Ejemplo de solución ejercicio prático N°2

Primero, cargamos los datos a trabajar.

```
# setwd("C:/Users/ProfeJLJara/Downloads")
setwd("/home/jljara/Dropbox/Cursos (parcial)/EI/2023-2/Ejercicios prácticos/EP02")
datos <- read.csv2("EP02 Datos.csv")
```

Pregunta 1

El Comité Olímpico cree que el mejor tiempo medio de los atletas de raza blanca después de ingresar al programa de entrenamiento es superior a 17,23 segundos. ¿Soportan los datos esta afirmación?

En este caso, debemos inferir acerca de la **población** de atletas de raza blanca que no han pasado por el entrenamiento. Filtramos los datos para obtener los necesitados.

```
muestra.1 <- datos %>%  filter(Raza == "Blanca") %>% pull(Previo)

n.1 <- length(muestra.1)
cat("Tamaño de la muestra:", n.1, "\n")
```

```
## Tamaño de la muestra: 26
```

Como la muestra es pequeña (menos de 30 observaciones) y no conocemos la desviación estándar de la población, sería adecuado usar la **prueba t de Student para una muestra**. Pero antes debemos verificar las condiciones.

Como se trata de 26 atletas diferentes y la muestra representa menos del 10% de la población, podemos asumir que las observaciones son independientes entre sí.

Ahora debemos verificar si las observaciones presentan una distribución cercana a la normal. Una forma de hacer esto es mediante la prueba de normalidad de Shapiro-Wilk.

```
normalidad.1 <- shapiro.test(muestra.1)
cat("Comprobación de la normalidad de los datos:\n")
```

```
## Comprobación de la normalidad de los datos:
```

```
print(normalidad.1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  muestra.1
## W = 0.97651, p-value = 0.7925
```

Puesto que el estadístico obtenido ($W = 0{,}977$) nos lleva a un valor p alto ($p = 0{,}793$), no podemos descartar que la muestra proviene de una distribución normal, por lo que podemos aplicar la prueba seleccionada.

Como no hay indicios de que tengamos que ser cautelosos con los resultados, fijamos el nivel de significación en $\alpha = 0{,}05$.

Ahora debemos formular las hipótesis:

H_0: antes del entrenamiento, la media de las mejores marcas de los atletas de raza blanca ($\mu_{\mathrm{B}}^{\mathrm{antes}}$) en 100 metros planos es de 17,23 segundos ($\mu_{\mathrm{B}}^{\mathrm{antes}} = 17{,}23$ [s]).

H_A: antes del entrenamiento, la media de las mejores marcas de los atletas de raza blanca en 100 metros planos es distinta de 17,23 segundos ($\mu_{\mathrm{B}}^{\mathrm{antes}} \neq 17{,}23$ [s]).

Efectuamos la prueba con estas condiciones:

```
alfa.1 <- 0.05
valor.nulo.1 <- 17.23

prueba.1 <- t.test(muestra.1, alternative = "two.sided", mu = valor.nulo.1, conf.level = 1 - alfa.1)

cat("Prueba de hipótesis pregunta 1:\n")
```

```
## Prueba de hipótesis pregunta 1:
```

```
print(prueba.1)
```

```
##
##  One Sample t-test
##
## data:  muestra.1
## t = -4.5168, df = 25, p-value = 0.0001304
## alternative hypothesis: true mean is not equal to 17.23
## 95 percent confidence interval:
##  16.09988 16.80773
## sample estimates:
## mean of x
##  16.45381
```

Ahora debemos interpretar lo que nos dice la prueba realizada.

Vemos que la prueba ($t(25) = 4{,}517$) nos indica que la verdadera media de las mejores marcas de atletas de raza blanca antes de entrar al programa de entrenamiento es distinta de 17,23 [s] ($p < 0{,}001$) y que con 95% de confianza se encuentra en el intervalo [16,100; 16,808].

Pregunta 2

¿Sugieren los datos que en promedio la mejor marca de los atletas de raza oriental se reduce en menos de 6,45 segundos tras el entrenamiento?

En este caso, debemos inferir acerca de la media de dos muestras pareadas (media de las diferencias). Obtengamos las muestras con las que trabajaremos.

```
anterior <- datos %>% filter(Raza == "Oriental") %>% pull("Previo")
posterior <- datos %>% filter(Raza == "Oriental") %>% pull("Posterior")
```

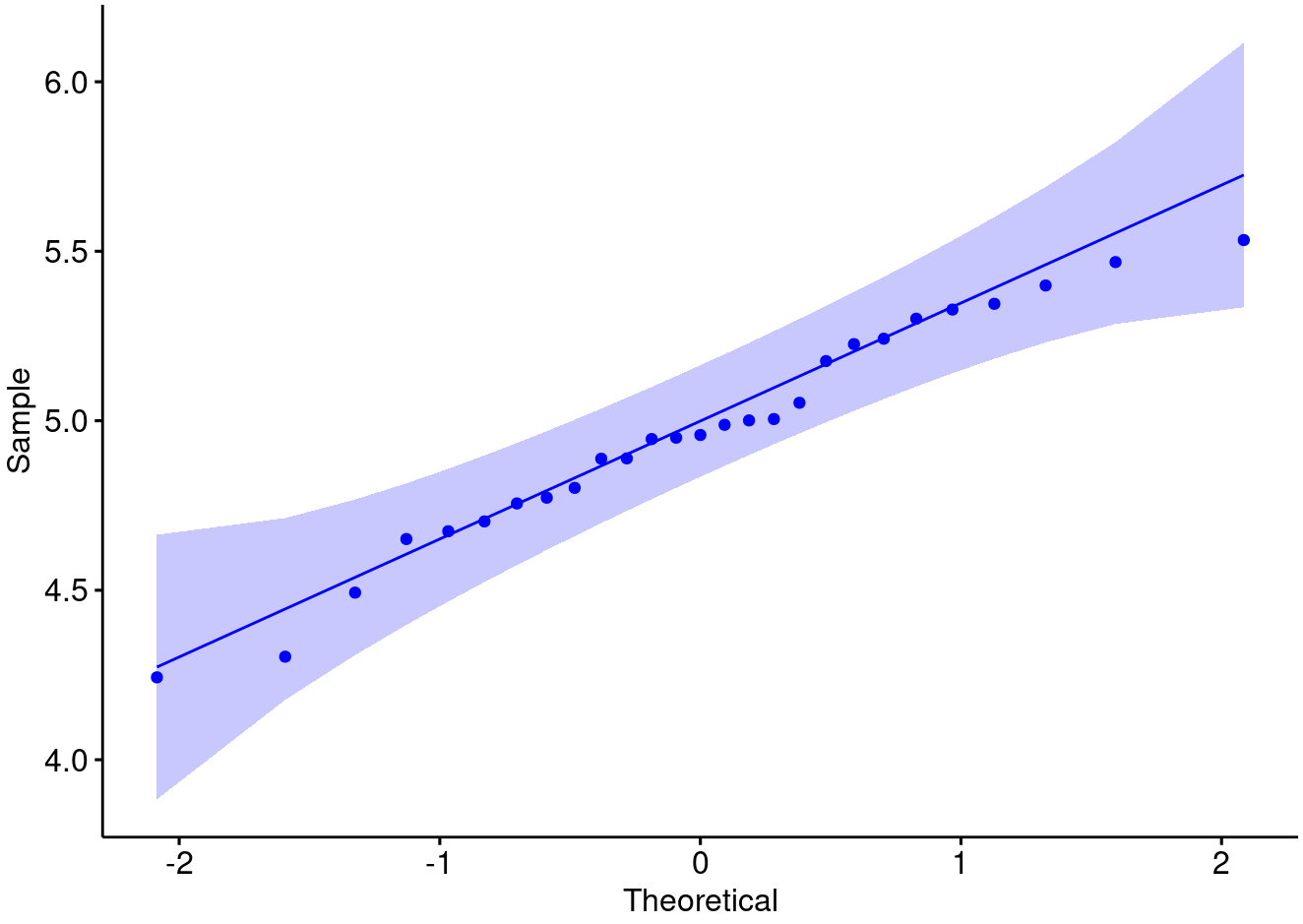
Veamos ahora el tamaño de las muestras (ambas tienen el mismo, al estar pareadas).

```
n.2 <- length(anterior)
cat("Tamaño de la muestra:", n.2, "\n")
```

```
## Tamaño de la muestra: 27
```

Como la muestra es pequeña (menos de 30 observaciones) y no conocemos la desviación estándar de la población, sería adecuado usar la **prueba t de Student para dos muestras pareadas**. Pero antes debemos verificar las condiciones. Como se trata de 27 atletas diferentes, menor al 10% de la población, podemos suponer que los pares de observaciones son independientes entre sí. Ahora debemos verificar si las diferencias presentan una distribución cercana a la normal. Una forma de hacer esto es mediante un gráfico Q-Q.

```
diferencias <- anterior - posterior
g2 <- ggqqplot(data.frame(diferencias), x = "diferencias", color = "blue")
print(g2)
```



La forma de los datos en el gráfico no se aleja tanto de una recta, y podemos ver que no hay evidencia de valores atípicos, pues no hay puntos fuera de la banda coloreada.

Como no hay indicios de que tengamos que ser cautelosos con los resultados, fijamos el nivel de significación en 0,05.

Ahora debemos formular las hipótesis:

H_0**: tras el entrenamiento, la media de las mejores marcas de los atletas de raza oriental en los 100 metros planos se reduce en 6,45 segundos (**$\mu_{\mathrm{O}}^{\mathrm{antes}} - \mu_{\mathrm{O}}^{\mathrm{despues}} = 6{,}45$ **[s] donde** $\mu_{\mathrm{O}}^{\mathrm{antes}}$ **y** $\mu_{\mathrm{O}}^{\mathrm{despues}}$ **las medias de los mejores registros antes y despúes del entrenamiento, respectivamente).**

H_A: **tras el entrenamiento, la media de las mejores marcas de los atletas de raza oriental en los 100 metros planos se reduce en menos de 6{,}45 segundos ($\mu_{\mathrm{O}}^{\mathrm{antes}} - \mu_{\mathrm{O}}^{\mathrm{despues}} < 6{,}45$ [s]).**

Debemos fijarnos bien en qué orden vamos a calcular las diferencias. Como escribimos las hipótesis, si se resta la media de los mejores tiempos después del entrenamiento a la media de los mejores tiempos antes del entrenamiento, esperamos que esta diferencia sea positiva (mayores tiempos antes del entrenamiento), por lo que el valor nulo también debe ser positivo y la hipótesis alternativa es que la verdadera diferencia promedio es menor (*less*) a este valor. Si hubiéramos considerado la diferencia como la media de los mejores tiempos después del entrenamiento menos la media de los mejores tiempos antes del entrenamiento, esperaríamos un valor negativo y, en consecuencia, el valor nulo ha de ser negativo y la hipótesis alternativa esperaría encontrar un valor mayor (*greater*) a este número hipotético. Efectuamos (solo por esta vez, por razones pedagógicas) ambas versiones de la prueba con estas consideraciones:

```
alfa.2 <- 0.05
valor.nulo.2a <- 6.45
valor.nulo.2b <- -6.45

prueba.2a <- t.test(x = anterior, y = posterior, alternative = "less", mu = valor.nulo.2a, paired = TRUE, c
onf.level = 1 - alfa.2)
prueba.2b <- t.test(x = posterior, y = anterior, alternative = "greater", mu = valor.nulo.2b, paired = TRU
E, conf.level = 1 - alfa.2)

cat("Prueba de hipótesis pregunta 2:\n")
```

```
## Prueba de hipótesis pregunta 2:
```

```
print(prueba.2a)
```

```
##
##  Paired t-test
##
## data:  anterior and posterior
## t = -23.058, df = 26, p-value < 2.2e-16
## alternative hypothesis: true mean difference is less than 6.45
## 95 percent confidence interval:
##      -Inf 5.076219
## sample estimates:
## mean difference
##        4.966481
```

```
print(prueba.2b)
```

```
##
##  Paired t-test
##
## data:  posterior and anterior
## t = 23.058, df = 26, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than -6.45
## 95 percent confidence interval:
##  -5.076219       Inf
## sample estimates:
## mean difference
##       -4.966481
```

Ahora podemos concluir,

> Las pruebas indican que debemos rechazar hipótesis nula en favor de la alternativa con el nivel de significación establecido ($t(26) = 23{,}058; p < 0{,}001$), por lo que podemos concluir, con 95% de confianza, que la mejor marca de los atletas de raza oriental en los 100 metros planos se redujo en promedio menos de 6,45 segundos (95% CI para la media de las diferencias: [$-\infty$; 5,076] [s]) tras el entrenamiento.

Pregunta 3

¿Es posible afirmar que, en promedio, los atletas de raza negra superan a los de raza blanca por más de 2 segundos después del entrenamiento?

En este caso, debemos inferir acerca de la diferencia entre las medias de dos muestras independientes (diferencia de las medias). Obtengamos las muestras con las que trabajaremos.

```
raza.negra <- datos %>% filter(Raza == "Negra") %>% pull(Posterior)
raza.blanca <- datos %>% filter(Raza == "Blanca") %>% pull(Posterior)
```

Veamos ahora el tamaño de las muestras.

```
n.raza.negra <- length(raza.negra)
n.raza.blanca <- length(raza.blanca)
cat("Tamaño de las muestras:", n.raza.negra, "y", n.raza.blanca, "\n")
```

```
## Tamaño de las muestras: 28 y 26
```

Como las muestras son pequeñas (menos de 30 observaciones) y no conocemos la(s) desviación(es) estándar(es) de la(s) población(es), sería adecuado usar la **prueba t de Student para dos muestras independientes**. Pero antes debemos verificar las condiciones. Como en el caso de ambas muestras se trata de 28 y 26 atletas diferentes, menor al 10% de la población respectiva, y la elección de una/o en particular no influye en ñla elección de otra/o, podemos suponer que las observaciones son independientes entre sí. Ahora debemos verificar si cada una de las muestras presenta una distribución cercana a la normal.

```
normalidad.raza.negra <- shapiro.test(raza.negra)
normalidad.raza.blanca <- shapiro.test(raza.blanca)
cat("\nComprobación de la normalidad de los datos:\n")
```

```
##
## Comprobación de la normalidad de los datos:
```

```
cat("Primera muestra:\n")
```

```
## Primera muestra:
```

```
print(normalidad.raza.negra)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  raza.negra
## W = 0.95201, p-value = 0.2225
```

```
cat("Segunda muestra:\n")
```

```
## Segunda muestra:
```

```
print(normalidad.raza.blanca)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  raza.blanca
## W = 0.94734, p-value = 0.2008
```

Podemos ver que, en ambos casos, las pruebas de normalidad resultan negativas (raza negra: $W = 0{,}952; p = 0{,}223$; raza blanca: $W = 0{,}947; p = 0{,}201$), por lo que es razonable suponer que ambas muestras provienen de una distribución cercana a la normal.

Como no hay indicios de que tengamos que ser cautelosos con los resultados, fijamos el nivel de significación en 0,05.

Ahora debemos formular las hipótesis:

H_0**: después del entrenamiento, en promedio, los atletas de raza negra superan a los de raza blanca por 2 segundos (**$|\mu_{\mathbf{N}}^{\mathbf{despues}} - \mu_{\mathbf{B}}^{\mathbf{despues}}| = \mathbf{2}$ **[s], donde** $\mu_{\mathbf{N}}^{\mathbf{despues}}$ **y** $\mu_{\mathbf{B}}^{\mathbf{despues}}$ **son, respectivamente, las medias de los mejores registros despúes del entrenamiento de los atletas de raza negra y de raza blanca).**

H_A**: tras del entrenamiento, en promedio, los atletas de raza negra superan a los de raza blanca por más de 2 segundos (**$|\mu_{\mathbf{N}}^{\mathbf{despues}} - \mu_{\mathbf{B}}^{\mathbf{despues}}| < \mathbf{2}$ **[s]).**

Realicemos la prueba con estas consideraciones y tendiendo cuidado de ser consistentes con si obtendremos una diferencia positiva o negativa y ejecutar la versión adecuada para muestras independientes (es decir, la prueba de Welch).

```
alfa.3 <- 0.05
valor.nulo.3 <- 2

prueba.3 <- t.test(x = raza.blanca, y = raza.negra, alternative = "greater",
                   mu = valor.nulo.3, paired = FALSE, conf.level = 1 - alfa.3)

cat("Prueba de hipótesis pregunta 3:\n")
```

```
## Prueba de hipótesis pregunta 3:
```

```
print(prueba.3)
```

```
##
##  Welch Two Sample t-test
##
## data:  raza.blanca and raza.negra
## t = -3.8383, df = 50.809, p-value = 0.9998
## alternative hypothesis: true difference in means is greater than 2
## 95 percent confidence interval:
##  0.3047431       Inf
## sample estimates:
## mean of x mean of y
##  13.36719  12.54732
```

Interpretemos estos resultados.

La prueba nos dice que no hay evidencia suficiente para rechazar hipótesis nula ($t(50{,}809) = 3{,}838; p > 0{,}999$), por lo que podemos concluir, con 95% de confianza, que en promedio las mejores marcas de los atletas de raza negra no supera a las de los atletas de raza blanca en más de 2 segundos (95% CI para la diferencia de las medias: [0,305; ∞] [s]) tras el entrenamiento.

Pregunta 4

¿Será cierto que hay más atletas de raza blanca que redujeron sus mejores marcas en menos de 2,8 segundos que atletas de raza negra que lo hicieron en menos de 1,5 segundos?

En este caso, no están preguntando por números absolutos, ya que los tamaños de las muestras son distintos. Debemos entonces inferir acerca de la **diferencia entre dos proporciones** a partir de **muestras independientes**.

Formulemos las hipótesis:

H_0: la proporción de atletas de raza blanca que redujeron sus mejores marcas en menos de 2,8 segundos ($\tilde{\mathbf{p}}_{\mathbf{B}}^{\mathbf{D<2,8}}$) no es distinta a la proporción de atletas de raza negra que redujeron sus mejores marcas en menos de 1,5 segundos ($\tilde{\mathbf{p}}_{\mathbf{N}}^{\mathbf{D<1,5}}$). Matemáticamente: $\tilde{\mathbf{p}}_{\mathbf{B}}^{\mathbf{D<2,8}} - \tilde{\mathbf{p}}_{\mathbf{N}}^{\mathbf{D<1,5}} = \mathbf{0}$.

H_A: la proporción de atletas de raza blanca que redujeron sus mejores marcas en menos de 2,8 segundos ($\tilde{\mathbf{p}}_{\mathbf{B}}^{\mathbf{D<2,8}}$) es mayor a la proporción de atletas de raza negra que redujeron sus mejores marcas en menos de 1,5 segundos ($\tilde{\mathbf{p}}_{\mathbf{N}}^{\mathbf{D<1,5}}$). Matemáticamente: $\tilde{\mathbf{p}}_{\mathbf{B}}^{\mathbf{D<2,8}} - \tilde{\mathbf{p}}_{\mathbf{N}}^{\mathbf{D<1,5}} > \mathbf{0}$.

Obtengamos los datos que necesitamos trabajar.

Primero filtramos para dejar los atletas de interés, luego creamos una columna con la disminución de tiempo logrado por el programa de entrenamiento y, finalmente, contamos los éxitos según lo indicado en el enunciado.

```
muestra.3 <- datos %>% filter(Raza == "Blanca" | Raza ==  "Negra")  %>%
  mutate(disminucion = abs(Posterior - Previo))

n.raza.blanca <- muestra.3 %>% filter(Raza == "Blanca") %>% nrow()
exitos.raza.blanca <- muestra.3 %>% filter(Raza == "Blanca") %>%
  filter(disminucion >= 2.9) %>% nrow()

n.raza.negra <- muestra.3 %>% filter(Raza == "Negra") %>% nrow()
exitos.raza.negra <- muestra.3 %>% filter(Raza == "Negra") %>%
  filter(disminucion >= 1.6) %>% nrow()
```

Corresponde revisar las condiciones para aplicar una prueba de la diferencia de dos proporciones de forma válida. Por un lado, las muestras son independientes pues, de acuerdo al enunciado, los atletas se eligieron de forma aleatoria. Faltaría entonces verificar que cada proporción, por separado, sigue el modelo normal.

```
cat("Condiciones:\n")
```

```
## Condiciones:
```

```
cat("Éxitos y fracasos en la muestra de raza blanca:",
    exitos.raza.blanca, ",", n.raza.blanca - exitos.raza.blanca, "\n")
```

```
## Éxitos y fracasos en la muestra de raza blanca: 16 , 10
```

```
cat("Éxitos y fracasos en la muestra de raza negra:",
    exitos.raza.negra, ",", n.raza.negra - exitos.raza.negra, "\n")
```

```
## Éxitos y fracasos en la muestra de raza negra: 14 , 14
```

Vemos que las frecuencias encontradas cumplen (justito) la condición de éxito-fracaso (se espera observar al menos 10 éxitos y al menos 10 fracasos). Podemos entonces continuar con la prueba de proporciones.

Nota:

Si la condición de éxito-fracaso no se cumpliera, se podrían seguir dos caminos. Primero, asegurarse de usar la corrección de continuidad de Yates al ejecutar la prueba (`correct = TRUE`) y considerar usar un nivel de significación exigente ($\alpha < 0{,}05$). La segunda alternativa sería explorar métodos exactos que han estado saliendo en los últimos años, aunque estos no están tan validados por la comunidad todavía. Por ejemplo:

Laurencelle, L. (2021). The exact binomial test between two independent proportions: A companion. The Quantitative Methods for Psychology, 17, 76-79.

Corresponde ahora realizar la prueba.

```
prueba.4 <- prop.test(c(exitos.raza.blanca, exitos.raza.negra),
                      c(n.raza.blanca, n.raza.negra),
                      alternative = "greater")

cat("Pruebas de hipótesis pregunta 4:\n")
```

```
## Pruebas de hipótesis pregunta 4:
```

```
print(prueba.4)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(exitos.raza.blanca, exitos.raza.negra) out of c(n.raza.blanca, n.raza.negra)
## X-squared = 0.33472, df = 1, p-value = 0.2814
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.142579  1.000000
## sample estimates:
##    prop 1    prop 2
## 0.6153846 0.5000000
```

Interpretemos estos resultados.

La prueba nos dice que no hay evidencia suficiente para rechazar hipótesis nula ($\chi^2(1) = 0{,}335; p = 0{,}281$). Así, no hay razones descartar que, tras el entrenamiento, la proporción de atletas de raza blanca que redujeron sus mejores marcas en menos de 2,8 [s] es igual a la proporción de atletas de raza negra que lo hicieron en menos de 1,5 [s] (95% CI para la diferencia de las proporciones: [-0,143; 1,000] [s]).