Maestría en Computo Estadístico
Inferencia Estadística
Tarea 9
13 de diciembre de 2020
Enrique Santibáñez Cortés
Repositorio de Git: Tarea 9, IE.

Pruebas de permutación

24 alumnos del ultimo año escolar, fueron divididos aleatoriamente en dos grupos de igual tamaño, esto con el fin de someterlos a un experimento en el cual a un grupo se le inscribiría en un curso pre universitario extra clase, como preparación para el ingreso a la universidad publica, y el otro grupo únicamente recibiría las clases del colegio. El archivo *resultados_examen.csv* guarda la información de los alumnos, su puntaje global en el examen de admisión a la universidad y el grupo al que pertenecía. Con esta información realice.

1. Una prueba de hipótesis por el método de pruebas de permutación, que ayude a identificar si el curso pre universitario si presento una mejoría en los resultados de los estudiantes.

RESPUESTA

Definición: 1 *Para poder realizar la prueba de permutación solo se debe seguir una serie de pasos que a continuación se hará mención:*

1. *Analizar el problema: Primero habrá que identificar la hipótesis nula y alternativa de interés. Normalmente se plantea la hipótesis que las poblaciones son iguales, sin presentar diferencias.*
2. *Escoger un estadístico de prueba: Se elegirá un estadístico de prueba que discrimine entre la hipótesis nula y la alternativa.*
3. *Calcular el valor observado del estadístico de prueba: Se deberá calcular el estadístico de prueba en el orden original de los registros obtenidos de la siguiente forma:*

$$t_{obs} = T(X_1, \cdots, X_m, Y_1, \cdots, Y_n)$$

4. *Permutar Aleatoriamente los datos: Se reorganizará las etiquetas (permutará) y volverá a calcular los estadísticos de prueba utilizando los datos permutados.*
5. *Repetir el paso anterior: Se repetirá el paso anterior $B-$veces, esto debido a que no es práctico evaluar los $N!$ de permutaciones, y dejar $T_1, \cdots, T_B$ denote los valores resultantes.*
6. *Tomar una decisión: Se aceptará o rechazará la hipótesis utilizando esta distribución como guía, la aproximación del $p-$valor es*

$$p - value = \frac{1}{B}\sum_{j=1}^{B} I(T_j > t_{obs}).$$

Para identificar si el curso pre universitario presentó una mejoría en los resultados de los estudiantes, ocupamos la prueba de permutación para comparar la media entre dos poblaciones de tamaño 24. Y se utilizan todas las combinaciones posibles:

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  2.1.3     v dplyr   0.8.1
## v tidyr   0.8.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(gtools)
set.seed(19970808)

# leemos los datos
calificaciones <- read.csv("resultados_examen.csv")

# dividimos los datos
pre_universitarios <- calificaciones %>% filter(preU=="SI")
sin_universitarios <- calificaciones %>% filter(preU=="NO")

# calculamos el estadístico para la muestra original
T1 <- mean(pre_universitarios$puntaje)-mean(sin_universitarios$puntaje)


# permutaciones
x <- c(pre_universitarios$puntaje, sin_universitarios$puntaje)
x <- sort(x)
grupo1 <- combinations(length(x), 12,v = x)
grupo2 <- matrix(0,nrow = nrow(grupo1),ncol = ncol(grupo1))
for (i in 1:nrow(grupo1)) {
  v = which(!(x %in% grupo1[i,]))
  grupo2[i,] = x[v]
}

medias_grupo1 <- apply(X = grupo1,MARGIN = 1,FUN = mean)
medias_grupo2 <- apply(X = grupo2,MARGIN = 1,FUN = mean)

#Calculo estadistico de prueba para todas las combinaciones
dif_medias <- medias_grupo2 - medias_grupo1
Tobs <- sum(dif_medias >= T1)

#p valor
p_valor = Tobs/length(dif_medias)
p_valor
```

```
## [1] 0.005224551
```

Por lo tanto, con un 95 % de confianza entonces podemos rechazar la hipótesis nula por lo que podemos concluir que si existe una mejoría en los resultadoas de los estudiantes que estuvieron en el curso pre

universitario.

2. Una prueba parametrica de su elección para el mismo contraste de hipótesis.

RESPUESTA

Definición: 2 *Sea $X_1, X_2, \cdots, X_{n_1}$ v.a. independientes con distribución Normal(μ_X, σ_X^2) (donde σ_X^2 es desconocida, y $Y_1, Y_2, \cdots, Y_{n_2}$ v.a. independientes con distribución Normal(μ_Y, σ_Y^2) (donde σ_Y^2 es desconocida, y X_i es independiente a Y_i. Entonces si queremos verificar que las medias de la población son distintas, entonces se plantean el juego de hipótesis:*

$$H_0: \ \mu_X = \mu_Y \quad vs. \quad H_1: \ \mu_X > \mu_Y.$$

Esta prueba tiene el estadístico de prueba

$$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

donde

$$s_p = \sqrt{\frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}}.$$

Y la región de rechazar la hipotesis nula H_0 con un nivel α es si $T > t_{n_1+n_2-2,1-\alpha}$.

Observemos que el tamaño de las muestras son pequeñas, por lo que no podemos utilizar el TLC para determinar un estadístico de prueba. Ahora, no conocemos ni la distribución de la población ni la varianza real, pero para este caso supondremos que la distribución es normal. Ocupemos la pureba de prueba de Shapiro−Wilk

```
shapiro.test(pre_universitarios$puntaje) # test shapiro
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pre_universitarios$puntaje
## W = 0.95521, p-value = 0.7139
```

```
shapiro.test(sin_universitarios$puntaje) # test shapiro
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sin_universitarios$puntaje
## W = 0.95748, p-value = 0.7473
```

Entonces podemos concluir que si se distribuyen como una normal. Ahora, nos interesa saber si el tratamiento tiene efecto sobre la media de la concentración de la glucosa en la sangre entonces podemos usar la prueba t−student (definición 2) para validar este resultado. Tenemos el siguiente juego de hipótesis

$$H_0: \ \mu_T = \mu_C \quad \text{vs.} \quad H_1: \ \mu_T \neq \mu_C.$$

donde μ_T es la media del grupo tratamiento y μ_C es la media del grupo control. Nuestro estadístico de prueba es (realizamos los cálculos en R)

```r
n_1 <- nrow(pre_universitarios) # tamaños
n_2 <- nrow(sin_universitarios)

x_bar <- mean(pre_universitarios$puntaje) # medias muestrales
y_bar <- mean(sin_universitarios$puntaje)

std2_x <- var(pre_universitarios$puntaje) # varianza muestral
std2_y <- var(sin_universitarios$puntaje)

sp <- sqrt(((n_1-1)*std2_x+(n_2-1)*std2_y)/(n_1+n_2-2)) # sp

t <- (x_bar-y_bar)/(sp*sqrt(1/n_1+1/n_2)) # estadístico prueba
t
```

```
## [1] 2.794469
```

$$t = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2{,}7944691,$$

```r
qt(0.975, 9)
```

```
## [1] 2.262157
```

Entonces con $\alpha = 0{,}05$ tenemos que el porcentil de la distribución t−student es $t_{n_1+n_2-2,\alpha/2} = t_{9,0,025} = 2{,}262157$. Por lo tanto, como $t = 2{,}794469 > 2{,}26 = t_{9,0,025}$ **podemos rechazar la hipotésis nula y por lo que podemos concluir que si existe una mejoría en los resultadoas de los estudiantes que estuvieron en el curso pre universitario.**

3. Compare los resultados

RESPUESTA
Comparando las concluciones de ambas pruebas observamos que en las dos concluímos lo mismo. Cabe aclara que en la prueba parámetrica antes de utilizarla tuvimos que comprobar que las m.a son normales creo que eso es una ventaja muy clara de las pruebas de permutación sobre las pruebas no parámetricas. ■.