Se mide el grado X de expresión de un gen en el tejido ovárico de 23 mujeres sanas[1], obteniéndose los siguientes datos:

0.51	0.52	0.62	0.67	0.67	0.70	0.76	0.76	0.79	0.81	0.81	0.84
0.89	0.94	1.01	1.09	1.15	1.15	1.16	1.27	1.35	1.37	2.63	

Se mide el grado de expresión del mismo gen en el tejido ovárico de 30 mujeres con cáncer de ovario, obteniéndose la siguiente muestra:

0.81	0.70	0.64	0.67	0.60	0.42	0.70	0.55	0.98	1.10
0.69	0.34	0.60	0.49	1.19	0.87	2.33	1.16	0.50	0.95
0.81	2.78	1.25	0.69	1.03	0.69	0.57	0.72	0.72	0.94

Utilizando el programa R, determinar si hay suficiente evidencia muestral de que el nivel esperado de expresión de ese gen es diferente en mujeres sanas y en pacientes con cáncer de ovario. ¿Se puede aceptar la igualdad de varianzas?

Solución: Sea Y el grado de expresión del gen en el tejido ovárico de una mujer con cáncer de ovario. Denotamos $\mu_1 = \mathbb{E}(X)$ y $\mu_2 = \mathbb{E}(Y)$. La información muestral de la que disponemos es

$$n_1 = 23 \quad \bar{x} = 0.9770 \quad s_1 = 0.4396$$

$$n_2 = 30 \quad \bar{y} = 0.8830 \quad s_2 = 0.5123$$

A nivel α, nos piden hacer el contraste

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

Para ello suponemos que $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$. Hacemos primero el contraste de homocedasticidad a nivel $\alpha = 0.1$

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

La región de rechazo de este contraste es $R = \{F = s_1^2/s_2^2 \notin (F_{22;29;0.95}, F_{22;29;0.05})\}$. Como $F = 0.7434641$ y $(F_{22;29;0.95}, F_{22;29;0.05}) \simeq (1/1.98, 1.92) = (0.51, 1.92)$, no hay suficiente evidencia para rechazar la hipótesis de homocedasticidad. De ahora en adelante suponemos que $\sigma_1 = \sigma_2 = \sigma$. Estimamos σ^2 mediante la varianza combinada:

$$s_p^2 = \frac{22s_1^2 + 29s_2^2}{51} = 0.2326.$$

Por ejemplo a nivel $\alpha = 0.05$, la región de rechazo del contraste de igualdad de medias es $R = \{|t| > t_{51;0.025} \simeq \frac{t_{40;0.025} + t_{60;0.025}}{2} = 2.01\}$, donde el estadístico del contraste es

$$t = \frac{0.977 - 0.883}{\sqrt{0.2326\left(\frac{1}{23} + \frac{1}{30}\right)}} = 0.703.$$

Por tanto, a nivel $\alpha = 0.05$ no hay evidencia de que los niveles esperados del gen sean diferentes en ambos grupos de mujeres. Entonces el p-valor es mayor que 0.05.

[1]Fuente de los datos: Pepe *et al.* (2003). Selecting Differentially Expressed Genes from Microarray Experiments. *Biometrics*, 59,133–142.

Con R:

```
X = c(0.51, 0.52, 0.62, 0.67, 0.67, 0.70, 0.76, 0.76, 0.79, 0.81, 0.81, 0.84,
    0.89, 0.94, 1.01, 1.09, 1.15, 1.15, 1.16, 1.27, 1.35, 1.37, 2.63)
Y = c(0.81, 0.70, 0.64, 0.67, 0.60, 0.42, 0.70, 0.55, 0.98, 1.10,
     0.69, 0.34, 0.60, 0.49, 1.19, 0.87, 2.33, 1.16, 0.50, 0.95,
     0.81, 2.78, 1.25, 0.69, 1.03, 0.69, 0.57, 0.72, 0.72, 0.94)

var.test(X,Y)
       F test to compare two variances

data:  X and Y
F = 0.73635, num df = 22, denom df = 29, p-value = 0.4637
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3375968 1.6790407
sample estimates:
ratio of variances
        0.7363538

t.test(X,Y,var.equal=T)
       Two Sample t-test

data:  X and Y
t = 0.70295, df = 51, p-value = 0.4853
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1743775 0.3622906
sample estimates:
mean of x mean of y
0.9769565 0.8830000
```