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 proporciona diferentes coeficientes 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 estas variables 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 estas 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 está memoria”.

En el presente trabajo se explicarán dos coeficientes de correlación, el coeficiente phi y el coeficiente tetracórico de Pearson. Se explicará

Los datos de dos variables dicotómicas suele presentarse en tablas de contingencia de 2×2.

Por último, conviene mencionar el hecho de que una fuerte correlación entre dos variables no debe interpretarse como un efecto de causalidad.

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 anterior 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 1562 pacientes que sí recibieron la vacuna y se recuperaron de la viruela.

Conviene generalizar para entender la teoría detrás y poder hacer los cálculos del coeficiente de correlación tetracórico, entonces una tabla de contingencia de 2×2 tiene la forma:

		$\boldsymbol{y}$		
				Total:
$\boldsymbol{x}$		a	b	$a+b$
		c	d	$c+d$
Total:		$a+c$	$b+d$	N

Tabla 2. Tabla de contintencia de 2×2.

Medida de asociación

Si dos variables son dependientes, entonces una intuitivamente proporciona información de la otra. Correlación o dependencia es cualquier relación estadística, causal o no, 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 la correlación se refiere a cualquiera de varios tipos específicos de operaciones matemáticas entre las variables y sus respectivos valores esperados.

[1] Ilustración VI de Pearson(1900)

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

Prueba de independencia χ^2 de Pearson

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. 1)

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:

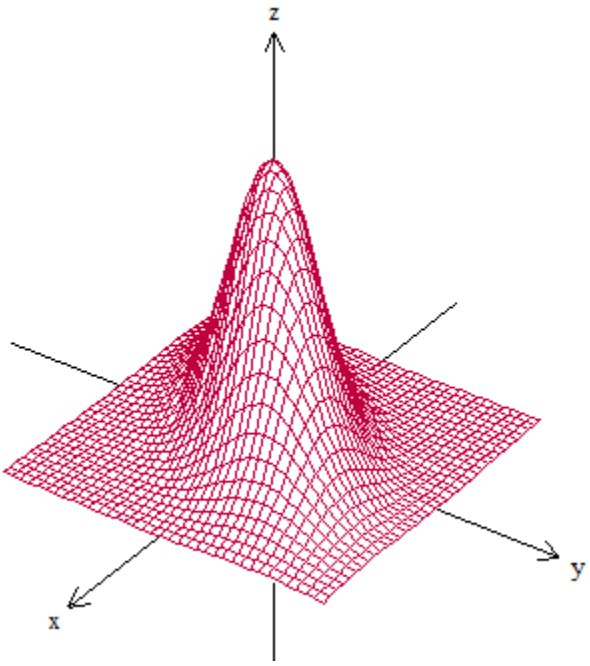

Figura 1. Superficie de frecuencia descrita por la ec. 1

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. 1) 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.

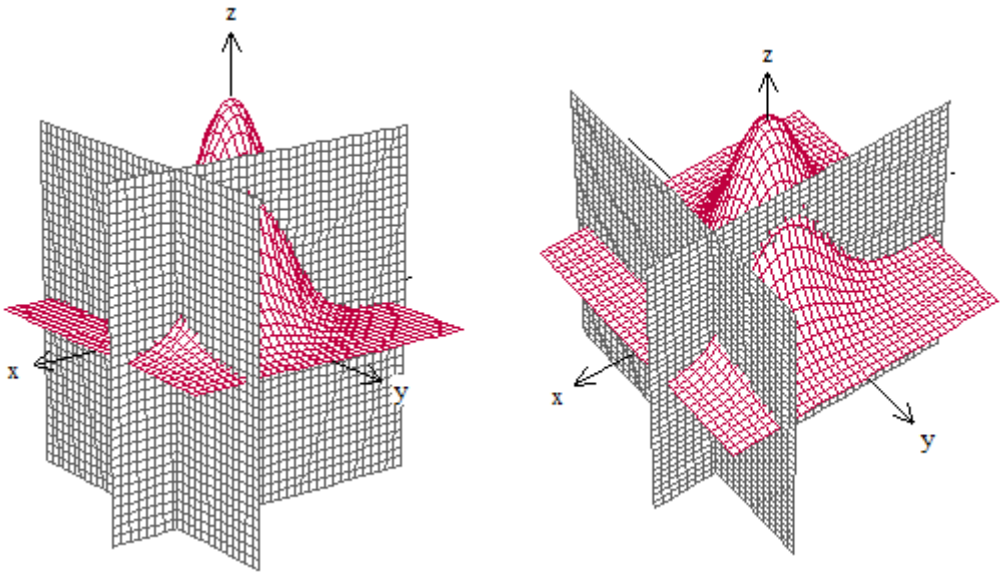

Figura 2. Campana de frecuencias intersectada por dos planos perpendiculares, vista desde dos perspectivas diferentes.

Esta es una forma de "dicotomizar" a las variables. Ahora, si vemos la Figura 2**¡Error! No se encuentra el origen de la referencia.** desde arriba de tal modo que se vea el plano xy tendríamos la siguiente gráfica:

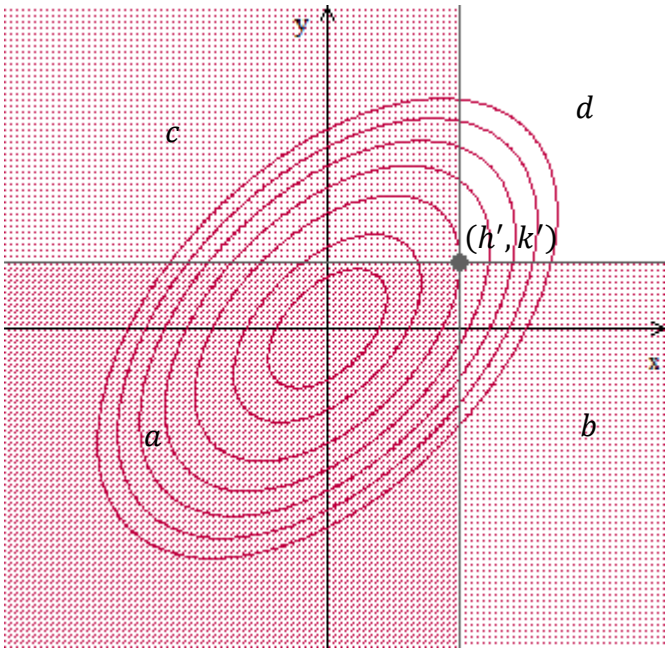

Figura 3. Plano xy cortado por las rectas $x = h'$ y $y = k'$.

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. 1).

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. 2)

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.

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

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

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. 3)

Del mismo modo:

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

(ec. 4)

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

(ec. 5)

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

(ec. 6)

Teniendo en cuenta que la figura es simétrica La diferencia entre (ec. 4) y (ec. 3) 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,

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

Entonces,

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

(ec. 7)

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. 8)

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. 2) 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. 9)

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. 10)

Y resolviendo esta ecuación se conoce el coeficiente de correlación tetracórico. Es importante mencionar que la serie de la (ec. 10) 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. 7) y (ec. 8). Luego, éstos se sustituyen en (ec. 9) para conocer H y K. Y, por último, se sustituyen todos los valores en (ec. 10) 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. 10) no es la única expresión que aproxima el valor de r. Pearson, en su mismo artículo obtiene otras dos expresiones diferentes para estimarlo. 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:

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

(ec. 11)

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.

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

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

(ec. 12)

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

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

(ec. 13)

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. 10) 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. 2), 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. 14)

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 $a + b > c + d$ [4]:

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

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

(ec. 15)

Donde

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

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

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

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

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 13. Pearson (1913) dijo: "Ahora bien, la fórmula anterior (refiriéndose a la (ec. 15)) 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.

Diferentes usos del error probable

Descripción de la metodología

Comparación/relación con otras técnicas similares

Conclusión

Apéndice

Desarrollo detallado de la **¡Error! No se encuentra el origen de la referencia.**:

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

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

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

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

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

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

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:

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

Referencias