PRÁCTICAS DE ESTADÍSTICA CON *R*

PRÁCTICA 4: INFERENCIA ESTADÍSTICA

4.1 Estimación de la media

Los conceptos que vamos a ir desarrollando a lo largo de la primera parte de la práctica hacen referencia a los dos ejercicios siguientes:

■Ejemplo 4-1-1: Se midieron en 10 días tomados al azar los niveles de cloro del agua que sale de una planta de tratamiento obteniéndose los valores:

2,2 – 1,9 – 1,7 – 1,6 – 1,7 – 1,8 – 1,7 – 1,9 – 2,0 – 2,0

Se pide: a) Dar una estimación puntual de la media y de la varianza de la variable aleatoria que representa el nivel de cloro. b) Intervalo de confianza al 95% para la media y la varianza. c) Ídem al 99%. Se supone normalidad en los datos.

■Ejemplo 4-1-2: Se quiere saber si existen diferencias significativas en la facturación de dos tiendas de joyería de una misma cadena. Para ello se eligieron al azar 11 días en los que se contabilizaron las ventas en la joyería A y otros 10 días en la joyería B. Los datos obtenidos fueron:

Ventas A	1320	1495	990	1250	12900	1900	1500	1100	1250	1100	1930
Ventas B	1110	1405	985	1290	1300	1705	1200	1105	1150	1210	

Obtener las estimaciones por punto y por intervalo de las medias, varianzas, diferencia de medias y cociente de varianzas. Se supone normalidad en los datos.

Vamos a obtener en primer lugar la estimación puntual de la media, o sea la media muestral, para el primer ejemplo.

```
> # Ejemplo 4-1-1
> nivclor<-c(2.2,1.9,1.7,1.6,1.7,1.8,1.7,1.9,2,2)
> nivclor
 [1] 2.2 1.9 1.7 1.6 1.7 1.8 1.7 1.9 2.0 2.0
> mean(nivclor)
[1] 1.85
```

Para el segundo ejemplo estimaremos las medias de las variables **VentasA** y **VentasB**. Se supone que los datos de este ejercicio están en el archivo "Ejemplo 6-2.txt", cuya ruta de acceso es: "C:/Ejemplo 4-1-1.txt".

```
> # Ejemplo 4-1-1
> # Leamos el archivo de datos
> datos<-read.table("C:\\Ejemplo 4-1-1.txt",header=T)
```

```
Error en scan(file, what, nmax, sep, dec, quote, skip,
nlines, na.strings,  :
        la línea 11 no tiene 2 elementos
```

Como se ve aparece un mensaje de error porque **"VentasB"** tiene un dato menos que **"VentasA"**. Por tanto, lo primero que debemos hacer es transformar el archivo "Ejemplo 4-1-2.txt" en el archivo "Ejemplo 4-1-2-nuevo.txt" poniendo NA (*not available)* en el último valor de las ventas de B. Así evitaremos problemas al leer el archivo:

```
> datos<- read.table("C:\\Ejemplo 4-1-2-
nuevo.txt",header=T)
> datos
  VentasA VentasB
1   1320 1110
2   1495 1405
3    990  985
4   1250 1290
5  12900 1300
6   1900 1705
7   1500 1200
8   1100 1105
9   1250 1150
10  1100 1210
11  1930   NA
> attach(datos)

> VentasA
 [1]  1320  1495   990  1250 12900  1900  1500  1100
1250  1100  1930
> VentasB
 [1] 1110 1405  985 1290 1300 1705 1200 1105 1150 1210
NA

> mean(VentasA,na.rm=T);mean(VentasB,na.rm=T)
[1] 2430.455
[1] 1246
> # Con la opción na.rm=T no tenemos en cuenta los
valores perdidos que pudiera haber. No obstante, otra
alternativa es eliminar tales valores, después de
haberlos detectado previamente

> is.na(VentasA);is.na(VentasB)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE  TRUE
> # El elemento que ocupa el lugar 11 en VentasB es
NA. Lo eliminamos
> VentasB.nuevo<-VentasB[-11]
```

```
> mean(VentasA);mean(VentasB.nuevo)
[1] 2430.455
[1] 1246
```

4.2 Estimación de la varianza y de la desviación típica

Los estimadores centrados de la varianza y de la desviación típica son, respectivamente, la cuasivarianza y la cuasidesviación típica. Calculemos esos valores para los ejemplos 4-1-1 y 4-1-2.

```
> var(nivclor)
[1] 0.03388889
> sd(nivclor)
[1] 0.1840894
> var(VentasA);var(VentasB.nuevo)
[1] 12151922
[1] 39993.33
> sd(VentasA);sd(VentasB.nuevo)
[1] 3485.961
[1] 199.9833
```

Como ya se ha indicado en la práctica 1, mediante la función **var** se obtiene la cuasivarianza muestral y no la varianza muestral. Así mismo, la función **sd** da la cuasidesviación típica muestral (*standard deviation*) en lugar de la desviación típica muestral.

Como es evidente, se obtienen los mismos valores anteriores si hacemos

```
> sqrt(var(nivclor))
[1] 0.1840894
> sqrt(var(VentasA));sqrt(var(VentasB.nuevo)
[1] 3485.961
[1] 199.9833
```

4.3 Intervalo de confianza para la media

La manera más cómoda, aunque indirecta, de obtener en *R* intervalos de confianza es a través de tests o contrastes de hipótesis, cuestión que veremos con más detalle en el capítulo siguiente.

Para conseguir un intervalo de confianza para la media del ejemplo 4-1-1, como no se conoce la varianza de la población, usamos la función **t.test**, que es el test de la t de Student para una muestra.

Veamos cómo se obtienen los intervalos de confianza, al nivel del 95% y del 99%, para la media de la variable **nivclor**:

```
> t.test(nivclor)$conf # Por defecto se obtiene el intervalo de confianza al nivel del 95%
95 percent confidence interval:
 1.718310 1.981690

> # Para obtener solo el IC al 99% hacemos
> t.test(nivclor,conf.level=0.99)$conf
 [1] 1.660814 2.039186
```

4.4 Intervalo de confianza para la varianza

Para construir un intervalo de confianza para la varianza del ejemplo 4-1-1 utilizamos la fórmula correspondiente (S es la *cuasidesviación típica*):

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right]$$

```
> # Redondeamos los valores a 4 decimales
> round(c(9*var(nivclor)/qchisq(0.975,9),9*var(nivclor)/qchisq(0.025,9)),4)
[1] 0.0160 0.1129

> round(c(9*var(nivclor)/qchisq(0.995,9),9*var(nivclor)/qchisq(0.005,9)),4)
[1] 0.0129 0.1758
```

Podríamos crear en *R* un *objeto* (función) que calcule el intervalo de confianza, objeto que podemos almacenar para usos posteriores:

```
> CI.para.la.varianza<-function(x,alfa)
round(c((length(x)-1)*var(x)/qchisq(1-alfa/2,length(x)-1),(length(x)-1)*var(x)/qchisq(alfa/2,length(x)-1)),4)

> CI.para.la.varianza(x=nivclor,alfa=0.05)
[1] 0.0160 0.1129
> CI.para.la.varianza(x=nivclor,alfa=0.01)
[1] 0.0129 0.1758
```

4.5 Intervalo de confianza para el cociente de varianzas

Para obtener un intervalo de confianza para el cociente de varianzas de una forma directa, sin usar la fórmula, debemos utilizar la función **var.test** sobre la que profundizaremos en el capítulo siguiente.

```
> var.test(VentasA,VentasB.nuevo)$conf
[1]   76.65465 1148.23288

> # Como el valor 1 no está incluido en el intervalo, deducimos que las varianzas de las dos poblaciones no pueden ser consideradas iguales
```

4.6 Intervalo de confianza para la diferencia de medias

Al igual que en casos anteriores, obtendremos el intervalo de confianza para la diferencia de medias de las variables **VentasA** y **VentasB** del ejercicio 4-1-2 de forma indirecta, utilizando un test de hipótesis.

```
> t.test(VentasA,VentasB.nuevo,var.equal=F,conf.level=0.95)$conf
[1] -1159.397  3528.307

> # De acuerdo al resultado del apartado anterior debemos elegir la opción var.equal=F para indicarle al programa que las varianzas no pueden ser consideradas iguales
> # Al ser las varianzas poblacionales desconocidas y desiguales el programa ha utilizado la aproximación de Welch
```

4.7 Contraste sobre la media y la varianza de una población normal

En relación a los datos del ejemplo 4-1-1 vamos a contrastar si se puede aceptar que provienen de una población normal de media 1.9. Como la varianza de la población es desconocida debemos realizar un contraste t de Student:

```
> # Ejemplo 6-1 (cont.)
> nivclor<-c(2.2,1.9,1.7,1.6,1.7,1.8,1.7,1.9,2,2)
> nivclor
 [1] 2.2 1.9 1.7 1.6 1.7 1.8 1.7 1.9 2.0 2.0
> t.test(nivclor,mu=1.9)

        One Sample t-test
```

```
data:  nivclor
t = -0.8589, df = 9, p-value = 0.4127
alternative hypothesis: true mean is not equal to 1.9
95 percent confidence interval:
 1.718310 1.981690
sample estimates:
mean of x
     1.85
```

La interpretación de esta salida es la siguiente: Se ha efectuado un test bilateral (véase la hipótesis alternativa) de la t de Student con 9 grados de libertad (df). El estadístico de contraste es -0.8589. Se da así mismo el intervalo de confianza al nivel 95%. El p-valor es 0.4127. Dado que este p-valor es muy alto, no se puede rechazar la hipótesis nula de que la media de la población es 1,9. La media muestral es 1.85.

A continuación se desea contrastar si la varianza de la población, cuyos datos son los del ejemplo 4-1-1, es 0,05. Para ello procederemos a construir una función que no está directamente implementada en *R*. Lo haremos de una forma simple, obteniendo la región de aceptación y posteriormente calcularemos el p-valor.

```
> # Definimos una función para efectuar el test sobre la varianza de una población normal de media desconocida

> var.test.una.población.normal<-function(x,conf.level=0.95) {
+ n=length(x)
+ alfa=1-conf.level
+ valcrit1=qchisq(1-alfa/2,n-1)
+ valcrit2=qchisq(alfa/2,n-1)
+ c((n-1)*var(x)/valcrit1,(n-1)*var(x)/valcrit2) }

> var.test.una.población.normal(nivclor)
[1] 0.01603342 0.11294667
> # Como el valor 0.05 pertenece al intervalo anterior no podemos rechazar la hipótesis de que la varianza de la población es 0.05

> # A continuación obtenemos el valor-p del test
> var(nivclor)
[1] 0.03388889
> pchisq(8*var(nivclor)/0.05,8)
[1] 0.2883590
> # Valor muy grande que da gran seguridad en aceptar la hipótesis nula
```

4.8 Contraste sobre la igualdad de varianzas de dos poblaciones normales

Aunque el objetivo del ejemplo 4-1-2 es comparar las ventas en las joyerías A y B, para lo cual debemos confrontar las ventas medias en ambos establecimientos, previamente efectuaremos un contraste sobre la igualdad de varianzas (desviaciones típicas), ya que el resultado que obtengamos será utilizado posteriormente en el contraste sobre las medias.

```
> #Efectuamos el contraste de igualdad de varianzas

> datos<- read.table("C:\\Ejemplo 4-1-2-
nuevo.txt",header=T)
> datos
   VentasA VentasB
1     1320    1110
2     1495    1405
3      990     985
4     1250    1290
5    12900    1300
6     1900    1705
7     1500    1200
8     1100    1105
9     1250    1150
10    1100    1210
11    1930      NA
> attach(datos)
> var.test(VentasA,VentasB)

        F test to compare two variances

data:  VentasA and VentasB.nuevo
F = 303.8487, num df = 10, denom df = 9, p-value =
7.775e-10
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
   76.65465 1148.23288
sample estimates:
ratio of variances
          303.8487
```

Interpretación del resultado anterior: en primer lugar se observa que se ha efectuado un test mediante la distribución F de Snedecor sobre la igualdad de varianzas de dos poblaciones (normales), cuyo estadístico de contraste F es 303.8487. Los grados de libertad del numerador y denominador son, respectivamente, 10 y 9. El valor-p resultante es muy pequeño por lo que se

rechaza la hipótesis nula de igualdad de varianzas. Como se observa, la hipótesis alternativa es que el cociente de varianzas no es igual a 1 (varianzas distintas).

Como se había comentado en el capítulo anterior, adicionalmente este contraste da información sobre el intervalo de confianza al 95% (valor por defecto) para el cociente de varianzas, siendo la correspondiente estimación puntual 303.8487.

4.9 Contraste de igualdad de medias de dos poblaciones normales

Ahora estamos en condiciones de contrastar la igualdad de medias de las ventas en las joyerías A y B del ejemplo 4-1-2. Al haber deducido del contraste anterior que las varianzas poblacionales de ambos establecimientos se pueden considerar diferentes, pondremos **var.equal=F** (*false*).

```
> t.test(VentasA,VentasB,var.equal=F)
        Welch Two Sample t-test
data:  ventasA and ventasB
t = 1.1249, df = 10.072, p-value = 0.2867
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -1159.397  3528.307
sample estimates:
mean of x mean of y
 2430.455  1246.000
```

Si deseamos incrementar el nivel del confianza hasta el 99% podemos hacer:

```
> t.test(VentasA,VentasB,var.equal=F,conf.level=0.99)
        Welch Two Sample t-test
data:  VentasA and VentasB
t = 1.1249, df = 10.072, p-value = 0.2867
alternative hypothesis: true difference in means is
not equal to 0
99 percent confidence interval:
 -2147.289  4516.198
sample estimates:
mean of x mean of y
 2430.455 1246.000
```

La salida de resultados se interpreta del siguiente modo: el programa ha efectuado una aproximación (Welch) al no poder considerarse iguales las varianzas poblacionales. El valor del estadístico de contraste es t=1.1249, los grados de libertad son 10.072 (se trata de una aproximación) y el valor-p es 0.2867. La hipótesis alternativa se refiere a que la diferencia de medias no es igual a cero o,

equivalentemente, que las medias no son iguales. Adicionalmente se obtienen los correspondientes intervalos de confianza al 95% (por defecto) y al 99%.

Como el valor-p es muy grande no podemos rechazar la igualdad de medias y concluimos que no hay diferencias significativas en las ventas de los establecimientos A y B.

4.10 Contraste sobre una proporción

■Ejemplo 4-10: Se ha encuestado a 110 personas sobre si están de acuerdo con la construcción del tren de alta velocidad, habiendo contestado 48 de ellas afirmativamente. ¿Respalda este resultado la hipótesis de que la proporción de opiniones afirmativas en la población es el 50%?

```
> prop.test(48,110,p=0.5)

        1-sample proportions test with continuity
correction

data:  48 out of 110, null probability 0.5
X-squared = 1.5364, df = 1, p-value = 0.2152
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3431002 0.5341288
sample estimates:
        p
0.4363636
```

Se ha efectuado un test para la proporción de una población con corrección por continuidad. En lugar de utilizar el habitual test que aplica la aproximación de la distribución binomial a la normal, se utiliza aquí otro en el que se emplea la distribución χ^2. El alto valor-p obtenido respalda la hipótesis nula de que la proporción en la población es del 50%.

4.11 Contrastes χ^2

En este apartado vamos a ver dos ejemplos de aplicación de contrastes χ^2: el primero de bondad de ajuste y el segundo de independencia.

■Ejemplo 4-11-1: Las letras que más frecuentemente aparecen en el idioma inglés son E, N, T, R y O. Cuando alguna de ellas se presenta en un texto, la probabilidad de que aparezca cada una viene dada en la tabla siguiente:

E	N	T	R	O
0.29	0.17	0.21	0.17	0.16

Esta información es útil en criptografía. Supongamos que en un cierto texto se han contabilizado estas cinco letras apareciendo cada una de ellas el número de veces que se indica en la tabla:

E	N	T	R	O
100	80	110	55	14

Efectuar un contrate χ^2 de bondad de ajuste de los datos muestrales a la distribución teórica.

```
> x<-c(100,80,110,55,14)
> prob.teoricas<-c(0.29,0.17,0.21,0.17,0.16)
> chisq.test(x,p=prob.teoricas)

        Chi-squared test for given probabilities

data:  x
X-squared = 55.3955, df = 4, p-value = 2.685e-11
```

Como el valor-p es muy pequeño, es muy improbable que el texto esté escrito en inglés.

■Ejemplo 4-11-2: Una compañía evalúa una propuesta para fusionarse con una corporación. Una muestra aleatoria simple de 250 accionistas proporciona la siguiente información:

Núm. de acciones por accionista	A favor	En contra	Indecisos
Menos de 200	38	29	9
Entre 200 y 1000	30	42	7
Más de 1000	32	59	4

¿Existe alguna razón para dudar de que la opinión con respecto a la propuesta es independiente del número de acciones que posee cada accionista?

```
> pequeño<-c(38,29,9)
> mediano<-c(30,42,7)
> grande<-c(32,59,4)
> chisq.test(data.frame(pequeño,mediano,grande))
```

```
        Pearson's Chi-squared test

data:  data.frame(pequeño, mediano, grande)
X-squared = 10.7957, df = 4, p-value = 0.02896
```

El valor-p obtenido en este test de independencia no es ni pequeño (<0.01) ni grande (>0.1), por lo que el resultado es algo ambiguo y lo más conveniente sería repetir la prueba aumentando, si es posible, el tamaño muestral.

PRÁCTICA 4: EJERCICIOS

■Ejercicio 4-1: Se quiere estimar el precio del metro cuadrado de vivienda nueva en el municipio de Getxo. Para ello se han tomado 12 viviendas al azar, obteniéndose los valores siguientes en miles de euros por metro cuadrado. Se supone normalidad.

4.01 3.87 4.68 2.83 3.88 4.92 4.46 5.64 4.91 2.35 4.12 1.11

■Ejercicio 4-2: Para el conjunto de datos **vitcap,** perteneciente al paquete **IswR,** estimar la diferencia entre las medias de la variable **vital.capacity** cuando la variable **group** toma los valores 1 y 3. Se supone normalidad.

■Ejercicio 4-3: Con objeto de comparar las varianzas de dos poblaciones normales se han tomado dos muestras de tamaños 6 y 10, obteniéndose para la primera los valores 6, 8, 5, 4, 9, 5 y para la segunda 6, 7, 6, 9, 6, 2, 9, 4, 6, 4. Contrastar al nivel α=0.05 si puede admitirse la igualdad de las varianzas poblacionales.

■Ejercicio 4-4: En una ciudad se implantó un plan para incentivar a los automóviles con dos o más ocupantes. Para ello se observaron 2000 vehículos antes del plan y 1500 después, obteniéndose 655 y 576 automóviles respectivamente con dos o más pasajeros. ¿Indican los datos que el plan consiguió su propósito? Nivel de significación α=0.05. Hallar el valor-p correspondiente a los datos muestrales del problema.

■Ejercicio 4-5: Se quiere estudiar si existen o no diferencias significativas entre tres institutos en relación a las calificaciones obtenidas por sus alumnos en la asignatura de matemáticas. Para ello se seleccionaron al azar 50 alumnos en cada uno de los tres centros, obteniéndose los siguientes resultados:

	Calificaciones		
	0-4	5-7	8-10
Instituto A	17	20	13
Instituto B	20	15	15
Instituto C	25	16	9

¿Existen diferencias significativas entre los tres institutos?