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

presenta:

Brenda Corina Cerezo Silva

DIRECTORA DE TESINA:

M. EN C. LETICIA EUGENIA GRACIA
MEDRANO VALDELAMAR

México, 2021

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.

Resumen

El coeficiente de correlación tetracórico se utiliza para conocer la asociación entre dos variables dicotómicas. Surge del supuesto de que las dos variables se distribuyen como una normal bivariada y propone un "discretización" a partir de cortes hechos con dos planos perpendiculares.

Índice general

Índice de figuras

Capítulo 1

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 lineal 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 éstas 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) está correlacionada con la variable resultado (dicotómica por el hecho de describir la cualidad de si el paciente *se recuperó* o *murió*). En psicología muchos trastornos sólo pueden ser medidos en términos de, digamos, *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* [9], 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 explicará el coeficiente de correlación recién mencionado, algunos comentarios relativos a su cálculo, junto con una metodología sugerida para su implementación.

Capítulo 2

Conocimientos preliminares

2.1. 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 o frecuencias para cada clasificación. Por ejemplo [1]:

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

Tabla 2.1: Datos de la viruela recuperados por Karl Pearson (1900)

En la Tabla 2.1 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 1562 pacientes que sí recibieron la vacuna y se recuperaron de la viruela.

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, 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.

[1] Ilustración VI de Pearson(1900). En la sección 4 del presente trabajo se analiza esta tabla de contingencia.

Ekström (2009) [4] enlista las propiedades que satisface una medida de asociación $S(X, Y)$, donde X y Y son dos variables aleatorias:

1. S está definida para cualquier par de variables aleatorias.
2. $S(X, Y) = S(Y, X)$.
3. $-1 \leq S(X, Y) \leq 1, S(X, X) = 1, S(X, -X) = -1$.
4. Si X y Y son independientes, entonces $S(X, Y) = 0$.
5. $S(-X, Y) = S(X, -Y) = -S(X, Y)$.
6. Si f y g son funciones casi seguramente, estrictamente crecientes, entonces $S\left(f(X), g(Y)\right) = S(X, Y)$.
7. Si (X, Y) y $\{X_n, Y_n\}_{n=1}^{\infty}$ 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 $lim_{n\to\infty} S(X_n, Y_n) = S(X, Y)$.

Las propiedades 3 y 4, 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 5 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.

Ahora bien, la correlación entre X y Y se define como

$$r = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}},$$

donde $cov(\cdot)$ y $var(\cdot)$ son la covarianza y la varianza de las variables aleatorias. En realidad no es una medida de asociación pues no satisface la propiedad 6, pero sí una variante de ésta en la que las funciones sean estrictamente un polinomio de primer grado. Por tal motivo, el coeficiente de correlación es comunmente referido como una medida de asociación lineal.

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 se ser cercana a -1).

2.3. Asociación no implica causalidad

Fernando Cortés en su artículo "Observación, causalidad y explicación causal" (2018) [3] 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, aunque 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) [3] 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) [3] 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".

2.4. Prueba de independencia χ^2 de Pearson

En la Tabla 2.1 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 de la Tabla 2.2.

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

Tabla 2.2: 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. La suma de los elementos de cada renglón, $a+b$ y $c+d$, suele representarse como n_{+i}, y la suma de los elementos de cada columna, $a+c$ y $b+d$, como n_{i+} con $i \in \{1,2\}$.

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 2.3: Probabilidades conjuntas de las variables x y y.

Donde $\{\pi_{ij}\}$, con $i,j \in \{1,2\}$, representa la probabilidad conjunta, es decir, $\pi_{ij} = P(X=i, Y=j) = \pi_{ij}$; $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$ con $i,j \in \{1,2\}$ representan la probabilidad marginal, es decir, $\pi_{i+} = P(X=i) = \sum_j P(X=i, Y=j)$ y $\pi_{+j} = P(Y=j) = \sum_i P(X=i, Y=j)$.

Si se supone independencia entre las variables x y y entonces:

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

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 las variables no son independientes.

Como π_{i+} y π_{+j} son desconocidas, para estimar las frecuencias esperadas, las probabilidades marginales se sustituyen por las proporciones muestrales [2]:

$$\widehat{\mu_{ij}} = Np_{i+}p_{+j} = N\left(\frac{n_{i+}}{N}\right)\left(\frac{n_{+j}}{N}\right) = \frac{n_{i+}n_{+j}}{N} \tag{2.1}$$

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 2.2, las frecuencias esperadas estimadas, bajo H_0, son:

$$\begin{aligned}
\widehat{a} &= \frac{(a+b)(a+c)}{N} \\
\widehat{b} &= \frac{(a+b)(b+d)}{N} \\
\widehat{c} &= \frac{(c+d)(a+c)}{N} \\
\widehat{d} &= \frac{(c+d)(b+d)}{N}
\end{aligned} \tag{2.2}$$

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

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij} - \widehat{\mu_{ij}})^2}{\widehat{\mu_{ij}}} \tag{2.3}$$

Donde χ^2 es una estadística que se distribuye $\chi^2_{(I-1)(J-1)}$, es decir, ji-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 2.2 y sustituyendo (2.2) en (2.3) el estadístico queda:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij} - \widehat{\mu_{ij}})^2}{\widehat{\mu_{ij}}} = \frac{(a-\widehat{a})^2}{\widehat{a}} + \frac{(b-\widehat{b})^2}{\widehat{b}} + \frac{(c-\widehat{c})^2}{\widehat{c}} + \frac{(d-\widehat{d})^2}{\widehat{d}} \sim \chi^2_{(1)} \tag{2.4}$$

Ahora bien, la prueba de que el estadístico χ^2 (2.3) se distribuye ji-cuadrada se puede consultar en cualquier libro de texto y parte del supuesto de que las variables se distribuyen multinomiales. En este caso se trata del estudio de variables dicotómicas así que el supuesto es que las frecuencias observadas siguen una distribución binomial, sin embargo, se sabe que si np y $n(1-p)$ son ambos grandes, es decir, que n sea grande y p no esté cercano a 1 ni a 0, asintóticamente esta distribución se aproxima a la normal, es decir, $Bin(n,p) \approx N\left(np, np(1-p)\right)$. 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.

[2]Conviene aclarar que esta forma de estimar los valores esperados resulta de una distribución asociada a la tabla de contingencia.

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 importante calcularla. Esto se retomará nuevamente en el capítulo 4.

Capítulo 3

Coeficiente de correlación tetracórico

3.1. 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:

$$z = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2rxy}{\sigma_1\sigma_2}\right)} \tag{3.1}$$

Donde x y y son dos variables continuas con desviación estándar σ_1 y σ_2, respectivamente, y correlación r. Si observamos bien, se trata de 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:

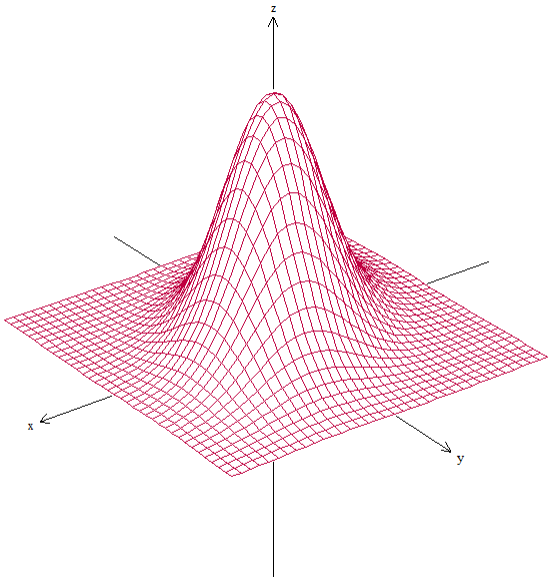

Figura 3.1: Superficie de frecuencia descrita por la ec. 3.1

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 3.2). Dicho de otro modo, la función z (ec. 3.1) gráficamente representa una campana de Gauss de volumen N, obsérvese que $z > 0 \quad \forall \quad x, y \in R$, 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.

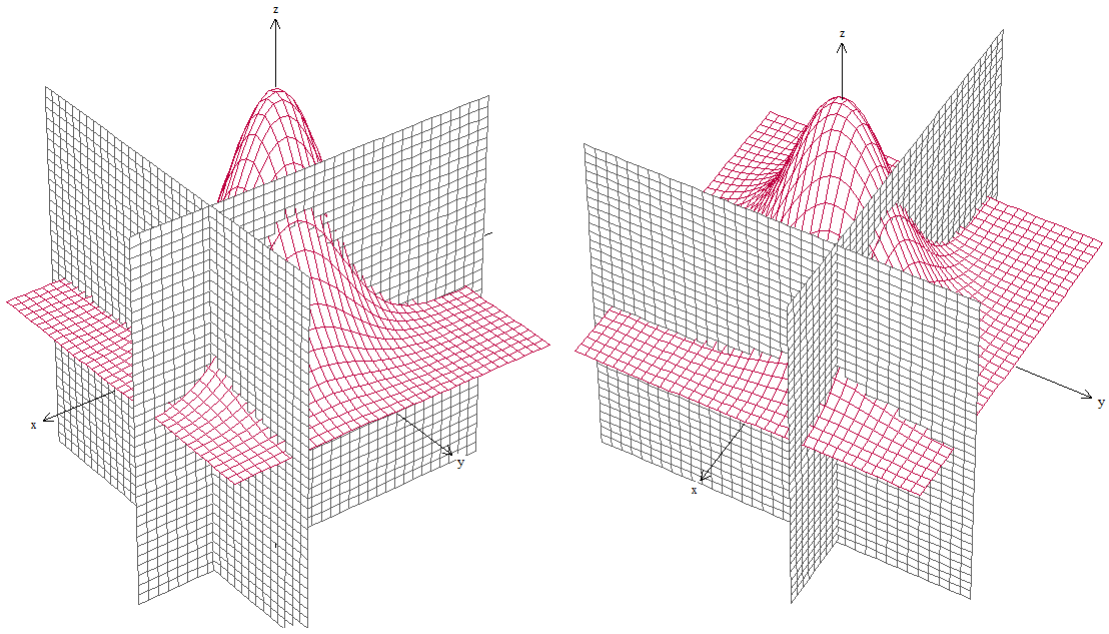

Figura 3.2: Campana de frecuencias intersectada por dos planos perpendiculares, vista desde dos perspectivas diferentes.

Esta es una forma de "dicotomizar" a las variables. Ahora, si vemos la Figura 3.2 desde arriba de tal modo que se vea el plano xy, resulta la Figura 3.3, 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 3.1: Tabla de contingencia de 2×2, con punto de corte (h', k') en la superficie de frecuencias de (ec. 3.1).

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

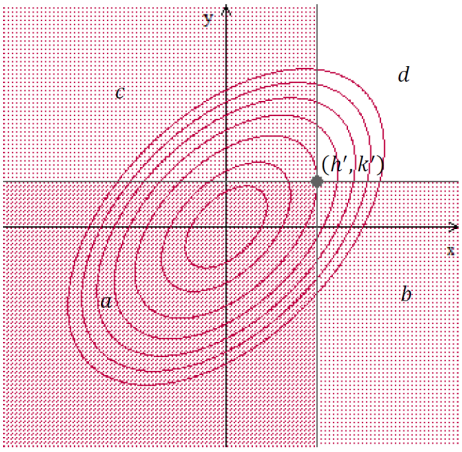

Figura 3.3: Plano xy cortado por las rectas $x = h'$ y $y = k'$.

$h = h'/\sigma_1$ y $k = k'/\sigma_2$ que "dicotomizó" a las variables y, en consecuencia, se puede conocer r en (ec. 3.1) como se explica en la sección 3.2.

3.2. Cálculo de r

El análisis comienza en cómo encontrar el punto (h, k), donde $h = h'/\sigma_1$ y $k = k'/\sigma_2$. Primeramente, el valor d se puede calcular de la siguiente forma (ver Tabla 3.1 y Figura 3.3):

$$\begin{aligned} d &= \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int_{h'}^{+\infty}\int_{k'}^{+\infty} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2rxy}{\sigma_1\sigma_2}\right)}\,dy\,dx \\ &= \frac{N}{2\pi\sqrt{1-r^2}} \int_{h}^{+\infty}\int_{k}^{+\infty} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(x^2+y^2-2rxy\right)}\,dy\,dx, \end{aligned} \tag{3.2}$$

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, lo que corrobora que es invariante a la escala de medición.

Ahora bien, obsérvese que

$$\begin{aligned} b + d &= \int_{h'}^{+\infty}\int_{-\infty}^{+\infty} \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2rxy}{\sigma_1\sigma_2}\right)}\,dy\,dx \\ &= \int_{h}^{+\infty}\int_{-\infty}^{+\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(x^2+y^2-2rxy\right)}\,dy\,dx, \end{aligned} \tag{3.3}$$

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

$$\begin{aligned}
&= \int_h^{+\infty} \int_{-\infty}^{+\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(y^2-2rxy+r^2x^2+x^2-r^2x^2\right)}\, dy\, dx, \\
&= \int_h^{+\infty} \int_{-\infty}^{+\infty} \frac{N}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left((y-xy)^2+x^2(1-r^2)\right)}\, dy\, dx, \\
&= \int_h^{+\infty} \frac{N\sqrt{2\pi}}{2\pi} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(x^2(1-r^2)\right)} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{(y-xy)^2}{1-r^2}}\, dy\, dx,
\end{aligned} \tag{3.4}$$

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:

$$b + d = \frac{N}{\sqrt{2\pi}} \int_h^{+\infty} e^{-\frac{1}{2}x^2}\, dx. \tag{3.5}$$

Del mismo modo:

$$a + c = \frac{N}{\sqrt{2\pi}} \int_{-\infty}^{h} e^{-\frac{1}{2}x^2}\, dx, \tag{3.6}$$

$$c + d = \frac{N}{\sqrt{2\pi}} \int_k^{+\infty} e^{-\frac{1}{2}y^2}\, dy, \tag{3.7}$$

$$a + b = \frac{N}{\sqrt{2\pi}} \int_{-\infty}^{k} e^{-\frac{1}{2}y^2}\, dy. \tag{3.8}$$

Teniendo en cuenta que la figura es simétrica la diferencia entre las ecuaciones (3.6) y (3.5) queda:

$$(a + c) - (b + d) = 2\frac{N}{\sqrt{2\pi}} \int_0^{h} e^{-\frac{1}{2}x^2}\, dx, \tag{3.9}$$

y, por lo tanto,

$$\frac{(a + c) - (b + d)}{2N} = \frac{1}{\sqrt{2\pi}} \int_0^{h} e^{-\frac{1}{2}x^2}\, dx. \tag{3.10}$$

Entonces

$$\frac{(a + c) - (b + d)}{2N} + \frac{1}{2} = \phi(h), \tag{3.11}$$

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

$$\frac{(a+b)-(c+d)}{2N}+\frac{1}{2}=\phi(k). \tag{3.12}$$

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. 3.2) 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:

$$H=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}h^2}, \qquad K=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}k^2}. \tag{3.13}$$

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

$$\epsilon=\frac{ad-bc}{N^2HK}=\sum_{n=1}^{\infty}\frac{r^n}{n!}v_{n-1}w_{n-1}, \tag{3.14}$$

donde $v_n=hv_{n-1}-(n-1)v_{n-2}$, $w_n=kw_{n-1}-(n-1)w_{n-2}$ con $v_0=w_0=1$, es decir,

$$\begin{aligned}\frac{ad-bc}{N^2HK}=r&+\frac{r^2}{2}hk+\frac{r^3}{6}(h^2-1)(k^2-1)+\frac{r^4}{24}h(h^2-3)k(k^2-3)\\&+\frac{r^5}{120}(h^4-6h^2+3)(k^4-6k^2+3)\\&+\frac{r^6}{720}h(h^4-10h^2+15)k(k^4-10k^2+15)\\&+\frac{r^7}{5040}(h^6-15h^4+45h^2-15)(k^6-15k^4+45k^2-15)\\&+\frac{r^8}{40320}h(h^6-21h^4+105h^2-105)k(k^6-21k^4+105k-105)+...\end{aligned} \tag{3.15}$$

Y resolviendo esta ecuación polinomial se conoce el coeficiente de correlación tetracórico. Es importante mencionar que la serie de la (ec. 3.15) siempre converge [1] si $r<1$, para cualesquiera valores de h y k .

Con respecto a esta serie, lo que Pearson hace en su artículo es trabajar con la expresión (3.2) de tal forma que se tiene

$$\begin{aligned}\frac{d}{N}&=\frac{1}{2\pi\sqrt{1-r^2}}\int_h^{+\infty}\int_k^{+\infty}e^{-\frac{1}{2}\frac{1}{1-r^2}\left(x^2+y^2-2rxy\right)}\,dy\,dx,\\&=\frac{1}{2\pi\sqrt{1-r^2}}\int_h^{+\infty}\int_k^{+\infty}U\,dy\,dx.\end{aligned} \tag{3.16}$$

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

Una vez definida U se hace la expansión de la serie de potencias de Taylor en $r = 0$, se obtiene el logarítmo de la serie (lo que conviene pues se simplifican las exponenciales) y después se distribuyen las integrales a cada término de la suma. Se invita al lector a consultar esta demostración en la página 6 del artículo de Pearson, 1900 [9].

En resumen, inicialmente se conocen las frecuencias observadas 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. 3.11) y (ec. 3.12). Luego, éstos se sustituyen en (ec. 3.13) para conocer H y K. Y, por último, se sustituyen todos los valores en (ec. 3.15) y se resuelve el polinomio 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 algunas líneas de código para poder obtenerlo y, sobre todo, darle uso dentro del análisis estadístico.

3.3. Comentarios sobre el cálculo de r

La ecuación (3.15) 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" [2].

A continuación, se presentan diferentes expresiones para la estimación del coeficiente de correlación, para evitar confusión r es el coeficiente estimado por (ec. 3.15) y, conforme se vayan presentando, se les irá definiendo como r_2, r_3, y así sucesivamente.

La segunda serie que propone es:

$$\frac{ad - bc}{N^2HK} = \theta + \frac{1}{2}hk\theta^2 - (h^2 + k^2 - h^2k^2)\frac{\theta^3}{6} + hk\left[h^2k^2 - 3(h^2 + k^2) + 5\right]\frac{\theta^4}{24} - ..., \quad (3.17)$$

y se define $r_2 = sen(\theta)$. 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) [1] 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:

$$r_3 = \frac{m}{\sqrt{1 + \theta m^2}}, \quad (3.18)$$

donde $\frac{a}{a+c} = \int_{-\infty}^{z_1} \varphi(w)\, dw$, $\frac{d}{b+d} = \int_{-\infty}^{z_2} \varphi(w)\, dw$ y $\frac{a+c}{N} = \int_{-\infty}^{x} \varphi(w)\, dw$ 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) [6] comparan en su artículo otras 7 expresiones (de varios autores) y, en su conclusión, recomiendan dos sobre los demás.

$$r_4 = cos\left(\frac{\pi}{\delta}\right), \quad (3.19)$$

[2] Se invita al lector a revisar la segunda sección del artículo de Pearson, 1900

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

$$r_5 = cos\left(\frac{180\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right). \tag{3.20}$$

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 ecuación (3.15) es preciso y mientras más términos de la serie se consideren mayor será la exactitud.

Ekström (2009) [4] menciona que actualmente se pueden ocupar métodos numéricos de optimización que ayuden a resolver la ecuación integral (3.2), 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()` utiliza el algoritmo propuesto por Kirk (1973) [7] para aproximar numéricamente el coeficiente de correlación tetracórico por método de Máxima Verosimilitud (MV), toma la idea central de cortar en cuatro secciones la superficie de frecuencias pero la estimación de r la trabaja a través de MV. Estas estimaciones son muy precisas y, en ocasiones, mejores que las de las series.

3.4. 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 [8]. 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:

$$E.P._r = 0.67449\sigma_r, \tag{3.21}$$

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

Pearson (1900) [9] determinó la expresión para calcular el $E.P._r$ 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 3.3)[3]:

$$E.P._r = \frac{0.67449}{\sqrt{N}\chi_0}\left[\frac{(a+d)(c+b)}{4N^2} + \psi_2^2\frac{(a+c)(b+d)}{N^2} + \psi_1^2\frac{(a+b)(c+d)}{N^2} + 2\psi_1\psi_2\frac{ad-bc}{N^2} - \psi_2\frac{ab-cd}{N^2} - \psi_1\frac{ac-bd}{N^2}\right]^{\frac{1}{2}} \tag{3.22}$$

donde

$$\begin{aligned} \beta_1 &= \frac{h-rk}{\sqrt{1-r^2}}, \qquad \beta_2 = \frac{k-rh}{\sqrt{1-r^2}}, \\ \psi_1 &= \frac{1}{\sqrt{2\pi}}\int_0^{\beta_1} e^{-\frac{1}{2}z^2}\,dz, \qquad \psi_2 = \frac{1}{\sqrt{2\pi}}\int_0^{\beta_2} e^{-\frac{1}{2}z^2}\,dz, \\ \chi_0 &= \frac{1}{2\pi}\cdot\frac{1}{\sqrt{1-r^2}}e^{-\frac{1}{2\pi}\cdot\frac{1}{1-r^2}(h^2+k^2-2rhk)}. \end{aligned} \tag{3.23}$$

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

Para poder concluir la ecuación (3.22) se obtuvieron primero los errores probables de h ($E.P._h$) y de k ($E.P._k$):

$$\begin{aligned} E.P._h &= \frac{0.67449}{H\sqrt{N}}\sqrt{\frac{(b+d)(a+c)}{N^2}}, \\ E.P._k &= \frac{0.67449}{K\sqrt{N}}\sqrt{\frac{(c+d)(a+b)}{N^2}}. \end{aligned} \tag{3.24}$$

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.

3.5. 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 sección 3.3. Pearson (1913) [10] dijo: "Ahora bien, la fórmula anterior (refiriéndose a la ecuación (3.22)) 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

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

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.

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

Se programó una función en R a la que se llamó `rhot` 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 (3.15) y (3.17), 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:

1. La función `rhot` requiere cuatro parámetros que son las frecuencias de la tabla de contingencias (en el orden de la Tabla 3.1).

2. 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.

3. La ecuación (3.15) se programó aumentada hasta el término r^{101} para mayor precisión, sin embargo, la ecuación (3.17) se dejó hasta el término r^4 porque, como se comentó, el siguiente término es sensible a los valores de h y k.

4. 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
```

```
rhot <- 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 el 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 ecuación (3.1) para describir las frecuencias, es decir, se tienen dos variables con distribución normal bivariada (x, y) con medias $\binom{0}{0}$, varianzas $\binom{\sigma_1}{\sigma_2}$ 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 familiarizarse con las funciones recién mencionadas. Los parámetros que se simularon fueron $N = 1000$ observaciones (aunque la función arrojó al final 999 por detalles de redondeo), vector de varianzas $\binom{\sigma_1}{\sigma_2} = \binom{5}{5}$, correlación $\rho = 0.4$ y punto de corte $(h', k') = (4, 1)$.

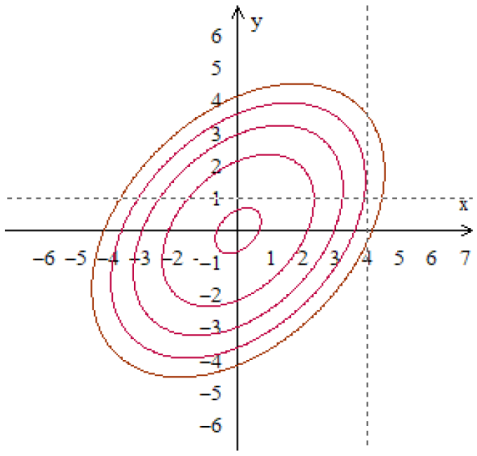

Figura 3.4: Elipses del contorno de la superficie de frecuencias (3.1) con $N = 1000$, $\sigma_1 = \sigma_2 = 5$, $\rho = 0.4$ y punto de corte $(4, 1)$.

En la Figura 3.4 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 3.2). El comando en R 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 3.2: Tabla de contingencia de 2×2, con punto de corte (h', k') en la superficie de frecuencias de (ec. 3.1) con parámetros $\binom{\sigma_1}{\sigma_2} = \binom{5}{5}$, $\rho = 0.4$ y $(h', k') = (4, 1)$.

```
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))
```

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 corresponen a $(h'/\sigma_1, k'/\sigma_2) = (4/5, 1/5) \approx (0.8022, 0.2008)$ 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 (3.15), extendida hasta el término r^{101}, y (3.17), 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 dar a conocer las diferentes opciones que se tienen para la estimación del coeficiente de correlación tetracórico.

3.7. 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 3.1.

- **Una celda vacía**. Con respecto a la estimación del coeficiente de correlación tetracórico en el caso de que 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 a analizar el motivo por el cual se observa 0 en una celda ya que, aunque evidentemente depende del fenómeno de estudio, es mucho más precisa la estimación si se tiene una tabla con observaciones en cada celda. Ahora bien:

  - 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 3.5, 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.
  - 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 esté 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 3.5 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.
  - En el caso de que $c = 0$ puede suceder que el punto de corte se encuentre en segundo cuadrante o en el primero pero muy arriba, si es el primer caso tampoco podrá estimarse el intervalo probable, obsérvese la ilustración de la Figura 3.5, en la gráfica inferior izquierda. Como se comentó en los puntos anteriores, no es estrictamente necesaria una correlación positiva para $c = 0$, pero sí lo facilita.
  - 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 el intervalo

probable y casos en los que no. Por ejemplo, en la ilustración de la Figura 3.5, en la gráfica inferior derecha, sí podría calcularse el error probable pues el punto de corte está en el primer cuadrante.

- **Dos celdas vacías**. Para estimar el coeficiente de correlación si hay dos celdas vacías, 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:

  - En el caso que $a = b = 0$ o $a = c = 0$, para ilustrarlo obsérvense las dos gráficas inferiores en la Figura 3.6, 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.
  - En el caso $c = d = 0$ o $b = d = 0$, para ilustrarlo obsérvense las dos gráficas superiores en la Figura 3.6, 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 $b = d = 0$.
  - 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.

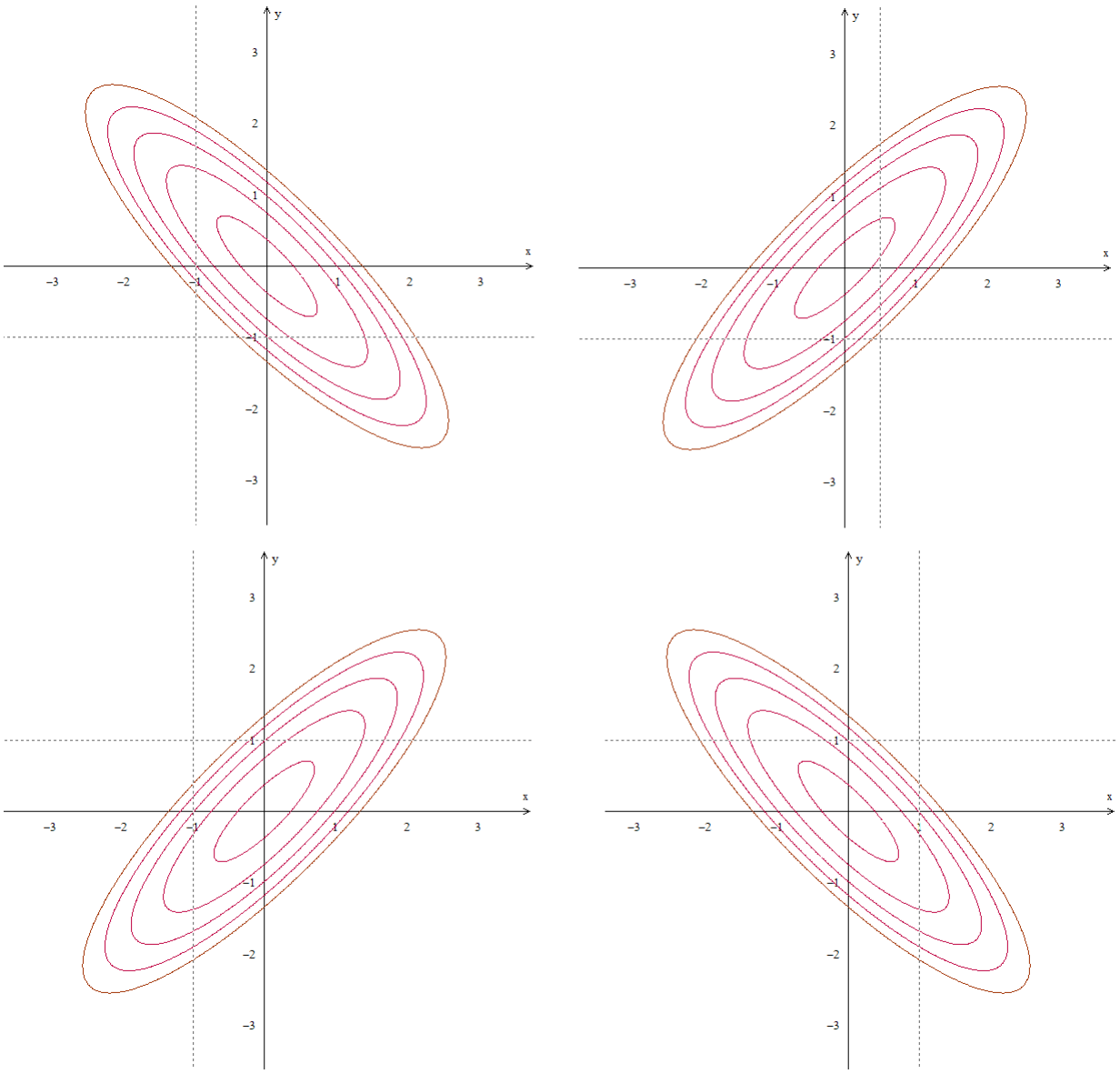

Figura 3.5: Ejemplos de elipses del contorno de la superficie de frecuencias (3.1) con $N = 1000$, $\sigma_1 = \sigma_2 = 1$. Izquierda superior: $\rho = -0.85$ y punto de corte $(-1, -1)$ lo que deriva en que $a = 0$. Derecha superior: $\rho = 0.85$ y punto de corte $(0.5, -1)$ lo que deriva en que $b = 0$. 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 $d = 0$.

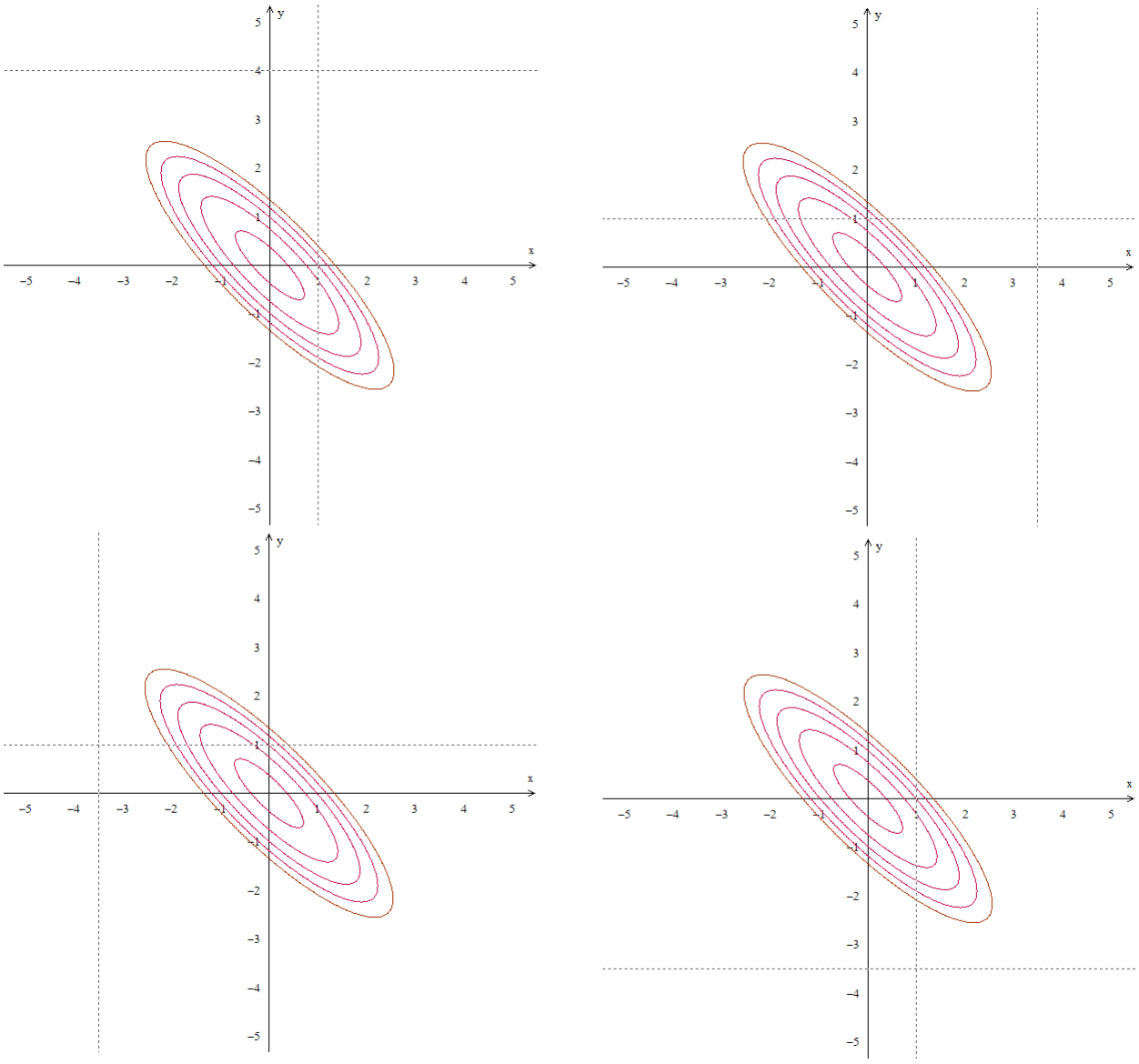

Figura 3.6: Ejemplos de elipses del contorno de la superficie de frecuencias (3.1) con $N = 1000$, $\sigma_1 = \sigma_2 = 1$ y $\rho = -0.9$ Izquierda superior: punto de corte $(1, 4)$ lo que deriva en que $c = d = 0$. Derecha superior: punto de corte $(3.5, 1)$ lo que deriva en que $b = d = 0$. Izquierda inferior: punto de corte $(-3.5, 1)$ lo que deriva en que $a = c = 0$. Derecha inferior: punto de corte $(1, -3.5)$ lo que deriva en que $a = b = 0$.

Capítulo 4

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 2.2, 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 [1]:
  1. Utilizar cualquiera de las expresiones matemáticas descritas en el presente documento, es decir, las ecuaciones (3.15), (3.17), (3.18), (3.19) y/o (3.20). Considérese los comentarios sobre la precisión de cada una.
  2. Ocupar la función `rhot` cuyo código se muestra en el subtema 3.6 que, como ya se mencionó, ocupa las ecuaciones (3.15) y (3.17).
  3. 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 (3.22) y (3.24). 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 2.3.

4.1. Ejemplos

Ejemplo 1. Efecto de vacunación. Como se mencionó en el subtema 2.1 esta tabla corresponde a la ilustración VI del artículo de Pearson (1900) [9]. Conviene primero

[1] Particularmente en los ejemplos se hará uso de 2. y 3.

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 4.1: 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
```

Se puede comparar con el comando de la librería `Pysch`:

```
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
```

```
# Se imprimen los valores con más decimales para poder comparar
# En el valor rho viene el coeficiente
(tetrachoric(matrix(c(1562,42,383,94),2,2))$rho)

## [1] 0.595824

# En el valor tau vienen (h,k)
(tetrachoric(matrix(c(1562,42,383,94),2,2))$tau)

## [1] 1.5113222 0.7414289
```

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 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 claramente 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-19 [2]:

		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

Tabla 4.2: Ejemplo 2. Casos de covid-19 por género en México, fecha de corte 08/abril/2021.

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:

[2]Según los datos abiertos de la Dirección General de Epidemiología

```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`.

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

```
##                  Estimacion              P.E.               l.lim
## Coef. Corr. 1  0.174581234         No aplica                 ---
## Coef. Corr. 2  0.173009845         No aplica                 ---
## h              1.335033874  1.18508112392143  0.149952749970854
## k             -0.001432426 0.845347854540841 -0.846780281018206
##                          u.lim
## Coef. Corr. 1              ---
## Coef. Corr. 2              ---
## h              2.52011499781372
## k             0.843915428063476
```

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 similiar a la estimación de la librería `Pysch`:

```r
(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 `rhot`) o 0.18 (con la función `tetrachoric`) se puede concluir que existe asociación entre las

primeras categorías de las variables dado que la correlación es positiva [3], que son ser mujer y recuperarse, o visto desde otra perspectiva, entre las segundas categorías de las variables, que son ser hombre y morir. 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 muestra que las variables no son independientes y, aunque la asociación no es sustanciosa, por el hecho de estar trabajando con grandes cantidades la diferencia entre hombres y mujeres resulta evidente.

[3]Obsérvese la Figura 3.3, al ser positiva la correlación la asociación lineal se da en las regiones correspondientes a a y d.

Capítulo 5

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 este documento) 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:

1. **Coeficiente de correlación r_φ (phi)**: este coeficiente parte del supuesto de que las dos variables en cuestión tienen distribución discreta de probabilidad. Ekström (2011) [5] 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".

   Ekström (2011) [5] prueba en su artículo que $r = 0 \Leftrightarrow r_\varphi = 0$ y si $r = 1 \Leftrightarrow r_\varphi = 1$. Concluye 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 r_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 y cualquiera que se elija, no derivará en conclusiones sustancialmente diferentes.

   La expresión de este coeficiente adaptada a la Tabla 3.1 de contingencia que se ha venido manejando en la presente monografía es:

$$r_\varphi = \frac{P(a) - P(a+b)P(a+c)}{\sqrt{P(a+b)P(a+c)P(c+d)P(b+d)}} \tag{5.1}$$

   donde P es la probabilidad de ocurrencia y se sustituye por la estimación muestral de esta probabilidad.

Aplicándolo al ejemplo 1 (Tabla 4.1) queda:

$$r_\varphi = \frac{\frac{1562}{2081} - \frac{1604}{2081} \times \frac{1945}{2081}}{\sqrt{\frac{1604}{2081}\frac{1945}{2081}\frac{477}{2081}\frac{136}{2081}}} \approx 0.3 \tag{5.2}$$

En este caso el coeficiente r_φ calcula una asociación menor en comparación con r.

Cabe mencionar que no siempre sucede que el coeficiente r_φ sea menor que r, Ekström (2011) [5] muestra que depende de los valores de las frecuencias observadas y en ocasiones sucede al revés.

2. **Coeficiente V de Cramér**: medida de asociación entre dos variables nominales (no necesariamente dicotómicas). Chen y Popovich (2002) [2] explican 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:

$$V = \sqrt{\frac{\chi^2}{N \times (min(r,c) - 1)}} = \sqrt{\frac{\chi^2}{N}} \tag{5.3}$$

Donde χ^2 es la estadística de independencia de Pearson sin corrección de Yates (2.4), 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 4.1) queda:

$$V = \sqrt{\frac{1751.76}{2081}} \approx 0.3 \tag{5.4}$$

De igual forma se obtiene una estimación menor que r.

Capítulo 6

Conclusiones

Para conocer la asociación entre dos variables dicotómicas una opción es el coeficiente de correlación tetracórico que propuso Pearson (1900) [9]. Éste puede ser estimado a través de la ecuación (3.15) donde valores cercanos a 1 o -1 indican una asociación fuerte, y valores cercanos a 0 indican no asociación o independencia entre las variables. También se presentó la expresión para determinar el error probable (3.22) que puede incorporarse en el reporte del análisis estadístico que se esté realizando.

En la sección 4 se propone: 1) primero, realizar la prueba χ^2 de Pearson para conocer si las variables son estadísticamente independientes (teniendo presente que esta prueba siempre rechaza H_0 si N es grande y, por tal motivo, se sugiere no tomarlo como único estadístico de referencia), 2) calcular el coeficiente de correlación tetracórico, para lo cual se presentaron las alternativas de la función `tetrachoric()` de la librería `psych` de R, o la función `rhot` presentada en la sección 3.6, 3) calcular los errores probables (si es que se cumple la condición explicada en la sección 3.4), y 4) concluir con una interpretación a los resultados teniendo cuidado de no hacer inferencias subjetivas ni ambiciosas, recordando también lo discutido en el subtema 2.3.

Se concluye presentando otros dos coeficientes distintos, el coeficiente de correlación r_φ y el coeficiente V de Crámer, que también sirven para conocer la asociación entre variables dicotómicas, con las diferencias de que el coeficiente de correlación r_φ parte del supuesto de que las variables tienen una distribución conjunta discreta, y no una distribución conjunta normal como lo supone el tetracórico; y que el coeficiente V de Crámer no es un coeficiente de correlación sino de asociación, pues toma valores en el intervalo $(0, 1)$. Sin embargo, se comentó que cualquiera de los tres es bueno y el investigador/analista puede ocupar libremente el que sea, posiblemente una ventaja del tetracórico es que puede calcularse también su intervalo probable.

Bibliografía

[1] J. N. Castellan. «On the estimation of the tetrachoric correlation coefficient». En: *Psychometrika* 31 (1966), págs. 67-73. DOI: 10.1007/BF02289458.

[2] P. Y. Chen y P. M. Popovich. «Correlation: Parametric and Nonparametric Measures». En: *Sage Publications, Inc* 139.1 (2002). URL: https://rufiismada.files.wordpress.com/2012/02/correlation__parametric_and_nonparametric_measures__quantitative_applications_in_the_social_sciences_.pdf.

[3] F. Cortés. «Observación, causalidad y explicación causal». En: *Perfiles Latinoamericanos* 26 (2018). URL: http://www.scielo.org.mx/pdf/perlat/v26n52/0188-7653-perlat-26-52-00014.pdf.

[4] J. Ekström. «Contributions to the Theory of Measures of Association for Ordinal Variables». En: *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Science* 50 (2009). URL: https://uu.diva-portal.org/smash/get/diva2:210896/FULLTEXT01.pdf.

[5] J. Ekström. «The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate». En: *UCLA: Department of Statistics* (2011). URL: https://escholarship.org/uc/item/7qp4604r#main.

[6] E. F. El-Hashash. «Methods for Determining the Tetrachoric Correlation Coefficient for Binary Variables». En: *Asian Journal of Probability and Statistics* (2018), págs. 1-12. URL: https://www.journalajpas.com/index.php/AJPAS/article/view/28782.

[7] D. B. Kirk. «On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient». En: *Psychometrika* 38 (1973), págs. 259-268. URL: https://link.springer.com/article/10.1007/BF02291118?error=cookies_not_supported&code=27202d8d-2cde-40cd-a143-66db4089d9b6.

[8] C. W. Odell. «The interpretation of the probable error and the coefficient of correlation». En: *University of Illinois at Urbana-Champaign* 32 (1926). URL: https://www.ideals.illinois.edu/bitstream/handle/2142/28259/interpretationof32odel.pdf?sequence=1&isAllowed=y.

[9] K. Pearson. «I. Mathematical contributions to the theory of evolution. —VII. On the correlation of characters not quantitatively measurable». En: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 195 (1900), págs. 1-47. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.1900.0022.

[10] K. Pearson. «On the Probable Error of a Coefficient of Correlation as found from a Fourfold Table». En: *Biometrika* 9 (1913), págs. 22-27. URL: `https://academic.oup.com/biomet/article-abstract/9/1-2/22/325854`.