

PRACTICA 9: Inferencia sobre dos poblaciones . Enunciados.

Grado Biología, UAH.

Los datos que aparecen en el fichero `Practica09-robles.csv` son los que mismos con los que ya hemos trabajado en algunas prácticas anteriores (al menos en la 1, 2, ve allí para más información).

1. En relación a la dispersión de la concentración de Potasio

- (a) ¿es distinta para cada variedad de roble? Usa un nivel de significación del 10%.

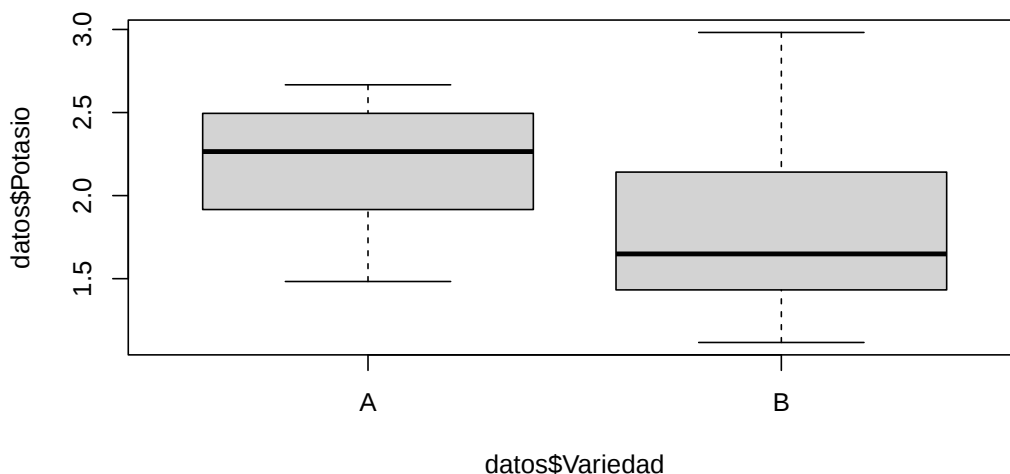
La pregunta es sobre el cociente de varianzas, en concreto, contrastar para la variable Potasio

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad H_1 : \sigma_A^2 \neq \sigma_B^2$$

para lo que se necesita la plantilla de cociente de varianzas, datos en bruto, dos poblaciones.

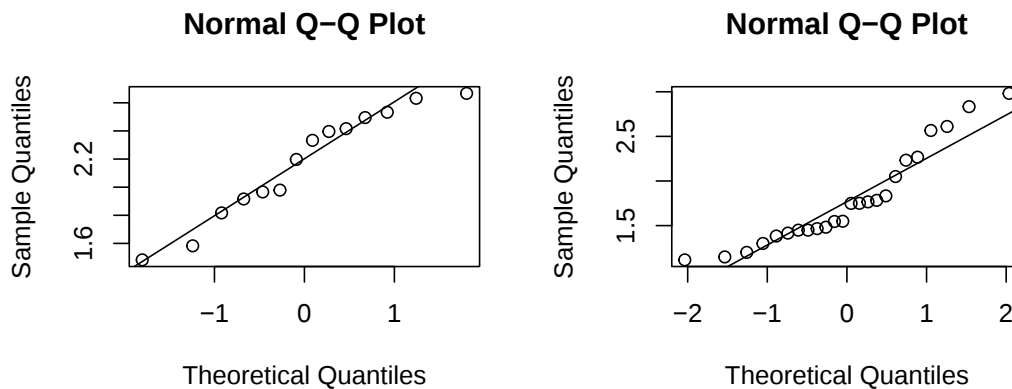
Una primera aproximación consiste en visualizar los boxplots para comparar, *grosso modo*, la dispersión de los datos

```
boxplot(datos$Potasio ~ datos$Variedad)
```



El boxplot parece indicar cierta diferencia, pero no si esa diferencia es estadísticamente significativa. Para decidir hay que hacer el contraste de varianzas y lo primero es evaluar la normalidad de los datos

```
## LECTURA DESDE FICHERO
tabla = read.table(file = "p1-robles.csv", sep = ";", header = T)
muestra1 = tabla$Potasio[tabla$Variedad == "A"]
muestra2 = tabla$Potasio[tabla$Variedad == "B"]
par(mfrow = c(1,2))
qqnorm(muestra1)
qqline(muestra1)
qqnorm(muestra2)
qqline(muestra2)
```



```
par(mfrow = c(1,1))
```

Ambas nubes de puntos están bastante alineadas, por lo que no dudamos de que los datos provienen de una población normal.

Se puede utilizar, además, el contraste de normalidad de Shapiro (no lo hemos usado antes, H_0 es que los datos provienen de una distribución normal)

```
shapiro.test(muestra1)
```

Shapiro-Wilk normality test

data: muestra1

W = 0.93455, p-value = 0.353

```
shapiro.test(muestra2)
```

Shapiro-Wilk normality test

data: muestra2

W = 0.90167, p-value = 0.02335

el primero arroja un p-valor compatible con asumir la normalidad de la muestra, no así el segundo (al nivel de confianza que plantea el ejercicio). Seguimos adelante (porque estamos en clase ;))

Así, contrastamos la diferencia de varianzas

```
(CHvar = var.test(x = muestra1, y = muestra2,
                  alternative = "two.sided",
                  conf.level = 0.9))
```

F test to compare two variances

data: muestra1 and muestra2

F = 0.51551, num df = 13, denom df = 23, p-value = 0.2164

alternative hypothesis: true ratio of variances is not equal to 1

90 percent confidence interval:

0.2369972 1.2520124

sample estimates:

ratio of variances

0.5155068

y el p-valor indica que la muestra no contradice H_0 , por lo que no se rechaza. Observa que el intervalo de confianza contiene al 1, lo que apoya esta decisión.

- (b) **Estima el cociente de ambas varianzas al nivel de confianza del 90%. Interpreta el resultado.**

Se pide interpretar el intervalo de confianza para el cociente de varianzas

```
CHvar$conf.int
## [1] 0.2369972 1.2520124
## attr(,"conf.level")
## [1] 0.9
```

es decir, σ_A^2 es entre 0.237 y 1.252 veces (redondeando) σ_B^2 .

2. Sobre la contaminación por Potasio

- (a) **¿Podemos decir que la contaminación por Potasio es distinta en ambas variedades? (indica la respuesta para diferentes niveles de significación: 0.1, 0.05, 0.01).**

El contraste en este caso es

$$\mu_A = \mu_B \quad \mu_A \neq \mu_B$$

En este caso hay que comparar las medias. Las muestras tienen tamaños

```
# para escribir menos crear esta variables
length(muestra1)

[1] 14
```

y

```
# para escribir menos crear esta variables
length(muestra2)

[1] 24
```

es decir, son muestras pequeñas. Del ejercicio anterior sabes que no hay motivo para dudar de que las muestras provengan de poblaciones normales, de modo que podemos usar la t de Student para comparar las medias. Por el ejercicio anterior sabes también que se puede considerar que las varianzas son iguales. Con la plantilla correspondiente se obtiene, para $\alpha = 0.1$

```
t.test(x = muestra1, y = muestra2, var.equal = TRUE, mu = 0,
       alternative = "two.sided", conf.level = .9)
```

Two Sample t-test

```
data: muestra1 and muestra2
t = 2.3464, df = 36, p-value = 0.02458
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 0.1076207 0.6597603
sample estimates:
mean of x mean of y
 2.172357  1.788667
```

para $\alpha = 0.05$

```
(CH.t = t.test(x = muestra1, y = muestra2, var.equal = TRUE, mu = 0,
               alternative = "two.sided", conf.level = .95))
```

Two Sample t-test

```
data:  muestra1 and muestra2
t = 2.3464, df = 36, p-value = 0.02458
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.0520573 0.7153237
sample estimates:
mean of x mean of y
 2.172357  1.788667
```

para $\alpha = 0.01$

```
t.test(x = muestra1, y = muestra2, var.equal = TRUE, mu = 0,
       alternative = "two.sided", conf.level = .99)
```

Two Sample t-test

```
data:  muestra1 and muestra2
t = 2.3464, df = 36, p-value = 0.02458
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
-0.06099864 0.82837959
sample estimates:
mean of x mean of y
 2.172357  1.788667
```

es decir, para $\alpha = 0.1, 0.05$ se rechaza H_0 mientras que para $\alpha = 0.01$ no hay evidencia muestral para rechazarla. Observa, y es importante, que uno de los extremos de los intervalos de confianza el muy próximo a 0, y que en el caso de para $\alpha = 0.05$ y para $\alpha = 0.01$ la diferencia es nimia, probablemente no sea significativa (relevante) desde el punto de vista científico.

- (b) **Estima la diferencia en la contaminación media por Potasio entre ambas variedades al nivel de confianza del 95%.** Se pide interpretar ahora el intervalo de confianza para la diferencia de medias

```
CH.t$conf.int
```

```
## [1] 0.0520573 0.7153237
## attr(,"conf.level")
## [1] 0.95
```

Como la diferencia es positiva, μ_A es mayor que μ_B entre 0.052 y 0.715 mg/Kg.

3. **Si dispusieras de muestras grandes (ambas muestras con más de 30 individuos), ¿cómo procederías?** No sería necesario comprobar la normalidad, y utilizaría una normal en lugar de una t de Student.

Los datos que contiene el fichero Practica09-ciudades.csv son los mismos que los de las prácticas 7 y 8, ve allí si necesitas información sobre la descripción de las variables.

4. **Cinco años después, en 1975, se vuelve a medir la variable SO2_{tr}, y se codifica como SO2_{tr75}. Se pide**
- (a) **¿Se puede decir, al 10% de nivel de significación, que el nivel de SO2_{tr} se incrementó entre 1970 y 1975?** En este caso los datos están pareados. Por tanto, hay que crear una nueva variable, que llamaremos **incremento**:

```
incremento = datos$SO2_tr75 - datos$SO2_tr
```

de nuevo habría que comprobar la normalidad de la nueva variable, pero se asume que lo es. A partir de la próxima práctica, que dispondrás de las herramientas adecuadas, habrá que hacer la comprobación, claro).

Hay que contrastar las hipótesis

$$H_0 : \mu_{\text{incremento}} \leq 0 \qquad H_1 : \mu_{\text{incremento}} > 0$$

La muestra es grande porque

```
length(incremento)
```

```
[1] 41
```

por lo que podemos usar un normal:

```
library(TeachingDemos)
muestra = incremento
mu0 = 0
(CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra),
                 alternative = "greater", conf.level = 0.9))
```

One Sample z-test

```
data: muestra
z = 1.7591, n = 41.00000, Std. Dev. = 2.13879, Std. Dev. of the sample
mean = 0.33402, p-value = 0.03928
alternative hypothesis: true mean is greater than 0
90 percent confidence interval:
 0.1594967      Inf
sample estimates:
mean of muestra
 0.5875637
```

y concluir que sí hubo un incremento estadísticamente significativo (aunque este hecho no indica de por sí que fuera un incremento científicamente significativo).

- (b) **Cuantifica la evolución de la concentración de SO₂ entre los años 1970 y 1975.** El intervalo de confianza lo obtuvimos como subproducto del contraste en el apartado anterior: observa que el hecho de que el intervalo no contenga al cero corrobora que una de ellas supera a la otra y, como es positivo, efectivamente ha habido un incremento en la concentración de SO₂. De hecho, este incremento es de entre 0.1595 y ∞ .
5. **Hemos comprobado que, sobre esta muestra, el 56% de las ciudades con más de 1 millón de habitantes (POP>100) tienen un plan de sostenibilidad medioambiental. Este mismo porcentaje en Canadá, medido sobre una muestra de 35 ciudades, es del 72%. ¿Podemos admitir, al 99% de confianza, que hay diferencias significativas a este respecto entre ambos países?** De nuevo hay que contrastar hipótesis. En este caso sobre proporciones. Las hipótesis son

$$H_0 : p_{USA} = p_{Canada} \qquad H_1 : p_{USA} \neq p_{Canada}$$

Veamos cuántas ciudades de la muestras supera el millón de habitantes

```
sum(datos$POP > 100)
```

```
[1] 39
```

esto es importante para poder consignar el tamaño de la muestra 1

```
n1 = sum(datos$POP > 100) # tamaño muestra 1
n2 = 35 # tamaño muestra 2
# k1 = # num de exitos 1
# k2 = # num de exitos 2

# Si tienes proporciones muestrales p1, p2, entonces
# comenta las dos líneas anteriores. Después, descomenta
# y usa estas cuatro líneas:
p1 = 0.56
p2 = 0.72
k1 = n1 * p1
k2 = n2 * p2
# Opciones para H1: greater / less / two.sided
(CH = prop.test(c(k1, k2), n = c(n1, n2), correct = FALSE,
                alternative = "two.sided",
                conf.level = 0.99))

2-sample test for equality of proportions without continuity
correction

data:  c(k1, k2) out of c(n1, n2)
X-squared = 2.039, df = 1, p-value = 0.1533
alternative hypothesis: two.sided
99 percent confidence interval:
 -0.4430829  0.1230829
sample estimates:
prop 1 prop 2
 0.56  0.72
```

El p-valor indica que las proporciones se pueden considerar iguales.

Si se quisiera cuantificar entre qué valores está la diferencia, basta con fijarse en el intervalo de confianza para la diferencia de proporciones:

```
[1] -0.4431  0.1231
```

es decir, $p_{USA} - p_{Canada} \in (-0.4431, 0.1231)$ lo que equivale a

$$p_{Canada} - 0.4431 < p_{USA} < p_{Canada} + 0.1231$$