Sobre el uso adecuado del coeficiente de correlación de Pearson: verificación de supuestos mediante un ejemplo aplicado a las ciencias de la salud

On the proper use of the Pearson correlation coefficient: checking assumptions through an example applied to health sciences

Juan Hernández-Lalinde, Mg.[1]; https://orcid.org/0000-0001-6768-1873, Jhon-Franklin Espinosa-Castro, Mg.[1]; https://orcid.org/0000-0003-2186-3000, Mariana-Elena Penaloza-Tarazona, Dr.[1]; https://orcid.org/0000-0002-3863-0580, Édgar Díaz-Camargo, Dr. (c)[2]; https://orcid.org/0000-0002-7349-3059, María Bautista-Sandoval, Dr. (c)[2]; https://orcid.org/0000-0001-6259-2812, Manuel E. Riaño-Garzón, Dr. (c)[2]; https://orcid.org/0000-0002-4476-9538, Oriana M. Chacón Lizarazo, Ps.[3]; https://orcid.org/0000-0003-0292-9713, Yudy Karina Chaparro-Suárez, Ps.[3]; https://orcid.org/0000-0003-4098-8925, Diego García Álvarez, Mg.[4]; https://orcid.org/0000-0002-9350-785X, Valmore Bermúdez-Pirela, Dr.[2]; https://orcid.org/0000-0003-1880-8887*

[1]Universidad Simón Bolívar, Departamento de Ciencias Sociales y Humanas, Cúcuta, Colombia.

[2]Universidad Simón Bolívar, Facultad de Ciencias Jurídicas y Sociales, Barranquilla, Colombia.

[3]Universidad Simón Bolívar, Semillero Psicoex, Facultad de Ciencias Jurídicas y Sociales, Cúcuta, Colombia.

[4]Instituto Universitario San Francisco de Asís. Punta del Este, Uruguay.

**Autor de correspondencia: Juan Hernández-Lalinde. Universidad Simón Bolívar, Departamento de Ciencias Sociales y Humanas. Calle 14 entre avenidas 4 y 5, Barrio La Playa. C. P.: 540006. Cúcuta, Colombia. Correo electrónico: j.hernandezl@unisimonbolivar.edu.co.*

Resumen

La comprobación de los supuestos en los que se sustenta el uso del coeficiente de correlación de *Pearson* suele ser una tarea en la que se cometen no pocos errores. Si bien es sencillo el proceso que lleva a su cálculo e interpretación, no resulta tan fácil la labor de verificar el cumplimiento de condiciones como la normalidad bivariada o la ausencia de datos atípicos, probablemente porque esto demanda la implementación de técnicas multivariantes. La presente revisión pretende servir de orientación a investigadores de las ciencias de salud y afines, los cuales seguramente se toparán con situaciones en las que deba emplearse esta herramienta estadística. El artículo gira en torno a un caso práctico derivado de un estudio de prevalencia de síndrome metabólico realizado en la ciudad de Maracaibo, Venezuela. El objetivo principal es el de mostrar mediante este ejemplo la manera adecuada de constatar las premisas vinculadas a este coeficiente, no sin olvidar el debido argumento teórico que las respalda. Se prescinde del aspecto matemático en favor del informático, para lo cual se utiliza el programa abierto *R-Studio* en todas y cada una de las actividades de procesamiento, diagramación y cómputo. Se proveen las bases de datos empleadas en el desarrollo del problema, a la vez de suministrar los *scripts* que activan las funciones del paquete con el propósito de que el lector pueda reproducir el análisis y comparar los resultados. Toda esta información puede ser consultada y descargada desde un repositorio de libre acceso.

Palabras clave: coeficiente de correlación, *Pearson*, supuestos, *R-Studio*, caso práctico, síndrome metabólico, Maracaibo.

Abstract

The checking of the assumptions on which the use of the Pearson correlation coefficient is based, is usually a task in which many errors are committed. Although the process that leads to its calculation and interpretation is simple, the task of verifying conditions such as bivariate normality or the absence of outlier is not so easy, probably because this requires the implementation of multivariate techniques. This review intends to serve as guidance to health sciences researchers, who will surely find situations in which this statistical tool should be used. The article is based on a prevalence study of metabolic syndrome carried out in the Maracaibo city, Venezuela. The main objective is to show by this example the appropriate way to verify the assumptions linked to this coefficient, not forgetting the due theoretical argument that supports them. The mathematical aspect is discarded in order to get the benefits of using computers power, for which the open source R-Studio program is used in each and every one of the processing, plotting and computation activities. The dataset used in the development of the problem are provided, as well as the scripts that activate the functions of the package with the purpose that the reader can reproduce the analysis and compare the results. All this information can be consulted and downloaded from an open access repository.

Keywords: correlation coefficient, Pearson, assumptions, R-Studio, practical case, metabolic, syndrome, Maracaibo.

Introducción

El coeficiente de correlación de *Pearson* es una medida ampliamente utilizada en diversas áreas de la investigación; desde la ciencia de los alimentos, en la que es usado para explorar la relación entre el oscurecimiento enzimático y la temperatura de productos como la pera, manzana o setas[1]; hasta campos más recientes como el aprendizaje automatizado, en el que sirve como soporte para mejorar la seguridad informática ante posibles ataques de piratas cibernéticos[2]. La medicina no es la excepción; de hecho, es quizá una de las ramas que más provecho ha obtenido de la implementación de esta herramienta junto a disciplinas similares como la psicología, odontología, enfermería o bioanálisis. Publicaciones como la de *He et al*.[3], en la que se examina la asociación entre la obesidad abdominal y la carga de síndrome metabólico en adolescentes de la Universidad de Pensilvania, así como la de *Korkmazer* y *Solak*[4], en la que los hallazgos sugieren una fuerte correlación entre la diabetes mellitus gestacional y los niveles de suero materno TNF-α, son solo algunos ejemplos en los que se demuestra la utilidad de este estadístico.

Contrario a lo que podría pensarse, la extensa popularidad del coeficiente *R* de *Pearson* trae consigo un uso que, en muchas ocasiones, es indebido. Resulta más común de lo que cabría esperar, que esta medida sea empleada en variables nominales u ordinales cuando su utilización está restringida a características de intervalo o de razón. Más frecuentes son los casos en los que se omite la evaluación del supuesto de normalidad bivariada por una revisión de la distribución marginal de cada variable, incurriendo así en una equivocación que podría invalidar los hallazgos del estudio. Premisas como la de la ausencia de datos atípicos multivariados pueden llegar a ser más complicadas de evaluar, toda vez que requieren de técnicas estadísticas más complejas; sin mencionar los problemas ligados al muestreo, que incluso dependen de que sea ejecutada una rigurosa planificación en las fases iniciales de la investigación para evitarlos.

Tomando en cuenta este escenario, se plantea la presente revisión, cuyo propósito es el de servir como guía a profesionales e investigadores que se desempeñen dentro del campo de la medicina y ciencias afines. La idea básica es la de proporcionar un documento que ilustre de forma práctica cómo verificar adecuadamente las suposiciones que validan el uso de este coeficiente. Para esto, se parte de un contexto investigativo auténtico en el que se estudia la prevalencia de síndrome metabólico, caso que a su vez sirve como plataforma para desarrollar todos los análisis del artículo. La fundamentación matemática se descarta y se reemplaza por un lenguaje intuitivo; se da protagonismo a la disponibilidad de programas de código abierto como *R-Studio* y se da acceso a las bases de datos empleadas en la revisión.

Descripción del problema

El problema que se utiliza como ejemplo en la presente revisión, se basa en el estudio del síndrome metabólico (SM) desarrollado en la ciudad de Maracaibo, Venezuela, del cual se han obtenido diversas publicaciones interesantes[5-7]. El objetivo principal de dicha investigación fue el de determinar la prevalencia de SM en la mencionada ciudad, basándose para ello en los criterios que la *International Diabetes Federation* establecía en 2005 (IDF-2005) y 2009 (IDF-2009), así como también en las nociones sugeridas por la *Adult Treatment Panel* en su tercera edición de 2005 (ATPIII-2005). Para alcanzar los objetivos de dicho estudio se llevó a cabo una investigación descriptiva, de corte transaccional, en la cual se seleccionaron 2230 sujetos mediante muestreo aleatorio polietápico[8]. Para profundizar en los detalles metodológicos, consúltese el artículo titulado *The Maracaibo city metabolic syndrome prevalence study: design and scope*[9].

Con base en estos antecedentes, se propone acá realizar la comprobación de los supuestos de un análisis de correlación a partir de variables como la resistencia a la insulina, el índice de masa corporal y la razón triglicéridos-colesterol de un grupo de sujetos seleccionados según muestreo aleatorio simple sin reposición (MASSR). Para esto, se extrajo de la base de datos original la subpoblación integrada por individuos con edad comprendida desde los 20 hasta los 49 años, de raza mestiza y cuya condición socioeconómica coincidiera con estratos de clase media. El motivo que impulsó esta delimitación se ciñe a la idea de obtener una población aproximadamente homogénea que permitiera realizar el muestreo ya señalado. Por otro lado, la escogencia de las variables obedece únicamente a intereses didácticos e ilustrativos, y no sugiere que exista o no relación entre tales características.

Para ahondar en el tema de las variables seleccionadas, se suministra en este momento una breve descripción de sus propiedades más importantes con el propósito de ubicar al lector no familiarizado dentro del contexto del artículo. La resistencia a la insulina se ha evaluado a través del *Homeostatic Model Assessment* en su segunda versión (HOMA2). El HOMA es un índice adimensional ampliamente validado, empleado para medir la resistencia a la insulina a partir de la glicemia en ayunas. Descrito por primera vez en 1985 por *Matthews*, *Hosker*, *Rudenski*, *Naylor*, *Treacher* y *Turner*[10], ha sufrido actualizaciones recientes que han derivado en la versión ya mencionada, la cual provee mediciones más precisas[11,12]. Por otro lado, el índice de masa corporal (IMC) resulta de dividir el peso del sujeto, expresado en kilogramos; sobre la altura o talla al cuadrado, medida en metros. Fue caracterizado por *Adolphe Quetelet* hacia mediados del siglo XIX cuando acuñó la frase *el crecimiento transversal del hombre es mayor que el vertical*; sin embargo, no fue hasta 1972 que *Ancel Keys* lo empleó por primera vez bajo el nombre con el que es conocido hoy en día[13,14]. Finalmente, la relación triacilglicéridos-colesterol (TAG/HDL) se obtiene de resultados de laboratorio en ayunas, al calcular el cociente entre la concentración de triglicéridos séricos y el valor del llamado *colesterol bueno* o *high-density lipoprotein cholesterol*, ambos expresados en miligramos sobre decilitros (mg/dL)[15-17].

Descripción de la base de datos usada como población

La población objetivo —definida a partir de la depuración anterior— quedó conformada por un total de $N = 478$ individuos, de los cuales, el 44.77 % (n=214) eran mujeres y el

55.23 % (*n*=264) eran hombres. La distribución de los grupos etarios fue la siguiente: el 39.33 % (*n*=188) de los participantes tenían entre 20 y 29 años, inclusive; el 28.45 % (*n*=136) de los analizados reportaban una edad localizada en el intervalo que va desde los 30 hasta los 39 años; y el 32.22 % (*n*=154) de los consultados registró edades que oscilaron desde los 40 hasta los 49 años.

Ahora bien, en lo que respecta a las variables de interés en la población, se halló una resistencia a la insulina de 2.24 ± 1.54 (CV = 60.63 %), con mínimo de 0.30 y máximo de 12.50. El índice de masa corporal fluctuó entre 14.22 kg/m^2 y 67.58 kg/m^2, con valores de 28.74 ± 6.53 kg/m^2 (CV = 22.72 %), mientras que la relación TAG/HDL exhibió cifras de 3.54 ± 3.99 (CV = 112.71 %), oscilando desde 0.34 hasta 49.05.

Descripción de la base de datos usada como muestra

Como ya se ha mencionado, la muestra fue extraída de la base de datos integrada por $N = 478$ sujetos, utilizando para ello algoritmos que permiten reproducir lo que sucede en un MASSR. El tamaño de la muestra fue de $n = 30 n = 30$ individuos, esto con la intención de mantener la fracción de muestreo por debajo del 10 % (n/N=30/478=6.28%)y omitir el factor de corrección de población finita, además de contar con un número que fuera lo suficientemente grande como para aplicar las condiciones del teorema de límite central[18,19]. Así pues, la proporción según sexo fue la siguiente: 63.33 % (*n*=19) fueron mujeres, en tanto que 36.67 % (*n*=11) fueron hombres. Del total de seleccionados, el 30.00 % (*n*=9) tenían edades comprendidas entre 20 y 29 años, inclusive; el 33.33 % (*n*=10) reportaron valores que fluctuaron desde los 30 hasta los 39 años; y el 36.67 % (*n*=11) se localizó en el intervalo que va desde los 40 hasta los 49 años.

En lo referente a las variables cuantitativas, se encontró una resistencia a la insulina cuyos valores promediaron los 2.55 ± 1.95 (CV = 76.47 %), con mínimos y máximos de 0.50 y 10.30, respectivamente. En cuanto al índice de masa corporal, las cifras redondearon los 30.71 ± 6.20 kg/m^2 (CV = 20.19 %), abarcando un rango comprendido entre 17.29 kg/m^2 y 43.53 kg/m^2. Finalmente, la razón TAG/HDL osciló desde 0.80 hasta 10.89, con media y desviación estándar de 3.28 ± 2.60 (CV = 79.27 %).

Con la intención de promover el uso de recursos compartidos, se anexa acá el enlace desde donde se podrán descargar las bases de datos antes descritas: **http://dx.doi.org/10.17632/x9tghxpdc3.1**. Además, se presentan en la **tabla 1** los descriptivos de los dos conjuntos con el objeto de validar la representatividad de la muestra.

Tabla 1

Variables	En la población		
	Parámetros	Estimadores puntuales	ICB 95 %
Sexo			
Masculino	264 (55.23 %)	19 (63.33 %)	45.71 % - 78.72 %
Femenino	214 (44.77 %)	11 (36.67 %)	21.28 % - 54.49 %
Grupos etarios			
20 – 29	188 (39.33 %)	9 (30.00 %)	16.00 % - 47.65 %
30 – 39	136 (28.45 %)	10 (33.33 %)	18.60 % - 51.11 %
40 – 49	154 (32.22 %)	11 (36.67 %)	21.28 % - 54.49 %
IMC (en kg/m^2)	28.74 ± 6.53	30.71 ± 6.20	28.40 – 33.03
TAG/HDL (adimensional)	3.54 ± 3.99	3.28 ± 2.60	2.31 – 4.25
HOMA2IR (adimensional)	2.24 ± 1.54	2.55 ± 1.95	1.82 – 3.28

Tabla 1. Descripción de las bases de datos usadas como población y como muestra. Se presentan descriptivos básicos para caracterizar la población. Asimismo, se ofrecen las estimaciones puntuales y por intervalos obtenidas de la muestra. Nótese que todos los intervalos bilaterales (ICB) contienen al parámetro que estiman bajo un nivel de confianza de 95 %.

Descripción de los paquetes de *R-Studio* utilizados

Para reproducir el análisis que se ofrece en esta publicación, será necesario instalar una versión de *R* que sea igual o superior a la 3.5.1, además de una edición de *R-Studio* que sea igual o mayor a la 1.1.456. Vale la pena aclarar que, aunque todas las tareas de procesamiento hayan sido efectuadas con *R-Studio*, es necesario tener el programa original instalado para que este pueda funcionar.

Con relación a aquellos paquetes adicionales a los que el *software* incorpora por defecto, se requerirán los siguientes: *MVN* (*multivariate normality tests)*, mismo que servirá para implementar pruebas de normalidad multivariada como las de *Mardia*, *Royston*, *Henze-Zirkler*, entre otras; *mvoutlier* (*multivariate outlier detection based on robust methods*), el cual será empleado al momento de detectar datos atípicos mediante las distancias robustas de *Mahalanobis*; y *randtests* (*testing randomness in R*), que servirá para implementar contrastes de aleatoriedad o independencia como el de *Wald-Wolfowitz* o el de *Bartel*. Todas estas extensiones pueden descargarse con facilidad desde la consola, únicamente ejecutando el comando *install.packages* ("*nombre del programa*").

Por último, se suministran también los *scripts* de *R-Studio* con el propósito de que estos sean usados en caso de repetir o verificar los resultados. Con ello, además, se eliminarán en las próximas secciones del artículo aquellas explicaciones innecesarias acerca del aspecto informático que podrían confundir al lector, facilitando así las tareas de interpretación. Se han preparado seis archivos de *R-Studio* en los que se exponen los siguientes procedimientos: análisis de correlación sin depurar los datos, verificación de la normalidad bivariada antes de la depuración, verificación de la existencia de datos atípicos, verificación de la normalidad bivariada luego de la depuración, análisis de correlación con los datos depurados y verificación del supuesto de independencia luego de

la depuración. Toda esta información puede encontrarse en: **http://dx.doi.org/10.17632/tncnnkcthr.1**.

Análisis de correlación mediante el coeficiente *R* de *Pearson*

A diferencia de lo que se logra con un análisis de regresión lineal, en el que se modela la relación entre una variable dependiente y una o varias independientes mediante la ecuación de una recta[19-24], en un análisis de correlación solo se estudia la *magnitud*, el *sentido* y la *significación* de la asociación lineal entre al menos dos variables[25-29]. Mientras que en la regresión se plantea que una de las características será explicada por y a través de un conjunto de predictores, en la correlación no se asume ningún tipo de direccionalidad entre las variables que se vinculan, propiedad que es conocida como *simetría*[28,29]. En consecuencia, un análisis de este tipo estará limitado a calcular el coeficiente *R* de *Pearson*, interpretar correctamente su magnitud y sentido, generar la prueba de hipótesis correspondiente, construir el intervalo de confianza respectivo y verificar los supuestos subyacentes.

Dicho esto, se presentan en este punto del trabajo los resultados derivados de evaluar la asociación entre las parejas de variables HOMA2IR-IMC y HOMA2IR-TAG/HDL. Para garantizar la comprensión cabal del contenido, sírvase descargar de los repositorios ya identificados, tanto la base de datos que sirve como muestra, como los *scripts*. Adicionalmente, se reafirma la aclaración hecha con anterioridad de que el aspecto matemático no será abordado de manera explícita; se asume que el lector tiene un conocimiento básico de estadística y que conoce las ecuaciones necesarias para llevar a cabo este procedimiento.

La mayoría de investigadores con experiencia coinciden en una recomendación: comenzar todo estudio con una exploración de los datos, sondeo en el que resultan de gran utilidad ciertas herramientas gráficas como histogramas, gráficos de caja y bigotes o diagramas de dispersión. En esta oportunidad, el gráfico de dispersión no solo sirve para vislumbrar qué tipo de relación conecta a las variables, sino también para identificar datos atípicos o patrones en la nube de puntos que pudieran alterar los resultados y generar confusión al momento de interpretar los hallazgos. Así pues, se presentan en la **figura 1** los diagramas de dispersión asociados a los pares de variables previamente mencionados. En el eje de las ordenadas se representa la resistencia a la insulina, mientras que en el eje de las abscisas se ubican el índice de masa corporal y la relación TAG/HDL. La primera impresión sustenta la hipótesis de que el HOMA2IR pudiera estar relacionado linealmente con el IMC y el índice TAG/HDL. Nótese que los puntos se distribuyen alrededor de las rectas sin exhibir comportamientos irregulares, exponenciales o curvilíneos, pero también obsérvese la existencia de tres datos potencialmente atípicos en el primer diagrama (**figura 1a**) y dos en el segundo (**figura 1b**). Esta situación no debería pasar desapercibida; de hecho, lo conveniente sería examinar estos valores para confirmar si en realidad son observaciones aberrantes, para descubrir la causa de tales inconsistencias, y en últimas, para tomar decisiones al respecto.

Figura 1

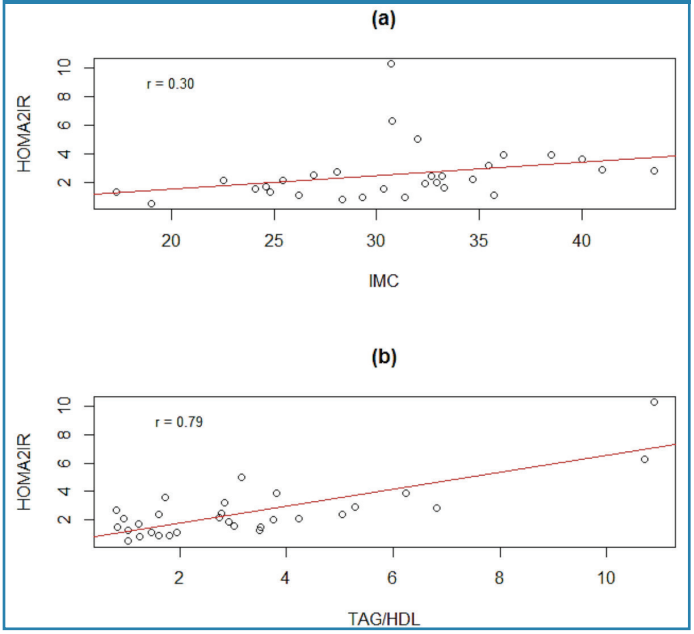

Figura 1. Diagramas de dispersión obtenidos a partir de la base de datos usada como muestra. Se presenta: (a), diagrama de dispersión con línea de ajuste para el par HOMA2IR-IMC incluyendo su coeficiente (0.30); y (b), diagrama de dispersión con línea de ajuste para el par HOMA2IR-TAG/HDL incluyendo su coeficiente (0.79). Nótese la presencia de tres puntos en el primer diagrama que podrían ser clasificados como datos atípicos, mientras que en el segundo diagrama se observan dos valores considerablemente alejados del resto de los datos.

Una vez hecho esto se procede con el análisis, omitiendo por ahora la posible eliminación de los valores inusuales y trabajando con la base de datos sin depurar. La **figura 2** muestra la salida del programa *R-Studio* copiada directamente desde la consola. Como puede apreciarse, no hay evidencia de correlación lineal significativa entre la resistencia a la insulina y el índice de masa corporal ($r = 0.30$, $t = 1.66$, gl = 28, $p = .107$, ICB 95 %: -0.07 a 0.60), pero sí entre la resistencia a la insulina y la relación TAG/HDL ($r = 0.79$, $t = 6.81$, gl = 28, $p = 2.11 \times 10^{-7}$, ICB 95 %: 0.60 a 0.90). Es importante recalcar que, casi con toda seguridad, las observaciones aberrantes tendrán efecto sobre estos resultados, influencia que será examinada en los siguientes apartados del trabajo.

Figura 2

```
Console ~/
> #Análisis de correlación para el par IMC-HOMA2IR#
> cor.test(IMC, HOMA2IR, alternative = "two.sided", method = "pearson", exact = TRUE, conf.level = 0.95

        Pearson's product-moment correlation

data:  IMC and HOMA2IR
t = 1.6644, df = 28, p-value = 0.1072
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06752043  0.59590116
sample estimates:
      cor
0.3000476

>
> #Análisis de correlación para el par TAG/HDL-HOMA2IR#
> cor.test(TAG_HDL, HOMA2IR, alternative = "two.sided", method = "pearson", exact = TRUE, conf.level =

        Pearson's product-moment correlation

data:  TAG_HDL and HOMA2IR
t = 6.8148, df = 28, p-value = 2.106e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6004440 0.8953427
sample estimates:
      cor
0.7898514
```

Figura 2. Salida del programa *R-Studio* generada a partir del análisis de correlación para las variables HOMA2IR-IMC y HOMA2IR-TAG/HDL. Se muestra el estadístico de prueba, los grados de libertad, el nivel de significación, la estimación puntual y el intervalo de confianza de 95 %. En la parte superior de la imagen se presentan los resultados de correlacionar al HOMA2IR con el IMC, mientras que en la inferior se exhiben los de asociar al HOMA2IR con el índice TAG/HDL.

Verificación de supuestos

La validez de las conclusiones que se extraen de la inferencia estadística estará supeditada al cumplimiento de los supuestos bajo los cuales se construyen los modelos que sirven para implementarla. Por tanto, nada se logra con la realización del análisis y su debida interpretación, si no se verifica que estos tengan asidero en un procedimiento bien ejecutado. En las próximas secciones se evaluará si los datos empleados en esta revisión *pasan* la prueba impuesta por siete premisas: nivel de medición, datos pareados, normalidad bivariada, ausencia de datos atípicos, linealidad, independencia de observaciones y condiciones del muestreo.

1. Nivel de medición

Cuando se usa el coeficiente de correlación de *Pearson*, se debe prestar especial atención al tipo de variable que se examina. Así, las características evaluadas deben poseer un nivel de medición que sea al menos de intervalo; en otras palabras, no es correcto utilizar esta medida si se está trabajando con variables nominales u ordinales[25,28,30-33]. Herramientas alternativas como el coeficiente biserial-puntual, que es una versión del coeficiente *R* implementada cuando una variable es intervalar y la otra dicotómica, pueden ser usadas toda vez que la condición acá analizada no se satisfaga[34,35]. También, por supuesto, existen alternativas no paramétricas como las de *Tau* de *Kendall* o *Rho* de *Spearman-Brown*, que admiten niveles ordinales, de intervalo o de razón[36-39].

Resultará sencillo para el lector constatar que se le da observancia a esta conjetura. El IMC es un factor que calcula cuánta masa hay en la superficie que ocupa el cuerpo de un individuo[14], de manera que bien podría clasificarse dentro del grupo de variables medidas en escala de razón: por un lado, el cero no es únicamente referencial e indicaría la ausencia de dicha propiedad; por el otro, las razones son válidas puesto que afirmaciones como *el sujeto A (30 kg/m²) tiene el doble de masa por metro cuadrado que el sujeto B (15 kg/m²)* tienen sentido y no entrañan inconsistencias con el atributo que expresa la variable[40-43]. En lo concerniente al HOMA2IR y al índice TAG/HDL, al ser cifras adimensionales cuyo dominio es el conjunto de números reales positivos sin incluir el cero, deberían ser catalogadas como variables de intervalo. Obsérvese que expresiones como *el sujeto A (HOMA2IR=4) tiene el doble de resistencia a la insulina que el sujeto B (HOMA2IR = 2)* no son correctas en un sentido exacto; esto solo implica que el *valor* de tal característica en el primer individuo dobla al del segundo, pero no significa que la *propiedad* que mide se comporte de la misma manera[40-43].

2. Datos pareados

El hecho de acometer análisis multivariados supone, necesariamente, la medición simultánea de más de una característica en cada elemento de la muestra. El problema que se estudia en esta revisión es un claro ejemplo de ello: cada una de las 2230 personas que participaron en la investigación original fueron encuestadas para obtener información acerca del sexo, edad, presión arterial sistólica, presión arterial diastólica, índice de masa corporal, resistencia a la insulina, entre otras[8,9]. Lo anterior permite construir una base de datos relacionada; una matriz en la que cada fila es un sujeto y cada columna es una variable.

En este orden de ideas, podrá corroborarse con facilidad que la noción de *datos pareados* o *datos relacionados* se cumple satisfactoriamente en esta revisión. Obsérvese que el archivo que sirve como muestra es una matriz de 30 filas por 7 columnas, y que; además, no existen valores perdidos. Si este no fuera el caso, habría que descartar por completo aquellos registros en los que se observaran casillas vacías, reduciendo así la dimensión del conjunto utilizado.

3. Normalidad bivariada

Este supuesto, junto al de la ausencia de datos atípicos, es el que entraña mayores dificultades debido a que requiere de la implementación de técnicas estadísticas multivariantes. Suele creerse que el cumplimiento de la normalidad marginal es condición suficiente para utilizar el coeficiente de correlación de *Pearson*. En términos estrictos, esto no es correcto; lo conveniente es que el investigador se cerciore de que la distribución conjunta de las variables se ajusta a un modelo normal bivariado; porque, aunque ambas características por separado se distribuyan normalmente, bien podrían no hacerlo de manera simultánea[44-49].

A partir de esta necesidad, se han formulado varias pruebas de bondad de ajuste que toman en consideración el nivel multivariado de los datos. Procedimientos como el de *Mardia*, *Royston* o *Henze-Zirkler* son los más conocidos, aunque de ellos, el primero es tal vez el que ha demostrado mayor fiabilidad y exactitud[44-46,50]. Por otro lado, contrastes como el de *Shapiro-Wilk* o el de *Kolmogorov-Smirnov-Lilliefors* son ampliamente utilizados a nivel univariado, toda vez que se desea identificar qué tanto se asemeja un conjunto de datos a una campana de *Gauss*, especialmente cuando el número de observaciones es pequeño (generalmente menos de 50).

Tomando en cuenta estos comentarios, se ofrecen ahora los resultados del análisis de normalidad de las combinaciones HOMA2IR-IMC y HOMA2IR-TAG/HDL. Se ha utilizado la prueba de *Mardia* para verificar la normalidad conjunta, en tanto que se ha empleado la de *Shapiro-Wilk* para confrontar la normalidad marginal. Remítase a las **figuras 3 y 4** que exhiben el procedimiento y los hallazgos de esta sección del estudio. Nótese que, en el caso del par HOMA2IR-IMC, tanto la asimetría como la curtósis de los datos reflejan alejamientos significativos de la normalidad bivariada ($M\text{-}Skewness = 38.26$, $p = 9.91 \times 10^{-8}$, $M\text{-}Kurtosis = 4.47$, $p = 7.90 \times 10^{-6}$). En lo referente al comportamiento marginal, se aprecia que el del HOM2IR tampoco es consistente con la hipótesis evaluada ($SW = 0.76$, $p < .001$), pero sí lo es el del IMC ($SW = 0.99$, $p = .992$). Ahora bien, al examinar las variables HOMA2IR-TAG/HDL, se advierte que ninguno de los estadísticos de forma *pasa* la prueba de *Mardia*, lo que se traduce en un rechazo de la suposición de normalidad bivariada ($M\text{-}Skewness = 31.95$, $p = 1.96 \times 10^{-6}$, $M\text{-}Kurtosis = 4.24$, $p = 2.28 \times 10^{-5}$). A nivel univariado sucede lo mismo, siendo que la distribución del TAG/HDL muestra diferencias significativas con respecto a la campana de *Gauss* ($SW = 0.81$, $p = <.000$).

Figura 3

```
Console ~/
> #Para instalar y activar el paquete MVN, necesario para ejecutar las pruebas#
> install.packages("MVN")
Installing package into 'C:/Users/USUARIO/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/MVN_5.5.zip'
Content type 'application/zip' length 385755 bytes (376 KB)
downloaded 376 KB

package 'MVN' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\USUARIO\AppData\Local\Temp\RtmpCOm3J9\downloaded_packages
> library(MVN)
sROC 0.1-2 loaded
Warning message:
package 'MVN' was built under R version 3.5.2
>
```

Figura 3. Salida del programa *R-Studio* generada a partir de la instalación y activación del paquete *MVN*, extensión necesaria para realizar las pruebas de normalidad multivariadas. Se recomienda copiar en la consola del programa el contenido de esta imagen o la información del segundo *script* que se suministra en esta revisión.

Figura 4

```
Console ~/
> #Para crear las matrices de datos bivariadas#
> X=data.frame(IMC, HOMA2IR)
> Y=data.frame(TAG_HDL, HOMA2IR)
>
> #Para ejecutar las pruebas de normalidad univariadas y bivariadas#
> mvn(X, mvnTest = "mardia", univariateTest = "SW")
$`multivariateNormality`
             Test         Statistic              p value Result
1 Mardia Skewness 38.2587583323866  9.9093819065216e-08     NO
2 Mardia Kurtosis 4.46782165300821 7.90201419209602e-06     NO
3             MVN              <NA>                 <NA>     NO

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk     IMC      0.9903  0.9924      YES
2 Shapiro-Wilk  HOMA2IR     0.7618  <0.001      NO

$Descriptives
        n     Mean  Std.Dev Median   Min   Max  25th   75th        Skew   Kurtosis
IMC     30 30.71467 6.201739 31.075 17.29 43.53 26.38 34.335 -0.08849571 -0.4392913
HOMA2IR 30  2.54800 1.953122  2.100  0.50 10.30  1.35  2.875  2.25040734  5.9637236

> mvn(Y, mvnTest = "mardia", univariateTest = "SW")
$`multivariateNormality`
             Test         Statistic              p value Result
1 Mardia Skewness 31.9501477697193 1.95850749124786e-06     NO
2 Mardia Kurtosis 4.23510926459419 2.28440600673618e-05     NO
3             MVN              <NA>                 <NA>     NO

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk  TAG_HDL     0.8064   1e-04      NO
2 Shapiro-Wilk  HOMA2IR     0.7618  <0.001      NO

$Descriptives
        n     Mean  Std.Dev Median Min   Max  25th  75th      Skew Kurtosis
TAG_HDL 30 3.281667 2.596449  2.805 0.8 10.89 1.495 3.805 1.573874 2.050884
HOMA2IR 30 2.548000 1.953122  2.100 0.5 10.30 1.350 2.875 2.250407 5.963724
```

Figura 4. Salida del programa *R-Studio* generada a partir del análisis de normalidad a las parejas HOMA2IR-IMC, denotada como *X*; y HOMA2IR-TAG/HDL, denotada como *Y*. En la parte superior se indica cómo construir las matrices bivariadas a partir del comando *data.frame*, esto debido a que la extensión *MVN* no trabaja con vectores individuales sino conjuntos. En los hallazgos se podrán encontrar los estadísticos de prueba, el valor-p y las medidas de resumen tradicionales. Lo más importante se encuentra en la columna *result*; nótese que el supuesto de normalidad bivariada ha sido rechazado para ambas matrices.

4. Ausencia de datos atípicos bivariados

Probablemente, la presencia de datos atípicos sea la más crucial de todas las suposiciones asociadas al análisis de correlación. En algunas situaciones, basta con que exista un solo punto que se distancie considerablemente del resto para que los resultados del estudio se vean seriamente afectados. Conceptualizándolo en términos simples, un *outlier* es un valor que se aleja significativamente del patrón que describen las demás observaciones[51]. Ahora bien, cuando se construyen modelos de regresión lineal, esta definición puede ser ligeramente distinta y merece la pena hacer la distinción entre lo que es un dato atípico, una observación con alto apalancamiento y un punto de influencia. Dentro de este contexto, un *outlier* es una observación cuya respuesta en la variable dependiente (*Y*) no sigue la línea de tendencia general formada por el grupo, mientras que un punto de apalancamiento (*high leverage observation*) es aquel que posee un valor extremo en la variable independiente (*X*) o en alguna combinación de los predictores (X_j). Finalmente, cualquier dato será un punto de influencia si su eliminación modifica ampliamente los resultados del análisis, manifestándose con cambios importantes en el coeficiente de correlación o de determinación, en las estimaciones de la pendiente y del intercepto, o en los resultados de la prueba de hipótesis. En consecuencia, un *outlier* y un *high leverage observation* tienen el potencial de ser clasificados como puntos de influencia, siempre y cuando se realice la verificación respectiva y se comprueben dichas sospechas[52].

Una revisión de la figura 1 podría facilitar el entendimiento de lo expuesto en el párrafo anterior. Obsérvese que en la parte (a) de la imagen hay tres datos atípicos, mismos que poseen valores en el eje *Y* (HOMA2IR) que se alejan sensiblemente de la recta. Por otro lado, en la parte (b) del gráfico pueden distinguirse dos observaciones que, aunque yacen próximas a la línea de regresión, se separan del resto de forma apreciable en el componente *X* (TAG/HDL) del diagrama. Siendo así, los casos de la figura 1a habrán de clasificarse como *outliers*, mientras que los de la figura 1b serán catalogados como puntos de apalancamiento.

Hechas las deliberaciones anteriores, se procederá a inspeccionar estas observaciones a fin de esclarecer si tienen o no impacto en los hallazgos de la investigación. Para esto, se emplearán en esta revisión las distancias de *Mahalanobis*[53] a través del paquete *mvoutlier* de *R-Studio*, extensión que utiliza estimaciones robustas que no se ven afectadas por la presencia de datos inusuales. En contextos más amplios como los de regresión, podrían emplearse otras medidas además de esta, tales como las de los residuales estandarizados, residuales estudentizados internos, residuales estudentizados externos, distancias de *Cook*, entre otras[54]. Las **figuras 5, 6 y 7** presentan los resultados de esta etapa del trabajo. Compruébese que, tanto para la combinación HOMA2IR-IMC, como para el par HOMA2IR-TAG/HDL, los mismos tres datos son clasificados como atípicos por parte del procedimiento. Una inspección detallada revela que tales casos corresponden a los sujetos ubicados en las posiciones 12, 13 y 18, así que, utilizando esta información, se podría realizar una revisión exhaustiva que permita identificar la causa de tal anomalía. Lo siguiente sería decidir si se conservan o eliminan tales registros de la base de datos, resolución que debe estar fundamentada no solo en cuestiones estadísticas, sino también en criterios disciplinares e investigativos.

Para efectos ilustrativos, en esta revisión se eliminarán los casos 12 y 18 por ser los más extremos. Se aclara que el criterio que impulsa tal depuración es intuitivo, aunque obviamente se basa en la experiencia y en lo que reflejan los diagramas de dispersión. En la práctica, el analista deberá probar varias alternativas e identificar cuál de estos valores es el que influye en mayor grado, realizando muchas veces procesos iterativos y repetitivos en los cuales el auxilio de la informática resulta vital. La **tabla 2** ofrece una comparación de los resultados obtenidos al incluir y al descartar los registros previamente mencionados. Nótese que, en efecto, tales observaciones resultaron puntos de influencia puesto que modificaron sustancialmente el coeficiente de correlación y los estadísticos del contraste de normalidad. De manera es-

pecífica, el *R* de *Pearson* evidenció un aumento relativo de 86.67 %, pasando de 0.30 a 0.56 y reportando una asociación lineal significativa entre el índice HOMA2IR y el IMC (.002). Por su parte, la alta correlación observada inicialmente en la combinación HOMA2IR-TAG/HDL se redujo en un 39.24 %, disminuyendo de 0.79 a 0.48 pero aun evidenciando valores significativos (.010). Por último, resalta el impacto considerable que la eliminación de los datos tiene en el supuesto de normalidad bivariada, suposición que pasa de ser rechazada en ambas parejas de variables, a una condición en la que es asumida como plausible.

Figura 5

```
Console ~/
> #Para crear las matrices de datos bivariadas#
> attach(Datos)
> X=data.frame(IMC, HOMA2IR)
> Y=data.frame(TAG_HDL, HOMA2IR)
>
>
> #Para instalar y activar el paquete mvoutlier#
> install.packages("mvoutlier")
Installing package into 'C:/Users/USUARIO/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/mvoutlier_2.0.9.zip'
Content type 'application/zip' length 802746 bytes (783 KB)
downloaded 783 KB

package 'mvoutlier' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\USUARIO\AppData\Local\Temp\Rtmp4iQXRo\downloaded_packages
> library(mvoutlier)
Loading required package: sgeostat
SROC 0.1-2 loaded

Attaching package: 'mvoutlier'

The following objects are masked _by_ '.GlobalEnv':

    X, Y

Warning message:
package 'mvoutlier' was built under R version 3.5.2
>
>
> #Para ejecutar el análisis de datos atípicos sobre X#
> aq.plot(X)
$`outliers`
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
[18]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

> aq.plot(Y)
$`outliers`
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
[18]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Figura 5. Salida del programa *R-Studio* generada a partir de la instalación del paquete *mvoutlier* y del análisis de datos atípicos multivariados mediante las distancias robustas de *Mahalanobis*. Los valores identificados con *false* no se consideran observaciones aberrantes. Nótese la presencia de tres datos inusuales en la combinación HOMA2IR-IMC y tres en la pareja HOMA2IR-TAG/HDL, hecho que corrobora las sospechas planteadas previamente.

Figura 6

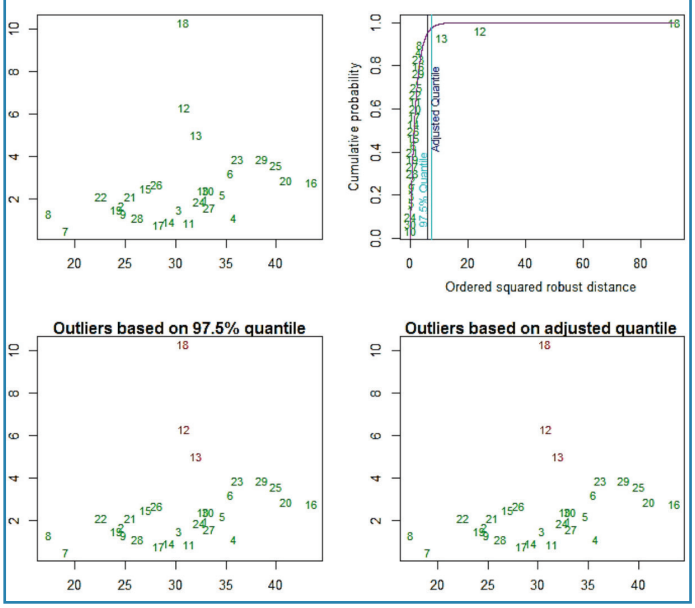

Figura 6. Salida del programa *R-Studio* para el análisis de datos atípicos a la combinación HOMA2IR-IMC. Los valores inusuales están identificados en rojo en los dos paneles inferiores. También se aprecian en el diagrama ubicado en la parte superior derecha, siendo que sobrepasan la línea azul de los cuantiles ajustados. Nótese que las observaciones atípicas son las que corresponden al orden 12, 13 y 18.

Figura 7

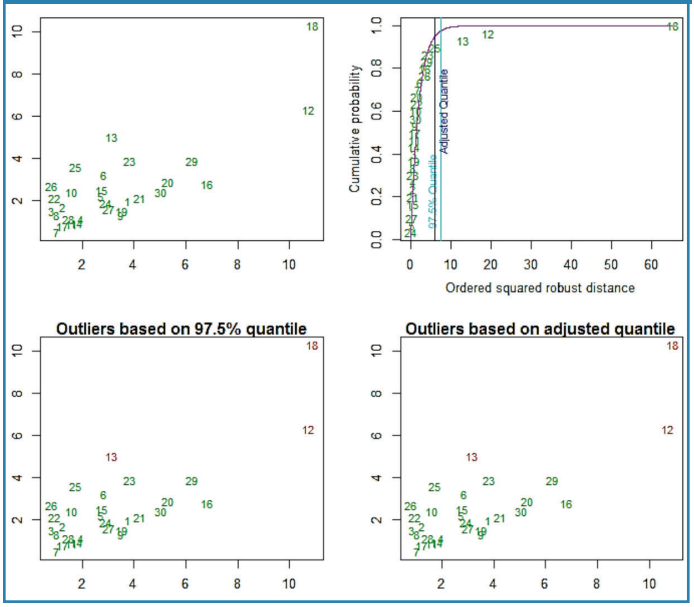

Figura 7. Salida del programa *R-Studio* para el análisis de datos atípicos a la combinación HOMA2IR-TAG/HDL. Los valores inusuales están identificados en rojo en los dos paneles inferiores. También se aprecian en el diagrama ubicado en la parte superior derecha, siendo que sobrepasan la línea azul de los cuantiles ajustados. Nótese que en este caso las observaciones atípicas también corresponden al orden 12, 13 y 18.

Tabla 2

Estadísticos	HOMA2IR-IMC		HOMA2IR-TAG/HDL	
	Todos los datos (n =30)	Sin casos 12 y 18 (n =28)	Todos los datos (n =30)	Sin casos 12 y 18 (n =28)
Correlación: r (p)	0.30 (.107)	0.56 (.002)	0.79 (<.001)	0.48 (.010)
Norm. biv.: *Ske* (p)	38.26 (9.91×10^{-9})	3.98 (.408)	31.95 (1.96×10^{-6})	9.21 (.056)
Norm. biv.: *Kur* (p)	4.47 (7.90×10^{-6})	-0.07 (.946)	4.24 (2.28×10^{-5})	-0.14 (.888)

Tabla 2. Comparación entre los resultados del análisis de correlación y de normalidad bivariada al incluir todos los datos y al eliminar los registros 12 y 18. Se muestran los estimadores puntuales, los estadísticos de prueba y el valor p. Obsérvese la influencia que tienen los casos descartados a través de los cambios significativos encontrados en la magnitud del coeficiente *R* de *Pearson* y en la normalidad.

5. Linealidad

El término *correlación* suele emplearse indistintamente para hablar de la relación entre variables, sin embargo, en el argot estadístico, suele asumirse que este concepto hace referencia únicamente al tipo de covariación que se representa mediante una línea recta. Técnicamente, esto es impreciso; de hecho, coeficientes como el de *Spearman-Brown* o *Tau* de *Kendall* también miden la correlación entre variables, pero cuando esta describe comportamientos monótonos. Una asociación monótona se da si, a medida que una de las variables crece, la otra crece o decrece de forma constante. En otras palabras: si *X* incrementa su valor y al mismo tiempo *Y* nunca decrece, se estará en presencia de una relación monótonamente creciente; por el contrario, si se observa que *X* aumenta, pero *Y* nunca lo hace, se hablará de una asociación monótona decreciente[55].

El coeficiente *R* de *Pearson* describe un caso especial de asociación monótona: la lineal. En tal sentido, toma relevancia el esfuerzo del investigador por determinar si los resultados se basan verdaderamente en una vinculación de este

tipo y no en relaciones que aparentan ser lineales, pero que bien podrían ser exponenciales, polinómicas, logarítmicas o de otro tipo. En consecuencia, se realiza ahora la verificación de este supuesto utilizando los mismos gráficos que se presentaron en la figura 1, pero a partir de la base de datos depurada. Como puede apreciarse en la **figura 8**, no hay razones para creer que se incumple la premisa de linealidad; más allá de la variabilidad esperada, la nube de puntos de ambos diagramas se distribuye a lo largo de las rectas sin mostrar comportamientos sistemáticos. Tampoco se distinguen observaciones inusualmente alejadas del centro de los datos.

Figura 8

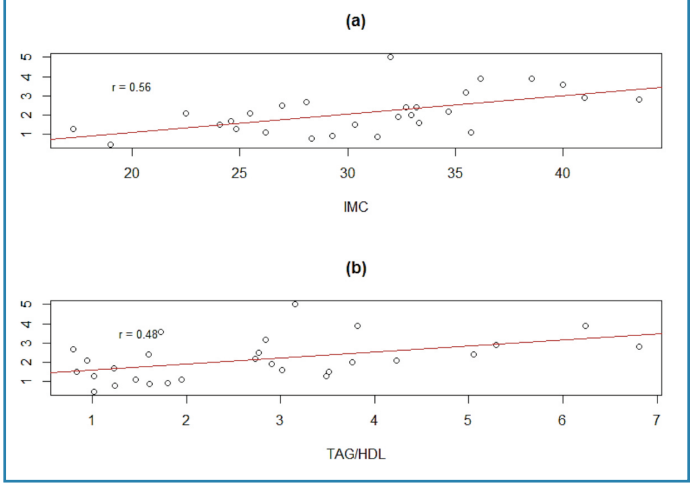

Figura 8. Diagramas de dispersión obtenidos a partir de la base de datos usada como muestra luego de la depuración. Se presenta: (a), diagrama de dispersión con línea de ajuste para el par HOMA2IR-IMC incluyendo su coeficiente corregido (0.56); y (b), diagrama de dispersión con línea de ajuste para el par HOMA2IR-TAG/HDL incluyendo su coeficiente corregido (0.48). Nótese que no hay puntos que destaquen de manera importante y que hagan sospechar de la existencia de valores atípicos. Obsérvese además que las observaciones se amoldan de manera aproximada a una línea recta.

6. Independencia de observaciones

Aunque pudiera parecer una contradicción, la inferencia estadística construida alrededor del coeficiente de correlación de *Pearson* exige que haya independencia de observaciones. Esto no implica la ausencia de relación entre las variables sino que exista independencia dentro del grupo; es decir, que el valor que un sujeto cualquiera reporta en una de las variables, esté relacionado únicamente con aquel que se evidencia en la otra variable, pero con ese mismo individuo y no con los demás integrantes de la muestra, sea cual sea la dirección que pudiera vincularlos[56,58]. Este tipo de conjetura es más una cuestión de diseño que un ejercicio estadístico, a pesar de que existen pruebas de aleatoriedad que aportan luces al respecto. Si el equipo de investigación vela por seleccionar aleatoriamente a los participantes, si se realiza un esfuerzo por asignar al azar a los sujetos y a los tratamientos en caso de llevar a cabo estudios experimentales, la suposición de independencia estaría garantizada al menos teóricamente.

Para continuar, se muestran en este apartado los resultados de aplicar la prueba de *Wald-Wolfowitz*. Como se detalla en la figura 9, el índice de masa corporal se distribuye de forma que sus observaciones son independientes ($WW = 0.77$, $p = .441$), situación que se repite en las otras dos características evaluadas. De manera concreta, la resistencia a la insulina reporta el mismo número de rachas por encima o por debajo de la mediana ($WW = -0.39$, $p = .700$), mientras que la razón triglicéridos-colesterol tampoco exhibe comportamientos que permitan descartar la aleatoriedad de los datos ($WW = 0.37$, $p = .710$).

7. Condiciones del muestreo

El postulado sobre el tipo de muestreo guarda relación estrecha con la suposición de independencia anteriormente considerada. Las ecuaciones que se utilizan en la estadística inferencial básica se apoyan en la noción de *variables aleatorias independientes e idénticamente distribuidas* (VAIID). Una muestra constituirá un conjunto de VAIID si todos los individuos de la población tienen la misma probabilidad de ser escogidos y si, además, la característica evaluada se distribuye bajo un mismo modelo estocástico[59-63]. Para darle cumplimiento a esto, la selección de los participantes supondría la reposición de cada unidad una vez extraída; aspecto que, evidentemente, no tiene mucho sentido en una situación real. De manera pues que el acatamiento estricto de este supuesto rara vez puede alcanzarse en investigaciones no experimentales, en las que es improbable encontrar poblaciones homogéneas que permitan implementar técnicas como el muestreo aleatorio simple (MAS), sin mencionar la dificultad que entraña la construcción del marco muestral necesario para la selección al azar de los sujetos.

Ahora bien, tal y como se explicó en la descripción del problema, en este artículo se ha propuesto un escenario figurado basado en una auténtica situación de investigación. En circunstancias así, sería sencillo demostrar que las condiciones del muestreo satisfacen en gran medida lo que se exige. Nótese en primer lugar, que antes de seleccionar la muestra con la que se ha venido trabajando la población fue fraccionada en estratos y esta fue obtenida de uno de ellos, hecho que avalaría la condición de uniformidad ya que la estratificación requiere de homogeneidad interna. Por otro lado, el restringir la fracción de muestreo a porcentajes menores que 5 % o 10 %, faculta al analista para prescindir del factor de corrección que ajusta los errores estándares de los estimadores, pudiendo así utilizar las fórmulas que emplearía en un análisis realizado sobre una población infinita, en el que se supone constante la probabilidad de inclusión[59].

Figura 9. Salida del programa *R-Studio* para la instalación del paquete *randtests* con el que se ejecuta la prueba de las rachas. Se muestran los estadísticos, las rachas, el tamaño muestral y el valor p. Adicionalmente, la extensión señala la hipótesis alternativa. Nótese que la suposición de independencia no es rechazada en ninguna de las tres variables de la investigación

Figura 9

```
Console ~/
> #Para instalar el paquete randtests#
> install.packages("randtests")
Installing package into 'C:/Users/USUARIO/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/randtests_1.0.zip'
Content type 'application/zip' length 75886 bytes (74 KB)
downloaded 74 KB

package 'randtests' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\USUARIO\AppData\Local\Temp\Rtmpo5ymO3\downloaded_packages
> library(randtests)
Warning message:
package 'randtests' was built under R version 3.5.2
>
>
> #Para verificar la aleatoriedad en la base de datos depuradas#
> attach(X1)
The following objects are masked from Datos:

    HOMA2IR, IMC

The following objects are masked from X1 (pos = 5):

    HOMA2IR, IMC

> runs.test(IMC)

        Runs Test

data:  IMC
statistic = 0.77033, runs = 17, n1 = 14, n2 = 14, n = 28, p-value = 0.4411
alternative hypothesis: nonrandomness

> runs.test(HOMA2IR)

        Runs Test

data:  HOMA2IR
statistic = -0.38516, runs = 14, n1 = 14, n2 = 14, n = 28, p-value = 0.7001
alternative hypothesis: nonrandomness

> runs.test(TAG_HDL)

        Runs Test

data:  TAG_HDL
statistic = 0.37161, runs = 17, n1 = 15, n2 = 15, n = 30, p-value = 0.7102
alternative hypothesis: nonrandomness
```

Referencias

1. Quevedo R, Pedreschi F, Bastias JM, Díaz O. Correlation of the fractal enzymatic browning rate with the temperature in mushroom, pear and apple slices. LWT - Food Sci Technol. enero de 2016;65:406-13.
2. Hospodar G, Gierlichs B, De Mulder E, Verbauwhede I, Vandewalle J. Machine learning in side-channel analysis: a first study. J Cryptogr Eng. diciembre de 2011;1(4):293-302.
3. He F, Rodríguez-Colon S, Fernández-Mendoza J, Vgontzas AN, Bixler EO, Berg A, et al. Abdominal Obesity and Metabolic Syndrome Burden in Adolescents—Penn State Children Cohort Study. J Clin Densitom. enero de 2015;18(1):30-6.
4. Korkmazer E, Solak N. Correlation between inflammatory markers and insulin resistance in pregnancy. J Obstet Gynaecol. 17 de febrero de 2015;35(2):142-5.
5. Rojas J, Bermúdez VJ, Añez RJ, Bello LM, Toledo A, Torres Y, et al. Comportamiento epidemiológico del síndrome metabólico en el municipio Maracaibo-Venezuela. Rev Síndr Cardiometabólico. 2013;3(2):13.
6. Bermúdez V, Pacheco M, Rojas J, Córdova E, Velázquez R, Carrillo D, et al. Epidemiologic Behavior of Obesity in the Maracaibo City Metabolic Syndrome Prevalence Study. Maedler K, editor. PLoS ONE. 18 de abril de 2012;7(4):e35392.
7. Salazar J, Bermúdez V, Olivar LC, Torres W, Palmar J, Añez R, et al. Insulin resistance indices and coronary risk in adults from Maracaibo city, Venezuela: A cross sectional study. F1000Research [Internet]. 9 de marzo de 2018 [citado 13 de enero de 2019];7. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107995/
8. Bermúdez V, Rojas J, Salazar J, Calvo MJ, Morillo J, Torres W, et al. The Maracaibo city metabolic syndrome prevalence study: primary results and agreement level of 3 diagnostic criteria. Rev Latinoam Hipertens. 2014;9(4):20-32.
9. Bermúdez V, Marcano RP, Cano C, Arráiz N, Amell A, Cabrera M, et al. The Maracaibo City Metabolic Syndrome Prevalence Study: Design and Scope: Am J Ther. mayo de 2010;17(3):288-94.
10. Matthews DR, Hosker JR, Rudenski AS, Naylor BA, Treacher DF, Turner RC, et al. Homeostasis model assessment: insulin resistance and fl-cell function from fasting plasma glucose and insulin concentrations in man. :8.
11. Levy JC, Matthews DR, Hermans MP. Correct Homeostasis Model Assessment (HOMA) Evaluation Uses the Computer Program. Diabetes Care. 1 de diciembre de 1998;21(12):2191-2.
12. Wallace TM, Levy JC, Matthews DR. Use and Abuse of HOMA Modeling. Diabetes Care. 1 de junio de 2004;27(6):1487-95.
13. Blackburn H, Jacobs D. Commentary: Origins and evolution of body mass index (BMI): continuing saga. Int J Epidemiol. 1 de junio de 2014;43(3):665-9.
14. Hall DMB. What use is the BMI? Arch Dis Child. 11 de enero de 2006;91(4):283-6.
15. Chiarpenello J, Bonino J, Pent MV, Baella AL. Índice triglicéridos/hdl colesterol en una población pediátrica de la ciudad de rosario y zona de influencia. 2018;5.
16. Roa Barrios M, Arata-Bellabarba G, Valeri L, Velázquez-Maldonado E. Relación entre el cociente triglicéridos/cHDL, índices de resistencia a la insulina y factores de riesgo cardiometabólico en mujeres con síndrome del ovario poliquístico. Endocrinol Nutr. febrero de 2009;56(2):59-65.
17. Belén L, Oliva ML, Maffei L, Rossi ML, Squillace C, Alorda MB, et al. Relación TG/HDL-C y resistencia a la insulina en mujeres adultas argentinas según su estado nutricional. Rev Esp Nutr Humana Dietética. 21 de noviembre de 2013;18(1):18-24.
18. Alf C, Lohr S. Sampling Assumptions in Introductory Statistics Classes. Am Stat. febrero de 2007;61(1):71-7.
19. Montgomery DC, Runger GC. Applied statistics and probability for engineers. 3rd ed. New York: Wiley; 2003. 706 p.
20. Weisberg S. Applied linear regression. 3rd ed. Hoboken, N.J: John Wiley & Sons, Ltd; 2005. 310 p. (Wiley series in probability and statistics).
21. Rawlings JO, Pantula SG, Dickey DA. Applied regression analysis: a research tool. 2nd ed. New York: Springer; 1998. 657 p. (Springer texts in statistics).
22. Samprit Chatterjee, Ali S. Hadi. Regression Analysis by Example. 4th ed. Hoboken, N.J: John Wiley & Sons, Ltd; 2006. 383 p. (Wiley series in probability and statistics).
23. Sedgwick P. Simple linear regression. BMJ. 12 de abril de 2013;346(apr12 1):f2340-f2340.
24. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. 2003;7(6):9.
25. Mukaka M. A guide to appropriate use of Correlation coefficient in medical research. Malawi Med J J Med Assoc Malawi. septiembre de 2012;24(3):69-71.
26. Ozer DJ. Correlation and the coefficient of determination. Psychol Bull. 1985;97(2):307-15.
27. Sedgwick P. Pearson's correlation coefficient. BMJ. 4 de julio de 2012;345(jul04 1):e4483-e4483.
28. Asuero AG, Sayago A, González AG. The Correlation Coefficient: An Overview. Crit Rev Anal Chem. enero de 2006;36(1):41-59.
29. Rodgers JL, Nicewander WA. Thirteen Ways to Look at the Correlation Coefficient. Am Stat. febrero de 1988;42(1):59.
30. Yeager K. LibGuides: SPSS Tutorials: Pearson Correlation [Internet]. [citado 18 de diciembre de 2018]. Disponible en: https://libguides.library.kent.edu/SPSS/PearsonCorr

31. Pearson Product-Moment Correlation - When you should run this test, the range of values the coefficient can take and how to measure strength of association. [Internet]. [citado 18 de diciembre de 2018]. Disponible en: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

32. Pearson's Product-Moment Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example. [Internet]. [citado 18 de diciembre de 2018]. Disponible en: https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php

33. Use and Misuse of Correlation Coefficients [Internet]. STAT 509. [citado 15 de enero de 2019]. Disponible en: https://newonlinecourses.science.psu.edu/stat509/node/160/

34. Point-Biserial Correlation in SPSS Statistics - Procedure, assumptions, and output using a relevant example. [Internet]. [citado 18 de diciembre de 2018]. Disponible en: https://statistics.laerd.com/spss-tutorials/point-biserial-correlation-using-spss-statistics.php

35. Gupta SD. Point biserial correlation coefficient and its generalization. Psychometrika. diciembre de 1960;25(4):393-408.

36. Spearman's Rank Order Correlation using SPSS Statistics - A How-To Statistical Guide by Laerd Statistics [Internet]. [citado 17 de enero de 2019]. Disponible en: https://statistics.laerd.com/spss-tutorials/spearmans-rank-order-correlation-using-spss-statistics.php

37. Zar JH. Spearman Rank Correlation. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2005 [citado 17 de enero de 2019]. Disponible en: http://doi.wiley.com/10.1002/0470011815.b2a15150

38. Kendall's Tau-b using SPSS Statistics - A How-To Statistical Guide by Laerd Statistics [Internet]. [citado 17 de enero de 2019]. Disponible en: https://statistics.laerd.com/spss-tutorials/kendalls-tau-b-using-spss-statistics.php

39. Kendall's Tau and Spearman's Rank Correlation Coefficient [Internet]. Statistics Solutions. [citado 17 de enero de 2019]. Disponible en: https://www.statisticssolutions.com/kendalls-tau-and-spearmans-rank-correlation-coefficient/

40. Triola MF, Pineda Ayala LE, Hernández Ramírez R. Estadística. 10.ª ed. México: Pearson/Educación; 2009.

41. Johnson R, Kuby P. Just the essentials of elementary statistics. 10.ª ed. Belmont, CA: Thomson Brooks/Cole; 2008.

42. Gary W. Heiman. Basic Statistics for the Behavioral Sciences. 6th ed. Belmont, CA: Wadsworth Cengage Learning; 2011. 504 p.

43. Ranjit Kumar. Research Methodology. 3rd ed. Los Angeles: Sage Publications; 2011.

44. Härdle W, Simar L. Applied multivariate statistical analysis. Fourth Edition. Berlin Heidelberg New York Dordrecht London: Springer; 2015. 580 p.

45. Timm NH. Applied multivariate analysis. New York: Springer; 2002. 693 p. (Springer texts in statistics).

46. Rencher AC. Methods of multivariate analysis. 2nd ed. New York: J. Wiley; 2002. 708 p. (Wiley series in probability and mathematical statistics).

47. Burdenski TK. Evaluating Univariate, Bivariate, and Multivariate Normality Using Graphical and Statistical Procedures. Am Educ Res Assoc. 2000;62.

48. Oppong FB, Agbedra SY. Assessing Univariate and Multivariate Normality, A Guide For Non-Statisticians. Math Theory Model. 2016;6(2):26-33-33.

49. Shao Y, Zhou M. A characterization of multivariate normality through univariate projections. J Multivar Anal [Internet]. noviembre de 2010 [citado 18 de enero de 2019];101(10). Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3837532/

50. Kankainen A, Taskinen S, Oja H. On Mardia's Tests of Multinormality. En: Hubert M, Pison G, Struyf A, Van Aelst S, editores. Theory and Applications of Recent Robust Methods [Internet]. Basel: Birkhäuser Basel; 2004 [citado 30 de diciembre de 2018]. p. 153-64. Disponible en: http://link.springer.com/10.1007/978-3-0348-7958-3_14

51. Charu C. Aggarwal. Outlier analysis. 2nd edition. New York, NY: Springer Science+Business Media; 2016. 481 p.

52. Distinction Between Outliers & High Leverage Observations [Internet]. STAT 501. [citado 18 de enero de 2019]. Disponible en: https://newonlinecourses.science.psu.edu/stat501/node/337/

53. Franklin S, Thomas S, Franklin S. Robust multivariate outlier detection using Mahalanobis' distance and modified Stahel-Donoho estimators. Semantic Sch. 2001;35.

54. Influential Points [Internet]. STAT 501. [citado 21 de enero de 2019]. Disponible en: https://newonlinecourses.science.psu.edu/stat501/node/336/

55. A comparison of the Pearson and Spearman correlation methods [Internet]. [citado 21 de enero de 2019]. Disponible en: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/

56. Editor MB. Common Assumptions about Data (Part 1: Random Samples and Statistical Independence) [Internet]. The MiniTab Blog. [citado 21 de enero de 2019]. Disponible en: http://blog.minitab.com/blog/quality-business/common-assumptions-about-data-part-1-random-samples-and-statistical-independence

57. Independent Observations Assumption [Internet]. Statistics Data Sciences. [citado 21 de enero de 2019]. Disponible en: http://sites.utexas.edu/sos/indobs/

58. Romano JL, Kromrey JD. What Are the Consequences If the Assumption of Independent Observations Is Violated in Reliability Generalization Meta-Analysis Studies? Educ Psychol Meas. junio de 2009;69(3):404-28.

59. William G. Cochran. Sampling Techniques. 3rd ed. New York, NY: John Wiley & Sons, Inc.; 1977. 442 p. (Wiley series in probability and mathematical statistics).

60. Rao PSRS. Sampling methodologies: with applications. Boca Raton, Fla: Chapman & Hall/CRC; 2000. 311 p. (Texts in statistical science).

61. Thompson SK. Sampling. 3rd ed. Hoboken, N.J: John Wiley & Sons, Inc.; 2012. 436 p. (Wiley series in probability and statistics).

62. Levy PS, Lemeshow S. Sampling of populations: methods and applications. 3rd ed. New York: John Wiley & Sons, Inc.; 1999. 525 p. (Wiley series in probability and statistics).

63. Sharon L. Lohr. Sampling: Design and Analysis. 2nd ed. Boston, MA: Brooks/Cole Cengage Learning; 2010. 609 p. (Advanced Series).