Análisis de dos muestras

Supongamos el siguiente ejemplo.

La resistencia a la rotura de un componente eléctrico constituye una característica importante de un cierto proceso. Un fabricante utiliza un material nuevo de fabricación frente al material clásico.

Se recoge una muestra de 10 elementos usando el primer componente y otra de 10 elementos usando el segundo componente.

Se pueden considerar a los dos procesos como dos tratamientos o dos niveles diferentes de un factor dado.

Componente Nuevo	**Componente Antiguo**
16.85	17.50
16.40	17.63
13.21	18.25
16.35	18.00
16.52	17.86
17.04	17.75
16.96	18.22
17.15	17.90
16.59	17.96
16.57	18.15

Se tiene que la media muestral del componente nuevo es $\bar{y}_1 = 16{,}76$ y la del componente antiguo es $\bar{y}_2 = 17{,}92$.

Se pretende averiguar si existen diferencias significativas entre ambos tratamientos a nivel de resistencia.

En este caso, se considera que los datos proceden de una *m.a.s* de una distribución normal, y que el diseño es completamente aleatorizado.

El contraste de hipótesis que se tiene que realizar es bilateral:

$$\begin{aligned} H_0 &\equiv \mu_1 = \mu_2 \\ H_1 &\equiv \mu_1 \neq \mu_2 \end{aligned}$$

Fijamos $\alpha = P\{\text{error tipo I}\} = P\{\text{Rechazar } H_0 | H_0 \text{ siendo cierta}\}$

Suponiendo normalidad y suponiendo que $\sigma_1^2 = \sigma_2^2$, se utiliza el estadístico siguiente:

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

donde $\bar{y}_1$ e $\bar{y}_2$ son las medias muestrales, n_1 y n_2 son el tamaño de cada muestra y

$$s_p = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Se compara el valor de este estadístico con el valor de una distribución t de Student $t_{\frac{\alpha}{2};n_1+n_2-2}$.

Si $|t_0| > t_{\frac{\alpha}{2};n_1+n_2-2}$ entonces se rechaza H_0.

Así, si H_0 es verdadera, t_0 se distribuye como una t de Student y es de esperar que un $100(1-\alpha)\,\%$ de los valores de t_0 estén entre $\left(-t_{\frac{\alpha}{2};n_1+n_2-2}; t_{\frac{\alpha}{2};n_1+n_2-2}\right)$. Una muestra concreta que produzca un valor fuera de este intervalo es *rara* si H_0 fuese cierta, lo que lleva a rechazar la hipótesis H_0.

En el ejemplo:

Componente nuevo	**Componente antiguo**
$\bar{y}_1 = 16{,}76$	$\bar{y}_2 = 17{,}92$
$s_1^2 = 0{,}10$	$s_2^2 = 0{,}061$
$s_1 = 0{,}316$	$s_2 = 0{,}247$
$n_1 = 10$	$n_2 = 10$

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} = \\ &= \frac{9 \cdot 0{,}10 + 9 \cdot 0{,}061}{10 + 10 - 2} = 0{,}081 \end{aligned}$$

y $s_p = 0{,}204$.

Así,

$$\begin{aligned} t_0 &= \frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \\ &= \frac{16{,}76 - 17{,}92}{0{,}284\sqrt{\frac{1}{10}+\frac{1}{10}}} = -9{,}13 \end{aligned}$$

Por otro lado, $t_{\frac{\alpha}{2};n_1+n_2-2} = t_{0{,}025;18} = 2{,}10$, y como $|t_0| = 9{,}13 > t_{0{,}025;18} = 2{,}10$, entonces no se acepta H_0, esto es, existen diferencias significativas entre las dos medias al nivel de confianza del 95 %.

También se podría haber considerado un intervalo de confianza, dado que el contraste es bilateral:

$$\left[(\bar{y}_1 - \bar{y}_2) \pm t_{\frac{\alpha}{2};n_1+n_2-2}\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

y ver si el valor 0 se encuentra en su interior.

En el caso en que las varianzas de las poblaciones de las que proceden las dos muestras son diferentes ($\sigma_1^2 \neq \sigma_2^2$) se introduce una modificación en los grados de libertad de la distribución t de Student (corrección de Welch). Así, se considera el estadístico

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

que se compara con una distribución t de Student, $t_{\frac{\alpha}{2},v}$ donde v son los grados de libertad aproximados:

$$v = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

Diseño de comparación por pares (apareado)

En el apartado anterior, se consideraba que los valores de ambas muestras no estaban correlacionados. Sin embargo, existen situaciones en las que un mismo objeto recibe dos tratamientos distintos (e.g. pruebas con dos tipos de medicamentos), de manera que no se pueden considerar observaciones independientes.

En otros casos se puede mejorar el diseño experimental eliminando fuentes de error.

Ejemplo.

Supongamos un instrumento de medida de la dureza de un cierto material (se mide la profundidad de la huella producida por la presión de una punta sobre una probeta). Supongamos que se dispone de dos tipos de puntas distintas y se quiere comprobar si existen o no diferencias entre ellas.

Un posible diseño experimental, sería tomar 20 probetas al azar y probar la mitad de ellas con una punta y la otra mitad con la otra. Se tendría, así, un diseño completamente aleatorizado y se utilizaría una prueba t de Student como en el apartado anterior.

Supongamos que existen diferencias entre las probetas, debidas a la distinta homogeneidad del material o a las diferentes condiciones de fabricación. Esto aumentaría el error de medida, que no sería controlable, y la diferencia entre las puntas podría resultar enmascarada.

Una posible forma de evitarlo sería el siguiente diseño.

Se divide en dos partes a la probeta y se asigna aleatoriamente una punta u otra a cada parte. El modelo estadístico que se considera es el siguiente:

$$y_{ij} = \mu_i + \beta_j + \varepsilon_{ij}$$

para $i = 1, 2$ y $j = 1, 2, \ldots, 10$. Donde

$y_{ij} \equiv$ Resultado de la punta i sobre la probeta j

$\mu_i \equiv$ Dureza media (poblacional) de la punta i.

$\beta_j \equiv$ Efecto sobre la dureza medida de la probeta j.

$\varepsilon_{ij} \equiv$ Error experimental (con media 0 y varianza σ_i^2).

Si se quiere eliminar el efecto no controlable de las diferentes probetas, se pueden considerar las diferencias entre las medidas:

$d_j = y_{1j} - y_{2j}$ para $j = 1, 2, \ldots, 10$.

El valor esperado de la diferencia

$$E(d_j) = E(y_{1j}) - E(y_{2j}) = \left(\mu_1 + \beta_j\right) - \left(\mu_2 + \beta_j\right) = \mu_1 - \mu_2$$

luego, si se denomina $\mu_d = \mu_1 - \mu_2$, entonces la hipótesis $H_0 \equiv \mu_1 = \mu_2$ equivale a $H_0 \equiv \mu_d = 0$.

El contraste que se realiza es

$$\begin{aligned} H_0 &\equiv \mu_d = 0 \\ H_1 &\equiv \mu_d \neq 0 \end{aligned}$$

y se emplea

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}}$$

donde

$$\begin{aligned} \bar{d} &= \frac{1}{n}\sum_{j=1}^{n} d_j \\ s_d^2 &= \frac{1}{n-1}\sum_{j=1}^{n}\left(d_j - \bar{d}\right)^2 \end{aligned}$$

La hipótesis $\mu_d = 0$ se rechaza cuando

$$|t_0| > t_{\frac{\alpha}{2},(n-1)}$$

En el ejemplo:

Punta 1	**Punta 2**	d_j
7	6	1
3	3	0
3	5	-2
4	3	1
8	8	0
3	2	1
2	4	-2
9	9	0
5	4	1
4	5	-1

$$
\begin{aligned}
\bar{d} &= -0{,}10 \\
s_d &= 1{,}20 \\
t_0 &= \frac{\bar{d}}{s_d/\sqrt{n}} = -0{,}26 \\
t_{\frac{\alpha}{2},(n-1)} &= t_{0{,}025,9} = 2{,}262
\end{aligned}
$$

Con lo cual se acepta H_0 ya que $|t_0| < 2{,}262$.

Observaciones.

Si se utiliza un diseño apareado, en realidad se está utilizando un diseño de tipo aleatorizado por *bloques*.

Un bloque es una unidad experimental relativamente homogénea (en el ejemplo anterior eran las probetas). La variabilidad *dentro* de los bloques es menor que *entre* los bloques.

Si se consideran bloques se elimina error experimental, pero se reduce el número de grados de libertad (la sensibilidad de una prueba aumenta cuando aumenta el número de grados de libertad), con lo cual, si la variabilidad dentro de los bloques es igual a la variabilidad entre los mismos no se debería considerar el diseño apareado (bloques).

Comparaciones entre varios grupos.

Supongamos que se tienen más de dos posibles grupos a comparar. La primera idea sería realizar contrastes de la t de Student por pares de grupos. Por ejemplo, si se tienen 5 grupos: 4 tratamientos y un control, podrían plantearse un total de $\binom{5}{2} = 10$ posibles pares de comparaciones. Haciendo esto produce el siguiente problema:

Si la probabilidad de aceptar H_0 correctamente es $(1-\alpha)$, e.g. 0.95, entonces la probabilidad de aceptar correctamente H_0 en las 10 pruebas es $0{,}95^{10} = 0{,}60$ si éstas son independientes. Es decir, aumenta mucho el error de tipo I. Entonces hay que utilizar una metodología diferente: ANOVA.

Aplicación con R

Se puede usar la librería **Rcmdr** de **R**:

```
> library(Rcmdr)
> Commander()
```

Desde allí se pueden importar los datos que han de estar colocados en dos columnas: la primera con los valores y la segunda con un factor (definido como objeto factor) que tome valores 1 y 2 correspondientes a cada grupo.

Después, se va a los menús siguientes:

Statistics -> **Means** -> **Single-Sample t-test**

Alternativamente, se puede hacer lo mismo con R pero mediante sentencias:

```
> CNuevo <- c(16.85, 16.40, 13.21, 16.35, 16.52, 17.04, 16.96, 17.15,
16.59, 16.57)
> CAntiguo <- c(17.50, 17.63, 18.25, 18.00, 17.86, 17.75, 18.22,
17.90, 17.96, 18.15)
```

Se comprueba la igualdad entre las varianzas de ambas muestras:

```
> var.test(CNuevo, CAntiguo)

        F test to compare two variances

data:  CNuevo and CAntiguo
F = 21.2113, num df = 9, denom df = 9, p-value = 0.0001013
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  5.268586 85.396550
sample estimates:
ratio of variances
          21.21130
```

Al ser distintas para cada grupo se toma la opción correspondiente del comando siguiente:

```
> t.test(CNuevo, CAntiguo, paired=F, var.equal=F)

        Welch Two Sample t-test

data:  CNuevo and CAntiguo
t = -4.2167, df = 9.847, p-value = 0.001843
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.3829934 -0.7330066
sample estimates:
mean of x mean of y
   16.364    17.922
```

Se puede hacer un procedimiento *Bootstrap* para las estimas.

```
> # Fijas el numero de iteraciones
>  iter <- 2000
>  n1 <- length(CNuevo)
>  n2 <- length(CAntiguo)
>  d <- NULL
>  for(i in 1:iter) {
+      xtemp <- sample(CNuevo,n1,replace=T)
+      ytemp <- sample(CAntiguo,n2,replace=T)
+      d[i] <- mean(xtemp) - mean(ytemp)
+     }

>  dbar <- mean(d)
>  se <- sqrt(var(d))
>  # Construyes el estadistico
>  z <- dbar/se
>  pvalor <- 2*(1-pnorm(abs(z)))

>  cbind(dbar,se,z,pvalor)

          dbar        se         z      pvalor
[1,] -1.567326 0.3546939 -4.418815 9.92437e-06

>  quantile(d,c(0.025, 0.975))

   2.5%   97.5%
-2.3552 -1.0480
```

Observaciones apareadas

```
> Punta1 <- c(7,3,3,4,8,3,2,9,5,4)
> Punta2 <- c(6,3,5,3,8,2,4,9,4,5)

> t.test(Punta1, Punta2, paired=T)

        Paired t-test

data:  Punta1 and Punta2
t = -0.2641, df = 9, p-value = 0.7976
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9564389  0.7564389
sample estimates:
mean of the differences
                   -0.1
```

Aplicación con SAS

```
options ls=75 nodate nonumber;
title 'Comparación entre dos grupos independientes';
data dgrup;
input medida grupo $;

/* En el caso de que los datos estén en un fichero externo (e.g.
datos.txt) se usa la orden infile:
infile 'c:\...\datos.txt';  */

cards;
16.85 1
16.40 1
13.21 1
16.35 1
16.52 1
17.04 1
16.96 1
17.15 1
16.59 1
16.57 1
17.50 2
17.63 2
18.25 2
18.00 2
17.86 2
17.75 2
18.22 2
17.90 2
17.96 2
18.15 2

proc ttest;
class grupo;
var medida;
run;
```

```
                 Comparación de dos grupos independientes

                           The TTEST Procedure

                               Statistics

                              Lower CL           Upper CL  Lower CL
Variable  grupo            N      Mean     Mean      Mean   Std Dev   Std Dev

medida    1               10    15.547   16.364    17.181    0.7854    1.1418
medida    2               10    17.745   17.922    18.099    0.1705    0.2479
medida    Diff (1-2)            -2.334   -1.558    -0.782    0.6243    0.8262
```

Statistics

Variable	grupo	Upper CL Std Dev	Std Err	Minimum	Maximum
medida	1	2.0845	0.3611	13.21	17.15
medida	2	0.4526	0.0784	17.5	18.25
medida	Diff (1-2)	1.2218	0.3695		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
medida	Pooled	Equal	18	-4.22	0.0005
medida	Satterthwaite	Unequal	9.85	-4.22	0.0018

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
medida	Folded F	9	9	21.21	<.0001

```
options ls=75 nodate nonumber;
title 'Comparación de diferencias entre dos grupos ';
data df2g;
input Punta1 Punta2;
dife = Punta1-Punta2;
cards;
7  6
3  3
3  5
4  3
8  8
3  2
2  4
9  9
5  4
4  5
proc means n mean stderr t prt;
var dife;
run;
```

```
              Comparación de diferencias entre dos grupos

                          The MEANS Procedure

                       Analysis Variable : dife

     N           Mean       Std Error     t Value     Pr > |t|
  ------------------------------------------------------------
    10     -0.1000000       0.3785939       -0.26       0.7976
  ------------------------------------------------------------
```