EP02_grupo3

Diego Fernandez, Víctor Duarte, Alonso Henriquez

2024-10-01

Obtención de datos:

```
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.4.1

## Cargando paquete requerido: ggplot2
```

```
datos <- read.csv2("EP02 Datos.csv")

RB <- datos %>% filter(Raza == "Blanca")
RN <- datos %>% filter(Raza == "Negra")
RA <- datos %>% filter(Raza == "Oriental")
```

Pregunta 1

1.- El Comité Olímpico cree que el mejor tiempo medio de los atletas de raza blanca después de ingresar al programa de entrenamiento es de 13,6 segundos. ¿Soportan los datos esta afirmación?

Respuesta:

Formulación de hipótesis:

Las hipótesis correspondientes son:

$H_0 : \mu = 13,6$, es decir que la media poblacional es igual a 13,6 [s].

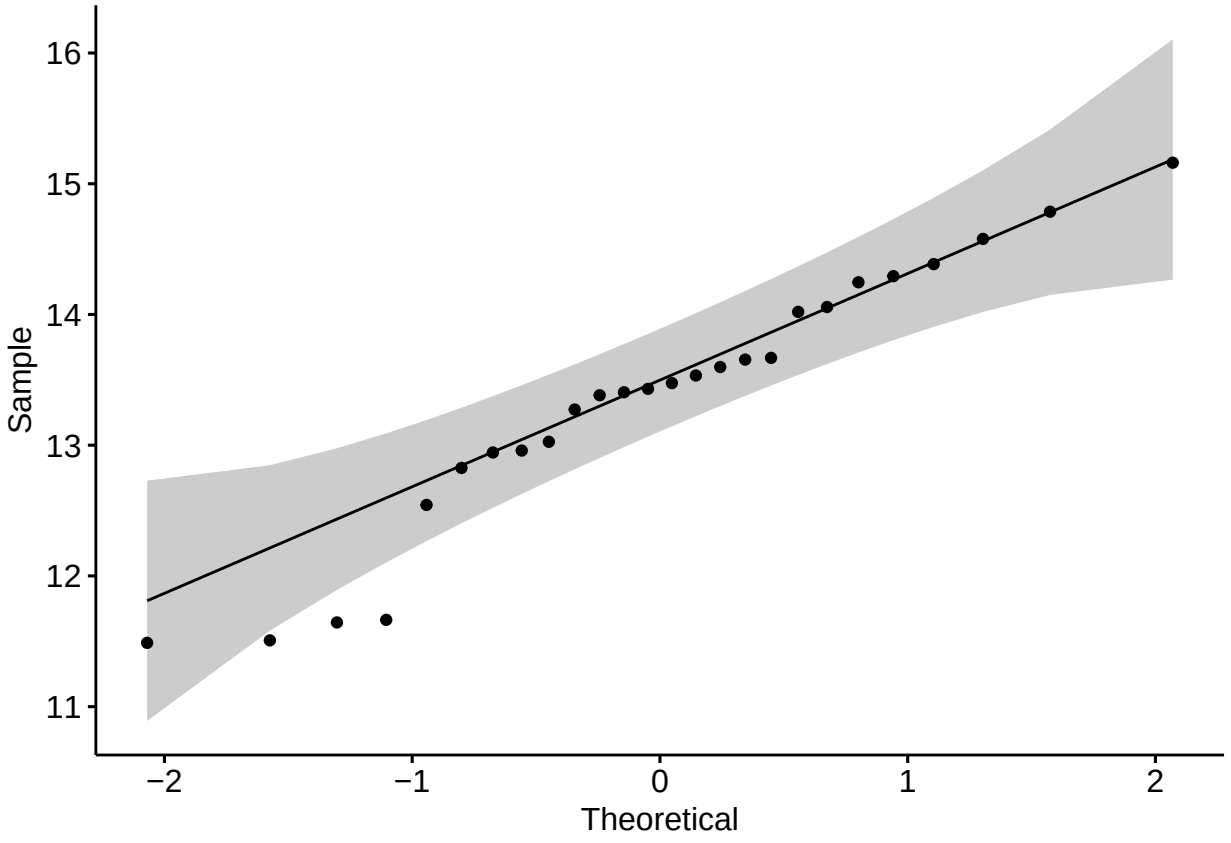
$H_1 : \mu \neq 13.6$ es decur que la media poblacional es distinta de 13,6 [s].

Selección de prueba:

Para esta pregunta se utilizará el t.test de una variable, ya que estamos comparando la media poblacional y el tamaño de la muestra es $n < 30$.

Comprobando condiciones necesarias:

```
ggqqplot(RB$Posterior)
```




```
shapiro.test(RB$Posterior)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  RB$Posterior
## W = 0.94734, p-value = 0.2008
```

Aplicación de la prueba:

```r
t.test(RB$Posterior, mu = 13.6, alternative = "two.sided", conf.level = 1 - 0.05)
```

```
##
##  One Sample t-test
##
## data:  RB$Posterior
## t = -1.1909, df = 25, p-value = 0.2449
## alternative hypothesis: true mean is not equal to 13.6
## 95 percent confidence interval:
##  12.96457 13.76981
## sample estimates:
## mean of x
##  13.36719
```

como $p > 0.05$ se falla en rechazar H_0, entonces, con un 95% de confianza no se puede decir que el promedio de tiempo de los atletas de raza blanca posterior al entrenamiento es distinto de 13,6 [s].

Pregunta 2

2.- ¿Sugieren los datos que la mejor marca de los atletas de raza oriental se reduce en promedio menos de 4,7 segundos tras el entrenamiento? t.test pareada.

Respuesta:

Definiendo variables aleatorias:

X : marcas de tiempo de los atletas de raza oriental previo al entrenamiento.

Y : marcas de tiempo de los atletas de raza oriental posterior al entrenamiento.

$d_0 := X - Y$

Selección de hipótesis:

$H_0 : \mu_{d_0} = 4,7$

$H_1 : \mu_{d_0} < 4,7$

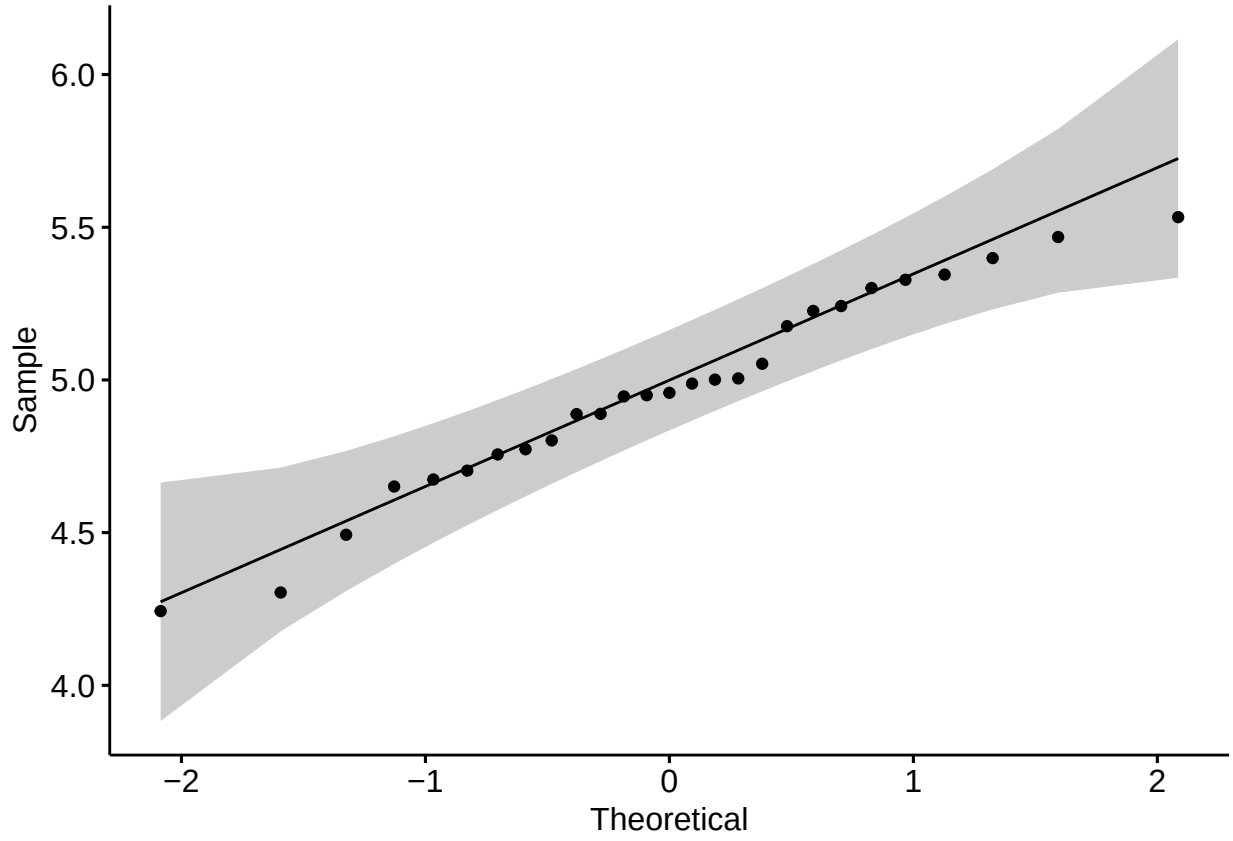
Selección de prueba:

Para esta pregunta se utilizará el t.test para muestras pareadas, ya que estamos comparando las medias y consiste en un mismo grupo de prueba en distintos tiempos.

Comprobación de condiciones necesarias:

```
dif = RA$Previo - RA$Posterior

ggqqplot(dif)
```




```
cat(shapiro.test(dif)$p.value)
```

```
## 0.6838767
```

Aplicación de la prueba:

```
t.test(x = RA$Previo, y = RA$Posterior, paired = TRUE, mu = 4.7, alternative = "less", conf.level = 1-0
```

```
##
##  Paired t-test
##
## data:  RA$Previo and RA$Posterior
## t = 4.1419, df = 26, p-value = 0.9998
## alternative hypothesis: true mean difference is less than 4.7
## 95 percent confidence interval:
```

```
##      -Inf 5.076219
## sample estimates:
## mean difference
##        4.966481
```

```r
t.test(dif, mu = 4.7, alternative = "less", conf.level = 1- 0.05)
```

```
##
##  One Sample t-test
##
## data:  dif
## t = 4.1419, df = 26, p-value = 0.9998
## alternative hypothesis: true mean is less than 4.7
## 95 percent confidence interval:
##      -Inf 5.076219
## sample estimates:
## mean of x
##  4.966481
```

Conclusión: como $p > 0.05$ se falla en rechazar H_0 es decir, no se puede asegurar que el promedio de las diferencias de tiempo en los atletas orientales es menor a 4,7 [s] con un 95% de confianza.

Pregunta 3

3.- ¿Es posible afirmar que, en promedio, los atletas de raza negra superaban a los de raza blanca por más de 2 segundos antes del entrenamiento?

Respuesta:

Formulación de hipótesis:

Las hipótesis a formular en este caso son:

H_0: La diferencia entre el promedio de tiempo entre la raza blanca y negra respectivamente antes del entrenamiento es menor o igual a 2.

H_1: Los atletas de raza blanca superaban a los de raza negra por más de 2 segundos antes del entrenamiento.

En lenguaje matemático:

Si μ_{negra} y μ_{blanca} son la media de tiempo antes del entrenamiento de los atletas antes del entrenamiento, entonces:

H_0: $\mu_{blanca} - \mu_{negra} = 2$

H_1: $\mu_{blanca} - \mu_{negra} > 2$

Selección de prueba:

Para esta pregunta se hizo uso de la prueba t de Student para dos muestras no pareadas, ya que las 2 muestras son los atletas de raza negra y blanca. Estos son no pareados ya que nos preguntan por el desempeño antes del entrenamiento, lo que provoca que no pueda estar pareado ya que solo se estudia el antes.

Verificación de condiciones:

Para hacer uso de este test deben cumplirse los siguientes requisitos:

1.- Cada muestra cumple con las condiciones para usar la distribución t.

2.- Las muestras son independientes entre sí.

```
library(ggpubr)
#Test de normalidad.

#Tiempos previos al entrenamiento.
Raza_Blanca_Previo <- RB$Previo

Raza_Negra_Previo <- RN$Previo

#Verificación si las muestras se distribuyen de manera cercana a la normal.

normalidad_Blanca_Previo <- shapiro.test(Raza_Blanca_Previo)
print(normalidad_Blanca_Previo)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Raza_Blanca_Previo
## W = 0.97651, p-value = 0.7925
```

```
normalidad_Negra_Previo <- shapiro.test(Raza_Negra_Previo)
print(normalidad_Negra_Previo)
```

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

En cuanto al supuesto de normalidad para cada muestra, al aplicar Shapiro-wilk se obtiene, $p = 0.7925$ y $p = 0.6631$. En ambos casos el valor $p > 0.05$, por lo que podemos asumir que ambas muestras provienen de poblaciones que se distribuyen de forma aproximadamente normal.

Ambas muestras son independientes entre sí, pues son de diferentes razas de atletas y fueron seleccionados de manera aleatoria dentro de sus respectivas razas.

Prueba estadística:

```
#Nivel de significación.

P3_alfa <- 0.05

#Aplicación de la prueba t para dos muestras independientes.

P3_prueba <- t.test(y = Raza_Negra_Previo,
```

```
                    x = Raza_Blanca_Previo,
                    paired =FALSE,
                    alternative = "greater",
                    mu = 2,
                    conf.level = 1- P3_alfa)

print(P3_prueba)
```

```
##
##  Welch Two Sample t-test
##
## data:  Raza_Blanca_Previo and Raza_Negra_Previo
## t = 1.2282, df = 50.106, p-value = 0.1126
## alternative hypothesis: true difference in means is greater than 2
## 95 percent confidence interval:
##  1.875807      Inf
## sample estimates:
## mean of x mean of y
##  16.45381  14.11304
```

Finalmente, de los resultados del t test para 2 muestras independientes con un $p > 0.05$, se falla en rechazar a la hipotesis nula, lo cual significa que con un 95% de confianza, no se puede asegurar que la diferencia entre los tiempos promedios de los atletas de raza blanca y negra es mayor a 2 [s] antes del entrenamiento.

Pregunta 4

4.- ¿Será cierto que hay más atletas de raza blanca que redujeron sus mejores marcas en al menos 2,9 segundos que atletas de raza negra que lo hicieron en al menos 1,9 segundos?

Respuesta:

Selección de hipótesis:

Las hipótesis correspondientes son:

p_0 := proporción de atletas de raza blanca que tengan una marca mayor o igual a 2,9[s].

p_1 := proporción de atletas de raza negra que tengan una marca mayor o igual a 1,9[s].

$H_0 : p_0 - p_1 = 0$

$H_1 : p_0 - p_1 > 0$

Cálculo de las proporciones de las muestras:

```
RBdif =  RB$Previo - RB$Posterior
RNdif =  RN$Previo - RN$Posterior

dataB = data.frame(RBdif)
RBdifMayor = dataB %>% filter(RBdif >= 2.9)
```

```r
RBn = length(RBdifMayor$RBdif)

dataN = data.frame(RNdif)
RNdifMayor = dataN %>% filter(RNdif >= 1.9)
RNn = length(RNdifMayor$RNdif)

PropRB = RBn/nrow(RB)
PropRN = RNn/nrow(RN)
```

Selección de la prueba:

Para esta pregunta se utilizará el test de Wilson para dos muestras, ya que estamos hablando de proporciones de distintas razas.

Comprobación de condiciones

Ambas muestras son de razas distintas, por lo tanto podemos asegurar que son independientes entre si. Condición éxito-fracaso.

$np \geq 10$ y $n(1-p) \geq 10$

```r
#Prop RB
exitoRB = nrow(RB)*PropRB
fracasoRB = nrow(RB)*(1-PropRB)

#Prop RN
exitoRN = nrow(RN)*PropRN
fracasoRN = nrow(RN)*(1-PropRN)
```

Si bien la condición de éxito-fracaso se cumple en los atletas de raza blanca, esta no se cumple en los altetas de raza negra. //Como no sabemos qué hacer, lo aplicaremos de igual forma.

Aplicación de la prueba:

```r
exitos = c(RBn,RNn)
n = c(nrow(RB),nrow(RN))
alpha = 0.05

prueba <- prop.test(exitos, n = n, alternative = "greater", conf.level = 1-alpha)

print(prueba)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  exitos out of n
## X-squared = 7.3991, df = 1, p-value = 0.003263
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.161778 1.000000
```

```
## sample estimates:
##    prop 1    prop 2
## 0.6153846 0.2142857
```

Como $p < 0.05$ se rechaza H_0 en favor de H_1, entonces podemos asegurar con un 95% de confianza que hay más atletas de raza blanca que redujeron sus mejores marcas en al menos 2,9 segundos que atletas de raza negra que lo hicieron en al menos 1,9 segundos.