USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

El coeficiente de correlación.

Desde la independencia lineal de Pearson a la independencia general de variables aleatorias

Lucía Souto Vázquez

2023/2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

El coeficiente de correlación.

Desde la independencia lineal de Pearson a la independencia general de variables aleatorias

Lucía Souto Vázquez

Julio, 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa
Título: El coeficiente de correlación. Desde la independencia lineal de Pearson a la independencia general de variables aleatorias.
Breve descripción del contenido
Se trata de revisar los conceptos más notables desarrollados en la literatura estadística y 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 bases de datos reales.
Recomendaciones
-
Otras observaciones
Se adjuntarán dos anexos: ▪ Anexo I: Demostraciones de algunos resultados relevantes. ▪ Anexo II: Código de  empleado para la ilustración de datos.

Índice

Resumen

Este trabajo constituye una revisión acerca de las medidas de dependencia más comunes usadas para describir las relaciones existentes entre variables aleatorias.

En el primer capítulo se presentan algunos conceptos básicos relacionados con la teoría de la probabilidad que pueden resultar de utilidad para introducir el concepto de dependencia en los sucesivos capítulos. Ya en el segundo capítulo haremos un breve recorrido a través de las distintas nociones de dependencia y presentaremos un conjunto de propiedades deseables para las medidas globales de asociación que trataremos más adelante.

En el tercer capítulo introduciremos la medida de dependencia más reconocida actualmente, el *coeficiente de correlación de Pearson*, explicando su marco histórico, propiedades y limitaciones. Dichas limitaciones nos llevarán a presentar los coeficientes de correlación basados en el estudio de los rangos de las variables, que proporcionarán una visión más amplia de la dependencia con respecto al de Pearson, como la ρ *de Spearman* o la τ *de Kendall*. En el cuarto capítulo estudiaremos un coeficiente naciente que surge nuevamente debido a las deficiencias de los anteriores: el *coeficiente de correlación de distancias*. Además, ampliaremos el estudio dando algunas nociones de dependencia desde la perspectiva de las *funciones cópula*, haciendo un recorrido previo por su definición y propiedades.

Por último, en el capítulo final interpretaremos los conceptos anteriores en base a un conjunto de datos reales. Para obtener los resultados usaremos el lenguaje de programación R y apoyaremos nuestras conclusiones en gráficas para facilitar la comprensión del estudio.

Palabras clave: *dependencia, correlación, Pearson, Spearman, Kendall, correlación de distancias, cópula.*

Abstract

This paper is a review of the most common dependence measures used to describe the relationships between random variables.

In the first chapter we present some basic concepts related to probability theory that may be useful to introduce the concept of dependence in the following chapters. Already in the second chapter we will make a brief tour through the different notions of dependence and present a set of desirable properties for the global measures of association that we will discuss later.

In the third chapter we will introduce the currently most recognized measure of dependence, *Pearson's correlation coefficient*, explaining its historical framework, properties and limitations. These limitations will lead us to present correlation coefficients based on the study of the ranks of the variables, which will provide a wider view of dependence with respect to Pearson's, such as *Spearman's* ρ or *Kendall's* τ. In the fourth chapter we will study a nascent coefficient that arises again due to the deficiencies of the previous ones: the *distance correlation coefficient*. In addition, we will extend the study by giving some notions of dependence from the perspective of *copula functions*, making a preliminary tour of their definition and properties.

Finally, in the final chapter we will interpret the above concepts based on a real dataset. To obtain the results we will use the R programming language and support our conclusions with graphs to facilitate the understanding of the study.

Keywords: *dependence, correlation, Pearson, Spearman, Kendall, distance correlation, copula.*

Introducción

El concepto de dependencia surge de forma natural en el medio que nos rodea abarcando desde fenómenos tan elementales como los causados por la propia naturaleza hasta manifestaciones de la misma en campos tan diversos como la medicina, ingeniería, política, economía, etc. La dependencia es un concepto considerado de carácter estocástico, pues gran parte de los sucesos que ocurren en el mundo están sujetos a una cierta incertidumbre.

La observación de un fenómeno no basta para determinar una dependencia; es necesaria la construcción de medidas que ayuden a cuantificar las relaciones existentes. Sin embargo, las medidas de dependencia no fueron estudiadas en profundidad hasta una época tardía por lo que el concepto de correlación que introdujo el estadístico Francis Galton en 1888 fue la única medida reconocida y utilizada a pesar de ser inapropiada en muchas ocasiones (Samuel et al., 2001).

Las investigaciones acerca de la dependencia resurgieron hace relativamente poco porque, aunque el coeficiente de correlación tal y como estaba definido era una herramienta simple, sus deficiencias en muchos aspectos obligaron a los investigadores a buscar nuevas maneras de medir la dependencia. La ρ de Spearman o la τ de Kendall ampliaron el conocimiento en este ámbito ya que sus creadores se basaron en el estudio de los rangos de las variables consiguiendo coeficientes que llegaran a medir dependencias monótonas, una ventaja inmensa con respecto al coeficiente de correlación clásico que únicamente era capaz de detectar dependencias lineales.

Sin embargo, todos estos coeficientes eran incapaces de recoger dependencias no monótonas y además solo eran útiles cuando tratábamos de relacionar dos variables. Para modelar dependencias mucho más complejas más allá de lo que permiten los coeficientes ya mencionados, aparecen las funciones cópula de la mano del matemático Abe Sklar en 1959. Esta no es la única manera de tratar con este tipo dependencias; muy recientemente, en el año 2005, el matemático Gábor J. Székely introdujo el coeficiente de correlación de distancias como medida de dependencia entre vectores con dimensiones superiores.

A pesar de que a nivel univariado es de gran importancia el estudio del comportamiento de las variables, estudiar las relaciones existentes entre ellas nos puede ayudar a hacer predicciones o definir patrones de comportamiento. Nuestro objetivo principal en este trabajo es ayudar al lector a comprender las distintas maneras que existen para medir dichas relaciones, darles una interpretación y entender que los resultados obtenidos para cada medida no son excluyentes entre sí sino que aportan información distinta e incluso complementaria sobre cualquier estudio.

Capítulo 1

Preliminares

Este capítulo será un breve repaso de algunos conceptos básicos de estadística que resultarán de interés en lo que resta de trabajo. Trataremos de aproximarnos a la noción de independencia desde la perspectiva probabilística para entender la idea de correlación en los sucesivos capítulos. Además, proporcionaremos algunas definiciones de conceptos que serán recurrentes a lo largo del trabajo.

Tanto los conceptos presentados a continuación como propiedades derivadas u otras nociones y demostraciones que omitiremos por ser menos relevantes para el propósito central de este proyecto se pueden encontrar en Wasserman, 2013, Bertsekas y Tsitsiklis, 2008 y DeGroot y Schervish, 2013.

La necesidad natural de encontrar asociaciones entre dos o más variables y de estudiar y generalizar sus comportamientos hace que surjan los vectores aleatorios. Por simplicidad, reduciremos los siguientes resultados al marco bidimensional, sin olvidar que se pueden generalizar para dimensiones mayores. Si se desea ver el comportamiento en este último caso, véase Corral, 2023 o Castañeda et al., 2012.

Definición 1.1 (Vector Aleatorio)**.** Sea $\Omega \neq \emptyset$ y sean X e Y variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, \mathrm{P})$. A la función $(X, Y) : \Omega \longrightarrow \mathbb{R}^2$ definida por

$$(X, Y)(w) := (X(w), Y(w))^{\mathsf{T}} = \begin{pmatrix} X(w) \\ Y(w) \end{pmatrix}, \quad \text{para cada } \omega \in \Omega,$$

se le llama **vector aleatorio bidimensional** y lo llamaremos simplemente vector aleatorio en adelante.

Por otra parte, para describir dos sucesos independientes, A y B, a través de la probabilidad, decimos que esto ocurre si y solo si $\mathrm{P}(A \cap B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$. En caso contrario, decimos que son dependientes. Extendamos el concepto de dependencia e independencia para las variables aleatorias.

La idea de independencia de dos variables aleatorias es que la observación de una de ellas no permite hacer predicciones de la otra. A continuación, presentamos la definición formal:

Definición 1.2 (Variables Aleatorias Independientes)**.** Sean X e Y dos variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, \mathrm{P})$. Si para cualquier par de conjuntos de Borel A y B de $\mathbb{R}$ tenemos que

$$\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A) \cdot \mathrm{P}(Y \in B),$$

entonces decimos que X e Y son **independientes**.

Como consecuencia de la *Definición 1.2* tenemos que para todo $x, y \in \mathbb{R}$,

$$F(x, y) = \mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x) \cdot \mathrm{P}(Y \leq y),$$

es decir,

$$F(x, y) = F_X(x) \cdot F_Y(y), \quad \text{para todo } \ x, y \in \mathbb{R}, \tag{1.1}$$

donde F denota la función de distribución conjunta del vector (X, Y) y F_X, F_Y las funciones de distribución marginales de X e Y, respectivamente.

Recíprocamente si se tiene la condición de la ecuación (1.1), entonces las variables serán independientes.

Equivalentemente, es posible determinar la independencia entre dos variables aleatorias a partir de su función de masa de probabilidad (en el caso discreto) o su función de densidad (en el caso continuo).

Corolario 1.3. *Sean X e Y dos variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, \mathrm{P})$. Se dice que X e Y son independientes para todo $x, y \in \mathbb{R}$, si y solo si:*

$$p(x, y) = p_X(x) \cdot p_Y(y),$$

donde p denota la función de masa de probabilidad conjunta del vector (X, Y) y p_X, p_Y las funciones de masa de probabilidad marginales de X e Y, respectivamente (caso discreto).

$$f(x, y) = f_X(x) \cdot f_Y(y),$$

donde f denota la función de densidad conjunta del vector (X, Y) y f_X, f_Y las funciones de densidad marginales de X e Y, respectivamente (caso continuo).

Si vamos un poco más allá, podemos extender la caracterización de independencia a la función característica:

Proposición 1.4. *Las componentes del vector aleatorio (X, Y) son independientes si y solo si*

$$\varphi_{X,Y}(t_1, t_2) = \varphi_X(t_1) \cdot \varphi_Y(t_2)\,, \quad \textit{para todo } \ t_1, t_2 \in \mathbb{R},$$

donde $\varphi_{X,Y}$ denota la función característica conjunta del vector (X, Y) y φ_X, φ_Y las funciones características de X e Y, respectivamente.

Otro concepto muy importante en estadística que describe relaciones entre variables es el de **variables aleatorias independientes e idénticamente distribuidas** (i.i.d.):

Definición 1.5 (Variables Aleatorias Independientes e Idénticamente Distribuidas)**.** Sean X e Y variables aleatorias que toman valores en $I \subseteq \mathbb{R}$. Sean F_X y F_Y las funciones de distribución de X e Y, respectivamente, y $F(x, y)$ su función de distribución conjunta. Se dice que X e Y son variables aleatorias **i.i.d.** si y solo si:

- X e Y cumplen la ecuación (1.1) $\forall\, x, y \in I$. *(Independientes)*
- X e Y cumplen que $F_X(x) = F_Y(x)\ \ \forall\, x \in I$. *(Idénticamente distribuidas)*

Observación 1.6. La *Definición 1.5* se puede extender de forma natural a n variables.

Lo más común en la práctica es encontrarse con fenómenos dependientes y posteriormente estudiar el tipo de dependencia que los relaciona. Entre las posibilidades que existen, la dependencia más simple que se puede dar entre dos variables aleatorias es la lineal. Algunos de los conceptos más importantes relacionados con este tipo de dependencia son la covarianza y el coeficiente de correlación de Pearson que introduciremos en el Capítulo 3, así como otros coeficientes que son más apropiados cuando las variables se desvían de la normalidad o se desconoce su distribución.

No debemos perder de vista que en cualquier análisis estadístico tanto los resultados como las conclusiones obtenidas deben ser representativas de la población sujeta al estudio. Es por esto que la elección de las variables y el conjunto de datos escogido juegan un papel crucial en el análisis de la relación de dependencia entre variables.

Definición 1.7. Se llaman **parámetros poblacionales** a las medidas descriptivas de toda una población. Son cantidades que se obtienen a partir de las observaciones de las variables y de sus probabilidades y determinan perfectamente la distribución de estas. Son constantes fijas y generalmente se representan mediante letras griegas.

Definición 1.8. Dada una variable aleatoria X con función de distribución F, una **muestra aleatoria simple** de X (m.a.s.) de tamaño n es un conjunto finito de n variables independientes $X_1, \ldots, X_n$ con la misma distribución de probabilidad o, equivalentemente, cumpliendo la *Definición 1.5* y, a su vez, con la misma distribución que X.

Definición 1.9. Se llaman **estadísticos muestrales** a los valores individuales más probables de los parámetros poblacionales en la muestra. Al contrario que estos últimos, los estadísticos no son constantes ya que dependen de la estructura de la muestra.

En la práctica, generalmente son utilizados como **estimadores puntuales** debido al desconocimiento del verdadero parámetro. Los conceptos que se presentan a continuación con relación a las propiedades de los estimadores han sido tomados de Espartero, 2012. Nos interesa que dichos estimadores sean lo más precisos y confiables posible y, para ello, una condición importante es que no tengan sesgo.

Definición 1.10. Decimos que un estimador puntual $\hat{\theta}$ es **insesgado** cuando el promedio de sus valores para todas las posibles muestras coincide con el valor del parámetro poblacional θ, es decir,

$$\text{Sesgo}(\hat{\theta}) = \text{E}\left[\hat{\theta}\right] - \theta = 0.$$

En caso contrario, se dice que el estimador $\hat{\theta}$ es sesgado.

Definición 1.11. Si un estimador $\hat{\theta}$ es sesgado, pero el sesgo tiende a cero cuando el número de observaciones n tiende a infinito, se dice que es **asintóticamente insesgado**.

$$\lim_{n\to\infty} \text{Sesgo}(\hat{\theta}) = 0.$$

Otra propiedad interesante que conviene que tenga un estimador es la consistencia:

Definición 1.12. Se dice que un estimador $\hat{\theta}$ es **consistente** si converge en probabilidad al parámetro estimado θ, es decir, para todo $\varepsilon > 0$,

$$\lim_{n\to\infty} \text{P}(|\hat{\theta} - \theta| < \varepsilon) = 1.$$

En la práctica, para comprobar si un estimador $\hat{\theta}$ es consistente, es suficiente con verificar que se cumplen las siguientes propiedades simultáneamente:

(1) $\hat{\theta}$ es insesgado o asintóticamente insesgado cumpliendo las *Definiciones 1.10* o *1.11*, respectivamente.

(2) La varianza de $\hat{\theta}$ tiende a cero cuando el número de observaciones n tiende a infinito, es decir,

$$\lim_{n\to\infty} \text{Var}(\hat{\theta}) = 0.$$

Este breve repaso a través de los conceptos anteriores será de gran utilidad en capítulos posteriores cuando introduzcamos el coeficiente de correlación y sus diversas propiedades tanto para una población como para una muestra.

Capítulo 2

Medidas de Dependencia

Una medida de dependencia indica, normalmente a través de un valor numérico, cómo de relacionadas están dos variables aleatorias. La dependencia engloba desde casos tan simples en los que las variables se asocian de manera completamente lineal hasta otros en los que encontramos una clara independencia mutua. El análisis de datos bivariado permite cuantificar estas relaciones a través del nivel de covarianza entre las variables. Este análisis estadístico facilita la construcción de una gran variedad de coeficientes que, mediante valores estimados, muestran la magnitud y el sentido de la dependencia.

En este contexto, el tipo de estadísticos que se pueden utilizar para garantizar la pertinencia y validez de los resultados está condicionado generalmente por el nivel de medición de las variables involucradas (véase Stevens, 1946), entendiendo por esto último, la forma en que esa variable se clasifica según su naturaleza matemática. Existen cuatro maneras de agrupar las variables según esto: escala nominal, ordinal, de intervalo y de razón, y es habitual que nos refiramos a estas dos últimas como escalas continuas (véase Ochoa y Molina, 2018 para más detalles).

En el caso de variables con escalas continuas o tamaño muestral considerable ($n > 30$), entre otras cosas, se utilizan estadísticos de tipo **paramétrico**. Por el contrario, los de tipo **no paramétrico** servirán cuando las variables tengan un nivel de medición de intervalo, ordinal o nominal. Más adelante, cuando introduzcamos los coeficientes de correlación, se hará notoria esta distinción.

A continuación, describiremos los conceptos más notables de dependencia total[1], tomados principalmente de Balakrishnan y Lai, 2009, e introduciremos las primeras nociones que dieron algunos autores para establecer medidas globales de dependencia convenientes. Ya en el Capítulo 3, se presentarán los coeficientes de correlación más comunes y se verificará si satisfacen los criterios mencionados en el presente capítulo.

[1] Supondremos que todas las funciones a las que se haga mención en la Sección 2.1 son Borel-medibles y sobreyectivas.

2.1. Tipos de Dependencia Total

Dependencia Mutua Completa

Si dos variables X e Y pueden predecirse a partir de la otra, intuitivamente, X se puede expresar en función de Y y viceversa, es decir, X e Y son dependientes entre sí. La siguiente definición será de gran utilidad para describir este aspecto de manera más formal.

Definición 2.1. Una variable aleatoria Y es *completamente dependiente* de X si existe una función b tal que

$$\mathrm{P}(Y = b(X)) = 1. \tag{2.1}$$

Definición 2.2 (Dependencia Mutua Completa)**.** X e Y son **completamente dependientes entre sí** si se cumple la ecuación (2.1) para alguna función b inyectiva.

Este concepto es la antítesis de la independencia estocástica. Como hemos visto, la dependencia mutua completa permite la predicción absoluta de cualquiera de las variables con respecto a la otra, mientras que la independencia estocástica conlleva que las variables no sirvan para predecirse mutuamente como vimos en el Capítulo 1.

Dependencia Monótona

Desde un punto de vista más analítico, algunos autores mostraron que la dependencia mutua completa no siempre es el opuesto perfecto de la independencia, motivando así la creación de un nuevo concepto de dependencia total denominado dependencia monótona.

Definición 2.3 (Dependencia Monótona)**.** Sean X e Y variables aleatorias continuas. Entonces Y es **monótonamente dependiente** de X si existe una función b estrictamente monótona tal que

$$\mathrm{P}(Y = b(X)) = 1.$$

Observación 2.4. Nótese que una función b inyectiva no tiene por qué ser monótona, lo que hace que la dependencia monótona sea más fuerte que la dependencia mutua.

A partir de la *Definición 2.3* se deduce lo siguiente:

Proposición 2.5. *Y es monótonamente dependiente de X si y solo si X es monótonamente dependiente de Y.*

Observación 2.6. Si la función b de la *Definición 2.3* es *creciente* se dice que X e Y son *directamente dependientes*; por el contrario, si b es *decreciente* se dice que X e Y son *inversamente dependientes*.

Dependencia Funcional e Implícita

Por último, presentaremos dos conceptos más débiles relacionados con la dependencia.

Definición 2.7 (Dependencia Funcional)**.** X e Y son **funcionalmente dependientes** si $X = a(Y)$ o $Y = b(X)$ para alguna función a y b. Esto es, X e Y son funcionalmente dependientes si alguna de las dos verifica la *Definición 2.1*.

Definición 2.8 (Dependencia Implícita)**.** X e Y son **implícitamente dependientes** si existen dos funciones a y b tales que $a(X) = b(Y)$, con $\mathrm{Var}(a(X)) > 0$. En otras palabras, puede no existir una función que conecte directamente X e Y y, sin embargo, estén relacionadas.

Todas estas nociones de dependencia total se pueden ordenar en función de su fuerza. Si lo hacemos desde la dependencia más fuerte a la más débil, estas quedarían organizadas de la siguiente manera:

$$\boxed{\text{Lineal} \gg \text{Monótona} \gg \text{Mutua Completa} \gg \text{Funcional} \gg \text{Implícita}}$$

2.2. Medidas Globales de Dependencia

La realidad es que a la hora de realizar cualquier estudio entre dos variables aleatorias no es tan sencillo encontrar una dependencia total entre ellas. En estos casos, suele ser de gran utilidad la búsqueda de cantidades que sean capaces de medir la fuerza o el grado de dependencia entre las mismas. Si podemos expresar tal cantidad mediante un escalar, solemos referirnos a ella como *índice* (también llamada *medida global* en Samuel et al., 2001).

En esta sección, presentaremos siete criterios propuestos por Gibbons y Chakraborti, 2003 (p. 401) que reflejan las propiedades más deseables que debe cumplir un índice para resultar de utilidad. Otros autores como Rényi, 1959 o Lancaster, 2004 propusieron versiones similares de estos axiomas que se pueden encontrar en Balakrishnan y Lai, 2009 (pp. $144 - 145$).[2]

Proposición 2.9. *Supongamos que una medida relativa de asociación adecuada es aquella que satisface los siguientes criterios:*

[2]La razón de haber elegido los criterios propuestos por Gibbons y Chakraborti, 2003 no es otra que el hecho de que se ajustan mejor que los propuestos por otros autores a los contenidos que trataremos en capítulos posteriores.

1. *Para cualesquiera dos pares independientes (X_i, Y_i) y (X_j, Y_j) de variables aleatorias que sigan una distribución bivariante, la medida será igual a $+1$ si la relación es directa y perfecta en el sentido de que*

   $$X_i < X_j \text{ siempre que } Y_i < Y_j \quad o \quad X_i > X_j \text{ siempre que } Y_i > Y_j.$$

   Esta relación se denominará concordancia (acuerdo) perfecta (perfecto).

2. *Bajo las mismas hipótesis del criterio 1, la medida será igual a -1 si la relación es inversa y perfecta en el sentido de que*

   $$X_i < X_j \text{ siempre que } Y_i > Y_j \quad o \quad X_i > X_j \text{ siempre que } Y_i < Y_j.$$

   Esta relación se denominará discordancia (desacuerdo) perfecta (perfecto).

3. *Si ni el criterio 1 ni el 2 se cumplen para todos los pares, la medida se situará entre los dos extremos -1 y $+1$. Es deseable que, en algún sentido, grados crecientes de concordancia se reflejen mediante valores positivos crecientes y grados crecientes de discordancia se reflejen mediante valores negativos crecientes.*

4. *La medida será igual a cero si X e Y son independientes.*

5. *La medida para (X, Y) es la misma que para (Y, X), $(-X, -Y)$ o $(-Y, -X)$, es decir, es simétrica.*

6. *La medida para $(-X, Y)$ o $(X, -Y)$ será el negativo de la medida para (X, Y).*

7. *La medida será invariante bajo todas las transformaciones de X e Y para las que se conserva el orden de magnitud.*

El punto principal de los axiomas expuestos en la *Proposición 2.9* es dar una idea aproximada sobre lo que entendemos por dependencia y lo que se requiere sobre la misma y proporcionar un criterio para estudiar las propiedades de las distintas medidas.

Las tres medidas globales más destacadas para estudiar la dependencia son:

- El coeficiente de correlación de Pearson (ρ).
- El coeficiente de correlación de Spearman (ρ_s).
- La tau de Kendall (τ).

Dedicaremos los siguientes capítulos a estudiar en profundidad estas tres medidas, así como otras tan novedosas como el coeficiente de correlación de distancias o la relación que tiene la correlación con las funciones cópula.

Capítulo 3

El coeficiente de correlación

3.1. Introducción al coeficiente de correlación

La noción de correlación se desarrolló a partir de la idea de que las variables pueden estar relacionadas entre sí de algún modo, es decir, cambios en una variable pueden estar asociados a cambios en otra. Este concepto se construyó sobre las bases establecidas por matemáticos como Pierre-Simon Laplace o Carl Friedrich Gauss como se explica en Pearson, 1920.

Gauss desempeñó un papel fundamental en el desarrollo de la teoría de la correlación, aunque el enfoque principal de sus investigaciones no se centró específicamente en ella. Pearson, 1920 señala que, para comprender el origen de esta teoría, es esencial revisar las investigaciones de Gauss acerca de los métodos estadísticos que desarrolló para modelar la variabilidad y la incertidumbre en los datos observados.

Él observó que los errores en las mediciones se distribuían según una curva normal y desarrolló una técnica que permitiera encontrar la mejor aproximación lineal a un conjunto de datos de manera que se minimizaran esos errores: *el método de mínimos cuadrados*.[1]

Las contribuciones significativas en este campo surgieron más adelante de la mano de autores como August Bravais, Francis Edgeworth y, mayoritariamente, Francis Galton. En 1888, Galton introdujo el término *"correlación"* y, gracias a sus aportaciones, abrió las puertas para futuras investigaciones a otros matemáticos como Karl Pearson o Charles Spearman.

Pearson fue el primero en tomar el testigo de Galton e introdujo, en el año 1895, el *coeficiente de correlación*. Su aplicación es tan extensa y variada que, aunque pueda parecer algo trivial, existen numerosos motivos por los que resulta muy fácil hacer un mal uso del mismo, ya sea por imprecisiones al verificar

[1] El método de mínimos cuadrados proporciona una estimación insesgada de la recta de regresión que será útil a la hora de interpretar el coeficiente de correlación.

las hipótesis que se deben cumplir o bien a la hora de interpretar los resultados. A pesar de esto, Pearson justifica el uso de este coeficiente en la sencillez de su cálculo y su capacidad de predecir la dependencia lineal entre sucesos.

En el análisis de variables aleatorias con distribución de probabilidad bivariante, determinar la existencia de asociación entre ellas se cuantifica estadísticamente mediante la covarianza: valor que refleja en qué cuantía varían estas de forma conjunta respecto a sus medias por lo que, esencialmente, la covarianza mide la relación lineal entre dos variables.

A lo largo de todo el capítulo supondremos que tanto las medias como las varianzas de las variables aleatorias X e Y existen y son finitas para poder garantizar que los resultados presentados a continuación puedan ser calculados.

Además de la definición clásica de covarianza, existe una forma alternativa para calcularla a partir de las distribuciones cuya expresión está íntimamente ligada al concepto de dependencia (véase Hernández y Caita, 2017 (p. 285)).

Teorema 3.1 (Identidad de Höffding)**.** *Sea (X, Y) un vector aleatorio con función de distribución conjunta F y funciones de distribución marginales F_X y F_Y, respectivamente. La identidad de Höffding establece que la covarianza entre X e Y viene dada por*

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (F(x, y) - F_X(x) F_Y(y))\, dxdy. \tag{3.1}$$

A continuación, presentaremos un importante resultado que relaciona directamente los conceptos de *covarianza* y *dependencia*:

Proposición 3.2. *Sean X e Y dos variables aleatorias definidas sobre el mismo espacio de probabilidad. Si X e Y son independientes, entonces* $\mathrm{Cov}(X, Y) = 0$.

Demostración. Gracias a las propiedades del operador $\mathrm{E}[\ ]$, si X e Y son independientes se tendrá que

$$\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]. \tag{3.2}$$

Empleando la ecuación (3.2) se llega a que

$$\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = \mathrm{E}[XY] - \mathrm{E}[XY] = 0.$$

□

Observación 3.3. Recordemos que si X e Y son independientes se verifica la ecuación (1.1) del Capítulo 1 para las distribuciones. En este caso, la identidad de Höffding expuesta en la ecuación (3.1) muestra que $\mathrm{Cov}(X, Y) = 0$, lo que es consistente con el enunciado de la *Proposición 3.2*.

Observación 3.4. El recíproco de la *Proposición 3.2* no se tiene en general.

Por otro lado, la *concordancia* a la que se hace referencia en la *Proposición 2.9* del Capítulo 2 también se manifiesta de una manera similar en la covarianza: si valores grandes (pequeños) de X están asociados a valores grandes (pequeños) de Y, la covarianza será grande y positiva. Por otro lado si la correspondencia es inversa, es decir, valores grandes (pequeños) de X están asociados a valores pequeños (grandes) de Y, entonces la covarianza será grande pero negativa.

A pesar de que la covarianza no es una medida de asociación relativa[2], algunos de los criterios expuestos en la *Proposición 2.9* se asemejan, en gran parte, a las propiedades que tiene la misma. Este hecho nos lleva a pensar que vamos por buen camino a la hora de encontrar una medida con las características adecuadas.

3.2. El Coeficiente de Correlación producto-momento de Pearson

Normalmente cuando consideramos dos variables para estudiar sus relaciones existe la posibilidad de que estas estén expresadas en distintas unidades dando lugar en numerosos casos a un resultado de la covarianza no interpretable. Este hecho originó el conocido coeficiente de correlación de Pearson como medida de asociación relativa (libre de escala) que no es más que una estandarización de la covarianza.

Definición 3.5 (Coeficiente de Correlación)**.** Sean X e Y dos variables aleatorias con medias y varianzas finitas. La versión poblacional del **coeficiente de correlación de Pearson** entre X e Y, que denotaremos por ρ, viene dada por

$$\rho(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}\,, \tag{3.3}$$

donde σ_X y σ_Y denotan las desviaciones estándar de las variables X e Y, respectivamente.

Observación 3.6. Al definir ρ como una medida de dependencia relativa, su característica más distintiva es la *adimensionalidad*.

La validez de las conclusiones que se pueden extraer a partir del cálculo del coeficiente de correlación está condicionada por el cumplimiento de siete premisas. Sin ellas, ρ puede proporcionarnos información errónea acerca de la dependencia de las variables del estudio. La información expuesta a continuación sobre estos supuestos ha sido tomada de Lalinde et al., 2018.

[2]La covarianza puede tomar valores entre $-\infty$ y $+\infty$.

Proposición 3.7. *Sean X e Y dos variables aleatorias definidas sobre el mismo espacio de probabilidad y ρ el coeficiente de correlación de Pearson. Para garantizar que ρ sea un coeficiente válido ha de cumplir estos siete supuestos:*

1. ***Nivel de medición de las variables:*** *ρ solo admite variables con escalas continuas, a las que ya se hizo referencia en el Capítulo 2, aunque no es necesario que ambas tengan el mismo nivel.*
2. ***Datos pareados:*** *para poder realizar el cálculo de ρ es necesario que existan datos emparejados en ambas variables, por lo que si hay ausencia de valores en alguna de ellas, sus registros deberán ser descartados del análisis.*
3. ***Distribución normal bivariante:*** *la validez de ρ se sustenta principalmente en que la distribución conjunta de X e Y sea una distribución normal.*
4. ***Ausencia de datos atípicos a nivel bivariado:*** *los datos atípicos, también llamados outliers, suelen afectar considerablemente a ρ puesto que este no es capaz de detectarlos.*
5. ***Linealidad:*** *ρ mide la fuerza y la dirección de la relación lineal existente entre dos variables X e Y. Una buena forma de verificar la linealidad es a través de los diagramas de dispersión; basarse exclusivamente en ρ puede no aportar la información adecuada, especialmente cuando hay outliers, relaciones no lineales o presencia de grupos en los datos.*
6. ***Independencia de observaciones:*** *Cada individuo debe aparecer una única vez y los valores obtenidos para un sujeto en una variable no deben estar relacionados con el resto de valores de esa misma variable. Por desgracia, no existe una forma eficaz de comprobar este hecho por lo que se recomienda que la elección de las observaciones sea completamente al azar.*
7. ***Condiciones del muestreo:*** *como consecuencia del supuesto 6, para garantizar unos buenos resultados, la muestra del vector (X, Y), dada por $(X_i, Y_i),\ i = 1, \ldots, n$, debe constituir un conjunto de variables aleatorias i.i.d., es decir, se debe seleccionar una muestra acorde a las características del muestreo aleatorio simple.*

Observación 3.8. En referencia al supuesto 1, cabe destacar que existen casos especiales en los que el nivel de medición de las variables puede variar entre los cuatro mencionados en el Capítulo 2. En ese caso, se utilizan otro tipo de coeficientes como el coeficiente φ o el biserial-puntual, entre otros. En la Sección 1.7 de Sulbarán, 2012 se puede consultar una breve descripción de cada uno de ellos si el lector lo desea. En secciones posteriores, estudiaremos dos alternativas que sí resultan de interés para este trabajo: el coeficiente de correlación de Spearman y la tau de Kendall.

Teorema 3.9 (Propiedades del Coeficiente de Correlación de Pearson)**.** *El coeficiente de correlación de Pearson verifica los seis primeros criterios expuestos en la Proposición 2.9.*

Observación 3.10. Las relaciones a las que se hace referencia en las propiedades *1* y *2* de la *Proposición 2.9* serán, en este caso, de dependencia lineal ya que ρ no es útil para detectar otro tipo de dependencia (ver supuesto 5 de la *Proposición 3.7*).

La demostración del *Teorema 3.9* es trivial (se puede consultar en el Anexo I).

Una medida idónea debería cumplir las siete propiedades expuestas en la *Proposición 2.9*. Cabe entonces preguntarse qué ocurre con la propiedad 7 de esta proposición para el coeficiente de correlación de Pearson. Desafortunadamente, ρ no es invariante bajo todas las transformaciones monótonas:

Proposición 3.11. *Sean X e Y variables aleatorias definidas sobre el mismo espacio de probabilidad y ρ el coeficiente de correlación de Pearson. Se tiene que ρ es invariante (salvo signo) únicamente en el caso de que a las variables se les apliquen transformaciones lineales (cambios de escala y localización).*

Demostración. Sean $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ dos funciones lineales tales que $f(X) = aX+b$ y $g(Y) = cY+d$, $a, c \in \mathbb{R}\backslash\{0\}$ y $c, d \in \mathbb{R}$. Entonces, por las propiedades de la varianza y de la covarianza, se tiene que

$$\rho(f(X), g(Y)) = \rho(aX + b, cY + d) = \frac{\mathrm{Cov}(aX + b, cY + d)}{\sqrt{\mathrm{Var}(aX + b)\,\mathrm{Var}(cY + d)}} =$$
$$= \frac{ac\,\mathrm{Cov}(X, Y)}{\sqrt{a^2c^2\,\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{ac\,\mathrm{Cov}(X, Y)}{|ac|\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \pm\rho(X, Y).$$

□

Observación 3.12. Exceptuando el caso anterior, dada una función monótona $T : \mathbb{R} \longrightarrow \mathbb{R}$ arbitraria[3] se tiene que $\rho(T(X), T(Y)) \neq \rho(X, Y)$ para dos variables aleatorias X e Y.

Definición 3.13. Se dice que dos variables aleatorias no constantes X e Y están **incorrelacionadas** si ρ o, equivalentemente la covarianza, toma el valor cero.

Recordemos que la propiedad 4 de la *Proposición 2.9*, tomando como ρ la medida a la que se hace referencia, nos decía que si dos variables son independientes siempre están incorrelacionadas, en virtud de la *Definición 3.13*. Sin embargo, que dos variables estén incorrelacionadas no significa que esa relación no exista, sino que simplemente no es lineal. Por eso, el recíproco de la propiedad 4 no se tiene en general.

Decimos "en general" ya que existe un caso en el que *incorrelación* e *independencia* sí son términos equivalentes: en condiciones de una distribución normal bivariante. Anteriormente, se hizo referencia,

[3] Recordemos que una función $T : \mathbb{R} \longrightarrow \mathbb{R}$ es monótona si conserva el orden, es decir, si y solo si

$$x \leq y \Rightarrow T(x) \leq T(y) \text{ (creciente), o bien, } x \leq y \Rightarrow T(x) \geq T(y) \text{ (decreciente).}$$

en el supuesto 3 de la *Proposición 3.7*, a la importancia de esta para el uso adecuado de ρ y es que este solamente tiene un comportamiento deseado cuando tratamos con este tipo de distribuciones.

A continuación, introduciremos de manera breve algunos conceptos relacionados con la distribución normal bivariante junto con sus principales características y sus distribuciones asociadas ya que es una herramienta fundamental en estadística debido a su capacidad para modelar dependencias lineales entre dos variables continuas. En particular, nos interesa porque proporciona una base teórica para calcular y entender la covarianza y el coeficiente de correlación de Pearson.

La **distribución normal bivariante** cuenta con propiedades matemáticas bien definidas y sencillas que la convierten en una distribución muy adecuada para realizar análisis de datos ya que las técnicas de estimación y pruebas de hipótesis son más precisas cuando se asume normalidad. La información relativa a este apartado se puede encontrar en Hernández y Caita, 2017 (pp. $286 - 287$) o en la Sección 10 del Capítulo 5 de DeGroot y Schervish, 2013.

Definición 3.14 (Distribución Normal Bivariante)**.** Se dice que el vector aleatorio (X, Y) sigue una **distribución normal bivariante** o *gaussiana* con vector de medias $\mu = (\mu_X, \mu_Y)^t \in \mathbb{R}^2$ y matriz de varianzas-covarianzas

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix} \in \mathcal{M}_{2\times 2} \text{ simétrica y definida positiva,} \tag{3.4}$$

o lo que es lo mismo, $(X, Y) \sim N_2(\mu, \Sigma)$, si su función de densidad conjunta ϕ viene dada por la expresión

$$\phi(x, y) := \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x - \mu_X,\ y - \mu_Y)\Sigma^{-1}\begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}\right\},\ x, y \in \mathbb{R}, \tag{3.5}$$

donde $|\Sigma|$ denota el determinante de la matriz Σ. En estas circunstancias, el vector μ representa la esperanza del vector (X, Y) y Σ la matriz de varianzas-covarianzas de cada una de sus componentes.

Únicamente en el caso bivariante, tomando como ρ el coeficiente de correlación de Pearson, se tiene que $\sigma_{XY} = \rho\sigma_X\sigma_Y = \rho\sigma_Y\sigma_X = \sigma_{YX}$ en virtud de la ecuación (3.3), de tal manera que la matriz dada por la ecuación (3.4) se puede reescribir como

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_Y\sigma_X & \sigma_Y^2 \end{pmatrix}.$$

Por tanto, para todo $x, y \in \mathbb{R}$, se tendrá que la función de densidad dada en la ecuación (3.5) será ahora

$$\phi(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2}Q\right\}, \tag{3.6}$$

donde $Q = \left\{ \frac{1}{(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) \right] \right\}$.

Proposición 3.15. *Sea (X, Y) un vector aleatorio tal que $(X, Y) \sim N_2(\mu, \Sigma)$. La distribución normal bivariante cumple que dados $a, b \in \mathbb{R}$, la variable aleatoria $aX + bY$ también sigue una distribución normal, es decir,*

$$aX + bY \sim N\left(a\mu_X + b\mu_Y,\ \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y} \right).$$

En particular, X e Y siguen una distribución normal.

Proposición 3.16. *Las marginales de una distribución normal bivariante son normales unidimensionales tales que*

$$X \sim N(\mu_X, \sigma_X) \quad e \quad Y \sim N(\mu_Y, \sigma_Y),$$

es decir,

$$\phi_X(x) = \int_{-\infty}^{\infty} \phi(x, y)dy = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu_X}{\sigma_X} \right)^2 \right\}$$

y

$$\phi_Y(y) = \int_{-\infty}^{\infty} \phi(x, y)dx = \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left\{ -\frac{1}{2} \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right\}.$$

Observación 3.17. Los detalles referentes al cálculo de las integrales se pueden encontrar en Ortega, 2014 (pp. $109 - 110$).

Teorema 3.18. *Sea (X, Y) un vector aleatorio que sigue una distribución normal con vector de medias μ y matriz de varianzas-covarianzas Σ. X e Y son variables aleatorias independientes si y solo si las variables están incorrelacionadas ($\rho = 0$), o lo que es lo mismo, si la matriz Σ es una matriz diagonal tal que $\Sigma_{12} = \Sigma_{21} = 0$, es decir,*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}, \quad \Sigma_{11}, \Sigma_{22} \geq 0.$$

La demostración de este teorema se puede encontrar en el Anexo I.

Por lo general, el parámetro ρ es desconocido. En la práctica, podemos estimarlo a partir de una muestra aleatoria simple de n variables bivariantes $(X_1, Y_1), \ldots, (X_n, Y_n)$ como se hace a continuación.

Definición 3.19 (Coeficiente de Correlación Muestral)**.** Si X e Y son variables aleatorias que toman valores $\{(X_i, Y_i)\}_{i=1}^n$ de una muestra aleatoria simple de tamaño n, se puede estimar el coeficiente de correlación de Pearson entre X e Y, r, mediante

$$r(X,Y) = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 \left(Y_i - \bar{Y}\right)^2}}, \tag{3.7}$$

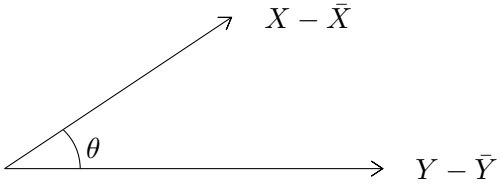
donde S_X y S_Y denotan las desviaciones estándar muestrales de las variables X e Y, respectivamente.

Observación 3.20. A pesar de las diferencias evidentes entre ρ y r debido a su naturaleza, los resultados y propiedades comentadas con anterioridad son válidos para ambos coeficientes.

Si nos fijamos en la definición de r en la ecuación (3.7), se puede ver fácilmente la idea geométrica que hay detrás sin más que considerar los "vectores centrados" $X - \bar{X}$ e $Y - \bar{Y}$. Utilizando entonces la relación del producto escalar para calcular el ángulo entre dichos vectores[4] se tiene que:

$$\cos\theta = \frac{(X - \bar{X}) \cdot (Y - \bar{Y})}{\|X - \bar{X}\| \cdot \|Y - \bar{Y}\|} = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}} = r(X,Y). \tag{3.8}$$

La ecuación (3.8) implica que r se puede interpretar como el coseno del ángulo que forman los vectores de las desviaciones de X e Y con respecto a sus medias.



Toda la información relativa a esta interpretación de r se puede encontrar en Ríus Díaz et al., 2012.

- $\theta = 0$ implicaría que ambos vectores son colineales con el mismo sentido, lo que se traduce en una relación perfecta y directa o, lo que es lo mismo, $r = 1$.
- $\theta = 90$ implicaría que ambos vectores son ortogonales y, por tanto, no existiría una dependencia lineal entre ellos, es decir, $r = 0$.
- $\theta = 180$ implicaría que ambos vectores son colineales con sentido inverso, lo que se traduce en una relación perfecta e inversa o, lo que es lo mismo, $r = -1$.

El coeficiente r es de gran importancia en los modelos de regresión lineal simple. Estos modelos son de la forma

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \text{ donde } \alpha, \beta \in \mathbb{R} \text{ y } \varepsilon_i \text{ representa el error aleatorio e } i \in \{1, \ldots, n\}. \tag{3.9}$$

[4] Sean $\mathbf{u}, \mathbf{v} \neq 0$ dos vectores. El producto escalar entre ellos es el producto de la magnitud de cada vector por el coseno del ángulo que forman, es decir,

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \cdot \cos\theta.$$

Sin embargo, en el contexto que nos interesa, nos centraremos en el modelo estimado a partir del dado en la ecuación (3.9), es decir, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, donde $\hat{\alpha}$ y $\hat{\beta}$ son las estimaciones de α y β, respectivamente, e $\hat{Y}_i$ los valores predichos de Y_i por el modelo, $\forall i \in \{1, \ldots, n\}$.

Es sencillo deducir que r toma el mismo signo que la pendiente estimada de la recta de regresión, $\hat{\beta}$. En efecto, utilizando el método de mínimos cuadrados que mencionábamos al inicio del capítulo, $\hat{\beta}$ resulta ser el cociente entre la covarianza de X e Y y la varianza de X, es decir,

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} = r\frac{S_Y}{S_X}.$$

No entraremos en detalle de cómo se obtiene esto último ya que no forma parte de los objetivos de nuestro estudio pero se puede consultar el desarrollo de dicho método en Camacho et al., 2006 (pp. $11 - 12$).

A partir de lo anterior podemos concluir que para $r > 0$ (dependencia lineal directa), el ajuste dará lugar a una recta de regresión creciente, para $r < 0$ (dependencia lineal inversa), a una recta de regresión decreciente y en el caso de que $r = 0$, la recta de regresión ajustada no tendrá pendiente (recta constante).

Lo importante de este modelo es que una de las medidas de bondad de ajuste más significativas a la hora de interpretar resultados está relacionada directamente con la correlación: el coeficiente de determinación, R^2 (véase Balakrishnan y Lai, 2009 (p. 148)).

Definición 3.21 (Coeficiente de Determinación)**.** Si X e Y son variables aleatorias que toman valores $\{(X_i, Y_i)\}_{i=1}^n$ bajo el modelo de regresión lineal simple descrito en la ecuación (3.9), el coeficiente de determinación, R^2, vendrá dado por

$$R^2 = \frac{\sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}, \tag{3.10}$$

donde $\hat{Y}_i$ es la predicción de Y_i calculada a partir de la ecuación de regresión estimada.

Este coeficiente determina la calidad del modelo para replicar los resultados y la proporción de la variabilidad total en la variable respuesta que puede ser explicada por la regresión lineal.

Observación 3.22. En regresión lineal simple, y solo en este caso, el coeficiente de determinación no es más que el cuadrado del coeficiente de correlación de Pearson por lo que se podrá reemplazar la expresión de la ecuación (3.10) por el cuadrado de la expresión de la ecuación (3.7), es decir,

$$R^2 = r^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2}.$$

Tanto r como R^2 son utilizados para evaluar las relaciones en un modelo de regresión lineal por lo que cabe preguntarse si el valor de r proporciona estimaciones precisas y fiables entre las variables X e Y o si solamente es consecuencia del azar, o lo que es lo mismo, si r es un buen estimador de ρ. Una buena manera de saberlo es verificar, entre otras cosas, si es insesgado y consistente.

Pues bien, en Shieh, 2010 podemos observar que tanto la esperanza como la varianza de r se pueden aproximar por:

$$\mathrm{E}[r] \doteq \rho - \frac{\rho(1-\rho^2)}{2(n-1)} \quad \text{y} \quad \mathrm{Var}(r) \doteq \frac{(1-\rho^2)^2}{n-1}.$$

Es claro que, por la *Definición 1.10*, $\mathrm{Sesgo}(r) = \mathrm{E}[r] - \rho \doteq -\dfrac{\rho(1-\rho^2)}{2(n-1)} \neq 0$ y, por tanto, r es un estimador sesgado de ρ. Sin embargo, r es un estimador **asintóticamente insesgado** ya que cumple que $\lim_{n\to\infty} \mathrm{Sesgo}(r) = 0$, es decir, r es un estimador más preciso cuanto mayor es la cantidad de datos en la muestra.

Ahora que hemos comprobado que r es un estimador asintóticamente insesgado, bastaría comprobar el límite de su varianza para comprobar su consistencia. En efecto,

$$\lim_{n\to\infty} \mathrm{Var}(r) \doteq \lim_{n\to\infty} \frac{(1-\rho^2)^2}{n-1} = 0,$$

y, por tanto, r es también un estimador **consistente**.

Observación 3.23. Cabe destacar que en muestras pequeñas con $\rho \neq 0$, r tiene un gran sesgo y una gran varianza. En estos casos, es recomendable el uso de otro tipo de coeficientes de correlación como los no paramétricos que introduciremos más adelante.

No solo el hecho de trabajar con muestras pequeñas supone una desventaja a la hora de interpretar la correlación. Confundir este término con el de *causalidad* es más habitual de lo que puede parecer. Sin embargo, hay veces que la correlación surge a partir de dos variables sin conexión lógica aparente que se ven afectadas por la presencia de un tercer factor no considerado denominado *factor de confusión*.

Como hemos visto a lo largo de todo el capítulo, existen muchos motivos por los que buscar otras alternativas al coeficiente de correlación de Pearson (no es apropiado con distribuciones que se desvían de la normalidad, no es invariante bajo todas las transformaciones monótonas, etc.). Para mitigar estas limitaciones suele ser habitual el uso de coeficientes basados en rangos o cuantiles.

3.3. El Coeficiente de Correlación de Rangos

La estadística *no paramétrica* se basa en métodos que no hacen suposiciones sobre la forma específica en la que están distribuidos los datos por lo que se puede decir que son métodos de "distribución libre".

Estos métodos son útiles ya que no requieren que los datos cumplan con ciertos criterios de distribución para obtener resultados válidos.

Normalmente cuando se usa la estadística no paramétrica es útil enfocarse en las magnitudes relativas (posición u orden de los datos en relación con otros datos) en lugar de basar nuestras conclusiones en magnitudes absolutas (valores exactos). Esto es así porque las magnitudes absolutas son más sensibles a la forma de la distribución, haciendo que medidas como la media o la desviación típica de un modelo puedan variar significativamente en función de cómo se distribuyen los datos, de la presencia de valores atípicos (outliers) o de transformaciones en los datos. En contraposición, las magnitudes relativas, al enfocarse únicamente en el orden de los datos, son más robustas y no se ven tan afectadas por estos factores de manera que son mucho más adecuadas a la hora de trabajar con estadística no paramétrica.

De hecho, como lo que nos interesa es encontrar una medida que satisfaga todos los criterios de la *Proposición 2.9*, nos basaremos precisamente en las relaciones de orden de las variables.

La *rho de Spearman* (ρ_s) y la *tau de Kendall* (τ) son los coeficientes de correlación de rangos más conocidos. Esencialmente, son medidas entre las clasificaciones, y no entre los valores reales, de dos variables X e Y.

3.3.1. La ρ de Spearman

El coeficiente de correlación de Spearman, también llamado coeficiente de correlación de rangos, recibe su nombre en honor al psicólogo inglés Charles Edward Spearman, quien introdujo en 1904 esta nueva medida de asociación entre variables con escalas al menos ordinales o que, a pesar de ser cuantitativas, no siguen un comportamiento normal.

Definición 3.24 (Coeficiente de Correlación de Spearman)**.** Sean X e Y dos variables aleatorias. La versión poblacional del **coeficiente ρ de Spearman** entre X e Y, que denotaremos por ρ_s, viene dada por

$$\rho_s(X, Y) = \rho(\text{rango}(X), \text{rango}(Y)),$$

donde ρ denota el coeficiente de correlación de Pearson.

Observación 3.25. Si consideramos que $\dfrac{1}{n} \cdot \text{rango}(X_i)$ converge a $F_X(X_i),\ i = 1, \ldots, n$ (análogamente para Y_i), entonces ρ_s puede expresarse como

$$\rho_s(X, Y) = \rho(F_X(X), F_Y(Y)),$$

donde F_X, F_Y denotan las funciones de distribución de X e Y, respectivamente (véase Yu y Hutson, 2024).

Observación 3.26. En general, cuando tratamos con el coeficiente de correlación de Spearman, se presta poca atención a su medida poblacional; lo habitual suele ser hablar de ρ_s en términos muestrales. Sin embargo, en alguna ocasión recurriremos a las expresiones anteriores ya que su estructura puede facilitarnos la comprensión de algunas de sus propiedades.

Ahora bien, los supuestos a los que se hacía referencia en la *Proposición 3.7* para garantizar que ρ fuese válido se han refinado para ρ_s, de manera que solo los apartados 2, 6 y 7 permanecerán inalterables. Con respecto al supuesto 1 sobre el nivel de medición de las variables, ahora ρ_s también admitirá variables con escalas ordinales ya que trabajará con la posición de los datos en la muestra (rangos). Este hecho hace que los supuestos 3, 4 y 5 ya no sean necesarios, puesto que al clasificar las observaciones a partir de sus rangos, ρ_s será más estable ante la forma de la distribución o los outliers y, de manera intuitiva, invariante por transformaciones que preservan el orden, lo que le permite ser capaz de captar dependencias monótonas no necesariamente lineales.

Antes de definir formalmente la versión muestral de la ρ de Spearman, vamos a dar una idea aproximada de su conceptualización. La información que sigue ha sido obtenida en su mayoría del Capítulo 11 de Gibbons y Chakraborti, 2003.

Sea $\{(X_i, Y_i)\}_{i=1}^n$ una muestra aleatoria de n pares de observaciones de las variables X e Y, respectivamente. Supongamos que dichas observaciones se clasifican independientemente de menor a mayor utilizando los números enteros $1, 2, \ldots, n$, esto es, a cada observación se le asigna una posición en función de su magnitud en relación con las demás de su grupo[5]. Los datos resultantes de hacer este procedimiento serán n conjuntos de rangos emparejados a partir de los que es posible calcular el coeficiente de correlación tal y como se define en la ecuación (3.7) de la *Definición 3.19*.

El estadístico resultante de hacer lo anterior se denomina *coeficiente de correlación de Spearman* (muestral) y lo denotaremos por r_s. No debemos perder de vista que r_s mide el grado de correspondencia entre las clasificaciones de las variables, en lugar de entre sus valores reales, para una población bivariante continua.

Para realizar los cálculos, consideraremos que no se producen empates en ninguna de las variables. El hecho de conocer el esquema de emparejamiento de las mismas, anteriormente mencionado, permite simplificar considerablemente la ecuación (3.7). Así, la expresión para r_s no será más que una adaptación de la fórmula para el cálculo de la r de Pearson para rangos.

Denotemos los respectivos rangos de las variables aleatorias en la muestra por

$$R_i = \text{rango}(X_i) \quad \text{y} \quad S_i = \text{rango}(Y_i), \qquad i = 1, \ldots, n.$$

[5]Bajo el supuesto de que las distribuciones marginales de X e Y sean continuas, teóricamente existirán conjuntos únicos de clasificaciones ya que la probabilidad de que haya dos o más observaciones con el mismo valor (empates) es casi inexistente.

Para simplificar la ecuación (3.7), veamos cada término en la expresión por separado. Como la muestra resultante de ordenar por rangos siempre estará formada por los enteros $\{1, 2, \ldots, n\}$[6] y, además, la suma es conmutativa, obtendremos valores constantes para cualquier muestra . Así,

$$\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \quad \text{y} \quad \bar{R} = \bar{S} = \frac{n+1}{2}. \tag{3.11}$$

$$\sum_{i=1}^{n} (R_i - \bar{R})^2 = \sum_{i=1}^{n} (S_i - \bar{S})^2 = \sum_{i=1}^{n} \left(i - \frac{n+1}{2} \right)^2 = \frac{n(n^2-1)}{12}. \tag{3.12}$$

Sustituyendo las variables aleatorias X e Y en la ecuación (3.7) por sus rangos asociados R y S, respectivamente y, en virtud de las ecuaciones (3.11) y (3.12), se tiene que:

$$r_s = r(R,S) = \frac{\sum_{i=1}^{n} \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right)}{\left[\sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2 \sum_{i=1}^{n} \left(S_i - \bar{S}\right)^2\right]^{1/2}} = \frac{\sum_{i=1}^{n} \left(R_i S_i - R_i \bar{S} - \bar{R} S_i + \bar{R}\bar{S}\right)}{\left[\left(n(n^2-1)/12\right)^2\right]^{1/2}} =$$

$$= \frac{12\left(\sum_{i=1}^{n} R_i S_i - 2 \cdot \frac{n(n+1)^2}{4} + \frac{n(n+1)^2}{4}\right)}{n(n^2-1)} = \frac{12\left(\sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4}\right)}{n(n^2-1)}.$$

Simplificando y agrupando los términos de la expresión anterior convenientemente, se llega finalmente a la siguiente igualdad para r_s:

$$r_s = \frac{12 \sum_{i=1}^{n} R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1}. \tag{3.13}$$

La expresión dada en la ecuación (3.13) no es la más habitual para r_s. Este suele presentarse comúnmente en términos de las diferencias entre los rangos D_i, donde $D_i = R_i - S_i = (R_i - \bar{R}) - (S_i - \bar{S})$. De este modo,

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (R_i - \bar{R})^2 + \sum_{i=1}^{n} (S_i - \bar{S})^2 - 2\sum_{i=1}^{n} \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right), \tag{3.14}$$

y, por tanto, el numerador de la expresión que estamos buscando para r_s es

$$\sum_{i=1}^{n} \left(R_i - \bar{R}\right)\left(S_i - \bar{S}\right) = \frac{n(n^2-1)}{12} - \frac{1}{2}\sum_{i=1}^{n} D_i^2.$$

Sustituyendo esto último en la expresión original llegaremos al resultado deseado. Todo lo anterior se puede sintetizar en la definición que se expone a continuación:

[6] Para una muestra aleatoria de tamaño n:

$$\text{mín}(R_i) = \text{mín}(S_i) = 1 \text{ y } \text{máx}(R_i) = \text{máx}(S_i) = n.$$

Definición 3.27 (Coeficiente de Correlación de Spearman Muestral)**.** Sean R y S los rangos de dos variables aleatorias X e Y que toman valores $\{(R_i, S_i)\}_{i=1}^n$ de una muestra aleatoria simple de tamaño n y $D_i = R_i - S_i,\ i = 1, \ldots, n$, las diferencias entre dichos rangos. Se puede definir el coeficiente de correlación de Spearman, r_s, entre R y S mediante

$$r_s = 1 - \frac{6\sum_{i=1}^n D_i^2}{n(n^2-1)}. \tag{3.15}$$

Observación 3.28. Sin pérdida de generalidad se puede asumir que los n pares de observaciones se etiquetan de acuerdo con la magnitud creciente de la componente X, es decir, $R_i = i$ para $i = 1, 2, \ldots, n$. S_i será el rango de la observación Y que está emparejada con el rango i de X y, en ese caso, $D_i = i - S_i$.

Teorema 3.29 (Propiedades del Coeficiente de Correlación de Spearman)**.** *El coeficiente de correlación de Spearman verifica todos los criterios expuestos en la Proposición 2.9 relativos a una buena medida de asociación.*

Observación 3.30. La demostración de las propiedades se hará en base a r_s ya que es la forma más habitual en la que se trabaja con este coeficiente. Sin embargo, no debemos olvidar que todas estas propiedades se tienen también para su versión poblacional, ρ_s.

Demostración. Sean R y S los respectivos rangos de dos variables aleatorias X e Y que toman valores $\{(i, S_i)\}$ y $D_i = i - S_i,\ i = 1, \ldots, n$, y sea r_s el coeficiente de correlación de Spearman.

1. Efectivamente sean (i, S_i) y (j, S_j) dos emparejamientos de rangos de variables aleatorias que siguen una distribución bivariante continua. Para que exista una concordancia perfecta entre rangos, la componente Y también debe ser creciente, o lo que es lo mismo, $S_i = i$ y, por tanto, $D_i = 0$ para $i = 1, 2, \ldots, n$, de manera que, atendiendo a la ecuación (3.15), $r_s = 1$.

2. Para que exista una discordancia perfecta entre rangos, la disposición de Y debe ser inversa a la disposición de X, es decir, Y debe ser decreciente, o equivalentemente, $S_i = n - i + 1$. Así,
$$\sum_{i=1}^n D_i^2 = \sum_{i=1}^n [i - (n - i + 1)]^2 = \sum_{i=1}^n \left[2\left(i - \frac{n+1}{2}\right)\right]^2 = 4\sum_{i=1}^n \left(i - \frac{n+1}{2}\right)^2 = \frac{n(n^2-1)}{3},$$
donde esta última igualdad deriva de la ecuación (3.12). Sustituyendo finalmente este resultado en la ecuación (3.15), se obtiene que $r_s = -1$.

3.∼6. Como la ecuación (3.15) es algebraicamente equivalente a la ecuación (3.7), es inmediato evidenciar que r_s estará limitado al intervalo $[-1, 1]$ para cualquier conjunto de pares numéricos. Además, al igual que se tenía para r, r_s será simétrico y la independencia de las variables implicará un coeficiente nulo, $r_s = 0$.

7. Dado que los rangos se conservan en todas las transformaciones que preservan el orden y r_s es una medida de asociación basada en rangos, este será invariante[7].

□

Al igual que ocurría para el coeficiente de correlación de Pearson, ρ, el recíproco de la propiedad 4 de la *Proposición 2.9* no se tiene en general, pues que dos variables tomen un coeficiente de correlación de Spearman nulo solamente implica que no existe una relación monótona entre ellas.

Sin embargo, una vez más el hecho de estar en condiciones de una distribución normal bivariante hace que la equivalencia entre $\rho_s = 0$ e independencia sea cierta. Basta con darse cuenta que ρ_s y ρ son algebraicamente equivalentes para obtener de manera inmediata este resultado que ya hemos probado en el *Teroema 3.18* de la sección anterior.

Además, bajo estas condiciones, existe una forma de relacionar ambos coeficientes:

Teorema 3.31. *Sea (X, Y) un vector aleatorio que sigue una distribución normal con vector de medias μ y matriz de varianzas-covarianzas Σ y sea ρ el coeficiente de correlación de Pearson. Entonces ρ_s se puede escribir en función de ρ como*

$$\rho_s = \frac{6}{\pi} \arcsin \frac{\rho}{2}. \tag{3.16}$$

Los detalles de la demostración del teorema anterior se pueden encontrar en RRL, s.f. y Weisstein, 2024.

El *Teorema 3.31* reafirma la idea anterior de que $\rho_s = 0$ también implica la independencia de las variables bajo la distribución normal bivariante. Esto es así porque en el caso de que $\rho = 0$, por la ecuación (3.16), $\rho_s = \frac{6}{\pi} \arcsin(0) = 0$ y, por tanto, por el *Teorema 3.18*, las variables serán independientes.

Por otro lado, tiene sentido preguntarse si r_s es un buen estimador de ρ_s. Para ello, veamos en primer lugar la forma de la esperanza y la varianza del estimador basándonos en la expresión de r_s dada en la ecuación (3.13). Así, si denotamos $s = \sum_{i=1}^{n} R_i S_i$ tendremos que

$$\mathrm{E}[r_s] = \frac{12}{n(n^2-1)} \mathrm{E}[s] - \frac{3(n+1)}{n-1} \quad \text{y} \quad \mathrm{Var}(r_s) = \frac{144}{n^2(n^2-1)^2} \mathrm{Var}(s).$$

El cálculo de estas expresiones sin plantear hipótesis adicionales puede llegar a resultar complicado y se escapa de los objetivos de este trabajo; sin embargo, se puede ver en Moran, 1948 que el cálculo de la

[7] Las transformaciones monótonas pueden afectar a las distancias entre las observaciones, pero en ningún caso alteran su orden. Estas funciones pueden hacer que acelere o disminuya el crecimiento de las variables cambiando su curvatura sin perder la monotonía (véase Pinilla y Rico, 2021).

esperanza de s suponiendo que las variables siguen una distribución normal es de la forma

$$\mathrm{E}[s] = \frac{n(n-1)}{2}\left[(n-2)\left(1-\frac{1}{\pi}\arccos\frac{\rho}{2}\right)+\left(1-\frac{1}{\pi}\arccos\rho\right)\right].$$

Empleando algunas identidades trigonométricas[8] y operando convenientemente se llega finalmente a que

$$\mathrm{E}[r_s] \doteq \frac{6}{\pi}\left(\frac{n-2}{n+1}\cdot\arcsin\frac{\rho}{2}+\frac{1}{n+1}\cdot\arcsin\rho\right).$$

De aquí se deduce que, claramente r_s no es un estimador insesgado de ρ_s puesto que $\mathrm{Sesgo}(r_s) = \mathrm{E}[r_s] - \rho_s \neq 0$. Sin embargo, este es un estimador **asintóticamente insesgado**, pues si hacemos el límite del sesgo cuando la muestra tiende a infinito se tiene que $\lim_{n\to\infty}\mathrm{Sesgo}(r_s) = \lim_{n\to\infty}\mathrm{E}[r_s] - \rho_s = \frac{6}{\pi}\arcsin\frac{\rho}{2} - \rho_s = 0$ por la ecuación (3.16) del *Teorema 3.31*.

Al igual que el coeficiente de correlación de Pearson, bajo las condiciones de la distribución normal bivariante, r_s resulta ser más preciso a medida que aumenta el tamaño de la muestra. Sin embargo, la principal ventaja de este coeficiente con respecto al de Pearson es que resulta también útil en casos en los que la muestra es pequeña.

Cuando una muestra es pequeña, es más probable que algunos de los supuestos que requiere la r de Pearson vistos en la *Proposición 3.7* como la normalidad o la linealidad no se lleguen a alcanzar. Es más, en esta situación, incluso un único valor atípico puede tener un gran impacto sobre él. No obstante, el coeficiente de correlación de Spearman es más robusto ante estas condiciones, lo que hace que, con muestras más reducidas, su uso sea más adecuado y proporcione un resultado más confiable acerca de la relación existente entre las variables.

Pinilla y Rico, 2021 destacan que r_s no se debe tomar como un reemplazo no paramétrico de r puesto que cada uno constituye un enfoque diferente del análisis (r estudia relaciones lineales y r_s monótonas), por lo que incluso, en algunos casos, podrían considerarse complementarios.

Por último, es interesante estudiar lo que ocurre con la expresión de r_s en el caso de que se produzcan empates. Aunque se podrían asignar al azar las posiciones de las observaciones empatadas y no cambiaría nada con respecto a lo explicado anteriormente, este enfoque introduce un factor de aleatoriedad poco conveniente.

[8] Se utilizarán las siguientes identidades:

$\pi - \arccos x = \arccos(-x)$.

$\arccos x = \frac{\pi}{2} - \arcsin x$.

$\arcsin(-x) = -\arcsin x$.

En estos casos, lo más habitual es asignar rangos iguales a los empates haciendo la media de las posiciones que les corresponderían a esas observaciones si no hubiera dichos empates. De esta forma, se consigue no modificar la suma de los rangos de cada variable dada en la ecuación (3.11). No obstante, la suma de cuadrados de la ecuación (3.12) ya no será la misma en presencia de empates y, por tanto, r_s tampoco lo será.

Aunque r_s se podría calcular directamente con el procedimiento habitual a partir de los nuevos rangos asignados a los empates, es posible encontrar una expresión análoga a la de la ecuación (3.15) en función de las diferencias D_i para utilizarla en presencia de dichos empates.

Teorema 3.32. *En presencia de empates, el coeficiente de correlación de Spearman, r_s, se define como*

$$r_s = \frac{n(n^2-1) - 6\sum_{i=1}^{n} D_i^2 - 6(u+v)}{\left[(n(n^2-1) - 12u)\,(n(n^2-1) - 12v)\right]^{1/2}},$$

donde u y v representan las correcciones por empates en las variables X e Y, respectivamente.

La demostración de este teorema se puede consultar en el Anexo I.

Observación 3.33. Es claro que, en el caso de no haber observaciones empatadas en ninguna de las variables ($u = 0$, $v = 0$), se obtendría la fórmula inicial expuesta en la ecuación (3.15).

Algunos autores recomiendan que esta corrección del coeficiente de Spearman se utilice solamente en el caso de haya un número considerable de empates en las observaciones (véase Morales y Rodríguez, 2016).

3.3.2. La τ de Kendall

El coeficiente de correlación de rangos de Kendall, conocido comúnmente como coeficiente tau de Kendall, recibe su nombre en reconocimiento al matemático inglés Maurice George Kendall, quien desarrolló en 1938 una medida de asociación por rangos alternativa a la de Spearman basada en la concordancia de las observaciones, término que ya introdujimos en la *Proposición 2.9*.

Recordemos que la búsqueda de un coeficiente que fuera invariante bajo todas las transformaciones que preservan el orden nos llevó a definir el coeficiente de correlación de Spearman. Pues bien, Kendall intentó refinar la idea de este último utilizando probabilidades y sin perder la condición de invarianza.

Dado que las probabilidades de sucesos que implican relaciones de desigualdad son invariantes bajo todas las transformaciones monótonas ya que no alteran la relación de orden entre los valores, una medida de asociación basada en las probabilidades de concordancia y discordancia, como es la tau de Kendall,

procurará satisfacer los siete criterios de la *Proposición 2.9*. La información relativa a este apartado ha sido consultada en el Capítulo 11 de Gibbons y Chakraborti, 2003.

Para describir el coeficiente tau de Kendall, es necesario definir previamente las probabilidades de concordancia y discordancia que denotaremos por p_c y p_d, respectivamente.

Definición 3.34. Sean (X_i, Y_i) y (X_j, Y_j), $i \neq j$, dos pares de observaciones del vector aleatorio continuo (X, Y). Se dice que los pares (X_i, Y_i) y (X_j, Y_j) son **concordantes** si $(X_j - X_i)(Y_j - Y_i) > 0$ y **discordantes** si $(X_j - X_i)(Y_j - Y_i) < 0$. De aquí, se obtiene que las probabilidades asociadas a dichos conceptos son:

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

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

Definición 3.35 (Coeficiente de Rangos de Kendall)**.** Sean X e Y dos variables aleatorias. La versión poblacional del **coeficiente tau de Kendall**, que denotaremos por τ, se define como la diferencia entre las probabilidades de concordancia y discordancia, es decir,

$$\tau = p_c - p_d. \tag{3.17}$$

Observación 3.36. Si consideramos X' e Y' dos variables aleatorias i.i.d. respecto a las variables X e Y. Entonces τ también puede expresarse como

$$\tau = \mathrm{P}[(X - X')(Y - Y') \geq 0] - \mathrm{P}[(X - X')(Y - Y') \leq 0].$$

Al igual que el coeficiente de correlación de Spearman, τ es un coeficiente mucho menos estricto que ρ en cuanto a los supuestos que deben cumplir las variables para que este funcione. Como ambos son coeficientes basados en rangos, τ también contará con las ventajas que anteriormente explicamos para ρ_s, entre ellas que no se verá afectado por la presencia de valores atípicos o por la forma de la distribución.

Como antes, consideraremos que las distribuciones marginales de X e Y son continuas eliminando así la probabilidad de que existan empates $X_i = X_j$ e $Y_i = Y_j$. De aquí se deduce que $\mathrm{P}[(X_j - X_i)(Y_j - Y_i) = 0] = 0$ dando lugar a la siguiente relación entre p_c y p_d:

$$p_c = \mathrm{P}[(X_j - X_i)(Y_j - Y_i) > 0] = 1 - \mathrm{P}[(X_j - X_i)(Y_j - Y_i) \leq 0] =$$
$$= 1 - \{\mathrm{P}[(X_j - X_i)(Y_j - Y_i) < 0] + \mathrm{P}[(X_j - X_i)(Y_j - Y_i) = 0]\} = 1 - p_d.$$

Así pues, sustituyendo esta igualdad en la ecuación (3.17), bajo las condiciones mencionadas τ puede expresarse como

$$\tau = 1 - 2p_d = 2p_c - 1. \tag{3.18}$$

Teorema 3.37 (Propiedades del Coeficiente de Rangos de Kendall)**.** *El coeficiente de rangos de Kendall verifica todos los criterios expuestos en la Proposición 2.9 relativos a una buena medida de asociación.*

Demostración. Sean X e Y dos variables aleatorias continuas y τ el coeficiente de rangos de Kendall.

1. Sean (X_i, Y_i) y $(X_j, Y_j),\ i \neq j$, dos pares de observaciones del vector aleatorio continuo (X, Y). Si la relación es de concordancia perfecta, entonces $p_d = 0$ y, por la ecuación (3.18), $\tau = 1$.

2. En las mismas condiciones de antes, si la relación es de discordancia perfecta, entonces $p_c = 0$ y, por la ecuación (3.18), $\tau = -1$.

3. Si no se cumplen ni 1 ni 2, se tiene que $0 < p_c,\ p_d < 1$, puesto que son probabilidades. Tomando, por ejemplo, $0 < p_c < 1$ llegamos a que

$$0 < p_c < 1 \overset{(\cdot\, 2)}{\Longrightarrow} 0 < 2p_c < 2 \overset{(-1)}{\Longrightarrow} -1 < 2p_c - 1 < 1 \overset{\text{ec. (3.18)}}{\Longrightarrow} -1 < \tau < 1.$$

4. Si X e Y son independientes, se tiene que $\mathrm{P}(X_i < X_j) = \mathrm{P}(X_i > X_j)$ y las probabilidades conjuntas en p_c y p_d son el producto de las probabilidades individuales, es decir,

$$\begin{aligned} p_c &= \mathrm{P}(X_i < X_j)\,\mathrm{P}(Y_i < Y_j) + \mathrm{P}(X_i > X_j)\,\mathrm{P}(Y_i > Y_j) = \\ &= \mathrm{P}(X_i > X_j)\,\mathrm{P}(Y_i < Y_j) + \mathrm{P}(X_i < X_j)\,\mathrm{P}(Y_i > Y_j) = p_d. \end{aligned}$$

   Sustituyendo esta igualdad en la ecuación (3.17) llegamos a que $\tau = 0$.

5.∼6. Debido a la propiedad conmutativa del producto, la intersección o la suma, es inmediato ver que las probabilidades de concordancia y discordancia, p_c y p_d, permanecen invariantes (salvo signo) y, por tanto, τ.

7. Dado que τ es una medida de asociación basada en rangos, y estos son invariantes ante todas las transformaciones monótonas, τ será invariante.

□

El coeficiente τ comparte algunas similitudes con los coeficientes de Pearson y Spearman y es que el recíproco de la propiedad 4 de la *Proposición 2.9* no se cumple en general a no ser que los datos pertenezcan a una población normal bivariante.

En estas condiciones, es posible relacionar τ con ρ del siguiente modo:

Teorema 3.38. *Sea (X, Y) un vector aleatorio que sigue una distribución normal con vector de medias μ y matriz de varianzas-covarianzas Σ y sea ρ el coeficiente de correlación de Pearson. Entonces τ se puede escribir en función de ρ como*

$$\tau = \frac{2}{\pi} \arcsin \rho. \tag{3.19}$$

La demostración de este teorema se puede consultar en el Anexo I.

Del *Teorema 3.38* se puede deducir fácilmente que, bajo la distribución normal bivariante, $\tau = 0$ también implica la independencia de las variables, pues si $\rho = 0$, por la ecuación (3.19), $\tau = \frac{2}{\pi} \arcsin(0) = 0$ y, por tanto, por el *Teorema 3.18*, las variables serán independientes.

En las práctica, para estimar el parámetro τ a partir de una muestra aleatoria de n pares $\{(X_i, Y_i)\}_{i=1}^n$ de una población bivariante, debemos calcular las estimaciones de los parámetros p_c y p_d. Para cada conjunto (X_i, Y_i), (X_j, Y_j) de pares de observaciones muestrales, se definen en primer lugar las siguientes variables indicadoras:

$$A_{ij} = sgn(X_j - X_i) sgn(Y_j - Y_i)\ , \tag{3.20}$$

donde

$$sgn(u) = \begin{cases} -1, & \text{si } u < 0, \\ 0, & \text{si } u = 0, \\ 1, & \text{si } u > 0. \end{cases}$$

Por tanto, los valores que podrá tener la variable A_{ij} son

$$a_{ij} = \begin{cases} 1, & \text{si los pares son concordantes,} \\ -1, & \text{si los pares son discordantes,} \\ 0, & \text{si no son concordantes ni discordantes debido a que } X_i = X_j \text{ o } Y_i = Y_j. \end{cases}$$

La distribución de probabilidad marginal de esta variable indicadora es:

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

Como se tiene que, por definición, $a_{ij} = a_{ji}$ y $a_{ii} = 0$, hay únicamente $\binom{n}{2}$ conjuntos de pares que debemos considerar. A partir de esto, es sencillo construir un estimador de τ como sigue:

Definición 3.39 (Coeficiente de Rangos de Kendall Muestral)**.** Si X e Y son variables aleatorias que toman valores $\{(X_i, Y_i)\}_{i=1}^n$ de una muestra aleatoria simple de tamaño n, se puede estimar el coeficiente de rangos de Kendall, T, mediante

$$T = \sum\sum_{1\leq i<j\leq n} \frac{A_{ij}}{\binom{n}{2}} = 2 \cdot \sum\sum_{1\leq i<j\leq n} \frac{A_{ij}}{n(n-1)}, \tag{3.21}$$

donde los A_{ij} designan las variables indicadoras dadas en la ecuación (3.20).

Observación 3.40. La ecuación (3.21) es comparable a la forma en la que previamente se define τ. Por ello, las propiedades mencionadas con anterioridad para la versión poblacional de este coeficiente también son válidas para T.

Cabe destacar que, por cómo están definidos los A_{ij} en la ecuación (3.20), si $X_i = X_j$ o $Y_i = Y_j$ entonces, $A_{ij} = 0$. Como el estadístico presentado en la ecuación (3.21) depende de estos términos A_{ij}, T es capaz de manejar adecuadamente observaciones empatadas ya que, en el caso de haberlas, contribuirían con valor cero al mismo, esto es, la estimación de τ no depende de la continuidad de la distribución de la población.[9]

La forma en la que se ha definido T en la ecuación (3.21) es la forma más sencilla para la deducción de sus propiedades teóricas. Sin embargo, otra forma común de expresar este estadístico es haciendo una clasificación de los pares de observaciones según sean concordantes o discordantes a partir del signo obtenido de calcular los A_{ij}. Si denotamos por C y D el número de pares concordantes (A_{ij} positivos) y discordantes (A_{ij} negativos) para $1 \leq i < j \leq n$, respectivamente, y por S la diferencia $C - D$, tendremos que

$$T = \frac{C - D}{\binom{n}{2}} = \frac{S}{\binom{n}{2}}. \tag{3.22}$$

Si no hay empates dentro de los grupos X o Y, es decir, $A_{ij} \neq 0$ para $i \neq j$, entonces $C + D = \binom{n}{2}$. Despejando C o D en esta expresión, podemos reescribir la ecuación (3.22) como

$$T = 1 - \frac{2D}{\binom{n}{2}} = \frac{2C}{\binom{n}{2}} - 1. \tag{3.23}$$

Estas dos igualdades de la ecuación (3.23) son claramente análogas a las obtenidas para el parámetro τ en la ecuación (3.18). Dicha correspondencia revela que $\frac{C}{\binom{n}{2}}$ y $\frac{D}{\binom{n}{2}}$ son los estimadores insesgados de p_c y p_d, respectivamente.

Observación 3.41. Dada una muestra de tamaño n, la cantidad C es posiblemente la más sencilla de calcular. Supongamos que los pares se escriben de menor a mayor según el valor de la componente X;

[9] El supuesto de continuidad en el contexto de la distribución bivariante implica que la probabilidad de que dos observaciones sean iguales es cero.

entonces C se calculará simplemente como el número de valores de $1 \leq i < j \leq n$ para los que $Y_j - Y_i > 0$ ya que, de esta forma, se tendrá que $a_{ij} = 1$.

Otra interpretación de T es como *coeficiente de desorganización*. Recibe este nombre ya que puede demostrarse que el número total de intercambios entre dos observaciones de Y consecutivas hasta transformar su disposición en el orden natural, de menor a mayor, sin dejar de asumir que la disposición de X también sigue el orden natural, es igual a D, o lo que es lo mismo, $\dfrac{\binom{n}{2}(1-T)}{2}$.

Por otro lado, cabe preguntarse si T es un buen estimador de τ. Por ser una variable discreta, la esperanza de los A_{ij} vendrá dada de la siguiente forma:

$$\mathrm{E}[A_{ij}] = \sum a_{ij} p_{A_{ij}}(a_{ij}) = 1p_c + (-1)p_d + 0(1 - p_c - p_d) = p_c - p_d = \tau.$$

Por la ecuación (3.21),

$$\mathrm{E}[T] = \frac{2}{n(n-1)} \sum \sum_{1 \leq i < j \leq n} \mathrm{E}[A_{ij}] = \frac{2}{n(n-1)} \frac{n(n-1)}{2} \cdot \tau = \tau. \tag{3.24}$$

De la ecuación (3.24) se obtiene que T es un estimador **insesgado** de τ. Ya que los empates no afectan al valor esperado del estadístico T, en consecuencia, no introducen sesgo en la estimación. Por tanto, incluso en presencia de empates, T seguirá siendo un estimador insesgado de τ en cualquier distribución bivariante.

También se puede demostrar que T es un estimador **consistente** de τ para cualquier distribución bivariante dado que su varianza se aproxima a cero a medida que el tamaño de la muestra se aproxima a infinito. Esto se traduce en que, para una cantidad significativa de datos, T proporcionará resultados muy cercanos al valor del verdadero parámetro. Para ver los detalles relativos a la demostración de esta propiedad, véase Gibbons y Chakraborti, 2003 (pp. $405 - 408$).

Si bien es cierto que T es capaz de manejar los empates de una manera adecuada sin introducir sesgo en la estimación, dependiendo de la naturaleza de dichos empates, hay situaciones en las que es preferible la utilización de una fórmula adaptada a ellos ya que pueden modificar ligeramente el denominador de la ecuación (3.22) dado que no se considerarían los empates para verificar la concordancia. Es posible que esto último afecte a la varianza del estimador y se subestime la fuerza de la relación al dividir por un número más grande que el que debería ser (véase Gibbons y Chakraborti, 2003 (pp. $418 - 420$)).

Teorema 3.42. *En presencia de empates, la tau de Kendall, T, se define como*

$$T = \frac{2S}{\left[\left(n(n-1) - \sum_{k=1}^{l} u_k(u_k-1)\right)\left(n(n-1) - \sum_{k=1}^{l} v_k(v_k-1)\right)\right]^{1/2}},$$

donde u_k y v_k denotan el número de observaciones empatadas en el grupo k de un total de l dentro de las muestras X e Y, respectivamente.

Capítulo 4

Nuevos puntos de vista

Hasta ahora solo hemos hablado de coeficientes que miden relaciones muy concretas (lineales o monótonas) para un par de variables. Cabe preguntarse si existen otros coeficientes que amplíen el conocimiento y permitan explorar otro tipo de dependencias de mayor complejidad y en dimensiones mayores. Sin embargo, las extensiones multivariantes de ciertos coeficientes, como por ejemplo los basados en rangos, son ineficaces para probar tipos no monótonos de dependencia. De ahí surge la correlación de distancias, introducida por el matemático Gábor J. Székely en el año 2005. Esta medida de dependencia se basa principalmente en otra medida de distancias acuñada por él mismo en 1980: la distancia de energía.

4.1. El Coeficiente de Correlación de Distancias

La correlación de distancias es una nueva medida de dependencia entre vectores aleatorios introducida formalmente en Székely et al., 2007. Para calcular esta medida, Székely y Rizzo, 2013 se basaron en el concepto de distancia de energía que se calcula a partir de la distancia entre las observaciones que, en este caso, será la euclidiana.

Notación 4.1. Denotaremos el producto escalar de dos vectores t y s por $\langle t, s\rangle$. Para funciones con valores complejos $\varphi(\cdot)$, denotaremos el conjugado complejo por $\overline{\varphi}$ y $|\varphi|^2 = \varphi\overline{\varphi}$. La norma euclidiana de $x \in \mathbb{R}^p$ se denotará por $|x|_p$. Por último, una muestra de $X \in \mathbb{R}^p$ se denotará por la matriz $\mathbf{X} \in \mathcal{M}_{n\times p}$ y sus filas estarán formadas por los vectores muestrales $(X_i)_{i=1}^n$.

La **distancia de energía** es una distancia estadística entre las distribuciones de vectores aleatorios que caracteriza la igualdad de dichas distribuciones. Se basa, como bien indica su nombre, en la noción de energía gravitatoria introducida por Newton, que no es más que una función de la distancia entre dos cuerpos.

Definición 4.2 (Distancia de Energía)**.** La **distancia de energía** entre variables aleatorias d-dimensionales independientes X e Y se define como

$$\mathcal{E}(X,Y) = 2\,\mathrm{E}\,|X-Y|_d - \mathrm{E}\,|X-X'|_d - \mathrm{E}\,|Y-Y'|_d,$$

donde $\mathrm{E}\,|X|_d < \infty$, $\mathrm{E}\,|Y|_d < \infty$ y X' e Y' son variables i.i.d. respecto de X e Y.

Por otro lado, en la práctica, es habitual el uso de estadísticos de energía. Los estadísticos de energía, denominados U-estadísticos y V-estadísticos, son funciones de distancias entre observaciones en espacios métricos por lo que, aunque las observaciones sean objetos complejos (funciones), el uso de las distancias reales no negativas resulta de gran ayuda. Estos son muy útiles y suelen ser más generales y a menudo más potentes que los estadísticos clásicos.

Székely y Rizzo, 2013 optan por utilizar los V-estadísticos ya que serán no negativos y será más sencillo interpretarlos como una distancia estadística. Los V-estadísticos están basados en distancias, es decir, para una muestra aleatoria d-dimensional $X_1, \ldots, X_n$ y una función núcleo $h : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$,

$$V_n = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} h(X_i, X_j),$$

donde h es una función simétrica de las distancias euclidianas entre los elementos de la muestra.

Más adelante veremos que la forma empírica del coeficiente de correlación de distancias presenta forma de V-estadístico.

La correlación de distancias, que denotaremos por $\mathcal{R}$, se puede ver como una extensión de la distancia de energía, ya que esta medirá la dependencia entre conjuntos de variables en lugar de simplemente comparar distribuciones. Dicha correlación se deriva de una serie de cantidades: la varianza de distancias, la desviación típica de distancias y la covarianza de distancias. Estas cantidades desempeñan las mismas funciones que la varianza, desviación típica y covarianza clásicas definidas para el coeficiente de correlación de Pearson.

Esta nueva medida es fácil de estimar y generaliza la idea de correlación proporcionando un nuevo enfoque al problema de probar la independencia conjunta de vectores aleatorios. A lo largo de esta sección consideraremos que $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$ son vectores aleatorios, donde p y q son enteros positivos con $p \neq q$, en general.

Si recordamos, la independencia también se puede medir a través de las funciones características, de modo que dos vectores X e Y son independientes si $\varphi_{X,Y} = \varphi_X \varphi_Y$. Es por esto que la covarianza de distancias, que denotaremos por $\mathcal{V}$, puede aplicarse para medir la distancia $\|\varphi_{X,Y} - \varphi_X\varphi_Y\|$ usando la norma euclidiana.

Definición 4.3. Para funciones complejas γ definidas en $\mathbb{R}^p \times \mathbb{R}^q$, la **norma en el espacio ponderado** de funciones en $\mathbb{R}^{p+q}$, que denotaremos por $\|\cdot\|_w$, se define como

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

donde $w(t,s)$ es una función de peso positiva arbitraria para la que existe la integral anterior.

Haciendo uso de la definición anterior y eligiendo una función de peso conveniente, se puede definir una medida de dependencia de la siguiente forma:

$$\mathcal{V}^2(X,Y;w) = \|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2 w(t,s)\, dtds.$$

Observación 4.4. Por las propiedades de la función característica, $\mathcal{V}^2(X,Y;w) = 0$ si y solo si X e Y son independientes. Por tanto, $\mathcal{V}$ será análogo al valor absoluto de la covarianza clásica y dividiendo

$$\frac{\mathcal{V}(X,Y;w)}{\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}},$$

donde $\mathcal{V}^2(X;w) = \mathcal{V}^2(X,X;w) = \int_{\mathbb{R}^{2p}} |\varphi_{X,X}(t,s) - \varphi_X(t)\varphi_X(s)|^2 w(t,s)\, dtds$, obtendremos un tipo de correlación absoluta que denotaremos por R_w.

Sin embargo, no todas las funciones de peso conducen a un R_w que resulte interesante. Buscamos un coeficiente que sea invariante con respecto a transformaciones de escala, es decir, $R_w(X,Y) = R_w(\varepsilon X, \varepsilon Y)$ para $\varepsilon > 0$, y positivo cuando las variables sean dependientes para que de esta forma R_w caracterice la independencia. Si la función de peso $w(t,s)$ es integrable, y tanto X como Y tienen varianza finita entonces se tiene que

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

Así, para w integrable, si $\rho = 0$, entonces R_w puede ser arbitrariamente cercano a cero incluso si X e Y son dependientes. Sin embargo, si aplicamos una función de peso no integrable, obtendríamos un R_w que es invariante respecto a transformaciones de escala y que no puede ser cero para X e Y dependientes. La elección de w no es única, pero la elección que se hace en Székely et al., 2007 para esta función garantiza que las fórmulas empíricas sean sencillas y fáciles de aplicar.

Lema 4.5. *Si $0 < \alpha < 2$, entonces para todo $x \in \mathbb{R}^d$ se tiene que*

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

donde

$$C(d,\alpha) = \frac{2\pi^{\frac{d}{2}}\Gamma\left(1-\frac{\alpha}{2}\right)}{\alpha 2^\alpha \Gamma\left(\frac{d+\alpha}{2}\right)} \quad y \quad \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\, dt \quad \textit{es la función gamma completa.}$$

Observación 4.6. Las integrales en $t = 0$ y $t = \infty$ se calculan como $\lim_{\varepsilon \to 0} \int_{\mathbb{R}^d \setminus \{\varepsilon B + \varepsilon^{-1} B^c\}}$, donde B es la bola unitaria centrada en 0 en $\mathbb{R}^d$ y B^c es el complementario de B.

En el caso más simple ($\alpha = 1$), la constante C expuesta en el *Lema 4.5* es

$$c_d = C(d,1) = \frac{\pi^{\frac{1+d}{2}}}{\Gamma\left(\frac{1+d}{2}\right)}.$$

En vista de esto, la elección natural para la función de peso será

$$w(t,s) = \frac{1}{c_p c_q |t|_p^{1+p} |s|_q^{1+q}}. \tag{4.1}$$

Observación 4.7. En lo que sigue $\|\cdot\| = \|\cdot\|_w$, $\mathcal{V}(\cdot,\cdot) = \mathcal{V}(\cdot,\cdot\,;w)$ y $\mathcal{V}(\cdot) = \mathcal{V}(\cdot\,;w)$, siendo w la función de peso definida en la ecuación (4.1). Además, en las integrales resultará útil utilizar el símbolo dw que se define como

$$dw = \frac{1}{c_p c_q |t|_p^{1+p} |s|_q^{1+q}}\, dtds.$$

Para garantizar la finitud de $\|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|^2$ basta que $\mathrm{E}\,|X|_p < \infty$ y $\mathrm{E}\,|Y|_q < \infty$. Si consideramos $U = \exp i\langle t, X\rangle - \varphi_X(t)$ y $V = \exp i\langle s, Y\rangle - \varphi_Y(s)$, por la desigualdad de Cauchy-Schwarz se tiene que

$$\begin{aligned} |\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2 = |\mathrm{E}\,[UV]|^2 \leq (\mathrm{E}[|U\|V|])^2 \leq \mathrm{E}[|U|^2|V|^2] = \\ = (1 - |\varphi_X(t)|^2)(1 - |\varphi_Y(s)|^2). \end{aligned} \tag{4.2}$$

Si $\mathrm{E}[|X|_p + |Y|_q] < \infty$, entonces por el *Lema 4.5*, el Teorema de Fubini y la desigualdad de la ecuación (4.2), se tiene que

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

donde X' e Y' son variables i.i.d. con respecto de X e Y.

Una vez explicado lo anterior, ya estamos en condiciones de introducir los conceptos asociados a la correlación de distancias.

Definición 4.8 (Covarianza de Distancias)**.** Sean X e Y dos vectores aleatorios con medias finitas. La **covarianza de distancias** (dCov) entre X e Y, que denotaremos por $\mathcal{V}(X,Y)$, se define como la raíz cuadrada positiva de

$$\mathcal{V}^2(X,Y) = \|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|^2 = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2}{|t|_p^{1+p} |s|_q^{1+q}}\, dtds.$$

Teorema 4.9 (Propiedades de la Covarianza de Distancias)**.** *Sea $\mathcal{V}(X,Y)$ la covarianza de distancias entre dos vectores aleatorios X e Y con medias finitas.*

1. $\mathcal{V}(X,Y) \geq 0$.
2. $\mathcal{V}^2(a_1+b_1\,\mathbf{C}_1\,X,\ a_2+b_2\,\mathbf{C}_2\,Y) = |b_1\,b_2|\,\mathcal{V}^2(X,Y)$, *para todos los vectores constantes $a_1 \in \mathbb{R}^p$, $a_2 \in \mathbb{R}^q$, los escalares $b_1, b_2 \in \mathbb{R}$ y las matrices ortonormales $\mathbf{C}_1 \in \mathcal{M}_{p\times p}, \mathbf{C}_2 \in \mathcal{M}_{q\times q}$.*
3. *Si los vectores aleatorios (X_1, Y_1) y (X_2, Y_2) son independientes, entonces*

$$\mathcal{V}(X_1 + X_2, Y_1 + Y_2) \leq \mathcal{V}(X_1, Y_1) + \mathcal{V}(X_2, Y_2).$$

   La igualdad se cumple si y solo si X_1 e Y_1 o X_2 e Y_2 son ambas constantes o si X_1, X_2, Y_1, Y_2 son mutuamente independientes.
4. $\mathcal{V}(X,Y) = 0$ *si y solo si X e Y son independientes.*

Análogamente a lo visto para la covarianza de distancias, se puede definir la varianza de distancias:

Definición 4.10 (Varianza de Distancias)**.** Sea X un vector aleatorio con media finita. La **varianza de distancias** (dVar) de X, que denotaremos por $\mathcal{V}(X)$, se define como la raíz cuadrada positiva de

$$\mathcal{V}^2(X) = \mathcal{V}^2(X,X) = \|\varphi_{X,X}(t,s) - \varphi_X(t)\varphi_X(s)\|^2.$$

Teorema 4.11 (Propiedades de la Varianza de Distancias)**.** *Sea $\mathcal{V}(X)$ la varianza de distancias del vector aleatorio X con media finita.*

1. $\mathcal{V}(X) = 0$ *implica que $X = \mathrm{E}[X]$, casi seguro.*
2. $\mathcal{V}(a + b\mathbf{C}X) = |b|\,\mathcal{V}(X)$, *para todos los vectores contantes $a \in \mathbb{R}^p$, los escalares $b \in \mathbb{R}$ y las matrices ortonormales $\mathbf{C} \in \mathcal{M}_{p\times p}$.*
3. $\mathcal{V}(X+Y) \leq \mathcal{V}(X) + \mathcal{V}(Y)$, *para dos vectores aleatorios independientes X e Y.*
   La igualdad se cumple si y solo si uno de los vectores aleatorios X o Y es constante.

Una vez definido lo anterior ya es posible definir el coeficiente de correlación de distancias:

Definición 4.12 (Coeficiente de Correlación de Distancias)**.** Sean X e Y dos vectores aleatorios con medias finitas. El **coeficiente de correlación de distancias** (dCor) entre X e Y, que denotaremos por $\mathcal{R}(X,Y)$, se define como la raíz cuadrada positiva de

$$\mathcal{R}^2(X,Y) = \begin{cases} \dfrac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \text{si } \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0, \\ 0, & \text{si } \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \tag{4.3}$$

Observación 4.13. Es evidente que la definición de $\mathcal{R}$ sugiere una analogía con el coeficiente de correlación de Pearson, ρ.

Como siempre, en la práctica debemos basarnos en un coeficiente sujeto a una muestra de tamaño finito, por lo que es importante definir el coeficiente de correlación de distancias en una muestra aleatoria simple $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i) : i = 1, \dots, n\}$ a partir de la distribución conjunta de los vectores aleatorios $X \in \mathbb{R}^p$ e $Y \in \mathbb{R}^q$. Para ello, se deben calcular las matrices de distancia transformadas A y B para los vectores X e Y, respectivamente.

Antes de definir formalmente cómo se calcula este coeficiente en una muestra, daremos unas nociones sobre el procedimiento a seguir. En primer lugar, necesitaremos calcular la covarianza de distancias para lo que se seguirán los siguientes pasos:

1. Se obtienen dos matrices $n \times n$ de distancias, una para cada vector. Para ello, se calculan las distancias euclidianas entre los pares de observaciones para cada uno de los vectores quedando así la formulación para las n^2 entradas de cada matriz:

$$a_{ij} = |X_i - X_j|_p \quad \text{y} \quad b_{ij} = |Y_i - Y_j|_q, \quad i, j = 1, \dots, n.$$

2. Se calcula la media de las filas y de las columnas de cada matriz de distancias obtenida en el paso anterior, así como la media total de ellas. Formularemos los conceptos para la matriz A correspondiente al vector X.

$$\bar{a}_{i\bullet} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \qquad \bar{a}_{\bullet j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \qquad \bar{a}_{\bullet\bullet} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij}.$$

   Análogamente se definen los conceptos $\bar{b}_{i\bullet}$, $\bar{b}_{\bullet j}$ y $\bar{b}_{\bullet\bullet}$ para la matriz correspondiente al vector Y.

3. Se realiza un doble centrado a cada elemento en las matrices, es decir, se le resta la media de su fila y de su columna y se le suma la media de toda la matriz.

$$A_{ij} = a_{ij} - \bar{a}_{i\bullet} - \bar{a}_{\bullet j} + \bar{a}_{\bullet\bullet} \quad \text{y} \quad B_{ij} = b_{ij} - \bar{b}_{i\bullet} - \bar{b}_{\bullet j} + \bar{b}_{\bullet\bullet}, \quad i, j = 1, \dots, n.$$

   De esta forma, nos aseguramos de que tanto las filas como las columnas sumen 0.

4. Por último, se calculan las covarianzas entre las n^2 distancias centradas.

A partir de esto, es sencillo obtener la forma empírica de la covarianza de distancias como sigue:

Definición 4.14 (Covarianza de Distancias Muestral)**.** Sea $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i)\}_{i=1}^n$ una muestra aleatoria simple de los vectores aleatorios X e Y. La covarianza de distancias muestral, que denotaremos por $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$, se define como la raíz cuadrada positiva de

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}.$$

Observación 4.15. La distancias calculadas en el paso 1 se pueden generalizar de manera que para todas las potencias α de las mismas, $|\cdot|^\alpha$, con $\alpha \in (0, 2)$, $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}; \alpha) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$. El intervalo en el que está definido α se podría intuir a partir de la elección en el *Lema 4.5*. Cabe destacar que, aunque $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}; \alpha = 2)$ también está definido, este no caracteriza la independencia; es más, se tiene que en el caso bivariado, $\mathcal{R}(X, Y; \alpha = 2)$ puede ser calculado de manera precisa a partir del coeficiente de correlación de Pearson.

Análogamente, $\mathcal{V}_n(\mathbf{X})$ será la varianza de distancias muestral, que se define como la raíz cuadrada positiva de

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^2.$$

Una vez más, estamos en condiciones de definir la forma empírica del coeficiente de correlación de distancias:

Definición 4.16 (Coeficiente de Correlación de Distancias Muestral)**.** Sea $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i)\}_{i=1}^n$ una muestra aleatoria simple de los vectores aleatorios X e Y. El coeficiente de correlación de distancias muestral, que denotaremos por $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$, se define como la raíz cuadrada positiva de

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

Teorema 4.17. *Si $(\boldsymbol{X}, \boldsymbol{Y})$ es una muestra a partir de la distribución conjunta de (X, Y). Entonces*

$$\mathcal{V}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \|\varphi_{X,Y}^n(t, s) - \varphi_X^n(t) \varphi_Y^n(s)\|^2,$$

donde $\varphi_{X,Y}^n(t, s) = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} \exp\{i\langle t, X_k\rangle + i\langle s, Y_k\rangle\}$ denota la función característica en una muestra $\{(X_i, Y_i)\}_{i=1}^n$ y $\varphi_X^n(t)$ y $\varphi_Y^n(s)$ las respectivas funciones características marginales muestrales de X e Y.

Teorema 4.18. *Si* $\mathrm{E}\,|X|_p < \infty$ *y* $\mathrm{E}\,|Y|_q < \infty$*, entonces*

$$\lim_{n\to\infty} \mathcal{V}_n(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{V}(X, Y), \quad \textit{casi seguro.}$$

Observación 4.19. Las demostraciones de estos teoremas son demasiado extensas y se escapan del objetivo principal de este trabajo por lo que si el lector lo desea puede consultar Székely et al., 2007 (pp. $2774 - 2778$).

Corolario 4.20. *Del Teorema 4.18 se deduce que si* $\mathrm{E}[|X|_p + |Y|_q] < \infty$*, entonces*

$$\lim_{n\to\infty} \mathcal{R}_n^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{R}^2(X, Y), \quad \textit{casi seguro.}$$

La definición de $\mathcal{R}$ sugiere que esta medida de dependencia basada en las distancias tiene propiedades análogas al coeficiente de correlación de Pearson, lo que nos lleva a pensar que ciertos resultados expuestos para la correlación y la varianza clásicas deberían ser válidos para $\mathcal{R}$ y $\mathcal{V}$.

Teorema 4.21 (Propiedades del Coeficiente de Correlación de Distancias)**.** *Sea* $\mathcal{R}(X, Y)$ *el coeficiente de correlación de distancias entre dos vectores aleatorios* X *e* Y *y* $\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y})$ *su correspondiente versión muestral. Entonces*

1. *Si* $\mathrm{E}[|X|_p + |Y|_q] < \infty$*, entonces* $0 \leq \mathcal{R}(X, Y) \leq 1$ *y*

   $$\mathcal{R}(X, Y) = 0 \ \textit{ si y solo si } X \textit{ e } Y \textit{ son independientes.}$$

2. $0 \leq \mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) \leq 1$.

3. *Si* $\mathcal{R}_n(\boldsymbol{X}, \boldsymbol{Y}) = 1$*, entonces existe un vector* a*, un número real* $b \in \mathbb{R}\backslash\{0\}$ *y una matriz ortogonal* $\mathbf{C}$ *tal que* $\boldsymbol{Y} = a + b\boldsymbol{X}\mathbf{C}$.

Demostración.

1. $\mathcal{R}(X, Y)$ existe siempre que X e Y tengan medias finitas y X e Y son independientes si y solo si el numerador

   $$\mathcal{V}^2(X, Y) = \|\varphi_{X,Y}(t, s) - \varphi_X(t)\varphi_Y(s)\|^2$$

   de $\mathcal{R}^2(X, Y)$ es cero. Para probar que $0 \leq \mathcal{R} \leq 1$ retomemos la desigualdad expuesta en la ecuación (4.2). De aquí, se obtiene que

   $$\int_{\mathbb{R}^{p+q}} |\varphi_{X,Y}(t, s) - \varphi_X(t)\varphi_Y(s)|^2 \, dw \leq \int_{\mathbb{R}^{p+q}} |(1 - |\varphi_X(t)|^2)(1 - |\varphi_Y(s)|^2)|^2 \, dw,$$

   y, por tanto, $0 \leq \mathcal{R}(X, Y) \leq 1$.

2. Se sigue por un razonamiento análogo a 1.

3. Sin pérdida de generalidad, supondremos que X e Y pertenecen al mismo espacio euclídeo y ambos están contenidos en $\mathbb{R}^p$. A partir de la desigualdad de Cauchy-Schwarz se puede comprobar fácilmente que $\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = 1$ si y solo si $A_{ij} = \varepsilon B_{ij}$ para algún ε. Supongamos que $|\varepsilon| = 1$, entonces

$$|X_i - X_j|_p = |Y_i - Y_j|_q + d_i + d_j,$$

para todo i y j y para ciertas constantes d_i, d_j. Considerando $i = j$ tenemos que $d_i = 0$ para todo i. Como ambas muestras son isométricas, se puede obtener Y en función de X a través de operaciones de traslación, rotación y reflexión. Por tanto, se puede expresar $\mathbf{Y} = a + b\mathbf{X}\mathbf{C}$ para un cierto vector a, $b = \varepsilon$ y una matriz ortogonal $\mathbf{C}$. En el caso de que $\varepsilon \neq 0$ y $|\varepsilon| \neq 1$, aplicamos nuevamente el argumento geométrico a $\varepsilon\mathbf{X}$ e $\mathbf{Y}$ y se sigue que $\mathbf{Y} = a + b\mathbf{X}\mathbf{C}$, donde $b = \varepsilon$.

□

En resumen, el coeficiente de correlación de distancias presenta una innumerable lista de propiedades que lo hacen destacar por encima de otros. Székely y Rizzo, 2009 muestra que la correlación de distancias verifica propiedades análogas a medidas de asociación que describimos anteriormente como que $\mathcal{R}(X, Y) = \mathcal{R}(Y, X)$ o $\mathcal{R}(aX + b, cY + d) = \mathcal{R}(X, Y)$.

$\mathcal{R}$ también se puede generalizar a cualquier espacio métrico aunque no en todos se caracteriza la independencia. En Lyons, 2013, se explica que para probar independencia es necesario y suficiente un espacio métrico de tipo *fuertemente negativo*. Por otro lado, no necesita que las variables sigan una distribución normal y es capaz de detectar relaciones no lineales. Maneja variables aleatorias multidimensionales que pueden estar en diferentes dimensiones y no tienen por qué ser escalables. Además, su estimación es sencilla y fácil de calcular ya que no requiere inversión de matrices o estimación de parámetros.

En algunos casos resulta interesante considerar medidas de dependencia que sean invariantes afines. A pesar de que la medida definida en la ecuación (4.3) no es invariante afín, se puede adaptar para que este coeficiente no se vea afectado por este tipo de transformaciones.

De este modo se pueden definir dos muestras aleatorias escaladas a partir de las muestras $\mathbf{X}$ e $\mathbf{Y}$ mediante

$$\mathbf{X}^* = \mathbf{X}S_X^{-\frac{1}{2}} \quad \text{e} \quad \mathbf{Y}^* = \mathbf{Y}S_Y^{-\frac{1}{2}},$$

donde S_X y S_Y son las matrices de covarianza muestrales de $\mathbf{X}$ e $\mathbf{Y}$, respectivamente.

Entonces, el estadístico para la correlación de distancias afín, que se denotará por $\mathcal{R}_n^*(\mathbf{X}, \mathbf{Y})$ entre dos muestras aleatorias $\mathbf{X}$ e $\mathbf{Y}$ es la raíz cuadrada positiva de

$$\mathcal{R}_n^{*2}(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^2(\mathbf{X}^*, \mathbf{Y}^*)}{\sqrt{\mathcal{V}_n^2(\mathbf{X}^*)\mathcal{V}_n^2(\mathbf{Y}^*)}}.$$

Observación 4.22. Las propiedades anteriormente comentadas son igualmente válidas para $\mathcal{R}_n^*$ ya que lo que se consigue con la transformación anterior es modificar la ecuación (4.1) referente a la función peso: $w(t,s) \longrightarrow w\left(S_X^{-1/2}t, S_Y^{-1/2}s\right)$.

Por último, siempre es interesante estudiar los coeficientes bajo las condiciones de la distribución normal bivariante ya que de esta manera somos capaces de ver qué cantidad representa la correlación de distancias en el caso de vectores aleatorios bivariados al relacionarla con ρ.

Sean X e Y dos variables que siguen una distribución normal estándar con $\mathrm{Cov}(X,Y) = \rho(X,Y) = \rho$. Definamos una nueva función

$$F(\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2 \frac{dt}{t^2}\frac{ds}{s^2}.$$

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

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}} = \frac{F(\rho)}{F(1)}. \tag{4.4}$$

Teorema 4.23. *Si X e Y siguen una distribución normal estándar con coeficiente de correlación $\rho = \rho(X,Y)$, entonces*

1. $\mathcal{R}(X,Y) \leq |\rho|$.
2. $\mathcal{R}^2(X,Y) = \dfrac{\rho \arcsin\rho + \sqrt{1-\rho^2} - \rho \arcsin\frac{\rho}{2} - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}}$.
3. $\inf\limits_{\rho\neq 0} \dfrac{\mathcal{R}(X,Y)}{|\rho|} = \lim\limits_{\rho\to 0} \dfrac{\mathcal{R}(X,Y)}{|\rho|} = \dfrac{1}{2(1+\pi/3-\sqrt{3})^{1/2}} \approxeq 0.89066.$

La demostración de este teorema se puede consultar en el Anexo I.

Por último, cabe destacar que el estadístico $\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})$ es un estimador sesgado de $\mathcal{V}^2(X,Y)$. Sin embargo, se puede ver que es **asintóticamente insesgado** en vista de su valor esperado:

$$\mathrm{E}[\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})] = \frac{(n-1)(n-2)^2}{n^3}\mathcal{V}^2(X,Y) + \frac{2(n-1)^2}{n^3}\gamma - \frac{(n-1)(n-2)}{n^3}\alpha\beta,$$

donde $\gamma = \mathrm{E}[|X-X'||Y-Y'|]$, $\alpha = \mathrm{E}\,|X-X'|$ y $\beta = \mathrm{E}\,|Y-Y'|$.

Haciendo entonces $\lim\limits_{n\to\infty}\left(\mathrm{E}[\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})] - \mathcal{V}^2(X,Y)\right) = \mathcal{V}^2(X,Y) - \mathcal{V}^2(X,Y) = 0$, se llega a la conclusión expuesta. En Székely y Rizzo, 2014 se da una expresión acerca de un estimador insesgado para $\mathcal{V}^2(X,Y)$.

4.2. El coeficiente de correlación visto desde la perspectiva de las funciones cópula

En esta sección definiremos lo que son las funciones cópula y proporcionaremos algunos resultados de interés que guardan gran relación con ellas. Antes de entrar en formalidades, daremos una breve e intuitiva definición de las mismas para ir familiarizándonos con el concepto. Esta información fue obtenida de Lloréns, s.f.

La cópula es lo que resta de una función de distribución al eliminar de la misma el comportamiento marginal de las variables y regula y determina la estructura de dependencia entre ellas. Dicho de otra forma, en la cópula reside la forma en la que unas variables dependen de otras.

La motivación para hablar de las cópulas surge de las limitaciones que presentan muchas medidas de dependencia en determinadas situaciones y, en esos casos en particular, es donde nos resultarán de utilidad las cópulas. A continuación, definiremos formalmente estas funciones basándonos en Hernández y Caita, 2017. Por simplicidad, consideraremos el caso bivariante aunque todos estos resultados se hacen extensibles al caso multivariante de forma natural.

Definición 4.24 (Cópula)**.** Una **cópula** C es la función de distribución de un vector aleatorio (U, V) sobre el rectángulo $\mathbb{I}^2 := [0, 1]^2$ cuyas funciones de distribución marginales siguen una uniforme $U(0, 1)$. Alternativamente, una cópula es una función $C : \mathbb{I}^2 \longrightarrow \mathbb{I}$ con las siguientes propiedades:

1. $C(u, v)$ es una función no decreciente en cada componente para todo $u, v \in \mathbb{I}$.
2. Para cualquier $u, v \in \mathbb{I}$, se tiene que
   - $C(u, 0) = 0 = C(0, v)$.
   - $C(u, 1) = u$.
   - $C(1, v) = v$.
3. Para cualesquiera $u_1, u_2, v_1, v_2 \in \mathbb{I}$ tales que $u_1 \leq u_2$ y $v_1 \leq v_2$, tenemos que
$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

Una vez definido formalmente el concepto de cópula, introduciremos la base sobre la que se fundamenta esta idea: el Teorema de Sklar, que lleva el nombre del matemático Abe Sklar, a quien se le atribuye el haber enlazado por medio de una cópula la función de distribución conjunta de un vector aleatorio y sus distribuciones marginales. Este se considera uno de los resultados más importantes en la teoría de cópulas y juega un papel muy importante ya que, además, proporciona un método para la construcción de cópulas dada una función de distribución bivariada.

Teorema 4.25 (Teorema de Sklar)**.** *Sea F una función de distribución conjunta bivariada con funciones de distribución marginales F_X y F_Y para las variables X e Y, respectivamente. Entonces existe una cópula $C : \mathbb{I}^2 \to \mathbb{I}$ tal que para cualesquiera $x, y \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$,*

$$F(x, y) = C(F_X(x), F_Y(y)). \tag{4.5}$$

Si F_X y F_Y son continuas, entonces C es única; en cualquier otro caso, C está determinada de forma única sobre el conjunto $\text{rango}(F_X) \times \text{rango}(F_Y)$.

Observación 4.26. Si C es una cópula y F_X y F_Y son funciones de distribución univariadas, entonces la función F vista en la igualdad (4.5), es una función de distribución conjunta bivariada con marginales F_X y F_Y.

Se puede ver que la expresión

$$C(u, v) = F(F_X^{-1}(u), F_Y^{-1}(v)), \tag{4.6}$$

es una cópula cuando se evalúan $x = F_X^{-1}(u)$ e $y = F_Y^{-1}(v)$ en la ecuación (4.5) y se suponen F_X y F_Y continuas. Esto quiere decir que podemos usar la expresión dada en la ecuación (4.6) para construir cópulas cuando se conoce la función de distribución y sus marginales.

Observación 4.27. Las propiedades expuestas a continuación se pueden consultar en el Capítulo 1 de Balakrishnan y Lai, 2009.

Teorema 4.28 (Propiedades de las Funciones Cópula)**.** *Sea C una cópula y $(u, v) \in [0, 1] \times [0, 1]$. Entonces*

1. *$C^-(u, v) \leq C(u, v) \leq C^+(u, v)$, donde $C^+(u, v) = \text{mín}(u, v)$ y $C^-(u, v) = \text{máx}(u+v-1, 0)$.*
2. *Para todo $v \in [0, 1]$, existe la derivada parcial $\dfrac{\partial C}{\partial u}$ para casi todo u y $0 \leq \dfrac{\partial}{\partial u} C(u, v) \leq 1$. Análogamente, $0 \leq \dfrac{\partial}{\partial v} C(u, v) \leq 1$.*
3. *Sea (U, V) un par de variables aleatorias independientes. Entonces su cópula asociada será $C(u, v) = uv$.*
4. *La combinación convexa de dos cópulas también es una función cópula.*
5. *La función cópula se conserva ante transformaciones estrictamente crecientes.*

4.2.1. Cópulas Arquimedianas

A continuación, introduciremos una familia de cópulas que tienen propiedades deseables que facilitan la manipulación y el modelado de los datos. La información relativa a este apartado se puede consultar en Díaz, 2013. Estas cópulas son capaces de capturar la dependencia de colas, es decir, cómo las variables se comportan de manera conjunta en los extremos de sus distribuciones. Además, son funciones invertibles a partir de las que se pueden calcular funciones de distribución conjuntas: las cópulas arquimedianas. Estas cópulas son de la forma $\varphi(C(u,v)) = \varphi(u) + \varphi(v)$, por lo que si queremos encontrar la definición explícita de C tendremos que definir previamente una función inversa que denotaremos por φ^{-1}.

Definición 4.29 (Pseudoinversa)**.** Sea $\varphi : [0,1] \longrightarrow [0,\infty]$ una función continua y estrictamente decreciente tal que $\varphi(1) = 0$. La **pseudoinversa** de φ es la función $\varphi^{[-1]} : [0,\infty] \longrightarrow [0,1]$ dada por

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

Observación 4.30. Nótese que $\varphi^{[-1]}(t)$ es continua y decreciente en $[0,\infty]$ y estrictamente decreciente en $[0,\varphi(0)]$. Sin embargo, $\varphi^{[-1]}(\varphi(t)) = t$ para todo $t \in [0,1]$ y

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

En el caso de que $\varphi(0) = \infty$, entonces se tiene que $\varphi^{[-1]} = \varphi^{-1}$.

Llegados a este punto, estamos en condiciones de definir las cópulas arquimedianas (lo haremos para el caso multidimensional).

Teorema 4.31. *Sea $\varphi : [0,1] \longrightarrow [0,\infty]$ una función continua y estrictamente decreciente tal que $\varphi(1) = 0$ y sea $\varphi^{[-1]}$ la pseudoinversa de φ. Si $C : [0,1]^n \longrightarrow [0,1]$ es una función dada por*

$$C(u_1, ..., u_n) = \varphi^{[-1]}(\varphi(u_1) + ... + \varphi(u_n)). \tag{4.7}$$

Entonces C es una cópula si y solo si φ es convexa.

Las cópulas de la forma vista en la ecuación (4.7) son llamadas **cópulas arquimedianas** y la función φ es el generador de la cópula. Si $\varphi(0) = \infty$, decimos que φ es un generador estricto y, en este caso, $C(u_1, ..., u_n) = \varphi^{-1}(\varphi(u_1) + ... + \varphi(u_n))$, es decir, es una cópula arquimediana estricta.

4.2.2. Cópulas Gaussianas

La cópula gaussiana es la distribución más popular en la práctica por su estructura sencilla y su facilidad para implementarla. Esta cópula es efectiva para capturar dependencias lineales y es especialmente útil cuando se trabaja con datos multivariados. Sin embargo, la cópula gaussiana no es capaz de captar adecuadamente la dependencia de colas. Esta no es una cópula arquimediana por lo que se deben tener en cuenta sus limitaciones a la hora de trabajar con ella.

Nos centraremos en un primer instante en el caso bivariante (véase Hernández y Caita, 2017) y después extenderemos los resultados a un número mayor de dimensiones.

Sea (X, Y) un vector aleatorio que sigue una función de distribución normal estándar bivariante, es decir, $(X, Y) \sim N_2(0, \mathrm{I}_2)$. Entonces, tal y como hemos visto en la ecuación (3.6) del Capítulo 3, su función de densidad, que seguiremos denotando por ϕ, vendrá dada por

$$\phi(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{-(x^2 - 2\rho xy + y^2)}{2(1-\rho^2)}\right\},$$

y, por lo tanto, su función de distribución conjunta Φ_2^ρ tendrá la siguiente expresión:

$$\Phi_2^\rho(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{x}\int_{-\infty}^{y} \exp\left\{\frac{-(y_1^2 - 2\rho y_1 y_2 + y_2^2)}{2(1-\rho^2)}\right\} dy_1 dy_2,$$

donde $\rho \in (-1, 1)$ es el coeficiente de correlación de Pearson entre las variables X e Y.

Como consecuencia de la ecuación (4.6), se define la cópula gaussiana como sigue:

$$C_\rho^{Ga}(u, v) = \Phi_2^\rho(\Phi^{-1}(u), \Phi^{-1}(v)),$$

donde Φ^{-1} denota la inversa de la función de distribución normal estándar univarida.

Dicha cópula está dada explícitamente por la expresión

$$C_\rho^{Ga}(u, v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u)}\int_{-\infty}^{\Phi^{-1}(v)} \exp\left\{\frac{-(s^2 - 2\rho st + t^2)}{2(1-\rho^2)}\right\} dsdt.$$

Se concluye entonces que las variables con función de distribución $C_\rho^{Ga}(\Phi(u), \Phi(v))$ siguen una normal estándar bivariada con coeficiente de correlación ρ.

Extendamos este resultado a una dimensión arbitraria n. Los siguientes resultados han sido tomados de Žežula, 2009. En este caso, reconsiderando la función de densidad de la distribución normal estándar

para dimensión n, tendríamos que es de la forma

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

donde $x = (x_1, ..., x_n)'$ y R representa la matriz de correlación.

Para cualquier distribución multivariante absolutamente continua con función de distribución conjunta F y funciones marginales F_i, con $i = 1, ...n$, la cópula C es una función de distribución en $[0, 1]^n$ que cumple:

$$F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)).$$

Dado que C es una función de distribución, derivándola obtendríamos su función de densidad correspondiente, a la que denotaremos por c. Si ahora f fuese la función de densidad conjunta correspondiente y f_i, con $i = 1, ...n$, las densidades marginales, se define la densidad de la cópula como sigue:

$$\frac{\partial^n C}{\partial F_1 ... \partial F_n}.$$

En consecuencia, podríamos expresar la función de densidad conjunta f como

$$f(x) = c(F_1(x_1), ..., F_n(x_n)) \prod_{i=1}^{n} f_i(x_i).$$

Como la densidad normal multivariante se puede reescribir de la forma:

$$\phi_n(x) = \frac{1}{|R|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}x'(R^{-1} - \mathrm{I}_n)x\right\} \prod_{i=1}^{n} \phi_i(x_i),$$

donde $x_i = \Phi^{-1}(F_i(x_i))$, la función de densidad de una cópula gaussiana vendrá dada por

$$c(x) = \frac{1}{|R|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}x'(R^{-1} - \mathrm{I}_n)x\right\}.$$

Las cópulas gaussianas son herramientas que resultan útiles en diversas situaciones ya que admiten cualquier distribución marginal y cualquier matriz de correlaciones definida positiva. A pesar de esto, las cópulas gaussianas consideran solamente la dependencia por pares entre componentes individuales de una variable aleatoria, lo que no abarca la profundidad total de estructuras de dependencia posibles.

Existen, además, otros problemas conectados con el uso práctico de las cópulas gaussianas. A continuación, mencionamos algunos de los más relevantes:

- Cuando la dimensión n es grande, R puede resultar difícil de estimar dado que hay demasiados parámetros.

- Las densidades gaussianas se basan en el uso del coeficiente de correlación de Pearson que, recordemos, no es invariante bajo transformaciones monótonas de las variables originales, pero dichas transformaciones son requeridas habitualmente.
- El coeficiente ρ no resulta una medida de dependencia apropiada en muchas situaciones y son normalmente estas las situaciones en las que nos interesaría usar la modelización de cópulas.

En ciertas ocasiones, el problema del exceso de parámetros puede ser evitado. De hecho, muchas veces nos encontramos con estructuras de correlación más sencillas, que pueden utilizarse para la construcción de cópulas gaussianas multivariantes simples con pocos parámetros a estimar. En lo que sigue, destacaremos las dos estructuras más importantes:

- Estructura de correlación uniforme. Este modelo está caracterizado por la matriz de correlación

$$R = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} = (1-\rho)\,\mathrm{I}_n + \rho \mathbf{11'},$$

  donde **1** denota el vector $(1,\ldots,1)^{\mathsf{T}} \in \mathbb{R}^n$ y $\rho \in \left[\frac{-1}{n-1},\ 1\right]$.

- Estructura de correlación serial. En este modelo la matriz de correlación es

$$R = \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix} = \mathrm{I}_n + \sum_{i=1}^{n-1} \rho^i (C_1^i + C_1^{i'}),$$

  donde $C_1 = \begin{pmatrix} 0_{n-1} & \mathrm{I}_{n-1} \\ 0 & 0'_{n-1}, \end{pmatrix}$ y $\rho \in [-1,1]$.

4.2.3. Correlación a través de cópulas

Tal y como hemos visto, el coeficiente de correlación de Pearson causa problemas en ciertas ocasiones. Por este motivo, a partir de las cópulas se plantean unos coeficientes de correlación alternativos que además ya son conocidos para nosotros: la τ de Kendall y la ρ de Spearman. Los veremos a continuación

dándoles un enfoque distinto al visto hasta el momento. Esta vez, comenzaremos la explicación para el coeficiente τ de Kendall y para finalizar el capítulo trataremos la ρ de Spearman, ρ_s. La información relativa a todo este apartado se puede consultar en el Capítulo 5 de Nelsen, 2006.

La τ de Kendall

Consideremos un par de variables aleatorias continuas (X, Y). Tal y como hemos visto en la ecuación (3.36), se puede definir la expresión del coeficiente τ como sigue:

$$\tau = \left\{ \mathrm{P}[(X - X')(Y - Y') > 0] - \mathrm{P}[(X - X')(Y - Y') < 0] \right\},$$

donde X' e Y' son variables i.i.d. respecto a las variables X e Y.

Teorema 4.32. *Sean F y F' las funciones de distribución conjunta de (X, Y) y (X', Y'), respectivamente, F_X la función marginal de X y X' y F_Y la función marginal de Y e Y'. Sean ahora C_1 y C_2 las cópulas de (X, Y) y (X', Y'), respectivamente, tales que $F(x, y) = C_1(F_X(x), F_Y(y))$ y $F'(x, y) = C_2(F_X(x), F_Y(y))$. Entones, dado que $\tau = \left\{ \mathrm{P}[(X - X')(Y - Y') > 0] - \mathrm{P}[(X - X')(Y - Y') < 0] \right\}$, se tiene que*

$$\tau = \tau(C_1, C_2) = 4 \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v) - 1.$$

La demostración de este teorema se puede consultar en el Anexo I.

Una consecuencia interesante que se desprende del *Teorema 4.32* es que si ahora tenemos un vector aleatorio (X, Y), cuya cópula asociada es C, se tendrá que

$$\tau = \tau(C, C) = 4 \int_0^1 \int_0^1 C(u, v) c(u, v) du dv - 1. \tag{4.8}$$

Además, se puede observar que la integral de la ecuación (4.8) puede ser interpretada como el valor esperado de la función $C(U, V)$ considerando U y V variables aleatorias que siguen una uniforme $U(0, 1)$ y cuya distribución conjunta es C, esto es,

$$\tau = 4\,\mathrm{E}[C(U, V)] - 1.$$

Por otra parte, si la cópula C es arquimediana y la función φ es el generador de dicha cópula, podríamos relacionar el coeficiente y la cópula como sigue:

$$\tau = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)}\, dt.$$

Observación 4.33. Se puede encontrar la demostración de este resultado en Nelsen, 2006 (p. 163).

La ρ de Spearman

A pesar de que en el Capítulo 3 dimos la versión poblacional del coeficiente de rangos de Spearman como se puede ver en la *Definición 3.24*, existe otra forma de expresar ρ_S basada en la concordancia y la discordancia de las observaciones.

Definición 4.34. Sean (X_1, Y_1), (X_2, Y_2) y (X_3, Y_3) tres pares de variables independientes con función de distribución conjunta F. Entonces ρ_s es proporcional a la probabilidad de concordancia menos la probabilidad de discordancia para los pares (X_1, Y_1) y (X_2, Y_3), es decir,

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

Teorema 4.35. *Sean X e Y dos variables aleatorias continuas cuya cópula es C. Entonces se tiene que la versión poblacional de la ρ de Spearman para X e Y viene dada por*

$$\rho_s = 12 \int_0^1 \int_0^1 C(u, v)\, dudv - 3. \tag{4.9}$$

Como antes, la integral de la ecuación (4.9) puede ser interpretada como un valor esperado para U y V, variables siguiendo una uniforme $U(0, 1)$, esto es,

$$\rho_s = 12\,\mathrm{E}[UV] - 3 = \frac{\mathrm{E}[UV] - \frac{1}{4}}{\frac{1}{12}}.$$

Observación 4.36. A partir de lo anterior se deduce que la correlación de rangos de Spearman entre X e Y es simplemente el coeficiente de correlación producto-momento de Pearson entre las variables uniformes U y V.

Observación 4.37. La demostración de este resultado se sigue de la del *Teorema 4.32*. El lector podrá encontrar la justificación en Nelsen, 2006 (p. 167).

Capítulo 5

Ilustración en base a datos reales

Para finalizar este trabajo, dedicaremos el capítulo final a ilustrar las distintas medidas de dependencia explicadas sobre un conjunto de datos reales. La base de datos escogida se ha obtenido del *UC Irvine Machine Learning Repository* (Janosi et al., 1988)[1], aunque también es posible encontrarla accediendo al dataset $heart$ que incorpora el paquete *"kmed"* (Budiaji, 2022) de  . Dicha base de datos consta de un total de 14 variables que se corresponden con factores para el diagnóstico de enfermedades cardíacas de 297 pacientes. Para el análisis de la correlación, se han seleccionado únicamente 6 variables de interés que serán suficientes para ilustrar cómo funciona cada coeficiente.

Variables	**Descripción**
`cp`	Cuatro tipos de dolor de pecho: (1) angina típica, (2) angina atípica, (3) dolor no anginoso y (4) asintomático $\rightarrow$ Categórica.
`age`	Edad (años) $\rightarrow$ Numérica.
`trestbps`	Presión arterial en reposo al ingresar en el hospital (mmHg) $\rightarrow$ Numérica.
`chol`	Nivel de colesterol en sangre (mg/dl) $\rightarrow$ Numérica.
`thalach`	Frecuencia cardíaca máxima alcanzada (lat/min) $\rightarrow$ Numérica.
`oldpeak`	Depresión del segmento ST en el electrocardiograma después de realizar ejercicio físico en comparación con el estado de reposo (mm) $\rightarrow$ Numérica.

[1]Licencia: https://creativecommons.org/licenses/by/4.0/

El objetivo será obtener las correlaciones entre los pares de variables numéricas y contrastarlas con las que se obtienen al separar la muestra por grupos según la variable categórica para ver si las relaciones están o no influenciadas por dichos grupos.

Para ello, se calcularán las matrices de correlación para Pearson, Spearman, Kendall y distancias y se presentarán algunos gráficos para ilustrar los resultados. Por último, interpretaremos los resultados obtenidos; no debemos perder de vista que este es un estudio muy general en el que solo nos centraremos en lo que respecta al coeficiente de correlación y, por tanto, no exploraremos en profundidad muchas otras características de los datos. Es por esto que no podremos extraer conclusiones sólidas acerca de la dependencia de las variables pero sí hacernos una idea mediante los gráficos y los valores numéricos de qué información es capaz de aportar cada coeficiente.

Todos los resultados expuestos a continuación se calcularán con el lenguaje de programación R y el código utilizado se podrá consultar en el Anexo II.

En primer lugar, restringiremos el estudio a las variables numéricas consideradas anteriormente. Antes de nada, es conveniente realizar un diagrama de dispersión para cada par de variables. Visualizar cómo se distribuyen las observaciones en cada caso es útil para detectar posibles tendencias lineales, monótonas u otro tipo de relaciones que nos servirán para respaldar los resultados numéricos.

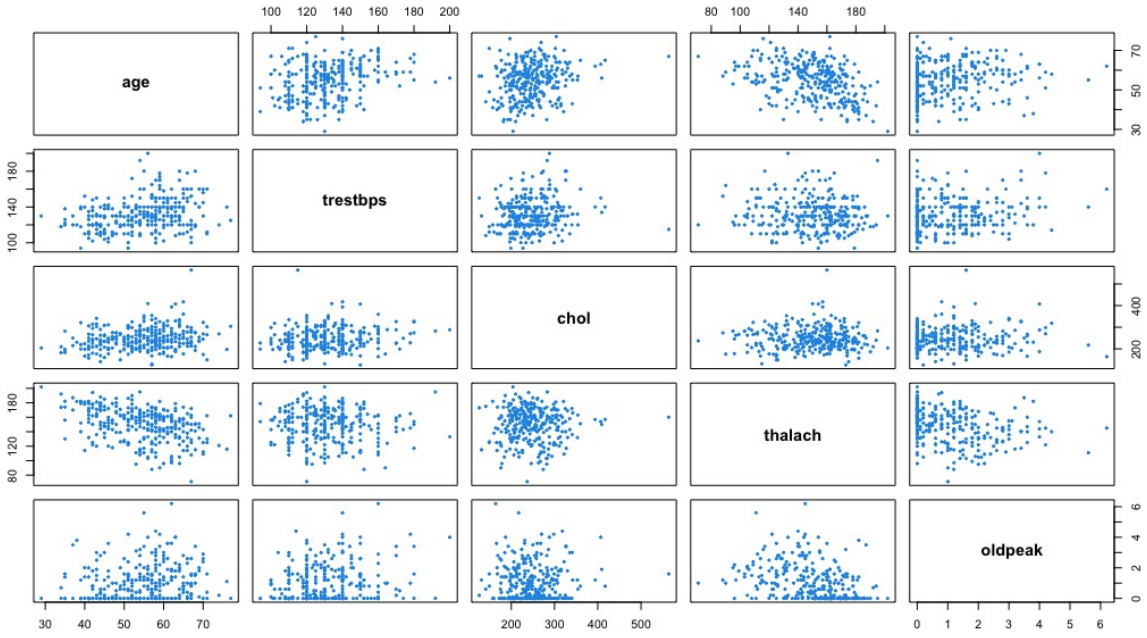


Figura 5.1: Gráfico de pares que refleja la nube de puntos entre las distintas variables dos a dos.

La Figura 5.1 muestra las relaciones bivariadas entre las variables numéricas del estudio a través de diagramas de dispersión en una matriz donde la diagonal indica los nombres de las variables y cada gráfico representa la relación de un par.

En dicha figura no se aprecian relaciones lineales fuertes ya que los puntos en los diagramas se encuentran bastante dispersos y no siguen una clara tendencia de línea recta. En general, las relaciones parecen débiles e incluso podrían llegar a revelar independencia entre las variables. Además, algunos datos están muy alejados del comportamiento esperado (datos atípicos) y podrían afectar a los resultados numéricos relativos al coeficiente de Pearson que, como ya explicamos, es sensible a ellos. Sin embargo, se pueden observar algunas tendencias monótonas, aunque no muy marcadas, que podrían ser clave para sustentar los resultados que expongamos a continuación.

Los puntos en los diagramas entre la edad y la frecuencia cardíaca presentan una trayectoria decreciente que nos indicaría que ambas variables son inversamente proporcionales. Este hecho parece lógico, pues que la capacidad máxima del corazón para latir sea menor es un fenómeno natural del envejecimiento.

Con menor intensidad, se puede apreciar una tendencia creciente de los puntos en los diagramas entre la edad y el nivel de colesterol o entre la edad y la presión arterial, aunque esta es muy leve. A partir de estas interpretaciones podríamos pensar que la edad es un factor relevante en lo que respecta a enfermedades cardíacas ya que afecta en mayor o menor medida a aspectos importantes de la salud cardiovascular.

Aún y todo, estas conclusiones se sustentan únicamente en la observación, por lo que no son más que conjeturas que deberemos apoyar con resultados analíticos.

Tras esta breve descripción de los datos, cabe preguntarse si la distribución de cada par de variables es conocida. En particular, nos interesa verificar si siguen una distribución normal bivariante[2]. Para ello, se empleará el test $mvn(\,)$ del paquete *"MVN"* (Korkmaz et al., 2014) de R. A continuación, se mostrará una tabla con los p-valores relativos al test para cada par de variables. Dado que los resultados son simétricos omitiremos el triángulo superior de dicha tabla.

	age	trestbps	chol	thalach	oldpeak
age	-	-	-	-	-
trestbps	$3{,}48 \times 10^{-03}$	-	-	-	-
chol	$1{,}58 \times 10^{-02}$	$1{,}83 \times 10^{-03}$	-	-	-
thalach	$4{,}04 \times 10^{-07}$	$3{,}13 \times 10^{-05}$	$1{,}57 \times 10^{-05}$	-	-
oldpeak	$3{,}40 \times 10^{-12}$	$7{,}49 \times 10^{-13}$	$1{,}14 \times 10^{-14}$	0	-

Tabla 5.1: p-valores relativos al test multivariante de Henze-Zirkler para cada par de variables.

[2]Recordemos que esta condición era de especial importancia para el adecuado funcionamiento del coeficiente de Pearson.

La Tabla 5.1 refleja p-valores extremadamente bajos. Es cierto que para un nivel de significación del $1\,\%$ no hay evidencias en contra de que la edad y el nivel de colesterol sigan una distribución normal bivariante, pero esta hipótesis se sigue rechazando para los otros niveles de significación habituales ($5\,\%$ y $10\,\%$). En el resto de casos, el p-valor indica que ninguno de los pares de variables considerados sigue una distribución normal bivariante, por lo que, salvando esa pequeña particularidad, asumiremos que la distribución de nuestros datos es desconocida.

Una vez analizadas estas características, ya podemos calcular la matriz de correlaciones de los datos[3]. Primero lo haremos usando el método de Pearson para el que se obtienen los siguientes resultados:

$$\mathcal{M}_r = \begin{bmatrix} 1{,}00 & 0{,}29 & 0{,}20 & -0{,}39 & 0{,}20 \\ 0{,}29 & 1{,}00 & 0{,}13 & -0{,}05 & 0{,}19 \\ 0{,}20 & 0{,}13 & 1{,}00 & 0{,}00 & 0{,}04 \\ -0{,}39 & -0{,}05 & 0{,}00 & 1{,}00 & -0{,}35 \\ 0{,}20 & 0{,}19 & 0{,}04 & -0{,}35 & 1{,}00 \end{bmatrix}$$

Algunos de los resultados obtenidos numéricamente refuerzan las conjeturas mencionadas anteriormente, pues la correlación más fuerte se corresponde con las variables relativas a la edad (`age`) y a la frecuencia cardíaca (`thalach`) con un valor de $-0{,}39$, que además es negativo, lo que indicaría una relación lineal inversa entre ellas. Otro par de variables digno de mención es el de la edad (`age`) y la presión arterial (`trestbps`), que en este caso presenta una correlación de $0{,}29$ siendo la más alta de las relaciones positivas seguido de otras como el colesterol (`chol`) y la edad (`age`) que también presentan una correlación positiva de $0{,}20$.

Cabe destacar que el par formado por la frecuencia cardíaca (`thalach`) y la depresión del segmento ST (`oldpeak`) aparentemente en el gráfico de la *Figura 5.1* no se detectaba una relación que pudiéramos pensar como lineal y, sin embargo, la matriz muestra un valor de $-0{,}35$ que es un valor relativamente significativo en comparación con el resto de valores calculados, algunos incluso muy cercanos a cero indicando una posible incorrelación.

Todas estas relaciones se pueden representar de innumerables formas. En este caso, lo haremos a partir de un gráfico que une cada variable con el resto mediante líneas más gruesas o más finas en función de la fuerza de la correlación y que representa los valores mediante colores. Para ello, necesitaremos utilizar el paquete *"correlation"* de  (Makowski et al., 2022)[4].

[3] Todas las matrices presentadas de ahora en adelante seguirán la misma estructura que la Tabla 5.1 en lo que a la ordenación de filas y columnas se refiere.

[4] La función relativa a la figura que presentamos a continuación se puede consultar en el Anexo II.

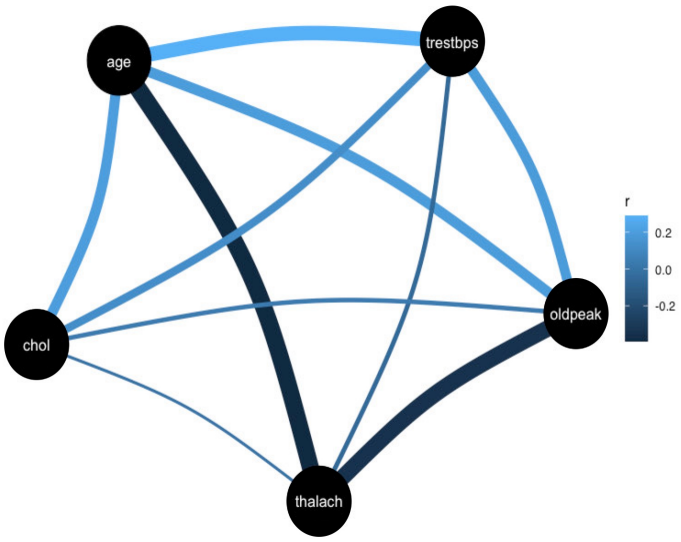


Figura 5.2: Gráfico de líneas relativo a las correlaciones obtenidas con el método de Pearson.

En general, las correlaciones de la matriz $\mathcal{M}_r$ son muy débiles (ninguna de ellas sobrepasa el valor absoluto de $0{,}4$), lo que es de esperar en vista de la gran dispersión que presentan los gráficos de la *Figura 5.1*. Sin embargo, estos resultados no son del todo adecuados ya que podrían verse afectados por la presencia de datos atípicos, por la distribución o por la presencia de los grupos que forma la variable categórica, factores que ya vimos en el Capítulo 3 que perjudicarían al coeficiente de correlación de Pearson.

Podríamos probar entonces la influencia de los datos atípicos a nivel bivariado sobre este coeficiente. Una vez más, el código para la detección de datos atípicos se puede encontrar en el Anexo II y necesitará nuevamente la utilización del paquete *"MVN"*. De este proceso, resultará un nuevo conjunto de datos que tendrá únicamente 93 observaciones. Realizando de nuevo el proceso mencionado anteriormente para verificar la normalidad de los pares, pero ahora sobre el nuevo conjunto, concluiremos que para algunos pares de variables aumenta significativamente el p-valor llegando a considerar varios de ellos normales para cualquiera de los niveles de significación habituales ($1\,\%$, $5\,\%$ y $10\,\%$). Como dicha eliminación de valores atípicos modifica la distribución levemente, también podría llegar a modificar el valor de la correlación. Veamos cómo sería la matriz de correlaciones resultante de dicho proceso, $\mathcal{M}_r^*$.

$$\mathcal{M}_r^* = \begin{bmatrix} 1{,}00 & 0{,}27 & 0{,}14 & -0{,}34 & 0{,}08 \\ 0{,}27 & 1{,}00 & 0{,}08 & 0{,}09 & 0{,}09 \\ 0{,}14 & 0{,}08 & 1{,}00 & 0{,}01 & 0{,}05 \\ -0{,}34 & 0{,}09 & 0{,}01 & 1{,}00 & -0{,}32 \\ 0{,}08 & 0{,}09 & 0{,}05 & -0{,}32 & 1{,}00 \end{bmatrix}$$

La falta de normalidad todavía presente en algunos pares podría ser un posible motivo de que los resultados no mejoren demasiado; es más, estos ni siquiera se han visto muy alterados para los pares en los que se ha conseguido la normalidad buscada, por lo que podríamos pensar que realmente existe una baja correlación lineal entre las variables, que parece lo esperable en vista de los diagramas de dispersión.

Sin embargo, este proceso de eliminación de datos atípicos no ha sido más que una comprobación de cómo influyen este tipo de datos y sus distribuciones al coeficiente de correlación de Pearson, pero este no es conveniente para nuestro estudio por diversas razones: no se han visto afectadas notablemente las correlaciones y además estaríamos reduciendo de un estudio inicial de 297 observaciones a uno de 93, es decir, se perdería aproximadamente un 69 % de la información relativa al estudio.

Por otra parte, recordemos que el coeficiente de Pearson es más preciso cuanto mayor es la muestra, por lo que eliminar esta gran cantidad de datos no solo no beneficiaría al coeficiente sino que perjudicaría en muchos aspectos a los resultados. Además, en lo sucesivo veremos coeficientes que son robustos a este tipo de observaciones y que no dependen de la distribución conjunta de los datos por lo que esto ya no será un problema al que debamos dar importancia.

Veamos por último para este coeficiente, meramente a título informativo, si estos resultados podrían mejorar considerando los grupos dados por la variable categórica (`cp`). Para visualizarlo, cargaremos en R los paquetes *"GGally"* y *"ggplot2"* (Schloerke et al., 2024 y Wickham, 2016, respectivamente) y emplearemos el comando $ggpairs(\ )$, que nos devuelve un gráfico que resume la información sobre la correlación por grupos para estas variables:

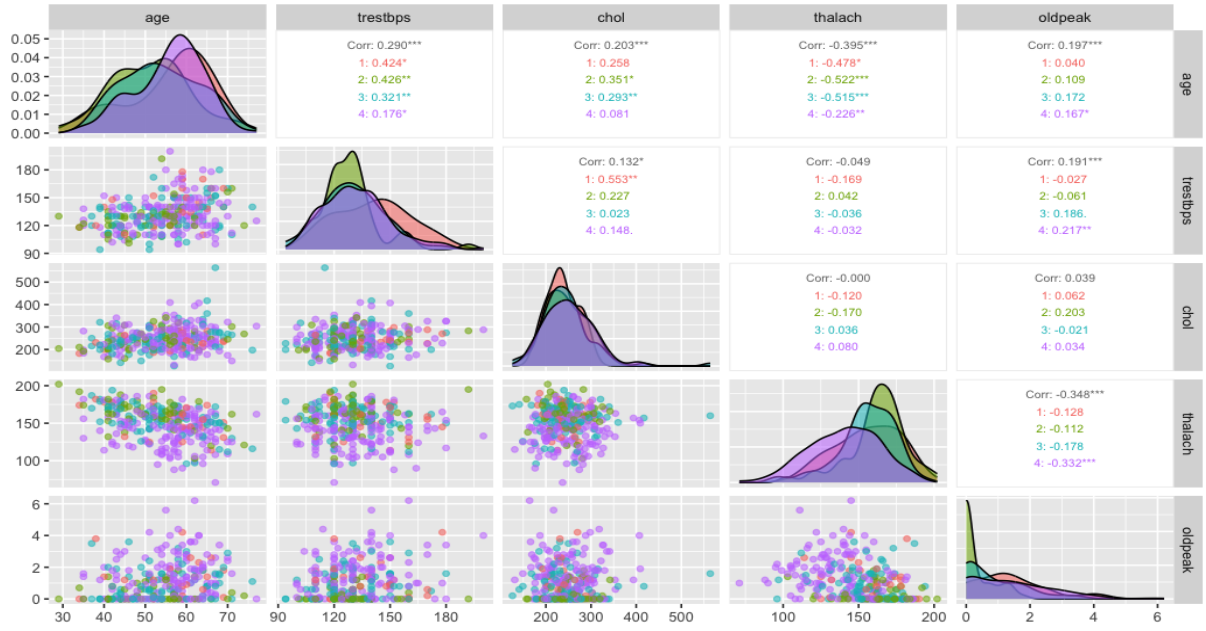


Figura 5.3: Gráfico mixto de correlaciones considerando los grupos dados por la variable categórica `cp`.

La *Figura 5.3* muestra en el triángulo inferior de la matriz los gráficos de dispersión relativos a los pares de variables por colores en función de los grupos de la variable categórica y en el triángulo superior los coeficientes de correlación de Pearson globales ya calculados en la matriz $\mathcal{M}_r$ junto con los distintos coeficientes si consideramos los grupos (cada uno en su respectivo color). Por último, en la diagonal, se muestra la forma de las funciones de densidad de las variables univariantes para los diferentes grupos.

En general, la *Figura 5.3* muestra un cambio en las relaciones entre variables si consideramos los grupos en comparación con la población general. Por ejemplo, la relación entre el colesterol y la presión arterial pasa a ser notablemente alta, de $0{,}132$ a $0{,}553$, cuando consideramos únicamente al grupo de personas que sufren de angina típica (grupo 1), lo que nos indicaría que estas dos variables presentan una mayor relación en personas con este tipo de dolencia. Lo mismo pasa con la relación entre la frecuencia cardíaca y la edad, que se vuelve más fuerte de forma inversa, de $-0{,}395$ a $-0{,}522$, para el grupo de personas que sufren de angina atípica, indicando que el envejecimiento de las personas que presentan dicha dolencia implica una mayor bajada en la frecuencia cardíaca. Al contrario sucede, por ejemplo, para el grupo de asintomáticos si consideramos la relación entre el colesterol y la edad, que presenta una relación mucho más débil en comparación con la población general.

Sin embargo, estos resultados se siguen pudiendo ver afectados tanto por el tamaño de la muestra para cada grupo, que es significativamente mayor para los asintomáticos, como por la variabilidad de los datos en dichos grupos, afectando directamente a la correlación. Es por esto, que en lo que sigue continuaremos el estudio para la totalidad de la muestra sin distinguir grupos ya que recurriremos al cálculo de coeficientes más adecuados al tipo de datos que estamos manejando.

Veamos entonces qué ocurre si calculamos los coeficientes de Spearman y Kendall basados en rangos que son más robustos a los datos atípicos y no necesitan de una distribución determinada. A continuación, análogamente a lo hecho para Pearson, mostraremos las matrices de correlación correspondientes que denotaremos por $\mathcal{M}_R$ (Spearman) y $\mathcal{M}_T$ (Kendall):

$$\mathcal{M}_R = \begin{bmatrix} 1{,}00 & 0{,}30 & 0{,}18 & -0{,}39 & 0{,}25 \\ 0{,}30 & 1{,}00 & 0{,}14 & -0{,}05 & 0{,}16 \\ 0{,}18 & 0{,}14 & 1{,}00 & -0{,}03 & 0{,}02 \\ -0{,}39 & -0{,}05 & -0{,}03 & 1{,}00 & -0{,}44 \\ 0{,}25 & 0{,}16 & 0{,}02 & -0{,}44 & 1{,}00 \end{bmatrix}, \quad \mathcal{M}_T = \begin{bmatrix} 1{,}00 & 0{,}21 & 0{,}13 & -0{,}28 & 0{,}18 \\ 0{,}21 & 1{,}00 & 0{,}10 & -0{,}03 & 0{,}11 \\ 0{,}13 & 0{,}10 & 1{,}00 & -0{,}02 & 0{,}02 \\ -0{,}28 & -0{,}03 & -0{,}02 & 1{,}00 & -0{,}31 \\ 0{,}18 & 0{,}11 & 0{,}02 & -0{,}31 & 1{,}00 \end{bmatrix}$$

El análisis de la correlación ofrece una gran cantidad de diagramas con diferentes formas y colores que apoyan y complementan la información del estudio. Para diversificar la visualización de los datos se mostrarán en las *Figuras II.I* y *II.II* del Anexo II dos diagramas correspondientes a los coeficientes de Spearman y Kendall: el primero referente a una matriz coloreada en función de los valores de las

correlaciones empleando el paquete *"correlation"* ya mencionado y el segundo utilizando elipses. Este último necesitará la utilización del paquete *"ellipse"* (Murdoch y Chow, 2023) y representará la fuerza y la dirección de la correlación a través del grosor y la inclinación de dichas elipses, respectivamente.

Analíticamente, los resultados obtenidos en las matrices $\mathcal{M}_R$ y $\mathcal{M}_T$ son muy semejantes debido a que ambos métodos se basan en los rangos de las variables. En vista de esto, compararemos entonces dichas matrices con la obtenida anteriormente para Pearson, $\mathcal{M}_r$.

Las interpretaciones obtenidas para la matriz de Pearson sugerían que las variables tenían un relación lineal débil o casi inexistente. Observando los resultados obtenidos para las matrices $\mathcal{M}_R$ y $\mathcal{M}_T$, los valores casi no se han visto modificados; el hecho de que estos coeficientes sean capaces de detectar otro tipo de relaciones (monótonas) y, sin embargo, los valores sigan siendo especialmente bajos podría indicar que las relaciones monótonas también son muy débiles, es decir, que las variables no disminuyen o aumentan de forma conjunta y consistente con un patrón claro.

Dado que cada coeficiente mide un tipo de dependencia distinto, una modificación en los valores no implica que ese coeficiente sea más o menos adecuado sino una diferente naturaleza de la relación.

Recordemos que los *Teoremas 3.31* y *3.38* relacionaban el coeficiente de correlación de Pearson con los de Spearman y Kendall, respectivamente, en condiciones de una distribución normal bivariante. Es interesante ver si las igualdades que presentaban son ciertas para los datos que estamos trabajando.

La Tabla 5.1 mostraba que, en general, ninguno de los pares verificaba este supuesto. Aún así, tomaremos como ejemplo las variables de la edad y el colesterol, que eran las únicas para las que se verificaba normalidad al 1 %, frente a la edad y la depresión del segmento ST, para las que no se verifica en ningún caso, con el fin de comparar los resultados. Empleando las relaciones dadas en los teoremas llegaremos a que

	`age-chol`	`age-oldpeak`
R	0.18	0.25
Ec. (3.16)	0.19	0.19

	`age-chol`	`age-oldpeak`
T	0.13	0.18
Ec. (3.19)	0.13	0.13

A la vista está que los resultados de la primera columna, donde están las variables que cumplen normalidad al 1 %, presentan resultados semejantes mientras que los de la segunda columna no, comprobando así que las igualdades mencionadas son eficaces.

En resumen, podríamos decir que los resultados vistos hasta ahora muestran variables débilmente relacionadas o incluso que pueden llegar a ser independientes. Veamos entonces, por último, lo que ocurre con el coeficiente de correlación de distancias, que es el único que caracteriza la independencia sin depender de la distribución y es capaz de detectar todo tipo de relaciones, tanto lineales como no lineales. Recor-

demos que dicho coeficiente solo toma valores positivos entre 0 y 1 y, por tanto, no indica la dirección de la relación sino únicamente la fuerza de la asociación.

Para calcular la matriz de correlación de distancias, que denotaremos por $\mathcal{M}_{\mathcal{R}}$, recurriremos al paquete *"energy"* de R (Rizzo y Szekely, 2022). El código se muestra en el Anexo II y se obtiene lo siguiente:

$$\mathcal{M}_{\mathcal{R}} = \begin{bmatrix} 1{,}00 & 0{,}31 & 0{,}19 & 0{,}37 & 0{,}24 \\ 0{,}31 & 1{,}00 & 0{,}15 & 0{,}10 & 0{,}17 \\ 0{,}19 & 0{,}15 & 1{,}00 & 0{,}11 & 0{,}09 \\ 0{,}37 & 0{,}10 & 0{,}11 & 1{,}00 & 0{,}40 \\ 0{,}24 & 0{,}17 & 0{,}09 & 0{,}40 & 1{,}00 \end{bmatrix}$$

En líneas generales, las correlaciones tampoco son elevadas, lo que podría sugerir que realmente estas variables no presentan relaciones fuertes de ningún tipo; incluso podríamos pensar que dichas relaciones son más bien monótonas o no lineales. La correlación entre el colesterol y la depresión del segmento ST presenta un valor de $0{,}09$ relativamente cercano a cero lo que podría indicarnos una posible independencia entre ambas. En los demás casos, no hay ninguna tan próxima a cero como la anterior como para suponer independencia pero sí relaciones muy débiles ya que los valores en los pares no han aumentados demasiado con respecto a los valores que se obtenían para Pearson.

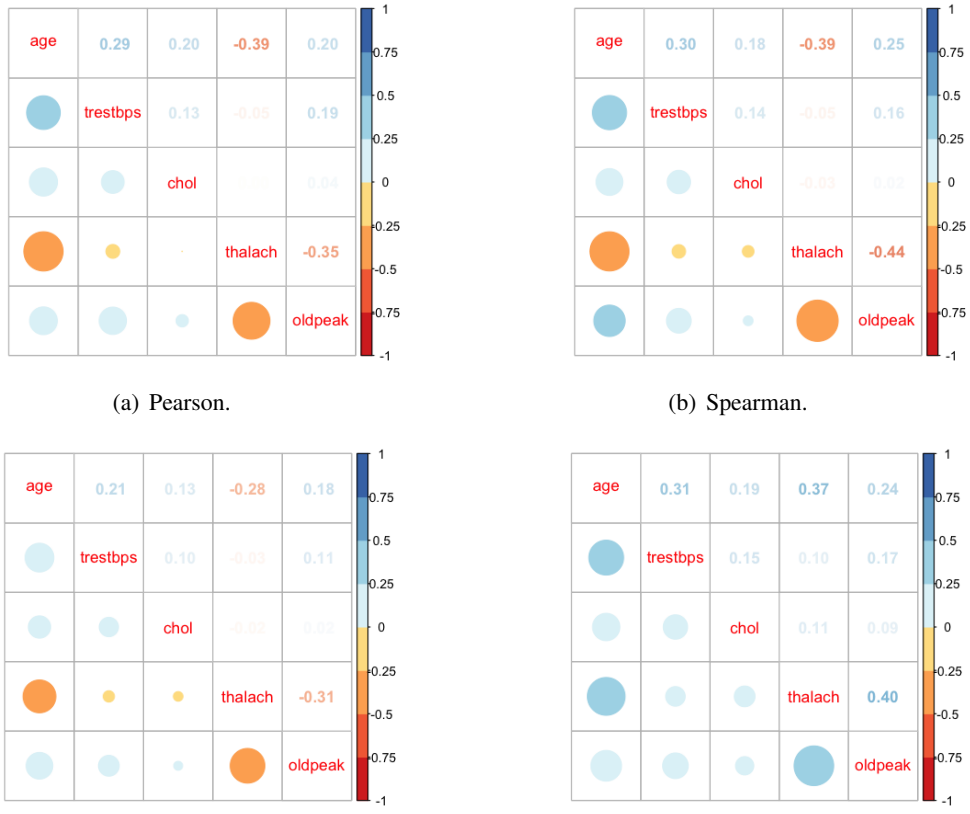
En la *Figura II.III* del Anexo II se podrá encontrar otro tipo de gráfico distinto a los ya vistos anteriormente pero que esencialmente recoge la misma información con los resultados obtenidos para la correlación de distancias. Para representar dicho gráfico se han empleado los paquetes *"RColorBrewer"* (Neuwirth, 2022) y *"corrplot"* (Wei y Simko, 2021) de R.

Por otro lado, análogamente a lo visto para Spearman y Kendall, el apartado 2 del *Teorema 4.23* muestra una forma de calcular la correlación de distancias a partir del coeficiente de Pearson bajo la distribución normal bivariante. En este caso, se obtienen los siguientes resultados:

	`age-chol`	`age-oldpeak`
$\mathcal{R}$	0.19	0.24
Th. 4.23 (2)	0.18	0.18

Esto, una vez más, prueba de manera práctica la igualdad que habíamos comprobado teóricamente en dicho teorema.

Para finalizar el estudio, presentaremos cuatro gráficos que muestran las correlaciones mencionadas a lo largo del capítulo: Pearson, Spearman, Kendall y distancias. Esta será un forma más clara y visual de ver los resultados obtenidos y comparar fácilmente los valores para cada par de variables en cada caso. Para representarlos, se han utilizado los paquetes *"RColorBrewer"* y *"corrplot"* ya mencionados.



(a) Pearson. (b) Spearman.

(c) Kendall. (d) Distancias.

Figura 5.4: Gráficos de correlación para los distintos métodos.

La *Figura 5.4* resume a la perfección lo que veníamos explicando a lo largo de todo el estudio. La diagonal superior muestra los valores de los coeficientes obtenidos en las matrices $\mathcal{M}_r$, $\mathcal{M}_R$, $\mathcal{M}_T$ y $\mathcal{M}_\mathcal{R}$, respectivamente y la diagonal inferior representa mediante círculos y colores la fuerza y dirección de la asociación.

A modo de conclusión, a pesar de que resulta difícil hacer una interpretación clara sobre la relación que presentan las variables ya que el estudio ha sido superficial y basado solo en los resultados de los coeficientes y la observación, podemos sospechar que las variables, en general, están poco relacionadas entre sí y posiblemente influidas por otros factores que no hemos considerado en el estudio.

ANEXOS

Anexo I

Demostraciones de resultados expuestos en la teoría

***Teorema 3.9* (pág. 12)**:

Demostración. Sean X e Y dos variables aleatorias y ρ el coeficiente de correlación de Pearson.

1. Efectivamente, si la relación entre X e Y es directa y perfecta, podremos expresar una en función de la otra como $Y = aX + b$ con $a, b \in \mathbb{R}$ y $a > 0$. Así, por cómo está definida la covarianza, se tiene que
$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, aX + b) = a\,\mathrm{Cov}(X, X) = a\,\mathrm{Var}(X).$$
Por otro lado,
$$\sqrt{\mathrm{Var}(Y)} = \sqrt{\mathrm{Var}(aX + b)} = \sqrt{a^2\,\mathrm{Var}(X)} = |a|\sqrt{\mathrm{Var}(X)}.$$
Como $a > 0$, $\dfrac{a}{|a|} = 1$ y sustituyendo todo lo anterior en la ecuación (3.3) de la *Definición 3.5* se obtendrá que:
$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathrm{Var}(X)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(X)}} = +1.$$

2. Esta demostración es análoga a la vista en 1 con la única diferencia de que ahora suponemos una relación inversa y perfecta entre las variables, es decir, $Y = aX + b$ con $a, b \in \mathbb{R}$ y $a < 0$. En este caso, $\dfrac{a}{|a|} = -1$ por lo que se obtendrá que la ecuación (3.3) será:
$$\rho(X, Y) = -\frac{\mathrm{Var}(X)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(X)}} = -1.$$

3. Esta propiedad se puede demostrar fácilmente a partir de la desigualdad de Cauchy-Schwarz. Se tiene que

$$(\mathrm{Cov}(X,Y))^2 = \left(\mathrm{E}\left[(X-\mathrm{E}[X])\,(Y-\mathrm{E}[Y])\right]\right)^2 \leq \left(\mathrm{E}\left[\,|(X-\mathrm{E}[X])\,(Y-\mathrm{E}[Y])|\,\right]\right)^2 \leq$$
$$\leq \mathrm{E}\left[(X-\mathrm{E}[X])^2\right]\cdot \mathrm{E}\left[(Y-\mathrm{E}[Y])^2\right] = \mathrm{Var}(X)\,\mathrm{Var}(Y).$$

Tomando raíces cuadradas a ambos lados de la expresión anterior obtenemos que

$$|\sigma_{XY}| = |\,\mathrm{Cov}(X,Y)| = \sqrt{(\mathrm{Cov}(X,Y))^2} \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} = |\sigma_X\sigma_Y|$$

y, por lo tanto,

$$|\rho(X,Y)| = \left|\frac{\sigma_{XY}}{\sigma_X\sigma_Y}\right| \leq 1 \quad \text{o, equivalentemente,} \quad \rho(X,Y)\in[-1,1].$$

4. Si X e Y son independientes, por la *Proposición 3.2* se tiene que el numerador de la ecuación (3.3) es cero y, por tanto, $\rho(X,Y)=0$.

5. Demostrar este criterio es inmediato haciendo uso de la propiedad simétrica de la covarianza. Comprobaremos solo una de las igualdades, siendo análogo para el resto de casos.

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

6. Procederemos del mismo modo que en 5. Una vez más, comprobaremos solo una de las igualdades ya que la otra es análoga.

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

□

Observación I.I. El valor de a en la demostración anterior se define en $\mathbb{R}\backslash\{0\}$ ya que, en el caso de que fuera cero, Y sería constante y ρ no estaría definido.

***Teorema 3.18* (pág. 15)**:

Demostración.

1. La implicación hacia la derecha ya ha sido probada en el punto 4 de la demostración del *Teorema 3.9* y es cierta aún sin estar bajo las condiciones de la distribución normal.

2. La implicación hacia la izquierda es fácil de comprobar. Veamos que si $\rho = 0$, la función de densidad conjunta expuesta en la ecuación (3.6) se podrá factorizar en el producto de las funciones de densidad marginales de X e Y. Así, sustituyendo $\rho = 0$ en dicha ecuación se obtiene

$$\begin{aligned}\phi(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y}\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} = \\ &= \frac{1}{2\pi\sigma_X\sigma_Y}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}\cdot\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}.\end{aligned}$$

   Agrupando los términos de un modo conveniente, llegamos a que

$$\begin{aligned}\phi(x,y) &= \left(\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\}\right)\cdot\left(\frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right) = \\ &= \phi_X(x)\cdot\phi_Y(y),\end{aligned}$$

   donde esta última igualdad viene dada por la *Proposición 3.16*.

   En virtud del *Corolario 1.3* del Capítulo 1, se tiene que X e Y son independientes.

□

***Teorema 3.32* (pág. 25)**:

Demostración. Sea l el número total de grupos con observaciones empatadas. Denotemos por p_k la última posición asignada antes de la aparición de un grupo de empates y por u_k el número de observaciones empatadas en el grupo k dentro de la muestra X, $k = 1, \ldots, l$. Se tiene lo siguiente para cada k:

Si estas u_k observaciones no estuvieran empatadas, se les asignarían los rangos usuales

$$p_k + 1,\ p_k + 2, \ldots,\ p_k + u_k,$$

y la suma de cuadrados de dichos rangos sería de la forma

$$S_N^k = \sum_{i=1}^{u_k}(p_k+i)^2 = u_k\left[p_k^2 + p_k(u_k+1) + \frac{(u_k+1)(2u_k+1)}{6}\right]. \tag{I.I}$$

Sin embargo, como hemos supuesto que dichas observaciones sí están empatadas, a todas se les asociará el mismo rango que vendrá dado por

$$\sum_{i=1}^{u_k} \frac{p_k + i}{u_k} = p_k + \frac{u_k + 1}{2},$$

y, análogamente, la suma de cuadrados para estos rangos empatados será

$$S_E^k = \sum_{i=1}^{u_k} \left(p_k + \frac{u_k + 1}{2} \right)^2 = u_k \left[p_k^2 + p_k(u_k + 1) + \frac{(u_k + 1)^2}{4} \right]. \tag{I.II}$$

De la diferencia entre las ecuaciones (I.I) y (I.II) se obtiene la corrección por empates para cada grupo k necesaria para modificar la suma de cuadrados de la ecuación (3.12), es decir,

$$S_N^k - S_E^k = \frac{u_k(u_k + 1)(2u_k + 1)}{6} - \frac{u_k(u_k + 1)^2}{4} = \frac{u_k(u_k^2 - 1)}{12},$$

y, por tanto,

$$\sum_{i=1}^{n} (R_i - \bar{R})^2 = \frac{n(n^2 - 1)}{12} - u,$$

donde $u = \sum_{k=1}^{l} \frac{u_k(u_k^2 - 1)}{12}$.

Análogamente, si denotamos por v la correspondiente suma sobre todos los grupos de empates en la muestra Y, obtendríamos que $\sum_{i=1}^{n} (S_i - \bar{S})^2 = \frac{n(n^2 - 1)}{12} - v$, dando lugar a la siguiente expresión para r_s:

$$r_s = \frac{12 \sum_{i=1}^{n} \left(R_i - \bar{R} \right) \left(S_i - \bar{S} \right)}{\left[(n(n^2 - 1) - 12u) \left(n(n^2 - 1) - 12v \right) \right]^{1/2}} = \frac{12 \left(\sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4} \right)}{\left[(n(n^2 - 1) - 12u) \left(n(n^2 - 1) - 12v \right) \right]^{1/2}}.$$

Una vez más, a partir de la ecuación (3.14), se puede obtener una expresión para r_s en función de D_i, pues

$$6 \sum_{i=1}^{n} D_i^2 = n(n^2 - 1) - 6(u + v) - 12 \sum_{i=1}^{n} \left(R_i - \bar{R} \right) \left(S_i - \bar{S} \right).$$

Y sustituyendo en lo anterior se llega al resultado deseado:

$$r_s = \frac{n(n^2 - 1) - 6 \sum_{i=1}^{n} D_i^2 - 6(u + v)}{\left[(n(n^2 - 1) - 12u) \left(n(n^2 - 1) - 12v \right) \right]^{1/2}},$$

□

***Teorema 3.38* (pág. 28)**:

Demostración. Supongamos que X e Y siguen una distribución normal bivariante con varianzas σ_X^2 y σ_Y^2 y coeficiente de correlación ρ. Entonces para dos pares independientes cualesquiera (X_i, Y_i) y (X_j, Y_j) de esta población, las diferencias

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

también siguen una distribución normal bivariante de media cero, varianza uno y covarianza igual a ρ. Por tanto, gracias a que ρ permanece invariante antes transformaciones lineales, $\rho(U, V) = \rho(X, Y)$.

Es claro que ahora p_c puede reescribirse en función de U y V como $p_c = \mathrm{P}(UV > 0)$. Calculando esta probabilidad de manera rigurosa, tenemos que

$$\begin{aligned} p_c &= \int_{-\infty}^{0}\int_{-\infty}^{0} \phi(x,y)dxdy + \int_{0}^{\infty}\int_{0}^{\infty} \phi(x,y)dxdy \\ &= 2\int_{-\infty}^{0}\int_{-\infty}^{0} \phi(x,y)dxdy = 2\Phi(0,0)\ , \end{aligned}$$

donde ϕ y Φ denotan la función de densidad y la función de distribución conjuntas, respectivamente, de una distribución normal bivariante estandarizada. Dado que se puede demostrar que $\Phi(0,0) = \dfrac{1}{4} + \dfrac{1}{2\pi}\arcsin\rho$, para la normal bivariante se obtiene finalmente que

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

Como estamos en las condiciones que nos garantizan que se cumple la ecuación (3.18), sustituyendo lo anterior tendremos que

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

□

***Teorema 4.23* (pág. 40)**:

Demostración. Sean X e Y dos vectores que siguen una distribución normal estándar con coeficiente de correlación ρ.

1. Por las propiedades de la distribución normal se tiene que

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

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

$$= \rho^2 G(\rho),$$

donde $G(\rho) = \sum_{k=1}^{\infty} \frac{2^{2k}-2}{(2k)!}\rho^{2(k-1)} \int_{\mathbb{R}^2} e^{-t^2-s^2}(ts)^{2(k-1)}\,dtds$ es una suma de términos no negativos no decreciente en ρ. Puesto que $F(\rho) = \rho^2 G(\rho)$ y $G(\rho) \leq G(1)$, se llega a que

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

Observación I.II. La igualdad se alcanza cuando $\rho = \pm 1$.

2. Es claro que $F(0) = F'(0) = 0$, por lo que $F(\rho) = \int_0^\rho \int_0^x F''(z)\,dzdx$. La segunda derivada de F es

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

donde

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

Aplicando ahora un cambio de variable usando el hecho de que los autovalores de la forma cuadrática $t^2 - s^2 - 2zts$ son $1 \pm z$ y $\int_{-\infty}^{\infty} e^{-t^2\lambda}\,dt = \sqrt{\frac{\pi}{\lambda}}$. Entonces

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

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

Sustituyendo en la expresión anterior $\rho = 1$, se obtiene que

$$F(1) = 4\pi\left(\frac{\pi}{2} - \frac{\pi}{6} - \sqrt{3} + 1\right) = 4\pi\left(1 + \frac{\pi}{3} - \sqrt{3}\right),$$

y usando la relación de la ecuación (4.4) se obtiene el resultado que queríamos probar en 2.

3. En la prueba de 1 se tiene que $\frac{\mathcal{R}}{|\rho|}$ es una función no decreciente de $|\rho|$. Aplicando este resultado junto con el apartado 2, llegamos a que

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

□

***Teorema 4.32* (pág. 47)**:

Demostración. Dado que las variables aleatorias son continuas, $\mathrm{P}[(X - X')(Y - Y') < 0] = 1 - \mathrm{P}[(X - X')(Y - Y') > 0]$ y, en consecuencia,

$$\tau = 2\,\mathrm{P}[(X - X')(Y - Y') > 0] - 1. \tag{I.III}$$

Por otra parte, se tiene que $\mathrm{P}[(X - X')(Y - Y') > 0] = \mathrm{P}[(X > X'), (Y > Y')] + \mathrm{P}[(X < X'), (Y < Y')]$, y estas probabilidades se pueden evaluar integrando sobre la distribución de uno de los vectores (X, Y) o (X', Y'). Tomando, por ejemplo, (X, Y), se llega a que

$$\begin{aligned}\mathrm{P}[(X > X'), (Y > Y')] &= \mathrm{P}[(X' < X), (Y' < Y)] = \\ &= \int\int_{\mathbb{R}^2} \mathrm{P}[(X' < x), (Y' < x)] dC_1(F_X(x), F_Y(y)) = \\ &= \int\int_{\mathbb{R}^2} C_2(F_X(x), F_Y(y)) dC_1(F_X(x), F_Y(y)).\end{aligned}$$

Empleando ahora las transformaciones de probabilidad $u = F_X(x)$ y $v = F_Y(y)$, obtendremos

$$\mathrm{P}[(X > X'), (Y > Y')] = \int\int_{\mathbb{I}^2} C_2(u, v) dC_1(u, v).$$

Análogamente,

$$\begin{aligned}\mathrm{P}[(X < X'), (Y < Y')] &= \int\int_{\mathbb{R}^2} \mathrm{P}[(X' > x), (Y' > x)] dC_1(F_X(x), F_Y(y)) = \\ &= \int\int_{\mathbb{R}^2} \left[1 - F_X(x) - F_Y(y) + C_2(F_X(x), F_Y(y))\right] dC_1(F_X(x), F_Y(y)) = \\ &= \int\int_{\mathbb{I}^2} \left[1 - u - v + C_2(u, v)\right] dC_1(u, v).\end{aligned}$$

Pero dado que C_1 es la función de distribución conjunta de un par (U, V) de variables aleatorias que siguen una uniforme $U(0, 1)$, $\mathrm{E}[U] = \mathrm{E}[V] = \dfrac{1}{2}$ y, en consecuencia,

$$\mathrm{P}[(X < X'), (Y < Y')] = 1 - \frac{1}{2} - \frac{1}{2} + \int\int_{\mathbb{I}^2} C_2(u, v) dC_1(u, v) = \int\int_{\mathbb{I}^2} C_2(u, v) dC_1(u, v).$$

Finalmente, se tiene que

$$\mathrm{P}[(X - X')(Y - Y') > 0] = 2 \int\int_{\mathbb{I}^2} C_2(u, v) dC_1(u, v),$$

y concluimos sustituyendo esto último en la ecuación (I.III). □

Anexo II

Código R

```r
### Lectura y analisis de los datos
library(kmed)
datos <- na.omit(heart[,c(3,1,4,5,8,10)]); attach(datos)
dim(datos)
head(datos)

summary(datos) # Resumen de las caracteristicas de las variables
class(datos) # Tipo de formato de los datos. En este caso, data.frame
lapply(datos,class) # Se obtiene la clase de cada una de las variables
# que conforman el fichero
isnum = unlist(lapply(datos,is.numeric)); isnum
which(isnum == FALSE) # Muestra las variables que no son numericas las
# cuales vamos a descartar para los proximos calculos

dat <- datos[,isnum]
rownames(dat) <- NULL; head(dat)

## Grafico de pares
pairs(dat, pch = 19, cex = 0.4, col = 4, font.labels = 2)
```

Figura 5.1

```r
## Verificar normalidad bivariante
library(MVN)
pnorm = matrix(NA, nrow = ncol(dat), ncol = ncol(dat))
rownames(pnorm) <- names(dat)
colnames(pnorm) <- names(dat)
for (i in 2:ncol(dat)){
```

```
    for (j in 1:(i-1)){
      pnorm[i,j] = mvn(data = dat[,c(i,j)])$multivariateNormality$p
    }
  }
  round(pnorm,3)
```

Tabla 5.1

```
  ### Matriz de correlaciones de Pearson
  r = cor(dat)
  round(r,2)

  ## Grafico relativo a r
  library(correlation)
  plot(visualisation_recipe(correlation(dat)))
```

Figura 5.2

```
  ## Analisis sin datos atipicos
  # Deteccion de outliers y eliminacion en el triangulo superior
  dat1 <- dat
  for (i in 1:(ncol(dat)-1)) {
    for (j in (i+1):ncol(dat)) {
      outliers <- mvn(dat[,c(i,j)], mvnTest = "hz",
          multivariateOutlierMethod = "quan", showOutliers = TRUE,
          showNewData = TRUE)$multivariateOutliers$Observation
      dat1 <- dat1[-as.integer(outliers),]
    }
  }
  dim(dat1)
  head(dat1)
  pairs(dat1)

  # Verificar normalidad bivariante
  pnorm1 = matrix(NA, nrow = ncol(dat1), ncol = ncol(dat1))
  rownames(pnorm1) <- names(dat1)
  colnames(pnorm1) <- names(dat1)
  for (i in 2:ncol(dat1)){
    for (j in 1:(i-1)){
      pnorm1[i,j] = mvn(data = dat1[,c(i,j)])$multivariateNormality$p
    }
  }
```

```
round(pnorm1,3)

# Matriz de correlacion de Pearson
r1 = cor(dat1)
round(r1,2)

# Porcentaje de informacion perdida
100 - dim(dat1)[1]/dim(dat)[1]*100

## Analisis de la presencia de grupos
library(GGally)
library(ggplot2)
ggpairs(datos, columns = 2:6, aes(color = cp, alpha = 1),
        upper = list(continuous = wrap("cor", size = 2.5)))
```

Figura 5.3

```
# Tamano de la muestra en cada grupo
table(cp)

### Matrices de correlaciones de Spearman, Kendall y distancias
## Matriz de correlaciones de Spearman
R = cor(dat, method = "spearman")
round(R,2)

# Grafico relativo a R
library(correlation)
rez <- correlation(datos, method = "spearman")
x <- cor_sort(as.matrix(rez))
layers <- visualisation_recipe(x)
plot(layers)
```

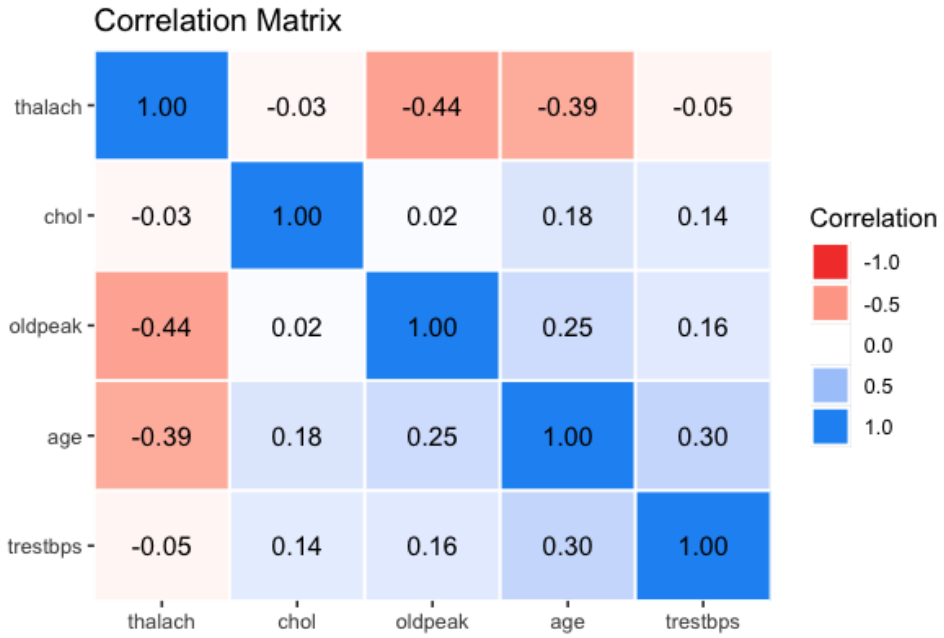


Figura II.I: Matriz de correlaciones de Spearman.

```r
# Relacion entre Spearman y Pearson cuando los datos siguen una normal
rho = cor(age,chol) # Variables normales al 1%
R[1,3]; 6/pi*asin(rho/2)
rho = cor(age,oldpeak) # Variables no normales
R[1,5]; 6/pi*asin(rho/2)

## Matriz de correlaciones de Kendall
t = cor(dat, method = "kendall")
round(t,2)

# Grafico relativo a t
library(RColorBrewer)
library(corrplot)
library(ellipse)
color <- brewer.pal(8, "Accent") <- colorRampPalette(color)(110)
ord <- order(t[1, ]); data_ord <- t[ord, ord]
plotcorr(data_ord, col = color[data_ord*50+50], mar = c(1,1,1,1))
```

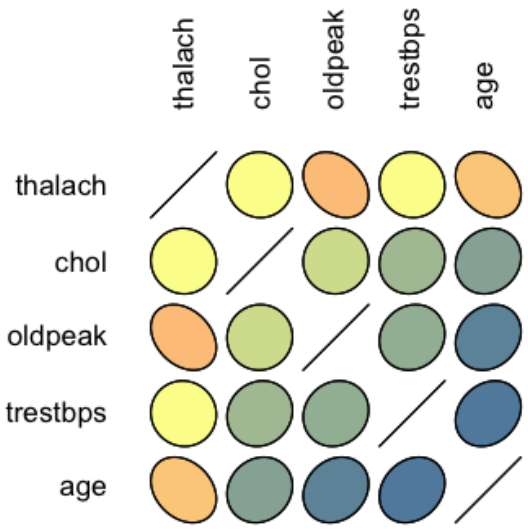


Figura II.II: Matriz de correlaciones de Kendall.

```r
# Relacion entre Kendall y Pearson cuando los datos siguen una normal
rho = cor(age,chol) # Variables normales al 1%
t[1,3]; 2/pi*asin(rho)
rho = cor(age,oldpeak) # Variables no normales
t[1,5]; 2/pi*asin(rho)

## Matriz de correlaciones de distancias
library(energy)
d = matrix(0, nrow = ncol(dat), ncol = ncol(dat))
rownames(d) <- names(dat)
colnames(d) <- names(dat)
for (i in 2:ncol(dat)){
  for (j in 1:(i-1)){
    d[1,1] = dcor(dat[,1],dat[,1])
    d[i,i] = dcor(dat[,i],dat[,i])
    d[i,j] = dcor(dat[,i],dat[,j])
    d[j,i] = dcor(dat[,j],dat[,i])
  }
}
round(d,2)

# Grafico relativo a d
library(RColorBrewer)
library(corrplot)
```

```
25  corrplot(d, method = "color", cl.pos = 'n', type = "lower", tl.srt =
        45)
```

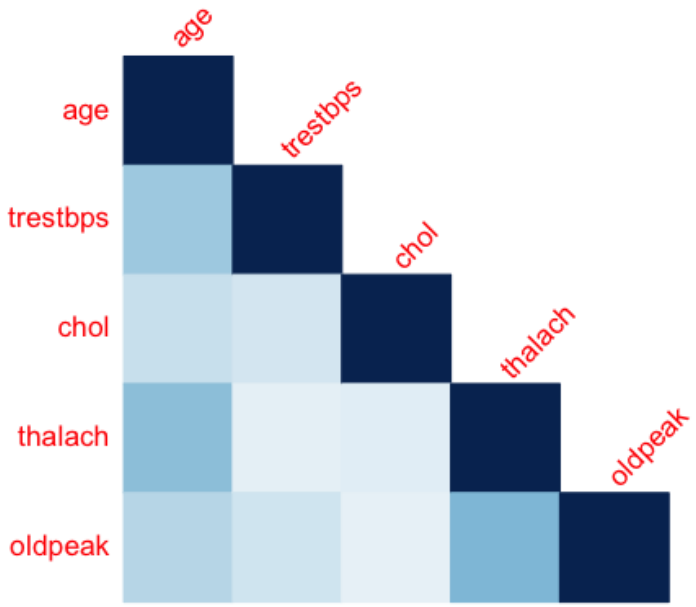


Figura II.III: Matriz de correlaciones de distancias.

```
1  # Relacion entre distancias y Pearson cuando los datos siguen una normal
2   rho = cor(age,chol) # Variables normales al 1%
3   num <- rho*asin(rho) + sqrt(1-(rho^2)) - rho*asin(rho/2) - sqrt(4-(rho
        ^2)) + 1
4   den <- 1 + pi/3 - sqrt(3)
5   d[1,3]; sqrt(num/den)
6   rho = cor(age,oldpeak) # Variables no normales
7   num <- rho*asin(rho) + sqrt(1-(rho^2)) - rho*asin(rho/2) - sqrt(4-(rho
        ^2)) + 1
8   den <- 1 + pi/3 - sqrt(3)
9   d[1,5]; sqrt(num/den)
10
11  ### Graficos para facilitar la comparacion de las correlaciones
12  library(RColorBrewer)
13  library(corrplot)
14  par(mfrow = c(2,2))
15
16  ## Pearson
17  corrplot(r, col = brewer.pal(n = 8, name = "RdYlBu"), method = "circle"
        , tl.pos = "d", tl.cex = 1, tl.col = "red")
```

```
18  corrplot(r, add = TRUE, type = "upper", method = "number", diag = FALSE
        , tl.pos = "n", cl.pos = "n")
19
20
21  ## Spearman
22  corrplot(R, col = brewer.pal(n = 8, name = "RdYlBu"), method = "circle"
        , tl.pos = "d", tl.cex = 1, tl.col = "red")
23  corrplot(R, add = TRUE, type = "upper", method = "number", diag = FALSE
        , tl.pos = "n", cl.pos = "n")
24
25  ## Kendall
26  corrplot(t, col = brewer.pal(n = 8, name = "RdYlBu"), method = "circle"
        , tl.pos = "d", tl.cex = 1, tl.col = "red")
27  corrplot(t, add = TRUE, type = "upper", method = "number", diag = FALSE
        , tl.pos = "n", cl.pos = "n")
28
29  ## Distancias
30  corrplot(d, col = brewer.pal(n = 8, name = "RdYlBu"), method = "circle"
        , tl.pos = "d", tl.cex = 1, tl.col = "red")
31  corrplot(d, add = TRUE, type = "upper", method = "number", diag = FALSE
        , tl.pos = "n", cl.pos = "n")
32
33  par(mfrow = c(1,1))
```

Figura 5.4

Bibliografía

Balakrishnan, N., & Lai, C. D. (2009). *Continuous bivariate distributions*. Springer Science+Business Media. https://doi.org/10.1007/b101765

Bertsekas, D., & Tsitsiklis, J. N. (2008). *Introduction to probability* (Vol. 1). Athena Scientific.

Budiaji, W. (2022). *kmed: Distance-Based k-Medoids* [R package version 0.4.2]. https://CRAN.R-project.org/package=kmed

Camacho, C., López, A., & Arias, M. (2006). Regresión lineal simple. https://personal.us.es/vararey/regresion-simple.pdf

Castañeda, L. B., Arunachalam, V., & Dharmaraja, S. (2012). *Introduction to probability and stochastic processes with applications*. John Wiley & Sons.

Corral, N. (2023). *Modelización Estadística: Apuntes sobre Vectores*. Consultado en 2024, desde https://bellman.ciencias.uniovi.es/norberto/ModelizacionEstadistica/Apuntes/vectores.pdf

DeGroot, M., & Schervish, M. (2013). *Probability and Statistics*. Pearson Education. https://books.google.es/books?id=hIPkngEACAAJ

Díaz, W. (2013). Contribuciones a la dependencia y dimensionalidad en cópulas.

Espartero. (2012). Tema 5: Estimación puntual I. Propiedades de los estimadores. Consultado en 2024, desde https://www5.uva.es/espartero/Tema5_2012.pdf

Gibbons, J. D., & Chakraborti, S. (2003). Nonparametric statistical inference fourth edition, revised and expanded. *Statistics textbooks and monographs*, *168*.

Hernández, J. A. M., & Caita, L. A. M. (2017). Una falacia en probabilidad ilustrada vía teoría de cópulas. *Comunicaciones en Estadística*, *10*(2), 281-295.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease. https://doi.org/https://doi.org/10.24432/C52P4X

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, *6*(2), 151-162. https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

Lalinde, J. D. H., Castro, F. E., Rodríguez, J. E., Rangel, J. G. C., Sierra, C. A. T., Torrado, M. K. A., Sierra, S. M. C., & Pirela, V. J. B. (2018). Sobre el uso adecuado del coeficiente de correlación

de Pearson: definición, propiedades y suposiciones. *Archivos venezolanos de Farmacología y Terapéutica*, *37*(5), 587-595.

Lancaster, H. (2004). Dependence, measures and indices of. *Encyclopedia of statistical sciences*.

Lloréns, L. L. (s.f.). *Cópulas: su concepto y utilidad en la medición de riesgos*. Consultado en 2024, desde https://tinyurl.com/44dm4cr5

Lyons, R. (2013). Distance covariance in metric spaces.

Makowski, D., Wiernik, B. M., Patil, I., Lüdecke, D., & Ben-Shachar, M. S. (2022, octubre). correlation: Methods for Correlation Analysis [Version 0.8.3]. https://CRAN.R-project.org/package=correlation

Morales, P., & Rodríguez, L. (2016). Aplicación de los coeficientes correlación de Kendall y Spearman. *Agrollanía*, *13*.

Moran, P. (1948). Rank correlation and product-moment correlation. *Biometrika*, *35*(1/2), 203-206.

Murdoch, D., & Chow, E. D. (2023). *ellipse: Functions for Drawing Ellipses and Ellipse-Like Confidence Regions* [R package version 0.5.0]. https://CRAN.R-project.org/package=ellipse

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes* [R package version 1.1-3]. https://CRAN.R-project.org/package=RColorBrewer

Ochoa, C., & Molina, M. (2018). Estadística. Tipos de variables. Escalas de medida. *Evid Pediatr*, *14*(29), 1-5.

Ortega, J. (2014). Material Didáctico: Capítulo 4. Consultado en 2024, desde https://www.cimat.mx/~jortega/MaterialDidactico/Prope2014/Cap4.pdf

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25-45. https://doi.org/10.2307/2331722

Pinilla, J. O., & Rico, A. F. O. (2021). ¿ Pearson y Spearman, coeficientes intercambiables? *Comunicaciones en estadística*, *14*(1), 53-63.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, *10*(3-4), 441-451.

Ríus Díaz, F., Barón López, F. J., Sánchez Font, E., & Parras Guijosa, L. (2012). Bioestadística: métodos y aplicaciones [Versión electrónica. Material de apoyo (node39.htm)]. Consultado en 2024, desde https://virtual.uptc.edu.co/ova/estadistica/docs/libros/ftp.bioestadistica.uma.es/libro/node39.htm

Rizzo, M., & Szekely, G. (2022). *energy: E-Statistics: Multivariate Inference via the Energy of Data* [R package version 1.7-11]. https://CRAN.R-project.org/package=energy

RRL. (s.f.). Find the correlation coefficient ρ_g of $(G(x), H(y))$ [Version: 2018-01-04]. Consultado en 2024, desde https://math.stackexchange.com/q/2550138

Samuel, M. D. D. K., Mari, D. D., & Kotz, S. (2001). *Correlation and dependence*. World Scientific.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2024). *GGally: Extension to 'ggplot2'* [R package version 2.2.1]. https://CRAN.R-project.org/package=GGally

Shieh, G. (2010). Estimation of the simple correlation coefficient. *Behavior Research Methods*, *42*(4), 906-917.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677-680.

Sulbarán, D. (2012). Análisis bivariado de datos: un resumen para el curso de Estadística II.

Székely, G. J., & Rizzo, M. L. (2009). Brownian distance covariance. *The annals of applied statistics*, 1236-1265.

Székely, G. J., & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, *143*(8), 1249-1272.

Székely, G. J., & Rizzo, M. L. (2014). Partial distance correlation with methods for dissimilarities.

Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769-2794. https://doi.org/10.1214/009053607000000505

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Wei, T., & Simko, V. (2021). *R package 'corrplot': Visualization of a Correlation Matrix* [(Version 0.92)]. https://github.com/taiyun/corrplot

Weisstein, E. W. (2024). *Bivariate Normal Distribution* [From MathWorld–A Wolfram Web Resource]. Consultado en 2024, desde https://mathworld.wolfram.com/BivariateNormalDistribution.html

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Yu, H., & Hutson, A. D. (2024). A robust Spearman correlation coefficient permutation test. *Communications in Statistics-Theory and Methods*, *53*(6), 2141-2153.

Žežula, I. (2009). On multivariate Gaussian copulas. *Journal of Statistical Planning and Inference*, *139*(11), 3942-3946. https://doi.org/10.1016/j.jspi.2009.05.039