Tarea E.I.

Stephan Paul, Alvaro Zamorano y Braulio Bravo

2024-10-01

```
# Leer datos
datos = read.csv2("EP02 Datos.csv", sep = ";")

# Mostrar datos
head(datos)
```

```
##   Id   Raza Previo Posterior
## 1  1 Blanca 16.274    14.057
## 2  2 Blanca 17.152    13.655
## 3  3 Blanca 15.925    12.826
## 4  4 Blanca 16.814    12.959
## 5  5 Blanca 16.911    14.293
## 6  6 Blanca 17.774    14.786
```

1. El Comité Olímpico cree que el mejor tiempo medio de los atletas de oriental después de ingresar al programa de entrenamiento es superior a 14,9 segundos. ¿Soportan los datos esta afirmación?

```
# Lectura de datos
datos = read.csv2("EP02 Datos.csv", sep=";")
# Mostrar datos
head(datos)
```

```
##   Id   Raza Previo Posterior
## 1  1 Blanca 16.274    14.057
## 2  2 Blanca 17.152    13.655
## 3  3 Blanca 15.925    12.826
## 4  4 Blanca 16.814    12.959
## 5  5 Blanca 16.911    14.293
## 6  6 Blanca 17.774    14.786
```

Hipotesis

Ho: Tiempo medio de los atletas posterior al entrenamiento es igual a 14.9, es decir, mu = 14,9

Ha: Tiempo medio de los atletas posterior al entrenamiento es superior a 14.9, es decir, mu>14,9

```
Oriental = datos %>% filter(Raza == "Oriental")
Tiempo_Posterior_Oriental = Oriental$Posterior

# Prueba Normalidad
shapiro.test(Tiempo_Posterior_Oriental)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Tiempo_Posterior_Oriental
```

```
## W = 0.9689, p-value = 0.573
```

```
# Numero de observaciones
cat("Numero de observaciones: ")
```

```
## Numero de observaciones:
```

```
print(nrow(Oriental))
```

```
## [1] 27
```

```
t.test(Tiempo_Posterior_Oriental, alternative="greater", mu=14.9, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  Tiempo_Posterior_Oriental
## t = -1.1826, df = 26, p-value = 0.8762
## alternative hypothesis: true mean is greater than 14.9
## 95 percent confidence interval:
##  14.37084      Inf
## sample estimates:
## mean of x
##  14.68333
```

Se usa shapiro Wilk para comprobar si los datos se comportan de manera similar a una distribucion normal, y el p-value es muy superior al nivel de significancia, lo que indica que podemos suponer con relativa confianza que la población donde proviene la muestra sigue efectivamente una distribucion normal.

Luego como la muestra de atletas se selecciono de manera aleatoria y por tanto las observaciones son independientes entre si, luego se cumplen las 2 condiciones necesarias para realizar un test t de student. La cual fue realizada debido a que el numero de observaciones es menor a 30 y no se conoce la desviacion estandar de la misma.

Se puede concluir luego de realizar el test que como el p-value es mucho mayor al nivel de significancia (5%), se puede decir que no hay suficiente evidencia para rechazar la hipotesis nula, por lo tanto con un 95% de confianza el tiempo medio de los atletas es igual a 14.9.

2. ¿Sugieren los datos que la mejor marca de los atletas de raza negra se reduce en promedio menos de 1,3 segundos tras el entrenamiento?

Hipotesis

Ho: Diferencia entre los tiempos es igual a 1.3, es decir, x1-x2 >= 1.3

Ha: Diferencia entre los tiempos es menor a 1.3, es decir, x1-x2 < 1.3

con x1 siendo el tiempo previo al entrenamiento y x2 el tiempo posterior a este.

```
# Filtrar por raza negra
negra = datos %>% filter(Raza == "Negra")

# Mostrar datos filtados
head(negra)
```

```
##   Id  Raza Previo Posterior
## 1 27 Negra 13.257    11.704
## 2 28 Negra 14.266    12.216
## 3 29 Negra 12.793    10.994
## 4 30 Negra 13.999    13.017
## 5 31 Negra 11.962    10.686
```

```
## 6 32 Negra 14.055    13.248
```

```
# Prueba Normalidad
diferencia_negra = negra$Previo - negra$Posterior
shapiro.test(diferencia_negra)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diferencia_negra
## W = 0.9781, p-value = 0.8027
```

```
# Numero de observaciones
cat("Numero de observaciones: ")
```

```
## Numero de observaciones:
```

```
print(nrow(negra))
```

```
## [1] 28
```

```
# Test
t.test(diferencia_negra, alternative="less", mu=1.3, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  diferencia_negra
## t = 3.3703, df = 27, p-value = 0.9989
## alternative hypothesis: true mean is less than 1.3
## 95 percent confidence interval:
##  -Inf  1.7
## sample estimates:
## mean of x
##  1.565714
```

Se usa shapiro Wilk para comprobar si las diferencias de los datos se comportan de manera similar a una distribucion normal, y el p-value es muy superior al nivel de significancia, lo que indica que podemos suponer con relativa confianza que la población donde proviene la muestra sigue efectivamente una distribucion normal. Por otro lado, como tenemos que comparar 2 grupos conectados de manera especial (antes y despues del entrenamiento) hay que hacer una prueba pareada y podemos usar la diferencia de los datos.

Luego como la muestra de atletas se selecciono de manera aleatoria y por tanto las observaciones son independientes entre si, luego se cumplen las 2 condiciones necesarias para realizar un test t de student. La cual fue realizada debido a que el numero de observaciones es menor a 30 y no se conoce la desviacion estandar de la misma.

Se puede concluir luego de realizar el test que como el p-value es mucho mayor al nivel de significancia (5%), se puede decir que no hay suficiente evidencia para rechazar la hipotesis nula, por lo tanto con un 95% de confianza la diferencia de los tiempos puede ser igual o mayor a 1.3.

3. ¿Es posible afirmar que, en promedio, los atletas de raza negra superaban a los de raza oriental por más de 5,8 segundos antes del entrenamiento?

Hipotesis

Ho: Diferencia entre la raza negra y la raza oriental es igual a 5.8, es decir, x1-x2 = 5.8

Ha: Diferencia entre la raza negra y la raza oriental es mayor a 5.8, es decir, x1-x2 > 5.8

x1 siendo la raza oriental y x2 la raza negra.

```r
# Prueba Normalidad
Negra_Previo = negra$Previo

Oriental_Previo = Oriental$Previo

shapiro.test(Oriental_Previo)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Oriental_Previo
## W = 0.98354, p-value = 0.932
```

```r
shapiro.test(Negra_Previo)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Negra_Previo
## W = 0.97301, p-value = 0.6631
```

```r
# numero de observaciones es el mismo de las preguntas anteriores
# (ambos grupos de datos tienen menos de 30 observaciones)

# Test
t.test(Negra_Previo, Oriental_Previo, alternative="greater", mu=-5.8, conf.level=0.95, paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Negra_Previo and Oriental_Previo
## t = 0.99025, df = 47.957, p-value = 0.1635
## alternative hypothesis: true difference in means is greater than -5.8
## 95 percent confidence interval:
##  -5.982612       Inf
## sample estimates:
## mean of x mean of y
##  14.11304  19.64981
```

Se usa shapiro Wilk para comprobar si las muestras de los tiempos se comportan de manera similar a una distribucion normal, y el p-value es muy superior al nivel de significancia en ambos casos, lo que indica que podemos suponer con relativa confianza que la población donde provienen las muestras sigue efectivamente una distribucion normal. Por otro lado, como tenemos que comparar 2 grupos independientes hay que hacer una prueba no pareada debido a que los datos provienen de muestras de razas distintas.

Luego como la muestras vienen de la misma seleccion de atletas de las preguntas 1 y 2 las observaciones son independientes entre si, luego se cumplen las 2 condiciones necesarias para realizar un test t de student.

Se puede concluir luego de realizar el test que como el p-value es mucho mayor al nivel de significancia (5%), se puede decir que no hay suficiente evidencia para rechazar la hipotesis nula, por lo tanto con un 95% de confianza la diferencia de los tiempos es igual 5.8.

4. ¿Será cierto que hay más atletas de raza oriental que redujeron sus mejores marcas en al menos 4,8 segundos que atletas de raza blanca que lo hicieron en al menos 3,2 segundos?

Hipotesis

Ho: Diferencia entre numero de atletas de raza oriental y la raza blanca que redujeron sus tiempos en 4.8

segundos y 3.2 segundos respectivamente es igual a 0, es decir, x1-x2 = 0

Ha: Diferencia entre numero de atletas de raza oriental y la raza blanca que redujeron sus tiempos en 4.8 segundos y 3.2 segundos respectivamente es igual a 0, es decir, x1-x2 > 0

con x1 siendo numero de atletas de raza oriental y x2 el numero de atletas de raza blanca.

```
library(dplyr)
datos = read.csv2("EP02 Datos.csv")

valor_nulo = 0

# Datos orientales
atletas_orientales = datos %>% filter(Raza == "Oriental")
diferencia_orientales = abs(atletas_orientales$Previo - atletas_orientales$Posterior)
# Los casos de exito en los orientales son los que redujeron su tiempo en al menos 4,8 segundos
exito_orientales = atletas_orientales  %>% filter(diferencia_orientales >= 4.8)
n_exito_orientales <- nrow(exito_orientales)


# Datos blancos
atletas_blancos = datos %>% filter(Raza == "Blanca")
diferencia_blancos = abs(atletas_blancos$Previo - atletas_blancos$Posterior)
# Los casos de exito en los blancos son los que redujeron su tiempo en al menos 3,2 segundos
exito_blancos = atletas_blancos  %>% filter(diferencia_blancos >= 3.2)
n_exito_blancos <- nrow(exito_blancos)

# comprobar normalidad blancos
shapiro.test(diferencia_blancos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diferencia_blancos
## W = 0.95709, p-value = 0.3375
```

```
total_orientales = nrow(atletas_orientales)
prop_oriental = n_exito_orientales/total_orientales
total_blancos = nrow(atletas_blancos)
prop_blancos = n_exito_blancos/total_blancos

prop_agrupada = (n_exito_blancos + n_exito_orientales) / (total_blancos + total_orientales)

error_orientales = (prop_agrupada * (1-prop_agrupada) ) / total_orientales
error_blanco = (prop_agrupada * (1-prop_agrupada) ) / total_blancos

error_est_hip = sqrt(error_blanco + error_orientales)
z = ((prop_oriental - prop_blancos) - valor_nulo ) / error_est_hip

p = pnorm(z,lower.tail = FALSE)
cat("p-value test de Wald: ")
```

```
## p-value test de Wald:
```

```
cat(p)
```

```
## 0.001969497
```

Se usa shapiro Wilk para comprobar si las muestras de la raza blanca se comportan de manera similar a una distribucion normal, y el p-value es muy superior al nivel de significancia en ambos casos, lo que indica que podemos suponer con relativa confianza que la población donde provienen la muestra sigue efectivamente una distribucion normal, junto con que la raza oriental ya se comprobo su similitud a la normalidad en las preguntas anteriores. Por otro lado, como tenemos que comparar proporciones de 2 grupos distintos usamos un test de Wald para un valor nulo 0.

Luego como la muestras vienen de la misma seleccion de atletas las observaciones son independientes entre si, luego se cumplen las 2 condiciones necesarias para realizar un test de Wald.

Se puede concluir luego de realizar el test que como el p-value es mucho mayor al nivel de significancia (5%), se puede decir que hay suficiente evidencia para rechazar la hipotesis nula en favor de la alternativa, por lo tanto con un 95% de confianza hay más atletas de raza oriental que redujeron sus mejores marcas en al menos 4,8 segundos que atletas de raza blanca que lo hicieron en al menos 3,2 segundos.