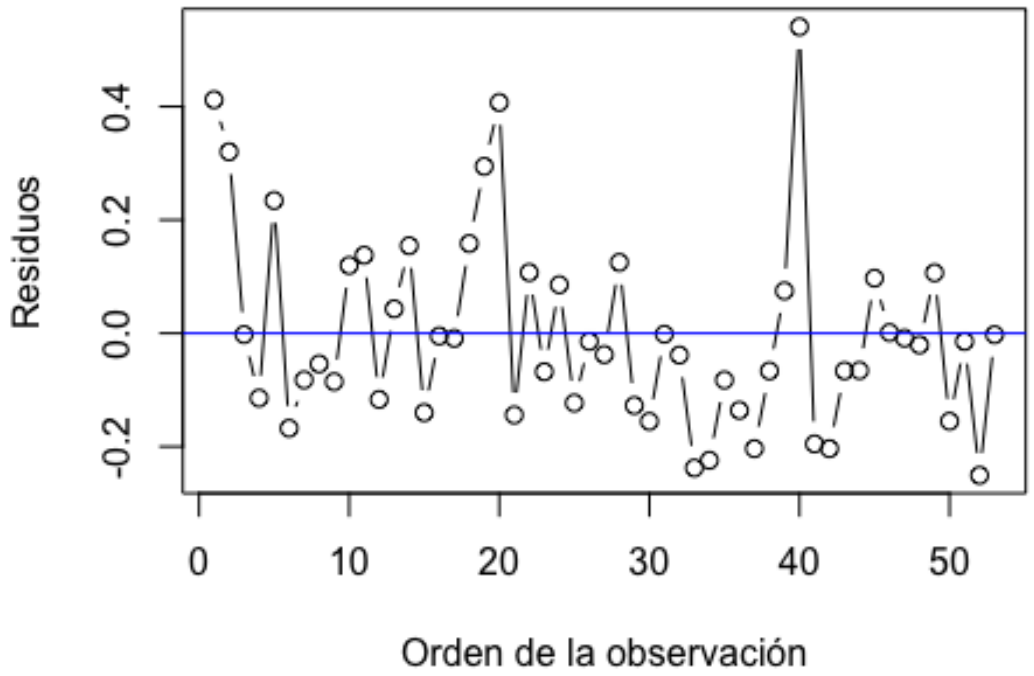

### Homocedastícidad

```
plot(anova$fitted.values, anova$residuals, ylab="Residuos", xlab="Valores
estimados")
abline(h=0, col="blue")
```




**Independencia**

```
plot(c(1:sum(n)), model$residuals, xlab="Orden de la observación",
ylab="Residuos", type = "b")
abline(h=0, col="blue")
```




Después de hacer el análisis de las varianzas, podemos observar que la edad de los peces, ya sean jóvenes o maduros, no afecta en la concentración de mercurio en los peces de los distintos lagos. Aunque pensemos que esto pueda estar influenciado ya que la cantidad de peces en la muestra favorece a los peces maduros, ya que son mucho más la cantidad que los peces jóvenes, pero aun con esta diferencia podemos observar que las concentraciones de mercurio son las mismas.

## Conclusiones

Después de hacer todo el análisis estadístico en base a la concentración de mercurio en los peces, podemos observar que:

- Gracias a la regresión lineal simple, podemos ver que cuando un lago tiene mayor PH, es decir, el agua es más alcalina, la concentración de mercurio en los peces va a disminuir, por el contrario si el agua es más ácida, entonces la concentración de mercurio en los peces va a ser mayor.
- Después de hacer el análisis de varianza en base a la edad de los peces y su concentración de mercurio, podemos responder otra de las preguntas, y es que NO habrá diferencia significativa entre la concentración de mercurio por la

edad de los peces, y esto lo podemos demostrar gracias a los intervalos de confianza.