PRACTICA 8: contraste de hipótesis paramétricos. Soluciones.

Grado en Biología Sanitaria, UAH, 2023/24.

Objetivos

- Determinar las condiciones de aplicabilidad de los contrastes de hipótesis paramétricos y el tipo de contraste a utilizar en cada situación.
- Realizar contrastes de hipótesis paramétricos para la media, la varianza y la proporción con una y dos poblaciones.
- Complementar la información que proporciona el p-valor con la del intervalo de confianza.
- Interpretar los resultados obtenidos.

Los datos que aparecen en el fichero `Practica08-robles.csv` se refieren a un estudio realizado sobre un robledal cercano a una planta industrial, parte de los cuales se visualizan en la tabla. Se han seleccionado robles de dos variedades (A y B) y ubicados en cuatro zonas distintas. Además la mitad de ellos han sido sometidos a cierto tratamiento (codificados con 1), los no tratados (codificados con 0). Sobre cada árbol se han medido las concentraciones (mg/kg) de ocho elementos químicos en sus hojas: hierro, manganeso y zinc (metales pesados); calcio y magnesio (metales alcalinotérreos); potasio (metal alcalino); y fósforo y nitrógeno (no metales).

Además

- Sipondremos que has guardado en la variable `robles` el contenido de `Practica08-robles.csv`
- Aunque en las soluciones se leen los datos con `read.table()`, tú puedes usar el botón `Import Dataset`.

```
robles = read.table(file = "Practica08-robles.csv", sep = ";", header = TRUE, dec = ".")
```

```
head(robles, 4)
```

```
  Num Hierro Manganeso   Zinc Calcio Magnesio Potasio Fosforo Nitrogeno Zona
1   1  0.058    0.0303 0.0089  2.365    0.400   2.632   0.145     2.776    1
2   2  0.060    0.0294 0.0109  2.745    0.432   2.495   0.161     2.918    1
3   3  0.058    0.0289 0.0090  2.513    0.349   2.396   0.169     4.826    1
4   4  0.059    0.0275 0.0090  2.361    0.349   1.979   0.155     4.893    1
  Variedad Tratamiento
1        A           0
2        A           0
3        A           0
4        A           0
```

Responde de forma concisa y razonadamente a las siguientes preguntas:

1. **Contrasta la hipótesis de que la media estimada para la población para la variable Magnesio es distinta de 0.41, al nivel de significación del 5%. Usa R para calcular el p-valor y el intervalo de confianza. Calcula, además, la región de rechazo. Comenta los resultados. En particular, en relación al tamaño del efecto, compara la relación entre el intervalo de confianza y μ_0, y entre la región de no rechazo y $\bar{X}$. Usa 3 cifras significativas**.

   La media de cualquier muestra nunca valdrá exáctamente 0.41, de modo que lo que puedes medir es si la media de tu muestra es demasiado diferente de 0.41, por lo que se trata de un contraste bilateral sobre la media. En concreto, contrastar

   $$H_0 : \mu_{Mg} = 0.41 \qquad H_1 : \mu_{Mg} \neq 0.41$$

   Como la población es grande ($n > 30$) se usa una normal, en concreto, la plantilla `media_1pob_T_o_Z_enBruto.R` para obtener el **p-valor** y el **intervalo de confianza**:

```r
library(MASS)
library(TeachingDemos)

# Ajusta los parametros de read.table.
robles = read.table(file = "Practica08-robles.csv",
                    sep = ";", header = TRUE, dec = ".")
muestra = robles$Magnesio

# ajustar para contrastar hipotesis
mu0 = 0.41

# Opciones para alternativa: greater / less / two.sided
# Elige el contraste adecuado y descomenta la linea correspondiente
# (CHconT = t.test(muestra, mu = mu0, alternative = "two.sided", conf.level = ))
(CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra),
                 alternative = "two.sided", conf.level = 0.95))
```

```
One Sample z-test

data:  muestra
z = -1.0402, n = 38.000000, Std. Dev. = 0.075792, Std. Dev. of the
sample mean = 0.012295, p-value = 0.2982
alternative hypothesis: true mean is not equal to 0.41
95 percent confidence interval:
 0.3731125 0.4213086
sample estimates:
mean of muestra
      0.3972105
```

En concreto

(a) **p-valor**

```r
signif(unname(CHconZ$p.value), 4)
```

```
[1] 0.2982
```

(b) **intervalo de confianza**

```r
signif(unname(CHconZ$conf.int)[1:2], 2)
```

```
[1] 0.37 0.42
```

Por otro lado, se pide la **región de rechazo**. Por la teoría sabemos que, en caso de ser cierta H_0, se tiene que (busca la desviación típica muestral y el tamaño de la muestra en la salida de R):

$$\bar{X}_{Mg} \sim N(\mu_0,\, s/\sqrt{n}) = N(0.41,\, 0.076/\sqrt{38})$$

por tanto, los extremos de la región de rechazo los marcan los valores de $\bar{X}_{Mg}$ que dejan por debajo y por encima de sí una probabilidad de 0.025 ($\alpha/2$), es decir

```r
signif(qnorm(c(0.025, 0.975), mean = 0.41, sd = 0.076/sqrt(38)), 2)
```

```
[1] 0.39 0.43
```

Un enfoque alternativo (pero totalmente equivalente) es usar la N(0,1) estandarizando la media muestral. La expresión, en este caso, es

$$\mu_{Mg} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

es decir, la región de no rechazo está entre

```
a  = mu0 - qnorm(0.025, lower.tail = F)*sd(muestra)/sqrt(length(muestra))
signif(a, 2)

[1] 0.39
```

y

```
b = mu0 + qnorm(0.025, lower.tail = F)*sd(muestra)/sqrt(length(muestra))
signif(b, 2)

[1] 0.43
```

Observa que la semi-amplitud de la región de no rechazo es $z_{\alpha/2}\frac{s}{\sqrt{n}}$, que es la misma que la del intervalo de confianza. La diferencia es que el intervalo de confianza está centrado en la media muestral y la región de no rechazo en la media poblacional. Así, es equivalente decir que

(a) No hay evidencia muestral para rechazar H_0 porque 0.41 (la media hipotetizada) está en el intervalo de confianza (0.37, 0.42), que es donde esperamos que esté con probabilidad 0.95.

(b) No hay evidencia muestral para rechazar H_0 porque la media observada 0.3972105 (la media muestral) está en la región de no rechazo (0.39, 0.43), que contiene el 95% de las medias muestrales más probables si no se rechaza H_0.

(c) El p-valor vale $0.2982 > \alpha = 0.05$, por lo que no se rechaza H_0.

Para calcular el **tamaño del efecto** es equivalente determinar

(a) La distancia del estadístico de contraste (la media muestral) y el extremo más próximo de la región de no rechazo.

```
[1] 0.011
```

(b) La distancia del valor de H_0 y el extremo más próximo del intervalo de confianza).

```
[1] 0.011
```

Muy importante: H_0 **nunca** se acepta, en todo caso, no se rechaza.

2. **¿Puede admitirse, al nivel de significación del 5%, que la contaminación media por Potasio está por encima de 1.5? Determina el tamaño del efecto. Calcula, además, la región de rechazo. Comenta los resultados.**

Recuerda que H_0 nunca se acepta; como mucho, diremos que no hay evidencias para rechazarla (que no es lo mismo). Por eso, para admitir que la concentración media es mayor que 1.5, hay que suponer lo contrario (H_0, que es menor o igual que 1.5) y, si la muestra contradice dicha hipótesis, al rechazarla nos quedamos con la alternativa. Es decir, podremos admitir que la concentración media de Potasio es superior a 1.5.

Se trata de un contraste de hipótesis unilateral sobre la media:

$$H_0 : \mu_K \leq 1.5 \qquad H_1 : \mu_K > 1.5$$

Como la población es grande ($n > 30$) se usa una normal, en concreto, la plantilla `media_1pob_T_o_Z_enBruto.R`.

```r
library(MASS)
library(TeachingDemos)

# Ajusta los parametros de read.table.
robles = read.table(file = "Practica08-robles.csv",
                    sep = ";", header = TRUE, dec = ".")
muestra = robles$Potasio

# ajustar para contrastar hipotesis
mu0 = 1.5

# Opciones para alternativa: greater / less / two.sided
# Elige el contraste adecuado y descomenta la linea correspondiente
# (CHconT = t.test(muestra, mu = mu0, alternative = "two.sided", conf.level = ))
(CHconZ = z.test(muestra, mu = mu0, stdev = sd(muestra),
                 alternative = "greater",
                 conf.level = 0.95))
```

```
One Sample z-test

data:  muestra
z = 5.1474, n = 38.000000, Std. Dev. = 0.514993, Std. Dev. of the
sample mean = 0.083543, p-value = 1.321e-07
alternative hypothesis: true mean is greater than 1.5
95 percent confidence interval:
 1.79261     Inf
sample estimates:
mean of muestra
       1.930026
```

Observa que, en este caso, el **intervalo de confianza**

$$(1.8, \infty)$$

tiene un aspecto raro, porque su extremo superior es $+\infty$. Este tipo de intervalos no suelen explicarse a nivel de primer curso de grado, pero se usan en *la realidad*. Esencialmente, nos dice que con una probablidad del 95%, a la vista de la muestra, la concentración de Potasio será de, al menos, 1.8mg/kg.

El tamaño del efecto es la distancia de $\mu = 1.5$ al extremo más próximo del intervalo de confianza (en este caso, el inferior):

```
[1] 0.29
```

Para la **región de rechazo**, la teoría dice que, en caso de ser cierta H_0, se tiene que (busca la desviación típica muestral y el tamaño de la muestra en la salida de R):

$$\bar{X}_K \sim N(1.5,\, 0.51/\sqrt{38})$$

Ahora la región de rechazo está formada por el 5% de los valores de $\bar{X}_K$ más contradictorios con H_0, es decir, el 5% de los que más se alejan de 1.5 por su derecha. Por tanto, la región de rechazo la forman los valores de la media muestral mayores que

```
h  = qnorm(0.05, 1.5, sd(muestra)/sqrt(length(muestra)), lower.tail = F)
signif(h, 2)

[1] 1.6
```

Observa que las regiones de rechazo/no rechazo, el intervalo de confianza y el p-valor son coherentes entre sí:

(a) La muestra sugiere rechazar H_0 porque 1.5 (la media hipotetizada) no está en el intervalo de confianza $(1.8, \infty)$, que es donde esperamos que esté con probabilidad 0.95.

(b) Hay evidencia muestral para rechazar H_0 porque la media observada 1.93 (la media muestral) está en la regi'on de rechazo $(1.6, +\infty)$, que contiene el 5% de las medias muestrales más contradictorias con H_0.

(c) El p-valor vale $1.3208149 \times 10^{-7} < \alpha = 0.05$, por lo que se rechaza H_0.

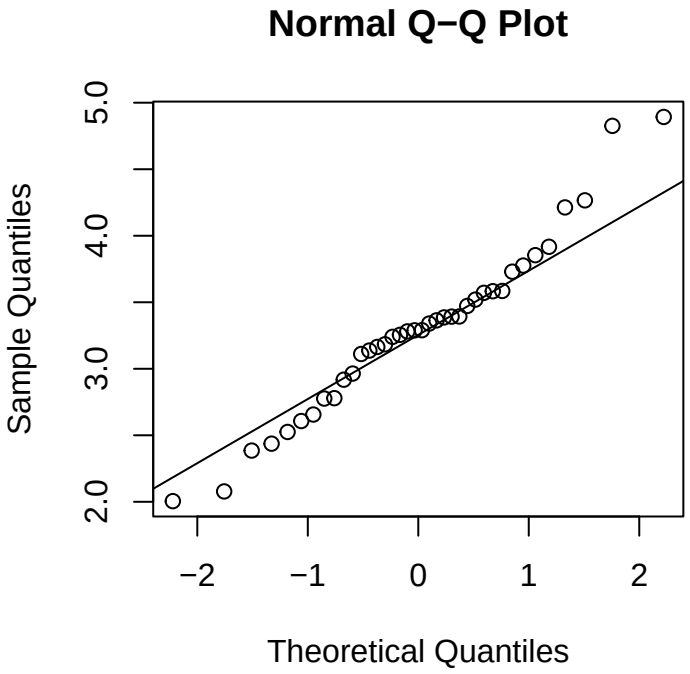
3. **¿Puede admitirse, al nivel de significación del 1%, que la varianza del Nitrógeno es inferior a 0.3?**

Se pide un contraste de hipótesis unilateral sobre la varianza:

$$H_0 : \sigma_N^2 \geq 0.3 \qquad H_1 : \sigma_N^2 < 0.3$$

Como se trata de la varianza, hay que comprobar la normalidad de los datos:

```
qqnorm(robles$Nitrogeno)
qqline(robles$Nitrogeno)
```



excepto por unos pocos datos en los extremos los puntos están bastante alineados, por lo que no rechazamos la normalidad de los datos. Ahora se usa la plantilla `varianza_1pob_enBruto.R`

```
library(MASS)
library(TeachingDemos)

robles = read.table(file = "Practica08-robles.csv",
                    sep = ";", header = TRUE, dec = ".")
muestra = robles$Nitrogeno

# ajustar para contrastar hipotesis
sigma0 = sqrt(0.3)
# Opciones para alternativa: greater / less / two.sided
(CH = sigma.test(muestra, sigma = sigma0,
                 alternative = "less", conf.level = 0.99))


One sample Chi-squared test for variance

data:  muestra
X-squared = 50.756, df = 37, p-value = 0.9346
alternative hypothesis: true variance is less than 0.3
99 percent confidence interval:
 0.000000 0.762856
sample estimates:
var of muestra
     0.4115347
```

por lo que no se rechaza la hipótesis nula; observa que la varianza muestral es mayor que 0.3, es decir, 0.3 está en el intervalo de confianza para la varianza (y no está cerca de sus extremos).

En este caso el tamaño del efecto se mide mediante un cociente:

(a) Se puede usar el cociente σ_0^2/(extremo más próximo del intervalo de confianza)

```
(s1 = signif( 0.3/0.762856, 3))

[1] 0.393
```

y se interpreta como que σ_0^2 podría ser hasta 0.393 veces mayor antes de ser rechazada H_0.

(b) Se puede usar el cociente (extremo más próximo del intervalo de confianza)/σ_0^2

```
(s2 = signif( 0.762856/0.3, 3))

[1] 2.54
```

e interpretarlo como que dispersión σ_0^2 es el 39.3% del máximo compatible con H_0.

Para que el intervalo de confianza sugiriera rechazar H_0 tendríamos que haber obtenido algo como (0, 0.2) (y un p-valor pequeño, claro). Si la varianza estuviera con probabilidad 0.99 entre (0, 0.2), rechazaríamos la posibilidad de que valiera 0.3 o más.

4. **¿Tienes motivos para dudar de que la concentración mediana de Nitrógeno es de 3mg/Kg, con un nivel de significación del 10%? Calcula el tamaño del efecto.**

Su la mediana es 3, la mitad de los robles tendrá más de 3 mg/Kg de Magnesio (y la otra mitad, menos). Si llamamos p a la proporción de robles con una concentración de Nitrógeno superior a 3, se trata de contrastar la hipótesis:

$$H_0 : p = 1/2 \qquad H_1 : p \neq 1/2$$

Se usa la plantilla `proporcion_1pob.R` en la que

```r
robles = read.table(file = "Practica08-robles.csv",
                    sep = ";", header = TRUE, dec = ".")
muestra = robles$Nitrogeno
n = length(muestra)  #num. elementos muestra
k = sum(muestra>3)  #num. exitos muestra
pMuestral = k / n

# ajustar para contrastar hipotesis
p0 = 0.5

# Opciones para alternativa: greater / less / two.sided
prop.test(k, n, conf.level =0.9 , p = p0, correct=FALSE, alternative = "two.sided")
```

```
1-sample proportions test without continuity correction

data:  k out of n, null probability p0
X-squared = 6.7368, df = 1, p-value = 0.009444
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.5787774 0.8142895
sample estimates:
        p
0.7105263
```

el p valor es menor que el nivel de significiación $\alpha = 0.1$, por lo que se rechaza H_0.

Ahora el intervalo de confianza no contiene el valor de la proporción muestra estipulado en H_0, lo que corrobora la decisión tomada. Además, $1/2$ está alejado el extremo inferior del intervalo, por lo que estamos seguros de la decisión tomada en el contrate.

En concreto, el tamaño del efecto es 0.07888 (es decir, 7.89 puntos porcentuales).

5. **Se supone que la concentración media de Manganeso es menor que** 0.01**. Se toma la decisión de rechazar esta hipótsis si se observa una concentración muestral mayor que** 0.013 **¿Qué nivel de significación está asociado con esa regla de decisión?**

   El contraste de hipótesis subyacente es $H_0 : \mu_{Mg} \leq 0.01$ frente a $H_1 : \mu_{Mg} > 0.01$, y el valor 0.013 marca el inicio de la región de rechazo. Así, la probabilidad

```r
pnorm(0.013, mean = 0.01,
      sd = sd(robles$Manganeso)/sqrt(length(robles$Manganeso)),
      lower.tail = FALSE)
```

```
[1] 0.0553651
```

nos da el nivel de significación pedido.