



**UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO**

**INSTITUTO DE INVESTIGACIONES EN
MATEMÁTICAS APLICADAS Y EN SISTEMAS
(IIMAS)**

**“MONOGRAFÍA SOBRE EL COEFICIENTE DE
CORRELACIÓN TETRACÓRICO”**

T E S I N A

QUE PARA OBTENER EL TÍTULO DE:

**ESPECIALISTA EN ESTADÍSTICA
APLICADA**

P R E S E N T A:

BRENDA CORINA CEREZO SILVA



**DIRECTORA DE TESINA:
M. EN C. LETICIA EUGENIA GRACIA
MEDRANO VALDELAMAR
2021**

Resumen

El coeficiente de correlación tetracórico se utiliza para conocer la asociación entre dos variables dicotómicas. Parte del supuesto que las dos variables se distribuyen normal vibariada y propone un “distretización” a partir de

A mi familia y seres queridos, quienes siempre me han brindado su apoyo y confianza incondicionales.

A la UNAM, por los conocimientos compartidos a través de sus profesores y alumnos.

A mi asesora M. en C. Leticia Gracia, por su apoyo, orientación y paciencia.

Contenido

Resumen	2
Introducción	5
Conocimientos preliminares	6
Tabla de contingencia.....	6
Medida de asociación	6
Asociación no implica causalidad.....	7
Prueba de independencia χ^2 de Pearson	8
Coefficiente de correlación tetracórico	11
Idea general	11
Cálculo de r	13
Comentarios sobre el cálculo de r	15
Error probable del coeficiente de r	17
Comentarios sobre el cálculo del error probable de r	18
Programación de r y su error probable (por series de Pearson).....	19
Casos especiales de tablas de contingencia para el cálculo de r y su error probable	24
Descripción de la metodología	28
Ejemplos	28
Comparación/relación con otras técnicas similares	32
Conclusión	34
Referencias	35

Introducción

El análisis de correlación entre dos variables es una de las principales metodologías que acompañan al análisis de datos. Las medidas de asociación no solo permiten inferir si existe alguna relación de dependencia entre variables, sino también permiten describir qué tan fuerte o débil es la relación.

Aunque existen variables que pueden medirse con extraordinaria precisión hasta incluso dar un valor con más de 18 decimales, como el tiempo, el peso, la distancia, la radiación, etc. (conocidas como variables continuas). Existen muchas otras cuya naturaleza no es tomar un valor numérico, sino categórico o cualitativo (conocidas como variables categóricas), por ejemplo: nacionalidad, sexo, grupo sanguíneo, etc. Cuando el valor de una variable solo puede ser alguna de dos categorías, se dice que es dicotómica. El presente trabajo explica el coeficiente de correlación tetracórico para determinar la correlación entre dos variables dicotómicas.

No debe subestimarse a las variables dicotómicas o pensar que el uso de estas limita la inferencia estadística. Simplemente es natural e inevitable que en ciertos estudios se presenten y más aún que sea el interés del investigador conocer la relación entre ellas. Las variables dicotómicas están presentes en una amplia gama de aplicaciones científicas. En consecuencia, la medida de asociación de éstas es muy útil en muchas situaciones. Por ejemplo, en el área de medicina muchos fenómenos sólo pueden ser medidos de forma fiable en términos de variables dicotómicas y resulta evidente el deseo de un investigador de saber, por ejemplo, si la variable vacuna (dicotómica por el hecho de describir la cualidad de si el paciente *recibió* o *no recibió* cierta vacuna) esta correlacionada con la variable resultado (dicotómica por el hecho de describir la cualidad de si el paciente *se recuperó* o *murió*). Otro ejemplo es en psicología, donde muchos trastornos sólo pueden ser medidos en términos de, por ejemplo, *diagnosticado* o *no diagnosticado*. Como último ejemplo, en las áreas sociales, en materia de discriminación de género, podría estudiarse la correlación entre la variable sexo (dicotómica por el hecho de describir la cualidad de *hombre* o *mujer*) y la variable aceptación (dicotómica por el hecho de describir la cualidad de cierto aspirante a una vacante en alguna empresa de ser *aceptado* o *rechazado*).

Actualmente existen varios coeficientes de correlación y numerosos artículos que discuten su eficiencia y su veracidad. Sin embargo, fue Karl Pearson quien, en 1990 a través de su séptimo artículo de la serie *Mathematical contributions to the theory of evolution*, presentó lo que hoy se conoce como el coeficiente de correlación tetracórico, aunque es interesante el hecho de que adoptó ese nombre tiempo después, pues Pearson solo se refiere a él como “el método que se presenta en esta memoria”.

En el presente trabajo se explicarán el coeficiente de correlación recién mencionado, algunos comentarios relativos a su cálculo, junto con una metodología sugerida para su implementación.

Conocimientos preliminares

Tabla de contingencia

Si se observan dos variables dicotómicas es común que la información se muestre en una tabla de contingencia de 2×2 , a cada individuo u objeto observado se le hace una clasificación cruzada y se cuentan los totales para cada clasificación, es decir, las frecuencias. Por ejemplo¹:

		Viruela		
		Se recuperó	Murió	Total:
Vacuna	Sí	1562	42	1604
	No	383	94	477
Total:		1945	136	2081

Tabla 1. Datos de la viruela recuperados por Karl Pearson (1900)

En la Tabla 1 se muestra la variable Vacuna cuyos posibles valores son sí o no, y la variable Viruela cuyos posibles valores son Se recuperó o Murió, entonces cada paciente observado recibe una doble clasificación, una por cada variable, así que, por ejemplo, hubo 1,562 pacientes que sí recibieron la vacuna y se recuperaron de la viruela.

Medida de asociación

Si dos variables son dependientes, entonces una intuitivamente proporciona información de la otra. Correlación o dependencia es cualquier relación estadística, entre dos variables aleatorias. La correlación es cualquier asociación estadística que comúnmente se refiere al grado en que un par de variables están relacionadas linealmente.

En lenguaje informal, la correlación es sinónimo de dependencia. Sin embargo, en sentido técnico se refiere a cualquiera de varios tipos específicos de operaciones matemáticas entre las variables y sus respectivos valores esperados.

Ekström (2009) enlista las propiedades que satisface una medida de asociación $S(X, Y)$, donde X y Y son dos variables aleatorias:

- I. S está definida para cualquier par de variables aleatorias.
- II. $S(X, Y) = S(Y, X)$.
- III. $-1 \leq S(X, Y) \leq 1$, $S(X, X) = 1$, $S(X, -X) = -1$.
- IV. Si X y Y son independientes, entonces $S(X, Y) = 0$.
- V. $S(-X, Y) = S(X, -Y) = -S(X, Y)$.
- VI. Si f y g son funciones casi seguramente estrictamente crecientes, entonces $S(f(X), g(Y)) = S(X, Y)$.

¹ Ilustración VI de Pearson(1900). En la sección Descripción de la metodología del presente trabajo se analiza esta tabla de contingencias.

- VII. Si (X, Y) y $\{(X, Y)\}_{n=1}^{\infty}$ son pares de variables aleatorias con función de distribución conjunta H y H_n , respectivamente, y si la secuencia $\{H_n\}$ converge a H , entonces $\lim_{n \rightarrow \infty} S(X_n, Y_n) = S(X, Y)$.

Las propiedades III y IV, implican que si existe una función creciente f tal que $f(X) = Y$ casi seguramente, entonces $S(X, Y) = 1$. Más aun, en combinación con la propiedad V se tiene que siempre que exista una función g estrictamente decreciente tal que $g(X) = Y$ casi seguramente, entonces $S(X, Y) = -1$.

En consecuencia, sencillamente se puede decir que una medida de asociación entre X y Y contiene información sobre el grado en que X y Y pueden ser representadas a través de una función estrictamente monótona de la otra.

Por último, cabe recordar que valores cercanos a 1 o a -1 indican una correlación fuerte, es decir, que valores grandes de la primera variable están asociados con valores grandes de la segunda (llamada correlación directa en caso de ser cercana a 1); y valores grandes de la primera variable están asociados con valores pequeños de la segunda (llamada correlación inversa en caso de ser cercana a -1).

Asociación no implica causalidad

Fernando Cortés en su artículo “Observación, causalidad y explicación causal” (2018) señala que no se puede inferir causalidad a partir de datos observacionales, y esto quiere decir que en realidad no es correcto hablar de causalidad refiriendo un modelo. Argumenta que los modelos son concreciones del pensamiento conceptual de quienes los diseñan o definen, esto quiere decir que los investigadores incluyen en el diseño de su investigación elementos apriorísticos, es decir, elementos que son extra científicos y que influyen en la investigación, por ejemplo, la motivación que tiene un investigador para indagar determinado fenómeno y no otro, su postura política o su postura filosófica... que aunque sabemos que no están contenidos directamente en la investigación, terminan afectando la manera en la que se presenta el objeto de estudio y la forma en la que se interpretan y comunican los resultados.

Siempre hay que tener presente que la realidad es muy compleja y esa complejidad no puede ser reducida a modelos de ningún tipo, y mucho menos a una tabla de contingencia de 2×2 , por eso cuando se estudian las frecuencias de dos variables dicotómicas no se debe olvidar que solo se está captando una pequeña parte de la realidad.

Me gustaría ilustrar esto con un ejemplo, supongamos que estamos estudiando la variable *Exposición al virus X* que tome valores sí o no, y la variable *Muerte* que indique si el paciente se recuperó o murió. Aunque observemos una alta frecuencia en las personas que están expuestas al virus y mueren, no se debe concluir “estar contagiado del virus X causa la muerte”, en realidad el virus *per se* no causa la muerte sino la exposición simplemente te hace propenso a posibles consecuencias cuyo desenlace puede ser fatal o recuperable. No se recomienda emitir una conclusión causal.

El investigador o la persona que reporta las conclusiones del análisis estadístico debe mantenerse humilde y decir que se está tratando de explicar el fenómeno pero que puede haber muchas otras variables o factores que no se estén considerando.

Son ampliamente debatidas las relaciones causales no solo en las ciencias sociales, sino también en las naturales. Esto es porque, aun que los datos no sean observacionales sino experimentales surge el cuestionamiento, ¿es posible controlar todos los factores relevantes que pueden influir sobre el vínculo causal? Cortés (2018) cita en su artículo: “No tenemos acceso directo al mundo externo. Lo captamos solamente a través de la expresión y la razón [...] La percepción y la acción median entre el mundo y nuestras ideas de él y nos dan la materia prima para la imaginación y el razonamiento. La elaboración resultante es un conjunto de ideas [...] Verificamos estas ideas acerca de la realidad comparándolas con datos empíricos, no con el mundo mismo”

Hay autores que afirman que es imposible fundar empíricamente el concepto de causalidad, porque si la relación causal que se quiere determinar parte de lo empírico no puede tener la dimensión de predicción. Porque el futuro es algo que no sabemos y que no vemos, por ejemplo, puede haber dos compuestos químicos que al juntarlos siempre hayan producido una reacción, pero nada asegura que mañana o en el futuro pase lo mismo porque el futuro es un terreno desconocido. Cortés (2018) dice que “no es posible derivar enunciados universales a partir de un cúmulo de enunciados particulares por numeroso que sea”.

En resumen, debemos evitar conclusiones tipo “ A causa B ” más bien se debe modular el léxico y ocupar expresiones tipo “si A entonces B es más propenso” o “existe una asociación de A con B ”.

Recordemos el dicho popular “los datos no mienten, pero se puede mentir con los datos”.

Prueba de independencia χ^2 de Pearson

En la Tabla 1 se ilustró una tabla de contingencia. Conviene generalizar para entender la teoría detrás y poder hacer los cálculos tanto para el coeficiente de correlación tetracórico como para el estadístico de independencia χ^2 . Una tabla de contingencia de 2×2 tiene la forma:

		x		
		1	2	Total:
y	1	a	b	$a + b$
	2	c	d	$c + d$
Total:		$a + c$	$b + d$	N

Tabla 2. Tabla de contingencia de 2×2 .

Es decir, que se observó a objetos/personas/ocurrencias cuando las variables x y y tomaron el valor de su primera categoría (de forma general suele usarse n_{11}), se observó b objetos/personas/ocurrencias cuando las variables x y y tomaron el valor de su primera y segunda categoría, respectivamente (de forma general suele usarse n_{12}),. Y análogamente los valores c y d (n_{21} y n_{22}),. A estos cuatro valores se les llama frecuencias observadas.

Luego, cada una de estas cuatro casillas tiene una probabilidad de ocurrencia asociada

		x		
		1	2	Total:
y	1	π_{11}	π_{12}	π_{1+}
	2	π_{21}	π_{22}	π_{2+}
Total:		π_{+1}	π_{+2}	1

Tabla 3. Probabilidades conjuntas de las variables x y y .

Donde $\{\pi_{ij}\}$, con $i, j \in \{1, 2\}$, representa la probabilidad conjunta. Si se supone independencia entre las variables x y y entonces:

$$H_0: \pi_{ij} = \pi_{i+}\pi_{+j}, \quad \forall i \text{ y } j$$

Lo que implica que las probabilidades marginales determinan las probabilidades conjuntas. Para probar H_0 se define $\mu_{ij} = N\pi_{i+}\pi_{+j}$ como las frecuencias esperadas. Aquí μ_{ij} es el valor esperado de la casilla ij asumiendo independencia. La hipótesis alternativa H_a simplemente establece que no hay independencia.

Como π_{i+} y π_{+j} son desconocidas, para estimar las frecuencias esperadas, las probabilidades marginales se sustituye por las proporciones muestrales:

$$\widehat{\mu}_{ij} = Np_{i+}p_{+j} = N \left(\frac{n_{i+}}{N} \right) \left(\frac{n_{+j}}{N} \right) = \frac{n_{i+}n_{+j}}{N}$$

Esto es el total por renglón multiplicado por el total por columna, dividido por el total de toda la muestra observada. Las $\{\widehat{\mu}_{ij}\}$ se llaman frecuencias esperadas estimadas. Retomando las frecuencias observadas de la Tabla 2, las frecuencias esperadas estimadas, bajo H_0 , son:

$$\begin{aligned} \hat{a} &= \frac{(a+b)(a+c)}{N} \\ \hat{b} &= \frac{(a+b)(b+d)}{N} \\ \hat{c} &= \frac{(c+d)(a+c)}{N} \\ \hat{d} &= \frac{(c+d)(b+d)}{N} \end{aligned}$$

(ec. 1)

Para probar la independencia en las tablas de contingencia, el estadístico χ^2 de Pearson es:

$$\chi^2 = \sum \frac{(n_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}}$$

(ec. 2)

Donde χ^2 se distribuye $\chi^2_{(I-1)(J-1)}$, es decir, chi-cuadrada con $(I-1)(J-1)$ grados de libertad, donde I y J son el total de categorías de la primera y segunda variable, respectivamente. Una vez más, retomando las frecuencias de la Tabla 2 y sustituyendo (ec. 1) en (ec. 2) el estadístico queda:

$$\chi^2 = \sum \frac{(n_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}} = \frac{(a - \hat{a})^2}{\hat{a}} + \frac{(b - \hat{b})^2}{\hat{b}} + \frac{(c - \hat{c})^2}{\hat{c}} + \frac{(d - \hat{d})^2}{\hat{d}} \sim \chi^2_{(1)}$$

(ec. 3)

Ahora bien, la prueba de que el estadístico χ^2 (ec. 2) se distribuye chi-cuadrada se puede consultar en cualquier libro de texto y parte del supuesto de que las variables se distribuyen multinomiales, como estamos bajo el estudio de variables dicotónicas el supuesto es que las frecuencias observadas siguen una distribución binomial, sin embargo, se sabe que asintóticamente esta distribución se aproxima a la normal. Esto es importante ya que para el cálculo del coeficiente de correlación tetracórico se supone que la tabla de contingencia sigue una distribución normal bivariada.

Entonces, si la prueba permite concluir que no hay evidencia para rechazar H_0 , que las dos variables son independientes, entonces su correlación es cero y, por lo tanto, no sería estadísticamente significativo calcularla. Esto se retomará nuevamente en la descripción de la metodología.

Coeficiente de correlación tetracórico

Idea general

La idea fundamental que introduce Pearson parte del hecho de que el total de individuos/objetos observados (en el ejemplo de la viruela $N = 2081$) sigue la siguiente superficie de frecuencia:

$$z = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{1}{1-r^2} \left(\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} - \frac{2rxy}{\sigma_1\sigma_2} \right)},$$

(ec. 4)

Donde x y y son dos variables continuas con desviación estándar σ_1 y σ_2 , respectivamente, y correlación r . Si observamos bien, es la función de densidad normal bivariada con medias cero y multiplicada por N , es decir, la campana está centrada en el origen y tiene N de volumen, como se muestra en la siguiente gráfica:

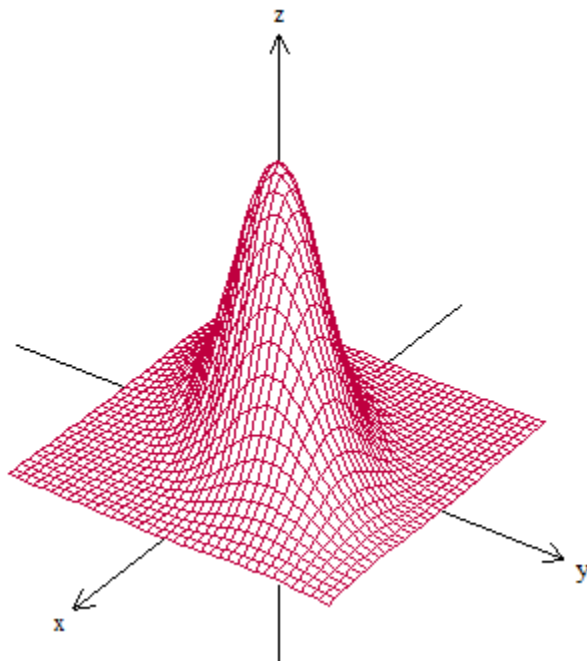


Figura 1. Superficie de frecuencia descrita por la ec. 1

Pearson propone intersectar a la campana con dos planos, uno paralelo a xz y el otro a yz , evidentemente perpendiculares entre sí, de tal forma que la campana quede cortada en cuatro secciones donde el volumen de cada una representa las frecuencias a, b, c y d observadas en la tabla de contingencia (ver Figura 2). Dicho de otro modo, la función z (ec. 4) gráficamente representa una campana de Gauss de volumen N , obsérvese que $z > 0 \forall x, y \in \mathbb{R}$, por lo que está por encima del plano xy , digamos el “piso”, ahora bien, si este “piso” tiene un punto de corte

(h', k') que lo divide en cuatro cuadrantes, el área bajo la curva de cada una de estas regiones representan las frecuencias observadas.

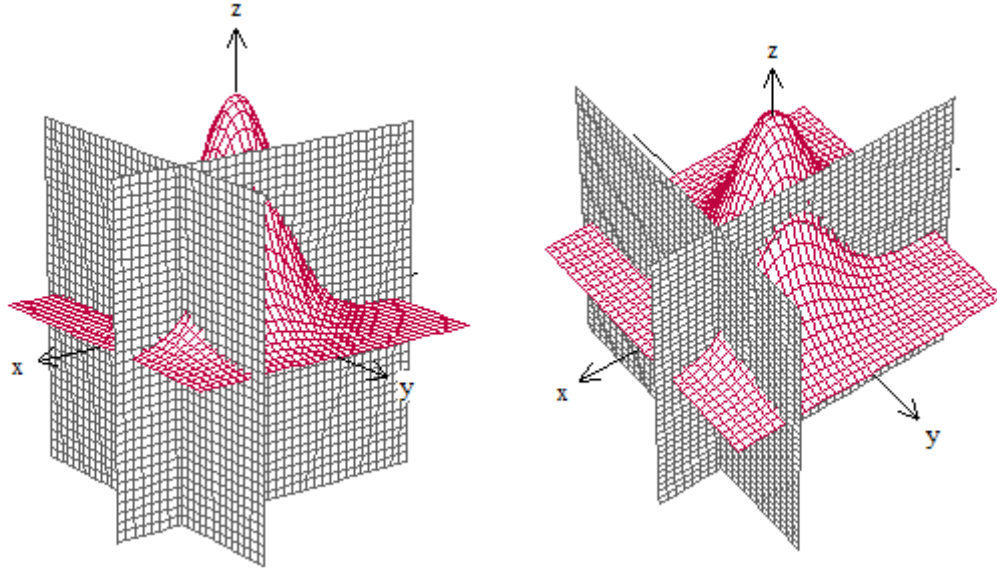


Figura 2. Campana de frecuencias intersectada por dos planos perpendiculares, vista desde dos perspectivas diferentes.

Esta es una forma de “dicotomizar” a las variables. Ahora, si vemos la Figura 2 desde arriba de tal modo que se vea el plano xy tendríamos la siguiente gráfica:

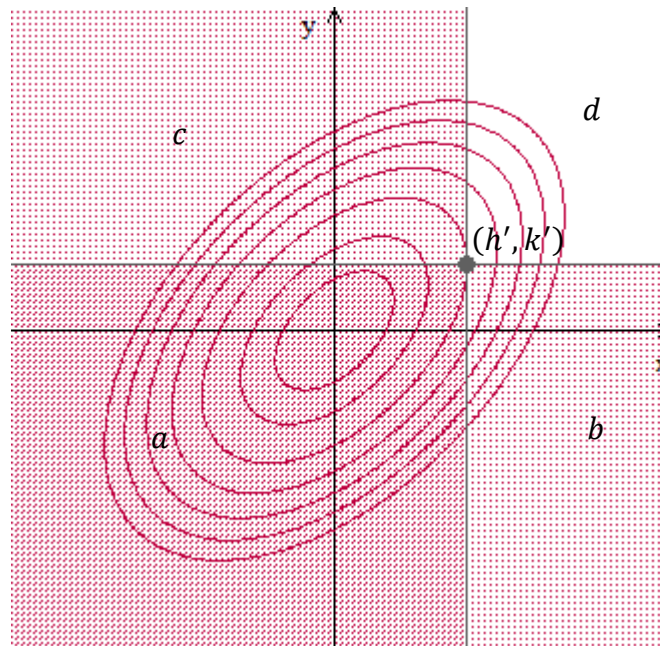


Figura 3. Plano xy cortado por las rectas $x = h'$ y $y = k'$.

Donde, $a + b + c + d = N$.

Es decir, el volumen de cada región es la frecuencia observada en la celda, de tal forma que la tabla de contingencia que se genera de esta “dicotomización” es:

		x		
		$x \leq h'$	$x > h'$	Total:
y	$y \leq k'$	a	b	$a + b$
	$y > k'$	c	d	$c + d$
Total:		$a + c$	$b + d$	N

Tabla 4. Tabla de contingencia de 2×2 , con punto de corte (h', k') en la superficie de frecuencias de (ec. 4).

La idea para conocer el coeficiente de correlación tetracórico es: si se tienen las frecuencias observadas, es decir, a, b, c y d entonces se puede conocer el punto (h, k) , donde $h = h'/\sigma_1$ y $k = k'/\sigma_1$ que “dicotomizó” a las variables y , en consecuencia, se puede conocer r en (ec. 4).

Cálculo de r

El análisis comienza en cómo encontrar el punto (h, k) , donde $h = h'/\sigma_1$ y $k = k'/\sigma_1$. Claramente:

$$\begin{aligned} d &= \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int_{h'}^{\infty} \int_{k'}^{\infty} e^{-\frac{1}{2} \frac{1}{1-r^2} \left(\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} - \frac{2rxy}{\sigma_1\sigma_2} \right)} dy dx, \\ &= \frac{N}{2\pi\sqrt{1-r^2}} \int_h^{\infty} \int_k^{\infty} e^{-\frac{1}{2} \frac{1}{1-r^2} (x^2 + y^2 - 2rxy)} dy dx, \end{aligned}$$

(ec. 5)

De lo anterior conviene comentar que con un ajuste sencillo se puede estandarizar a las variables y , sin embargo, seguir teniendo el mismo valor de correlación, lo que corrobora que es invariante a la escala de medición.

Ahora bien, obsérvese que:

$$\begin{aligned} b + d &= \int_{h'}^{\infty} \int_{-\infty}^{\infty} \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{1}{1-r^2} \left(\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} - \frac{2rxy}{\sigma_1\sigma_2} \right)} dy dx \\ &= \int_h^{\infty} \int_{-\infty}^{\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{1}{1-r^2} (x^2 + y^2 - 2rxy)} dy dx \end{aligned}$$

Pues $h = h'/\sigma_1$ y $k = k'/\sigma_1$, entonces

$$\begin{aligned} &= \int_h^{\infty} \int_{-\infty}^{\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{1}{1-r^2} (y^2 - 2rxy + r^2x^2 + x^2 - r^2x^2)} dy dx \\ &= \int_h^{\infty} \int_{-\infty}^{\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{1}{1-r^2} ((y-rx)^2 + (1-r^2)x^2)} dy dx \\ &= \int_h^{\infty} \frac{N\sqrt{2\pi}}{2\pi} e^{-\frac{1}{2} \frac{1}{1-r^2} (1-r^2)x^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-r^2}} e^{-\frac{1}{2} \frac{(y-rx)^2}{1-r^2}} dy dx \end{aligned}$$

Donde la función de la integral con respecto a y resulta ser la densidad normal $N(rx, \sqrt{1-r^2})$ y, por lo tanto, integra 1. Así que:

$$b + d = \frac{N}{\sqrt{2\pi}} \int_h^\infty e^{-\frac{1}{2}x^2} dx \quad (ec. 6)$$

Del mismo modo:

$$a + c = \frac{N}{\sqrt{2\pi}} \int_{-\infty}^h e^{-\frac{1}{2}x^2} dx, \quad (ec. 7)$$

$$c + d = \frac{N}{\sqrt{2\pi}} \int_k^\infty e^{-\frac{1}{2}y^2} dy, \quad (ec. 8)$$

$$a + b = \frac{N}{\sqrt{2\pi}} \int_{-\infty}^k e^{-\frac{1}{2}y^2} dy. \quad (ec. 9)$$

Teniendo en cuenta que la figura es simétrica La diferencia entre (ec. 7) y (ec. 6) queda:

$$(a + c) - (b + d) = 2 \frac{N}{\sqrt{2\pi}} \int_0^h e^{-\frac{1}{2}x^2} dx,$$

Y, por lo tanto,

$$\frac{(a + c) - (b + d)}{2N} = \frac{1}{\sqrt{2\pi}} \int_0^h e^{-\frac{1}{2}x^2} dx.$$

Entonces,

$$\frac{(a + c) - (b + d)}{2N} + \frac{1}{2} = \Phi(h), \quad (ec. 10)$$

Donde Φ es la función de distribución $N(0,1)$. Y del mismo modo

$$\frac{(a + b) - (c + d)}{2N} + \frac{1}{2} = \Phi(k). \quad (ec. 11)$$

Por lo tanto, cuando se conocen a, b, c y d , h y k pueden ser encontradas a través de la función de probabilidad acumulada de una normal estándar.

Ahora bien, si observamos (ec. 5) vemos que el único valor desconocido es r , pero resulta que no se puede despejar. Pearson, a través de sucesiones logra llegar a una expresión para aproximar su valor. Para ello, primeramente, propone:

$$H = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h^2}, \quad K = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}. \quad (ec. 12)$$

Y, finalmente, la serie que permite obtener r es:

$$\epsilon = \frac{ad - bc}{N^2 HK} = \sum_{i=1}^{\infty} \frac{r^i}{i!} v_{i-1} w_{i-1}$$

(ec. 13)

Donde

$$\begin{aligned} v_n &= h v_{n-1} - (n-1) v_{n-2}, \text{ con } v_0 = 1 \\ w_n &= h w_{n-1} - (n-1) w_{n-2}, \text{ con } w_0 = 1 \end{aligned}$$

Es decir²:

$$\begin{aligned} \frac{ad - bc}{N^2 HK} &= r + \frac{r^2}{2} h k + \frac{r^3}{6} (h^2 - 1)(k^2 - 1) + \frac{r^4}{24} h(h^2 - 3)k(k^2 - 3) \\ &+ \frac{r^5}{120} (h^4 - 6h^2 + 3)(k^4 - 6k^2 + 3) \\ &+ \frac{r^6}{720} h(h^4 - 10h^2 + 15)k(k^4 - 10k^2 + 15) \\ &+ \frac{r^7}{5040} (h^6 - 15h^4 + 45h^2 - 15)(k^6 - 15k^4 + 45k^2 - 15) \\ &+ \frac{r^8}{40320} h(h^6 - 21h^4 + 105h^2 - 105)k(k^6 - 21k^4 + 105k^2 - 105) + \dots \end{aligned}$$

(ec. 14)

Y resolviendo esta ecuación se conoce el coeficiente de correlación tetracórico. Es importante mencionar que la serie de la (ec. 14) siempre converge si $r < 1$, para cualesquiera valores de h y k^3 .

En resumen, inicialmente se conocen las frecuencias a, b, c y d de la tabla de contingencia, donde $a + b + c + d = N$. Primero, se obtienen los valores de h y k a través de (ec. 10) y (ec. 11). Luego, éstos se sustituyen en (ec. 12) para conocer H y K . Y, por último, se sustituyen todos los valores en (ec. 14) y se resuelve para conocer el valor del coeficiente de correlación tetracórico.

Sobra comentar que, anteriormente, el uso de este coeficiente era poco común pues su cálculo no es sencillo, sin embargo, actualmente basta con unas pequeñas líneas de código para poder obtenerlo y, sobre todo, darle uso dentro del análisis estadístico.

Comentarios sobre el cálculo de r

La (ec. 14) no es la única expresión que aproxima el valor de r . Pearson, en su mismo artículo obtiene una segunda serie, por diferente metodología, e incluso introduce un tercer método, que no concluye por su complejidad pues comenta “Parece posible que se puedan deducir desarrollos interesantes para [...] esta expresión”⁴. La segunda serie que propone es:

² se invita al lector a consultar esta demostración en la página 6 del artículo de Pearson, 1900.

³ Se invita al lector a consultar esta demostración en el tercer apartado del artículo de Pearson, 1900.

⁴ Se invita al lector a revisar la segunda sección del artículo de Pearson, 1900.

$$\frac{ad - bc}{N^2 HK} = \theta + \frac{1}{2}hk\theta^2 - (h^2 + k^2 - h^2k^2)\frac{\theta^3}{6} + hk[h^2k^2 - 3(h^2 + k^2) + 5]\frac{\theta^4}{24} + \dots \quad (ec. 15)$$

Donde $r = \text{sen}(\theta)$. Con respecto a esta expresión Pearson comenta que sugiere utilizar la serie hasta el término de θ^4 pues el siguiente es muy sensible a los valores de h y k .

Castellan (1966) explica y compara 7 expresiones diferentes (tres propuestas, de hecho, por Pearson y las otras cuatro por diversos autores) y, en la conclusión de su artículo, selecciona la siguiente de ellas como la mejor:

$$r_2 = \frac{m}{\sqrt{1 + \theta m^2}}, \quad (ec. 16)$$

Donde $\frac{a}{a+c} = \int_{-\infty}^{z_1} \varphi(w)dw$, $\frac{d}{b+d} = \int_{-\infty}^{z_2} \varphi(w)dw$, $\frac{a+c}{N} = \int_{-\infty}^x \varphi(w)dw$, con φ la función de densidad normal estándar, es decir, z_1 , z_2 y x son los valores de la abscisa en la curva normal univariada, $m = \frac{(a+c)(b+d)}{N^2} \cdot \frac{z_1+z_2}{\varphi(x)}$ y $\theta \cong 0.6$.

Hashash y El-Absy (2018) comparan en su artículo otras 7 expresiones (de varios autores) y, en su conclusión, recomiendan dos sobre los demás.

$$r_3 = \cos\left(\frac{\pi}{\delta}\right), \quad (ec. 17)$$

Donde $\delta = 1 + \sqrt{\frac{ad}{bc}}$. Y también,

$$r_4 = \cos\left(\frac{180^\circ \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}\right). \quad (ec. 18)$$

Ante tantas expresiones para estimar el coeficiente de correlación tetracórico surgen dos dudas, en primer lugar, el porqué de la existencia de tantas y, en segundo, cómo saber cuál escoger. El porqué de tantas expresiones se debe a que originalmente el cálculo era muy complejo y, entonces, se buscaron alternativas cuyas expresiones fueran más sencillas pero que, en consecuencia, pierden precisión o agregan supuestos o condiciones a las frecuencias observadas.

Cabe mencionar que el objeto de esta monografía no es el de comparar las diferentes expresiones que aproximan el valor del coeficiente de correlación tetracórico y seleccionar aquella que sea más exacta. El cálculo de la (ec. 14) es preciso y mientras más términos de la serie se consideren mayor será la exactitud.

Ekström (2009) menciona que actualmente se pueden ocupar métodos numéricos de optimización que ayuden a resolver la ecuación integral (ec. 5), por lo que encuentra obsoleto el método de expansión de series.

En la paquetería de R la función `poly.mat()` y, de la librería `Psych`, la función `tetrachoric()` ocupan el algoritmo propuesto por Kirk (1973) para aproximar numéricamente el coeficiente de correlación tetracórico por método de Máxima Verosimilitud

(MV), toma la idea central de cortar en cuatro secciones las superficie de frecuencias pero la estimación de r la trabaja a través de MV. Estas estimaciones son muy certeras y en ocasiones mejores que las de las series.

Error probable del coeficiente de r

En estadística, el error probable define el intervalo alrededor de un punto central de la distribución, de modo que la mitad de los valores de la distribución estarán dentro del intervalo y la otra mitad fuera. Por lo tanto, para una distribución simétrica es equivalente a la mitad del rango intercuartílico, o la desviación absoluta a la mediana.

El error probable del coeficiente de correlación tetracórico ($E.P._r$) se define como:

$$E.P._r = 0.67449 \sigma_r, \quad (ec. 19)$$

Donde σ_r es la desviación estándar de r y el valor 0.67449 equivale a $\Phi(3/4)$.

Pearson (1900) determinó la expresión para calcular el error probable del coeficiente de correlación tetracórico ($E.P._r$) que aplica sólo si $a + c > b + d$ y $a + b > c + d$, esta condición sucede si (h', k') es un punto en el primer cuadrante (obsérvese Figura 3)⁵:

$$E.P._r = \frac{0.67449}{\sqrt{N}\chi_0} \left[\frac{(a+d)(c+b)}{4N^2} + \psi_2^2 \frac{(a+c)(b+d)}{N^2} + \psi_1^2 \frac{(a+b)(c+d)}{N^2} \right. \\ \left. + 2\psi_1\psi_2 \frac{ad-bc}{N^2} - \psi_2 \frac{ab-cd}{N^2} - \psi_1 \frac{ac-bd}{N^2} \right]^{1/2}, \quad (ec. 20)$$

Donde

$$\beta_1 = \frac{h - rk}{\sqrt{1 - r^2}}, \quad \beta_2 = \frac{k - rh}{\sqrt{1 - r^2}}, \\ \psi_1 = \frac{1}{\sqrt{2\pi}} \int_0^{\beta_1} e^{-\frac{1}{2}z^2} dz, \quad \psi_2 = \frac{1}{\sqrt{2\pi}} \int_0^{\beta_2} e^{-\frac{1}{2}z^2} dz, \\ \chi_0 = \frac{1}{2\pi} \cdot \frac{1}{\sqrt{1 - r^2}} e^{-\frac{1}{2} \frac{1}{1 - r^2} (h^2 + k^2 - 2rhk)}.$$

Esta misma expresión puede ser utilizada para encontrar cualquier intervalo probable, no únicamente del $\alpha = 50\%$, basta con reemplazar el 0.67449 con $\Phi(1 - \alpha/2)$.

Para poder concluir la (ec. 20) obtuvo primero los errores probables de h ($E.P._h$) y de k ($E.P._k$):

⁵ Se invita al lector a consultar esta demostración en el cuarto apartado del artículo de Pearson, 1900.

$$E.P._h = \frac{0.67449}{H\sqrt{N}} \sqrt{\frac{(b+d)(a+c)}{N^2}}$$

$$E.P._k = \frac{0.67449}{K\sqrt{N}} \sqrt{\frac{(c+d)(a+b)}{N^2}}$$

(ec. 21)

Ahora bien, no debe confundirse la interpretación del intervalo probable con el de un intervalo de confianza o un intervalo de probabilidad, pues no se está llegando a ninguna distribución para el estimador ni tampoco se está asumiendo una distribución para el parámetro, más bien se toma una región que cubra al 50% de los *datos* y a partir de ahí se estima un intervalo para (h, k) y, en consecuencia, un intervalo para r .

Comentarios sobre el cálculo del error probable de r

También existen varios artículos que presentan diferentes expresiones que determinan el $E.P._r$ y los motivos son los mismos expuestos en la página 15. Pearson (1913) dijo: “Ahora bien, la fórmula anterior (refiriéndose a la (ec. 20)) para el error probable de r es ciertamente laboriosa de usar. He intentado de muchas maneras, conservando toda su precisión, darle una forma que implique cálculos menos laboriosos; sin embargo, no he logrado ninguna reducción sensible en su complejidad, tal que mantenga su completa generalidad”.

Afortunadamente, la complejidad en los cálculos no es algo que ahora nos obligue a sacrificar precisión, pues fácilmente se pueden programar las ecuaciones en una computadora y obtener los valores.

Programación de r y su error probable (por series de Pearson)

Se programó una función en R a la que se llamó `rhott` que estima el coeficiente por medio de las series propuestas por Pearson, ya que en R están las que lo estiman por MV. Ésta devuelve las estimaciones de r utilizando la (ec. 14) y (ec. 15), referidas como “Coef. Corr. 1” y “Coef. Corr. 2” en la salida, respectivamente; devuelve también las estimaciones de h y k , y de cada una de estas cuatro estimaciones se calcula y muestra el error y el intervalo probables.

Cabe mencionar lo siguiente:

- La función `rhott` requiere cuatro parámetros que son las frecuencias de la tabla de contingencias (en el orden de la Tabla 4).
- Se propone una corrección de 0.5 si el valor de la celda es cero. Esto para evitar errores al momento de encontrar las raíces de las ecuaciones.
- La (ec. 14) se programó aumentada hasta el término r^{101} para mayor precisión, sin embargo, la (ec. 15) se dejó hasta el término r^4 porque, como se comentó el siguiente término es sensible a los valores de h y k .
- El intervalo probable se muestra sólo si se cumple la condición de que (h', k') esté en el primer cuadrante, es decir, que $a + c > b + d$ y $a + b > c + d$.

```
# Función para calcular el coeficiente de correlación tetracórico -----
# Teniendo una tabla de contingencia de dos por dos del tipo
#          var 1
# -----
# var2 | cat1 | cat2
# -----
# cat1 |  a  |  b
# -----
# cat2 |  c  |  d
# Pearson (1900)
# Coef. Corr. 1 y Coef. Corr. 2 se refieren al coeficiente de correlación
# tetracórico obtenido por las series (xix) y (xxiv), respectivamente.
# Los intervalos probables se muestran sólo si a+c>b+d y a+b>c+d
rhott <- function(a,b,c,d){
  set.seed(2021)
  #Ajuste de celdas
  a = ifelse(a==0,0.5,a)
  b = ifelse(b==0,0.5,b)
  c = ifelse(c==0,0.5,c)
  d = ifelse(d==0,0.5,d)
  #Cálculo de parámetros
  N = a+b+c+d
  h = qnorm(0.5 + ((a+c)-(b+d))/(2*N))
  k = qnorm(0.5 + ((a+b)-(c+d))/(2*N))
  H = (1/sqrt(2*pi))*exp(-0.5*h^2)
  K = (1/sqrt(2*pi))*exp(-0.5*k^2)
  eps = (a*d-b*c)/(N*N*H*K)
  #Serie (xix) de Pearson, 1900, expandida hasta e término r^999
```

```

v <- vector()
w <- vector()
coef <- vector()
v[1] = h
w[1] = k
v[2] = h*v[1]-1
w[2] = k*w[1]-1
coef[1] = -1*eps
coef[2] = 1
coef[3] = v[1]*w[1]/factorial(2)
for (i in 3:101) {
  v[i] = h*v[i-1] - (i-1)*v[i-2]
  w[i] = k*w[i-1] - (i-1)*w[i-2]
  coef[i+1] = v[i-2]*w[i-2]/factorial(i)
}
r1 <- polyroot(coef)
r1 <- Re(r1[which(abs(Im(r1))<1e-12)])
r1 <- r1[which(abs(r1)<1)]
#Serie (xxiv) de Pearson, 1900, hasta el término theta^4
coef2 <- vector()
coef2[1] = -1*eps
coef2[2] = 1
coef2[3] = h*k/2
coef2[4] = (h^2+k^2-(h^2)*(k^2))/6
coef2[5] = h*k*((h^2)*(k^2)-3*(h^2+k^2)+5)/24
r2 <- polyroot(coef2)
r2 <- Re(r2[which(abs(Im(r2))<1e-12)])
r2 <- r2[which(abs(r2)<1)]
# Errores probables de h, k y r
peh <- qnorm(3/4)/(H*sqrt(N))*sqrt((b+d)*(a+c)/(N^2))
pek <- qnorm(3/4)/(K*sqrt(N))*sqrt((c+d)*(a+b)/(N^2))
per <- function (r){
  beta1 <- (h-r*k)/sqrt(1-r^2)
  beta2 <- (k-r*h)/sqrt(1-r^2)
  psi1 <- pnorm(beta1)-0.5
  psi2 <- pnorm(beta2)-0.5
  chi0 <- (1/(2*pi))*(1/sqrt(1-r^2))*exp(-0.5*(1/(1-r^2))*(h^2+k^2-2*r*h*k))
  return(qnorm(3/4)/(sqrt(N)*chi0*N)*sqrt(((a+d)*(c+b))/(4)+psi2^2*((a+c)*(b+d))+
psi1^2*((a+b)*(c+d))+2*psi1*psi2*(a*d-b*c)-psi2*(a*b-c*d)-psi1*(a*c-b*d)))
}
if(length(r1)==0 & length(r2)==0){
  return("No se pudo calcular.")
}
else{
  if(a+c>b+d & a+b>c+d){
    if(length(r1)==1 & length(r2)==1){
      x <- data.frame(Estimacion = c(r1,r2,h,k),
        P.E. = c(per(r1),per(r2),peh,pek),
        l.lim = c(r1-per(r1),r2-per(r2),h-peh,k-pek),
        u.lim = c(r1+per(r1),r2+per(r2),h+peh,k+pek),
        row.names = c("Coef. Corr. 1","Coef. Corr. 2","h","k"))

      return(x)
    }
    if(length(r1)==1 & length(r2)!=1){

```

```

    x <- data.frame(Estimacion = c(r1,h,k),
                    P.E. = c(per(r1),peh,pek),
                    l.lim = c(r1-per(r1),h-peh,k-pek),
                    u.lim = c(r1+per(r1),h+peh,k+pek),
                    row.names = c("Coef. Corr. 1","h","k"))

    return(list(x, "La serie 2 obtiene los siguientes valores:",r2))
  }
  if(length(r1)!=1 & length(r2)==1){
    x <- data.frame(Estimacion = c(r2,h,k),
                    P.E. = c(per(r2),peh,pek),
                    l.lim = c(r2-per(r2),h-peh,k-pek),
                    u.lim = c(r2+per(r2),h+peh,k+pek),
                    row.names = c("Coef. Corr. 2","h","k"))

    return(list(x, "La serie 1 obtiene los siguientes valores:",r1))
  }
  else{
    return(list("La serie 1 obtiene:",r1,"La serie 2 obtiene:",r2))
  }
}
else{
  if(length(r1)==1 & length(r2)==1){
    x <- data.frame(Estimacion = c(r1,r2,h,k),
                    P.E. = c("No aplica","No aplica",peh,pek),
                    l.lim = c("---","---",h-peh,k-pek),
                    u.lim = c("---","---",h+peh,k+pek),
                    row.names = c("Coef. Corr. 1","Coef. Corr. 2","h","k"))

    return(x)
  }
  if(length(r1)==1 & length(r2)!=1){
    x <- data.frame(Estimacion = c(r1,h,k),
                    P.E. = c("No aplica",peh,pek),
                    l.lim = c("---",h-peh,k-pek),
                    u.lim = c("---",h+peh,k+pek),
                    row.names = c("Coef. Corr. 1","h","k"))

    return(list(x, "La serie 2 obtiene los siguientes valores:",r2))
  }
  if(length(r1)!=1 & length(r2)==1){
    x <- data.frame(Estimacion = c(r2,h,k),
                    P.E. = c("No aplica",peh,pek),
                    l.lim = c("---",h-peh,k-pek),
                    u.lim = c("---",h+peh,k+pek),
                    row.names = c("Coef. Corr. 2","h","k"))

    return(list(x, "La serie 1 obtiene los siguientes valores:",r1))
  }
  else{
    return(list("La serie 1 obtiene:",r1,"La serie 2 obtiene:",r2))
  }
}
}
}
}

```

Además, se programó una función que permita simular tablas de contingencia. Cada tabla de contingencia ocupa la (ec. 4) para describir las frecuencias, es decir, se tienen dos variables con distribución normal bivariada (x, y) con medias $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, varianzas $\begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix}$ y correlación ρ que proporciona la probabilidad de ocurrencia, ahora, recordemos que se propone un punto de corte (h', k') que divide en cuatro secciones a la superficie y proporciona cuatro probabilidades que suman uno. Luego, cada una de estas cuatro probabilidades se multiplica por N para tener la frecuencia de cada celda de la tabla de contingencia de 2×2 . La ventaja de trabajar con tablas simuladas es que se conoce de antemano el parámetro que se está estimando (ρ). A continuación, se muestra el código:

```
frecuencias <- function(N,sigma1,sigma2,p,h,k){
  z <- function(x,y){
    N/(2*pi*sigma1*sigma2*sqrt(1-rhot^2))*exp(-0.5*(1/(1-rhot^2))*((x/sigma1)^2+(y/
sigma2)^2-2*rhot*x*y/(sigma1*sigma2)))
  }
  aux1 <- pbivnorm::pbivnorm(x=c(h/sigma1), y=c(k/sigma2), rho=p)
  a <- N*aux1
  aux2 <- pbivnorm::pbivnorm(x=c(Inf), y=c(k/sigma2), rho=p)
  b <- N*(aux2-aux1)
  aux3 <- pbivnorm::pbivnorm(x=c(-1*h/sigma1), y=c(-1*k/sigma2), rho=p)
  d <- N*(aux3)
  c <- N-(a+b+d)
  return(c(round(a),round(b),round(c),round(d)))
}
```

Ejemplo. Se muestra una simulación que servirá para familiarizar al lector con las funciones recién mencionadas. Los parámetros que se simularon fueron $N = 1,000$ observaciones (aunque la función arrojó al final 999 por detalles de redondeo), vector de varianzas $\begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$, correlación $\rho = 0.4$ y punto de corte $(h', k') = (4, 1)$.

En la Figura 4 se muestran los contornos de nivel y el punto de corte, las elipses no son tan excéntricas por que la correlación es débil (compare con la Tabla 5). El comando en R es:

```
N <- 1000
s1 <- 5
s2 <- 5
p <- 0.4
h_prima <- 4
k_prima <- 1
(ns <- frecuencias(N,s1,s2,p,h_prima,k_prima))
## [1] 504 75 284 136

sum(ns)
## [1] 999
```

Es decir, la tabla simulada es:

		x		
		$x \leq h'$	$x > h'$	Total:
y	$y \leq k'$	504	75	579
	$y > k'$	284	136	420
Total:		788	211	999

Tabla 5. Tabla de contingencia de 2×2 , con punto de corte (h', k') en la superficie de frecuencias de (ec. 4) con parámetros $\begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$, $\rho = 0.4$ y $(h', k') = (4, 1)$.

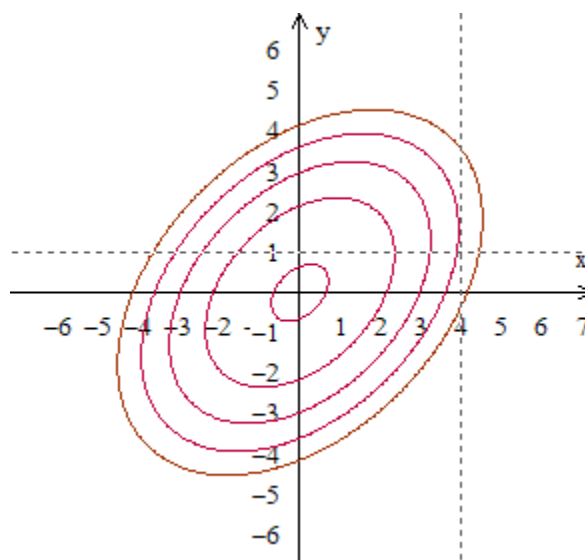


Figura 4. Elipses del contorno de la superficie de frecuencias (ec. 4) con $N = 1000$, $\sigma_1 = \sigma_2 = 5$, $\rho = 0.4$ y punto de corte $(4, 1)$.

Luego se ejecutó el siguiente comando:

```
rhot(ns[1],ns[2],ns[3],ns[4]) #Mi función

##          Estimación      P.E.      l.lim      u.lim
## Coef. Corr. 1  0.4008761 0.03349095 0.3673851 0.4343670
## Coef. Corr. 2  0.4005529 0.03349532 0.3670576 0.4340482
## h            0.8022257 0.03012104 0.7721046 0.8323467
## k            0.2008181 0.02694251 0.1738755 0.2277606
```

Se puede observar un cálculo de r casi exacto, pues el verdadero coeficiente de correlación es 0.4 y el valor estimado por ambas series es muy cercano, además de que se encuentra dentro del intervalo probable. Los valores de h y k que arroja corresponden a $\left(\frac{h'}{\sigma_1}, \frac{k'}{\sigma_2}\right) = \left(\frac{4}{5}, \frac{1}{5}\right) \approx (0.8022, 0.2008)$ que son el punto de corte escalado. Conviene recordar que “Coef. Corr. 1” y “Coef. Corr. 2” se refieren a las estimaciones por medio de las ecuaciones (ec. 14), extendida hasta el término r^{101} , y (ec. 15), respectivamente.

Por último, comparemos con:

```
(tetrachoric(matrix(c(ns[1],ns[2],ns[3],ns[4]),2,2))) #Función de la Librería psych

## Call: tetrachoric(x = matrix(c(ns[1], ns[2], ns[3], ns[4]), 2, 2))
## tetrachoric correlation
## [1] 0.4
##
```

```
## with tau of  
## [1] 0.8 0.2
```

Donde el primer valor que devuelve es la estimación de r y los siguientes dos son las coordenadas (h, k) .

Nuevamente, el objetivo de este trabajo no es el comparar funciones sino el de hacer del conocimiento del lector las diferentes opciones que tiene para la estimación en el momento de hacer su análisis.

Casos especiales de tablas de contingencia para el cálculo de r y su error probable

Antes de pasar a la aplicación del coeficiente tetracórico en ejemplos de datos reales, conviene discutir ciertas tablas de contingencia que causan problemas en el cálculo de r y del intervalo probable, recuérdese la posición de las frecuencias a, b, c y d en la Tabla 4. Tabla de contingencia de 2×2 , con punto de corte (h', k') en la superficie de frecuencias de (ec. 4). Tabla 4.

- **Una celda vacía.** Con respecto a la estimación del coeficiente de correlación tetracórico en el caso de haya una celda con frecuencia observada de 0 es más preciso, aunque con error considerable, el método MV, es decir, la función `tetrachoric()`, que las series de Pearson. Igualmente se invita al investigador a analizar el motivo por el cual se observa 0 en una celda ya que, aunque evidentemente que depende del fenómeno de estudio, es mucho más precisa la estimación si se tiene una tabla con observaciones en cada celda. Ahora bien:
 - i. En el caso de que $a = 0$ no puede estimarse el intervalo probable pues no se cumple la condición de que el punto de corte esté en el primer cuadrante, como se puede ver en la gráfica superior izquierda de la Figura 5, dado que se parte del supuesto de que la superficie de frecuencias se distribuye normal centrada no es posible que con un punto de corte en el primer cuadrante se tenga cero de volumen en la región de a . Conviene mencionar que no es estrictamente necesario que la correlación sea negativa para que suceda $a = 0$, pero sí lo facilita.
 - ii. En el caso de que $b = 0$ puede suceder que el punto de corte este muy a la derecha o muy arriba o ambas, es decir, que este en el primer o cuarto cuadrante, por este motivo habrá ocasiones en las que sí se pueda estimar el intervalo probable y otras en las que no. Por ejemplo, en la ilustración que se muestra en la parte superior derecha de la Figura 5 no se cumple la condición para estimar el error probable. Al igual que se mencionó en el punto anterior, no es estrictamente necesario que la correlación sea positiva para que $b = 0$, pero sí lo facilita.

- iii. En el caso de que $c = 0$ tampoco podrá estimarse el intervalo probable, obsérvese la ilustración de la Figura 5, en la gráfica inferior izquierda, no es posible que un punto en el primer cuadrante haga que $c = 0$. Como se comentó en los puntos anteriores, no es estrictamente necesaria una correlación positiva para $c = 0$, pero sí lo facilita.
- iv. En el caso de que $d = 0$ puede suceder que el punto esté muy arriba o muy a la derecha, es decir, habrá casos en los que sí se pueda estimar y casos en los que no. Por ejemplo, en la ilustración de la Figura 5, en la gráfica inferior derecha, sí podría calcularse el error probable pues el punto de corte está en el primer cuadrante.

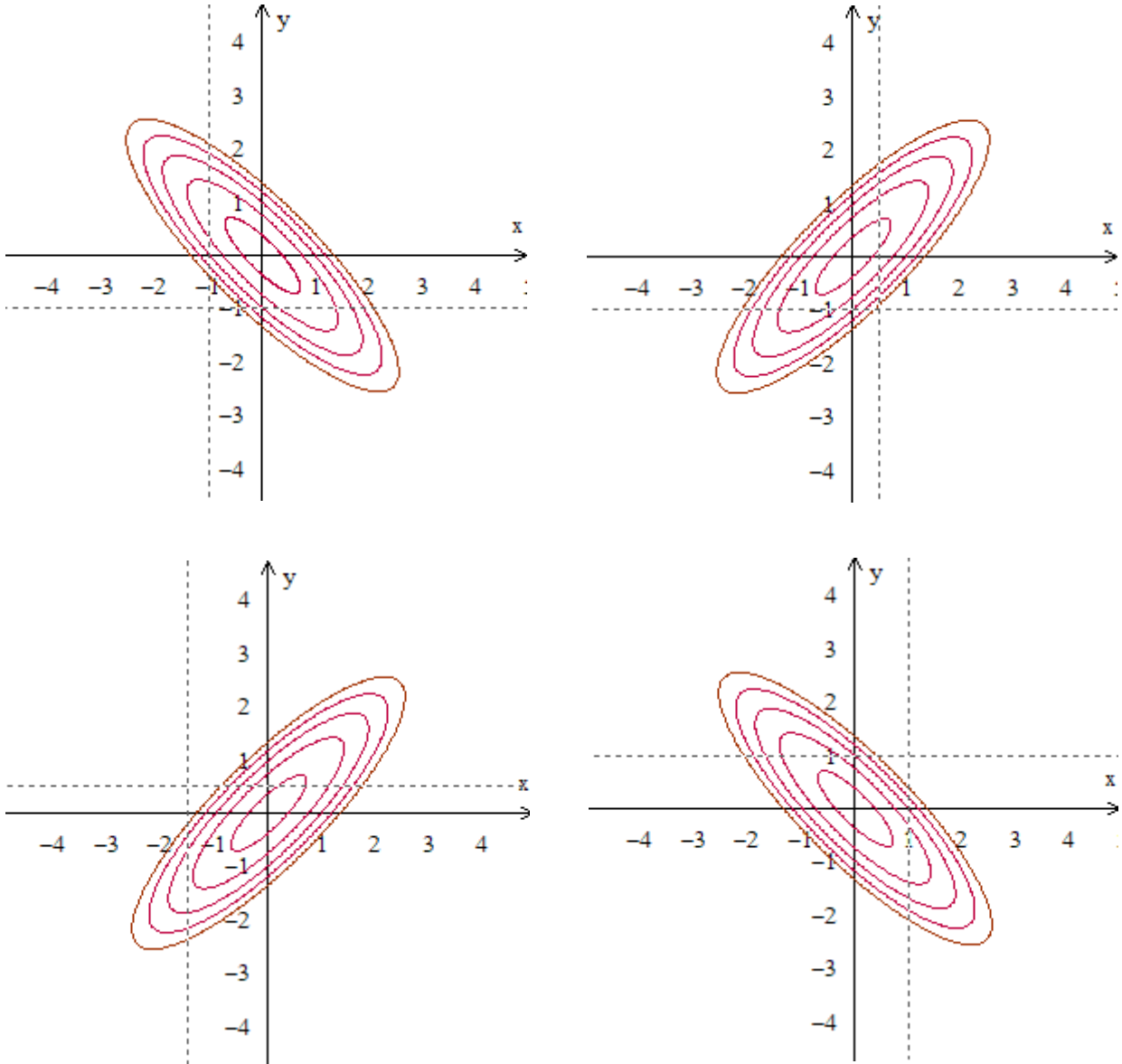


Figura 5. Ejemplos de elipses del contorno de la superficie de frecuencias (ec. 4) con $N = 1000$, $\sigma_1 = \sigma_2 = 1$. Izquierda superior: $\rho = -0.85$ y punto de corte $(-1, -1)$ lo que deriva en que $a = 0$. Derecha superior: $\rho = 0.85$ y punto de corte $(0.5, -1)$ lo que deriva en que $b = 0$. Izquierda inferior: $\rho = 0.85$ y punto de corte $(-1, 1)$ lo que deriva en que $c = 0$. Derecha inferior: $\rho = -0.85$ y punto de corte $(1, 1)$ lo que deriva en que $d = 0$.

- **Dos celdas vacías.** Para estimar el coeficiente de correlación, tanto las series de Pearson como el estimador MV funcionan igual, es tan ambigua la información en la tabla de contingencia con dos celdas sin observaciones que en la mayoría de los casos la estimación es lejana al parámetro original. En consecuencia, se invita al analista/investigador a revisar sus datos. Ahora bien:
 - i. En el caso que $a = b = 0$ o $a = c = 0$, para ilustrarlo obsérvense en la Figura 6 las dos gráficas inferiores, para que esto sea posible el punto de corte no puede estar en el primer cuadrante y, por lo tanto, no puede estimarse el intervalo probable.
 - ii. En el caso $c = d = 0$ o $b = d = 0$, para ilustrarlo obsérvense en la Figura 6 las dos gráficas superiores puede suceder que el punto de corte esté en el primer cuadrante o no, por ejemplo, en el caso $c = d = 0$, gráfica superior derecha, si $a > b$ justo como se observa, entonces sí se satisface la condición para estimar el error probable pues el punto de corte está en el primer cuadrante, pero si $a < b$, sucede que está en segundo. Análogamente sucede en el caso $b = d = 0$.
 - iii. Cabe mencionar que, si se cumple el supuesto de normalidad, entonces no es posible una tabla de contingencia con $a = d = 0$ o $c = b = 0$, si el analista/investigador está trabajando en una tabla con alguna de estas características convendrá buscar otra metodología pues el coeficiente de correlación no hará una estimación adecuada.

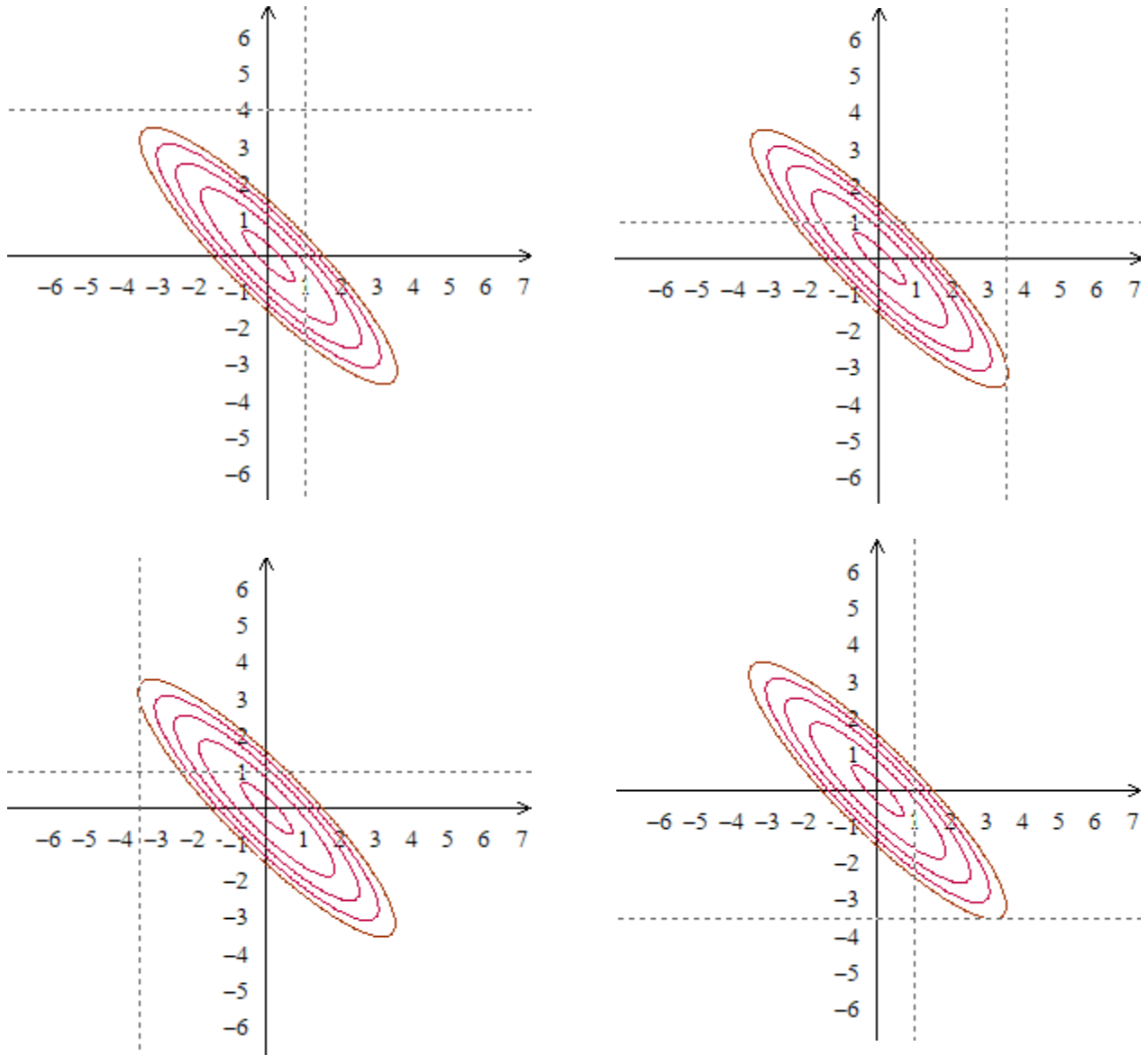


Figura 6. Ejemplos de elipses del contorno de la superficie de frecuencias (ec. 4) con $N = 1000$, $\sigma_1 = \sigma_2 = 1$ y $\rho = -0.9$. Izquierda superior: punto de corte $(1, 4)$ lo que deriva en que $c = d = 0$. Derecha superior: punto de corte $(3.5, 1)$ lo que deriva en que $b = d = 0$. Izquierda inferior: punto de corte $(1, -3.5)$ lo que deriva en que $a = b = 0$. Derecha inferior: punto de corte $(-3.5, 1)$ lo que deriva en que $a = c = 0$.

Descripción de la metodología

A continuación, se describe una propuesta para el análisis de asociación de dos variables dicotómicas:

- Una vez teniendo las frecuencias observadas a, b, c y d , como se muestra en la Tabla 2, se sugiere realizar la prueba χ^2 de Pearson para conocer si las variables son estadísticamente independientes, sin embargo, teniendo presente que esta prueba rechaza H_0 para valores grandes de N , donde surge la pregunta que todo estadístico se ha hecho en algún momento “¿cuándo N es grande?”. Por tal motivo, se invita al analista a realizar la prueba y reportarla, pero no tomarla como única estadística de referencia.
- Luego, calcular el coeficiente de correlación tetracórico, para esto se presentan las siguientes alternativas:
 - i. Utilizar cualquiera de las expresiones matemáticas descritas en la presente monografía, es decir, las ecuaciones (ec. 14), (ec. 15), (ec. 16), (ec. 17) y/o (ec. 18). Considérese los comentarios sobre la precisión de cada una.
 - ii. Ocupar la función `rho.t` cuyo código se muestra en el subtema *Programación de r y su error probable* que como ya se mencionó ocupa las ecuaciones (ec. 14) y (ec. 15).
 - iii. Ocupar la función `tetrachoric()` de la librería `psych` de R. Particularmente en los ejemplos se hará uso de ii y iii.
- Luego, calcular los errores probables para incluir en el reporte los intervalos. Para esto se ocupan las ecuaciones (ec. 20) y (ec. 21). Igualmente, la función `rho.t` las calcula.
- Por último, proporcionar una interpretación a los resultados teniendo cuidado de no hacer inferencias subjetivas ni ambiciosas, recordando también lo discutido en el subtema *Asociación no implica causalidad*.

Ejemplos

Ejemplo 1. Efecto de vacunación. Como se mencionó en el subtema *Tabla de contingencia* esta tabla corresponde a la ilustración VI del artículo de Pearson. Conviene primeramente observar el hecho de que las variables difícilmente pueden ser medidas en escalas cuantitativas. Pearson comenta que los únicos datos que se le proporcionaron fueron el número de recuperados y muertos, y la presencia o ausencia de cicatriz de la vacuna.

		Viruela		
		Se recuperó	Murió	Total:
Vacuna	Sí	1562	42	1604
	No	383	94	477
Total:		1945	136	2081

Tabla 6. Ejemplo 1. Efecto de vacunación, Pearson (1900)

Primero, se realiza la prueba de independencia χ^2 de Pearson con el siguiente comando en R

```
chisq.test(matrix(c(1562,42,383,94),2,2))
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: matrix(c(1562, 42, 383, 94), 2, 2)
## X-squared = 172.97, df = 1, p-value < 2.2e-16
```

Por lo que hay evidencia estadística para rechazar la hipótesis de independencia h_0 ya que el p – $value < 0.05$.

Luego, se calcula el coeficiente de correlación tetracórico:

```
rhot(1562,42,383,94) #Mi función

##               Estimacion      P.E.      l.lim      u.lim
## Coef. Corr. 1  0.5685753 0.02782675 0.5407486 0.5964021
## Coef. Corr. 2  0.5659664 0.02789037 0.5380760 0.5938568
## h              1.5113222 0.02869930 1.4826229 1.5400215
## k              0.7414289 0.02050633 0.7209225 0.7619352
```

Que resulta lo mismo que el comando de la librería Psych:

```
library(psych)
(tetrachoric(matrix(c(1562,42,383,94),2,2))) #función de La librería psych

## Call: tetrachoric(x = matrix(c(1562, 42, 383, 94), 2, 2))
## tetrachoric correlation
## [1] 0.6
##
## with tau of
## [1] 1.51 0.74
```

En conclusión, el coeficiente de correlación tetracórico estimado es 0.569 con un intervalo probable de (0.541,0.596), lo que permite concluir que existe una fuerte asociación entre los recuperados y la presencia de la cicatriz de la vacuna. Además, Pearson concluye: “Si bien la correlación es muy sustancial e indica el carácter protector de la vacunación, incluso después de que se incurre en la viruela, es, quizás, menor de lo que algunos partidarios fervientes de las vacunas nos hubieran hecho creer”. Lo que evidentemente nos recuerda el hecho de emitir conclusiones objetivas, realistas y no ambiciosas.

Ejemplo 2. Hasta el día 08/abril/2021 existen en México 2,267,019 personas que se han contagiado de Covid-19⁶

		Covid-19		
		Se recuperó	Murió	Total:
Género	Mujer	1,055,087	77,127	1,132,214
	Hombre	1,005,786	129,019	1,134,805
Total:		2,060,873	206,146	2,267,019

Para mejor manejo de los comandos se asignan los valores de la tabla a las variables a, b, c y d :

```
# Frecuencias observadas
```

```
a <- 1055087
```

```
b <- 77127
```

```
c <- 1005786
```

```
d <- 129019
```

```
N <- a+b+c+d
```

Se realiza la prueba de independencia χ^2 de Pearson con el siguiente comando de R:

```
# Prueba de independencia
```

```
chisq.test(matrix(c(a,b,c,d),2,2))
```

```
##
```

```
## Pearson's Chi-squared test with Yates' continuity correction
```

```
##
```

```
## data: matrix(c(a, b, c, d), 2, 2)
```

```
## X-squared = 14238, df = 1, p-value < 2.2e-16
```

Evidentemente la prueba indica rechazar la hipótesis de independencia debido a que se está trabajando con números muy grandes.

Luego se calcula el coeficiente de correlación, debido a problemas de memoria conviene escalar los datos, esto no afecta el valor del coeficiente como se podrá comprobar después con la función de la librería *psych*.

```
rhot(a/N,b/N,c/N,d/N) #Mi función
```

```
##          Estimacion          P.E.          l.lim          u.lim
## Coef. Corr. 1 0.174581234      No aplica      ---      ---
## Coef. Corr. 2 0.173009845      No aplica      ---      ---
## h          1.335033874 1.18508112392143 0.149952749970854 2.5201149978137
## k          -0.001432426 0.845347854540841 -0.846780281018206 0.8439154280634
```

⁶ Según los datos abiertos de la Dirección General de Epidemiología.

Se estima una correlación cercana a 0.2, es decir, existe una asociación débil entre ser mujer y recuperarse de covid-19. No se cumplen las condiciones para calcular el error probable.

Podemos comprobar que es similar a la estimación de la librería `Psych`:

```
(tetrachoric(matrix(c(a,b,c,d),2,2))) #Función de La Librería psych  
## Call: tetrachoric(x = matrix(c(a, b, c, d), 2, 2))  
## tetrachoric correlation  
## [1] 0.18  
##  
## with tau of  
## [1] 1.3350 -0.0014
```

En conclusión, recuperarse de covid-19 no es un evento independiente de ser hombre o mujer, es decir, al obtenerse una correlación estimada de 0.17 (con la función `rho`) o 0.18 (con la función `tetrachoric`) existe asociación entre las primeras dos categorías de las variables, estas son ser mujer y recuperarse. Dado que se sabe que la prueba de independencia χ^2 de Pearson rechaza H_0 cuando N es grande, el cálculo del coeficiente de correlación tetracórico permite concluir que las variables no son independientes.

Comparación/relación con otras técnicas similares

Existen muchos otros coeficientes que pueden ocuparse para obtener la correlación de dos categóricas, algunos de ellos para categorías nominales (como el coeficiente que se estudia en esta monografía) y otros para categorías ordinales. A continuación, se presentan dos coeficientes de asociación de variables dicotómicas y algunos comentarios acerca de sus cálculos y supuestos:

- **Coeficiente de correlación ϕ (phi):** este coeficiente parte del supuesto de que las dos variables en cuestión tienen distribución *discreta* de probabilidad. Ekström (2011) explica que Yule, alumno de Pearson, cuestionó el hecho de que el coeficiente de correlación tetracórico use como superficie de frecuencias la distribución normal bivariada diciendo “aquellos que no se vacunaron son igualmente no vacunados y, de igual forma, aquellos que murieron de viruela están igualmente muertos”, dicho de otro modo, con el coeficiente de correlación ϕ se propone que la superficie de frecuencias sean cuatro planos horizontales (a distintas alturas) y cuyo volumen integre N .

Ekström (2011) prueba en su artículo que $r = 0 \Leftrightarrow \phi = 0$, y si $r = 1 \Leftrightarrow \phi = 1$. Concluye su artículo comentando que ambos coeficientes son muy similares en el sentido de que los dos implican una distribución conjunta en el plano real, con la diferencia de que r asume la distribución normal y ϕ cualquier discreta. También comenta que el investigador/analista no debe sentirse ansioso por cuál de los dos coeficientes usar pues ambos tienen un principio teórico parecido u cualquiera que se elija, no dará derivará en conclusiones sustancialmente diferentes.

La expresión de este coeficiente adaptada a la Tabla 4 de contingencia que se ha venido manejando en la presente monografía es:

$$r_{\phi} = \frac{P(a) - P(a + b)}{\sqrt{P(a + b)P(a + c)P(c + d)P(b + d)}}$$

(ec. 22)

Donde P es la probabilidad de ocurrencia.

Aplicándolo al ejemplo 1 (Tabla 6) queda:

$$r_{\phi} = \frac{\frac{1562}{2081} - \frac{1604}{2081}}{\sqrt{\frac{1604}{2081} \times \frac{1945}{2081} \times \frac{477}{2081} \times \frac{136}{2081}}} \approx 0.3$$

Podemos ver que, aunque permite concluir que los eventos no son independientes, en este caso el coeficiente ϕ calcula una asociación mucho menor.

- **Coeficiente V de Cramér:** medida de asociación entre dos variables nominales (no necesariamente dicotómicas). Chen (2002) explica que este coeficiente refleja la razón de

la estadística observada y el máximo de la estadística. Como resultado toma valores entre 0 y 1. La expresión para calcularla es:

$$V = \sqrt{\frac{\chi^2}{N \times (\min(r, c) - 1)}} = \sqrt{\frac{\chi^2}{N}}$$

(ec. 23)

Donde χ^2 es la estadística de independencia de Pearson sin corrección de Yates (ec. 3), y $\min(r, c)$ es el menor número de columnas o renglones, como se está trabajando con variables dicotómicas este mínimo siempre será 2, por lo tanto se puede simplificar la expresión.

Aplicándolo al ejemplo 1 (Tabla 6) queda:

$$V = \sqrt{\frac{175.76}{2081}} \approx 0.3$$

De igual forma resulta diferente de cero lo que permite concluir asociación entre las variables, sin embargo, se obtiene una estimación menor que r .

Conclusión

Referencias

Brown, M. B. (1977, 1 septiembre). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, 42.

https://link.springer.com/article/10.1007/BF02293655?error=cookies_not_supported&code=8caa2f8b-d04f-4ea5-8ac4-c7f02e04893e

Castellan, J. N. (1966, 1 marzo). On the estimation of the tetrachoric correlation coefficient. *Psychometrika*, 31.

https://link.springer.com/article/10.1007/BF02289458?error=cookies_not_supported&code=bfdad515-5f1c-4e5f-a3d5-95bde6a94aae

Chen, P. Y., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric Measures: 139* (1.^a ed.). Sage Publications, Inc.

Choi, J. (2010, 30 noviembre). ERIC - EJ902636 - Correlational Analysis of Ordinal Data: From Pearson's «r» to Bayesian Polychoric Correlation, Asia Pacific Education Review, 2010-Dec. *Asia Pacific Educ. Rev.*, 11. <https://eric.ed.gov/?id=EJ902636>

Ekstrom, J. (2009, 15 mayo). Contributions to the Theory of Measures of Association for Ordinal Variables. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Science*, 50(2009). <https://uu.diva-portal.org/smash/get/diva2:210896/FULLTEXT01.pdf>

Ekström, J. (2011, 25 octubre). The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate. *UCLA Department of Statistics Papers*.

<https://escholarship.org/uc/item/7qp4604r>

El-Hashash, E. F. (2018, 31 diciembre). Methods for Determining the Tetrachoric Correlation Coefficient for Binary Variables | Asian Journal of Probability and Statistics. *Asian Journal of Probability and Statistics*.

Asian Journal of Probability and Statistics.

<https://www.journalajpas.com/index.php/AJPAS/article/view/28782>

Kirk, D. B. (1973, 1 junio). On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient. *Psychometrika*, 38.

https://link.springer.com/article/10.1007/BF02291118?error=cookies_not_supported&code=27202d8d-2cde-40cd-a143-66db4089d9b6

Odell, C. W. (1926). The interpretation of the probable error and the coefficient of correlation. *University of Illinois at Urbana-Champaign*, 32.

<https://www.ideals.illinois.edu/bitstream/handle/2142/28259/interpretationof32odel.pdf?sequence=1&isAllowed=y>

Pearson, K. (1900, 1 marzo). VII. On the correlation of characters not quantitatively measurable. *Royal Society Publishing*.

<https://royalsocietypublishing.org/doi/10.1098/rsta.1900.0022>

Pearson, K. (1913, 1 marzo). On the Probable Error of a Coefficient of Correlation as found from a Fourfold Table. *Biometrika*, 9. [https://academic.oup.com/biomet/article-](https://academic.oup.com/biomet/article-abstract/9/1-2/22/325854)

[abstract/9/1-2/22/325854](https://academic.oup.com/biomet/article-abstract/9/1-2/22/325854)