En vista de los datos `Datos-mercurio.txt` ¿hay suficiente evidencia estadística para afirmar que el nivel medio de contaminación por mercurio en los dos ríos es diferente? Contrastar la hipótesis de igualdad de varianzas.
Indicar, en cada caso, las suposiciones previas necesarias para garantizar la validez de los procedimientos empleados.

Suponemos que $X =$ Nivel de contaminación por mercurio en un pez (de la especie *large mouth bass*) elegido al azar en el río Lumber e $Y =$ Nivel de contaminación por mercurio en un pez (de la misma especie) del río Wacamaw son v.a. independientes y siguen una distribución normal: $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$.
Contrastemos primero la hipótesis de igualdad de varianzas a nivel α:

$$\begin{aligned} H_0 &: \quad \sigma_1 = \sigma_2 \\ H_1 &: \quad \sigma_1 \neq \sigma_2. \end{aligned} \tag{1}$$

La región de rechazo es $R = \{s_1^2/s_2^2 \notin [F_{n_1-1;n_2-1;1-\alpha/2}, F_{n_1-1;n_2-1;\alpha/2}]\}$.

```
X = read.table('Datos-mercurio.txt')
ContHg = X$V5
Rio = X$V1
ContHgL = ContHg[Rio==0]
ContHgW = ContHg[Rio==1]
s2L = var(ContHgL)
s2W = var(ContHgW)
s2L/s2W
[1] 0.6119333
alpha = 0.1
n1 = length(ContHgL)
n2 = length(ContHgW)
c(qf(alpha/2,n1-1,n2-1),qf(alpha/2,n1-1,n2-1,lower.tail=F))
[1] 0.690974 1.430908
```

Por tanto, a nivel $\alpha = 0{,}1$ no podemos considerar las varianzas iguales.

```
alpha = 0.05
c(qf(alpha/2,n1-1,n2-1),qf(alpha/2,n1-1,n2-1,lower.tail=F))
[1] 0.6432225 1.5328961
```

A nivel $\alpha = 0{,}05$ tampoco.
Entonces la región de rechazo del contraste

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 \\ H_1 &: \quad \mu_1 \neq \mu_2 \end{aligned} \tag{2}$$

a nivel de significación α es

$$R = \left\{ |\bar{x} - \bar{y}| \geq t_{f;\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\},$$

donde $f = 169$ es el entero más próximo a

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} = 168{,}57. \tag{3}$$

Como $|\bar{x}-\bar{y}| = 0{,}198$ y $t_{169;0{,}025}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = 0{,}223$, no tenemos suficiente evidencia estadística para rechazar $H_0 : \mu_1 = \mu_2$.

Con R podemos hacer *t-tests* (contrastes en los que el estadístico del contraste sigue una distribución t) de la siguiente manera:

```
t.test(ContHg ~ Rio, alternative = "two.sided", mu = 0, paired = FALSE, var.equal
    = FALSE, conf.level = 0.95)
```

o equivalentemente

```
t.test(ContHgL, ContHgW, alternative = "two.sided", mu = 0, paired = FALSE, var.
    equal = FALSE, conf.level = 0.95)
```

Obtenemos como resultado

```
        Welch Two Sample t-test

data:  ContHgL and ContHgW
t = -1.7547, df = 168.57, p-value = 0.08114
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.42150363 0.02481087
sample estimates:
mean of x mean of y
 1.078082 1.276429
```

El valor `t` es el del estadístico del contraste

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}},$$

y `df` es el valor de la expresión (3). El intervalo de confianza es $\mathrm{IC}_{0,95}(\mu_1-\mu_2)$.

Con `t.test` también podemos hacer contrastes para una sola muestra (es decir, contrastes acerca de la media de una $N(\mu,\sigma)$ con σ desconocido). Por ejemplo, si quisiéramos contrastar $H_0 = \mu_1 \geq 1$ frente a $H_1 : \mu_1 < 1$ escribiríamos:

```
t.test(ContHgL, alternative = "less", mu = 1, conf.level = 0.95)
```

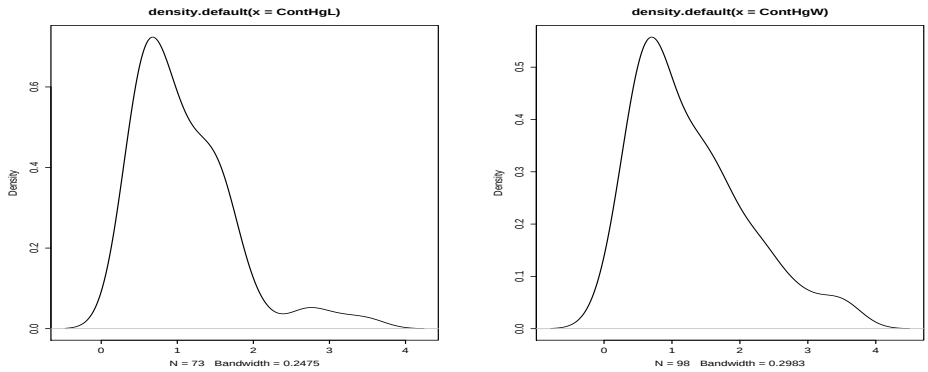
Y para hacer el contraste (1) de igualdad de varianzas

```
> var.test(ContHgL, ContHgW, ratio = 1, alternative = "two.sided", conf.level =
    0.95)

        F test to compare two variances

data:  ContHgL and ContHgW
F = 0.6119, num df = 72, denom df = 97, p-value = 0.0294
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3992008 0.9513555
sample estimates:
ratio of variances
         0.6119333
```

Otra posibilidad para hacer el contraste (2) sin suponer normalidad de X e Y (ver figura)



es utilizar que, por el TCL, $\bar{X} \overset{\text{aprox}}{\sim} N(\mu_1, \sigma_1/\sqrt{n_1})$ e $\bar{Y} \overset{\text{aprox}}{\sim} N(\mu_2, \sigma_2/\sqrt{n_2})$. Si X e Y son independientes entonces

$$\bar{X} - \bar{Y} \overset{\text{aprox}}{\sim} N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

y, por el Teorema de Slustky, si $H_0 : \mu_1 = \mu_2$ es cierta entonces

$$\bar{X} - \bar{Y} \overset{\text{aprox}}{\sim} N\left(0, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right).$$

A nivel $\alpha = 0{,}05$ no podemos rechazar la hipótesis nula (2) porque $|\bar{x} - \bar{y}| = 0{,}198 < z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 0{,}222$, pero sí podemos rechazar a nivel $\alpha = 0{,}1$.
Observemos que, como el tamaño muestral es grande, las regiones de rechazo suponiendo normalidad y utilizando la aproximación del TCL son prácticamente iguales.