Universidad de Los Andes

Proyecto de grado

El problema multivariado de las dos muestras: una aproximación desde métodos no paramétricos basados en grafos.

Andrés Felipe Hernández López

Asesor: José Ricardo Arteaga B.

Bogotá, Colombia
2022

Resumen

Andrés Felipe Hernández López

El problema multivariado de las dos muestras:
una aproximación desde métodos no paramétricos basados en
grafos.

La estadística no paramétrica estudia los modelos necesarios para poder hacer inferencias de una o varias muestras de datos cuya distribución subyacente es desconocida ex-ante, y de esta manera no se ajusta a los usuales criterios y supuestos que permiten usar los modelos paramétricos. Por otro lado, el problema de las dos muestras evalúa si las distribuciones de dos muestras de datos distintas son semejantes en algún criterio, tal como la ubicación o la escala. En este proyecto de grado estudiaremos algunos métodos no perimétricos para el problema de las dos muestras que se basan en grafos. El primero de estos es la generalización del test usual de las corridas de Wald-Wolfowitz propuesta por Friedman y Rafsky en 1979; el segundo es la generalización planteada al test de Kolmogorov-Smirnov en ese mismo artículo.; la tercera es el test de los k-vecinos más cercanos propuesto por Schilling en 1986; y el último es el nuevo test propuesto por Chen y Friedman en el 2017. Finalmente se comparará las potencias de dichos métodos cambiando la cantidad de datos de la muestra, el número de dimensiones y la cantidad k de vecinos más cercanos. Así se pondrá en evidencia las ventajas y debilidades de cada método en las distintas situaciones propuestas, con lo que se verá la versatilidad del método de Chen-Friedman frente a las alternativas de escala y ubicación.

Agradecimientos

Agradezco sobre todo a mi familia que siempre estuvo para mi, así como a mis amigos que escucharon mis retahílas de cuentos y datos curiosos de temas que muchas veces no eran sus favoritos. También quiero agradecer a Adolfo Quiroz por presentarme este gran curso de estadística no paramétrica y a José Ricardo por acompañarme en este proceso a pesar de que no era en una de sus áreas de experiencia.

Índice general

Introducción

El problema de las dos muestras es un problema clásico de la estadística, que nace del deseo de algún investigador de determinar si dos poblaciones distintas difieren en algún parámetro de interés. Por dar un ejemplo sencillo, si se está midiendo la altura de los niños y niñas de 14 años y se quiere determinar si los hombres son en promedio más altos que las mujeres. Por supuesto lo ideal sería medirlos a todos y responder inmediatamente la pregunta, pero en la practica esto puede presentar dificultades. Por ello lo que se hará es tomar una pequeña muestra de ambos grupos y hacer el experimento con estos. Pero el problema reside en ¿Qué hubiese sucedido si se tomaba una muestra distinta a la utilizada para el experimento? ¿Cambiaría el resultado? ¿Qué tan comparable es este resultado parcial contra el utópico resultado de medirlos a todos? Así, el problema de las dos muestras busca responder estas preguntas determinando si la diferencia entre las alturas de las muestras son estadísticamente significativas.

Otro problema muy usual es el de saber cuánto se alejan los datos de esta media. Volviendo a nuestro ejemplo, podríamos querer ver si la altura entre hombres presenta una mayor varianza que la altura de las mujeres. De esta manera un parámetro de ubicación como la media ya no aporta información relevante, por lo que tendríamos que buscar alguna medida de escala como la desviación estándar que nos hable de la dispersión de ambas muestras. A estos se les conoce como el problema de las dos muestras de ubicación y de escala respectivamente.

Existen métodos muy famosos para resolver estas preguntas. Por ejemplo el test t de Student el cual es muy bueno para detectar cambios en la media de dos muestras independientes cuándo los datos se comportan de una manera normal. De esta manera, se podría identificar con facilidad si la diferencia entre las medias observadas de la altura entre niños y niñas se debe a una razón estructural o fue una casualidad particular de las poblaciones estudiadas.

Pero ¿Qué pasa cuando queremos saber la diferencia de altura entre niños y niñas, pero teniendo en cuenta su edad, situación socio-económica o su nacionalidad? Este aumento de variables por cada uno de los niños analizados causaría que el problema pasa a tener varias dimensiones convirtiéndolo en uno uno de naturaleza multivariada. De esta manera los tests clásicos como la T de Student pierden sentido al vivir en un espacio unidimensional por lo que se generaron tests nuevos que solucionen esta necesidad como lo es la distribución T^2 de Hotelling, la cual surge de una generalización natural de la T de Student al caso multivariado.

No obstante hay casos en los que estos supuestos exigidos por las pruebas estadísticas clásicas no son fáciles de satisfacer. Para ilustrar, si se le pide evaluar a un grupo de estudiantes de la carrera de matemáticas la calidad de las materias cursadas o si se les pregunta qué tan felices los hace tener clase a las 6:30 am los resultados no van a ser de una naturaleza exacta. Ello en la medida que ninguno va a responder con una cifra exacta como .ªEstuve 78.7 % satisfecho con la clase.º "Mi felicidad de tener una clase así de temprano es 3.45 sobre 10". Para estos casos en los que la cuantificación se complejiza, o en los que se mezclan varias distribuciones, las pruebas clásicas se vuelven prácticamente obsoletas. Por ello, nacen los métodos no paramétricos que son nuestro objeto de estudio. Estos se especializan en aquellas situaciones en las que las distribuciones no son conocidas para aquel investigador que está intentando solucionar su pregunta con base en la estadística. No sobra decir que cuándo se tienen condiciones óptimas como dos muestras normales (o muy parecidas a distribuciones normales) los métodos que suponen distribuciones normales van a ser profundamente más eficientes y precisos que aquellos que no la suponen.

Similarmente al caso paramétrico, surgen métodos que estudian el problema de las dos muestras en el caso unidimensional y que así permiten hacerse una idea de si dos distribuciones, desconocidas para el investigador, son iguales o distintas. Algunos de estos son el de Wald-Wolfowitz y el de Kolmogorov-Smirnov, los cuales serán posteriormente estudiados en el capítulo de preliminares. Pero, volviendo al tema desarrollados en lineas anteriores, nuevamente aparece la necesidad de estudiar qué sucede cuando las preguntas se llevan a varias dimensiones. Así se tiene que Friedman y Rafsky logran en 1979 generalizar estos dos tests al caso multivariado utilizando grafos, los cuales inspiran a distintos autores a generar más métodos no paramétricos basados en distintos tipos de grafos que de alguna manera expresan la similitud de ambas muestras.

Este proyecto de grado está dirigido sobretodo a personas cuyos conocimientos de estadística no sean particularmente avanzados. Se espera que un lector con un conocimiento básico de la estadística clásica pueda entender la naturaleza y motivación detrás de estos métodos que se piensan presentar para motivarlo a estudiar más a profundidad esta área tan interesante de la estadística.

Capítulo 1

Preliminares

1.1. El problema

Ubicación: Sean $X_1, X_2 \ldots, X_m$ y $Y_1, \ldots, Y_n$ dos muestras independientes con distribuciones continuas $F(x)$ y $F(x - \Delta)$ respectivamente. Queremos probar la hipótesis nula $\mathbf{H_0} : \Delta = 0$ contra la hipótesis alternativa $\mathbf{H_a} : \Delta \neq 0$. Por simplicidad podemos asumir que $\mathbf{H_a} : \Delta > 0$.

Escala: Sean $X_1, \ldots, X_m$ y $Y_1, \ldots, Y_n$ dos muestras independientes con distribuciones continuas $F(x)$ y $F(\eta x)$ respectivamente para algún $\eta > 0$. Queremos probar la hipótesis nula $\mathbf{H_0} : \eta = 1$ contra la alternativa $\eta \neq 1$.

1.2. Distribución Libre

Sean $X_1, \ldots, X_n$ variables aleatorias con una distribución conjunta F, donde F pertenece a alguna familia $\mathcal{A}$ de posibles distribuciones conjuntas para $X_1, \ldots, X_n$. Sea $T(X_1, \ldots, X_n)$ un estadístico al que por simplicidad le diremos T. Randles y Wolfe definen en su libro *Introduction to the Theory of Nonparametric Statistics* una distribución libre sobre $\mathcal{A}$ como:

Definición 1.1 (Distribución libre)**.** Se dice que T es de distribución libre sobre $\mathcal{A}$ si la distribución de T es la misma para cada distribución conjunta en $\mathcal{A}$.

El siguiente ejemplo es tomado de este mismo libro y sirve para ilustrar de una manera sencilla este concepto.

Ejemplo 1.2. Sean $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ los estadísticos de orden de una muestra de tamaño n de una distribución normal con media 0 y varianza $0 < \sigma^2 < \infty$. Se considera el estadístico Q definido como:

$$Q = \frac{(X_{(n)} + X_{(1)})/2}{X_{(n)} - X_{(1)}}$$

Esto es, el punto medio dividido por el rango, se tiene que si tomamos $Z_1, \ldots, Z_n$ como una muestra i.i.d. con distribución normal estándar entonces

$$\left(\frac{X_1}{\sigma}, \cdots, \frac{X_n}{\sigma}\right) \overset{\mathcal{D}}{=} (Z_1, \ldots, Z_n)$$

Y por lo tanto se tiene que si lo ordenamos

$$\left(\frac{X_{(1)}}{\sigma}, \cdots, \frac{X_{(n)}}{\sigma}\right) \overset{\mathcal{D}}{=} (Z_{(1)}, \ldots, Z_{(n)})$$

Finalmente notamos que si multiplicamos por σ/σ obtenemos

$$Q = \frac{(X_{(n)} + X_{(1)})/2}{X_{(n)} - X_{(1)}} = \frac{(X_{(n)}/\sigma + X_{(1)}/\sigma)/2}{X_{(n)/\sigma} - X_{(1)}/\sigma} = \frac{(Z_{(n)} + Z_{(1)})/2}{Z_{(n)} - Z_{(1)}}$$

Por lo que la distribución de Q no depende de la varianza σ^2. Esto es que el estadístico Q es de distribución libre sobre la clase de distribuciones normales con media 0 y varianza común $\sigma^2 \in \mathbb{R}^+$.

Este es un ejemplo de un estadístico de distribución libre sobre alguna clase $\mathcal{A}$ que solo tiene un estilo de distribución conjunta. Por otro lado, cuando no se tiene la anterior situación, y $\mathcal{A}$ tiene varios estilos de distribuciones conjuntas, si se tiene un estadístico T que es de distribución libre sobre $\mathcal{A}$ se dice que este es no-paramétrico.

1.2.1. Estadísticos de conteo

Entre los estadísticos de distribución libre probablemente de los más fáciles de entender son los de conteo. Si se tiene una muestra de variables aleatorias independientes $X_1, \ldots, X_n$ con funciones de distribución $F_i(x)$,$i = 1, \ldots, n$, tales que $F_i(\theta) = p_0$ para algún $p_0 \in (0, 1)$ y un θ desconocido definimos el estadístico:

$$\Psi_i = \Psi(X_i - \theta_0) \tag{1.1}$$

Donde $\Psi(x) = 1, 0$ si $x >, \leq 0$ y θ_0 es un número real conocido. El siguiente es un teorema de Randles y Wolfe (1979).

Teorema 1.3. *Sean $\Psi_1, \ldots, \Psi_n$ como en (1.1) y sea $S(\Psi_1, \ldots, \Psi_n)$ un estadístico basado únicamente en $\Psi_1, \ldots, \Psi_n$. Entonces si $\theta = \theta_0$ se tiene que las $\Psi_1, \ldots, \Psi_n$ son variables aleatorias i.i.d. Bernoulli con parámetro $1 - p_0$ y $S(\Psi_1, \ldots, \Psi_n)$ es de distribución libre sobre la clase no-paramétrica $\mathcal{A}$ que consiste de todas las distribuciones conjuntas de variables aleatorias independientes y continuas para las cuáles el cuantil p_0 es igual a θ_0.*

Demostración. Como $\Psi_1, \ldots, \Psi_n$ son funciones de variables aleatorias independientes, entonces las funciones Ψ_i también deben ser independientes. Por otro lado, dada la definición de las Ψ_i

estas serán Bernoulli con parámetro $1 - F_i(\theta_0) = 1 - F_i(\theta) = 1 - p_0$. Entonces se tendrá por construcción que cualquier estadístico $S(\Psi_1, \ldots, \Psi_n)$ va a estar definido sobre variables Bernoulli con parámetro $1 - p_0$, esto dado que la distribución conjunta pertenezca a $\mathcal{A}$. □

Una función útil para implementar estos estadísticos de conteo es la de la suma

$$B = S(\Psi_1, \ldots, \Psi_n) = \sum_{i=1}^{n} \Psi_i$$

Este estadístico al ser una suma de variables Bernoulli está distribuido como una binomial con parámetros n y $(1 - p_0)$. Cuando $p_0 = 1/2$ B cuenta cuántos elementos de la muestra están sobre la mediana, por lo que se le llama el test de signos. A continuación mostramos un ejemplo de (Randles y Wolfe, 1979) utilizando este test.

Ejemplo 1.4. En 1969 (Barry) llevó a cabo un experimento de parapsicología para determinar si algunos individuos pueden usar su pensamiento para retrasar el crecimiento de unos hongos. Se usaron 10 individuos distintos para el experimento en el cual para cada individuo había un cultivo de control y otro experimental. Los sujetos no tenían permitido tocar los cultivos pero se le pedía que se concentraran sobre el cultivo experimental para retardar su crecimiento. Para esto entonces se define

$$X_i = U_i - V_i$$

Donde U_i y V_i son el crecimiento total para los grupos experimentales y de control. La hipótesis nula sería entonces que el proceso mental no tuvo ningún efecto por lo que X_i estaría distribuida simétricamente sobre 0, por lo que $\theta_0 = 0$ y $p_0 = 1/2$. Se obtuvieron como resultados que 9 de 10 individuos retrasaron el crecimiento de los cultivos experimentales de hongos, por lo que si usamos el test de signos obtenemos que $B = \sum_{i=1}^{10} \Psi_i$, por lo que obtenemos que dada H_0 la probabilidad de que $B \geq 9$ es:

$$P(B \geq 9 | H_0) = \sum_{i=9}^{10} \binom{10}{j} \left(\frac{1}{2}\right)^{10} = \frac{11}{1024} = 0{,}01074$$

Por lo que hay evidencia suficiente para rechazar la hipótesis nula a favor de la alternativa con un valor de significancia $\alpha = 0{,}02$ que indica que el pensamiento puede influir en el crecimiento de los hongos.

1.2.2. Estadísticos de rango

Sea $X_1, \ldots, X_n$ una muestra aleatoria con función de distribución continua $F(x)$, y sean $X_{(1)} \leq \cdots \leq X_{(n)}$ los respectivos estadísticos de orden. Dado este supuesto de continuidad en la distribución tenemos que $\mathbf{P}(X_i = X_{i+1}) = 0$ por lo que podemos asumir que lo siguiente está bien definido.

Definición 1.5 (Rango)**.** Se dice que X_i tiene rango R_i^* si $X_i = X_{(R_i^*)}$.

Sea $\mathbf{R}^* = (R_1^*, \ldots, R_n^*)$ el vector de rangos de $X_1, \ldots, X_n$ donde R_i^* es el rango de X_i entre los $X_1, \ldots, X_n$.

Teorema 1.6 (Randles y Wolfe)**.** *El vector de rangos está uniformemente distribuido en las permutaciones de* $(1, ..., n)$.

Demostración. Al haber $n!$ permutaciones de n basta mostrar que $\mathbf{R}^*$ es cada valor con probabilidad $1/n!$. Sea $r = (r_1, \ldots, r_n)$ una permutación cualquiera. Entonces

$$P(R^* = r) = P(Z_{d_1} < \ldots < Z_{d_n})$$

donde d_i es la posición de i en r. Por otro lado, como para una permutación α de $(1, \ldots, n)$ se cumple que $(Z_1, \ldots, Z_n) \overset{\mathcal{D}}{=} (Z_{\alpha_1}, \ldots, Z_{\alpha_n})$, por lo que se tiene que $P(Z_{d_1} < \ldots < Z_{d_n}) = P(Z_1 < \ldots < Z_n)$. Como r es arbitraria se obtiene el resultado. □

Corolario 1.7 (Randles y Wolfe)**.** *Si* $S(\mathbf{R}^*)$ *es un estadístico con base en* $X_1, \ldots, X_n$ *únicamente a través de* $\mathbf{R}^*$ *entonces* $S(\mathbf{R}^*)$ *es de distribución libre sobre la clase de las distribuciones conjuntas de n variables aleatorias i.i.d. continuas. A S se le llama un estadístico de rango.*

1.2.3. Estadístico de Wilcoxon

Ahora vamos a construir dos de los tests clásicos de la estadística no-paramétrica para el problema de las dos muestras que son el estadístico de la suma de los rangos y el de los rangos signados, ambos de Wilcoxon.

Sean $X_1, \ldots, X_m$ y $Y_1, \ldots, Y_n$ dos muestras aleatorias, $N = n + m$ y $Q = (Q_1, \ldots, Q_m), R = (R_1, \ldots, R_n)$ los vectores de rangos de cada muestra respectivamente con respecto a la muestra combinada de N observaciones. Note que en este caso si sabemos uno de estos vectores de rangos automáticamente sabemos el otro, por lo que basta estudiar únicamente uno de ellos. El test propuesto por Wilcoxon para este caso consiste en la suma de los rangos, es decir:

$$W = \sum_{i=1}^{n} R_i$$

Así al actuar únicamente a través de los rangos tenemos por el corolario que este es de distribución libre sobre las distribuciones conjuntas continuas. Por otro lado tenemos que por conteo la distribución discreta de W está dada para $w = \frac{n(n+1)}{2}, \frac{n(n+1)}{2} + 1, \ldots, \frac{n(2m+n+1)}{2}$

$$P(w) = \frac{t_{m,n}(w)}{\binom{N}{n}}$$

Donde $t_{m,n}(w)$ es el número subconjuntos sin ordenar de $[N]$ con n elementos tales que estos sumen w. De esta manera, un intervalo de confianza se construirá con base en esta distribución de W.

1.3. El problema de las dos muestras

Acá ya empezamos con los métodos que intentan solucionar nuestro problema principal: el de las dos muestras. A continuación presentaremos algunos estadísticos que precisamente son la base para la generalización que hacen Friedman y Rafsky (1979) al caso multivariado.

1.3.1. Estadístico de Wald-Wolfowitz

Para este estadístico se tienen nuevamente dos muestras $X_1, \ldots, X_m$ y $Y_1, \ldots, Y_n$ con distribuciones continuas F_X, F_Y respectivamente. Llamamos Z a la muestra combinada $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ y de esta manera asignamos etiquetas a $Z_{(1)}, \ldots, Z_{(N)}$ de la siguiente manera

$$Z_{(i)} = \begin{cases} X & \text{si } Z_{(i)} \in X \\ Y & \text{si } Z_{(i)} \in Y \end{cases}$$

De esta manera un arreglo con $m = 3$ y $n = 2$ podría quedar de la forma $XXYYX$, note que al ser continuas esto está bien definido pues la probabilidad de empate es 0.

Definición 1.8 (Corrida)**.** Se le llama una corrida a una sucesión de la misma etiqueta que está precedida y sucedida por una etiqueta distinta o por ninguna.

La motivación de la corrida es darse una idea de que tan homogénea es la mezcla entre ambas muestras, a un mayor número de corridas se tendrá una mayor mezcla. Por ejemplo siguiendo con $m = 3$ y $n = 2$ podríamos tener $XYXYX$ que tiene 5 corridas mientras que $XXXYY$ tiene 2. Cabe notar que ambos $XXXYY$ y $YYXXX$ tienen 2 corridas. De esta manera vemos que las corridas mientras que nos dan una idea de si ambas muestras tienen una media distinta, no nos brindan información alguna sobre cuál de ellas es mayor. Así tenemos una idea de cómo será el estadístico de Wald-Wolfowitz, el cual contará el número de corridas obtenidas dentro de los estadísticos de orden de la muestra combinada; y a este número se le llamará R. Precisamente, si el número de corridas es menor favorecerá la hipótesis alternativa se tendrá que rechazar la hipótesis nula con un nivel de significancia α cuando $R < c_\alpha$, donde c_α es el mayor punto tal que $P(R \leq c_\alpha | H_0) \leq \alpha$ (Gibbons y Chakraborti, 2011).

Más adelante se verá que el valor esperado y la varianza de R son:

$$\mu_R = \frac{2mn}{N} + 1$$
$$\sigma_R^2 = \frac{2mn(2mn - N)}{N^2(N-1)}$$

Lema 1.9 (Friedman y Rafsky, 1979)**.** *Se tiene que con* $N = n + m$

$$\frac{R - \mu_R}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Esta es consistente si $0 < \lim_{m,n\to\infty} m/n < \infty$

1.3.2. El estadístico de Kolmogorov-Smirnov

El test de Smirnov compara la diferencia entre las funciones de distribución empíricas de dos muestras $X_1, \ldots, X_m$ y $Y_1, \ldots, Y_n$ (Gibbons y Chakraborti, 2011). La función de distribución empírica, como bien dice su nombre, es la proporción de los datos de la muestra que son menores a un valor concreto. De esta manera si llamamos respectivamente $S(x)$ y $R(x)$ a las funciones de distribución empírica (fde) de las dos muestras y los estadísticos de orden son $X_{(1)}, \ldots, X_{(m)}$ y $Y_{(1)}, \ldots, Y_{(n)}$ tenemos que:

$$S(x) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ k/m & \text{si } X_{(k)} < x < X_{(k+1)} \\ 1 & \text{si } X_{(m)} < x \end{cases}$$

$$R(x) = \begin{cases} 0 & \text{si } x < Y_{(1)} \\ k/n & \text{si } Y_{(k)} < x < Y_{(k+1)} \\ 1 & \text{si } Y_{(n)} < x \end{cases}$$

Así las distribuciones empíricas nos dan un buen estimado de sus respectivas fda y por lo tanto si la hipótesis nula $H_0 : F_X = F_Y$ es cierta estas no deberían variar considerablemente la una con respecto a la otra. Así, si tomamos la muestra combinada $Z_1, \ldots, Z_N$ y definimos $s_i = S(Z_i), r_i = R(Z_i)$ se tiene una noción de cercanía entre ambas muestras, que es precisamente la que nos propone este test, dada por:

$$D = \max_{1\le i\le N} |s_i - r_i|$$

Hay que notar que, similarmente al caso anterior, la hipótesis alternativa es $H_A : F_X \neq F_Y$ por lo que es absoluta. En cambio, la zona de rechazo será la cola superior siendo el caso que para valores pequeños el test indica una similitud de ambas distribuciones. De esta manera, para un nivel de significancia α esta cola es $D > c_\alpha$. La distribución de D ha sido tabulada y el mismo Smirnov presentó que su distribución asintótica es (Gibbons y Chakraborti, 2011):

$$\lim_{m,n\to\infty} P\left(\sqrt{\frac{mn}{N}} D \le d\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$

1.4. Teoría de grafos

En esta sección se presentaran las definiciones necesarias de teoría de grafos para poder avanzar al tema que nos va a ocupar, que es la generalización de los tests anteriores y la presentación de otros para el caso multivariado. Estas definiciones son sacadas de «Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests». Así tenemos que:

Definición 1.10 (Grafo)**.** Un grafo G consiste de un conjunto de nodos N y un conjunto de pares de nodos $A \subset N \times N$ llamadas aristas.

Definición 1.11 (Arista)**.** Decimos que una arista une a dos nodos y que incide en ambos.

Definición 1.12 (Grado)**.** El grado de un nodo es la cantidad de aristas que inciden en él.

Definición 1.13 (Camino)**.** Un camino entre dos nodos es una sucesión alternante de nodos y aristas, la cual empieza y termina dichos nodos, todo el resto de nodos distintos y cada arista uniendo a los nodos que la preceden y la suceden.

Definición 1.14 (Longitud de un camino)**.** La longitud de un camino es la cantidad de aristas que contenga.

Definición 1.15 (Grafo conexo)**.** Si para todo par de nodos distintos del grafo existe un camino entre ellos.

Definición 1.16 (Ciclo)**.** Un ciclo es un camino que empieza y termina por el mismo nodo.

Definición 1.17 (Árbol)**.** Un árbol es un grafo conexo que no tiene ciclos.

Definición 1.18 (Subgrafo)**.** Un subgrafo de un grafo dado es un grafo cuyos nodos y aristas pertenecen en su totalidad al grafo dado.

Definición 1.19 (Subárboles)**.** Un subgrafo conexo de un árbol es un subárbol.

Definición 1.20 (Subgrafos disjuntos)**.** Dos subgrafos son disjuntos si no tienen ningún nodo en común.

Definición 1.21 (Subgrafo de cobertura)**.** Un subgrafo de cobertura de un grafo dado es un subgrafo cuyo conjunto de nodos es idéntico al del grafo dado.

Definición 1.22 (Árbol de cobertura)**.** Es un subgrafo de cobertura que a su vez es un árbol.

Definición 1.23 (Grafo de aristas ponderadas)**.** Es un Grafo cuyas aristas tienen asignado un número real.

Definición 1.24 (Mínimo árbol de cobertura)**.** El mínimo árbol de cobertura de un grafo de aristas ponderadas es un árbol de cobertura de dicho grafo cuya suma de los pesos de cada arista es mínima.

Definición 1.25 (Excentricidad)**.** La excentricidad de un nodo es la cantidad de aristas en el camino más largo que comience por ese nodo.

Definición 1.26 (Antípoda)**.** El nodo al final del camino más largo es la antípoda del nodo inicial.

Definición 1.27 (Diámetro)**.** El camino entre un nodo con excentricidad máxima y su antípoda es llamado el diámetro.

Definición 1.28 (Centro)**.** Un nodo para el cual la excentricidad sea mínima.

Definición 1.29 (Árbol enraizado)**.** Es un árbol que tiene designado uno de sus nodos como la raíz.

Definición 1.30 (Profundidad)**.** La profundidad de un nodo en un árbol enraizado es la longitud del único camino entre él y la raíz.

Definición 1.31 (Altura)**.** Es la máxima profundidad de cualquier nodo del árbol.

Definición 1.32 (Padre)**.** El padre de un nodo dado es el penúltimo nodo encontrado en el camino desde la raíz hasta el nodo dado. Todos los nodos salvo la raíz tienen un padre.

Definición 1.33 (Hijas)**.** Las hijas de un nodo son aquellos nodos que están unidos a él pero no son el padre.

Definición 1.34 (Ancestros)**.** Los ancestros de un nodo dado son todos aquellos nodos que se encuentran en el camino entre la raíz y él, excluyéndolo a él mismo.

Definición 1.35 (Descendientes)**.** Los descendientes de un nodo dado son todos aquellos para los cuales él es un ancestro.

Capítulo 2

Métodos multivariados basados en grafos

Vistos los preliminares, podemos proceder a estudiar el tema central de este proyecto: los métodos no paramétricos basados en grafos para el problema de las dos muestras. Empezaremos por las generalizaciones que le hicieron Friedman y Rafsky al test de Wald-Wolfowitz y al de Smirnov en (Friedman y Rafsky, 1979), continuaremos por el método basado en los k-vecinos más cercanos propuesto por Schilling y terminaremos por el nuevo test propuesto por Chen y Friedman en el 2017. En el capíyulo 3 se mediran las respectivas potencias de estos métodos y en el Apéndice B se pueden encontrar las implementaciones de estos métodos.

2.1. El método de Wald-Wolfowitz [Friedman y Rafsky, 1979]

En el caso univariado, para el test de Wald-Wolfowitz se ordena la muestra combinada Z. Esto se puede ver como un orden con base en minimizar la distancia entre los puntos de la muestra en $\mathbb{R}$. Así que para pasar este método al caso multivariado, Friedman y Rafsky proponen usar la generalización natural de esta idea: el árbol de mínima cobertura entre los puntos muestrales. Se construye de esta manera para mantener la noción de cercanía en $\mathbb{R}^n$, y en el caso de $n = 1$ el árbol sería el orden usual. Así el test de Wald-Wolfowitz se puede ejecutar mediante los siguientes pasos:

1. Construya el árbol de mínima cobertura de los puntos de la muestra combinada.
2. Borre las aristas que conecten puntos de distintas muestras.
3. Cuente la cantidad R de subárboles restantes, estos van a ser uno más que la cantidad de aristas borradas.

Así este método restringido al caso unidimensional se reduce al test de Wald-Wolfowitz original visto anteriormente.

De esta manera cuando vamos a calcular la esperanza y la varianza de R, se encuentra que la esperanza es la misma que en el caso univariado pero nos topamos que a la hora de encontrar la varianza esta depende de la topología del árbol. Procedemos a mostrar los resultados y luego a hacer los cálculos que también se pueden encontrar en Friedman y Rafsky, 1979.

Teorema 2.1 (Friedman y Rafsky, 1979)**.** *La media y la varianza de R son:*

$$\mu_{(R)} = \frac{2mn}{N} + 1$$

$$\sigma^2_{(R|C)} = \frac{2mn}{N(N-1)}\left[\frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)}(N(N-1)-4mn+2)\right]$$

Donde C es el número de parejas de aristas que comparten un nodo en común.

Demostración. Sean $a_1, \ldots, a_{N-1}$ las aristas del árbol de mínima cobertura, se define ζ_i para $1 \leq i \leq N-1$ como

$$\zeta_i = \begin{cases} 1 & \text{si } a_i \text{ une nodos de muestras distintas.} \\ 0 & \text{de lo contrario.} \end{cases}$$

Por lo tanto

$$\begin{aligned} R &= 1 + \sum_{i=1}^{N-1} \zeta_i \\ \mu_{(R)} &= 1 + \sum_{i=1}^{N-1} \mu_{(\zeta_i)} \end{aligned} \tag{2.1}$$

Al ser ζ_i Bernoulli notamos que $\mu_{(\zeta_i)} = P(\zeta_i = 1)$, lo que viene siendo la probabilidad de que los dos nodos que definen la arista tengan etiquetas distintas. Esto es

$$P(\zeta_i = 1) = 2\frac{mn}{N(N-1)}$$

Pues son igual de probables las etiquetas XY a YX. Esto junto con (2.1) nos da que:

$$\mu_{(R)} = \frac{2mn}{N} + 1 \tag{2.2}$$

Similarmente la varianza se puede calcular como:

$$\sigma^2_{(R)} = \sum_{i=1}^{N-1} \sigma^2_{(\zeta_i)} + \sum_{i<j} Cov[\zeta_i, \zeta_j] \tag{2.3}$$

De nuevo tenemos que, al ser ζ_i Bernoulli,

$$\sum_{i=1}^{N-1} \sigma^2_{(\zeta_i)} = \frac{2mn}{N} - \frac{4m^2n^2}{N^2(N-1)} \tag{2.4}$$

Por otro lado se tiene que

$$Cov[\zeta_i, \zeta_j] = E[\zeta_i\zeta_j] - E[\zeta_i]^2 \tag{2.5}$$

Para el primer término de esta resta esto no es más que $P(\zeta_i\zeta_j = 1)$, lo que significa que ambas aristas unan nodos de distintas muestras. Si estas aristas comparten un nodo en común esto significa que la etiqueta correspondiente debería ser XYX o YXY, mientras que si no comparten un nodo se puede tener 4 posibles etiquetas: XY,XY; XY,YX; YX,XY; YX,YX. Así se tiene que:

$$\mu_{(\zeta_i\zeta_j|\text{nodo común})} = \frac{mn}{N(N-1)} \tag{2.6}$$

$$\mu_{(\zeta_i\zeta_j|\text{ sin nodo común})} = \frac{4mn(m-1)(n-1)}{(N-1)(N-2)(N-3)} \tag{2.7}$$

Llamaremos a la cantidad de pares de aristas que comparten un nodo C, mientras que la cantidad total de aristas es $\binom{N-1}{2}$. Si combinamos esto con (2.2-2.7) obtenemos que:

$$\begin{aligned}\sigma^2_{(R|C)} =& \frac{2mn}{N} - \frac{4m^2n^2}{N^2(N-1)} + 2\Bigg\{C\left[\frac{mn}{N(N-1)} - \left(\frac{2mn}{N}+1\right)^2\right] \\ &+ \left(\frac{(N-2)(N-1)}{2} - C\right)\left[\frac{4mn(m-1)(n-1)}{(N-1)(N-2)(N-3)} - \left(\frac{2mn}{N}+1\right)^2\right]\Bigg\}\end{aligned} \tag{2.8}$$

Simplificando (2.8) obtenemos:

$$\sigma^2_{(R|C)} = \frac{2mn}{N(N-1)}\left[\frac{2mn-N}{N} + \frac{C-N+2}{(N-2)(N-3)}(N(N-1)-4mn+2)\right] \tag{2.9}$$

□

2.2. El método de Kolmogorov-Smirnov [Friedman y Rafsky, 1979]

El método de Kolmogorov-Smirnov depende fuertemente de la noción de orden en $\mathbb{R}$. Recordemos que este se basaba en la funciones de distribución empíricas de ambas X, Y. Así para el caso multidimensional se necesita ampliar esta idea y de esta manera poder de alguna manera extender la noción de orden para nuestros puntos muestrales convenientemente. Es aquí que Friedman y Rafsky proponen el pre-orden dirigido por altura (HDP por sus siglas en inglés). Este es un algoritmo que se encarga de recorrer el árbol, enraizándolo en un nodo con excentricidad máxima, de tal manera que pasa primero por los sub-árboles que tengan una menor altura , asignándole así un rango mayor a elementos que estén en árboles más altos. De esta manera se le aplica el

test univariado a los rangos obtenidos por estos métodos.
El método de Kolmogorov-Smirnov es conocido por ser bueno en ubicación pero no en escala, si se quiere ganar potencia en escala lo que se hace en una dimensión es ordenarlo por la distancia entre el punto y la mediana en la muestra combinada. Este es llamado el test radial de Kolmogorov-Smirnov, que a pesar de ganar potencia en alternativas de escala sufre la perdida de gran parte de su potencia en las alternativas de ubicación. La generalización natural de este método es la de asignar como raíz a un nodo central del árbol de mínima cobertura y asignar rangos a los puntos dependiendo de su profundidad, así los puntos que tengan una mayor profundidad con respecto al centro (esto es, los nodos periféricos) tendrán un rango mayor. El algoritmo para recorrer un árbol por medio del HDP se define recursivamente como sigue (Friedman y Rafsky, 1979):

1. Visite la raíz
2. Recorra por HDP en orden ascendente de altura los subárboles enraizados en las hijas de la raiz. (Se resuelven los empates visitando primero aquellas hijas que estén más cercanas en distancia euclidiana a la raíz).

Se puede notar que en la generalización de ambos métodos la lista de rangos obtenida no es única, depende de la raíz escogida con la que se inicia el algoritmo. En el método clásico se pueden escoger al menos dos pues en algún camino que sea diámetro ambas antípodas tienen excentricidad máxima, mientras que en el método radial pueden haber hasta dos nodos centrales. Friedman y Rafsky dicen que en su experiencia la elección de la raíz no cambia significativamente el resultado, esto pues se hace en esencia lo mismo sin importar la raíz.

2.3. El método de los k-vecinos más cercanos [Schilling, 1986b]

El método de las corridas de Wald-Wolfowitz es conocido por ser de los menos potentes (Friedman y Rafsky, 1979), esto se debe a que a pesar de ser muy natural e intuitivo trabaja con muy poca información sobre los datos. Es así que el método de los k-vecinos más cercanos propuesto por Schilling en 1986b intenta no usar únicamente el árbol de mínima cobertura, que por su estructura no permite que hayan ciclos, sino usar un grafo con una topología más compleja que pueda proveer una mayor cantidad de información al tener más aristas. Esto se ilustra en el Apéndice A. con la figura de un árbol de mínima cobertura contra un grafo de k-vecinos más cercanos.

Para conseguir esto, primero se necesita de alguna métrica $||\cdot||$ en el espacio para definir al k-vecino más cercano a algún punto.

Definición 2.2 (K-vecino más cercano (Schilling, 1986b))**.** Sea $Z = Z_1, \ldots, Z_N$ la muestra combinada. El k-vecino más cercano a Z_i, al cual denotaremos como $NN_i(k)$, se define como el Z_j que satisface que $||Z_{j'} - Z_i|| < ||Z_j - Z_i||$ para exactamente $k-1$ valores de j' $(1 \leq j' \leq N, j' \neq i, j)$.

La motivación del test es una muy natural, medir la proporción de elementos cercanos que se encuentren en la misma muestra. Bajo H_0 se puede esperar que la muestra combinada este de alguna manera "mezclada", así la proporción de vecinos que pertenezcan a la misma muestra va a ser similar a la proporción de vecinos que pertenecen a la otra muestra. En cambio bajo H_a se puede esperar que las distribuciones estén "separadas"la una de la otra, de esta manera la proporción de los vecinos que estén dentro de la misma muestra será alta.

Ahora definimos la siguiente función indicatriz con base en si el r-vecino pertenece a la misma muestra que Z_i, esto es:

$$I_i(r) = \begin{cases} 1 & \text{si } NN_i(r) \text{ pertenece a la misma muestra que } Z_i. \\ 0 & \text{de lo contrario.} \end{cases}$$

Así se define el test de los k-vecinos más cercanos como:

$$T_{k,N} = \frac{1}{Nk} \sum_{i=i}^{N} \sum_{r=1}^{k} I_i(r)$$

De esta manera $T_{k,N}$ es la proporción deseada de k-vecinos más cercanos a un punto pertenecientes a la misma muestra que pertenecen a la misma muestra que ese punto.

Bajo los supuestos que $\lambda_1 = \lim_{m\to\infty} \frac{m}{N}$ y $\lambda_2 = \lim_{n\to\infty} \frac{n}{N}$ existen. Se consideran los siguientes eventos:

1. $NN_1(r) = Z_2$, $NN_2(s) = Z_1$
2. $NN_1(r) = NN_2(s)$

Se dice que Z_1 y Z_2 son vecinos mutuos si para algún r, s se cumple el primer caso y que comparten un vecino en caso de que el segundo ocurra. Las probabilidades de que ocurran los eventos 1 y 2 se les llama $p_{1\{2\}}(r, s)$ respectivamente. Estos valores son por lo general dificiles de computar, pero los valores $Np_i(r, s)$, a los que llamaremos $p_i'(r, s)$, tienen límites que son independientes de H_0, los cuales son complejos pero computables y se estudian más a fondo en «Mutual and Shared Neighbor Probabilities: Finite- and Infinite-Dimensional Results» (Schilling, 1986a). Así tenemos que el promedio de estos valores es:

$$\overline{p}_i' = \frac{1}{k^2} \sum_{r,s=1}^{k} p_i'(r, s)$$

Uno de los resultados más importantes de Schilling,(1986b), es que la distribución asintótica de $T_{k,n}$ depende únicamente de $k, \lambda_1, \lambda_2, \overline{p}_1'$ y $\overline{p}_2'$ de la siguiente manera:

Teorema 2.3 (Schilling, 1986b)**.** *Si λ_1, λ_2 existen, entonces se tiene que:*

$$(\sqrt{Nk})\frac{T_{k,N} - \mu_k}{\sigma_k} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

Donde se tiene que:

$$\begin{aligned}
\mu_k &= \lim_{N\to\infty} E[T_{k,N}] \\
&= \lambda_1^2 + \lambda_2^2 \\
\sigma_k &= \lim_{N\to\infty} (Nk) var(T_{k,N}) \\
&= \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2 k\overline{p}_1' - \lambda_1\lambda_2(\lambda_1 - \lambda_2)^2 k(1 - \overline{p}_2')
\end{aligned}$$

Demostración. La demostración de este teorema se puede encontrar en el apéndice de Schilling, 1986b. □

Se da referencia de dos teoremas más que se estudian en Schilling, 1986a. Estos hablan del comportamiento asintótico de $\overline{p}_1'$ y $\overline{p}_2'$ y sus demostraciones se pueden encontrar en ese mismo artículo.

Teorema 2.4 (Schilling, 1986a)**.** *Para todo d, el límite $\lim_{k\to\infty} k\overline{p}_1'$ existe y es 1*

Teorema 2.5 ('Schilling, 1986a)**.** *Para todo par de enteros positivos r, s, el límite $\lim_{d\to\infty} p_2'(r,s)$ existe y es igual a 1*

De esta manera obtenemos que para una k no muy pequeña

$$\sigma^2 = \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2\left[1 - \binom{2k}{k}2^{-2k}\right] = \lambda_1\lambda_2 + 4\lambda_1^2\lambda_2^2$$

2.4. El método de Chen Friedman

El último método que se va a presentar es el propuesto por Chen y Friedman en el 2017. Los métodos anteriores presentaban problemas con respecto a la ubicación o a la escala, así que se propone uno que tenga una buena potencia con respecto a ambas situaciones.

En este nuevo método se usa el sentido usual de cercanía que se venía usando para construir un grafo de similitud, pero se incorpora un patrón común que se había pasado por alto anteriormente (Chen y Friedman, 2017). Cuándo la alternativa es de ubicación, el número de aristas entre la misma muestra va a ser más alto que su valor esperado mientras que en alternativas de escala se va a tener que para la muestra con una menor escala va a haber más aristas entre esa misma muestra mientras que para la muestra con la escala mayor va a haber menos aristas entre la misma muestra, este es un efecto que solo se amplifica al subir el número de dimensiones pues el volumen de un espacio d-dimensional aumenta exponencialmente con d (Chen y Friedman, 2017).

Así el nuevo test busca atrapar estas dos desviaciones y así tener una buena potencia sin importar si la alternativa es de escala o ubicación. Así si $a = (i, j)$ es una arista en nuestro grafo de similitud G, se definen:

$$J(a) = \begin{cases} 0 & \text{si } Z_i \text{ y } Z_j \text{ pertenecen a muestras distintas} \\ 1 & \text{si ambas } Z_i \text{ y } Z_j \text{ pertenecen a la muestra } X \\ 2 & \text{si ambas } Z_i \text{ y } Z_j \text{ pertenecen a la muestra } Y \end{cases}$$

$$R_k = \sum_{a \in G} I(J(a) = k)$$

Donde $I(x)$ es la función indicatriz. Entonces se tiene que R_0 es la cantidad de aristas que tienen nodos en muestras distintas, R_1 es la cantidad de aristas que tienen ambos nodos en la muestra X y R_2 es la cantidad de aristas que tienen ambos nodos en la muestra Y. Con esto se define el nuevo test como:

$$S = (R_1 - \mu_{(R_1)}, R_2 - \mu_{(R_2)})\Sigma^{-1}\begin{pmatrix} R_1 - \mu_{(R_1)} \\ R_2 - \mu_{(R_2)} \end{pmatrix} \tag{2.10}$$

Donde $\mu_{(R_1)}, \mu_{(R_2)}$ son los valores esperados de R_1, R_2 y Σ es la matriz de covarianza de $(R_1, R_2)^T$ bajo la distribución nula de permutación. Así se tiene que en caso de ser una alternativa de ubicación la cantidad de aristas con nodos pertenecientes a la misma muestra va a ser mayor a su valor esperado por lo que S sería más grande. Por otro lado, en caso de ser una alternativa de escala la muestra que tenga la varianza menor tendrá una mayor cantidad de aristas con nodos dentro de la misma muestra que su valor esperado, mientras que para la muestra con la varianza mayor la cantidad será menor. De esta manera tendremos que S va a ser más grande en caso de una alternativa de ubicación o escala por lo que será sensible a ambas.

Además se tiene que bajo ciertas condiciones sobre el tamaño del grafo de similitud y su topología se tiene que:

$$S = (R_1 - \mu_{(R_1)}, R_2 - \mu_{(R_2)})\Sigma^{-1}\begin{pmatrix} R_1 - \mu_{(R_1)} \\ R_2 - \mu_{(R_2)} \end{pmatrix} \xrightarrow{\mathcal{D}} \chi^2_2$$

Finalmente se puede obtener los valores de μ_1, μ_2 y Σ con algo de combinatoria, la demostración de este lema es tomada del Apéndice A.1 de Chen y Friedman, 2017.

Lema 2.6 (Chen y Friedman, 2017)**.** *Si R_1 y R_2 son definidos como arriba y $|G|$ denota la cantidad de aristas del grafo y C la cantidad de aristas que comparten un nodo, se tiene que:*

$$\mu_{(R_1)} = |G|\frac{m(m-1)}{N(N-1)}$$
$$\mu_{(R_2)} = |G|\frac{n(n-1)}{N(N-1)}$$
$$\Sigma_{11} = \mu_{(R_1)}(1-\mu_{(R_1)}) + 2C\frac{m(m-1)(m-2)}{N(N-1)(N-2)} + (|G|(|G|-1)-2C)\frac{m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)}$$
$$\Sigma_{22} = \mu_{(R_2)}(1-\mu_{(R_2)}) + 2C\frac{n(n-1)(n-2)}{N(N-1)(N-2)} + (|G|(|G|-1)-2C)\frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)}$$
$$\Sigma_{12} = |G|(|G|-1)-2C)\frac{mn(m-1)(n-1)}{N(N-1)(N-2)(N-3)} - \mu_{(R_1)}\mu_{(R_2)}$$
$$\Sigma_{21} = \Sigma_{12}$$

Demostración. Bajo la distribución nula de permutación se tiene que por la naturaleza Bernoulli de R_1:

$$\mu_{(R_1)} = \sum_{a\in G} P(J(a)=1) = \sum_{(i,j)\in G} P(Z_i \in X, Z_j \in X) = |G|\frac{m(m-1)}{N(N-1)} \tag{2.11}$$

Como en la demostración de 2.1, nuevamente se tiene que dependemos de la cantidad de pares de aristas C que comparten un nodo para el siguiente cálculo.

$$\begin{aligned}
\mu_{(R_1^2)} &= \sum_{a_1,a_2\in G} P(J(a_1)=1, J(a_2)=1) \\
&= \sum_{(i,j)\in G} P(Z_i \in X, Z_j \in X) + \sum_{\substack{(i,j),(i,k)\in G \\ i\neq k}} P(Z_i \in X, Z_j \in X, Z_k \in X) \\
&+ \sum_{\substack{(i,j),(k,l)\in G \\ i,j,k,l \text{ todos diferentes}}} P(Z_i \in X, Z_j \in X, Z_k \in X,\ Z_l \in X) \\
&= \mu_{(R_1)} + 2C\frac{m(m-1)(m-2)}{N(N-1)(N-2)} + (|G|(|G|-1)-2C)\frac{m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)}
\end{aligned} \tag{2.12}$$

Usando 2.11 y 2.12 obtenemos que

$$\Sigma_{11} = \mu_{(R_1^2)} - \mu_{(R_1)}^2 \tag{2.13}$$

Similarmente a esto obtenemos que

$$\begin{aligned}\mu_{(R_1R_2)} &= \sum_{a_1,a_2\in G} P(J(a_1)=1, J(a_2)=2) \\ &= \sum_{\substack{(i,j),(k,l)\in G \\ i,j,k,l \text{ todos diferentes}}} P(Z_i\in X, Z_j\in X, Z_k\in Y,\ Z_l\in Y) \\ &= (|G|(|G|-1)-2C)\frac{m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)} \qquad (2.14)\end{aligned}$$

Finalmente se tiene que

$$\Sigma_{12} = \mu_{(R_1R_2)} - \mu_{(R_1)}\mu_{(R_2)} \qquad (2.15)$$

□

Capítulo 3

Potencias

Finalmente se evaluaran las potencias de los métodos estudiados, estos se presentaron para poder cubrir una variedad alta de posibles alternativas. En el cuadro 3.1 se tiene que se tabuló las potencias con respecto a una alternativa de ubicación de los métodos que están basados en el árbol de mínima cobertura. En 3.2 se tiene un cuadro de las potencias de los grafos basados en los k-vecinos más cercanos con respecto a una alternativa de escala. Para ambos, se usaron datos sacados de una distribución normal multivariada con media 0 y matriz de covarianza I, se usaron como dimensiones a $(2, 10, 30, 50, 70, 90, 100)$ con respectivas desviaciones a la media de $\Delta = (0{,}6, 0{,}8, 1{,}1, 1{,}4, 1{,}7, 2, 2)$, donde las medias difieren Δ con respecto a la distancia usual. También se tiene que se implementaron con n=m=50,100,250 para representar la potencia en medida que aumentan los datos.

	d	2	10	30	50	70	90	100
Método	N \ Δ	0,6	0,8	1,1	1,4	1,7	2	2
Wald Wolfowitz	50	16	20	23	36	43	56	51
	100	36	30	50	62	79	88	88
	250	60	66	86	97	100	100	100
Smirnov	50	36	13	9	12	15	20	19
	100	61	26	17	15	18	26	28
	250	94	38	20	34	40	51	50
Smirnov Radial	50	19	32	45	45	57	52	59
	100	22	11	8	7	9	23	22
	250	39	52	89	91	92	90	93
S de Chen Friedman(MST)	50	11	13	13	20	34	39	36
	100	14	27	21	41	67	70	68
	250	37	48	32	85	93	100	100

CUADRO 3.1: Tabla de potencias: cantidad de intentos sobre 100 que generaron un p-valor menor a 0.05 con respecto a una alternativa de ubicación

Se puede ver como en dimensiones bajas el test de Wald Wolfowitz es insuficiente, mientras que el de Smirnov lo domina. Pero, esto se revierte en dimensiones altas. Ahí el test de Wald-Wolfowitz se vuelve considerablemente más fuerte mientras que el de Smirnov cae. El test de Chen y Friedman

por ahora se comporta similarmente al de la corridas de Wald Wolfowitz, lo cual no es sorpresivo dado que se construyen desde el mismo grafo.
Para la siguiente tabla se usaron datos sacados de la misma distribución, y se varió de igual manera la escala, el número de datos y las dimensiones. Por otro lado, se consideraron los grafos construidos con $(3, 5, 10, 20)$-vecinos y se midieron sus respectivas potencias.

			2	10	30	50	70	90	100
Método	N \Δ	k	0,6	0,8	1,1	1,4	1,7	2	2
K-Vecinos más cercanos	50	3	30	19	32	46	64	75	71
		5	28	24	38	59	71	86	86
		10	36	40	42	28	82	92	93
		20	42	40	57	71	90	92	92
	100	3	47	53	64	75	95	100	93
		5	48	49	74	83	100	100	98
		10	64	71	86	96	100	99	100
		20	77	76	87	100	100	100	100
S de Chen y Friedman (KNN)	50	3	26	21	26	46	61	69	74
		5	31	39	25	36	67	77	78
		10	41	37	49	59	72	93	89
		20	38	48	53	71	89	99	97
	100	3	41	47	49	81	96	99	96
		5	53	51	61	73	89	96	96
		10	65	63	73	94	94	99	93
		20	75	77	90	98	99	100	99

CUADRO 3.2: Tabla de potencias: cantidad de intentos sobre 100 que generaron un p-valor menor a 0.05 con respecto a una alternativa de ubicación

En esta se evidencia que el test de los k-vecinos más cercanos no solo tiene una buena potencia en dimensiones bajas sino que mejora rápidamente mientras crecen el número de vecinos o de dimensiones dada la alternativa de ubicación. Se hace notar el aumento de la información disponible de este grafo contra el árbol de mínima cobertura que logra abarcar menos. Se puede notar que ambos métodos se comportan parecido independiente de la dimensión o del k escogido.

Después se evaluaron los distintos métodos contra alternativas netamente de escala. Así que nuevamente se tomaron los datos de una distribución normal multivariada con media 0 y matriz de covarianza I. Esta vez se multiplicó la matriz de covarianza de una de las muestras por una variable σ para evaluar las potencias de dichos métodos. Se usaron $(2, 5, 10, 20, 50, 75, 100)$ dimensiones y se ajustó la covarianza respectivamente para cada caso con unos factores de $(1{,}4, 1{,}25, 1{,}2, 1{,}15, 1{,}15, 1{,}12, 1{,}1)$ respectivamente.

Con la tabla 3.3 se hace evidente que en este contexto el método de las corridas de Wald-Wolfowitz y el método clásico de Smirnov presentan una caída considerable en su desempeño, convirtiéndolos en casi inútiles. Por otro lado, el método radial de Smirnov gana potencia y se vuelve muy confiable sobre todo en dimensiones altas. Similarmente vemos que a diferencia del test de las corridas de

	d	2	5	10	20	50	75	100
Método	N \ σ	1,4	1,25	1,2	1,15	1,15	1,12	1,1
Wald Wolfowitz	50	7	6	7	5	8	4	4
	100	5	9	11	3	8	6	6
	250	14	11	4	11	8	2	4
Smirnov	50	3	2	4	7	6	7	1
	100	8	4	3	2	1	4	4
	250	8	5	3	4	7	6	6
Smirnov Radial	50	7	38	48	63	73	78	75
	100	21	15	23	16	19	23	20
	250	50	51	90	95	100	98	97
S de Chen Friedman(MST)	50	7	8	16	22	39	41	29
	100	7	10	16	25	47	48	43
	250	7	22	39	56	99	85	87

CUADRO 3.3: Tabla de potencias: cantidad de intentos sobre 100 que generaron un p-valor menor a 0.05 con respecto a una alternativa de escala

Wald-Wolfowitz el métodos propuesto por Chen y Friedman conserva potencia sobre todo en dimensiones más altas. Lo que lo hace un método confiable independientemente de la alternativa a tratar.

A continuación se revisó la alternativa de escala pero usando los métodos que están basados en el grafo de los k-vecinos más cercanos.

			2	5	10	20	50	75	100
Método	N \ σ	k	1,4	1,25	1,2	1,15	1,15	1,12	1,1
K-Vecinos más cercanos	50	3	7	12	7	6	5	5	1
		5	8	6	4	1	0	0	0
		10	7	9	5	2	1	2	4
		20	7	2	2	4	2	2	0
	100	3	21	4	7	4	0	1	3
		5	11	10	8	3	6	3	2
		10	14	10	5	6	4	1	3
		20	13	6	3	4	3	6	4
S de Chen y Friedman (KNN)	50	3	21	51	71	81	33	22	19
		5	23	44	61	84	80	66	61
		10	28	43	57	70	99	96	95
		20	35	45	61	68	93	87	91
	100	3	24	47	81	75	10	3	1
		5	26	54	73	95	31	15	5
		10	25	53	58	89	92	71	70
		20	29	52	82	88	100	99	99

CUADRO 3.4: Tabla de potencias: cantidad de intentos sobre 100 que generaron un p-valor menor a 0.05 con respecto a una alternativa de escala

En esta última tabla 3.4 podemos observar cómo el método de los k-vecinos de Schilling sufre al cambiar a una alternativa de escala y sobre todo al aumentar la dimensión. Este método que era tan potente para las alternativas de ubicación pierde su sentido, tal y cómo lo decían Chen y Friedman, en dimensiones altas cuando hay un cambio de escala. Por otro lado, cuando se usa el método de Chen y Friedman basandonos en el mismo grafo podemos observar cómo la potencia es considerablemente alta cuando se tiene un k no tan bajo, este también va mejorando a medida que suben las dimensiones. De esta manera tenemos que al comparar estos dos métodos tenemos un desempeño similar a la hora de evaluar alternativas de ubicación y tenemos un desempeño superior del nuevo método frente a las alternativas de escala.

Conclusión

Cada problema al que se enfrenta la estadística tiene una naturaleza distinta y por esto se han creado y refinado diversos métodos; para que estos puedan ser solventados de la mejor manera posible. A lo largo de este proyecto de grado se estudiaron algunos de los distintos métodos basados en grafos que han surgido para atacar precisamente estos problemas. Se analizaron sobre todo sus construcciones y se señaló su comportamiento asintótico. Así también, se evaluaron los distintos métodos y se buscaron cuáles eran sus potencias contra las alternativas de ubicación y escala. Encontramos que, por lo general, los métodos suelen ser potentes en alguna de las dos situaciones pero no en ambas. No obstante, se halló que el nuevo método, propuesto por Chen y Friedman, no cumple esta premisa, dado que resultó ser potente en ambas situaciones. Aparte del test mencionado anteriormente, se estudiaron también: (1) el test de Wald Wolfowitz; (2) el test clásico de Smirnov; (3) el test de los k-vecinos más cercanos de Schilling; y (4) el test radial de Smirnov. Teniendo en cuenta el enfoque original del proyecto de grado, se puede concluir que el nuevo test de Chen y Friedman provee los mejores resultados - dado que arroja las potencias más altas - en todas las situaciones evaluadas en este trabajo, convirtiéndolo en el más completo de los tests analizados.

Apéndice A

Figuras

En este apéndice se mostrará la representación gráfica del árbol de mínima cobertura y el de los k-vecinos más cercanos para una muestra de tamaño $N = 16$.

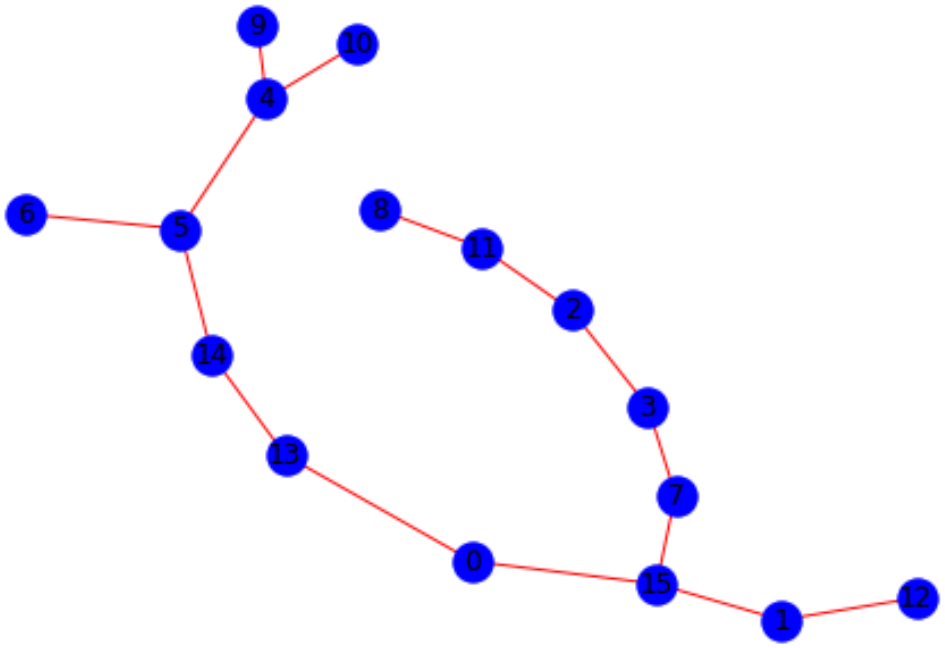

FIGURA A.1: Representación del árbol de mínima cobertura para una muestra de 16 datos.

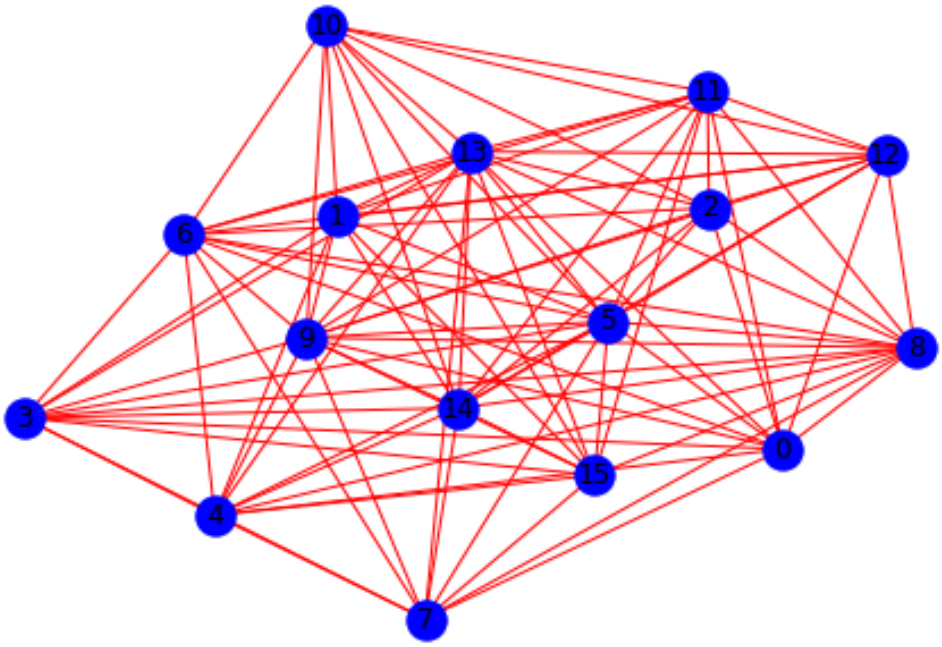

FIGURA A.2: Representación de los k-vecinos más cercanos para la misma muestra

Apéndice B

Implementación computacional de los métodos

Los paquetes que se utilizaron fueron los siguientes:

```
import math
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import numpy.random as rd
import networkx as nx
import operator
from sklearn.neighbors import kneighbors_graph
from scipy.sparse.csgraph import minimum_spanning_tree
```

Para la implementación del estadístico de Wald-Wolfowitz se implementó el código de Monaco, 2015. Ideas de este código se utilizaron para los demás métodos que si son de la autoría del autor.

```
def mst_edges(V, k):
    """
    Construct the approximate minimum spanning tree from vectors V
    :param: V: 2D array, sequence of vectors
    :param: k: int the number of neighbor to consider for each vector
    :return: V ndarray of edges forming the MST
    """

    # k = len(X)-1 gives the exact MST
    k = min(len(V) - 1, k)

    # generate a sparse graph using the k nearest neighbors of each point
    G = kneighbors_graph(V, n_neighbors=k, mode='distance')
```

```
    # Compute the minimum spanning tree of this graph
    full_tree = minimum_spanning_tree(G, overwrite=True)

    return np.array(full_tree.nonzero()).T\\
def ww_test(X, Y):
    """
    Multi-dimensional Wald-Wolfowitz test
    :param X: multivariate sample X as a numpy ndarray
    :param Y: multivariate sample Y as a numpy ndarray
    :param k: number of neighbors to consider for each vector
    :return: W the WW test statistic, R the number of runs
    """
    m, n = len(X), len(Y)
    N = m + n

    XY = np.concatenate([X, Y]).astype(np.float)

    # XY += np.random.normal(0, noise_scale, XY.shape)

    edges = mst_edges(XY, 10)

    labels = np.array([0] * m + [1] * n)

    c = labels[edges]
    runs_edges = edges[c[:, 0] == c[:, 1]]

    # number of runs is the total number of observations minus edges within each run
    R = N - len(runs_edges)

    # expected value of R
    e_R = ((2.0 * m * n) / N) + 1

    # variance of R is _numer/_denom
    _numer = 2 * m * n * (2 * m * n - N)
    _denom = N ** 2 * (N - 1)

    # see Eq. 1 in Friedman 1979
    # W approaches a standard normal distribution
    W = (R - e_R) / np.sqrt(_numer/_denom)

    return (W,stats.norm.cdf(W))
```

Para el método de Smirnov el código utilizado fue el siguiente:

```
def order_height(G,succ_list,root):
    temp=G
    for succ in succ_list:
        temp.remove_edge(root,succ)
    sg = [temp.subgraph(c) for c in nx.connected_components(temp)]
    heights=[(succ,max(nx.shortest_path_length(g,succ).values())) for g in sg for succ
        in succ_list if succ in g.nodes]
    for succ in succ_list:
        temp.add_edge(root,succ)
    return heights
def hdp(MST, root, order):
    order.append(root)
    succesors=[node for node,length in nx.shortest_path_length(MST,root).items() if
        (length==1 and node not in order)]
    if len(succesors) == 0:
        return order
    if len(succesors) == 1:
        for succ in succesors:
            return hdp(MST,succ,order)
    else:
        heights=order_height(MST,succesors,root)
        heights.sort(key=operator.itemgetter(1))
        for succ in heights:
            hdp(MST,succ[0],order)
        return order
def order_depth(G, root):
    order=[root]
    for node,length in nx.shortest_path_length(G,root).items():
        if length!=0:
            order.append(node)
    return order
def met_smirnov(X,Y):
    XY = np.concatenate([X, Y]).astype(np.float)
    MST=nx.from_edgelist(mst_edges(XY,10))
    ecc=list(nx.eccentricity(MST).values())
    nodes=list(nx.eccentricity(MST).keys())
    root=nodes[ecc.index(max(ecc))]
    lista=hdp(MST,root,[])
    x=[]
    y=[]
    for i in range(len(XY)):
        if XY[lista[i]] in X:
            x.append(i)
        elif XY[lista[i]] in Y:
```

```
            y.append(i)
    D=stats.ks_2samp(x,y)
    return (D[0],D[1])
def met_smirnov_rad(X,Y):
    XY = np.concatenate([X, Y]).astype(np.float)
    MST=nx.from_edgelist(mst_edges(XY,10))
    ecc=list(nx.eccentricity(MST).values())
    nodes=list(nx.eccentricity(MST).keys())
    root=nodes[ecc.index(min(ecc))]
    lista=order_depth(MST, root)
    x=[]
    y=[]
    for i in range(len(XY)):
        if XY[lista[i]] in X:
            x.append(i)
        elif XY[lista[i]] in Y:
            y.append(i)
    D=stats.ks_2samp(x,y)
    return (D[0],D[1])
```

El código utilizado para el método de Schilling es:

```
def met_schilling(k,X,Y):
    XY = np.concatenate([X, Y]).astype(np.float)
    T=0
    km=kneighbors_graph(XY, n_neighbors=k, mode='distance')
    m,n=len(X),len(Y)
    N=n+m
    l1=m/N
    l2=n/N

    edges=np.array([(i,ind) for i in range(N) for ind in km[i].indices])

    labels = np.array([0] * m + [1] * n)

    c = labels[edges]

    bw_sample_edges=edges[c[:, 0] == c[:, 1]]

    T=len(bw_sample_edges)+1

    T_nk=T/(N*k)
    meanT=l1**2+l2**2
    varT=l1*l2+(4*l1**2*l2**2)*(1-math.comb(2*k,k)*2**(-2*k))
```

```
    z=np.sqrt(N*k)*(T_nk-meanT)/np.sqrt(varT)
    return (z, 1-stats.norm.cdf(z))
```

Finalmente para el método de Friedman y Chen se tienen los siguientes códigos, estos dependiendo de si se usa el árbol de mínima cobertura o el de los k-vecinos.

```
def met_cfmst(X,Y):
    m, n = len(X), len(Y)
    N = m + n

    XY = np.concatenate([X, Y]).astype(np.float)

    edges = mst_edges(XY, 10)

    labels = np.array([0] * m + [1] * n)

    c = labels[edges]

    bw_sample_edges=edges[c[:, 0] == c[:, 1]]

    R1=len([i for i in bw_sample_edges if i[0]<m])
    R2=len(bw_sample_edges)-R1

    C=ww_C(nx.from_edgelist(mst_edges(XY,10)).degree)

    g=len(XY)

    m1=g*(m/N)*((m-1)/(N-1))
    m2=g*(n/N)*((n-1)/(N-1))

    s11=m1*(1-m1)+2*C*(m/N)*((m-1)/(N-1))*((m-2)/(N-2)) \
        +(g*(g-1)-2*C)*(m/N)*((m-1)/(N-1))*((m-2)/(N-2))*((m-3)/(N-3))
    s22=m2*(1-m2)+2*C*(n/N)*((n-1)/(N-1))*((n-2)/(N-2))
        +(g*(g-1)-2*C)*(n/N)*((n-1)/(N-1))*((n-2)/(N-2))*((n-3)/(N-3))
    s12=(g*(g-1)-2*C)*(m/N)*(n/(N-1))*((m-1)/(N-2))*((n-1)/(N-3))-m1*m2
    s21=s12

    R=np.array([(R1-m1),(R2-m2)])
    sigma=np.array([[s11,s12],
           [s21,s22]])
    sigma_inv=np.linalg.inv(sigma)

    S=R.dot(sigma_inv).dot(R)
    return (S,1-stats.chi2.cdf(S,2))
```

```
def met_cfknn(k,X,Y):

    m, n = len(X), len(Y)
    N = m + n

    XY = np.concatenate([X, Y]).astype(np.float)

    km=kneighbors_graph(XY, n_neighbors=k, mode='distance')
    edges=np.array([(i,ind) for i in range(N) for ind in km[i].indices])
    labels = np.array([0] * m + [1] * n)
    c = labels[edges]

    bw_sample_edges=edges[c[:, 0] == c[:, 1]]

    R1=len([i for i in bw_sample_edges if i[0]<m])
    R2=len(bw_sample_edges)-R1


    C=ww_C(nx.from_scipy_sparse_matrix(km, create_using=nx.Graph()).degree)

    g=len(edges)

    m1=g*(m/N)*((m-1)/(N-1))
    m2=g*(n/N)*((n-1)/(N-1))

    s11=m1*(1-m1)+2*C*(m/N)*((m-1)/(N-1))*((m-2)/(N-2)) \
        +(g*(g-1)-2*C)*(m/N)*((m-1)/(N-1))*((m-2)/(N-2))*((m-3)/(N-3))
    s22=m2*(1-m2)+2*C*(n/N)*((n-1)/(N-1))*((n-2)/(N-2)) \
        +(g*(g-1)-2*C)*(n/N)*((n-1)/(N-1))*((n-2)/(N-2))*((n-3)/(N-3))
    s12=(g*(g-1)-2*C)*(m/N)*(n/(N-1))*((m-1)/(N-2))*((n-1)/(N-3))-m1*m2
    s21=s12

    R=np.array([(R1-m1),(R2-m2)])
    sigma=np.array([[s11,s12],
           [s21,s22]])
    sigma_inv=np.linalg.inv(sigma)

    S=R.dot(sigma_inv).dot(R)
    return (S,stats.chi2.cdf(S,2))
```

Bibliografía

Barry, J. (1968). *General and comparative study of the psychokynetic effect on a fungus culture.* J. Parapsychology.

Chen, H. & Friedman, J. H. (2017). A New Graph-Based Two-Sample Test for Multivariate and Object Data. *Journal of the American Statistical Association*, *112*(517), 397-409. https://doi.org/10.1080/01621459.2016.1147356

Friedman, J. H. & Rafsky, L. C. (1979). Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annals of Statistics*, *7*(4), 283-298.

Gibbons, J. D. & Chakraborti, S. (2011). *Nonparametric Statistical Inference* (5.ª ed.). Taylor; Francis Group.

Monaco, V. (2015). *runs_test.py.* https://gist.github.com/vmonaco/e9ff0ac61fcb3b1b60ba/revisions

Randles, R. H. & Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* Krieger Publishinng Company.

Schilling, M. F. (1986a). Mutual and Shared Neighbor Probabilities: Finite- and Infinite-Dimensional Results. *Advances in Applied Probability*, *18*(2), 388-405. Consultado el 21 de mayo de 2022, desde http://www.jstor.org/stable/1427305

Schilling, M. F. (1986b). Multivariate Two-Sample Tests Based on Nearest Neighbors. *Journal of the American Statistical Association*, *81*(395), 799-806. Consultado el 21 de mayo de 2022, desde http://www.jstor.org/stable/2289012