CONTRASTE DE DOS MUESTRAS

Selección de actividades resueltas

© Jose Fco. Martínez Boscá, Arnau Mir Torres, Lluís M. Pla Aragonés, Àngel J. Gil Estallo (Autors) & Ángel A. Juan (Editor)

UOC Universitat Oberta de Catalunya

Introducción

En este *módulo*, se pretende calcular e interpretar aquellos contrastes sobre la diferencia de medias y la diferencia de proporciones para dos poblaciones, que permita tomar decisiones acerca de qué población hay que tener en cuenta en comparación con la otra.

Además de calcular intervalos de confianza (rango de valores dentro del que se espera encontrar un determinado parámetro de la población), se realizará lo que llamaremos prueba de hipótesis acerca de una afirmación sobre un parámetro de la población. Para poner de manifiesto sus aplicaciones en la vida real, pondremos ejemplos de actividades en el ámbito económico-empresarial y en el informático.

Hasta ahora, habíamos utilizado una sola muestra aleatoria, comparando su media con un valor supuesto de la media poblacional, es decir, nos planteábamos si era posible que muestra con una media dada pudiera provenir de una población la media propuesta. En este caso, extenderemos la idea anterior a dos muestras, preguntándonos si las medias de ambas son iguales o no, es decir, el planteamiento será razonar si es posible que las dos medias muestrales puedan provenir de dos poblaciones idénticas.

Por ejemplo, en una empresa informática se desea medir la eficiencia de dos servidores Web. Para ello, miden el tiempo de espera del cliente entre la petición que éste hace y la respuesta que le da el servidor. Se puede pedir:

- Contrastar si la variabilidad del tiempo de espera es más grande en el servidor A que en el B. Tomad $\alpha = 0{,}1$. Hallad el p-valor del contraste.
- Contrastar si podemos considerar que el servidor A es menos eficiente que el servidor B. Tomad $\alpha = 0{,}01$. Hallad el p-valor del contraste.
- Hallar un intervalo de confianza para la diferencia de tiempos de espera entre ambos servidores al 99% de confianza.
- Diremos que el tiempo de espera es aceptable si es menor que 9 milisegundos. ¿Podemos decir que la proporción de peticiones con tiempo de espera aceptable es distinta para los dos servidores? Tomad $\alpha = 0{,}05$. Hallad el p-valor del contraste.

Diferencia entre muestras independientes y dependientes

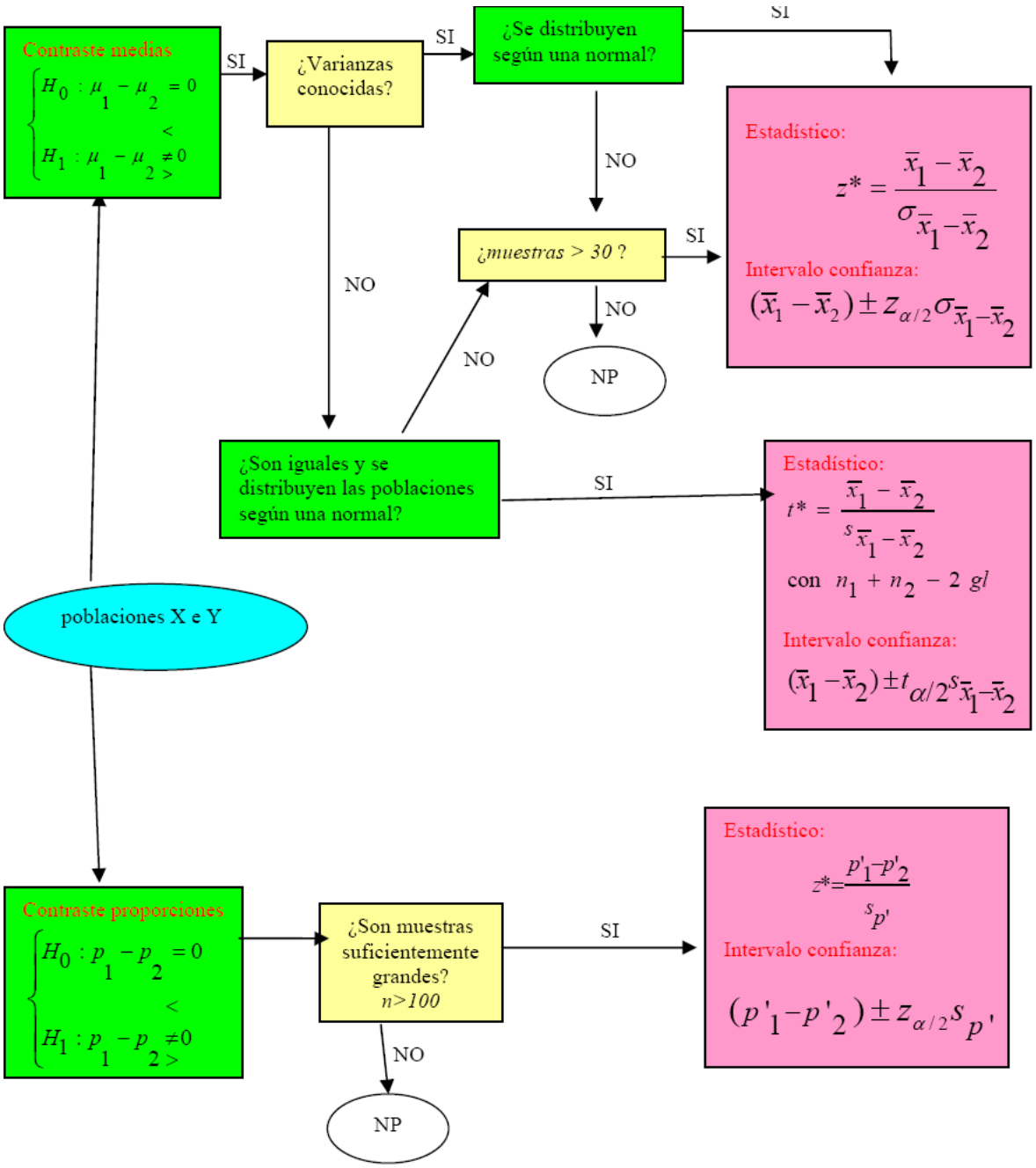
Dos muestras son independientes o dependientes entre sí, en función de si las observaciones de las muestras se han obtenido de los mismos individuos u objetos o no. Si ambas muestras se obtienen de distintos individuos, máquinas, empresas, objetos, etc…no hay nada en común en dichas muestras lo que hace que ambas sean ***"independientes".*** Sin embargo, si las observaciones o valores de ambas muestras se obtienen de los mismos individuos, empresas, agentes, etc., diremos que hay algo en común en dichas muestras por lo que serán muestras ***"dependientes"*** **o** ***"no independientes".***

Supongamos que queremos comparar los beneficios empresariales del sector de las telecomunicaciones entre el año 2001 y el año 2002. Para ello podemos tomar una muestra aleatoria formada por 50 empresas de telecomunicaciones de todo el país y medimos sus beneficios en el año 2001. A continuación, para poder comparar los beneficios del sector con el año 2002, se toma otra muestra aleatoria distinta con otras 30 empresas de telecomunicaciones y analizamos sus beneficios en el año 2002. En este caso se trata de muestras *"independientes"* puesto que las observaciones de ambas muestras se toman de distintos individuos, en este caso distintas empresas. Sin embargo, si en el año 2002 observamos los beneficios de las mismas 50 empresas de telecomunicaciones de la muestra del año 2001, estaríamos por tanto ante muestras *"dependientes"*, o *"aparejadas".*

Aunque el contraste sobre varianzas es del módulo siguiente haremos alguno para saber si son iguales o diferentes. Así podemos decidir el tipo de contraste que haremos sobre las medias y el estadístico correspondiente.

Mapa conceptual

CONTRASTE DE DOS MUESTRAS



NP significa que tenemos que hacer servir métodos. No Paramétricos (fuera del contenido del curso)

Actividad 1: En una empresa informática se desea medir la eficiencia de dos servidores Web. Para ello, miden el tiempo de espera del cliente entre la petición que éste hace y la respuesta que le da el servidor. Los tiempos de espera (en milisegundos) de ambos servidores (TA y TB) para 50 peticiones son:

TA	TB
9,67	6,45
9,62	9,64
9,50	8,53
10,88	9,20
8,94	4,55
10,59	8,51
9,81	12,11
9,46	7,65
9,26	8,85
9,02	8,45
8,61	8,80
9,42	8,82
10,86	9,85
10,01	6,94
10,55	10,47
11,26	8,47
10,64	7,42
10,23	7,48
11,63	11,01
8,91	9,56
10,27	6,80
9,49	8,99
8,99	7,48
10,09	12,57
9,11	7,97
9,47	8,62
8,08	12,11
9,98	12,55
10,30	7,98
7,05	10,20
11,79	11,28
9,59	6,53
10,88	8,14
9,83	8,99
10,92	10,01
10,98	8,14
9,54	9,69
10,17	7,03
10,32	8,59
10,01	10,31
9,96	10,83
9,28	8,41
10,30	9,15
11,08	7,06
10,05	8,04
9,74	11,70
11,14	10,56
9,44	7,82
9,17	6,01
10,86	8,82

a) Contrastad si podemos considerar que el servidor A es menos eficiente que el servidor B. Tomad $\alpha = 0,01$. Hallad el p-valor del contraste.

b) Hallad un intervalo de confianza para la diferencia de tiempos de espera entre ambos servidores al 99% de confianza. Suponed que las varianzas son iguales.
c) Diremos que el tiempo de espera es aceptable si es menor que 9 milisegundos. ¿Podemos decir que la proporción de peticiones con tiempo de espera aceptable es distinta para los dos servidores? Tomad $\alpha = 0,05$. Hallad el p-valor del contraste. Indicación: Para hacer este problema, calculad una nueva variable con R que valga 1 si el tiempo de espera es menor que 9 milisegundos y 0 en caso contrario. Para calcular esta variable, podemos utilizar las instrucciones "for" e "if" de R.

Solución

Apartado a)

1) Hemos de hacer un contraste para comparar dos medias. Dado que el enunciado nos pregunta "si el servidor A es menos eficiente que el servidor B", considerando que un servidor es menos eficiente si es más lento, entonces hemos de contrastar si la media del tiempo de espera del servidor A es más grande que la media del tiempo de espera del servidor B. Así pues, tenemos que considerar una hipótesis alternativa unilateral.

   Hemos de considerar el caso de no normalidad pero con muestras grandes (superior a 30 observaciones).

2) Las hipótesis nula y alternativa son: $H_0: \mu_A - \mu_B = 0$ $H_1: \mu_A - \mu_B > 0$

3) Fijamos $\alpha = 0,01$.

4) El estadístico de contraste es: . La distribución de es la de N(0,1).
Para resolver este apartado con R hacemos lo siguiente:

Para importar los datos a R haremos lo siguiente (si los datos están en un fichero servidors.txt, dónde indicamos que el separador decimal es la ,):

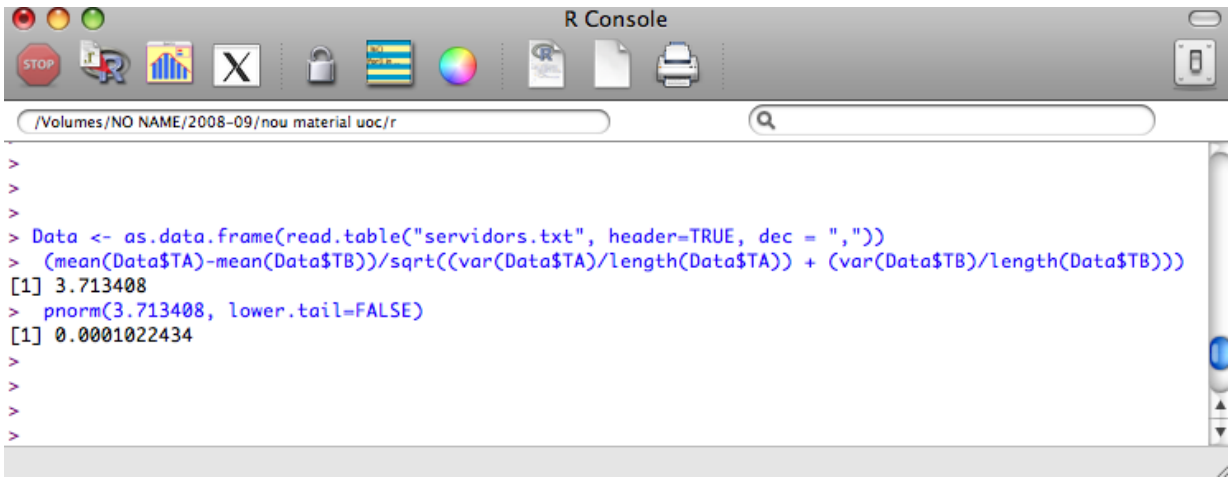
```
> Data <- as.data.frame(read.table("servidors.txt", header=TRUE,
dec = ","))
```

```
> (mean(Data$TA)-mean(Data$TB))/sqrt((var(Data$TA)/length(Data$TA)) +
(var(Data$TB)/length(Data$TB)))
[1] 3.713408
```

p-valor = P(Z > 3.713)

```
> pnorm(3.713408, lower.tail=FALSE)
[1] 0.0001022434
```

Este es directamente el aspecto de la *Rconsola:*



5) Tenemos que el p-valor es 0,000 de manera que rechazamos la hipótesis nula en favor de la hipótesis alternativa y concluimos que las medias de tiempo de espera son diferentes en cada servidor.

Como el p-valor es más pequeño que el nivel de significación, rechazamos la hipótesis nula y damos por buena la alternativa; el servidor A es menos eficiente que el servidor B.

Apartado b)

Tenemos que hacer un contraste bilateral:

```
> t.test(Data$TA,Data$TB,var.equal=TRUE,mu=0,conf.level=0.99)
Two Sample t-test
data: Data$TA and Data$TB
t = 3.7134, df = 98, p-value = 0.0003399
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
0.3020033 1.7623967
sample estimates:
mean of x mean of y
9.9350 8.9028
```

Por tanto, el intervalo es (0.302 , 1.762).

Apartado c)

Lo primero que se tiene que hacere es calcular las dos nuevas variables (una para cada servidor):

Para el servidor A:

```
> for (i in 1:length(Data$TA)) {
+ if(Data$TA[i]<9) Data$PA[i]<-1
+ else Data$PA[i]<-0
+ }
> Data$PA
[1] 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0
0 0 0 0 0
[39] 0 0 0 0 0 0 0 0 0 0 0 0
```

Para el servidor B:

```
> for (i in 1:length(Data$TB)) {
+ if(Data$TB[i]<9) Data$PB[i]<-1
+ else Data$PB[i]<-0
+ }
> Data$PB
[1] 1 0 1 0 1 1 0 1 1 1 1 1 0 1 0 1 1 1 0 0 1 1 1 0 1 1 0 0 1 0 0 1 1
1 0 1 0 1
[39] 1 0 0 1 0 1 1 0 0 1 1 1
```

Para saber la proporción se puede hacer lo siguiente, ya que la proporción se puede ver como la media en una población binaria:

```
> summary(Data$PA)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 0.00 0.00 0.12 0.00 1.00
```

```
> summary(Data$PB)
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 0.00 1.00 0.62 1.00 1.00
```

O hacerlo directamente:

```
> a<-0
> for (i in 1:length(Data$PA)) {
+ if(Data$PA[i]==1) a<-a+1
+ }
> a/length(Data$PA)
[1] 0.12
> b<-0
> for (i in 1:length(Data$PB)) {
+ if(Data$PB[i]==1) b<-b+1
+ }
> b/length(Data$PB)
[1] 0.62
```

Las dos proporciones muestrales: 0,12 para el servidor A y 0,62 para el servidor B.

Tendremos que hacer un contraste bilateral sobre diferencia de proporciones con nivel de significación 0,05. Las hipótesis son:

H0: p0 = p1
H0: p0 != p1

Primero calculamos la estimación de la proporción poblacional común:

```
> pcom<-(length(Data$TA)*0.12 +length(Data$TB)*0.62)/(length(Data$TA)
+length(Data$TB))
> pcom
[1] 0.37
```

Ahora el error estándard:

```
> sp <- sqrt(pcom*(1-pcom)*(1/length(Data$TA) + 1/length(Data$TB)))
> sp
[1] 0.09656086
```

I finalmente el estadístico de contraste:

```
> z <- (0.12-0.62)/sp
> z
[1] -5.178081
```

Cuyo p-valor es p = 2·P(Z>|z|):

```
> 2*pnorm(5.178081,lower.tail=FALSE)
[1] 2.241799e-07
```

Como el p-valor 0.0000 es menor que el nivel de significación (0.05), rechazamos la hipótesis nula y concluimos que **la proporción de peticiones con tiempos de espera aceptable es diferente para los dos servidores.**

Actividad 2: En una empresa se utilizan para hacer las copias de seguridad cintas magnéticas de dos tipos diferentes A y B. La calidad de las cintas depende del número de partículas magnéticas (NPM) por micra cuadrada en la superficie de la cinta, de forma que

cuántas más partículas, más calidad tiene la cinta. Para estudiar la calidad de las cintas, se han tomado algunas medidas del NPM de 10 cintas de tipos A y 7 cintas de tipos B y se obtuvo: $\bar{x}_A = 122,23$, $\bar{x}_B = 124,2857$, $s_A^2 = 337,938$ y , $s_B^2 = 472,231$ suponiendo que las varianzas del NPM de los dos tipos de cintas son iguales, ¿podemos pensar que los dos tipos de cintas tienen, en promedio, la misma calidad? (Tomad α=0,05). Encontrad el p-valor del contraste.

Solución

Debemos hacer un contraste de igualdad de medias de dos muestras con varianzas iguales. El contraste de hipótesis es: $H_0 : m_A = m_B$, $H_1 : m_A \neq m_B$.

El estadístico de contraste vale: $t = \frac{\bar{x}_A - \bar{x}_B}{s\sqrt{\frac{1}{10} + \frac{1}{7}}}$, dónde $\bar{x}_A$ y $\bar{x}_B$ son las medias de los dos tipos de cintas que valen $\bar{x}_A = 122.23, \bar{x}_B = 124.2857$, y s es la desviación típica común $s = \sqrt{\frac{9 \cdot s_A^2 + 6 \cdot s_B^2}{15}} \approx 19.79$.

Los cálculos son:

```
> scom<-sqrt((9*337.938+6*472.231)/15)
> scom
[1] 19.79028
> t<-(122.23-124.2857)/(scom*sqrt((1/10)+(1/7)))
> t
[1] -0.2107814
```

El p-valor es:

```
> 2*pt(-0.2107814,df=15)
[1] 0.8358945
```

Y el valor crítico:

```
> qt(0.025,df=15,lower.tail=FALSE)
[1] 2.131450
```

y por lo tanto aceptamos la hipótesis nula y concluimos que los dos tipos de cintas tienen, en promedio la misma calidad.

Actividad 3: Queremos comparar la rapidez de dos impresoras al imprimir una fotografía. Para lo cual escogemos 10 fotos al azar y las imprimimos en las dos impresoras, observando cuánto tiempo tarda cada una de ellas en imprimir cada foto. Los resultados obtenidos aparecen en la siguiente tabla:

Impresora A	47,2	72,54	59	46,0	74,1	72,5	46,9	78,6	42	63,3
Impresora B	46,1	82,8	63,6	43,9	76,4	82,5	50,4	77,2	40	69,8

Suponiendo que el tiempo de impresión de una foto se distribuye normalmente y que las varianzas poblacionales son desconocidas e iguales.

a) Calculad un intervalo de confianza para la media de cada una de las dos poblaciones al nivel de confianza del 90%. Comentar los resultados.

b) Calculad un intervalo de confianza para la diferencia de medias. Utilizando este intervalo contrastad la hipótesis de que las medias en los dos grupos no son diferentes (nivel de confianza del 90%).
c) Plantead las hipótesis a contrastar para la diferencia de medias y la fórmula del estadístico del contraste así como su valor y su distribución.
d) A partir del criterio del p-valor, ¿a qué conclusión se llega? ¿Qué error de equivocar-nos tendríamos que estar dispuestos a asumir?
e) ¿Cuál habría sido el resultado del contraste si suponemos que, en lugar de utilizar las mismas fotos para el estudio, se han utilizado 10 para una impresora y otras 10 diferentes para la otra? Entonces, plantea las hipótesis a contrastar sobre la igualdad de las medias y la conclusión a la que se llegaría con el criterio del p-valor. Compara los resultados con los del apartado anterior

Solución:

Introducimos los datos en la hoja de trabajo de R

a) Dado que desconocemos la varianza de la población, utilizaremos la t-student. Seleccionamos ***Estadísticos/Medias/Test t para una muestra***. El resultado es el siguiente:

```
> with(impresoras, (t.test(A, alternative='two.sided', mu=0.0, conf.level=.90)))

One Sample t-test
data: A
t = 13.753, df = 9, p-value = 2.391e-07
alternative hypothesis: true mean is not equal to 0
90 percent confidence intervalo:
52.18832 68.23968
sample estimates:
mean of x
60.214

> with(impresoras, (t.test(B, alternative='two.sided', mu=0.0, conf.level=.90)))

One Sample t-test
data: B
t = 11.916, df = 9, p-value = 8.177e-07
alternative hypothesis: true mean is not equal to 0
90 percent confidence intervalo:
53.53638 73.00362
sample estimates:
mean of x
63.27
```

Si nos fijamos en los dos intervalos de confianza, estos se solapan. Además, las medias muestrales son también muy parecidas. Esto hace pensar que estas medias poblacionales, es decir, las medias de las impresoras, pueden ser iguales.

b) Para calcular un intervalo de confianza para la diferencia de medias con datos emparejados no hay que apilar los datos, seleccionamos directamente ***Estadísticos / Medias / t test para datos apareados*:**
El resultado es el siguiente:

```
> with(impresoras, (t.test(A, B, alternative='two.sided', conf.level=.95, paired=TRUE)))

Paired t-test
Data: A and B
t = -2.0278, df = 9, p-value = 0.07319
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence intervalo:
-5.8185811 -0.2934189
```

sample estimates:
mean of the differences
-3.056

El valor cero no está en el intervalo de confianza, entonces podemos concluir que las medias son diferentes. También nos podemos fijar en el p-valor que es 0.074< 0.1, por lo tanto rechazamos la H0

c) Se realiza un contraste de medias en dos muestras emparejadas:

$$\left.\begin{array}{l} H_0: d_{A-B} = 0 \\ H_1: d_{A-B} \neq 0 \end{array}\right\}$$

con *dA-B* representando la diferencia media del tiempo de impresión de una foto utilizando una impresora y la otra .
El enunciado dice que las varianzas poblacionales son desconocidas pero iguales.
El estadístico de contraste:

$$t = \frac{\bar{x}_{A-B}}{\frac{s}{\sqrt{n}}},$$

es una t de Student con 9 grados de libertad, donde *s* es la desviación típica de las diferencias. El valor del estadístico de contraste para este problema vale *t* = -2.028.

d) El p-valor vale 0.074, como es más pequeño que 0,10 rechazamos la hipótesis nula, por lo tanto aceptamos que las medias no son iguales. El error de equivocarnos al rechazar la hipótesis de igualdad de medias es de 0.074.

e) Para hacer un contraste para la diferencia de medias con dos muestras independientes, seleccionamos ***Datos / Conjunto de datos activo / Apilar variables del conjunto de datos activo*** y seleccionamos ***Estadísticos / Medias / CH para dos muestras independientes*:**

Lo primero que haremos es crear una variable apilada, es decir, de 2 variables que tenemos, crearemos una sola variable con 10+10=20 observaciones y otra variable que es el tipo de impresora (A, B). El conjunto de datos es:

Apilar

	variable	Impresoras
1	47.20	A
2	72.54	A
3	59.00	A
4	46.00	A
5	74.10	A
6	72.50	A
7	46.90	A
8	78.60	A
9	42.00	A
10	63.30	A
11	46.10	B
12	82.80	B
13	63.60	B
14	43.90	B
15	76.40	B
16	82.50	B
17	50.40	B
18	77.20	B
19	40.00	B
20	69.80	B

El resultado es el siguiente:

```
> Apilar <- stack(impresoras[, c("A","B")])
> names(Apilar) <- c("variable", "Impresoras")
> t.test(variableImpresoras, alternative='two.sided', conf.level=.90, var.equal=TRUE,
data=Apilar)
```

TTwo Sample t-test

Data: variable by Impresoras
t = -0.44405, df = 18, p-value = 0.6623
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence intervalo:
-14.989996 8.877996
sample estimates:
mean in group A mean in group B
60.214 63.270

El valor cero está en el intervalo de confianza, entonces podemos concluir que las medias no son diferentes. También nos podemos fijar en el p-valor que es muy superior al nivel de significación, 10%, por tanto aceptamos la H0.

Las hipótesis del contraste son:
$H0{:}\mu A - \mu B{=}0$
$H1{:}\mu A - \mu B = 0$

Donde tenemos el tiempo medio de impresión de una foto utilizando la impresora A y el tiempo medio de impresión de una foto utilizando la impresora B.
El estadístico de contraste es

$$t = \frac{\bar{x}_A - \bar{x}_B}{s\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}},$$

con una distribución t de Student con 18 grados de libertad, donde *s* es la desviación típica común:

$$s = \sqrt{\frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{n_A+n_B-2}}.$$

El valor del estadístico de contraste per este problema vale *t*=-0.444. El p-valor vale 0.662, como es más grande que 0,10 no podemos rechazar la hipótesis nula, por tanto aceptamos la igualdad de medias. Así la conclusión no es la misma que la obtenida con datos emparejados, de hecho es totalmente diferente, antes hemos rechazado la hipótesis nula y ahora la hemos aceptado. También podemos observar que el error de equivocar nos al rechazar la hipótesis de igualdad de medias (diferencias igual a 0) es de 0.66 2 muy superior al 0.074 anterior.

Actividad 4: Consideramos el fichero de R "vendes_pac1_P_15_3". Importar este fichero en R teniendo en cuenta que el separador decimal es la coma "," y el de campos es el ";". Recordar que las variables del fichero son:

- m2: superficie
- Ubi: ubicación (1 Centro ciudad, 2 Centro comercial, 3 Calle peatonal, 4 Barrios, 5 Extrarradio)
- PreuAm2: precio del alquiler por m2 antes de hacer reformas
- PreuDm2: precio del alquiler por m2 después de hacer reformas
- AugmentFact: Aumento de la facturación durante el último año

a) Definir las variables "Precio del alquiler de los pisos antes de hacer reformas con menos de 300 m2" y "Precio del alquiler de los pisos antes de hacer reforma con más de 300 m2". Llamadlas *Precio.Alquiler.Antes.Menos.300* y *Precio.Alquiler.Antes.Mas.300* respectivamente.

b) Contrastar a un nivel de confianza del 95% si el precio del alquiler antes de hacer reformas de los pisos con menos de 300 m2 es mayor que el precio del alquiler antes de hacer reforma de los pisos con más de 300 m2. Suponer que las variables consideradas son normales con la misma varianza.

c) Contrastar si el precio del alquiler de los pisos después de hacer reformas es el mismo que el precio de alquiler antes de hacerlas.

d) Hallar un intervalo de confianza al 90% de confianza de la diferencia del precio de los pisos antes y después de las reformas. Según este último intervalo, ¿diríais que el precio de los pisos ha subido o no después de hacer reformas? Razonar la respuesta.

Solución

Primero cargamos los datos:

a) Las variables serán las siguientes:

```
Precio.Alquiler.Antes.Menos.300 = ventas$PreuAm2[ventas$m2 <300]
Precio.Alquiler.Antes.Mas.300 = ventas$PreuAm2[ventas$m2 > 300]
```

b) El contraste es el siguiente:

```
Precio.Alquiler.Antes.Menos.300 = ventas$PreuAm2[ventas$m2 <300]
Precio.Alquiler.Antes.Mas.300 = ventas$PreuAm2[ventas$m2 > 300]
t.test(Precio.Alquiler.Antes.Menos.300,Precio.Alquiler.Antes.Mas
.300,
       alternative = "greater",var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  Precio.Alquiler.Antes.Menos.300 and
Precio.Alquiler.Antes.Mas.300
## t = 0.85844, df = 97, p-value = 0.1964
## alternative hypothesis: true difference in means is greater
than 0
## 95 percent confidence interval:
##  -0.3814182        Inf
## sample estimates:
## mean of x mean of y
##  11.80435  11.39623
```

Cómo que el p-valor es mayor que el nivel de significación 0.05, concluimos que no tenemos suficientes indicios para afirmar que el precio antes de hacer reformas es mayor para los pisos con menos de 300 m2.

c) El constraste pedido es:

```
Precio.Alquiler.Antes = ventas$PreuAm2
Precio.Alquiler.Despues = ventas$PreuDm2
t.test(Precio.Alquiler.Antes,Precio.Alquiler.Despues,paired =
TRUE)

##
##  Paired t-test
##
## data:  Precio.Alquiler.Antes and Precio.Alquiler.Despues
## t = 0.66342, df = 98, p-value = 0.5086
## alternative hypothesis: true difference in means is not equal
to 0
## 95 percent confidence interval:
##  -0.1810264  0.3628446
## sample estimates:
## mean of the differences
##              0.09090909
```

El p-valor también es más grande que el nivel de significación. Por lo tanto, tampoco tenemos suficientes indicios para afirmar que el precio después de hacer reformas es mayor para los pisos con menos de 300 m2.

d) El intervalo pedido es:

```
Dif.precio=ventas$PreuDm2 - ventas$PreuAm2
t.test(Dif.precio)

##
##  One Sample t-test
##
## data:  Dif.precio
## t = -0.66342, df = 98, p-value = 0.5086
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.3628446  0.1810264
## sample estimates:
##   mean of x
## -0.09090909
```

Cómo el valor 0 está dentro del intervalo, podemos decir que no tenemos suficientes indicios para poder afirmar que el precio de los pisos ha subido.

> t.test(Dif.precio, conf.level=.90)

One Sample t-test

data: Dif.precio

t = -0.66342, df = 98, p-value = 0.5086

alternative hypothesis: true mean is not equal to 0

90 percent confidence interval:

-0.3184577 0.1366395

sample estimates:

mean of x

-0.09090909

Actividad 5: Consideremos el fichero de R "vendes_pac1_P_15_4". Importar este fichero en R teniendo en cuenta que el separador decimal es la coma "," y el de campos es ";". Recordar que las variables del fichero son:

- m2: superficie
- Ubi: ubicación (1 Centro ciudad, 2 Centro comercial, 3 Calle peatonal, 4 Barrios, 5 Extrarradio)
- PreuAm2: precio del alquiler por m2 antes de hacer reformas
- PreuDm2: precio del alquiler por m2 después de hacer reformas
- AugmentFact: Aumento de la facturación durante el último año

a) Definir las variables "Precio del alquiler de los locales comerciales ubicados en el centro de la ciudad antes de hacer reformas" y "Precio del alquiler de los locales comerciales ubicados en los barrios antes de hacer reformas". Llamadlas *Precio.Alquiler.Antes.Centro* y *Precio.Alquiler.Antes.Barrios*, respectivamente.

b) Contrastar a un nivel de confianza del 95% si el precio del alquiler antes de hacer reformas de los locales comerciales situados en el centro de la ciudad es mayor

que el precio del alquiler antes de hacer reforma de los locales comerciales situados en los barrios. Suponer que las variables consideradas son normales con la misma varianza.

c) Hacer lo mismo que en los apartados a) y b) pero ahora en lugar de considerar el precio del alquiler antes de las reformas, considerar el precio del alquiler después de las reformas. O sea, contrastar a un nivel de confianza del 95% si el precio del alquiler después de hacer reformas de los locales comerciales situados en el centro de la ciudad es mayor que el precio del alquiler después de hacer reforma de los locales comerciales situados en los barrios. Suponer que las variables consideradas son normales con la misma varianza. Llamad a las nuevas variables *Precio.Alquiler.Después.Centro* y *Precio.Alquiler.Después.Barrios*.

d) Hallar un intervalo de confianza al 95% de confianza de la diferencia del precio de los locales comerciales situados en el centro de la ciudad antes y después de las reformas. Según este último intervalo, ¿diríais que el precio de los locales comerciales ha subido o no después de hacer reformas en los locales comerciales situados en el centro de la ciudad? Razonar la respuesta.

Solución

Primero cargamos los datos:

```
ventas=read.csv("vendes_pac1_P_15_4.csv",sep=";",dec=",")
```

a) Las variables serán las siguientes:

```
Precio.Alquiler.Antes.Centro = ventas$PreuAm2[ventas$Ubi==1]
Precio.Alquiler.Antes.Barrios = ventas$PreuAm2[ventas$Ubi==4]
```

b) El contraste es el siguiente:

```
t.test(Precio.Alquiler.Antes.Centro,Precio.Alquiler.Antes.Barrios,
       alternative = "greater",var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  Precio.Alquiler.Antes.Centro and
Precio.Alquiler.Antes.Barrios
## t = 0.54363, df = 37, p-value = 0.295
## alternative hypothesis: true difference in means is greater
than 0
## 95 percent confidence interval:
##  -1.004955       Inf
## sample estimates:
## mean of x mean of y
##  11.36667  10.88889
```

Cómo el p-valor es mayor que el nivel de significación 0.05, concluimos que no tenemos suficientes indicios para afirmar que el precio del alquiler de los locales comerciales antes de hacer reformas situados al centro de la ciudad sea mayor que el precio del alquiler de los locales comerciales antes de hacer reformas situados en los barrios.

c) Las variables seran las siguientes:

```
Precio.Alquiler.Después.Centro = ventas$PreuDm2[ventas$Ubi==1]
Precio.Alquiler.Después.Barrios = ventas$PreuDm2[ventas$Ubi==4]
```

El contraste es el siguiente:

```
t.test(Precio.Alquiler.Después.Centro,Precio.Alquiler.Después.Barrios,
       alternative = "greater",var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Precio.Alquiler.Después.Centro and
Precio.Alquiler.Después.Barrios
## t = 1.1854, df = 37, p-value = 0.1217
## alternative hypothesis: true difference in means is greater
than 0
## 95 percent confidence interval:
##  -0.4890769        Inf
## sample estimates:
## mean of x mean of y
##  11.60000  10.44444
```

Cómo el p-valor es mayor que el nivel de significación 0.05, concluimos que no tenemos suficientes indicios para afirmar que el precio del alquiler de los locales comerciales después de hacer reformas situados en el centro de la ciudad sea mayor que el precio del alquiler de los locales comerciales después de hacer reformas situados en los barrios.

d) El intervalo pedido es:

```
t.test(Precio.Alquiler.Antes.Centro,Precio.Alquiler.Después.Cent
ro,conf.level = 0.95,paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  Precio.Alquiler.Antes.Centro and
Precio.Alquiler.Después.Centro
## t = -0.89323, df = 29, p-value = 0.3791
## alternative hypothesis: true difference in means is not equal
to 0
## 95 percent confidence interval:
##  -0.7675947  0.3009281
## sample estimates:
## mean of the differences
##              -0.2333333
```

Cómo el valor 0 está dentro del intervalo, podemos decir que no tenemos suficientes indicios para poder afirmar que el precio de los locales comerciales situados en el centro de la ciudad haya cambiado.

Direcciones de interés

http://fltbw2.rug.ac.be/iloapp/Applets/Ap6b.html
Applet interactivo de constraste de hipótesis con dos muestras.

http://e-stadistica.bio.ucm.es/mod_contraste/contraste_applet.html
Applet sobre contraste de hipótesis para muestras independientes.

http://ftp.medprev.uma.es/libro/node126.htm
Éste texto es la versión electrónica del manual de la Universidad de Málaga y habla sobre el contraste de hipótesis sobre la diferenciad proporciones y medias con dos muestras.

http://kitchen.stat.vt.edu/~sundar/java/applets/
Aplicaciones estadísticas con JAVA.

http://www.udc.es/dep/mate/recursos.html
Selección de recursos en Internet para la enseñanza-aprendizaje de la Estadística.

http://halweb.uc3m.es/esp/Personal/personas/stefan/ESP/applet.htm
Conjunto de applets interactivos de Estadística básica.

http://www.uoc.edu/in3/e-math/docs/CH_2Pob.pdf
Math-block del proyecto e-math sobre contraste de 2 poblaciones con teoría y ejemplos con y sin Minitab.