Inferencia Estadística

Profesor(es): Jarnishs Beltran
Ayudante: Pablo Rivera

Pauta ayudantía Nº9
Primavera 2020

Ejercicio en R

i) Un equipo de atletismo ha decidido contratar a un nuevo entrenador. Para decidir si al cabo de un año mantienen su contrato se selecciona aleatoriamente a 10 miembros del equipo y se cronometran sus tiempos en 100 metros lisos al inicio del año, al final del año se volverá a cronometrar a esos mismos 10 corredores. En vista de los datos obtenidos ¿Hay diferencia significativa entre el rendimiento de los corredores tras un año de entrenar con el nuevo instructor?

Los rendimientos fueron los siguientes:

antes:	12.9	13.5	12.8	15.6	17.2	19.2	12.6	15.3	14.4	11.3
después:	12.7	13.6	12.0	15.2	16.8	20.0	12.0	15.9	16.0	11.1

Realice un análisis con los distintos pasos que éste conlleva.

0.1. Solución Manual

Solution: Se trata de un caso de estudio en el que las mediciones se realizan sobre los mismos individuos bajo dos condiciones distintas, se trata de datos pareados.

```
datos <- data.frame(
          corredor = c(1:10),
          antes = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3),
          despues = c(12.7, 13.6, 12.0, 15.2, 16.8, 20.0, 12.0, 15.9, 16.0, 11.1)
        )
head(datos, 4)
```

	corredor <int>	antes <dbl>	despues <dbl>
1	1	12.9	12.7
2	2	13.5	13.6
3	3	12.8	12.0
4	4	15.6	15.2

4 rows

Al tratarse de datos pareados, interesa conocer la diferencia en cada par de observaciones.

```
diferencia <- datos$antes - datos$despues
datos      <- cbind(datos, diferencia)
head(datos,4)
```

	corredor <int>	antes <dbl>	despues <dbl>	diferencia <dbl>
1	1	12.9	12.7	0.2
2	2	13.5	13.6	-0.1
3	3	12.8	12.0	0.8
4	4	15.6	15.2	0.4

4 rows

```
colMeans(datos[,-1])
```

```
##      antes    despues diferencia
##      14.48      14.53      -0.05
```

1. Establecer las hipótesis

   H_0: no hay diferencia entre el tiempo medio de los corredores al inicio y al final del año. El promedio de las diferencias es cero ($\mu_d = 0$).

   H_a: sí hay diferencia entre el tiempo medio de los corredores al inicio y al final del año. El promedio de las diferencias no es cero ($\mu_d \neq 0$).

2. Establecer el estadístico (parámetro estimado) que se va a emplear

   El estadístico es el valor que se calcula a partir de la muestra y que se quiere extrapolar a la población de origen. En este caso es el promedio de las diferencias entre cada par de observaciones $\bar{d} = 0{,}5$.

3. Determinar el tipo de test, una o dos colas

   Los test de hipótesis pueden ser de una cola o de dos colas. Si la hipótesis alternativa emplea "¿" o "¡" se trata de un test de una cola, en el que solo se analizan desviaciones en un sentido. Si se la hipótesis alternativa es del tipo "diferente de" se trata de un test de dos colas, en el que se analizan posibles desviaciones en las dos direcciones. Solo se emplean test de una cola cuando se sabe con seguridad que las desviaciones de interés son en un sentido y únicamente si se ha determinado antes de observar la muestra, no a posteriori.

   En este caso solo se va a contratar al entrenador si el rendimiento ha mejorado, por lo que la hipótesis alternativa sería más exacta si se considera como ($\mu_d > 0$), es decir, que la media de tiempos al final del año es menor que al inicio.

4. Determinar el nivel de significancia

   $$\alpha = 0{,}05$$

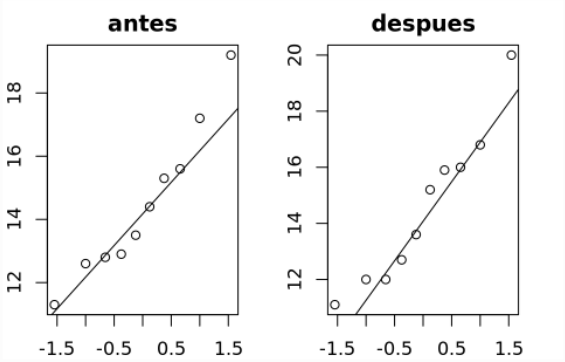
   El nivel de significancia α determina la probabilidad de error que se quiere asumir a la hora de rechazar la hipótesis nula. Se emplea como punto de referencia para determinar si el valor de p-value obtenido en el test de hipótesis es suficientemente bajo como para considerar significativas las diferencias observadas y por lo tanto rechazar H_0. A menor valor de alpha, menor probabilidad de rechazar la hipótesis nula. Por ejemplo, si se considera $\alpha = 0{,}05$, se rechazará la hipótesis nula en favor de la hipótesis alternativa si el p-value obtenido es menor que 0,05, y se tendrá una probabilidad del 5 % de haber rechazado H_0 cuando realmente es cierta. En nivel de significancia debe establecerse en función de que error sea más costoso:

- Error tipo I: Error de rechazar la H_0 cuando realmente es cierta.
- Error tipo II: Error de considerar como cierta H_0 cuando realmente es falsa.

5. Condiciones para comparar dos medias independientes mediante t-test

   Método gráfico:

```
par(mar = c(2, 2, 2, 2))
par(mfrow = c(1, 2))
qqnorm(datos$antes, xlab = "", ylab = "", main = "antes")
qqline(datos$antes)
qqnorm(datos$despues, xlab = "", ylab = "", main = "despues")
qqline(datos$despues)
```



   Método analítico:

```
shapiro.test(datos$antes)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$antes
## W = 0.94444, p-value = 0.6033
```

```
shapiro.test(datos$despues)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos$despues
## W = 0.93638, p-value = 0.5135
```

   Los gráficos *qqnorm* indican que las muestras se asemejan a los esperado en una población normal y los test de Saphiro-Wilk no muestran evidencias para descartar que las muestras procedan poblaciones sean normales (para un alpha de 0,05).

6. Cálculo de p-value

   Parámetro estimado $\bar{d} = mean(diferencia) = -0{,}05$

grados de libertad $= 10 - 1 = 9$

SE(promedio de las diferencias) $= \frac{\hat{S}diferencia}{\sqrt{n}} = \frac{0{,}7412452}{\sqrt{10}} = 0{,}2344023$

- $T_{calc} = \frac{\bar{d}}{SE} = \frac{-0{,}05}{0{,}2344023} = 0{,}2133085$
- $pvalue = P(t_{df} = 9 < 0{,}2133085) + P(t_{df} = 9 > 0{,}2133085)$

```
pt(q = -0.2133085, df = 9) + (1 - pt(q = 0.2133085, df = 9))
```

```
## [1] 0.83584
```

7. Tamaño del efecto

   En el caso particular de los t-test dependientes solo es posible aplicar la d de Choen.

$$d = \frac{|mediadelasdiferencias|}{sd(diferencias)}$$
$$d = \frac{|-0{,}05|}{0{,}7412} = 0{,}068$$

8. Conclusión

   $p-value > \alpha$, no hay evidencias significativas para rechazar H_0 en favor de H_A. No se pude considerar que el rendimiento de los atletas haya cambiado.

0.2. Solución mediante R

R contiene la función $t.test()$ que realiza un t.test con datos pareados si se le indica en el argumento. R calcula automáticamente las diferencias para cada evento, asumiendo que se las posiciones de cada vector se corresponden a los datos de un mismo individuo. Está función calcula además el intervalo de confianza para la diferencia de medias.

```
t.test(x = datos$antes, y = datos$despues, alternative = "two.sided",
       mu = 0, paired = TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  datos$antes and datos$despues
## t = -0.21331, df = 9, p-value = 0.8358
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5802549  0.4802549
## sample estimates:
## mean of the differences
##                   -0.05
```

```r
library(effsize)
cohen.d(d = datos$antes, f = datos$despues, paired = TRUE)
```

```
##
## Cohen's d
##
## d estimate: -0.0169815 (negligible)
## 95 percent confidence interval:
##      lower      upper
## -0.1842481  0.1502851
```