

COMPARACIÓN DE TRES ALGORITMOS DE PRUEBA DE PERMUTACIÓN APLICADOS A LAS MEDIAS MULTIVARIADAS DE DOS MUESTRAS INDEPENDIENTES

JAIME CARLOS PORRAS CERRÓN*
Universidad Nacional Agraria La Molina

Recibido: 12 de marzo del 2019 / Aprobado: 30 de mayo del 2019

doi: 10.26439/ing.ind2019.n037.4545

RESUMEN: El objetivo de esta investigación fue comparar tres algoritmos de prueba de permutación. Se propusieron escenarios de datos obtenidos mediante simulación de Monte Carlo y en cada uno se aplicaron los algoritmos propuestos. Los resultados muestran una potencia de prueba superior a 0,85. El primer algoritmo basado en la T^2 de Hotelling presentó la mayor potencia de prueba. La implementación de los algoritmos fue realizada con el programa R.

Palabras clave: prueba de permutación / simulación de Monte Carlo / prueba T^2 de Hotelling / Prueba t / prueba de Wilcoxon-Mann-Whitney / potencia de prueba

COMPARISON OF THREE PERMUTATION TEST ALGORITHMS APPLIED TO THE MULTIVARIATE MEANS OF TWO INDEPENDENT SAMPLES

ABSTRACT: The objective of this research was to compare three permutation test algorithms. Data scenarios obtained through the Monte Carlo simulation were proposed and the suggested algorithms were applied to each of them. The results showed a test power greater than 0,85. The first algorithm based on the Hotelling's T^2 presented the highest test power. The implementation of the algorithms was carried out using the R statistical program.

Keywords: permutation test / Monte Carlo simulation / Hotelling's T^2 test / t test / Wilcoxon-Mann-Whitney test / test power

* jaimepc@lamolina.edu.pe

1. INTRODUCCIÓN

En investigaciones de diferentes áreas, uno de los objetivos puede ser comparar las medias de dos poblaciones con la finalidad de determinar si estas difieren. Para cumplir con este propósito se puede hacer uso de procedimientos paramétricos univariados o multivariados. En el caso univariado se puede utilizar la prueba Z o la prueba t de Student (si se desconocen las varianzas poblacionales). En el caso multivariado, si las matrices de covarianza son homogéneas, se puede hacer uso de la prueba T^2 de Hotelling para comparar las medias de dos poblaciones.

Tanto para el caso univariado como para el multivariado se requiere verificar previamente que las muestras provengan de una distribución normal. Sin embargo, esta situación no siempre se cumple. Ante esta falla de las pruebas paramétricas tradicionales, los métodos no paramétricos, como la prueba de permutación, han surgido como una buena alternativa, debido a que no será necesario verificar que los datos provengan de una distribución teórica conocida.

Existen investigaciones donde se estudia la prueba de permutación. Butar y Park (2008) realizaron un estudio para comparar dos poblaciones mediante pruebas de permutación univariada. Concluyeron que las pruebas basadas en remuestreo no son complicadas de implementar y que al realizar todas las permutaciones posibles se puede obtener un nivel de significancia exacto, mientras que cuando el número de permutaciones es muy grande se puede obtener un nivel de significancia aproximado.

Chung y Romano (2011) presentaron una prueba de permutación para dos muestras independientes basadas en el estadístico de Wilcoxon, y obtuvieron sus resultados mediante un estudio de simulación de Monte Carlo.

Einsporn y Habtzghi (2013) propusieron una prueba de permutación para datos de pares incompletos. Esta situación surge tanto en estudios observacionales como experimentales cuando algunos de los datos están en forma de muestra pareada y el resto de los datos comprende dos muestras independientes. Se demostró mediante simulación de Monte Carlo que el método propuesto funciona mejor que los métodos tradicionales.

Samuh (2017) propuso un algoritmo para comparar dos muestras independientes a través de pruebas de permutación univariadas y el criterio *Ranked Set Sampling* mediante simulación. Concluyó que los resultados afirman que los niveles de potencia de la prueba de permutación son más altos que los niveles de potencia de la prueba t clásica.

Blair, Higgins, Karniski y Kromrey (2010) analizaron pruebas de permutación multivariada para una muestra que pueden sustituir a la T^2 de Hotelling. Las comparaciones de potencia se realizaron entre las pruebas de permutación y la prueba T^2 de Hotelling bajo una variedad de modelo de efecto de tratamiento, estructura de correlación y combinaciones de diferentes cantidades de variables. Los resultados muestran que las

pruebas de permutación tienen importantes ventajas de potencia con respecto a la T^2 de Hotelling. Algunas ventajas de las pruebas de permutación: pueden ser utilizadas cuando el número de variables es mayor al número de observaciones o cuando se desconoce la distribución de probabilidad conjunta de las variables en estudio.

Chung y Romano (2013) usaron pruebas de permutación para comparar parámetros multivariados de dos poblaciones mediante estudios de simulación de Monte Carlo. Utilizaron una T^2 modificada de Hotelling, así como pruebas basadas en la máxima diferencia absoluta estudentizada.

En la presente investigación, el desarrollo de la metodología de prueba de permutación multivariada para dos muestras independientes se realizó mediante la elaboración de procedimientos obtenidos con ayuda del programa estadístico R.

El objetivo de la presente investigación es comparar la potencia de prueba de tres algoritmos de prueba de permutación para la diferencia entre medias multivariadas en dos muestras independientes. Estos se aplicarán a diferentes escenarios (conjuntos de datos) obtenidos con la ayuda de la simulación de Monte Carlo.

2. MATERIALES Y MÉTODOS

2.1 Materiales

Para realizar la aplicación del presente trabajo de investigación, se utilizó el programa estadístico R que es muy útil para implementar los procedimientos de comparación. Esto se debe a que los programas estadísticos comerciales no cuentan con las pruebas que se consideraron en la presente investigación.

Se utilizaron diferentes funciones del programa:

- La función `t.test` del paquete `base stats` permite realizar la prueba t de Student para comparar las medias univariadas de dos muestras independientes.
- La función `wilcox.exact` del paquete `ExactRankTests` permite comparar las medianas univariadas mediante la prueba de Wilcoxon (corregida por empates).
- La función `hotelling.test` del paquete `Hotelling` permite realizar la prueba T^2 para la diferencia de medias multivariada de dos muestras independientes.

Lo primero que se hizo fue generar diferentes conjuntos de datos simulados provenientes de una distribución normal multivariada, los cuales deben presentar distintas características que incluyan: el tamaño de las dos muestras independientes (n_1 y n_2), el número de variables a analizar (p) y la variabilidad total de los datos. Es decir, por ejemplo, se fijan las siguientes características de un conjunto de datos: $n_1=50$ y $n_2=50$

observaciones, 8 variables y una variabilidad total igual a 6. Este tipo de conjunto de datos se genera por simulación de Monte Carlo, se repite r veces y en cada una de las repeticiones se evalúan los diferentes algoritmos de interés.

Se elaboraron funciones que permitieran desarrollar los algoritmos propuestos y evaluar su potencia.

En la tabla 1 se presenta una breve descripción de los conjuntos de datos que serán simulados. Para cada escenario propuesto se indican los tamaños de las muestras (n_1 y n_2) y el número de variables (p).

Tabla 1
Descripción de la estructura de los datos simulados

Escenario	n_1	n_2	p
1	20	20	3
2	20	40	3
3	100	100	3
4	100	150	3
5	200	200	5
6	200	250	5
7	500	500	7

Elaboración propia

En cada escenario se estimará la potencia de prueba de los diferentes algoritmos propuestos.

2.2 Métodos

Antes de definir los métodos y algoritmos que se desarrollaron, se presentan los términos utilizados:

k : Número de muestras independientes (en la presente investigación $k = 2$).

p : Número de variables cuantitativas del conjunto de datos.

n_1 : Tamaño de la muestra de la población 1.

n_2 : Tamaño de la muestra de la población 2.

$n = n_1 + n_2$: Tamaño total de la muestra.

\bar{X}_1 : Promedio de la muestra 1.

\bar{X}_2 : Promedio de la muestra 2.

S_1^2 : Varianza de la muestra 1.

S_2^2 : Varianza de la muestra 2.

S: Matriz de covarianza ponderada.

S_j: Matriz de covarianza de dimensión p x p del grupo j, j = 1,2,...,k.

A continuación se describen los aspectos teóricos presentados en Efron y Tibshirani (2011) y Higgins (2004) para poder entender los algoritmos desarrollados en la presente investigación.

2.2.1 Prueba t

En el caso univariado, para determinar si las medias poblacionales son iguales, es decir, $H_0: \mu_1 = \mu_2$, se puede hacer uso de la siguiente estadística, siempre que se cumplan los siguientes supuestos:

Las varianzas de las dos poblaciones son iguales (homogeneidad de varianzas).

La variable de interés en ambas poblaciones tiene distribución normal.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

Donde S_p es la desviación estándar ponderada definida por:

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

Con respecto al supuesto de normalidad, si n_1 y n_2 son suficientemente grandes, se cumple el Teorema Central del Límite y se puede proceder como si las poblaciones fueran normales. La prueba t es bastante robusta para las violaciones de la suposición de normalidad siempre que cada población sea relativamente simétrica con respecto a su media.

Una prueba equivalente puede ser utilizando el estadístico t^2 de la siguiente manera:

$$t^2 \sim F_{(1, n_1+n_2-2)}$$

2.2.2 Prueba de Wilcoxon-Mann-Whitney

En el caso univariado se puede utilizar para determinar si las medianas poblacionales son iguales, es decir, $H_0: Me_1 = Me_2$. A diferencia de la prueba t, no requiere del cumplimiento de normalidad para las muestras, pero sí que las muestras provengan de la misma distribución teórica.

$$Z = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$$

Donde:

U puede ser calculado mediante las siguientes expresiones:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \text{ y } U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

R_1 y R_2 son las sumas de los rangos de la primera y segunda muestra respectivamente.

Se elige el menor valor entre U_1 o U_2 para reemplazarlo en la prueba de la aproximación a la normal.

$$E(U) = \frac{n_1 n_2}{2}$$

$$V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

$$V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \left[\sum_{i=1}^g \frac{t_i^3 - t_i}{(n_1 + n_2)(n_1 + n_2 - 1)} \right]$$

Donde g es el número de grupos empatados y t_i el número de elementos empatados dentro de grupo i .

La primera expresión de $V(U)$ se debe utilizar cuando no hay empates (valores iguales) en los datos observados y la segunda expresión de $V(U)$ debe ser usada cuando hay empates.

2.2.3 Prueba T^2 de Hotelling

En el caso multivariado, para determinar si las p medias poblacionales son iguales, es decir, $H_0: \mu_1 = \mu_2$, donde ahora μ_1 y μ_2 son vectores de longitud p , se puede hacer uso del siguiente estadístico, si n_1 y n_2 son suficientemente grandes:

$$T^2 = (\bar{X}_1 - \bar{X}_2)^T \left[S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]^{-1} \quad (\bar{X}_1 - \bar{X}_2) \sim \chi^2_{(p)}$$

Donde S es la matriz de covarianza muestral ponderada

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 - 1) + (n_2 - 1)}$$

Si n_1 y n_2 no son suficientemente grandes, entonces es mejor utilizar el siguiente estadístico:

$$F = \frac{(n - p)}{p(n - 1)} T^2 \sim F_{(p, n_1 + n_2 - p - 1)}$$

2.2.4 Prueba de permutación

Una prueba de hipótesis empieza con un estadístico de prueba $\hat{\theta}$. Si la hipótesis nula, H_0 no es verdadera, esperamos observar grandes valores de $\hat{\theta}$ en comparación con H_0 verdadera.

Si se ha observado $\hat{\theta}$, el nivel de significancia alcanzado de la prueba (ASL por sus siglas en inglés) es definido como la probabilidad de observar al menos un valor grande cuando la hipótesis nula es verdadera, de tal manera que se puede definir así:

$$ASL = Prob_{H_0} \{ \hat{\theta}^* \geq \hat{\theta} \}$$

La cantidad $\hat{\theta}$ en la expresión del ASL es fijada como un valor observado (en el caso univariado se puede utilizar $\hat{\theta} = \bar{X}_1 - \bar{X}_2$), la variable aleatoria $\hat{\theta}$ tiene la distribución de la hipótesis nula, la distribución de $\hat{\theta}$ si H_0 es verdadera. La notación de * marca la diferencia entre la observación actual y un hipotético $\hat{\theta}^*$ generado de acuerdo a H_0 .

2.2.5 Algoritmo de la prueba estadística de permutación para dos muestras

Seleccionar B matrices independientes $g^*(1), g^*(2), \dots, g^*(B)$, cada una compuesta por n_1 y n_2 observaciones seleccionadas aleatoriamente de un conjunto de todos los $\binom{n}{n_1}$ posibles vectores.

Evaluar las replicaciones de permutación de $\hat{\theta}$ correspondiente a cada matriz de permutación,

$$\hat{\theta}^*(b) = S(g^*(b), v) \quad b = 1, 2, \dots, B$$

Aproximadamente ASL_{perm} es

$$ASL_{perm} = \#\{\hat{\theta}^*(b) \geq \hat{\theta}\} / B$$

En el caso de prueba de permutación multivariada para dos muestras, vectores multivariados son permutados entre los grupos de igual manera que las observaciones son permutadas entre grupos en una prueba de permutación univariada.

Por otro lado, la diferencia entre los algoritmos propuestos radica en la definición de $\hat{\theta}$ (también conocido como umbral), el cual permite el cálculo del ASL.

Como en la presente investigación se consideraron tres algoritmos para evaluar la diferencia significativa entre vectores de medias, los valores de o umbrales propuestos son:

- El estadístico T^2 de Hotelling.
- El máximo en valor absoluto de los estadísticos t .

$$t_{\max abs} = \max(|t_1|, |t_2|, \dots, |t_k|)$$

Es decir, se debe calcular el estadístico t de Student para cada una de las variables y considerar como umbral al máximo de ellos (Higgins, 2004).

- El máximo en valor absoluto de los estadísticos Z de Wilcoxon

$$Z = \max(|Z_1|, |Z_2|, \dots, |Z_k|)$$

Es decir, se debe calcular el estadístico Z de Wilcoxon-Mann-Whitney para cada una de las variables y considerar como umbral al máximo de ellos (Higgins, 2004).

En resumen la metodología de la presente investigación fue la siguiente:

- a) Generar diferentes conjuntos de datos (escenarios), considerando distintos tamaños de muestra (n_1 y n_2) y cantidad de variables (p).
- b) Aplicar los tres algoritmos a cada uno de los diferentes conjuntos de datos.
- c) Calcular la potencia de prueba de cada uno de los algoritmos propuestos. La potencia de prueba es obtenida repitiendo cada escenario $r = 50$ veces.

3. RESULTADOS Y DISCUSIONES

La tabla 2 muestra el número total de permutaciones para diferentes tamaños de muestra.

Tabla 2
Total de permutaciones para diferentes tamaños de n_1 y n_2

Escenario	n1	n2	Total de permutaciones
1	20	20	1,378465e + 11
2	20	40	4,191845e + 15
3	100	100	9,054851e + 58
4	100	150	6,063025e + 71
5	200	200	1,029525e + 119
6	200	250	6,798544e + 132
7	500	500	2,702882e + 299

Elaboración propia

Obtener todas las permutaciones posibles implicaría un gasto computacional excesivo, por lo que lo primero que se debe hacer es elegir un número (B) de permutaciones necesarias para optimizar el tiempo de procesamiento. Como la prueba de permutación utiliza en su procedimiento simulación de Monte Carlo, el ASL no es fijo. Ello implica un cierto nivel de variabilidad entre sus resultados. Se desea que esta variabilidad sea la mínima posible, por ejemplo, máximo 5 %. Para determinar el número óptimo de permutaciones se utilizará la función nm, donde se pueden evaluar diferentes valores de B (50, 100, 200, 500, 1000) para una cierta cantidad de repeticiones (r). En el presente trabajo se utilizó $r = 50$ repeticiones.

Se puede observar en los tres algoritmos que, a medida que el número de permutaciones utilizadas aumenta, la variabilidad disminuye. Esto ocurre en todos los escenarios propuestos. En la figura 1 se presenta lo mencionado para el primer escenario. Se puede deducir que un número adecuado de permutaciones suficiente es de 500. Utilizar más permutaciones implicaría un gasto computacional innecesario pues no reduciría significativamente la variabilidad entre resultados propuesta.

En la tabla 3 se presentan los diferentes valores de potencia de prueba. Para obtener estos resultados se utilizó simulación de Monte Carlo para generar $r = 50$ repeticiones diferentes en cada uno de los escenarios propuestos. Se definieron parámetros adecuados para la obtención de las matrices de datos que contenían a los números pseudoaleatorios. La potencia de prueba fue calculada como el número de veces que se rechazaba la hipótesis nula sobre la cantidad de repeticiones (r) realizadas.

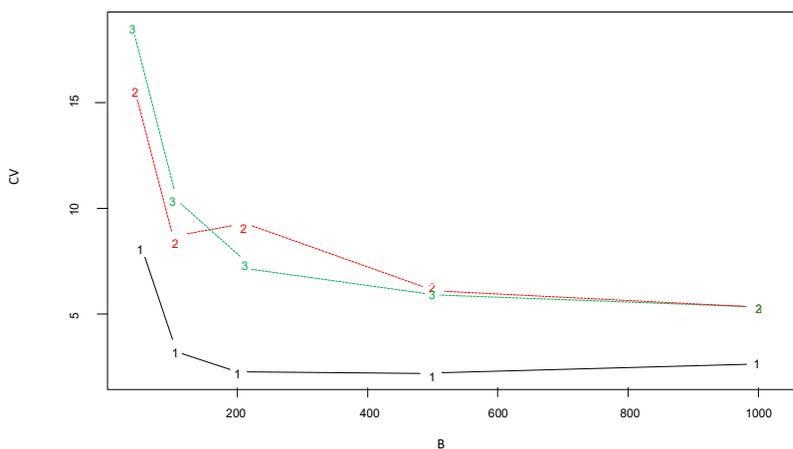


Figura 1. Número óptimo de permutaciones para $n_1 = 20$ y $n_2 = 20$

Elaboración propia

Se puede observar que el algoritmo basado en la prueba T^2 de Hotelling presenta una mayor potencia de prueba con respecto a los otros dos algoritmos propuestos. Esto se verificó estadísticamente a un nivel de significación de 0,05, mediante la Prueba de Friedman (p -valor = 0,002), donde se considera a los algoritmos en comparación como tratamientos y a los diferentes escenarios como bloques. Asimismo, al realizar las pruebas de comparación se encuentra que hay diferencia significativa entre los tres algoritmos.

Tabla 3

Potencia de prueba para los diferentes escenarios propuestos

Escenario	Algoritmo 1	Algoritmo 2	Algoritmo 3
1	0,97	0,87	0,85
2	0,95	0,88	0,86
3	0,96	0,89	0,89
4	0,94	0,91	0,90
5	0,95	0,92	0,93
6	0,96	0,93	0,92
7	0,98	0,94	0,93

Elaboración propia

En la figura 2 se puede visualizar que para los tres algoritmos propuestos las potencias de prueba se incrementan a medida que el número de variables y los tamaños de muestra aumentan.

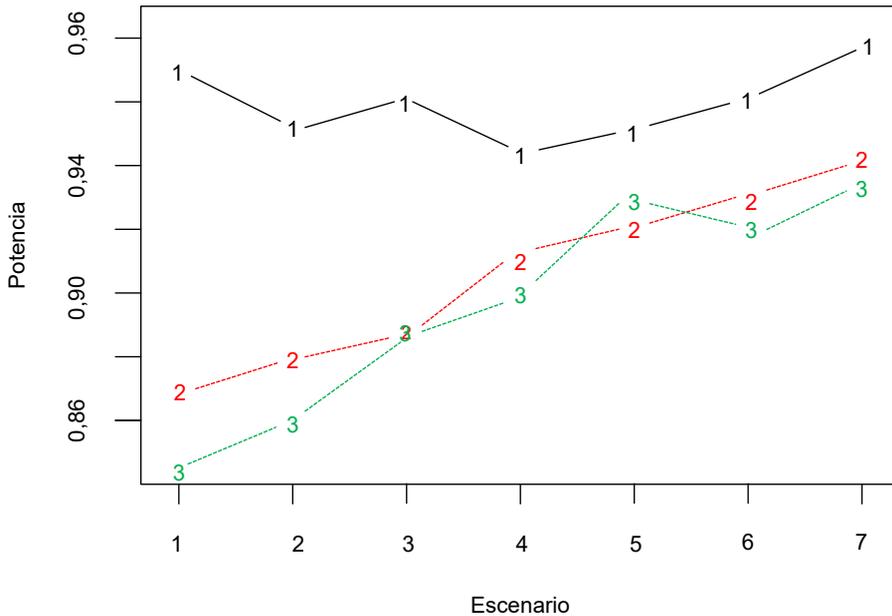


Figura 2. Potencia de prueba de los diferentes escenarios

Elaboración propia

4. CONCLUSIONES

El presente trabajo constituye un aporte en la investigación sobre la comparación de tres algoritmos para comparar las medias multivariadas de dos muestras independientes.

La prueba de permutación constituye un procedimiento alternativo a las pruebas paramétricas tradicionales que exigen el cumplimiento de ciertos supuestos que en la práctica no se cumplen.

Para investigaciones de este tipo es necesario generar diferentes escenarios que involucren criterios como: el tamaño de muestra, el número de variables consideradas en el estudio y la variabilidad representada en la matriz de covarianza. En esta investigación se utilizaron siete escenarios y solo una cantidad de permutaciones posibles para evitar un gasto computacional innecesario. Un número óptimo de permutaciones para obtener los resultados con una variabilidad máxima de 5 % es 500.

Se ha demostrado que de los tres algoritmos considerados, el que utiliza la prueba T^2 de Hotelling como criterio para determinar las diferencias significativas entre las medias poblacionales es el que presenta mayor potencia de prueba. Esto fue verificado estadísticamente mediante la prueba de Friedman.

Cabe resaltar que cuando se trabaja con una mayor cantidad de variables y los tamaños de muestra aumentan, la potencia de prueba se incrementa para los tres algoritmos.

Por otro lado, es necesario que en posteriores investigaciones similares se utilicen más escenarios para determinar con mayor precisión si para ciertos conjuntos de datos algunos de los algoritmos propuestos son más potentes que otros.

REFERENCIAS

- Blair, C., Higgins, J., Karniski, W. y Kromrey, J. (2010). A Study of Multivariate Permutation Tests Which May Replace Hotelling's T₂ Test in Prescribed Circumstances. *Journal Multivariate Behavioral Research* 29(2), pp. 141-163. doi: 10.1207/s15327906mbr2902_2
- Butar, F. y Park, J. (2008). Permutation Test for Comparing Two Populations. *Journal of Mathematical Sciences & Mathematics Education* 3(2), pp. 19-30.
- Chung, E. y Romano, J. (2011). *Asymptotically valid and exact permutation tests based on two-sample U-statistics*. (Technical report No. 2011-09). Stanford: Stanford University. Recuperado de <https://statistics.stanford.edu/sites/default/files/2011-09.pdf>
- Chung, E. y Romano, J. (2013). *Multivariate and Multiple Permutation Test*. (Technical report No.2013-05). Stanford: Stanford University. Recuperado de https://statistics.stanford.edu/sites/g/files/sbiybj6031/f/2013-05_0.pdf
- Efron, B. y Tibshirani, R. (2011): *An Introduction to the Bootstrap*. Nueva York: Chapman & Hall/CRC.
- Einsporn, R. y Habtzghi, D. (2013): Combining Paired and Two-Sample Data Using a Permutation. *Journal of Data Science* 11, pp. 767-779.
- Higgins, J. (2004). *An introduction to modern nonparametric statistics*. Londres: Thomson Learning.
- Samuh, M. (2017). Ranked Set Two Sample Permutation Test. *Statistica* 3, pp. 237-249.
- The R Project for Statistical Computing (3.6) [Software]. (2019). Recuperado de <https://www.r-project.org/>