USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

El coeficiente de correlación. Desde la independencia lineal a la independencia general de variables aleatorias.

Adrián Blanco Seijas

2019/2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

El coeficiente de correlación. Desde la independencia lineal a la independencia general de variables aleatorias.

Adrián Blanco Seijas

Julio 2020

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estadística e Investigación Operativa
Título: El coeficiente de correlación. Desde la independencia lineal a la independencia general de variables aleatorias.
Breve descrición do contido
Se trata de revisar los conceptos más notables ligados a la correlación de variables aleatorias. Como guión aproximado del estudio: 1) El coeficiente de correlación de Pearson. Propiedades. 2) El coeficiente de correlación de Spearman. Propiedades. 3) El coeficiente de correlación de distancias. Propiedades. 4) El coeficiente de correlación visto desde la perspectiva de las funciones cópula. 5) Ilustración en base a datos reales.
Recomendacións
Haber cursado las asignaturas de Probabilidade e Estatística e Inferencia Estatística.
Outras observacións

Índice general

Resumen

En este trabajo haremos un repaso al concepto de coeficiente de correlación. Empezaremos viendo el coeficiente de correlación de Pearson y sus propiedades. Tras estudiar este coeficiente, veremos que sus limitaciones nos llevan a buscar nuevos coeficientes que sean capaces de superarlas. Así surgirá el coeficiente de correlación de Spearman, que nos lleva al estudio de los rangos para descubrir las relaciones de dependencia entre variables. Posteriormente, el siguiente coeficiente, la τ de Kendall, tendrá una gran cantidad de variantes que le permitirá adaptarse a las necesidades que tengamos. Destacamos también la capacidad de la τ de Kendall de adaptar el concepto del coeficiente de correlación de rangos a condiciones poblacionales. El último caso de coeficientes de correlación que veremos es el coeficiente de correlación de distancias. Este último, que nace para solucionar los problemas existentes en los casos multidimensionales, es el más reciente y nos obligará a hacer una breve revisión a la teoría.

Finalmente, daremos unas nociones de las funciones cópula. Veremos sus propiedades y definiremos grupos de las funciones cópula arquimedianas y las funciones cópula gaussianas. Si bien las funciones cópula no están directamente influenciadas por la correlación, existen diversas relaciones, principalmente a través de las cópulas gaussianas. Acabaremos con unas breves observaciones realizadas sobre el uso de estas funciones en diversos ámbitos.

Abstract

Our intention in this essay is to review the concept of the correlation coeficient. We will start by taking a look on Pearson's correlation coeficient and it's properties. After that, the own coeficient's limitations will force us to keep searching new coeficients that are able to overcome this limitations. By this search the Spearman's correlation coefficient will appear, guiding us to discover dependence relations between variables by the study of ranks. Following, we will review the next coeficient, Kendall's τ, which will have a large

number of variations that will allow it's adaptation to our own needs. Will be worth to remark that the Kendall's τ is able to applay the rank's correlation coefficient concept to population conditions. The last case of correlation coefficients that will be studied will be the distance correlation coeficient. This last one, which it's defined to solve the problems that exist in multidimensional situations, it's the most recent of all and will force us to make a slight review of it's theory.

To give the essay an end, we will give some notions about the copula functions. We will see it's properties and define the groups of arquimedean copula functions and gaussian copula functions. Although copula functions are not directly infuenced by correlation, there are several relations, mostly through gaussian copulas. We will finish by giving some short observations about the use of this functions.

Capítulo 1

El primer coeficiente de correlación: Pearson

1.1. Introducción a los coeficientes de correlación

1.1.1. Motivación histórica de los coeficientes de correlación

La estadística representa uno de los ámbitos de conocimiento más reconocidos dentro de las matemáticas. Es una herramienta indispensable para una gran cantidad de estudios, ya que no se ve restringida a un único ámbito. La estadística se puede adaptar a condiciones muy diversas, como la matemática financiera, estudios biomédicos o estudios demográficos. Este alto grado de utilidad es en gran parte debido al concepto que estudiaremos durante este trabajo: El coeficiente de correlación.

Para poder entender el concepto de correlación, deberemos definir antes que entendemos por la independencia entre variables, ya que es esencial en su motivación.

Definición 1.1. Decimos que dos variables son **estadísticamente independientes** cuando al conocer el valor que toma una de ellas no tenemos ninguna información sobre la otra variable. En caso de que no sean independientes, existirá una relación entre ambas por lo que diremos que son variables **dependientes**.

Históricamente, los primeros pasos hacia el estudio de la correlación se debieron a la aparición de la distribución normal. Una vez definida esta curva, Gauss desarrolló una ley de distribución de errores motivado por el procedimiento del método de mínimos cuadrados. Sin embargo, estos errores no son observables y deben ser sustituidos por una estimación de los mismos, los residuos, que pierden la independencia. Posteriormente, Pearson destaca que Gauss presupone que las variables observadas son independientes, mientras que

el propio Pearson opina se deben considerar dependientes en un inicio [5]. Esta última es la mentalidad que se continúa siguiendo actualmente en la estadística. Tras estos avances, llegamos a una duda que definirá este trabajo: ¿Cómo medimos esas relaciones entre variables?

1.1.2. Aproximación a los coeficientes de correlación

Los coeficientes de correlación nacen de esta necesidad de ser capaces de medir la relación existente entre dos variables. Medir el nivel de asociación entre ellas es esencial a la hora de analizar datos. Esencialmente, estos coeficientes serán los encargados de darnos una respuesta a la pregunta anterior midiendo la dependencia o independencia de dichas variables. Gracias a esto, son esenciales a la hora de la realización de estudios científicos. Por ello, dos grandes impulsores de estas teorías son los estudios psicológicos y fisiológicos.

Definición 1.2. Dadas X e Y dos variables en un cierto espacio de probabilidad, diremos que X e Y son **variables independientes** si y solo si

$$P(X \text{ verifica el suceso A e } Y \text{ verifica en suceso B}) =$$

$$= P(X \text{ verifica el suceso A})P(Y \text{ verifica en suceso B}).$$

En caso de trabajar con funciones de distribuciones, entenderemos dicha independencia como

$$f_{XY} = f_X f_Y,$$

siendo f_{XY} la función de distribución conjunta y f_X y f_Y las funciones marginales de cada variable.

Entre los coeficientes de correlación deberemos distinguir dos tipos: poblacionales y muestrales. Los coeficientes de correlación serán poblacionales, mientras que serán sus estimadores las versiones muestrales de dichos coeficientes y lo que usaremos en los casos prácticos. En caso de trabajar con las versiones muestrales, sustituiremos las variables poblacionales X e Y por las observaciones (X_i, Y_i), que generarán en general una muestra aleatoria simple.

Definición 1.3. Diremos que un parámetro es un **parámetro poblacional** si depende únicamente de distribuciones. Por el contrario, diremos que es un **parámetro muestral** si únicamente depende de los datos observados.

Existen muchas y muy diversas formas de medir las relaciones entre variables, sin embargo no todas cumplen unas condiciones adecuadas para ser utilizadas para el estudio.

Para que estas medidas de asociación relativa sean razonablemente útiles deberíamos pedirles que sean capaces de verificar algunas propiedades. Podemos encontrar la siguiente lista de buenas propiedades en [2]:

Propiedades 1.4. *Tomamos X e Y como variables aleatorias.*

1. *Para dos pares de observaciones independientes (X_i,Y_i) y (X_j,Y_j) cualesquiera, la medida tendrá valor +1 si la relación entre variables es directa y perfecta. De esta manera tendremos dos opciones:*

   a) $X_i < X_j$ *siempre que* $Y_i < Y_j$

   b) $X_i > X_j$ *siempre que* $Y_i > Y_j$

   En estas circunstancias diremos que las dos variables tienen una relación de **concordancia perfecta**.

2. *En el caso opuesto, dadas las mismas condiciones, la medida tomará valor -1 si inversa y perfecta. Es decir, si:*

   a) $X_i < X_j$ *siempre que* $Y_i > Y_j$

   b) $X_i > X_j$ *siempre que* $Y_i < Y_j$

   En este caso, diremos que la relación entre las variables es de **discordancia perfecta**.

3. *Si no se cumplen los criterios anteriores para todos los pares de observaciones, entonces la medida tomará valores entre -1 y +1.*

4. *La medida tomará el valor 0 si las dos variables son independientes.*

5. *La medida para el par de variables (X, Y) debe coincidir con la medida de (Y, X), la de $(-X, -Y)$ y la de $(-Y, -X)$.*

6. *La medida para el par de variables $(-X, Y)$ o $(X, -Y)$ será la medida de (X, Y) con signo opuesto.*

7. *La medida debería ser invariante ante todas aquellas transformaciones de X e Y que mantengan el orden de magnitud.*

1.2. El coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson no es propiamente la primera medida de asociación entre variables. Dicho coeficiente surge de una extensa discusión entre diversos matemáticos ([5]), a pesar de tener también sus raíces en estudios de sociales debido a su uso. En el artículo mencionado anteriormente, Karl Pearson repasa los inicios de la medición de la correlación entre variables, señalando los avances realizados por Gauss y destacando el debate en el que matemáticos como August Bravais, Francis Galton o Francis Edgeworth, junto con el propio Pearson, se vieron inmersos.

El coeficiente de correlación de Pearson, ρ, precisa del cálculo de la covarianza y las varianzas de las variables, por lo que será necesario que podamos calcularlas. Para ello, la única necesidad existente para que podamos obtener este coeficiente es que las medias y varianzas de las variables sean calculables y, por tanto, finitas.

Definición 1.5. Consideremos dos variables aleatorias, X e Y, con medias y varianzas finitas. Diremos que el **coeficiente de correlación de Pearson** entre las variables es:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

.

También podemos escribirlo en función de la desviación típica:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

.

Una vez tenemos definido el coeficiente, debemos ver que propiedades posee. Será prioritario que verifique las vistas anteriormente.. Sabiendo esto, tenemos que verifica algunas de estas propiedades.

Proposición 1.6. *El coeficiente de correlación de Pearson verifica las propiedades 3, 4, 5 y 6 definidas en Propiedades 1.4. Además, en el caso de dependencia lineal, también verifica las propiedades 1 y 2.*

Demostración 1.7. *Procedemos ahora a demostrar las propiedades:*

1. *Consideramos en este caso que existe dependencia lineal entre las variables aleatorias X e Y respecto a la variable aleatoria Z. Las podremos definir como $X = \alpha Z + \beta$, $Y = \gamma Z + \theta$.*

   Por hipótesis, el signo de α y γ debe coincidir. Así pues,

$$Cov(X,Y) = Cov(\alpha Z + \beta, \gamma Z + \theta) = \alpha\gamma Cov(Z,Z) = \alpha\gamma Var(Z).$$

Una vez tenemos esto, podemos sustituír en la expresión del coeficiente de Pearson.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{\alpha\gamma Var(Z)}{[|\alpha|\sigma(Z)][|\gamma|\sigma(Z)]} = \frac{\alpha\gamma Var(Z)}{|\alpha||\gamma|\sigma(Z)^2} = +1.$$

2. *Esta propiedad se demustra de manera similar a 1. La única diferencia es que α y γ tienen signo distinto, por lo que*

$$\rho(X,Y) = \frac{\alpha\gamma Var(Z)}{|\alpha||\gamma|Var(Z)} = -1.$$

3. *Dada la observación anterior, tenemos que $|Cov(X,Y)| \leq \sigma(X)\sigma(Y)$. Por tanto,*

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \in [-1,1].$$

4. *Por hipótesis, las variables son independientes, por lo que $Cov(X,Y) = 0$. Así pues, $\rho(X,Y) = \frac{0}{\sigma(X)\sigma(Y)} = 0$.*

5. *Debemos ver que $\rho(X,Y) = \rho(Y,X) = \rho(-X,-Y) = \rho(-Y,-X)$.*

$$\rho(Y,X) = \frac{Cov(Y,X)}{\sigma(Y)\sigma(X)} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \rho(X,Y).$$

$$\rho(-X,-Y) = \frac{Cov(-X,-Y)}{\sigma(-X)\sigma(-Y)} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \rho(X,Y).$$

$$\rho(-Y,-X) = \frac{Cov(-Y,-X)}{\sigma(-Y)\sigma(-X)} = \frac{Cov(-X,-Y)}{\sigma(-X)\sigma(-Y)} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \rho(X,Y).$$

6. *Por último, veamos que $\rho(-X,Y) = \rho(X,-Y) = -\rho(X,Y)$.*

$$\rho(-X,Y) == \frac{Cov(-X,Y)}{\sigma(-X)\sigma(Y)} == \frac{-Cov(X,Y)}{\sigma(X)\sigma(Y)} = -\rho(X,Y).$$

$$\rho(X,-Y) == \frac{Cov(X,-Y)}{\sigma(X)\sigma(-Y)} == \frac{-Cov(X,Y)}{\sigma(X)\sigma(Y)} = -\rho(X,Y).$$

Demostrado que cumple las seis propiedades expuestas, es lógico preguntarse si el coeficiente verificará la última de ellas. Es decir, si el coeficiente es invariante ante transformaciones que mantengan el orden de magnitud. Aún cumpliendo las anteriores propiedades, el coeficiente de Pearson no verifica esta última.

No verificar esta última propiedad no es el único problema del coeficiente de correlación de Pearson, ya que dicho coeficiente también variará en otras circunstancias. Podemos

tomar como ejemplo la transformación logarítmica, muy utilizada en el ámbito de la regresión. Pero antes, debemos definir el coeficiente de correlación de Pearson muestral para poder aplicarlo a las simulaciones.

Definición 1.8. Consideremos dos variables aleatorias, X e Y, con medias y varianzas finitas y n pares de observaciones de la forma (X_i,Y_i). Definimos el **coeficiente de correlación de Pearson muestral**, r, entre X e Y como:

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{(X_i - \overline{X})^2(Y_i - \overline{Y})^2}}.$$

Ejemplo 1.9. Para llevar a cabo el ejemplo de la regresión logarítmica, hemos simulado una variable uniforme en el intervalo $[0{,}1, 1{,}5]$, de donde hemos tomado 50 datos que conformarán la variable $x1$. Lo hemos enfrentado a la variable $y1$ que formamos a partir de la variable $x1$ añadiéndole unos errores normales de media 0 y varianza 0.15. Al aplicarles la transformación logarítmica, definimos las varibles $lx1$ y $ly1$, que serán respectivamente $log(x1)$ y $log(y1)$.

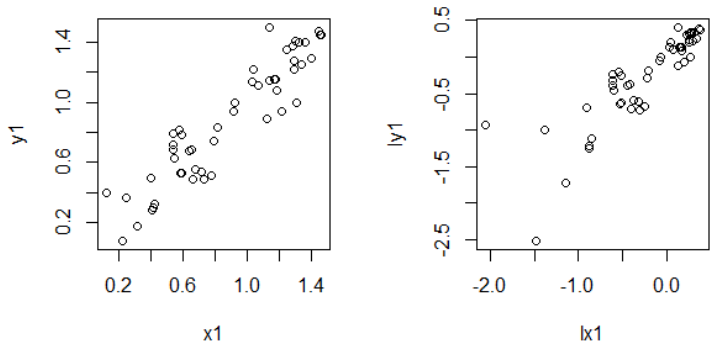

Figura 1.1: Gráfica de los datos enfrentados.

En la Figura 1.1, a la izquierda tenemos el caso de las variables originales. A la derecha, los el caso de la aplicación del logaritmo a las variables.

Si calculamos el coeficiente de correlación de Pearson en ambos casos, tenemos que $r(x1, y1) = 0{,}9266395$, pero $r(lx1, ly1) = 0{,}8628166$. Por tanto, claramente no se mantiene el coeficiente.

El problema con la regresión logarítmica y la no invarianza del coeficiente de correlación de Pearson ante las transformaciones mencionadas se debe al principal punto débil de esta

medida de asociación de variables. Dicho punto débil es que este coeficiente de correlación no tiene la capacidad de recoger otro tipo de relaciones entre variables que no sea lineal. Esta limitación que sufre el coeficiente de Pearson por solo ser capaz de medir relaciones lineales tiene grandes repercusiones. Esto nos generará problemas en múltiples ocasiones, ya que gran cantidad de variables no tienen dependencia lineal pese a que si poseen otros tipos de dependencia.

Ejemplo 1.10. Consideremos en este caso una variable X que sigue una distribución uniforme en el intervalo [0,20]. Estudiaremos entonces con el coeficiente de correlación de Pearson la relación entre X y las variables $Y = X^2$ y $Z = X^{10}$, tomando una muestra de 100 observaciones. Definimos la distribución uniforme en un intervalo de números positivos para no tener problemas con los cambios de signo de las potencias en los números negativos, algo que será visto más adelante.

Lógicamente podemos intuir que existe una clara dependencia entre la variable X y las variables Y y Z, ya que son simplemente potencias de los mismos datos. Sin embargo, el coeficiente de correlación de Pearson no será capaz de indicarnos tal relación adecuadamente. Es más, a medida que se incrementa la curva en la gráfica (es decir, en puntos mayores que 1, aumenta la potencia) menor será la relación dada por el coeficiente de correlación.

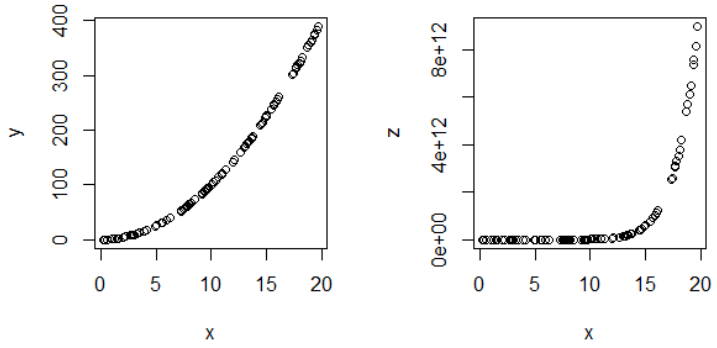

Figura 1.2: Gráfica de los datos de potencias enfrentadas. A la izquierda X frente a X^2. A la derecha X frente a X^{10}

Tenemos ahora representados unos datos simulados en la Figura 1.2. Al aplicarle sobre estos datos el coeficiente de correlación de Pearson hemos obtenido que

$$r(X, Y) = 0{,}9670344 \text{ y que } r(X, Z) = 0{,}6790979.$$

Si bien parece que se aproxima mucho a 1 la medición de la relación entre X e Y, esto no

ocurre en el caso de la relación entre X y Z. La relación real entre estas dos variables es de absoluta dependencia, pero dado que el coeficiente de correlación de Pearson es capaz de registrar únicamente relaciones lineales, falla a la hora de encontrar la relación entre estas variables ya que es una relación monótona (siempre que estemos en un intervalo donde todos los valores tienen el mismo signo).

Visto esto, llegamos a la conclusión de que el coeficiente de correlación de Pearson está especialmente diseñado para medir la dependencia lineal entre variables. Por ello, sigue siendo una de las mediciones de asociación entre variables más utilizadas y su uso está muy extendido, siendo principalmente usado en la regresión estadística.

Capítulo 2

La correlación derivada del coeficiente de Pearson

2.1. El Coeficiente de Spearman

A partir del coeficiente de correlación de Pearson surgen otra serie de coeficientes, que pretenden cubrir la necesidad de una medida de la relación entre variables que nos permitiese trabajar con asociaciones que no fuesen únicamente lineales. Este avance se dio en 1904, cuando el psicólogo Charles Spearman definió una nueva medida de asociación entre variables. Dicha medida es conocida actualmente como el coeficiente de correlación de Spearman.

Conocido también como el coeficiente de correlación de rangos, el coeficiente de Spearman ordena los datos obtenidos en cada variable de menor a mayor dándoles un valor a cada observación (rango). De esta manera, definimos unas nuevas variables que tendremos en cuenta para este coeficiente. Dichas variables pasarán a ser R_i=rank(X_i) y S_i=rank(Y_i). Consideraremos ahora que las observaciones son de la forma (r_1, s_1), (r_2, s_2),..., (r_n, s_n), siendo n el número de observaciones realizadas.

Observación 2.1. En estas condiciones, calcular diversos valores que nos serán útiles en la transformación del coeficiente:

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \qquad \overline{r} = \overline{s} = \frac{n+1}{2},$$

$$\sum_{i=1}^{n} (r_i - \overline{r})^2 = \sum_{i=1}^{n} (s_i - \overline{s})^2 = \sum_{i=1}^{n} (i - \frac{n+1}{2})^2 = \frac{n(n^2-1)}{12}.$$

Tras la obtención de los resultados expuestos en la anterior observación, ya tenemos los datos necesarios para intentar definir el coeficiente de correlación de Spearman. Si

sustituimos las variables X e Y por sus rangos asociados (R y S, respectivamente) en la expresión general del coeficiente de correlación de Pearson, obtendremos la siguiente definición del coeficiente de correlación de Spearman.

Definición 2.2. El **coeficiente de correlación de rangos** o de Spearman lo denotaremos por ρ_s y lo definimos como:

$$\rho_s(R,S) = \frac{12\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{n(n^2-1)}$$

También podemos definir el coeficiente de manera diferente. Podemos considerarlo en función de las diferencias entre los rangos de las variables. Estas diferencias las podremos definir como $D_i = R_i - S_i$. Además, tenemos que $D_i = (R_i - \overline{R}) - (S_i - \overline{S})$.

Teniendo en cuenta que

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n}(R_i - \overline{R})^2 + \sum_{i=1}^{n}(S_i - \overline{S})^2 - 2\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S}).$$

Sustituyendo en la expresión del coeficiente, llegamos a que lo podemos expresar como

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}.$$

Observación 2.3. El coeficiente de correlación de Spearman es el único de los principales coeficientes a estudiar que está definido como un parámetro muestral. El resto de los principales coeficientes requieren el cálculo de un estimador para su aplicación sobre los datos observados.

Ahora que tenemos definido el coeficiente de correlación de Spearman podemos ver que realmente es una mejora del coeficiente de correlación de Pearson. Uno de los avances dados por Spearman con la definición de este coeficiente es que el coeficiente de correlación de rangos es capaz de recoger dependencias no solo lineales, sino que también tiene la capacidad de trabajar en casos de relaciones monótonas.

Tras este avance en cuanto a los coeficientes, surge la pregunta de si este coeficiente verificará las propiedades vistas en el capítulo anterior.

Proposición 2.4. *El coeficiente de correlación de rangos verifica las propiedades descritas en Propiedades 1.4.*

Demostración 2.5. *1. Tomemos la versión del coeficiente de rangos que considera las diferencias entre el rango de cada observación. Entonces tenemos que* $D_i = R_i - S_i = 0, \forall i \in 1 = 1, ..., n$.

2. *En este caso, tomamos* $R_i = i$ *y* $S_i = n - i + 1$. *Por tanto, tenemos lo siguiente:*

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} [i - (n - i + 1)]^2 = 4 \sum_{i=1}^{n} (i - \frac{n+1}{2})^2 = \frac{n(n^2-1)}{3}.$$

Así pues, si sustituímos en la expresión del coeficiente, obtenemos que $\rho_s = -1$

3.∼6. *Por ser el coeficiente de Spearman algebraicamente equivalente al de Pearson (tal y como está definido), verifica las propiedades 3, 4, 5 y 6.*

7. *Debido a que las transformaciones que tenemos en cuenta mantienen el orden de magnitud, los rangos no se ven afectados por ellas. Es por eso que el coeficiente de Spearman sí verifica esta propiedad.*

Ejemplo 2.6. En este caso, podemos utilizar el Ejemplo 1.9 dado anteriormente. En este ejemplo concluíamos que el coeficiente de correlación de Pearson no verificaba la propiedad 7. Sin embargo, acabamos de demostrar que el coeficiente de correlación de rangos sí lo verifica. Si lo comprobásemos con dicho ejemplo y comparásemos con el coeficiente de correlación de Pearson, tendríamos que:

Coeficiente de correlación de Pearson	$r(x1, y1) = 0{,}9266395$	$r(lx1, ly1) = 0{,}8628166$
Coeficiente de correlación de Spearman	$\rho_s(x1, y1) = 0{,}9005042$	$\rho_s(lx1, ly1) = 0{,}9005042$

Así pues, vemos que en el ejemplo donde no se conservaba el coeficiente de correlación de Pearson sí se conserva el coeficiente de correlación de Spearman.

Que en este último ejemplo el coeficiente de correlación de rangos funcione mejor que coeficiente de correlación de Pearson se debe a que el coeficiente de Pearson solo es capaz de estudiar las relaciones lineales entre variables. Por su parte, el coeficiente de correlación de Spearman es capaz de recoger relaciones monótonas entre variables, por lo que las transformaciones que mantienen el orden no varían su valor (como está demostrado previamente).

Ejemplo 2.7. Para ilustrar la capacidad del coeficiente de Spearman de estudiar relaciones monótonas de variables, recuperamos también el Ejemplo 1.10, donde veíamos que el coeficiente de Pearson no nos daba una dependencia perfecta entre las variables X e $Y = X^2$ y todavía funcionaba peor en la relación entre X y $Z = X^{10}$. Comparando los resultados como hemos hecho en el ejemplo anterior, vemos que los problemas surgidos en ese caso se han solucionado con la definición del coeficiente de correlación de Spearman.

Coeficiente de correlación de Pearson	$r(X, Y) = 0{,}9670344$	$r(X, Z) = 0{,}6790979$
Coeficiente de correlación de Spearman	$\rho_s(X, Y) = 1$	$\rho_s(X, Z) = 1$

Podemos considerar que siempre que podamos ordenar y contar las observaciones a estudiar, podremos utilizar el coeficiente de correlación de rangos. No necesitaremos otras condiciones sobre las variables a considerar. Por ello, podemos decir que tiene unas hipótesis más débiles que el coeficiente de correlación de Pearson, ya que este requiere que medias y varianzas de ambas variables sean finitas y calculables para poder construirlo.

A mayores de que las hipótesis sean más débiles, otra diferencia entre ambos coeficientes es que el coeficiente de correlación de Spearman es no paramétrico. Es decir, este coeficiente de correlación no depende de la muestra ni de la distribución de cualquiera de las variables observadas. Esta condición tiene ciertas ventajas, como llevar a cabo una reducción a la cantidad esencial de la información de las variables o que prácticamente se puede utilizar en cualquier situación.

Podemos destacar otra propiedad que posee este coeficiente de correlación es que es robusto ([7]) a la presencia de outliers (valores atípicos). Es decir, permite ciertos desvíos del patrón normal en la relación entre las variables que estamos estudiando. Además, los valores atípicos pueden ser indicativos de datos que pertenecen a una población diferente del resto de las muestras establecidas.

Si bien parece que es un coeficiente lógico y sencillo de aplicar, no está libre de errores en su uso. Esto podemos verlo en [9], donde Wissler explica que el coeficiente fue utilizado de manera indiscriminada en sus inicios cometiendo ciertos errores al no tener en cuenta las posibles limitaciones del coeficiente. Estas limitaciones se encuentran en relaciones geométricas entre variables, puesto que el coeficiente de Spearman solo es capaz de trabajar con relaciones monótonas.

Ejemplo 2.8. Si bien hemos visto que en un intervalo de números positivos la relación entre X y X^2 es de concordancia perfecta (Ejemplo 2.7), la relación entre las variables X y X^2 no será de concordancia perfecta si se ven envueltos números negativos en el intervalo de la variable X. Consideramos ahora X variable con distribución uniforme en el intervalo [-10,10] y definimos $Y = X^2$. Realizamos 100 observaciones y obtenemos la Figura.

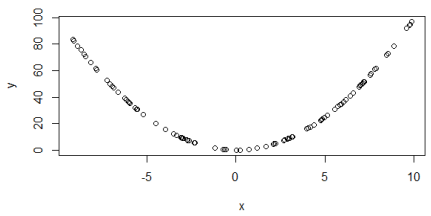

Figura 2.1: Gráfica entre las variables X $Y = X^2$.

A pesar de que, como ya dijimos antes, la dependencia entre las variables X y X^2 es clara, en estas condiciones el coeficiente de correlación de Spearman no es capaz de darnos tales datos. De hecho, en esta simulación hemos obtenido que $\rho_s = 0{,}1381338$. Es un valor suficientemente cercano a 0 como para que nos planteásemos la posibilidad de que ambas variables sean independientes.

Por último, debemos considerar también que ocurre en caso de que se den empates a la hora de ordenar las observaciones de cualquiera de las variables. Esto genera un problema con las definiciones previas, ya que esta circunstancia no está considerada. Dicho problema se solucionará haciendo la media de los rangos considerados. Es decir, si hubiese un empate de cuatro observaciones en el segundo puesto del orden, el valor del rango sería $r_i = \frac{2+3+4+5}{4} = 3'5$. Generalizando, si $r_i = ... = r_j$, definimos el nuevo $r'_{i\sim j} = \frac{\sum_{t=i}^{j} t}{j-i}$.

Si bien queda solucionado el orden de las observaciones, el cambio hecho afecta ahora el cuadrado de las diferencias se ve afectado, por lo que debemos redefinir el coeficiente para esta situación. En el caso de los empates, la media de los cuadrados es

$$\sum_{t=i}^{j}(r'_t)^2 = (j-i)r'^2_i,$$

cuando en caso de que no hubiese empates sería

$$\sum_{t=i}^{j}(r_t)^2 = (j-i)r_t^2 + (j-i)r_t(j-i+1) + \frac{(j-i)(j-i+1)(2(j-i)+1)}{6}.$$

Sabiendo esto, tenemos que la diferencia entre los cuadrados será

$$\sum_{t=i}^{j}(r_t)^2 - (r'_t)^2 = \frac{(j-i)[(j-i)^2-1]}{12}.$$

Denotaremos esta diferencia por u. Así pues, la media de las observaciones quedará de la forma

$$\sum_{i=1}^{n}(s_i - \overline{s})^2 = \frac{n(n^2-1)}{12} - u.$$

Consideremos u_X y u_Y como las diferencias generadas por los empates en las variables X e Y, respectivamente. Procedemos ahora a reescribir el coeficiente de Spearman en caso de empates.

Definición 2.9. En caso de existencia de observaciones empatadas, definimos el coeficiente de correlación de rangos de la siguiente manera:

$$\rho_s = \frac{12[\sum_{i=1}^{n} r_i s_i - \frac{1}{4}n(n+1)^2]}{([n(n^2-1) - 12u_X][n(n^2-1) - 12u_Y])^{\frac{1}{2}}}$$

Observación 2.10. Análogamente al caso sin empates, podemos dar una expresión equivalente para calcular el coeficiente con la diferencia entre rangos. Como el cuadrado de las diferencias se ve afectado también, pasará a ser

$$\sum_{i=1}^{n} D_i^2 = \frac{n(n^2-1)}{6} - u_X - u_Y - 2\sum_{i=1}^{n}(r_i - \overline{r})(s_i - \overline{s})$$

Por tanto, definiremos ρ_s como

$$\rho_s = \frac{n(n^2-1) - 6\sum_{i=1}^{n} D_i^2 - 6(u_X + u_Y)}{([n(n^2-1) - 12u_X][n(n^2-1) - 12u_Y])^{\frac{1}{2}}}$$

2.2. La τ de Kendall

La τ de Kendall surge a partir de los fundamentos del coeficiente de Spearman. Definiéndose como una nueva medida de correlación de rangos ([3]), este nuevo coeficiente refina la idea de coeficiente de Spearman utilizando la probabilidad. Esto hace posible que sea utilizable en casos en los que tengamos en cuenta una población, incluso simplemente trabajar con distribuciones.

Para poder definir la τ de Kendall, necesitamos definir previamente lo que llamaremos probabilidad de concordancia, p_c, y probabilidad de discordancia, p_d.

Definición 2.11. Consideramos dos variables aleatorias, X e Y, y dos pares de observaciones independientes, (X_i, Y_i) y (X_j, Y_j). Definimos las probabilidades de concordancia (p_c) y discordancia (p_d) como:

$$p_c = P\{[(X_i < X_j) \cap (Y_i < Y_j)] \cup [(X_i > X_j) \cap (Y_i > Y_j)]\} = P[(X_j - X_i)(Y_j - Y_i) > 0]$$

$$p_d = P\{[(X_i < X_j) \cap (Y_i > Y_j)] \cup [(X_i > X_j) \cap (Y_i < Y_j)]\} = P[(X_j - X_i)(Y_j - Y_i) < 0]$$

Definición 2.12. Sean X e Y dos variables aleatorias. Tomando p_c y p_d como las probabilidades de concordancia y discordancia entre ambas variables, el coeficiente de correlación τ **de Kendall** se define como

$$\tau = p_c - p_d.$$

Observación 2.13. Consideremos X' e Y' son variables independientes idénticamente distribuidas respecto a las variables X e Y. En estas condiciones, también podemos definir la τ de Kendall como

$$\tau = P[(X - X')(Y - Y') \geq 0] - P[(X - X')(Y - Y') \leq 0],$$

ya que tenemos que

$$P[(X - X')(Y - Y') \geq 0] - P[(X - X')(Y - Y') \leq 0] = p_c - p_d.$$

Observación 2.14. Podemos observar en estas definiciones varias cosas.

- La τ de Kendall no tiene ningún inconveniente con los casos de observaciones empatadas, ya que tanto p_c como p_d están definidos sobre desigualdades estrictas. Esto descartará la posibilidad de los empates del cálculo del coeficiente, permitiendo un perfecto funcionamiento.
- La τ de Kendall verifica las Propiedades 1.4.
- En caso de que las variables X e Y sean continuas, la probabilidad de empate será 0. Por tanto, tenemos que $p_c = 1 - p_d$. Así pues, podremos definir el coeficiente de Kendall como
$$\tau = 2p_c - 1 = 1 - 2p_d.$$
- Además, Kendall presenta este coeficiente como un coeficiente que surge de manera natural ([3]). Destaca que nace de la comparación de objetos y sus cualidades y por ello lo considera un coeficiente sencillo y lógico.

La τ de Kendall posee un estimador insesgado, llamado τ de Kendall muestral. Este estimador también puede ejercer como coeficiente de correlación y su construcción conlleva unos cálculos que detallaremos a continuación.

Consideremos X e Y variables aleatorias y (X_i,Y_i) y (X_j,Y_j) un par de observaciones independientes. Primero debemos estimar los valores de p_c y p_d. Definimos la matriz

$$A_{ij} = \text{ signo}(X_i - X_j)\ \text{signo}(Y_i - Y_j),$$

de manera que los coeficientes a_{ij} de la matriz pueden tomar valor 1, -1 o 0, considerando los tres casos posibles:

$$a_{ij} = \begin{cases} 1, & \text{si los pares son concordantes.} \\ -1, & \text{si los pares son discordantes.} \\ 0, & \text{en cualquier otro caso.} \end{cases}$$

Así pues, podemos construir la distribución marginal de estos elementos, obteniendo

$$f_{Aij}(a_{ij}) = \begin{cases} p_c, & \text{si } a_{ij} = 1. \\ p_d, & \text{si } a_{ij} = -1. \\ 1 - p_c - p_d, & \text{si } a_{ij} = 0. \end{cases}$$

y llegando a que la media de A_{ij} es

$$E(A_{ij}) = 1p_c + (-1)p_d = p_c - p_d = \tau.$$

Dado que $a_{ij} = a_{ji}$ y que $a_{ii} = 0$, tenemos en la matriz $\binom{n}{2}$ elementos distintos. Ahora podemos proceder a definir la τ de Kendall muestral.

Definición 2.15. El estimador insesgado de τ, la τ **de Kendall muestral**, se define como

$$\tau_m = \sum_{i=1}^{n}\sum_{j>i}^{n} \frac{A_{ij}}{\binom{n}{2}} = 2\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{A_{ij}}{n(n-1)}$$

La τ de Kendall muestral también se tiene definido de forma equivalente realizando una agrupación de los valores de la matriz A_{ij} por su signo. Por tanto, si tomamos C como el número de elementos de la matriz A_{ij} que tienen signo positivo y Q como el número de dicha matriz con signo negativo, podremos redefinir la τ de Kendall muestral. Entonces podemos definir la τ de Kendall muestral como

$$T = \frac{C - Q}{\binom{n}{2}},$$

siendo n el número de observaciones.

Tenemos redefinido el coeficiente, pero aún podemos simplificarlo. Esto se debe a que en caso de que no existan empates en las variables, tendremos que $A_{ij} \neq 0$ salvo en los elementos de la diagonal de la matriz (dado que es empate obligado). En estas condiciones tenemos que $C+Q = \binom{n}{2}$, de manera que podemos escribir la expresión anterior aplicándole estos cambios y obteniendo una nueva expresión. Sería

$$T = \frac{2C}{\binom{n}{2}} - 1 = 1 - \frac{2Q}{\binom{n}{2}}$$

Alcanzando esta expresión, podemos notar algo familiar en ella. Si retrocedemos un poco en el trabajo, en concreto a la Observación 2.13, entenderemos la razón. Previamente teníamos definida una forma de calcular la τ de Kendall como

$$\tau = 2p_c - 1 = 1 - 2p_d.$$

Estudiando un poco más a que se debe esta relación, llegamos a la conclusión de que $\frac{C}{\binom{n}{2}}$ y $\frac{Q}{\binom{n}{2}}$ son los estimadores insesgados de p_c y p_d respectivamente.

Por último, además de las propiedades y variaciones del coeficiente de correlación de Kendall que acabamos de ver, deberíamos valorar si existe una relación entre este coeficiente de correlación y los vistos anteriormente. Ya hemos dicho anteriormente que lo podríamos considerar una transformación del coeficiente de correlación de rangos. Así pues, también cabe destacar que existe una relación entre el coeficiente de Pearson y la τ de Kendall ([2]).

Proposición 2.16. *En condiciones de una normal bivariante, con variables X e Y, tenemos que se da la siguiente igualdad entre el coeficiente de correlación de Pearson y la τ de Kendall:*

$$\tau = \frac{2}{\pi} \arcsin \rho.$$

Demostración 2.17. *Se puede llegar a dicha expresión a través de una transformación sobre la distancia entre observaciones. Sean (X_i,Y_i) y (X_j,Y_j) un par de observaciones independientes, estandarizamos sus diferencias obteniendo*

$$U = \frac{X_i - X_j}{\sqrt{2}\sigma_X},\ V = \frac{Y_i - Y_j}{\sqrt{2}\sigma_Y}.$$

Tras estas transformaciones tendremos que $\rho(X,Y) = \rho(U,V)$. Para el cálculo de la τ de Kendall, nos basaremos en la expresión

$$\tau = 2p_c - 1,$$

por lo debemos encontrar previamente la expresión de p_c en estas condiciones. Consideremos la función $\varphi(x,y)$ la función densidad de una normal estándar bivariante. Dadas estas condiciones, podemos definir p_c como $p_c = P(UV) > 0$, entonces tendremos que su expresión será

$$p_c = \int_{-\infty}^{0}\int_{-\infty}^{0} \varphi(x,y)dxdy + \int_{0}^{\infty}\int_{0}^{\infty} \varphi(x,y)dxdy.$$

Por simetría de la función $\varphi(x,y)$, tenemos que

$$\int_{-\infty}^{0}\int_{-\infty}^{0} \varphi(x,y)dxdy + \int_{0}^{\infty}\int_{0}^{\infty} \varphi(x,y)dxdy = 2\int_{-\infty}^{0}\int_{-\infty}^{0} \varphi(x,y)dxdy.$$

Por tanto,

$$p_c = 2\int_{-\infty}^{0}\int_{-\infty}^{0} \varphi(x,y)dxdy = 2\Phi(0,0),$$

siendo $\Phi(x,y)$ la función de distribución acumulativa de la normal estándar bivariante. En este caso, podemos sustituir el valor de $\Phi(0,0)$ dada la igualdad

$$\Phi(0,0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin \rho,$$

lo que nos lleva a ver que

$$p_c = \frac{1}{2} + \frac{1}{\pi} \arcsin \rho.$$

Finalmente, si sustituimos en la expresión dada anteriormente, podremos ver que ciertamente

$$\tau = 2\left(\frac{1}{2} + \frac{1}{\pi} \arcsin \rho\right) - 1 = \frac{2}{\pi} \arcsin \rho.$$

Capítulo 3

El coeficiente de correlación de distancias

3.1. Correlación entre vectores multidimensionales

Hasta ahora hemos visto la correlación aplicada al caso de dos variables aleatorias, que es equivalente a aplicarla a un vector bidimensional, así que procederemos a ver lo que ocurre en el caso multidimensional. Tendremos varias opciones de estudio para estos casos: estudio por pares de variables, de una variable respecto a las restantes, entre dos variables eliminando el efecto de las demás o entre todas las variables. Procederemos a dar unas nociones de todos los casos (todos son estudiados en más profundidad en [6]).

3.1.1. Pares de variables sin excluir las restantes

Empezaremos por estudiar la relación lineal entre dos variables del vector X. Para ello usaremos el coeficiente de correlación de Pearson, por lo que denotaremos por r_{ij} al coeficiente de Pearson que mide la relación entre las variables X_i y X_j.

Una vez tenemos todos los coeficientes calculados, estamos en disposición de formar la matriz de correlación. Esta matriz será la encargada de darnos la relación lineal existente entre las variables del vector y será cuadrada y simétrica. Construimos dicha matriz, que denotaremos por $\mathbf{R}$, como

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}.$$

Considerando $\mathbf{S}$ la matriz de covarianzas del vector y $\mathbf{D}$ la diagonal de dicha matriz (que contendrá las varianzas de las variables), obtendremos algo similar a la expresión del coeficiente de correlación de Pearson muestral pero aplicado a matrices:

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{D}^{-\frac{1}{2}}.$$

3.1.2. Una variable frente al resto

Además de estudiar la relación entre pares de variables podemos estudiar la relación entre una variable y todas las demás. Este caso es más conocido como regresión múltiple. Denotamos por Y la variable que pretendemos explicar y por X_i, con $i = 1, ..., p$, a cada una de las p variables restantes. Definiremos un estimador lineal de y_i a través de las variables restantes como

$$\hat{y_i} = \overline{y_i} + \hat{\beta}_1(x_{i1} - \overline{x}) + \hat{\beta}_p(x_{ip} - \overline{x}), \text{ con } i = 1, ..., n,$$

siendo $\hat{\beta}_i$ los coeficientes que proporcionan una mejor aproximación respecto a los y_i. Para poder asegurar que es una aproximación adecuada, usaremos el método de mínimos cuadrados sobre los errores.

Consideremos $\mathbf{X_R}$ la matriz de datos de las variables que pretendemos usar para explicar Y y siendo $\mathbf{y}$ el vector que contiene los datos de la variable a explicar. Partiendo de esto, tenemos que

$$\mathbf{X'_R y} = \mathbf{X'_R X_R}\hat{\beta},$$

lo que nos llevará a poder calcular $\hat{\beta}$ como

$$\hat{\beta} = (\mathbf{X'_R X_R})\mathbf{X'_R y} = \mathbf{S_p^{-1} S_{XY}},$$

siendo $\mathbf{S_p}$ la matriz de covarianzas de las p variables explicativas y $\mathbf{S_{XY}}$ el vector que contiene las covarianzas entre Y y el resto de variables.

Definiendo los errores como $e_i = y_i - \hat{y}_i$ podemos definir el coeficiente de correlación múltiple, también conocido como coeficiente de determinación. Sacaremos este coeficiente a través de la identidad

$$y_i - \overline{y} = \hat{y}_i - \overline{y} + e_i$$

.

Definición 3.1. Definimos el **coeficiente de determinación** $R^2_{y,1,...,p}$ como

$$R^2_{y,1,...,p} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{\sum e_i^2}{\sum(y_i - \overline{y})^2}.$$

Si consideramos Y como parte de X, $Y = X_j$, denotaremos el coeficiente como $R^2_{j,1,...,r}$ siendo r la nueva dimensión de X.

3.1.3. Pares de variables eliminando el efecto de las restantes

La correlación entre dos variables eliminando el efecto del resto de ellas se mide a través del coeficiente de correlación parcial. Antes de definir el coeficiente, deberemos calcular los residuos sobre la regresión respecto a las variables sobrantes. También calcularemos el coeficiente de correlación simple entre estos dos residuos, que obtendremos estandarizando los datos de la matriz inversa de la matriz de covarianzas de los residuos, $\mathbf{S^{-1}}$. Dado todo esto, podemos definir ahora el coeficientes de correlación parcial.

Definición 3.2. Sean X el conjunto de las variables $(x_1, ..., x_n)$ y s^{ij} los elementos de la matriz $\mathbf{S^{-1}}$. Definimos el **coeficiente de correlación parcial** entre x_i y x_j como

$$r_{ij,1,\dots,n} = \frac{s^{ij}}{\sqrt{s^{ii}s^{jj}}}$$

También tendremos la relación siguiente

$$1 - r^2_{12,3,\dots,n} = \frac{1 - R^2_{1,2,\dots,n}}{1 - R^2_{1,3,\dots n}},$$

considerando que buscamos medir la correlación entre las variables x_1 y x_2. Pese a esto, no se pierde la generalidad.

Definición 3.3. Podemos definir la **matriz de correlaciones parciales**, $\mathbf{P}$, como aquella que contiene los coeficientes de correlación parcial entre pares de variables eliminando el efecto de las restantes. Por tanto, la podemos expresar como:

$$\mathbf{P} = \begin{bmatrix} 1 & r_{12,3,\dots,n} & \cdots & r_{1n,2,\dots,n-1} \\ r_{21,3,\dots,n} & 1 & \cdots & r_{2n,1,\dots,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1,2,\dots,n-1} & r_{n2,1,\dots,n-1} & \cdots & 1 \end{bmatrix}.$$

3.1.4. Todas las variables

Para obtener una medida conjunta de la dependencia entre las variables podemos utilizar el determinante de la matriz de correlación, que mide el alejamiento del conjunto de variables de la situación de perfecta dependencia lineal. Esta medida verificará las siguientes propiedades:

- $0 \leq |\mathbf{R}| \leq 1$.
- Si las variables a estudiar son incorreladas, entonces $\mathbf{R}$ es una matriz diagonal y por tanto $|\mathbf{R}| = 1$.

- Si al menos una de las variables es combinación lineal del resto, $\mathbf{R}$ es una matriz singular y por tanto $|\mathbf{R}| = 0$.
- En el caso general, $|\mathbf{R}_p| = (1 - R^2_{p,1,\dots,p-1})(1 - R^2_{p-1,1,\dots,p-2}) \cdots (1 - R^2_{2,1})$.

Se puede considerar la siguiente medida de dependencia lineal global:

$$D(\mathbf{R}_p) = 1 - |\mathbf{R}_p|^{\frac{1}{p-1}}.$$

A partir de esta expresión podremos dar un coeficiente de correlación que denominaremos coeficiente de correlación promedio, $\overline{\rho}$. Este será el encargado de darnos la relación general entre las variables.

Definición 3.4. Definimos el **coeficiente de correlación promedio** como

$$\overline{\rho}(\mathbf{R}_p) = D(\mathbf{R}_p)^{\frac{1}{2}} = \sqrt{1 - |\mathbf{R}_p|^{\frac{1}{p-1}}}$$

Todos los métodos de estudio de la correlación en vectores multidimensionales que acabamos de ver coinciden en su uso de matrices. Este detalle será un problema grave una vez empecemos a trabajar en altas dimensiones, ya que la cantidad de datos no será manejable. Además, las mediciones que hemos dado para los casos multidimensionales se basan en el coeficiente de correlación de Pearson, siendo únicamente capaces de medir relaciones lineales entre las variables. Para solucionar este problema surgirá el coeficiente que estudiaremos a continuación: El coeficiente de correlación de distancias.

3.2. Correlación de distancias

3.2.1. Aproximación a la correlación de distancias

El coeficiente de correlación de distancias, $\mathcal{R}$, surge como una medida de dependencia que tiene como objetivo los casos de vectores de alta dimensión. El coeficiente de correlación de distancias es el más recientes de los que estudiaremos (fue definido por primera vez en [8]). Su enfoque está principalmente centrado en una de las grandes áreas de la estadística actualmente, el big-data.

Este coeficiente tiene unas hipótesis similares a las de Pearson. En este caso, el coeficiente de correlación de distancias requiere que todas las distribuciones involucradas tengan medias finitas. Si se cumplen estas condiciones, el coeficiente nos asegura:

1. $\mathcal{R}(X, Y)$ está definido para $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$, con p y q dimensiones arbitrarias.
2. $\mathcal{R}(X, Y)$=0 solo si las variables X e Y son independientes.

Dados estos vectores, queremos ver si existe una dependencia entre ellos. Para ello empezaremos considerando las funciones características de cada uno de ellos (f_X,f_Y), a mayores de la función característica conjunta ($f_{X,Y}$). La covarianza de distancias ($\mathcal{V}$), en la que se basa este coeficiente de correlación, se aplicará sobre la distancia dada por $\|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|$. Esta norma mide la distancia entre las funciones y nos da una idea de la posible correlación existente, ya que si X e Y son independientes, entonces $f_{X,Y} = f_X f_Y$. De esta manera vemos que se verifica el argumento 2. dado antes y da pie a hipótesis de independencia aplicables:

$$H_0 \colon f_{X,Y} = f_X f_Y$$

$$H_1 \colon f_{X,Y} \neq f_X f_Y$$

Para llevar a cabo la medición necesaria de la distancia entre funciones características, debemos definir la norma que utilizaremos.

Definición 3.5. Para cualquier función compleja γ definida sobre el espacio $\mathbb{R}^p$x$\mathbb{R}^q$ definimos la norma $\|\cdot\|_w$ como

$$\|\gamma(t,s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(t,s)|^2 w(t,s)dtds.$$

En la definición anterior, el espacio $\mathbb{R}^{p+q}$ en el que se calcula la integral debe ser un espacio L_2. A mayores, $w(t,s)$ se considera una función peso arbitraria y positiva para la que exista la integral. Será especialmente importante la elección de esta función, ya que no todas generan un coeficiente $\mathcal{R}_w$ interesante. Por ejemplo, al estimador deberíamos pedirle que sea invariante ante cambios de escala.

Una vez tenemos definida la norma, pasamos a definir una medida general de dependencia entre las variables. Por ello, no consideraremos todavía una función peso específica.

Definición 3.6. Tomemos $\|\cdot\|_w$ definida en 3.1. Considerando una función peso $w(t,s)$ adecuada, definimos una medida de dependencia como

$$\mathcal{V}^2(X,Y;w) = \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_w^2 =$$

Entonces, será equivalente

$$\mathcal{V}^2(X,Y;w) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)dtds$$

Observación 3.7. Si bien esta última definición nos dirige hacia la covarianza de distancias, podemos ver que la varianza de distancias surgirá de la idea de considerar las dos variables idénticas:

$$\mathcal{V}^2(X;w) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|_w^2$$

$$= \int_{\mathbb{R}^{p+q}} |f_{X,X}(t,s) - f_X(t)f_X(s)|^2 w(t,s)dtds$$

Observación 3.8. Consideremos que la función peso $w(t,s)$ es una función integrable y tanto x y Y tienen varianzas finitas. Aplicando una transformación de escala de la forma $(X,Y) \mapsto (\epsilon X, \epsilon Y)$ y realizando el límite cuando $\epsilon \to 0$ obtenemos que

$$\lim_{\epsilon \to 0} \frac{\mathcal{V}^2(\epsilon X, \epsilon Y; w)}{\sqrt{\mathcal{V}^2(\epsilon X; w)\mathcal{V}^2(\epsilon Y; w)}} = \rho^2(X,Y).$$

Esto presenta un problema, ya que demuestra que una función peso integrable no nos asegura un coeficiente de correlación que sea invariante a transformaciones de escala. Para solventar este problema, debemos dar varios pasos, siendo clave la siguiente proposición (Lemma 1. en [8]).

Proposición 3.9. *Si* $0 < \alpha < 2$*, entonces para todo* $x \in \mathbb{R}^d$ *tenemos que*

$$\int_{\mathcal{R}}^{d} \frac{1 - cos\langle t, x\rangle}{|t|_d^{d+\alpha}} dt = C(d,\alpha)|x|^{\alpha},$$

donde

$$C(d,\alpha) = \frac{2\pi^{d/2}\Gamma(1-\frac{\alpha}{2})}{\alpha 2^{\alpha}\Gamma(\frac{d+\alpha}{2})}$$

y $\Gamma(\cdot)$ *es la función gamma. Las integrales en 0 y* ∞ *se deberán calcular como* $\lim_{\mathbb{R}^d \setminus [\epsilon B + \epsilon^{-1}B^c]}$*, siendo* B *la bola unitaria centrada en 0 de* $\mathbb{R}^d$ *y* B^c *el complementario de* B*.*

Gracias a esta proposición, llegamos a la función peso utilizada para el coeficiente de distancias. Dicha función será

$$w(t,s) = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1},$$

siendo $c_d = C(d,1)$, definido en la proposición anterior. A su vez, por estar trabajando con integrales, definimos dw que será

$$dw = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1} dtds$$

Dadas las hipótesis del coeficiente, junto con la elección de la función peso, tenemos que por la desigualdad de Cauchy-Schwartz

$$|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 \leq (1 - |f_X(t)|^2)(1 - |f_Y(s)|^2).$$

Utilizando también la Proposición 3.9 y el Teorema de Fubini obtenemos que

$$\int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 dw \leq E|X - X'|_p E|Y - Y'|_q < \infty,$$

siendo X y X' variables independientes idénticamente distibuídas. Lo mismo ocurre con Y e Y'.

Ahora ya estamos en condiciones de definir la varianza y covarianza de distancias.

3.2.2. Definiendo la correlación de distancias

Definición 3.10. Definimos la **covarianza de distancias** entre dos vectores aleatorios X e Y con medias finitas como el $\mathcal{V}(X,Y)$ no negativo obtenido de

$$\mathcal{V}^2(X,Y) = \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2 =$$

$$= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} dtds$$

Análogamente, definimos la **varianza de distancias** como la raíz cuadrada positiva de

$$\mathcal{V}^2(X) = \mathcal{V}^2(X,X)\|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2$$

Tras esta definición quedan sentadas las bases para poder alcanzar el coeficiente de correlación de distancias. Procedemos a definirlo.

Definición 3.11. Definimos el **coeficiente de correlación de distancias** entre dos vectores aleatorios X e Y con medias finitas como el número no negativo $\mathcal{R}(X,Y)$ definido por

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \text{si } \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \text{si } \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases}$$

Una vez tenemos definidos el coeficiente, la covarianza y la varianza teóricos, faltan por definir los equivalentes empíricos. Previamente a que podamos definirlos, debemos definir otros elementos que ayudarán en el cálculo de los mismos.

Dada una muestra (X,Y)=(X_i,Y_i),$i = 1,...,n$, siendo $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$ definimos

$$a_{kl} = |X_k - X_l|_p, \quad \overline{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \quad \overline{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl} \ ,$$

$$\overline{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \quad \text{y } A_{kl} = a_{kl} - \overline{a}_{k\cdot} - \overline{a}_{\cdot l} + \overline{a}_{\cdot\cdot}$$

para $k,l = 1,...,n$. Análogamente, definimos $b_{kl} = |Y_k - Y_l|_q$ y $B_{kl} = b_{kl} - \overline{b}_{k\cdot} - \overline{b}_{\cdot l} + \overline{b}_{\cdot\cdot}$. Una vez tenemos estos elementos definidos, pasamos a definir los casos empíricos.

Definición 3.12. Definimos la **covarianza de distancias empírica**, $\mathcal{V}_n(X,Y)$, como la raíz cuadrada positiva de

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}.$$

Por tanto, **la varianza empírica**, $\mathcal{V}_n(X)$, quedará definida como la raíz cuadrada positiva de

$$\mathcal{V}_n^2(X) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2.$$

Y, finalmente, el **coeficiente de correlación de distancias empírico**,

$$\mathcal{R}_n(X,Y)$$

, será la raíz cuadrada positiva dada por

$$\mathcal{R}_n^2(X,Y) = \begin{cases} \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \text{si } \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0 \\ 0, & \text{si } \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0 \end{cases}$$

3.2.3. Propiedades y relaciones

El coeficiente de correlación de distancias posee diversas propiedades interesantes. En el siguiente teorema veremos unas cuantas.

Teorema 3.13. *Propiedades del dCor.*

1. *Si $E(|X|_p + |Y|_q) < \infty$, entonces $0 \leq \mathcal{R} \leq 1$, y además $\mathcal{R}(X,Y)$=0 si y solo si X e Y son independientes.*
2. $0 \leq \mathcal{R}_n \leq 1$
3. *Si $\mathcal{R}_n(X,Y)$=1, entonces existe un vector α, un número no nulo β y una matriz ortogonal Θ tales que $Y = \alpha + \beta X\Theta$*

Demostración 3.14. *1. $\mathcal{R}$ existe siempre que X e Y tengan media finita. Además, X e Y son independientes si y solo si el numerador de $\mathcal{R}^2(X,Y)$ es 0. Es decir, si*

$$\mathcal{V}^2(X,Y) = \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2 = 0.$$

Consideremos ahora $U = \exp i\langle t, X\rangle - f_X(t)$ y $V = \exp i\langle s, Y\rangle - f_Y(s)$. Entonces tenemos que

$$|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 = |E[UV]|^2 \leq (E[|U||V|])^2 \leq E[|U|^2|V|^2] =$$

$$= (1 - |f_X(t)|^2)(1 - |f_Y(s)|^2).$$

Por tanto,

$$\int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 dw \leq \int_{\mathbb{R}^{p+q}} |(1 - |f_X(t)|^2)(1 - |f_Y(s)|^2)|^2 dw.$$

Por tanto, $0 \leq \mathcal{R}_n \leq 1$.

2. *La demostración de este apartado es análoga al del apartado 1.*

3. *Supongamos, sin pérdida de generalidad que X e Y pertenecen al mismo espacio euclídeo y que ambos esán contenidos en $\mathbb{R}^p$. Por la desigualdad de Cauchy-Schwartz, podemos ver que $\mathcal{R}_n(X,Y) = 1$ si y solo si $A_{kl} = \epsilon B_{kl}$ para un cierto ϵ. Supongamos que $|\epsilon| = 1$, entonces*

$$|X_k - X_l|_p = |Y_k - Y_l|_q + d_k + d_l, \forall k, l = 1, ..., n$$

siendo d_k y d_l constantes. Considerando $k = l$, tenemos que $d_k = 0$ para todo k. Ahora debemos aplicar un argumento geométrico. Ambas muestras son isométricas, por lo que podemos obtener Y en función de X a través de operaciones de traslación, rotación y reflexión. Por tanto, podemos expresar la igualdad $Y = \alpha + \beta X\Theta$, para un cierto vector α, $\beta = \epsilon$ y una matriz ortogonal Θ. En caso de que $|\epsilon| \neq 1$ y $\epsilon \neq 0$, aplicamos el criterio geométrico a ϵX e Y, obteniendo la misma expresión.

Otra característica importante del coeficiente de correlación de distancias es que es sensible a la dependencia no monótona entre variables. Esto es claramente una mejora respecto a los coeficientes de Spearman y Kendall, que recogían dependencias monótonas y que a su vez superaban la dependencia lineal medida por el coeficiente de Pearson.

Además de estas propiedades, el coeficiente de correlación de distancias es invariante para transformaciones de la forma

$$X \mapsto a + bCX,$$

$$Y \mapsto a' + b'C'X$$

donde los a, a' son vectores arbitrarios, b, b' son números no nulos arbitrarios y C, C' son matrices ortogonales arbitrarias.

También cabe destacar que la correlación de distancias es adaptable para que sea invariante por grupos afines. Esta propiedad es interesante, ya que nos permite asegurar que no se verá afectado en transformaciones afines de los datos. Para ello, definimos:

$X^* = XS_X^{-\frac{1}{2}}$, siendo $S_X$la matriz de covarianza muestral de X.

$Y^* = YS_Y^{-\frac{1}{2}}$, siendo $S_Y$la matriz de covarianza muestral de Y.

Ahora podemos definir de manera equivalente a los anteriores coeficientes de esta correlación un nuevo estadístico, que llamaremos estadístico de correlación de distancias afín. Este estadístico lo denotaremos por $\mathcal{R}_n^{*2}(X,Y)$ y será la raíz cuadrada positiva de

$$\mathcal{R}_n^{*2}(X,Y) = \begin{cases} \frac{\mathcal{V}_n^2(X^*,Y^*)}{\sqrt{\mathcal{V}_n^2(X^*)\mathcal{V}_n^2(Y^*)}}, & \text{si } \mathcal{V}_n^2(X^*)\mathcal{V}_n^2(Y^*) > 0 \\ 0, & \text{si } \mathcal{V}_n^2(X^*)\mathcal{V}_n^2(Y^*) = 0 \end{cases}$$

Finalmente, será interesante ver si esta medida de correlación tiene alguna relación con las vistas anteriormente. En este caso, tenemos un teorema que nos relaciona el coeficiente de correlación de distancias con el coeficiente de correlación de Pearson.

Teorema 3.15. *Si X e Y son normales estándar con correlación $\rho = \rho(X,Y)$, entonces se verifica que:*

1. $\mathcal{R}(X,Y) \leq |\rho|$
2. $\mathcal{R}^2(X,Y) = \frac{\rho \arcsin \rho + \sqrt{1-\rho^2} - \rho \arcsin \frac{\rho}{2} - \sqrt{4-\rho^2} + 1}{1 + \frac{\pi}{3} - \sqrt{3}}$
3. $\text{ínf}_{\rho \neq 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = \lim_{\rho \to 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = \frac{1}{2(1+\frac{\pi}{3}-\sqrt{3})^{\frac{1}{2}}} \cong 0{,}89066$

Observación 3.16. Antes de demostrar el teorema, debemos tener en cuenta las condiciones en las que nos encontramos. Por ello, debemos definir una nueva función:

$$F(\rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f_{XY}(t,s) - f_X(t) f_Y(s)|^2 \frac{dt}{t^2} \frac{ds}{s^2}.$$

Entonces, $\mathcal{V}^2(X,Y) = \frac{F(\rho)}{c_1^2} = \frac{F(\rho)}{\pi^2}$, y por tanto

$$\mathcal{R}_n^2(X,Y) = \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X,X)\mathcal{V}_n^2(Y,Y)}} = \frac{F(\rho)}{F(1)}.$$

Con esto ya estamos en condiciones de demostrar el teorema.

Demostración 3.17 (descrita en [8]). *1. Si X e Y son normales estándar con correlación ρ, entonces*

$$F(\rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |e^{-(t^2+s^2)/2-\rho ts} - e^{-t^2/2} e^{-s^2/2}|^2 \frac{dt}{t^2} \frac{ds}{s^2} =$$

$$= \int_{\mathbb{R}^2} e^{-t^2-s^2} (1 - 2e^{-\rho ts} + e^{-2\rho ts}) \frac{dt}{t^2} \frac{ds}{s^2} =$$

$$= \int_{\mathbb{R}^2} e^{-t^2-s^2} \sum_{n=2}^{\infty} \frac{2^n - 2}{n!} (-\rho ts)^n \frac{dt}{t^2} \frac{ds}{s^2} =$$

$$= \int_{\mathbb{R}^2} e^{-t^2-s^2} \sum_{k=1}^{\infty} \frac{2^{2k} - 2}{(2k)!} (-\rho ts)^{2k} \frac{dt}{t^2} \frac{ds}{s^2} =$$

$$= \rho^2 \left[\sum_{k=1}^{\infty} \frac{2^{2k} - 2}{(2k)!} (-\rho ts)^{2k} \rho^{2(k-1)} \int_{\mathbb{R}^2} e^{-t^2-s^2} (ts)^{2(k-1)} dt ds \right].$$

Entonces tenemos que $F(\rho) = \rho^2 G(\rho)$*, donde* $G(\rho)$ *es un sumatorio con todos los términos no negativos. Trivialmente, la función* $G(\rho)$ *es no decreciente en* ρ *y* $G(\rho) \leq G(1)$*. Por tanto,*

$$\mathcal{R}^2(X,Y) = \frac{F(\rho)}{F(1)} = \frac{G(\rho)}{G(1)} \leq \rho^2.$$

Dada esta desigualdad, tenemos que se verifica el apartado 1.

2. *Conocemos que* $F(0) = F'(0) = 0$*, por lo que*

$$F(\rho) = \int_0^\rho \int_0^x F''(z)dzdx.$$

Se define la segunda derivada de F *como*

$$F''(z) = \frac{d^2}{dz^2} \int_{\mathbb{R}^2} e^{-t^2-s^2}(1 - 2e^{-zts} + e^{-2zts})\frac{dt}{t^2}\frac{ds}{s^2} = 4V(z) - V(\frac{z}{2}),$$

donde está definida la función $V(z)$ *como*

$$V(z) = \int_{\mathbb{R}^2} e^{-t^2-s^2-2zts}dtds = \frac{\pi}{\sqrt{1-z^2}}.$$

Realizando ahora un cambio de variable basado en el hecho de que los autovalores asociados a $t^2 + s^2 + 2zts$ *son* $1 \pm z$ *y* $\int_{-\infty}^{\infty} e^{-t^2\lambda}dt = (\frac{\pi}{\lambda})^{\frac{1}{2}}$*. Entonces tenemos que*

$$F(\rho) = \int_0^\rho \int_0^x \left(\frac{4\pi}{\sqrt{1-z^2}} - \frac{2\pi}{\sqrt{1-z^2/4}}\right) dzdx =$$

$$= 4\pi \int_0^\rho (\arcsin(x) - \arcsin(\frac{x}{2}))dx =$$

$$4\pi(\rho \arcsin \rho + \sqrt{1+\rho^2} - \rho \arcsin \frac{\rho}{2} - \sqrt{4-\rho^2} + 1).$$

Si unimos esto con lo obtenido en la observación previa, vemos que se verifica el enunciado

3. *Tenemos demostrado en 1. que* $\frac{\mathcal{R}}{|\rho|}$ *es una función no decreciente de* $|\rho|$*. Aplicando esto junto con el apartado 2., llegamos a que*

$$\lim_{|\rho|\to 0} \frac{\mathcal{R}(X,Y)}{|\rho|} = \frac{1}{2\sqrt{1+\frac{\pi}{3}-\sqrt{3}}}$$

Así pues, queda demostrado el apartado 3.

Capítulo 4

Las funciones cópula

4.1. Introducción al las funciones cópula

Las funciones cópula son funciones que buscan relacionar distribuciones multivariantes con las distribuciones marginales de cada una de las variables. Por ello, las funciones cópula son distribuciones multivariantes cuyas marginales son uniformes en el intervalo (0,1). Son una herramienta que está sufriendo un gran avance actualmente. Surge antes de la aparición del coeficiente de correlación de distancias para cubrir la necesidad existente de alguna medida que fuese capaz de medir dependencias no lineales entre variables. Actualmente se usan principalmente en la matemática financiera, siendo importantes en la toma de decisiones.

Para ser funciones cópula deben verificar ciertas propiedades:

Propiedades 4.1. *Consideremos $C(u, v)$ una función cópula bivariante. Entonces se verifican:*

- *Para todo $u, v \in (0, 1)$, tenemos que*

$$C(u, 0) = 0 = C(0, v),\ C(u, 1) = u,\ C(1, v) = v.$$

- $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$
- *La cópula es continua en u y v. De hecho, verifica las condiciones fuertes de Lipschitz.*

$$|C(u_2, v_2) - C(u_1, v_1)| \leq |u_2 - u_1| + |v_2 - v_1|.$$

- *Para $0 \leq u_1 < u_2 \leq 1$ y $0 \leq v_1 < v_2 \leq 1$,*

$$P(u_1 \leq U \leq u_2, v_1 \leq V \leq v_2) =$$

$$= C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) > 0.$$

Observación 4.2. Dadas estas propiedades, podemos asegurar que las siguientes funciones son válidas como funciones cópula:

- $C^+(u,v) = \text{mín}(u,v)$.
- $C^-(u,v) = \text{máx}(u+v-1,0)$.
- $C^0(u,v) = uv$.

A parte de estas propiedades, la base de las funciones cópula es el Teorema de Sklar. Este teorema nos da la función que cumplen en la relación entre las distribuciones bivariantes y sus marginales univariantes. Se considera el caso bivariante por ser más sencillo, pero es extensible a casos multivariante.

Teorema 4.3 (Teorema de Sklar)**.** *Sea H una función distribución conjunta con distribuciones marginales F y G. Entonces, existe una cópula C tal que $\forall x, y \in [-\infty, \infty]$, tenemos que*

$$H(x,y) = C(F(x), G(y)).$$

Si F y G son continuas, entonces la cópula C es única. En otro caso, C está definida únicamente determinada en el espacio (rango(F) x rango(G)). Recíprocamente, si C es una función cópula y F y G son funciones de distribución univariante, entonces H es una función de distribución conjunta con F y G como distribuciones marginales.

Sin embargo, estas no son las únicas propiedades aplicables a las funciones cópula. Estas son otras de las que verifican:

Propiedades 4.4.

- *Para cada cópula C y cualquier $(u,v) \in [0,1] \times [0,1]$,*

  $$C^-(u,v) \leq C(u,v) \leq C^+(u,v),$$

  tomando $C^+(u,v) = \text{mín}(u,v)$ y $C^-(u,v) = \text{máx}(u+v-1,0)$.

- *Para todo $v \in [0,1]$, existe la derivada parcial $\frac{\partial C}{\partial u}$ para casi todo u, y $0 \leq \frac{\partial}{\partial u} C(u,v) \leq 1$. Análogamente, $0 \leq \frac{\partial}{\partial v} C(u,v) \leq 1$*

- *$C(u,v) = uv$ es la cópula asociada al par de variables aleatorias independientes (U,V).*

- *La combinación convexa de dos cópulas es también una función cópula.*

- *Las cópula asociada con la función densidad de una normal estándar bivariante tiene una función densidad*

  $$c(u,v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{\rho^2}{2(1-\rho^2)} ([\Phi^{-1}(u)]^2 + [\Phi^{-1}(v)]^2) + \frac{\rho}{1-\rho^2} \Phi^{-1}(u)\Phi^{-1}(v) \right].$$

- *La función cópula se conserva ante transformaciones estrictamente crecientes sobre las variables propias de las distribuciones marginales.*

4.2. Cópulas arquimedianas

Existe una gran diversidad de funciones cópula, que se agrupan en diversos grupos. Veremos ahora diversos tipos, empezando por la cópula arquimediana.

Definición 4.5. Llamamos **cópulas arquimedianas** a aquellas que son de la forma

$$\varphi(C(u,v)) = \varphi(u) + \varphi(v).$$

De manera equivalente, podemos reescribir la expresión anterior como

$$\varphi(H(x,y)) = \varphi(F(x)) + \varphi(G(y)).$$

Como buscamos la función cópula en las expresiones anteriores, necesitamos calcular la inversa de φ, a la que denotaremos por $\varphi^{[-1]}$. Para ello necesitamos la siguiente definición.

Definición 4.6. Sea $\varphi : [0,1] \to [0,\infty]$ una función continua y estrictamente decreciente y tal que $\varphi(1) = 0$. Definimos la **función pseudoinversa de** φ es la función $\varphi^{[-1]}$, que tiene como dominio el intervalo $[0,\infty]$ y como rango el intervalo [0,1]. La definiremos como

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t) & 0 \leq t \leq \varphi(0) \\ 0 & \varphi(0) \leq t \leq \infty \end{cases}$$

Observación 4.7. Observemos que si $\varphi(0) = \infty$, entonces $\varphi^{[-1]}(t) = \varphi^{-1}(t)$, siendo entonces

$$C(u,v) = \varphi^{-1}(\varphi(u) + \varphi(v))$$

una cópula si y solo si la función pseudoinversa es una función convexa decreciente.

La función φ se llama generador de la cópula. Además, si $\varphi(0) = \infty$ entonces diremos que es un generador estricto y $C(u,v) = \varphi^{-1}(\varphi(u) + \varphi(v))$ se definirá como una cópula arquimediana estricta.

4.3. Cópulas gaussianas

Las cópulas gaussianas son las más utilizadas en general, principalmente en el ámbito de la toma de riesgos, donde creó graves problemas en la economía por su aplicación. Para definirlas, denotaremos por $\Phi(x)$ a la función distribución de la normal estándar univariante y por $\varphi(x)$ a la función densidad de la normal estándar univariante. Entonces,

siendo $u = \frac{x-\mu}{\sigma}$, la distribución normal general tendrá una función densidad de la forma $f_{(1)}(x) = \frac{1}{\sigma}\varphi(u)$ y una función de distribución acumulada de la forma $F_{(1)}(x) = \Phi(u)$. Si consideramos ahora el caso una normal general de dimensión p, entonces tenemos que

$$f_{(p)}(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \prod_{i=1}^{p} \sigma_i |R|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}u'R^{-1}u\right\},$$

siendo $u = (u_1, ..., u_p)'$, $u_i = \frac{x_i-\mu_i}{\sigma_i}$ y R la matriz de correlación.

Tomemos ahora la función densidad de la normal multivariante, que podemos expresar como

$$f_{(p)}(x) = \frac{1}{|R|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}u'(R^{-1}-1)u\right\} \prod_{i=1}^{p} \frac{1}{\sigma_i}\varphi(x).$$

Siendo $u_i = \Phi^{-1}(F_i(x_i))$, la función densidad de una cópula gaussiana será de la forma

$$c(x) = \frac{1}{|R|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}u'(R^{-1}-1)u\right\},$$

donde F_i es una función distribución arbitraria y continua.

Definidas las cópulas gaussianas, debemos ser conscientes también de sus limitaciones. Entre ellas, podemos destacar las siguientes (vistas en [11]):

- En caso de grandes dimensiones, R puede ser difícil de estimar debido a la gran cantidad de parámetros involucrados.
- Las cópulas gaussianas se basan en el coeficiente de correlación de Pearson, por lo que no serán invariantes a transformaciones monótonas.
- El coeficiente de correlación de Pearson no es una herramienta adecuada en diversos casos, lo que nos impedirá usar eficientemente las cópulas gaussianas en esos mismos casos.

4.4. Correlación a través de cópulas

4.4.1. τ de Kendall

Dadas las propiedades de las funciones cópulas, podemos expresar ciertas medidas de dependencia en términos de funciones cópula. Veremos las expresiones del coeficiente de correlación de Spearman y la τ de Kendall. Empezaremos por esta última.

Consideremos (x_i, y_i) y (x_j, y_j) dos observaciones obtenidas a partir de las variables continuas (X, Y). Diremos que el par será concordante si $(x_i - x_j)(y_i - y_j) > 0$ y será discordante si $(x_i - x_j)(y_i - y_j) < 0$.

Como hemos visto antes, podemos definir la τ de Kendall como

$$\tau = P[(X - X')(Y - Y') \geq 0] - P[(X - X')(Y - Y') \leq 0],$$

siendo X' e Y' son variables independientes idénticamente distribuidas respecto a las variables X e Y.. Además, la τ de Kendall originalmente fue definido ([3]) como

$$\tau = \frac{c - d}{c + d} = \frac{c - d}{n},$$

siendo c el número de pares concordantes y siendo d el número de pares discordantes dentro de una muestra de n observaciones de (X, Y).

La expresión de la τ de Kendall como una cópula es

$$\tau = 4 \int_0^1 \int_0^1 C(u, v)c(u, v)dudv - 1 = 4E(C(U, V)) - 1.$$

Observación 4.8. Si la función cópula C es una cópula arquimediana generada por la función φ, entonces tenemos que podemos expresar la relación entre la τ de Kendall y la cópula C como

$$\tau = 4E(C(U, V)) - 1 = 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt.$$

4.4.2. Coeficiente de Spearman

Tenemos ahora vista la expresión de la τ de Kendall como una función cópula, así que procederemos ahora a ver la expresión del coeficiente de correlación de Spearman.

Consideremos $(X_1, Y_1), (X_2, Y_2)$ y (X_3, Y_3) tres pares independientes de observaciones de variables aleatorias con H como función de distribución conjunta. Entonces el coeficiente de correlación de Spearman, ρ_s, es proporcional a la probabilidad de que un par de observaciones, (X_1, Y_1) y (X_2, Y_3), sea concordante menos la probabilidad de que sea, es decir

$$\rho_s = 3 \left(P[(X_1 - X_2)(Y_1 - Y_3) > 0]) - P[(X_1 - X_2)(Y_1 - Y_3) < 0]) \right).$$

Si reescribimos esta expresión, obtenemos que el coeficiente de correlación de Spearman en términos de funciones cópula se puede escribir como

$$\rho_s = 12 \int_0^1 \int_0^1 C(u, v)dudv - 3 = 12E(U, V) - 3.$$

Si volvemos a reescribir esta última expresión tenemos que

$$\rho_s = \frac{E(U, V) - \frac{1}{4}}{\frac{1}{12}}.$$

Podemos deducir que el coeficiente de correlación de Spearman entre las variables X e Y se puede considerar símplemente un coeficiente de correlación de Pearson entre las variables uniformes U y V. Esta es una relación que no habíamos sido capaces de descubrir durante el estudio de ambos coeficientes en los capítulos anteriores.

4.4.3. Coeficiente de Gini

Sin embargo, estas no son las únicas maneras de medir correlación a través de las funciones cópula. Podemos destacar que debido a su uso en las matemáticas financieras es habitual unir las funciones cópula con el coeficiente de Gini. Dicho coeficiente está íntimamente ligado con la correlación de rangos y suele usarse como medida de la desigualdad de ingresos.

Para poder definir este coeficiente, consideraremos (X_i, Y_i), con $i = 1, ..., n$ y siendo n pares independientes de las variables X e Y. Reordenamos los datos a través de los rangos, es decir crearemos los pares $(X_{(i)}, Y_{[i]})$ ordenando como $X_{(1)} < X_{(2)} < ... < X_{(n)}$ y $Y_{[1]} < Y_{[2]} < ... < Y_{[n]}$. Cambiando las posiciones de X e Y obtendremos los pares $(Y_{(i)}, X_{[i]})$.

Definición 4.9 (en [10])**.** Definimos el **coeficiente de Gini** $\gamma(X, Y)$ como

$$\gamma(X,Y) = \frac{\frac{1}{n(n-1)} \sum_{i=1}^{n} (2i - 1 - n) Y_{[i]}}{\frac{1}{n(n-1)} \sum_{i=1}^{n} (2i - 1 - n) Y_{(i)}}.$$

Análogamente, el **coeficiente de Gini** $\gamma(Y, X)$ se definirá por

$$\gamma(Y,X) = \frac{\frac{1}{n(n-1)} \sum_{i=1}^{n} (2i - 1 - n) X_{[i]}}{\frac{1}{n(n-1)} \sum_{i=1}^{n} (2i - 1 - n) X_{(i)}}.$$

Podemos ver que estas mediciones de la relación entre variables no verifican propiedades importantes de los coeficientes de correlación. Podemos demostrar sencillamente que $\gamma(X, Y) \neq \gamma(Y, X)$. Por ello debemos definir una nueva versión del coeficiente de Gini que sí se adapte a las necesidades propias de un coeficiente de correlación.

Definición 4.10 (en [10])**.** Definimos el **coeficiente de Gini simétrico**, γ^*, de dos variables X e Y como

$$\gamma^*(X,Y) = \frac{1}{2} \left[\gamma(X,Y) + \gamma(Y,X) \right].$$

Una vez lo tenemos definido de manera general, podemos ver su definición unido a las cópulas. Para este caso, usaremos el coeficiente de Gini simétrico y tendremos que está relacionado de manera directa con la función cópula usada para la unión de las variables X e Y.

Proposición 4.11. *Sean X e Y variables aleatorias. Siendo C una función cópula que une ambas variables, entonces tenemos que podemos expresar el coeficiente de Gini como*

$$\gamma_C = 2\int_0^1 \int_0^1 ([u+v-1]-[u-v])dC(u,v).$$

De manera equivalente, podemos describir también el coeficiente como

$$\gamma_C = 2E(|U+V-1|-|U-V|).$$

4.5. Observaciones y críticas

Pese a estas condiciones y la gran utilidad de las funciones cópula, existen críticas y observaciones sobre las mismas. Veamos alguna de ellas (Apartado 1.16 de [1]):

- El uso de las funciones cópula no tiene ninguna ventaja en particular. Bastaría usar una distribución multivariante adecuada para el problema y a sobre la que podamos aplicar las técnicas estadísticas que consideremos convenientes.
- Las distribuciones marginales y la función cópula de una distribución multivariante están íntimamente relacionadas. Eso nos lleva a un estudio sesgado de la dependencia.
- La gran cantidad de modelos de funciones cópula genera de que se seleccionen los modelos por conveniencia.
- Las funciones cópula se consideran una alternativa a los modelos gaussianos en casos no gaussianos. Sin embargo, ya que las cópulas pueden generar cualquier distribución, la cantidad de funciones cópula aplicables es demasiado grande para ser utilizables e incluso entendibles.
- Casi no existe una base teórica en cuanto a la utilización estadística de las funciones cópula. Por ahora no están diseñados los procesos de estimación o los test de bondad de ajuste que se les pudiese aplicar.
- Las funciones cópula no ayudan a entender qué ocurre en los extremos de la distribución multivariante.
- Las funciones cópula no tienen encaje en las teorías actuales de los procesos estocásticos y las series de tiempo.

Bibliografía

[1] Balakrishnan, Narayanaswamy and Lai, Chin Diew *Continuous bivariate distributions*, Springer Science & Business Media, 2009.

[2] Gibbons, Jean Dickinson and Chakraborti, Subhabrata, *Nonparametric statistical inference fourth edition, revised and expanded*, Statistics Textbooks and Monographs, Marcel Dekker AG, Vol 168, 2003.

[3] Kendall, Maurice G, *A new measure of rank correlation*, Biometrika, Vol 30, nº1/2, 81–93, 1938.

[4] Koch, Inge *Analysis of multivariate and high-dimensional data*, Cambridge University Press, Vol 32, 2013.

[5] Pearson, Karl, *Notes on the history of correlation*,Biometrika, Vol 13, nº1, 25–45, 1920.

[6] Peña, Daniel *Análisis de datos multivariantes*, McGraw-Hill España, 2013.

[7] Restrepo, Luis F and González, Julián, *From pearson to Spearman*, Revista Colombiana de Ciencias Pecuarias, Universidad de Antioquia, Vol 20, nº2, 183–192, 2007

[8] Székely, Gábor J and Rizzo, Maria L and Bakirov, Nail K and others, *Measuring and testing dependence by correlation of distances*, The annals of statistics,Institute of Mathematical Statistics, Vol 35, nº6, 2769–2794, 2007.

[9] Wissler, Clark, *The Spearman correlation formula*, Science, Vol 22, nº558, 309–311, 1905.

[10] Xu W. and Hung Y. S. and Niranjan M. and Shen M., *Asymptotic Mean and Variance of Gini Correlation for Bivariate Normal Samples*, IEEE Transactions on Signal Processing, Vol 58, nº2, 522–534, 2010.

[11] *Žežula*, Ivan, *On multivariate Gaussian copulas*, Journal of Statistical Planning and Inference, Vol 139, nº11, 3942–3946, 2009.