|  |  |  |  |
| --- | --- | --- | --- |
| UNAM - Centro de Ciencias de la Atmósfera | **UNIVERSIDAD NACIONAL AUTÓNOMA**  **DE MÉXICO** | | |
|  | **INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS (IIMAS)** | | |
|  |  | | |
| **“MONOGRAFÍA SOBRE EL COEFICIENTE DE CORRELACIÓN TETRACÓRICO”** | | |
|  | **TESINA** |  |
|  | QUE PARA OBTENER EL TÍTULO DE: |  |
|  | **ESPECIALISTA EN ESTADÍSTICA APLICADA** |  |
|  | PRESENTA: |  |
|  | **BRENDA CORINA CEREZO SILVA** |  |
| IIMAS - UNAM - YouTube |  | **DIRECTORA DE TESINA:**  **M. EN C. LETICIA EUGENIA GRACIA MEDRANO VALDELAMAR**  **2021** |  |

Resumen

Pequeño comentario ☺

Reconocimientos

Índice

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.

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

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

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 el hecho de que la información se muestre en una tabla de contingencia de , 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]](#footnote-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 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 tiene la forma:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | |  |
|  |  |  |  | **Total:** |
|  |  | a | b | a+b |
|  | c | d | c+d |
| Total: | | a+c | b+d | N |

Tabla 2. Tabla de contintencia de .

Medida de asociación

Prueba de independencia 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 ) sigue la siguiente superficie de frecuencia:

(ec. 1)

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

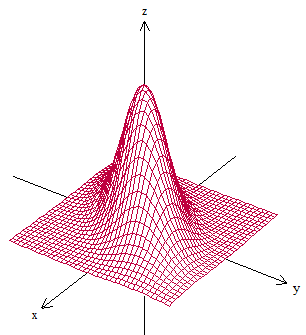
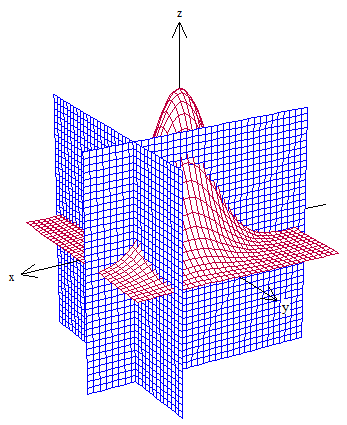
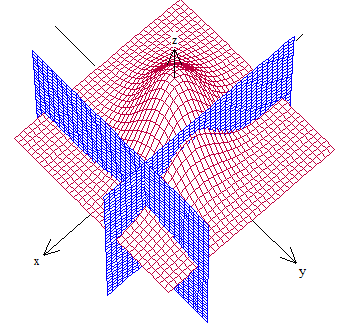


Figura 1. Superficie de frecuencia descrita por la ec. 1

Pearson propone intersectar a la campana con dos planos, uno paralelo a y el otro a , evidentemente perpendiculares entre sí, de tal forma que la campana quede cortada en cuatro secciones donde el volumen de cada una representa las frecuencias y observadas en la tabla de contingencia (ver Figura 2). Dicho de otro modo, la función (ec. 1) gráficamente representa una campana de Gauss de volumen , obsérvese que , por lo que está por encima del plano , digamos el “piso”, ahora bien, si este “piso” tiene un punto de corte 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 tendríamos la siguiente gráfica:

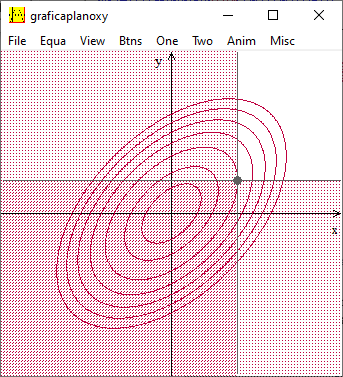


Figura 3. Plano cortado por las rectas y .

Donde, .

La idea para conocer el coeficiente de correlación es: si se conocen las frecuencias observadas, es decir, y entonces se puede conocer el punto que “dicotomizó” a las variables y, en consecuencia, se puede conocer en ec. 1.

**Cálculo de**

El análisis comienza en cómo encontrar el punto , claramente:

(ec. 2)

Si y .

Y también se tiene que[[2]](#footnote-2):

(ec. 3)

Del mismo modo:

(ec. 4)

(ec. 5)

(ec. 6)

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

Y, por lo tanto,

Entonces,

(ec. 7)

Donde es la función de distribución . Y del mismo modo

(ec. 8)

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

**Error probable del coeficiente de correlación**

Descripción de la metodología

Comparación/relación con otras técnicas similares

Conclusión

Apéndice

Referencias

1. Ilustración VI de Pearson(1900) [↑](#footnote-ref-1)
2. En el apéndice se muestra el paso a paso de (ec. 3) y de forma análoga se calculan de (ec. 4) a la (ec. 6). [↑](#footnote-ref-2)