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

Reconocimientos

Contenido

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 sétimo 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 ejemplo[1]:

		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 $S\big(f(X), g(Y)\big) = S(X, Y)$.

[1] Ilustración VI de Pearson(1900)

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

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 se ser cercana a 1); y valores grandes de la primera variable están asociados con valores pequeños de la segunda (llamada correlación inversa en caso se 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:

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

Tabla 2. Tabla de contintencia 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

		y		
		1	2	**Total:**
x	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:

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

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:

$$\hat{a} = \frac{(a+b)(a+c)}{N}$$
$$\hat{b} = \frac{(a+b)(b+d)}{N}$$
$$\hat{c} = \frac{(c+d)(a+c)}{N}$$
$$\hat{d} = \frac{(c+d)(b+d)}{N}$$

(ec. 1)

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

$$\chi^2 = \sum \frac{\left(n_{ij} - \widehat{\mu_{ij}}\right)^2}{\widehat{\mu_{ij}}}$$

(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:

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

(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 mulinomiales, 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:

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

(ec. 4)

Donde x y y son dos variables contínuas 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:

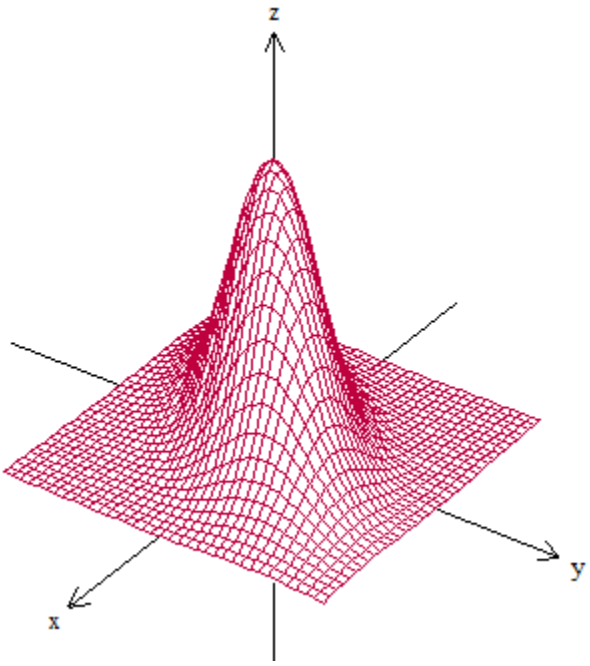

Figura 1. Superficie de frecuencia descrita por la ec. 1

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.

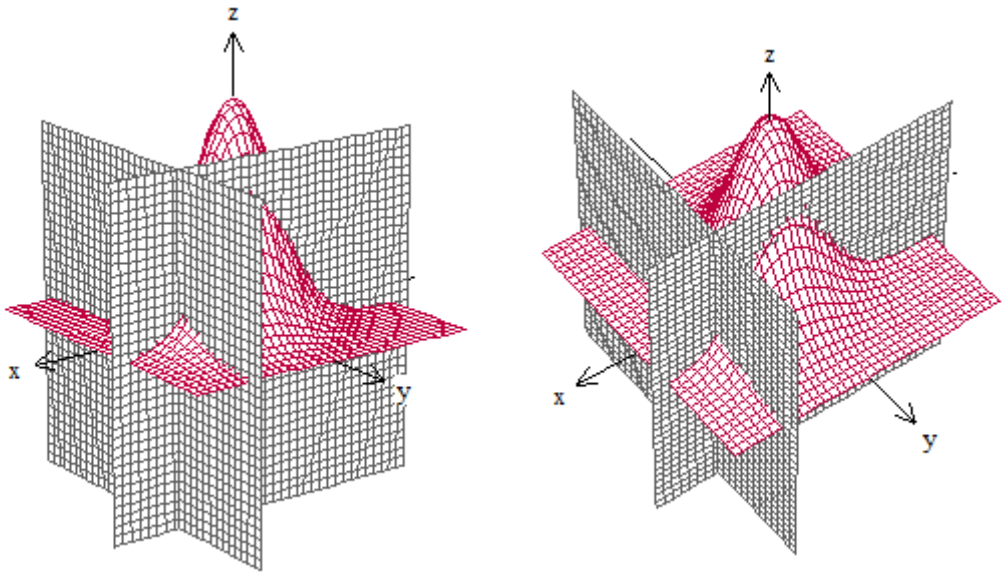

Figura 2. Campana de frecuencias intersectada por dos planos perpendiculares, vista desde dos perspectivas diferentes.

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:

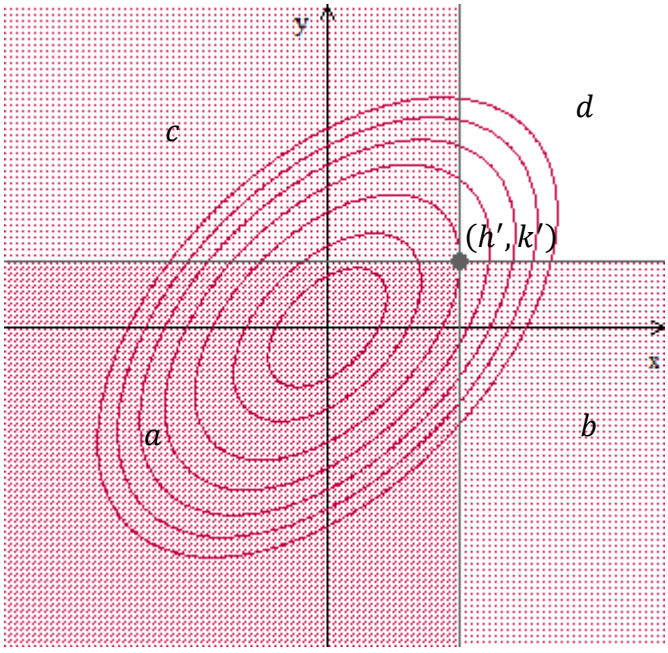

Figura 3. Plano xy cortado por las rectas $x = h'$ y $y = k'$.

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

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:

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

$$= \frac{N}{2\pi\sqrt{1-r^2}} \int_{h}^{\infty} \int_{k}^{\infty} e^{-\frac{1}{2}\cdot\frac{1}{1-r^2}\left(x^2+y^2-2rxy\right)} dy\, dx,$$

(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:

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

$$= \int_{h}^{\infty} \int_{-\infty}^{\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2}\cdot\frac{1}{1-r^2}\left(x^2+y^2-2rxy\right)} dy\, dx$$

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

$$= \int_{h}^{\infty} \int_{-\infty}^{\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2}\cdot\frac{1}{1-r^2}\left(y^2-2rxy+r^2x^2+x^2-r^2x^2\right)} dy\, dx$$

$$= \int_{h}^{\infty} \int_{-\infty}^{\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2}\cdot\frac{1}{1-r^2}\left((y-rx)^2+(1-r^2)x^2\right)} dy\, dx$$

$$= \int_{h}^{\infty} \frac{N\sqrt{2\pi}}{2\pi} e^{-\frac{1}{2}\cdot\frac{1}{1-r^2}(1-r^2)x^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{(y-rx)^2}{1-r^2}} dy\, dx$$

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

$$b + d = \frac{N}{\sqrt{2\pi}} \int_{h}^{\infty} e^{-\frac{1}{2}\cdot x^2} dx$$

(ec. 6)

Del mismo modo:

$$a + c = \frac{N}{\sqrt{2\pi}} \int_{-\infty}^{h} e^{-\frac{1}{2} \cdot x^2} dx,$$

(ec. 7)

$$c + d = \frac{N}{\sqrt{2\pi}} \int_{k}^{\infty} e^{-\frac{1}{2} \cdot y^2} dy,$$

(ec. 8)

$$a + b = \frac{N}{\sqrt{2\pi}} \int_{-\infty}^{k} e^{-\frac{1}{2} \cdot y^2} dy.$$

(ec. 9)

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

$$(a + c) - (b + d) = 2 \frac{N}{\sqrt{2\pi}} \int_{0}^{h} e^{-\frac{1}{2} \cdot x^2} dx,$$

Y, por lo tanto,

$$\frac{(a + c) - (b + d)}{2N} = \frac{1}{\sqrt{2\pi}} \int_{0}^{h} e^{-\frac{1}{2} \cdot x^2} dx.$$

Entonces,

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

(ec. 10)

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

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

(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:

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

(ec. 12)

Y finalmente[2]:

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

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

(ec. 13)

Y resolviendo esta ecuación se conoce el coeficiente de correlación tetracórico. Es importante mencionar que la serie de la (ec. 13) 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. 13) 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. 13) 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:

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

(ec. 14)

Donde $r = sen\,(\theta)$. 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:

[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.

$$r_2 = \frac{m}{\sqrt{1+\theta m^2}},$$

(ec. 15)

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.

$$r_3 = \cos\left(\frac{\pi}{\delta}\right),$$

(ec. 16)

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

$$r_4 = \cos\left(\frac{180°\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right).$$

(ec. 17)

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. 13) 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 librería PSYCH ocupa el algoritmo propuesto por Kirk (1973) para aproximar numéricamente el coeficiente de correlación tetracórico.

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:

$$E.P._r = 0.67449\ \sigma_r,$$

(ec. 18)

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]:

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

(ec. 19)

Donde

$$\beta_1 = \frac{h - rk}{\sqrt{1-r^2}}, \qquad \beta_2 = \frac{k - rh}{\sqrt{1-r^2}},$$

$$\psi_1 = \frac{1}{\sqrt{2\pi}}\int_0^{\beta_1} e^{-\frac{1}{2}z^2}dz, \qquad \psi_2 = \frac{1}{\sqrt{2\pi}}\int_0^{\beta_2} e^{-\frac{1}{2}z^2}dz,$$

$$\chi_0 = \frac{1}{2\pi}\cdot\frac{1}{\sqrt{1-r^2}}e^{-\frac{1}{2}\cdot\frac{1}{1-r^2}(h^2+k^2-2rhk)}.$$

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

Para poder concluir la (ec. 19) obtuvo primero los errores probables de h ($E.P._h$) y de k ($E.P._k$):

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

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

(ec. 20)

Ahora bien, no debe confundirse la interpretación de un intervalo de probabilidad con el de un intervalo de confianza, más adelante se retomará este tema y se explicarán los posibles usos e interpretaciones que se dan al intervalo de probabilidad.

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. 19)) para el error probable de r es ciertamente laboriosa de usar. He intentado de muchas maneras, conservando toda su precisión, darle una forma que

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

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.

Diferentes usos del error probable

Programación de r y su error probable

Se programaron las ecuaciones (ec. 13) y (ec. 14), y sus errores de probabilidad, en una función a la que llamé `rhot`. Ésta devuelve las estimaciones del coeficiente de correlación tetracórico utilizando la (ec. 13), referida como "Serie 1", y también la (ec. 14), referida como "Serie 2", devuelve también las estimaciones de h y k, de cada una de estas cuatro estimaciones se calcula el error probable y el intervalo de probabilidad.

Cabe mencionar lo siguiente:

- Se propone una corrección de 0.5 si el valor de la celda es cero. Esto porque se observó que producía error con la función `uniroot.all.`
- Hay casos en los que las series de las ecuaciones recién mencionadas tienen más de una solución en el intervalo $(-1,1)$, en estos casos la función `rhot` devuelve todos los valores pero entonces ya no da intervalo de probabilidad. Estos casos se discuten en los ejemplos.
- Para poder ocupar la función se requiere la librería `rootSolve`

```
rhot <- function(a,b,c,d){
  set.seed(2021)
  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)
  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)
  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))
  eps = (a*d-b*c)/(N*N*H*K)
  coef1 = 1
  coef2 = h*k/factorial(2)
  coef3 = (h^2-1)*(k^2-1)/factorial(3)
  coef4 = h*(h^2-3)*k*(k^2-3)/factorial(4)
  coef5 = (h^4-6*h^2+3)*(k^4-6*k^2+3)/factorial(5)
  coef6 = h*(h^4-10*h^2+15)*k*(k^4-10*k^2+15)/factorial(6)
  coef7 = (h^6-15*h^4+45*h^2-15)*(k^6-15*k^4+45*k^2-15)/factorial(7)
  coef8 = h*(h^6-21*h^4+105*h^2-105)*k*(k^6-21*k^4+105*k^2-105)/factorial(8)
  serie1 <- function(x){
    return(-1*eps+coef1*x+coef2*x^2+coef3*x^3+coef4*x^4+coef5*x^5+coef6*x^6+coef7*x
^7+coef8*x^8)
  }
  r1 <-  uniroot.all(serie1, c(-1,1))
  serie2 <- function(x){
    return(-1*eps+ x + h*k/2*x^2-(h^2+k^2-(h^2)*(k^2))/6*x^3 + h*k*((h^2)*(k^2)-3*(
h^2+k^2)+5)/24*x^4)
```

```
  }
  tet <-  uniroot.all(serie2, c(-1,1))
  r2 <- sin(tet)
  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)==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))
  }
  if(length(r1)==0 & length(r2)==0){
    return("No se pudo calcular.")
  }
  else{
    return(list("La serie 1 obtiene:",r1,"La serie 2 obtiene:",r2))
  }
}
```

Para mostrar los resultados de la función `rhot` se simularon varias tablas de contingencia. La ventaja de trabajar con tablas simuladas es que se conoce de antemano el parámetro que se está estimando. Para esto se ocupó la superficie de frecuencias descrita en (ec. 2), dando valores arbitrarios para las constantes. A continuación, se muestra el código:

```
frecuencias <- function(N,sigma1,sigma2,rhot,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/s1), y=c(k/s2), rho=p)
  a <- N*aux1
  aux2 <- pbivnorm::pbivnorm(x=c(Inf), y=c(k/s2), rho=p)
  b <- N*(aux2-aux1)
  aux3 <- pbivnorm::pbivnorm(x=c(h/s1), y=c(Inf), rho=p)
  c <- N*(aux3-aux1)
  d <- N*(1-aux2-aux3+aux1)
  return(c(round(a),round(b),round(c),round(d)))
}
```

El uso de estas dos funciones se muestra en los ejemplos del subtema Descripción de la metodología.

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 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. 13), (ec. 14), (ec. 15), (ec. 16) y/o (ec. 17). Considérese los comentarios sobre la precisión de cada una.
    ii. Ocupar la función `rhot` cuyo código se muestra en el subtema *Programación de r y su error probable* que como ya se mencionó ocupa las ecuaciones (ec. 13) y (ec. 14). Particularmente en los ejemplos se hará uso de esta función.
    iii. Ocupar la función `tetrachoric()` de la librería `psych` de R.
- Luego, calcular los errores probables para incluir en el reporte los intervalos. Para esto se ocupan las ecuaciones (ec. 19) y (ec. 20). Igualmente, la función `rhot` 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. Herencia del color del pelaje en caballos. Lo siguiente representa la distribución de los toros y potras en 1050 cajas de caballos de carreras de pura sangre, la agrupación se hizo en todos los colores de pelaje clasificados como "castaño y más oscuro", "castaño y más claro".

		Semental		
		Marrón y más oscuro	*Castaño y más claro*	**Total:**
Yegua	*Marrón y más oscuro*	631	125	756
	Castaño y más claro	147	147	294
Total:		778	272	1050

Comparación/relación con otras técnicas similares

Conclusión

Apéndice

Referencias