XS-2110 Métodos Estadísticos Prof. Gilbert Brenes

Medidas de Asociación.

En estadística, frecuentemente queremos estudiar si dos variables están relacionadas entre sí, aún cuando no podamos determinar si una es la causa de la otra.

Corrientemente se utiliza el término correlación para referirse a la relación estadística entre dos variables. Sin embargo, entre estadísticos, la palabra correlación se tiene que restringir a ciertas medidas que “miden” la relación entre variables.

Un estadístico debería utilizar la palabra más general “asociación” para referirse a la relación entre variables.

Para ligar mejor el tema de medidas de asociación con el tema de bondad de ajuste, empezaremos con las medidas de asociación para variables cualitativas, y después seguiremos con las medidas de asociación para variables continuas.

**Medidas de Asociación entre variables categóricas y pruebas de Independencia en Tablas de Contingencia o Tablas Cruzadas.**

Supóngase que se desea analizar la asociación entre dos variables nominales. Las distribución conjunta de las dos variables nominales se puede expresar con una tabla de contingencia:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Variable X |  |  |
| Variable Y | A | B | C | Total |
| G |  |  |  |  |
| H |  |  |  |  |
| Total |  |  |  |  |

Supóngase además que si esta tabla fuera la de una población de unidades estadísticas, y tenemos la distribución de probabilidades marginales de cada variable, la tabla la podríamos expresar así, sólo con las marginales:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Variable X |  |  |
| Variable Y | A | B | C | Total |
| G |  |  |  | PG |
| H |  |  |  | PH |
| Total | PA | PB | PC | 1.00 |

Si suponemos independencia entre las variables X y Y (algo equivalente a no asociación), cómo se calcularían las probabilidades conjuntas, o sea, las probabilidades a lo interno de la tabla?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Variable X |  |  |
| Variable Y | A | B | C | Total |
| G | PA\*PG | PB\*PG | PC\*PG | PG |
| H | PA\*PH | PB\*PH | PC\*PH | PH |
| Total | PA | PB | PC | 1.00 |

Si uno tiene una tabla de contingencia, con una distribución cruzada, uno puede plantear una prueba de hipótesis de bondad de ajuste para determinar si los datos en la tabla cruzada “tienen un buen ajuste” a una distribución conjunta de dos variables independientes.

A esta prueba se le llama prueba de independencia, porque asume que las dos variables nominales que conforman la distribución son independientes entre sí.

Si uno define que una de las variables (X) es una variable independiente, y la variable Y es una variable dependiente, uno puede también llamar a esta prueba como prueba de homogeneidad (de proporciones).

Nótese que si mantenemos la idea de una prueba de bondad de ajuste, se pueden usar tanto X2 como G2. Ambos estadísticos tienen una distribución χ2, con n-# parámetros-1 grados de libertad. ¿Cuántos parámetros hay en una prueba de independencia?

La distribución marginal de X está compuesta por las probabilidades PA, PB, y PC. Dado que la suma de estas probabilidades es 1, hay 2 probabilidades no redundantes de las otras, mientras que la tercera es una combinación lineal.

Igual, la distribución marginal de Y está compuesta por las probabilidades PG y PH. O sea, como PG+PH=1, hay sólo una probabilidad no redundante, mientras que la otra es una combinación lineal de la otra.

Si consideramos a las probabilidades no redundantes como parámetros, tendríamos 3 parámetros (dos probabilidades no redundantes de la variable X y una probabilidad no redundante de la variable Y). Así, que los grados de libertad de la tabla serían:

n-(# categorías de X –1)-(# categorías de Y – 1)-1= (# filas-1)\*(# columnas-1)

*Ejemplo.*

Se seleccionó una muestra de partos ocurridos en la maternidad Carit, y se desea estudiar la asociación en Costa Rica entre la condición de fumado de las madres y el bajo peso al nacer:

Cuadro 1. Muestra de partos en la maternidad Carit, por categorías de peso al nacer, según condición de fumadora de la madre.

|  |  |  |  |
| --- | --- | --- | --- |
| Peso al nacer | Madre fumadora | Madre no fumadora | Total |
| Bajo(menos de 2.5 kg)  Normal (de 2.5 kg a menos de 3.0 kg)  Alto (más de 3 kg) | 36 | 23.07  48 | 51.57  11 | 20.36 | 15 | 27.93  66 | 62.43  34 | 24.64 | 51  114  45 |
| Total | 95 | 115 | 210 |

¿Son independientes estas dos variables?



ó

Igual, tenemos las mismas restricciones que aquellas para cualquier prueba de bondad de ajuste:

1. El 75% de las Fi deberían ser mayores a 5
2. Para usar G2 se debería de tener una muestra razonablemente grande (digamos, más que 100), más grande que la que se puede usar para X2 (digamos que 40 ó más).

H0: Las variables bajo peso al nacer y condición de fumado de la madre son independientes entre sí.

H1: Las variables no son independientes entre sí.

Si usamos un α=0.05,

χ21-0.05, (filas-1)\*(col-1)g.l.= χ21-0.05, (3-1)\*(2-1)g.l.= χ21-0.05, 2g.l.=5.995

Realice el ejemplo, pero uniendo las categorías normal y alto peso.

Cuadro 1. Muestra de partos en la maternidad Carit, por categorías de peso al nacer, según condición de fumadora de la madre.

|  |  |  |  |
| --- | --- | --- | --- |
| Bajo peso al nacer (menos de 2.5 kg) | Madre fumadora | Madre no fumadora | Total |
| Sí  No | 36  59 | 15  100 | 51  159 |
| Total | 95 | 115 | 210 |

Las hipótesis nula y alternativa son las mismas.

Con un α=0.05, el estadístico de prueba χ20.05, (2-1)\*(2-1)g.l.= χ20.05, 1g.l.=3.8415

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Coeficiente de contingencia**

Es uno coeficiente de asociación que se puede utilizar en variables de cualquier tipo de escala. Si son variables cuantitativas, se recomienda agrupar los valores.

La fórmula es:  ,

donde χ2 es el estadístico de prueba en la prueba de independencia y n es el tamaño de muestra.

Siempre es mayor a 0 y menor a 1. Cuanto más cercano a 0, más débil es la asociación. Cuanto más cercano a 1, mayor es la asociación. El único inconveniente, es que el C nunca llega a ser uno, por lo que es difícil interpretar cuán fuerte es la asociación.

**V de Cramer:**

Es un coeficiente de asociación similar a C, pero el denominador es generalmente más grande. Su fórmula es:



donde c es el menor entre el total de columnas y el total de filas.

V comparte las mismas ventajas y limitaciones que C. Puede ser usado con cualquier tipo de variables, pero su valor máximo nunca llega a ser 1.

**Medidas de asociación epidemiológica: Odds ratio, razón de odds ó razón de ventajas OR**:

Este es uno de los estadísticos más utilizados para medir asociación en el campo de la epidemiología. También se utiliza en tablas 2X2, aunque si hay tiempo, veremos que con paquetes estadísticos, podemos calcular OR para una variable continua y otra nominal.

Si el OR es mayor a uno, significa que la categoría en el numerador es más probable que la categoría en el denominador. Si el OR es menor a uno, significa que la categoría en el numerador es menos probable que la categoría en el denominador, y si es cercana a uno, significa que no hay asociación entre las variables.

Su fórmula es:



Si se calcula el logaritmo natural del OR, entonces la no asociación se alcanza con valores iguales a 0.

Este coeficiente es muy utilizado en epidemiología, sobre todo en estudios de caso control, porque su valor no se ve afectado por el diseño de investigación que se utilice.

**Riesgo relativo, o RR.**

El riesgo relativo también es muy utilizado en epidemiología, y se interpreta como cuantas veces es una categoría tan probable como la otra. Su fórmula es:

Se puede demostrar fácilmente que cuando P(Y=1|X=2) y P(Y=1|X=1) tienden a cero (o sea, son muy pequeños), OR→RR

**Prueba exacta de Fisher-Irwin:**

Dentro de las pruebas exactas, la más conocida es la prueba exacta de Fisher. Se le llama prueba exacta porque se determinan todas las posibles tablas que se puedan generar a partir de las marginales, y se calculan las probabilidades exactas de ver dichas tablas.

Es una prueba exacta basada la distribución hipergeométrica. En esta prueba, se requiere el criterio del investigador para determinar los casos extremos.

Las hipótesis nula y alternativa de la prueba de Fisher-Irwin se pueden plantear de la siguiente forma:

H0: La variable A no está asociada a la variable B

H1: La variable A está asociada a la variable B.

Podemos también plantear las hipótesis nula y alternativa usando OR:

H0: ϕ=1

H1: ϕ<1, ó ϕ>1 (prueba de una cola)

H1: ϕ≠1, (prueba de dos colas)

donde ϕ es el OR poblacional.

Se tienen que definir categorías extremas, y se calcula la suma de las probabilidades pk de observar esas categorías extremas, donde:



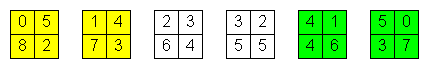
Algunos autores consideran que la prueba exacta de Fisher-Irwin es demasiado conservadora.

La prueba exacta de Fisher calcula la probabilidad de obtener exactamente las frecuencias observadas o “más extremas”, condicional a que las frecuencias marginales observadas se mantienen fijas. Por “más extremas”, nos referimos a configuraciones con probabilidades de ocurrencia menores en una dirección (prueba de una cola) o en ambas direcciones (prueba de dos colas).

Supongamos que tenemos la siguiente tabla 2x2:



Todas las configuraciones posibles son:



Nótese que las marginales son iguales:

Las probabilidades correspondientes son:

fisher3

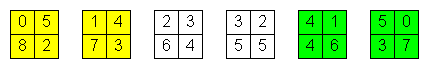
Las tablas en amarillo corresponden a configuraciones más extremas en la misma dirección. Las configuraciones más extremas en la misma dirección se calculan encontrando la frecuencia menor, y restándole uno progresivamente a esa frecuencia menor, reconfigurando la tabla pero manteniendo las mismas marginales.

Por ejemplo, en la tabla, la frecuencia menor es O11=2. Se le resta 1, por lo que en la siguiente tabla O11=1, y se recalculan las demás celdas (O12, O21, O22).

Las tablas en verde son las tablas “más extremas” en la dirección opuesta.

Otra forma es calcular la siguiente fórmula. Este valor está centrado en 1, y si la tabla base tiene este coeficiente mayor a 1, entonces las tablas en la misma dirección tienen también el coeficiente mayor a 1, mientras las tablas en la dirección opuesta tienen este coeficiente menor a 1.

El coeficiente se calcula: 

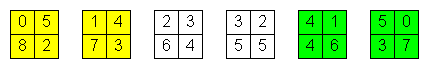


Los valores del coeficiente serían:

0 0.107 0.44 1.5 6 Indefinido

El coeficiente de la tabla base es 0.44, por lo que las tablas en la misma dirección son las que tienen este coeficiente menor a 1, mientras que las tablas en la dirección opuesta tienen valores mayores a 1.

Nuevamente, las configuraciones posibles son:



Las probabilidades correspondientes son:

fisher3

Entonces, las probabilidades de una cola (p-values) serían:

.326 + .093 + .007 = .426

y las probabilidades de dos colas serían:

.326 + .093 + .007 + .163 + .019 = .608

Nota[[1]](#footnote-1)

Ejercicio:

Con base en la tabla de bajo peso al nacer, escriba todas las tablas posibles, y realice la prueba de Fisher-Irwin, con una significancia del 0.05.

T1.=6 T2.=16 T.1=10 T.2=12



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 6 | 0 |  | 5 | 1 |  | 4 | 2 |  | 3 | 3 |  | 2 | 4 |  | 1 | 5 |  | 0 | 6 |
| 4 | 12 |  | 5 | 11 |  | 6 | 10 |  | 7 | 9 |  | 8 | 8 |  | 9 | 7 |  | 10 | 6 |
| p= |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.00281 | |  | 0.04052 | |  | 0.18576 | |  | 0.35383 | |  | 0.29854 | |  | 0.10615 | |  | 0.01238 | |
|  | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |
| OR= | |  |  | |  |  | |  |  | |  |  | |  |  | |  |  | |
| Indef | |  | 11.000 | |  | 3.333 | |  | 1.286 | |  | 0.500 | |  | 0.156 | |  | 0.000 | |

**Coeficiente de correlación lineal de Pearson, ó Coeficiente producto-momento de Pearson:**

r: en la muestra

ρ: en la población

Desde el punto de vista teórico, el coeficiente de correlación lineal es:



La fórmula de la estimación muestral de ρ es:



El coeficiente de correlación lineal de Pearson varía entre -1 y 1. Si:

|  |  |
| --- | --- |
| r negativo | Correlación negativa: Conforme aumenta x, disminuye y |
|  |  |
| r positivo | Correlación positiva: Conforme aumenta x, aumenta y |
|  |  |
| r=0 | No hay correlación |

* Si r es cercano a -1, la correlación es fuerte y negativa
* Si r es cercano a 1, la correlación es fuerte y positiva
* Si r es cercano a 0, la correlación es débil o no existe.

Cuál es el tamaño apropiado para determinar si una correlación es fuerte o no? Depende del área del conocimiento en el que trabajen.

En Ingeniería y Física, ρ=0.8 ya son débiles.

En Ciencias de la Salud y en biología, ρ=0.5 ya son débiles.

En Economía (series de tiempo), ρ=0.7 ya son débiles.

En otras ciencias sociales (sociología, ciencias políticas), ρ=0.4 pueden ser consideradas fuertes.

Recuérdese que r ó ρ mide asociación lineal, por lo que dos variables pueden estar asociadas, pero si la forma de su asociación es curvilínea, r puede ser cercano a 0.



r=0.91 r=0.46

Gráficamente, la correlación se puede representar con una elipse. Cuanto más angosta sea la elipse que envuelve la nube de puntos, mayor la correación.



Si r es un estimador de ρ, se puede realizar una prueba de hipótesis para determinar si, dado un valor de r en una prueba, hay suficiente evidencia estadística como para rechazar que ρ=0.

H0: ρ=0

H1: ρ≠0

El estadístico de prueba es:



Bajo la hipótesis nula, el estadístico tc tiene una distribución t-Student con n-2 g.l.

Si la hipótesis nula plantea otro valor para ρ:

H0: ρ=0.6

H1: ρ≠0.6

entonces el estadístico de prueba sería:



El estadístico z tiene una distribución normal estándar, aunque requiere que las muestras sean grandes, para poder utilizar la z.

Para cualquiera de las dos pruebas de hipótesis, se necesita suponer que las dos variables tienen conjuntamente una distribución normal bivariada.

**Coeficiente de correlación de rangos de Spearman.**

También se le denomina r ó ρ, pero para diferenciarla del coeficiente producto momento, se le usa un subíndice. O sea, rs ó ρs.

Se usa cuando:

1. la relación es un poco curvilínea (una asociación que se puede describir con una curva exponencial)
2. Se viola el supuesto de normalidad bivariada, pero aún así se quiere realizar la prueba de hipótesis
3. al menos una variable es de escala ordinal y la otra puede ser ordinal ó cuantitativa.
4. Si se tienen valores extremos

Se le denomina coeficiente de rangos, porque hay que ordenar los datos, y en lugar de usar los valores observados, se utiliza el rango de cada valor.

Las fórmulas son las siguientes:

, si menos del 25% de los datos son empates

, si más del 25% de los datos son empates

donde:

n = Tamaño de muestra

di = rango de xi –rango de yi

Tx ó Ty, son 

Con o sin empates, el coeficiente de correlación de Spearman converge al coeficiente de correlación de Pearson, aplicado a los rangos, cuando el tamaño de muestra es razonablemente grande, por lo que se puede aplicar la misma fórmula.

Para probar la hipótesis nula de que ρs=0, si n<30, se usa la tabla G, y si n≥30, se usa:

, y se contrasta con una t con n-2 g.l.

**Coeficiente tao (τ) de Kendall.**

Igual que con el coeficiente de rangos de Spearman, el tao de Kendall se usa cuando al menos una de las variables es ordinal. Existen varias taos: τa, τb, y τc, dependiendo de las correcciones incluidas en la fórmula. Así la primera fórmula que se les suministra es τa porque no corrige por empates, mientras que la segunda τc, sí corrige por empates. Tao-b se usa generalmente para tablas 2X2, lo cual no lo vamos a hacer en este curso.

Como dice en el libro de Gutiérrez (p.72), el procedimiento es el siguiente:

* se calculan los rangos de las variables X y Y
* se ordenan los valores según la variable X
* se observan los valores de los rangos de la variable Y.
* Para Y(i), se cuentan cuantos Y(j>i) son mayores que Y(i) y se le restan cuantos Y(j>i) son menores que Y(i).
* Se suman después estas diferencias, y esto es S, en la fórmula:





donde Tx y Ty son los mismos definidos anteriormente para el coeficiente de Spearman.

Para probar la hipótesis nula de que τ=0, si n≤10, usar la tabla H, y si no, con muestras grandes, se puede aproximar el estadístico de prueba tiene una distribución normal estándar:



Tanto el rs como el τ varían de -1 a 1, y se interpretan similar a un coeficiente de correlación lineal de Pearson

**Razón de correlación, o eta (η)**

Se utiliza para medir la asociación entre una variable nominal, y una variable cuantitativa (intervalo o de razón). η varía entre 0 y 1, donde 0 indica que no hay asociación, y 1 indica que hay asociación.

La fórmula es equivalente a la raíz cuadrada de la razón de la Suma de Cuadrados Entre Grupos y la Suma de Cuadrados Total, en un Análisis de Variancia de una vía.

La fórmula sería:



donde:

ni: # observaciones en el grupo i

: promedio de observaciones en el grupo i

: promedio de todas las observaciones

k: # de grupos

n: Total de muestra

Si los individuos dentro de cada grupo fueron asignados al azar, el siguiente estadístico de prueba tiene una distribución F con k-1 y n-k g.l.

Si H0:η=0, se puede utilizar el siguiente estadístico de prueba para la prueba de hipótesis.



1. Tomado de: http://www.people.ku.edu/~preacher/fisher/fisher.htm [↑](#footnote-ref-1)