Doble Grado en
Matemáticas y Estadística

TRABAJO FIN DE GRADO

Medidas de Asociación para Variables Nominales y Ordinales

Alumna: Gloria Vizcaíno Castaño
Tutor: Juan Manuel Muñoz Pichardo

Sevilla, Junio de 2022

Índice general

Resumen . . . III
Abstract . . . IV
Índice de Figuras . . . V
Índice de Tablas . . . VI

1. Introducción **1**
1.1. Definiciones . . . 1
1.1.1. Definición de Medición . . . 2
1.1.2. Definición de Asociación . . . 2
1.2. Dimensiones de la asociación . . . 2
1.2.1. Nivel de Medición . . . 3
1.2.2. Simetría y Asimetría . . . 3
1.2.3. Asociación unidireccional y bidireccional . . . 3
1.2.4. Clasificación cruzada . . . 3
1.2.5. Correlación, asociación y concordancia . . . 3
1.3. Criterios para las medidas de asociación . . . 4
1.4. Grado de Asociación . . . 5

2. Medidas para Variables Nominales **6**
2.1. Valores de probabilidad hipergeométrica . . . 6
2.2. Medidas λ_a y λ_b de Goodman y Kruskal . . . 9
2.3. Medidas t_a y t_b de Goodman y Kruskal . . . 11
2.4. Una prueba asimétrica de homogeneidad . . . 13
2.5. Medidas de concordancia . . . 16
2.5.1. Medida de concordancia de Robinson . . . 17
2.5.2. Medida de concordancia π de Scott . . . 19
2.5.3. Medida de concordancia κ de Cohen . . . 19
2.5.4. Aplicación con Varios Jueces . . . 21
2.6. Test Q de McNemar para el cambio . . . 23
2.7. Test Q de Cochran para el cambio . . . 24
2.8. Una medida sobre el Tamaño del Efecto para el test Q de Cochran . . . 25
2.8.1. Una medida del Tamaño del Efecto corregida por el azar . . . 27
2.8.2. Ventajas de la medida $\mathfrak{R}$ del tamaño del efecto . . . 28
2.9. Medida de asociación d_N^c de Leik y Gove . . . 28
2.9.1. Tabla de Contingencia Observada . . . 30
2.9.2. Tabla de Contingencia Esperada . . . 30
2.9.3. Tabla de Contingencia Maximal . . . 31
2.9.4. Cálculo de la d_N^c de Leik y Gove . . . 32

2.9.5. Un test de permutación para d_N^c . . . 33
2.10. Un problema de ocupación de la matriz . . . 33
2.11. Test exacto de Fisher . . . 35
2.11.1. Análisis exacto de Fisher con una tabla 2×2 . . . 35
2.11.2. Análisis exacto de Fisher con otras tablas de contigencia . . . 37
2.11.3. Análisis de tablas 2×2×2 . . . 38

3. Medidas para Variables Ordinales (Parte 1) 41
3.1. Principales medidas de Asociación Ordinal por pares . . . 41
3.1.1. Medida τ_a de Kendall de Asociación Ordinal . . . 43
3.1.2. Medida τ_b de Kendall de Asociación Ordinal . . . 43
3.1.3. Medida τ_c de Stuart de Asiciación Ordinal . . . 44
3.1.4. Medida γ de Goodman y Kruskal . . . 45
3.1.5. Medidas d_{yx} y d_{xy} de Somers . . . 45
3.2. Métodos estadísticos de permutación . . . 46
3.3. Distribuciones de las Frecuencias Marginales . . . 47
3.4. Otras medidas de Asociación Ordinal por pares . . . 48
3.4.1. Medidas $d_{y.x}$ y $d_{x.y}$ de Kim . . . 48
3.4.2. Medida e de Asociación Ordinal de Wilson . . . 48
3.4.3. Medida S de Whitfield para una Variable Binaria y una Variable Ordinal . . . 49
3.4.4. Coeficiente de correlación rango-biserial del Cureton . . . 50
3.4.4.1. Relaciones entre las diferentes medidas . . . 51

4. Medidas para Variables Ordinales (Parte 2) 53
4.1. Coeficiente de correlación de rango de Spearman . . . 53
4.2. Medida de concordancia de la regla del pie de Spearman . . . 55
4.2.1. Probabilidad de la regla del pie de Spearman . . . 57
4.2.2. Rangos Múltiples . . . 57
4.3. Coeficiente de Concordancia . . . 58
4.3.1. Procedimientos relacionados . . . 59
4.4. Medida de acuerdo u de Kendall . . . 60
4.5. Medida Kappa de Cohen . . . 60
4.5.1. Comparación de la ponderación lineal y cuadrática . . . 62
4.5.2. Kappa ponderado con múltiples jueces . . . 63
4.6. Análisis Ridit . . . 65
4.6.1. Cálculo . . . 66
4.6.1.1. Procedimientos de permutación exacta . . . 66
4.6.1.2. Procedimientos de permutación de remuestreo . . . 66

A. Apéndice: Título del Apéndice 68
A.1. Primera sección . . . 68

B. Apéndice: Título del Apéndice 69
B.1. Primera sección . . . 69

Bibliografía 74

Resumen

La evaluación de la relación y la intensidad de la misma entre dos o más variables aleatorias es un objetivo presente en cualquier estudio estadístico. Para ello se han propuesto una amplia gama de medidas estadísticas adaptadas a la naturaleza de las variables analizadas.

El objetivo de este trabajo se centra en recopilar las medidas de asociación entre variables nominales, incluyendo la definición teórica e inferencia estadística sobre las mismas.

Asimismo, el trabajo incluye su implementación en R y/o el uso de librerías de R, con una ilustración sobre datos reales. Además, con objeto de ilustrar la aplicabilidad de las mismas, incluiye referencias sobre trabajos científicos recientes en los que se han utilizado estas medidas.

Abstract

The evaluation of the relationship and its intensity between two or more random variables is an objective present in any statistical study. For this purpose, a wide range of statistical measures adapted to the nature of the variables analyzed have been proposed.

The aim of this project is to compile measures of association between nominal variables, including their theoretical definition and statistical inference.

Also, the report includes their implementation in R and/or the use of R libraries, with an illustration on real data. In addition, in order to illustrate their applicability, it includes references to recent scientific studies in which these measures have been used.

Índice de figuras

2.1. Representación gráfica de una tabla de contingencia 2x2x2 38

Índice de tablas

2.1. Notación tablas de contingencia 2x2 7
2.2. Notación tablas de contingencia 4x3 7
2.3. Notación para la clasificación cruzada de dos variables categóricas, A y B, con c y r categorías respectivamente 9
2.4. Notación para la clasificación cruzada de dos variables categóricas, A y B, con g y r categorías respectivamente 14
2.5. Clasificación cruzada cxc con proporciones en celdas 20
2.6. Notación de una tabla de clasificación cruzada 2x2 para el test Q de McNemar 23
2.7. Notación de una tabla de contingencia 2x2 35

3.1. Notación para la clasificación cruzada de dos variables categóricas, X e Y. 42
3.2. Ejemplo para la medida S de Whitfield 49

4.1. Notación para una tabla de validación cruzada de N objetos por 2 jueces en c categorías disjuntas y ordenadas 61

Capítulo 1

Introducción

Aunque existen una gran cantidad de métodos para medir la magnitud de la asociación entre dos variables, hay grandes dificultades para interpretar y comparar las distintas medidas, ya que a menudo difieren en su estructura, lógica e interpretación.

Así pues, las distintas medidas de asociación desarrolladas a lo largo de los años constituyen una mezcla de enfoques lógicos, estructurales y de interpretación.

Es conveniente clasificar las distintas medidas de asociación por el nivel de medición para el que fueron diseñadas originalmente y para el que son más apropiadas, reconociendo que algunas medidas pueden ser adecuadas para más de un nivel de medición, especialmente las numerosas medidas originalmente diseñadas para el análisis de tablas de contingencia 2×2, en las que el nivel de medición es a veces irrelevante.

Además de la consideración de la estructura, la lógica y la interpretación, un inconveniente importante de las medidas de asociación es la determinación del p-valor de la medida obtenida bajo la hipótesis nula.

Existen dos enfoques principales para determinar los p-valores de las medidas de asociación: el modelo poblacional de Neyman-Pearson y el modelo de permutación de Fisher-Pitman.

El modelo poblacional está plagado de suposiciones que rara vez se cumplen en la práctica y que, algunas veces, son inapropiadas; por ejemplo, independencia, normalidad, homogeneidad de la varianza...

De aquí en adelante se usará casi exclusivamente el modelo de permutación ya que está libre de cualquier suposición de distribución, no requiere un muestreo aleatorio, es completamente dependiente de los datos, proporciona valores de probabilidad exactos y es ideal para el análisis de muestras pequeñas.

Por tanto, en este Trabajo Fin de Grado, se usará el enfoque de permutación para la medición de la asociación estadística, definida ampliamente para incluir medidas de correlación, asociación y concordancia.

1.1. Definiciones

Dado que el título de este Trabajo de Fin de Grado es “Medidas de Asociación para Variables Nominales”, es conveniente, en primer lugar, definir “Medición” y “Asociación”.

1.1.1. Definición de Medición

Según Cowles, la medición es la mejor manera de describir con precisión los acontecimientos y las relaciones entre ellos [Cowles, 2001]. La medición ha sido una característica fundamental de la civilización humana desde sus inicios. Así, la medición es la aplicación de las matemáticas a los acontecimientos, el uso de números para designar objetos y acontecimientos, y sus relaciones.

Más formalmente, la medición es el proceso de la asignación a los fenómenos empíricos de un sistema numérico.

Se pueden distinguir cuatro niveles o escalas de medición: nominal, ordinal, de intervalo y de razón:

- El nivel **nominal** de medición no mide cantidades, simplemente clasifica los acontecimientos en una serie de categorías no ordenadas y se agrupan los acontecimientos que tienen características comunes. Ejemplos de clasificaciones nominales son el género, el tipo de sangre o el estado civil.
- La esencia del nivel **ordinal** de medición es que emplea las características de "mayor que" ($>$) o "menor que" ($<$). Las relaciones ($>$) y ($<$) no son reflexivas ni simétricas, pero sí transitivas. Ejemplos de escalas ordinales son el orden de nacimiento, el rango académico o las escalas Likert (Muy de acuerdo, De acuerdo, Neutral, En desacuerdo, Totalmente en desacuerdo).
- Las escalas de nivel de **intervalo** introducen otra dimensión en el proceso de medición y ordenan los eventos en intervalos de igual apariencia. En las escalas de intervalo, no hay un punto cero absoluto: si hay un valor de cero, éste se define arbitrariamente. Las temperaturas medidas en grados Fahrenheit o Centígrados son ejemplos tradicionales de medición de intervalos.
- Las escalas de **razón** son escalas que no sólo incorporan todas las características de una escala de intervalo, sino que tienen puntos cero absolutos, lo que permite la construcción de relaciones significativas. Ejemplos de escalas de intervalo son el tiempo, la edad, la altura o los grados Kelvins (0 Kelvins es el cero absoluto, definido como la ausencia de movimiento molecular).

Desde el punto de vista estadístico, las mediciones a nivel de intervalo y de razón suelen tratarse juntas y, en general, se denominan simplemente mediciones de nivel de intervalo.

1.1.2. Definición de Asociación

Aunque hay muchas formas de definir la asociación, quizás la más sencilla y útil sea: *se dice que dos variables están asociadas cuando la distribución de los valores de una variable difiere para diferentes valores de la otra variable.*

Además, si un cambio en la distribución de los valores de una variable no provoca un cambio en la distribución de los valores de la otra variable, se dice que las variables son independientes.

1.2. Dimensiones de la asociación

Hay que tener en cuenta varias dimensiones a la hora de medir la asociación:

1.2.1. Nivel de Medición

Como hemos visto anteriormente, pueden ser: variables de nivel nominal (categóricas), de nivel ordinal (clasificadas) y de nivel de intervalo. Además, en algunos casos, se consideran mezclas de los tres niveles de medición: variables de nivel nominal y ordinal, de nivel nominal y de intervalo, y de nivel ordinal e intervalo.

1.2.2. Simetría y Asimetría

Una medida de asociación puede ser asimétrica, con variables independientes y dependientes bien definidas, dando lugar a dos índices que miden la fuerza de la asociación dependiendo de la variable que se considere dependiente; o simétrica, dando lugar a un único índice de fuerza de asociación.

1.2.3. Asociación unidireccional y bidireccional

Las medidas de asociación pueden cuantificar la asociación unidireccional entre variables basándose en la medida en que una variable implica a la otra, pero no a la inversa. Por otro lado, la asociación bidireccional o mutua se refiere a la medida en que las dos variables se implican mutuamente. Todas las medidas asimétricas son medidas de asociación unidireccional, y algunas medidas simétricas son medidas de asociación unidireccional.

1.2.4. Clasificación cruzada

Las medidas de asociación se han construido históricamente para datos clasificados en tablas de contingencia de doble entrada o, alternativamente, en simples listas bivariadas de medidas de respuesta. Además, algunas medidas suelen calcularse para ambos casos.

1.2.5. Correlación, asociación y concordancia

Las medidas de **asociación** pueden medir de diversas maneras la correlación, la asociación o la concordancia. Muchos autores han tratado de distinguir entre los conceptos de correlación y asociación. Hay dos ámbitos correspondientes al término "asociación":

- El más general incluye todos los tipos de medidas de asociación entre dos variables en todos los niveles de medición.
- El más restrictivo está reservado a las medidas diseñadas específicamente para medir el grado de relación entre dos variables en los niveles de medición nominal y ordinal.

Así pues, en este trabajo, la asociación se usará de dos maneras. En primer lugar, como un concepto global que incluye medidas de correlación, asociación y concordancia; y en segundo lugar, se utilizará más específicamente como una medida de relación entre dos variables de nivel nominal, dos variables de nivel ordinal o alguna combinación de ambas.

En general, la **correlación** suele referirse a las medidas de covariación derivadas de las ecuaciones de regresión basadas en el método de mínimos cuadrados ordinarios. A menudo, pero no siempre, la correlación simple mide la relación entre dos variables a nivel de intervalo de medida, donde las dos variables se etiquetan normalmente como X e Y. La medida de correlación más usada es el coeficiente de correlación de Pearson al cuadrado.

Las medidas de **concordancia** intentan determinar la identidad de dos variables en cualquier nivel de medición, es decir, $X_i = Y_i$ para todo i. Algunos ejemplos de medidas de concordancia son la medida π de Scott, la medida A de Robinson, la medida de la regla de Spearman y los coeficientes kappa ponderados y no ponderados de Cohen.

La correlación y la concordancia se suelen confurdir, a continuación se muestra un ejemplo para entententender las diferencias:

Supongamos que un investigador desea establecer la relación entre los valores observados y los predichos por la regresión, y e $\hat{y}$, respectivamente. La concordancia implica que la relación funcional entre y e $\hat{y}$ puede describirse mediante una la recta $x = y$. Si por ejemplo obtenemos los pares (1,1), (3,3), (8,8), el coeficiente de correlación de Pearson al cuadrado es $r^2_{y,\hat{y}} = 1$ y el porcentaje de concordancia es del 100 %, es decir, los elementos de los tres pares $(y, \hat{y})$ son iguales. En este contexto, el coeficiente de correlación de Pearson al cuadrado, $r^2_{y,\hat{y}}$, también se ha utilizado como medida de concordancia. Sin embargo, $r^2_{y,\hat{y}} = 1,00$ implica una relación lineal entre y e $\hat{y}$, donde tanto la el corte con el eje de ordenadas como la pendiente son arbitrarias. Así, aunque la concordancia perfecta se describe con un valor de 1,00, también es cierto que $r^2_{y,\hat{y}} = 1,00$ describe una relación lineal que puede o no reflejar una concordancia perfecta, por ejemplo para los valores $(y, \hat{y})$: (2, 4), (4, 5), (6, 6), (8, 7), y (10, 8), el coeficiente de correlación de Pearson es $r^2_{y,\hat{y}} = 1,00$, y el porcentaje de concordancia es del 20 %, es decir, sólo un par valores coinciden [Berry et al., 2010].

((aquí podría meter gráficos para visualizar el ejemplo))

1.3. Criterios para las medidas de asociación

Varios investigadores han escrito sobre criterios importantes para las medidas de asociación, sobre todo Costner [Costner, 1965] y Goodman y Kruskal[Goodman and Kruskal, 1954]. Sin embargo, esta sección se basa principalmente en los criterios que Weiss [Weiss, 1968]consideraba más importantes.

Los criterios importantes para las medidas de asociación incluyen la normalización adecuada, la interpretación, la independencia de las frecuencias marginales y la magnitud (grado o fuerza) de la asociación:

- Normalización: Idealmente, los valores de una medida de asociación deberían cubrir el mismo rango que los valores de probabilidad, es decir, de 0 a 1. Además, la medida de asociación debe ser cero cuando las variables son independientes y uno cuando hay una asociación perfecta. Cuando sea conveniente considerar la asociación inversa, entonces menos uno debe representar la asociación negativa perfecta.
- Interpretación: Una medida de asociación debe tener una interpretación significativa, como la reducción proporcional del error probable, la proporción de la varianza explicada o la proporción por encima de lo que cabría esperar por azar. Muchas medidas de asociación carecen notablemente de este aspecto. De hecho, muchas medidas no permiten ninguna interpretación, excepto que un valor más alto indica más asociación que un valor más bajo, e incluso eso es a menudo cuestionable.
- Independencia de las frecuencias marginales: Idealmente, una medida de asociación no debería cambiar con un aumento (disminución) de los totales de frecuencia de filas o columnas; es decir, la medida de asociación debería ser independiente de los

totales de frecuencia marginal. Algunas medidas de asociación tienen esta propiedad, como las diferencias porcentuales y los odds ratio, pero muchas otras no.

- Grado de asociación: Los valores de una medida de asociación deben aumentar (disminuir) con el aumento (disminución) de los grados de asociación. Así, cuando las frecuencias de las celdas de una tabla de contingencia indican cambios en la asociación, la medida de asociación debería cambiar de forma acorde.

1.4. Grado de Asociación

Las diferentes medidas de asociación evalúan el grado de asociación de diversas maneras. Entre las diversas formas de medir la fuerza de la asociación se encuentran la desviación de la independencia, la magnitud de las diferencias de los subgrupos, las comparaciones por pares, la correspondencia incremental y la concordancia entre variables:

- Desviación de la independencia: Las medidas de asociación que se basan en la desviación de la independencia plantean cómo serían los datos si las dos variables fueran independientes, es decir, que no hubiera asociación, y luego miden el grado en que los datos observados se apartan de la independencia.
- Comparaciones por pares: Algunas medidas de asociación se basan en comparaciones por pares donde las diferencias entre las medidas de respuesta se calculan entre todos los pares de mediciones posibles y se dividen en pares concordantes y discordantes. Un par concordante es aquel en el que la dirección de la diferencia con una variable coincide con la dirección de la diferencia con la segunda variable. Un par discordante es aquel en el que la dirección de la diferencia con una variable no es igual a la dirección de la diferencia con la segunda variable.
- Correspondencia incremental: El grado de asociación se basa en la medida en que un aumento (disminución) incremental en una variable va acompañado de un aumento (disminución) en la otra variable. Este enfoque se denomina convencionalmente "correlación" en lugar de "asociación".
- Concordancia entre variables: El grado de asociación se mide por el grado en que los valores de una variable discrepan de los valores de la otra variable, por encima de lo esperado por el mero azar.

Capítulo 2

Medidas para Variables Nominales

Este capítulo se centra en las medidas de asociación diseñadas para las variables de nivel nominal, pero indagando en los métodos estadísticos de permutación exactos y de Monte Carlo para las medidas de asociación nominal que se basan en criterios distintos del estadístico de prueba chi-cuadrado de Pearson.

En primer lugar, se describen dos medidas asimétricas de asociación de nivel nominal propuestas por Goodman y Kruskal en 1954, λ y t. A continuación, el coeficiente kappa no ponderado de Cohen, κ, que proporciona una introducción a la medición de la concordancia, en contraste con las medidas de asociación. También se incluyen en el capítulo las pruebas Q de McNemar y Cochran, que miden el grado en que las medidas de respuesta cambian con el tiempo, la medida d_N^c de Leik y Gove de asociación nominal, y una solución al problema de ocupación de la matriz propuesta por Mielke y Siddiqui.

La prueba de probabilidad exacta de Fisher es la prueba de permutación ideal para las tablas de contingencia. Mientras que la prueba exacta de Fisher suele limitarse a las tablas de contingencia de 2×2, para las que se concibió originalmente, en este capítulo la prueba exacta de Fisher se extiende a las tablas de contingencia de $2 \times c$, 3×3, $2 \times 2 \times 2$ y otras más grandes.

Algunas medidas diseñadas para variables de nivel ordinal también sirven como medidas de asociación para variables de nivel nominal cuando r (número de filas) $= 2$ y c (número de columnas) $= 2$, es decir, una tabla de contingencia 2×2. Otras medidas se diseñaron originalmente para tablas de contingencia 2×2 con variables de nivel nominal, entre estas medidas de asociación están las diferencias porcentuales, las medidas Q e Y de Yule, los odds ratio y las medidas asimétricas de Somers, d_{yx} y d_{xy}.

2.1. Valores de probabilidad hipergeométrica

Los métodos estadísticos de permutación exacta, especialmente cuando se aplican a las tablas de contingencia, dependen en gran medida de los valores de probabilidad hipergeométricos. En esta sección, se hace una breve introducción a los valores de probabilidad hipergeométricos ilustrando su cálculo e interpretación. Para las tablas de contingencia de 2×2, el cálculo de los valores de probabilidad hipergeométricos es fácil de demostrar. Consideremos la siguiente tabla de contingencia 2×2:

Tabla 2.1: Notación tablas de contingencia 2x2

	A_1	A_2	**Total**
B_1	n_{11}	n_{12}	R_1
B_2	n_{21}	n_{22}	R_2
Total	C_1	C_2	N

donde n_{11}, . . . , n_{22} denotan las cuatro frecuencias absolutas, R_1 y R_2 denotan los totales de las frecuencias marginales de cada fila, C_1 y C_2 denotan los totales de las frecuencias marginales de cada columna y

$$N = \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij}$$

Dado que la tabla de contingencia que aparece en la *Tabla 2.1* es una tabla 2×2 y, en consecuencia tiene sólo un grado de libertad, la probabilidad de cualquier frecuencia de celda constituye la probabilidad de toda la tabla de contingencia. Por lo tanto, el valor de la probabilidad del punto hipergeométrico para la celda que contiene n_{11} viene dada por:

$$p(n_{11}|R_1, C_1, N) = \left(\begin{array}{c} C_1 \\ n_{11} \end{array} \right) \left(\begin{array}{c} C_2 \\ n_{12} \end{array} \right) \left(\begin{array}{c} N \\ R_1 \end{array} \right)^{-1} =$$

$$= \left(\begin{array}{c} R_1 \\ n_{11} \end{array} \right) \left(\begin{array}{c} R_2 \\ n21 \end{array} \right) \left(\begin{array}{c} N \\ C_1 \end{array} \right)^{-1} = \frac{R_1!R_2!C_1!C_2!}{N!n_{11}!n_{12}!n_{21}!n_{22}!}$$

El cálculo de los valores de probabilidad hipergeométricos para las tablas de contingencia $r \times c$ es más complejo que para las tablas de contingencia simples de 2×2. Consideremos la tabla de contingencia 4×3:

Tabla 2.2: Notación tablas de contingencia 4x3

	A_1	A_2	A_3	**Total**
B_1	n_{11}	n_{12}	n_{13}	R_1
B_2	n_{21}	n_{22}	n_{23}	R_2
B_3	n_{31}	n_{32}	n_{33}	R_3
B_4	n_{41}	n_{42}	n_{43}	R_4
Total	C_1	C_2	C_3	N

donde n_{11}, . . . , n_{43} denotan las 12 frecuencias de las celdas frecuencias absolutas, R_1..., R_4 denotan los totales de frecuencia marginal de las cuatro filas, C_1, C_2, y C_3 denotan los totales de las frecuencias marginales de las tres columna y

$$N = \sum_{i=1}^{4} \sum_{j=1}^{3} n_{ij}$$

Cuando sólo hay dos filas, como en el ejemplo anterior de 2×2, cada valor de probabilidad

de columna es binomial, pero con cuatro filas cada valor de probabilidad de columna es multinomial. Es bien sabido que un valor de probabilidad multinomial puede obtenerse de una serie interconectada de expresiones binomiales. Por ejemplo, para la columna A_1 de la *Tabla 2.2*:

$$\left(\begin{array}{c} C_1 \\ n_{11} \end{array} \right) \left(\begin{array}{c} C_1 - n_{11} \\ n_{21} \end{array} \right) \left(\begin{array}{c} C_1 - n_{11} - n_{21} \\ n_{31} \end{array} \right) =$$

$$= \frac{C_1!}{n_{11}!(C_1 - n_{11})!} \times \frac{(C_1 - n_{11})!}{n_{21}!(C_1 - n_{11} - n_{21})!} \times \frac{(C_1 - n_{11} - n_{21})!}{n_{31}(C_1 - n_{11} - n_{21} - n_{31})!} = \frac{C_1!}{n_{11}!n_{21}!n_{31}!n_{42}!}$$

Para la columna A_2:

$$\left(\begin{array}{c} C_2 \\ n_{12} \end{array} \right) \left(\begin{array}{c} C_2 - n_{12} \\ n_{22} \end{array} \right) \left(\begin{array}{c} C_2 - n_{12} - n_{22} \\ n_{32} \end{array} \right) =$$

$$= \frac{C_2!}{n_{12}!(C_2 - n_{12})!} \times \frac{(C_2 - n_{12})!}{n_{22}!(C_2 - n_{12} - n_{22})!} \times \frac{(C_2 - n_{12} - n_{22})!}{n_{32}(C_2 - n_{12} - n_{22} - n_{32})!} = \frac{C_2!}{n_{12}!n_{22}!n_{32}!n_{42}!}$$

y para la distribución de frecuencias marginales de las filas:

$$\left(\begin{array}{c} N \\ R_1 \end{array} \right) \left(\begin{array}{c} N - R_1 \\ R_2 \end{array} \right) \left(\begin{array}{c} N - R_1 - R_2 \\ R_3 \end{array} \right) =$$

$$= \frac{N!}{R_1!(N - R_1)!} \times \frac{(N - R_1)!}{R_2!(N - R_1 - R_2)!} \times \frac{(N - R_1 - R_2)!}{R_3!(N - R_1 - R_2 - R_3)!} = \frac{N!}{R_1!R_2!R_3!R_4!}$$

Por consiguiente, para tablas de contingencia rxc se tiene:

$$p(n_{ij}|R_i, C_j, N) = \frac{(\prod_{i=1}^{r} R_i!)(\prod_{j=1}^{c} C_j!)}{N! \prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!}$$

De esta forma, la ecuación anterior puede generalizarse fácilmente a tablas de contingencia multidireccionales más complejas [Mielke and Berry, 1988a].

Mientras que esta sección ilustra el cálculo de un valor de probabilidad puntual hipergeométrico, para una prueba de permutación exacta de una tabla de contingencia $r \times c$ es necesario calcular la medida de asociación seleccionada para las frecuencias de celdas observadas y, a continuación, enumerar exhaustivamente todos los posibles ordenamientos igualmente probables de los N objetos en las rcceldas, dadas las distribuciones de frecuencia marginal observadas.

Para cada ordenamiento en el conjunto de referencia de todas las permutaciones de las frecuencias de las celdas, se calcula una medida de asociación, T , y el valor exacto de

la probabilidad puntual hipergeométrica, $p(n_{ij}|R_i, C_j, N)$ para $i = 1,\ldots ,r$ y $j = 1, \ldots ,c$. Sea T_0 el valor del estadístico de prueba observado, es decir, la medida de asociación, entonces, el valor de probabilidad exacto de T_0 es la suma de la probabilidad de los puntos hipergeométricos asociados a los valores de T calculados en todos los posibles ordenamientos de las frecuencias de las celdas que son iguales o mayores que T_0.

Cuando el número de disposiciones posibles de las frecuencias es muy grande, las pruebas exactas son poco prácticas y se hacen necesarios los métodos estadísticos de permutación de Monte Carlo. Dadas las distribuciones de frecuencias marginales observadas, los métodos estadísticos de permutación de Monte Carlo generan una muestra aleatoria de todas las posibles ordenaciones de las frecuencias de las celdas, extraídas con reemplazo. El valor de la probabilidad a dos colas del remuestreo es simplemente la proporción de los valores T calculados en los ordenamientos seleccionados aleatoriamente que son iguales o mayores que T_0. En el caso del remuestreo de Monte Carlo los valores de probabilidad hipergeométricos no están implicados, simplemente se necesita la proporción de los valores de las medidas de asociación (valores T) iguales o mayores que el valor de la medida de asociación observada (T_0).

2.2. Medidas λ_a y λ_b de Goodman y Kruskal

Un problema común al que se enfrentan muchos investigadores es el análisis de una tabla de clasificación cruzada en la que ambas variables son categóricas, ya que las variables categóricas no suelen contener tanta información como las variables de nivel ordinal o de intervalo. Las medidas basadas en la chi-cuadrado, como la ϕ^2 de Pearson, la T^2 de Tschuprov, la V^2 de Cramér y la C de Pearson, han demostrado ser menos que satisfactorias debido a las dificultades de interpretación.

En 1954, Leo Goodman y William Kruskal propusieron varias medidas nuevas de asociación [Goodman and Kruskal, 1954]. Entre las medidas se encontraban dos medidas de predicción asimétricas de reducción proporcional en el error para los análisis de una muestra aleatoria de dos variables categóricas: λ_a, para cuando se considera que A es la variable dependiente, y λ_b, para cuando se considera que B es la variable dependiente.

Sea una tabla de contingencia $r \times c$ como la representada a continuación:

Tabla 2.3: Notación para la clasificación cruzada de dos variables categóricas, A y B, con c y r categorías respectivamente

B\A	A_1	A_2	$\ldots$	A_c	Total
B_1	n_{11}	n_{12}	$\ldots$	n_{1c}	$n_{1.}$
B_2	n_{21}	n_{22}	$\ldots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.c}$	N

donde A_j con $j = 1, ..., c$ representan las c categorías de la variable dependiente A, B_i con $i = 1, ..., r$ denotan las r categorías para la variable independiente B, n_{ij} la frecuencia absoluta de celda para $i = 1, ...r$ y $j = 1, ...c$, y N es el total observaciones. Denotamos

por un punto (.) la suma parcial de todas las filas o todas las columnas, según la posición del (ů) en la lista de subíndices. Si el (ů) está en la primera posición del subíndice la suma es sobre todas las filas y si el (ů) está en la segunda posición de subíndice, la suma es sobre todas las columnas. Por lo tanto, $n_{i.}$ denota el total de la frecuencia marginal de la i-ésima fila, $i = 1, ...r$, sumada en todas las columnas, y $n_{.j}$ indica la frecuencia marginal de la j-ésima columna, $j = 1, ...c$, sumada en todas las filas.

Teniendo en cuenta la notación de la *tabla 3*, se definen

$$W = \sum_{i=1}^{r} max(n_{i1}, n_{i2}, ..., n_{ic}) \qquad y \qquad X = max(n_{.1}, n_{.2}, ..., n_{.c})$$

Entonces, λ_a (siendo A la variable dependiente) viene dado por:

$$\lambda_a = \frac{W - X}{N - X}$$

De la misma manera, se definen

$$Y = \sum_{j=1}^{c} max(n_{1j}, n_{2j}, ..., n_{rj}) \qquad y \qquad Z = max(n_{1.}, n_{2.}, ..., n_{r.})$$

E igualmente, λ_b (siendo B la variable dependiente) viene dado por:

$$\lambda_b = \frac{Y - Z}{N - Z}$$

Tanto λ_a como λ_b son medidas de reducción proporcional del error. Consideremos λ_a y dos posibles casos:

- Caso 1: Sólo son conocidas las categorías disjuntas de la variable dependiente A.
- Caso 2: Son conocidas tanto las categorías disjuntas de la variable A como las categorías disjuntas de la variable independiente B.

En el caso 1, es conveniente que el investigador adivine la categoría de la variable dependiente A que tiene la mayor frecuencia marginal total (moda), que en este caso es $X = max(n_{.1}, ..., n_{.c})$. Entonces, la probabilidad de error es $N - X$; definimos esto como "errores del primer tipo" o E_1. En el caso 2, es conveniente que el investigador adivine la categoría de la variable dependiente A que tiene la mayor frecuencia absoluta (moda) en cada categoría de la variable independiente B, que en este caso es

$$W = \sum_{i=1}^{r} max(n_{i1}, n_{i2}, ..., n_{ic})$$

La probabilidad de error es entonces $N - W$; y lo definimos como "errores del segundo tipo" o E_2. Entonces, λ_a puede expresarse como

$$\lambda_a = \frac{E_1 - E_2}{E_1} = \frac{N - X - (N - W)}{N - X} = \frac{W - X}{N - X}$$

Como señalaron Goodman y Kruskal en 1954, se observó inmediatamente un problema con las interpretaciones tanto de λ_a como de λ_b. Dado que ambas medidas estaban basadas en los valores modales de las categorías de la variable independiente, cuando los valores modales ocurrían todos en la misma categoría de la variable dependiente, λ_a y λ_b serían cero. Así, mientras que λ_a y λ_b sean iguales a cero bajo independencia, λ_a y λ_b también podían ser iguales a cero para casos distintos de independencia. Esto hace que tanto λ_a como λ_b sean difíciles de interpretar; en consecuencia, λ_a y λ_b rara vez se encuentran en estudios de la asociación de variables categóricas.

Así, como explicaron Goodman y Kruskal en 1954:

1. λ_a es indeterminada si y sólo si la población se encuentra en una columna; es decir, aparece en una categoría de la variable A.
2. En caso contrario, el valor de λ_a se encuentra entre 0 y 1.
3. λ_a es 0 si y sólo si el conocimiento de la clasificación B no ayuda a predecir la clasificación A.
4. λ_a es 1 si y sólo si el conocimiento de un objeto de la categoría B especifica completamente su categoría A, es decir, si cada fila de la tabla de clasificación cruzada contiene como máximo un valor distinto de cero.
5. En el caso de la independencia estadística, λ_a, cuando está determinada, es cero. Lo contrario no tiene por qué ser cierto: λ_a puede ser cero sin que haya independencia estadística.
6. λ_a no cambia con ninguna permutación de filas o columnas.

2.3. Medidas t_a y t_b de Goodman y Kruskal

Como se ha señalado anteriormente, en 1954 Leo Goodman y William Kruskal propusieron varias medidas de asociación nuevas. Entre las medidas había una medida de predicción asimétrica de reducción proporcional del error, t_a, para el análisis de una muestra aleatoria de dos variables categóricas [**?**]. Se consideran dos variables desordenadas cruzadas, A y B, con la variable A como variable dependiente y la variable B como variable independiente. La *tabla 2.3* proporcionaba la notación para la clasificación cruzada.

El estadístico t_a de Goodman y Kruskal es una medida de la reducción relativa del error de predicción en la que se definen dos tipos de errores. El primer tipo es el error de predicción basado únicamente en el conocimiento de la distribución de la variable dependiente, denominado "error del primer tipo" (E_1) y que consiste en el número esperado de errores al predecir las c categorías de la variable dependiente$(a_1, ..., a_c)$ a partir de la distribución observada de las marginales de la variable dependiente $(n_{.1}, ..., n_{.c})$. El segundo tipo es el error de predicción basado en el conocimiento de las distribuciones de las de la variable independiente y de la dependiente, denominado "error del segundo tipo" (E_2) y que consiste en el número o los errores esperados al predecir las c categorías de la variable dependiente $(a_1, ..., a_c)$ a partir del conocimiento de las r categorías de la variable independiente $(b_1, ..., b_r)$.

Para ilustrar los dos tipos de error, se considera la predicción de la categoría A_1 a partir de sólo el conocimiento de su distribución marginal, $n_{.1}, ...n_{.c}$. Claramente, $n_{.1}$ de los N casos totales están en la categoría a_1, pero se desconoce exactamente qué $n_{.1}$ de los N

casos. La probabilidad de identificar incorrectamente uno de los N casos de la categoría a_1 viene dada por:

$$\frac{N - n_{.1}}{N}$$

Dado que se requieren $n_{.1}$ clasificaciones de este tipo, el número de clasificaciones incorrectas esperadas esperadas es

$$\frac{n_{.}1(N - n_{.}1)}{N}$$

y, para todas las c categorías de la variable A, el número de errores esperados del primer tipo viene dado por:

$$E_1 = \sum_{j=1}^{c} \frac{n_{.}j(N - n_{.}j)}{N}$$

Asimismo, para predecir $n_{11}, ..., n_{1c}$ a partir de la categoría independiente B_1, la probabilidad de clasificar incorrectamente uno de los $n_{1.}$ casos de la celda n_{11} es:

$$\frac{n_{1.} - n_{11}}{n_{1.}}$$

Dado que se requieren n_{11} clasificaciones de este tipo, el número de clasificaciones incorrectas es

$$\frac{n_{11}(n_{1.} - n_{11})}{n_{1.}}$$

y, para todas las cr celdas, el número de errores esperados del segundo tipo viene dado por:

$$E_2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{n_{ij}(n_{i.} - n_{ij})}{n_{i.}}$$

El estadístico ta de Goodman y Kruskal se define, por tanto, como:

$$t_a = \frac{E_1 - E_2}{E_1}$$

Una forma de cálculo eficiente para la t_a de Goodman y Kruskal viene dada por:

$$t_a = \frac{N \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{c} n_{.j}^2}{N^2 - \sum_{j=1}^{c} n_{.j}^2}$$

Un valor calculado de t_a indica la reducción proporcional del error de predicción dado el conocimiento de la distribución de la variable independiente B por encima del conocimiento de la distribución de la variable dependiente A. Como se define, t_a es un estimador puntual del parámetro poblacional τ_a de Goodman y Kruskal para la población de la que se obtuvo la muestra de N casos. Si la variable B se considera la variable dependiente y la variable A la variable independiente, entonces el estadístico de prueba de Goodman y Kruskal t_b y el parámetro poblacional asociado τ_b se definen análogamente.

Como el parámetro t_a toma valores de 0 a 1, posee una interpretación clara y significativa de reducción proporcional del error [Costner, 1965], y se caracteriza por una alta validez intuitiva [Hunter, 1973]; el estadístico de prueba t_a plantea dificultades si la hipótesis nula es que $H_0 : t_a = 0$ [Margolin and Light, 1974]. El problema es que la distribución muestral de t_a no es asintóticamente normal bajo la hipótesis nula. En consecuencia, la aplicabilidad de t_a de Goodman y Kruskal a las pruebas típicas de hipótesis nulas se ha visto muy limitada.

Aunque la t_a fue desarrollada por Goodman y Kruskal en 1954, no fue hasta 1963 que se estableció la normalidad asintótica para t_a y hasta 1972 no se obtuvo la varianza asintótica correcta para t_a, pero sólo para $0 < \tau_a < 1$ [Goodman and Kruskal, 1963].

En 1971, Richard Light y Barry Margolin desarrollaron R^2, una técnica de análisis de la varianza para las variables de respuesta categóricas, llamada CATANOVA para CATegorical ANalysis Of VAriance [Light and Margolin, 1971]. Light y Margolin aparentemente no sabían que el R^2 era idéntico al t_a de Goodman y Kruskal y que habían resuelto el viejo problema de probar $H_0 : \tau_a = 0$. La igualdad entre R^2 y t_a fue reconocida por primera vez por Särndal en 1974 [Särndal, 1974] y posteriormente discutida por Margolin y Light [Margolin and Light, 1974], donde demostraron que $t_a(N-1)(r-1)$ se distribuía según una chi-cuadrado con $(r-1)(c-1)$ grados de libertad bajo $H_0 : \tau_a = 0$ cuando $N \to \infty$ [Berry and Mielke, 1985].

2.4. Una prueba asimétrica de homogeneidad

A veces, se necesita determinar si las proporciones de elementos en un conjunto de categorías mutuamente excluyentes son las mismas para dos o más grupos. Cuando se extraen muestras aleatorias independientes de cada uno de los $g \geq 2$ grupos y luego se clasifican en $r \geq 2$ categorías mutuamente excluyentes, la prueba adecuada es una prueba de homogeneidad de las distribuciones g. En una prueba de homogeneidad, una de las distribuciones marginales se conoce antes de recoger los datos, es decir, los totales de frecuencia marginal de fila o columna que indican el número de elementos de cada uno de los g grupos. Esto se denomina muestreo multinomial de *producto*, ya que la distribución de muestreo es el producto de g distribuciones multinomiales y la hipótesis nula es que las g distribuciones multinomiales son idénticas.

Una prueba de homogeneidad es bastante diferente de una prueba de independencia, en la que se extrae una única muestra y se clasifica en ambas variables. En una prueba de independencia, ambos conjuntos de totales de frecuencias marginales se conocen sólo después de que se hayan recogido los datos. Esto se denomina muestreo multinomial simple, ya que la distribución del muestreo es una distribución multinomial[Böhning and Holling, 1989]. La prueba de homogeneidad más utilizada es la prueba chi-cuadrado de Pearson con $gl = (r-1)(g-1)$ grados de libertad. La prueba chi-cuadrado de homogeneidad

de Pearson supone la hipótesis nula de que no hay diferencia en las proporciones de sujetos en un conjunto de categorías mutuamente excluyentes entre dos o más poblaciones [Marascuilo and McSweeney, 1977].

La prueba de homogeneidad chi-cuadrado de Pearson es una prueba simétrica, que produce un solo valor para una tabla de contingencia $r \times g$. Por el contrario, una prueba asimétrica arroja dos valores dependiendo de la variable que se considere dependiente. Como señala Berkson, si las diferencias son todas en una dirección, una prueba simétrica como la chi-cuadrado es insensible a este hecho [Berkson, 1938].

Una prueba simétrica de homogeneidad, por su naturaleza, excluye la información conocida sobre los datos: qué variable es la independiente y qué variable es la dependiente. Aunque a veces es necesario reducir el nivel de medición cuando no se pueden cumplir los requisitos de distribución, en general no es aconsejable utilizar una prueba estadística que descarte información importante.

Se consideran las variables A y B, con B la variable dependiente. Sean $B_1, ..., B_r$ las $r \geq 2$ categorías de la variable dependiente, $A_1, ..., A_g$ las $g \geq 2$ categorías de la variable independiente, n_{ij} la frecuencia de celda en la i-ésima fila y j-ésima columna, $i = 1, ., r$ y $j = 1, ..., g$, y N el tamaño total de la muestra. Sean $n_{1.}, ...n_{r.}$ las frecuencias marginales de la variable B y $n_{.1}, \ldots, n_{.g}$ las frecuencias marginales totales de la variable A. La clasificación cruzada de las variables A y B se muestra en la tabla siguiente:

Tabla 2.4: Notación para la clasificación cruzada de dos variables categóricas, A y B, con g y r categorías respectivamente

B\A	A_1	A_2	$\ldots$	A_g	Total
B_1	n_{11}	n_{12}	$\ldots$	n_{1g}	$n_{1.}$
B_2	n_{21}	n_{22}	$\ldots$	n_{2g}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
B_r	n_{r1}	n_{r2}	$\ldots$	n_{rg}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.g}$	N

Aunque nunca se ha propuesto como prueba de homogeneidad, la prueba asimétrica t_b, introducida por primera vez por Goodman y Kruskal en 1954 [Goodman and Kruskal, 1954], es una alternativa atractiva a la prueba simétrica de homogeneidad, la chi-cuadrado. El estadístico de la prueba viene dado por:

$$t_b = \frac{N \sum_{j=1}^{g} \sum_{i=1}^{r} \frac{n_{ij}^2}{n_{.j}} - \sum_{i=1}^{r} n_{i.}^2}{N^2 - \sum_{i=1}^{r} n_{i.}^2}$$

donde B es la variable dependiente y el parámetro poblacional asociado se denota como τ_b. Si se considera que la variable dependiente es A, el estadístico de la prueba viene dado por dado por:

$$t_a = \frac{N \sum_{i=1}^{r} \sum_{j=1}^{g} \frac{n_{ij}^2}{n_{i.}} - \sum_{j=1}^{g} n_{.j}^2}{N^2 - \sum_{j=1}^{g} n_{.j}^2}$$

y el parámetro poblacional asociado es τ_a.

El estadístico de prueba t_b toma valores entre 0 y 1; t_b es 0 si y sólo si hay homogeneidad sobre las r categorías de la variable dependiente,B, para todos los g grupos, y t_b es 1 si y sólo si el conocimiento de la variable A_j, con $j = 1, ..., g$, determina completamente el conocimiento de la variable B_i para $i = 1, ..., r$. Del mismo modo, el estadístico de prueba t_a es 0 si y sólo si existe homogeneidad sobre las g categorías de la variable dependiente,A, para todos los r grupos, y t_a es 1 si y sólo si el conocimiento de la variable B_i, con $i = 1, ..., r$, determina completamente el conocimiento de la variable A_j para $j = 1, ..., g$.

Aunque no existe una equivalencia general para los estadísticos de prueba t_b o t_a con χ^2, existen ciertas relaciones en condiciones especiales: Si $g = 2$, entonces $\chi^2 = Nt_b$, y si $g > 2$ y $n_{.j} = N/g$ para $j = 1, ..., g$, entonces $\chi^2 = N(g-1)t_b$. Del mismo modo, si $r = 2$, $\chi^2 = Nt_a$, y si $r > 2$ y $n_{i.} = N/r$ para $i = 1, ..., r$, entonces $\chi^2 = N(r-1)t_a$. De lo anterior, se deduce que si $r = g = 2$, entoces $t_b = t_a = \chi^2/N$, que es el coeficiente de contingencia al cuadrado de la media de Pearson, ϕ^2. Por último, cuando $N \to \infty$, se tiene que $t_b(N-1)(r-1)$ y $t_a(N-1)(g-1)$ se distribuyen según una chi-cuadrado con $(r-1)(g-1)$ grados de libertad.

Existen tres métodos para determinar el valor de la probabilidad del estadístico de prueba t_b o t_a calculado: procedimientos exactos, de remuestreo de Montecarlo y asintóticos. Explicaciones considerando sólo t_b, pero los métodos son análogos para t_a:

- Valores exactos de probabilidad: Bajo la hipótesis nula, $H_0 : \tau_b = 0$, cada uno de los M posibles ordenamientos de los N elementos sobre las categorías rg de la tabla de contingencia es igualmente probable con distribuciones marginales fijas. Para cada ordenamiento de los datos observados en el conjunto de referencia de todos los ordenamientos posibles, se calcula el estadístico de prueba deseado. El valor de la probabilidad exacta de un estadístico de prueba t_b observado es la suma de los valores de probabilidad puntuales hipergeométricos asociados a valores mayores o iguales a t_b.
- Valores de probabilidad de remuestreo: Una prueba exacta no es práctica desde el punto de vista computacional, excepto para muestras bastante pequeñas. Un método alternativo que evita las exigencias computacionales de una prueba exacta es una aproximación de permutación por remuestreo. Bajo la hipótesis nula, $H_0 : \tau_b = 0$, las pruebas de permutación por remuestreo generan y estudian un subconjunto aleatorio de Monte Carlo de todos los posibles ordenamientos igualmente probables de los datos observados. Para cada ordenamiento seleccionado al azar de los datos observados, se calcula el estadístico de prueba deseado. El valor de la probabilidad de remuestreo de Montecarlo de un estadístico de prueba t_b observado es simplemente la proporción de los valores seleccionados aleatoriamente de t_b iguales o mayores que el valor observado de t_b.
- Valores de probabilidad asintótica: Bajo la hipótesis nula, $H_0 : \tau_b = 0$, como $N \to \infty$, $t_b(N-1)(g-1)$ se distribuye según una chi-cuadrado con $(r-1)(g-1)$ grados

de libertad. El valor de la probabilidad asintótica es la proporción de la chi-cuadrado apropiada igual o mayor que el valor observado de $t_b(N-1)(g-1)$.

2.5. Medidas de concordancia

La medición de la concordancia es un caso especial de medición de la asociación entre dos o más variables. Una serie de problemas de investigación estadística requieren medir la concordancia, en lugar de la asociación o la correlación. Los índices de concordancia miden el grado en que un conjunto de medidas de respuesta son idénticas a otro conjunto, es decir, concuerdan.

Su uso surge se asignan objetos a un conjunto de categorías desordenadas y disjuntas. En 1957, Robinson explicó que la concordancia estadística requiere que los valores emparejados sean idénticos, mientras que la correlación sólo requiere que los valores emparejados estén empatados por alguna función matemática [Robinson, 1957]. Por tanto, la concordancia es una medida más restrictiva que la correlación. Robinson argumentó que la distinción entre acuerdo y correlación lleva a la conclusión de que una estimación lógicamente correcta de la fiabilidad de una prueba viene dada por el coeficiente de correlación intraclase en lugar del coeficiente de correlación de Pearson (interclase) y que el concepto de acuerdo, en lugar de correlación, es la base adecuada de la teoría de la fiabilidad.

Según Berry, una medida de acuerdo entre evaluadores debería, como mínimo, incorporar siete atributos básicos [Berry and Mielke, 1988]:

- Una medida de concordancia debe ser corregida por el azar, es decir, cualquier coeficiente de concordancia debe reflejar la cantidad de concordancia que excede lo que se esperaría por el azar. Varios investigadores han defendido las medidas de concordancia corregidas por el azar, como Brennan y Prediger [Brennan and Prediger, 1981] o Cicchetti, Showalter y Tyrer [Cicchetti et al., 1985]. Aunque algunos otros han argumentado en contra, por ejemplo, Armitage, Blendis y Smyllie [Armitage et al., 1966]o Goodman y Kruskal [Goodman and Kruskal, 1954], aunque los partidarios de las medidas de concordancia corregidas por el azar superan ampliamente a los detractores.
- Una medida de concordancia entre evaluadores posee una ventaja añadida si es directamente aplicable a la evaluación de la fiabilidad. Robinson, en particular, hizo hincapié en que la fiabilidad no podía medirse simplemente mediante alguna función de la correlación producto-momento de Pearson, y argumentó que el concepto de concordancia debería ser la base de la teoría de la fiabilidad, no la correlación [Robinson, 1957].
- Varios investigadores han comentado la simplicidad de la distancia Euclídea para las medidas de concordancia entre evaluadores, señalando que la elevación al cuadrado de las diferencias entre los valores de las escalas es, en el mejor de los casos, cuestionable, aunque reconocen que las diferencias al cuadrado permiten interpretaciones más claras de los coeficientes [Fleiss, 1973][Krippendorff, 1970]. Además, Graham y Jackson señalaron que la elevación al cuadrado de las diferencias entre los valores, es decir, la ponderación cuadrática, da como resultado una medida de asociación, no de concordancia [Graham and Jackson, 1993]. Por lo tanto, la distancia Euclídea es una propiedad deseada para las medidas de concordancia entre evaluadores.

- Toda medida de acuerdo debe tener una base estadística. Una medida de concordancia sin una prueba de significación adecuada está muy limitada en su aplicación a situaciones prácticas de investigación. Los análisis asintóticos son interesantes y útiles, en condiciones de muestras grandes, pero suelen tener una utilidad práctica limitada cuando el tamaño de las muestras es pequeño.
- Una medida de concordancia que sirva para datos multivariantes tiene una ventaja decisiva sobre las medidas de acuerdo univariantes. Así, si un observador localiza un conjunto de datos en un espacio r-dimensional, una medida de concordancia multivariante puede determinar el grado en que un segundo observador localiza el mismo conjunto de datos en el espacio r-dimensional definido.
- Una medida de concordancia debe ser capaz de analizar los datos en cualquier nivel de medición. La medida kappa de Cohen para la concordancia entre evaluadores es, actualmente, la medida de concordancia más utilizada. Se han establecido extensiones de la kappa de Cohen a datos clasificados de forma incompleta por Iachan [Iachan, 1984] y a datos categóricos continuos por Conger [Conger, 1985]. Una extensión de la medida de acuerdo kappa de Cohen a datos ordinales totalmente clasificados y a datos de intervalo fue proporcionada por Berry y Mielke en 1988 [Berry and Mielke, 1988].
- Una medida de concordancia debe ser capaz de evaluar la información de más de dos calificadores o jueces. Fleiss propuso una medida de concordancia para múltiples calificadores en una escala nominal [Fleiss, 1971]. Landis y Koch consideraron la concordancia entre varios calificadores en términos de una opinión mayoritaria [Landis and Koch, 1977]. Light se centró en una extensión de la medida kappa de Cohen [Cohen, 1960] de concordancia entre calificadores a múltiples calificadores que se basaba en la media de todos los valores kappa por pares [Light, 1971]. Lamentablemente, la medida propuesta por Fleiss dependía de la proporción media de calificadores que estaban de acuerdo en la clasificación de cada observación y la formulación de Landis y Koch se vuelve computacionalmente prohibitiva si el número de observadores o el número de categorías de respuesta es grande. Además, la extensión de kappa propuesta por Fleiss no se redujo a una kappa de Cohen cuando el número de calificadores era de dos.

2.5.1. Medida de concordancia de Robinson

Una de las primeras medidas de concordancia máxima corregida fue desarrollada por W.S. Robinson en 1957 [Robinson, 1957]. Supongamos que hay $k = 2$ jueces y califican independientemente N objetos. Robinson argumentó que la correlación producto-momento de Pearson (interclase) calculada entre las calificaciones de dos jueces era una medida inadecuada de concordancia porque mide el grado en que los valores emparejados de las dos variables son proporcionales, cuando se expresan como desviaciones de sus medias, en lugar de ser idénticos. Robinson propuso una nueva medida de concordancia basada en el coeficiente de correlación intraclase que denominó A. Sean dos conjuntos de valoraciones con N pares de valores. Robinson definió A como

$$A = 1 - \frac{D}{D_{max}},$$

donde D (de Desacuerdo) viene dado por:

$$D = \sum_{i=1}^{N}(X_{1i} - \bar{X}_i)^2 + \sum_{i=1}^{N}(X_{2i} - \bar{X}_i)^2$$

y X_{1i} = valor de X_1 para el i-ésimo par de valoraciones, X_{2i} = valor de X_2 para el i-ésimo par de valoraciones, $\bar{X}_i$ = la media de X_1 y X_2 para el i-ésimo par de valoraciones.

Robinson observó que, por sí misma, D no es una medida muy útil porque implica X_1 y X_2. Para encontrar una medida de concordancia relativa, más que absoluta, Robinson estandarizó D por su rango de variación posible, dado por:

$$D_{max} = \sum_{i=1}^{N}(X_{1i} - \bar{X})^2 + \sum_{i=1}^{N}(X_{2i} - \bar{X})^2,$$

donde la media viene dada por:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_{1i} + \sum_{i=1}^{N} X_{2i}}{2N}$$

Coeficiente de correlación intraclase

Es bien sabido que el coeficiente de correlación intraclase (r_I) entre N pares de observaciones sobre dos variables es, por definición, el momento producto ordinario de Pearson (interclase) entre $2N$ pares de observaciones, de los cuales los primeros N son las observaciones originales, y los segundos N las observaciones originales con X_{1i} sustituyendo a X_{2i} y viceversa, con $i = 1, ...N$ [Fisher, 1934]:

$$r_I = \frac{N\sum_{i=1}^{N} X_{1i}X_{2i} - \sum_{i=1}^{N} X_{1i}\sum_{i=1}^{N} X_{2i}}{\sqrt{[N\sum_{i=1}^{N} X_{1i}^2 - (\sum_{i=1}^{N} X_{1i})^2][N\sum_{i=1}^{N} X_{2i}^2 - (\sum_{i=1}^{N} X_{2i})^2]}}$$

Para el caso de dos variables, las relaciones entre el coeficiente de concordancia de Robinson y el coeficiente de correlación intraclase vienen dadas por:

$$r_I = 2A - 1; A = \frac{r_I + 1}{2}$$

Por tanto, en el caso de dos variables, la correlación intraclase es una función lineal simple del coeficiente de concordancia.

Para $k > 2$ conjuntos de valoraciones, las relaciones entre el coeficiente de correlación intraclase y la A de Robinson no son tan simples y vienen dadas por:

$$r_I = \frac{kA - 1}{k - 1}; A = \frac{r_I(k - 1) + 1}{k}.$$

De las expresiones anteriores se observa que el valor del coeficiente intraclase no depende sólo de A sino también de k, el número variables. El rango de A de Robinson siempre

incluye los valores desde 0 hasta 1, independientemente del número de observaciones. Por lo tanto, las comparaciones entre los coeficientes de concordancia basados en diferentes números de variables son equiparables. El límite superior del coeficiente de correlación intraclase es siempre la unidad, pero su límite inferior es $-1/(k-1)$. Para $k = 2$ variables, el límite inferior de r_I es -1, pero para $k = 3$ variables es $-1/2$, para $k = 4$ es $-1/3$, para $k = 5$ el límite inferior es $-1/4$, y así sucesivamente.

2.5.2. Medida de concordancia π de Scott

Una de las primeras medidas de concordancia corregida por el azar fue introducida por William Scott en 1955 [Scott, 1955]. Supongamos que dos jueces o calificadores clasifican independientemente cada una de las N observaciones en una de las c categorías. Las clasificaciones resultantes pueden mostrarse en una tabla de contingencia $c \times c$, con las frecuencias absolutas en cada celda. Sea $n_{i.}$ la frecuencia marginal de la i-ésima fila, $i = 1, ...r$; sea $n_{.j}$ la frecuencia marginal de la j-ésima columna, $j = 1, ...c$ y sea

$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$$

frecuencia total de la tabla. El coeficiente de concordancia de Scott para los datos de nivel nominal viene dado por:

$$\pi = \frac{p_o - p_e}{1 - p_e}$$

donde

$$p_o = \frac{1}{N} \sum_{i=1}^{c} n_{ij} \qquad y \qquad p_e = \frac{1}{4N^2} \sum_{k=1}^{c} (n_{.k} + n_{k.})^2$$

En esta configuración, p_o es la proporción observada de observaciones en las que los jueces están de acuerdo, p_e es la proporción de observaciones para las que se espera una concordancia por azar, $p_o - p_e$ es la proporción de concordancia más allá de la esperada por el azar, $1 - p_e$ es la máxima proporción posible de concordancia más allá de la esperada por el azar, y π de Scott es la proporción de concordancia entre los dos jueces, una vez eliminada la concordancia por azar.

Aunque la π de Scott es interesante desde una perspectiva histórica, esta medida ha caído en desuso. Basada en proporciones conjuntas, la π de Scott asume que los dos jueces tienen la misma distribución de respuestas. La medida κ de Cohen no hace esta suposición y, en consecuencia, ha surgido como la medida preferida de concordancia entre evaluadores corregida por el azar para dos jueces/calificadores.

2.5.3. Medida de concordancia κ de Cohen

Actualmente, la medida más popular de concordancia entre dos jueces o calificadores es la medida corregida por el azar, propuesta por primera vez por Jacob Cohen en 1960 y denominada kappa [Cohen, 1960]. La kappa de Cohen mide la magnitud de la concordancia entre $b = 2$ observadores en la asignación de N objetos a un conjunto de c categorías

disjuntas y desordenadas. En 1968, Cohen propuso una versión de kappa que permitía ponderar las c categorías [Cohen, 1968]. Mientras que el kappa original (no ponderado) no distinguía entre magnitudes de desacuerdo, el kappa ponderado incorporaba la magnitud de cada desacuerdo y proporcionaba un crédito parcial para las discordancias cuando la concordancia no era completa. El enfoque habitual consiste en asignar pesos a cada par de desacuerdos, con pesos mayores que indican un mayor desacuerdo.

Tanto en el caso no ponderado como en el ponderado, kappa es igual a +1 cuando se produce una concordancia perfecta entre dos o más jueces, 0 cuando la concordancia es igual a la esperada en condiciones de independencia, y negativo cuando la concordancia es inferior a la esperada por azar. En esta sección se estudiará únicamente la kappa no ponderada, ya que es la que se usa normalmente para datos categóricos no ordenados.

Supongamos que dos jueces o calificadores clasifican independientemente cada una de las N observaciones en una de las c categorías desordenadas, exhaustivas y mutuamente excluyentes. Las clasificaciones resultantes pueden mostrarse en una clasificación cruzada $c \times c$ con proporciones para las entradas de las celdas:

Tabla 2.5: Clasificación cruzada cxc con proporciones en celdas

Fila\Columna	**1**	**2**	...	**c**	**Total**
1	p_{11}	p_{12}	...	p_{1c}	$p_{1.}$
2	p_{21}	p_{22}	...	p_{2c}	$p_{2.}$
⋮	⋮	⋮	⋱	⋮	⋮
c	p_{c1}	p_{c2}	...	p_{cc}	$p_{c.}$
Total	$p_{.1}$	$p_{.2}$	...	$p_{.c}$	$p_{..}$

Donde, $p_{i.}$ denota la proporción marginal de la i-ésima fila, $i = 1, ...c$; $p_{.j}$ denota la proporción marginal de la j-ésima columna, $j = 1, ...c$ y $p_{..} = 1,00$. Usando esta notación, el coeficiente kappa no ponderado de Cohen para datos de nivel nominal viene dado por:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

donde

$$p_o = \sum_{i=1}^{c} p_{ii} \qquad y \qquad p_e = \sum_{i=1}^{c} p_{i.} p_{.i}$$

El kappa de Cohen también puede definirse en términos de valores de frecuencia absoluta, lo que hace que los cálculos sean algo más sencillos. Así,

$$\kappa = \frac{\sum_{i=1}^{c} O_{ii} - \sum_{i=1}^{c} E_{ii}}{N - \sum_{i=1}^{c} E_{ii}},$$

donde O_{ii} denota el valor de frecuencia de celda observado en la diagonal principal de la tabla de concordancia $c \times c$, E_{ii} denota un valor de frecuencia de celda esperado en la diagonal principal, y

$$E_{ii} = \frac{n_{i.} n_{.i}}{N}, \qquad i = 1, .., c.$$

En la configuración de la tabla anterior, p_o es la proporción observada de observaciones en las que los jueces están de acuerdo, p_e es la proporción de observaciones para las que se espera una concordancia es la proporción de concordancia esperada por el azar, $p_o - p_e$ es la proporción de concordancia más allá de la esperada por azar, $1 - p_e$ es la proporción máxima posible de concordancia más allá de lo esperado por el azar, y el estadístico de la prueba kappa de Cohen es la proporción de concordancia entre los dos jueces, una vez eliminada la concordancia por azar.

2.5.4. Aplicación con Varios Jueces

La medida κ de Cohen de la concordancia entre evaluadores corregida por el azar fue originalmente diseñada sólo para $b = 2$ jueces. En esta sección, se introduce un procedimiento para calcular κ no ponderado con múltiples jueces. Aunque el procedimiento es apropiado para cualquier número de $c \geq 2$ categorías disjuntas y desordenadas y $b \geq 2$ jueces, la descripción del procedimiento se limita a $b = 3$ jueces independientes.

Sean $b = 3$ jueces que clasifican independientemente N objetos en c categorías disjuntas y desordenadas. La clasificación puede conceptualizarse como una tabla de contingencia $c \times c \times c$ con c filas, c columnas y c ¿cortes?. Sean n_{ijk} , R_i , C_j y S_k las frecuencias esperadas de las celdas y las frecuencias marginales de las filas, las columnas y los ¿cortes? para $i, j, k = 1, ..., c$. La frecuencia total viene dada por:

$$N = \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} n_{ijk} \ .$$

El estadístico de la prueba kappa no ponderada de Cohen para una tabla de contingencia de tres vías está dado por:

$$\kappa = 1 - \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} n_{ijk}}{\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} R_i C_j S_k} \qquad (2.1),$$

donde w_{ijk} son los "pesos" de desacuerdo asignados a cada celda para $i, j, k = 1, ., c$. Para el kappa no ponderado, los pesos de desacuerdo vienen dados por:

$$w_{ijk} = \begin{cases} 0 & i = j = k \\ 1 & c.c \end{cases} .$$

Dada una tabla de contingencia $c \times c \times c$ con N objetos clasificados de forma cruzada por $b = 3$ jueces independientes, una prueba de permutación exacta consiste en generar

todos los ordenamientos posibles, igualmente probables, de los N objetos en las c^3 celdas, preservando las distribuciones marginales de frecuencia. Para cada ordenamiento de las frecuencias de las celdas, se calcula el estadístico kappa no ponderado, κ, y el valor exacto de la probabilidad puntual hipergeométrica bajo la hipótesis nula, $p(n_{ijk}|R_i, C_j, S_k, N)$, donde

$$p(n_{ijk}|R_i, C_j, S_k, N) = \frac{(\prod_{i=1}^{c} R_i!)(\prod_{j=1}^{c} C_j!)(\prod_{k=1}^{c} S_k!)}{(N!)^{b-1} \prod_{i=1}^{c}\prod_{j=1}^{c}\prod_{k=1}^{c} n_{ijk}!} .$$

Si κ_o denota el valor del estadístico kappa no ponderado observado, el valor de probabilidad κ_o exacto bajo la hipótesis nula viene dado por:

$$P(\kappa_o) = \sum_{l=1}^{M} \Psi_l(n_{ijk}|R_i, C_j, S_k, N) ,$$

donde

$$\Psi_l(n_{ijk}|R_i, C_j, S_k, N) = \begin{cases} p(n_{ijk}|R_i, C_j, S_k) & si \quad \kappa \geq \kappa_o \\ 0 & c.c \end{cases}$$

y M denota el número total de posibles ordenamientos de frecuencia de celdas igualmente probables en el conjunto de referencia de todos los ordenamientos posibles de las frecuencias de las celdas, dado las distribuciones de frecuencias marginales observadas. Cuando M es muy grande, como es típico de contingencia multidireccional, las pruebas exactas no son prácticas y es necesario recurrir a los procedimientos de remuestreo de Monte Carlo. En estos casos, una muestra aleatoria de los M posibles ordenamientos igualmente probables de las frecuencias de las celdas proporciona una comparación de los estadísticos de la prueba κ calculados en L tablas aleatorias multidireccionales con el estadístico de la prueba κ calculado en la tabla de contingencia multidireccional observada.

Un algoritmo eficiente de remuestreo de Monte Carlo para generar ordenaciones aleatorias de frecuencias de celdas para tablas de contingencia multidireccionales con distribuciones de frecuencias marginales fijas fue desarrollado por Mielke, Berry y Johnston en 2007 [Mielke et al., 2007]. Para una tabla de contingencia de tres dimensiones con r filas, c columnas y s ¿cortes? el algoritmo de remuestreo se presenta en 12 sencillos pasos:

PASO 1. Construir una tabla de contingencia $r \times c \times s$ a partir de los datos observados.

PASO 2. Obtener los totales de frecuencia marginal fija $R_1, ..., R_r, C_1, ..., C_c, S_1, ...S_s$, y el total de frecuencias N. Establecer un contador de remuestreo $JL = 0$, y fijar L igual al número de muestras deseado.

PASO 3. Establecer el contador de remuestreo $JL = JL + 1$.

PASO 4. Establecer los contadores de frecuencia marginal $JR_i = R_i$ para $i = 1, ., r$; $JC_j = C_j$ para $j = 1, ..., c$; $JS_k = S_k$ para $k = 1, ...s$, y $M = N$.

PASO 5. Establecer $n_{ijk} = 0$ para $i = 1, ..., r, j = 1, ...c$, y $k = 1, ..., s$, y establecer los contadores de filas, columnas y ¿cortes? IR, IC e IS iguales a cero.

PASO 6. Crear las distribuciones de probabilidad acumulada PR_i , PC_j y PS_k a partir de los totales de frecuencia marginal ajustados JR_i , JC_j y JS_k para $i = 1, ., r, j = 1, ...c$, y $k = 1, ..., s$, donde $PR_1 = JR_1/M$ y $PR_i = PR_{i-1} + JR_i/M$ para $i = 1, ..., r$; $PC_1 = JC_1/M$ y $PC_j = PC_{j-1} + JC_j/M$ para $j = 1, ..., c$ y $PS_1 = JS_1/M$ y $PS_k = PS_{k-1} + JS_k/M$ para $k = 1, ..., s$.

PASO 7. Generar tres números pseudoaleatorios uniformes U_r , U_c y U_s sobre $[0, 1)$ y establecer los índices de fila, columna y ¿corte? $i = j = k = 1$, respectivamente.

PASO 8. Si $U_r \leq PR_i$, entonces $IR = i$, $JR_i = JR_i - 1$, e ir al PASO 9; en caso contrario $i = i + 1$ y se repite el PASO 8.

PASO 9. Si $U_c \leq PC_j$, entonces $IC = j$, $JC_j = JC_j - 1$, e ir al PASO 10; en caso contrario, $j = j + 1$ y se repite el PASO 9.

PASO 10. Si $U_s \leq PS_k$, entonces $IS = k$, $JS_k = JS_k - 1$, e ir al PASO 11; en caso contrario, $k = k + 1$ y se repite el PASO 10.

PASO 11. Establecer $M = M - 1$ y $n_{IR,IC,IS} = n_{IR,IC,IS} + 1$. Si $M > 0$, ir al PASO 4; en caso contrario, obtener el estadístico de prueba requerido.

PASO 12. Si $JL < L$, ir al PASO 3; en caso contrario, STOP.

Al terminar el procedimiento de remuestreo, se obtiene la κ de Cohen, como se indica en la ecuación (2.1), para cada una de las L tablas de contingencia aleatorias de tres dimensiones, dadas las distribuciones de frecuencia marginal fijas. Si κ_o denota el valor observado de κ, entonces bajo la hipótesis nula, el valor de la probabilidad aproximada del remuestreo para κ_o viene dado por:

$$P(\kappa_o) = \frac{1}{L} \sum_{l=1}^{L} \Psi_l(\kappa),$$

donde

$$\Psi_l(\kappa) = \begin{cases} 1 & si \quad \kappa \geq \kappa_o \\ 0 & c.c \end{cases} .$$

2.6. Test Q de McNemar para el cambio

En 1947, el psicólogo Quinn McNemar propuso una prueba de cambio derivada de la prueba t de pares emparejados para proporciones [McNemar, 1947]. Una aplicación típica es analizar respuestas binarias, codificadas con 0s y 1s, en $g = 2$ periodos de tiempo para cada uno de los $N \geq 2$ sujetos, como Éxito y Fracaso, Sí y No, De acuerdo y Desacuerdo, o A favor y en contra. Si las cuatro celdas se identifican como en la siguiente tabla:

Tabla 2.6: Notación de una tabla de clasificación cruzada 2x2 para el test Q de McNemar

Tiempo 1\Tiempo 2	**Favor**	**Contra**	**Total**
Favor	A	B	$A + B$
Contra	C	D	$C + D$
Total	$A + C$	$B + D$	N

entonces la prueba de McNemar para el cambio viene dada por:

$$Q = \frac{(B-C)^2}{B+C} \ ,$$

donde $N = A + B + C + D$ y B y C representan las dos celdas de cambio, es decir, de Favor a Contra y de Contra a Favor.

Alternativamente, la prueba Q de McNemar puede considerarse como un test de bondad de ajuste de una chi-cuadrado con dos categorías, donde las frecuencias observadas, O_1 y O_2, corresponden a las celdas B y C, respectivamente, y las frecuencias esperadas, E_1 y E_2, vienen dadas por $E_1 = E_2 = \frac{B+C}{2}$, es decir se espera que la mitad de los sujetos cambien en una dirección (por ejemplo, de Favor a Contra) y la otra mitad en la otra dirección (por ejemplo, de Contra a Favor), bajo la hipótesis nula de que no hay cambios del Tiempo 1 al Tiempo 2.

Sea

$$E = \frac{B+C}{2}$$

el valor esperado en el que, por azar, la mitad de los cambios son de Favor a Contra y la otra mitad son de Contra a Favor. Entonces, la bondad de ajuste de una chi-cuadrado para las dos categorías de cambio viene dada por:

$$\chi^2 = \frac{(B-E)^2}{E} + \frac{(C-E)^2}{E} = \frac{B^2}{E} + \frac{C^2}{E} + 2E - 2B - 2C \ .$$

Sustituyendo E por $(B+C)/2$ se obtiene:

$$\chi^2 = \frac{2B^2}{B+C} + \frac{2C^2}{B+C} + B + C - 2B - 2C = \frac{2B^2}{B+C} + \frac{2C^2}{B+C} - B - C =$$

$$= \frac{2B^2 + 2C^2 - B(B+C) - C(B+C)}{B+C} = \frac{B^2 - 2BC + C^2}{B+C} \Rightarrow$$

$$\Rightarrow \chi^2 = \frac{(B-C)^2}{B+C}$$

2.7. Test Q de Cochran para el cambio

La variable dicotómica desempeña un gran papel y tiene muchas aplicaciones en la investigación y la medición. Convencionalmente, se asigna un valor de uno a cada elemento de la prueba que un sujeto responde correctamente y un cero a cada respuesta incorrecta.

En 1950, William Cochran publicó un artículo sobre "La comparación de porcentajes en muestras emparejadas" [Cochran, 1950]. En este breve pero formativo artículo, Cochran describió una prueba de igualdad de proporciones emparejadas que ahora se utiliza ampliamente en la investigación educativa y psicológica. El emparejamiento puede basarse en las características de diferentes sujetos o en los mismos sujetos bajo diferentes condiciones.

El test Q de Cochran puede considerarse una extensión del test de McNemar a tres o más condiciones de tratamiento. Supongamos que se observa una muestra de $N \geq 2$ sujetos en una situación en la que en cada sujeto actúa individualmente bajo cada una de las $k \geq 1$ condiciones experimentales diferentes. La actuación se puntúa como un éxito (1) o como un fracaso (0). La pregunta de investigación evalúa si la proporción real de éxitos es constante a lo largo de los k períodos de tiempo.

El estadístico del test Q de Cochran para el análisis de k condiciones de tratamiento (columnas) y N sujetos (filas) viene dado por:

$$Q = \frac{(k-1)(k\sum_{j=1}^{k} C_j^2 - A^2)}{kA - B}$$

donde

$$C_j = \sum_{i=1}^{N} x_{ij}$$

es el número de 1s en la j-ésima de las k columnas,

$$R_i = \sum_{j=1}^{k} x_{ij}$$

es el número de 1s en la i-ésima de las N filas,

$$A = \sum_{i=1}^{N} R_i \ , \qquad B = \sum_{i=1}^{N} R_i^2 \ ,$$

y x_{ij} indica la entrada de la celda de 0 ó 1 asociada a la i-ésima de las N filas y la j-ésima de las k columnas. La hipótesis nula estipula que cada una de los

$$M = \prod_{i=1}^{N} \left(\begin{array}{c} k \\ R_i \end{array} \right)$$

ordenamientos distinguibles de 1s y 0s dentro de cada una de las N filas ocurren con igual probabilidad, dado que los valores de $R_1, \ldots, R_N$ son fijos.

2.8. Una medida sobre el Tamaño del Efecto para el test Q de Cochran

Las medidas del tamaño del efecto son cada vez más importantes a la hora de informar sobre los resultados de la investigación. Lamentablemente, no existen medidas del tamaño del efecto para muchas de las pruebas estadísticas más comunes. En esta sección, se presenta una medida del tamaño del efecto corregida por el azar para el test Q de Cochran para proporciones relacionadas.

Consideremos un enfoque alternativo al test Q de Cochran en el que se aplican g tratamientos de forma independiente a cada uno de los N individuos, con el resultado de cada aplicación del tratamiento registrado como 1 o 0, representando cualquier dicotomización adecuada de de los resultados del tratamiento, es decir, un diseño de bloques aleatorios en el que los individuos son los bloques y los resultados del tratamiento se registran como 1 o 0. Sean x_{ij} las medidas de respuesta 1 y 0 registradas para $i = 1, ...N$ y $j = 1, ..., g$. Entonces, el estadístico del test de Cochran puede definirse como

$$Q = \frac{g-1}{2\sum_{i=1}^{N} p_i(1-p_i)}[2(\sum_{i=1}^{N} p_i)(N - \sum_{i=1}^{N} p_i) - N(N-1)\delta] \ ,$$

donde

$$\delta = [g \begin{pmatrix} N \\ 2 \end{pmatrix}]^{-1} \sum_{k=1}^{g} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |x_{ik} - x_{jk}|$$

y

$$p_i = \frac{1}{g} \sum_{j=1}^{g} x_{ij} \qquad para \ \ i = 1, ..., N \ ,$$

es decir, la proporción de valores 1 para el i-ésimo individuo. Se observa que en esta representación la variación de Q depende totalmente de δ.

En 1979, Acock y Stavig [Acock and Stavig, 1979]propusieron un valor máximo para Q dado por:

$$Q_{max} = N(g-1)$$

El valor máximo de Q de Acock y Stavig fue empleado por Serlin, Carr y Marascuilo [Serlin et al., 1982] para proporcionar una medida del tamaño del efecto de la Q de Cochran dada por:

$$\hat{\eta}_Q^2 = \frac{Q}{Q_{max}} = \frac{Q}{N(g-1)} \ ,$$

que estandarizó la Q de Cochran por un valor máximo. Lamentablemente, el valor de $Q_{max} = N(g-1)$ defendido por Acock y Stavig sólo se alcanza cuando cada sujeto g-tupla es idéntico y hay al menos un 1 y un 0 en cada g-tupla. Así $\hat{\eta}_Q^2$ es una medida de "máxima corrección" del tamaño del efecto y $0 \leq \hat{\eta}_Q^2 \leq 1$ bajo estas singulares condiciones.

Supongamos que $0 < p_i < 1$ para $i = 1, ..., N$ ya que $p_i = 0$ y $p_i = 1$ no son informativos. Si p_i es constante para $i = 1, ...N$, entonces $Q_{max} = N(g-1)$. Sin embargo, para la gran mayoría de los casos en los que $p_i \neq p_j$ para $i \neq j$, $Q_{max} < N(g-1)$. Por lo tanto, el uso rutinario de establecer $Q_{max} = N(g-1)$ es problemático y conduce a resultados cuestionables.

También hay que señalar que $\hat{\eta}^2_Q$ es un miembro de la familia V de medidas de asociación nominal basada en el estadístico de prueba V^2 de Cramér dado por:

$$V^2 = \frac{\chi^2}{\chi^2_{max}} = \frac{\chi^2}{N[min(r-1, c-1)]} \ ,$$

donde r y c denotan el número de filas y columnas de una tabla de contingencia $r \times c$. Las dificultades para interpretar V^2 se extienden a las de $\hat{\eta}^2_Q$.

Wickens observó que la V^2 de Cramér carece de una interpretación intuitiva que no sea como un escalado de chi-cuadrado, lo que limita su utilidad [Wickens, 1989]. Asimismo, Costner observó que V^2 y otras medidas basadas en la chi-cuadrado de Pearson carecen de interpretación para valores distintos de 0 y 1, o el máximo, dadas las distribuciones de frecuencia marginal observadas [Costner, 1965]. Agresti y Finlay también señalaron que la V^2 de Cramér es muy difícil de interpretar y recomendaron otras medidas [Agresti and Finlay, 1997]. Blalock señaló que "todas las medidas basadas en el chi cuadrado son de alguna manera arbitrarias por naturaleza, y sus interpretaciones dejan mucho que desear... todas ellas dan mayor peso a las columnas o filas que tienen las marginales más pequeñas en lugar de que a las que tienen las marginales más grandes" [Blalock, 1958]. Ferguson discutió el problema de utilizar frecuencias marginales idealizadas [Ferguson, 1981], y Guilford señaló que medidas como la ϕ^2 de Pearson, la T^2 de Tschuprov y la V^2 de Cramér necesariamente subestiman la magnitud de la asociación presente [Guilford, 1950]. Dado que $\hat{\eta}^2_Q$ es simplemente un caso especial del V^2 de Cramér, presenta los mismos problemas de interpretación.

2.8.1. Una medida del Tamaño del Efecto corregida por el azar

Las medidas de tamaño del efecto corregidas por el azar tienen mucho que ver con las medidas corregidas por el máximo. Una medida del tamaño del efecto corregida por el azar es una medida de concordancia entre los N individuos sobre g tratamientos, corregida por el azar. Varios investigadores han defendido las medidas del tamaño del efecto corregidas por el azar, como Brennan y Prediger [Brennan and Prediger, 1981], Cincchetti, Showalter y Tyrer [Cicchetti et al., 1985], Conger [Conger, 1985] o Krippendorff [Krippendorff, 1970]. Una medida corregida por el azar es 0 en condiciones de azar, 1 cuando la concordancia entre los N sujetos es perfecta, y negativa en condiciones de desacuerdo. Algunas medidas bien conocidas corregidas por el azar son el coeficiente de concordancia entre evaluadores de Scott, la medida de concordancia u de Kendall y Babington Smith, los coeficientes no ponderados y ponderados de concordancia entre evaluadores de Cohen y la medida de la regla de pie de Spearman. Bajo ciertas condiciones, el coeficiente de correlación de orden de rango de Spearman es también una medida de concordancia corregida por el azar [Spearman, 1904], es decir, cuando las variables x e y toman valores de 1 a N sin valores empatados, o cuando la variable x incluye valores empatados y la variable y es una permutación de la variable x, entonces el coeficiente de correlación de orden de rango de Spearman es tanto una medida de correlación como una medida de concordancia corregida por el azar.

Sean x_{ij} las medidas de respuesta $(0, 1)$ para $i = 1, ...N$ bloques y $j = 1, ..., g$ tratamientos, entonces

$$\delta = [g \begin{pmatrix} N \\ 2 \end{pmatrix}]^{-1} \sum_{k=1}^{g} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |x_{ik} - x_{jk}|$$

Bajo la hipótesis nula de que la distribución de δ asigna igual probabilidad a cada una de las $M = (g!)^N$posibles asignaciones de las g medidas de respuesta dicotómicas a las g posiciones de tratamiento para cada uno de los N individuos, el valor medio de δ viene dado por

$$\mu_\delta = \frac{2}{N(N-1)}[(\sum_{i=1}^{N} p_i)(N - \sum_{i=1}^{N} p_i) - \sum_{i=1}^{N} p_i(1-p_i)] \ ,$$

donde

$$p_i = \frac{1}{g} \sum_{i=1}^{g} x_{ij} \qquad para \ \ i = 1, ..., N \ .$$

Entonces, una medida del tamaño del efecto corregida por el azar puede definirse como

$$\mathfrak{R} = 1 - \frac{\delta}{\mu_\delta} \ .$$

2.8.2. Ventajas de la medida $\mathfrak{R}$ del tamaño del efecto

Las medidas del tamaño del efecto corregidas por el azar, como $\mathfrak{R}$, poseen claras ventajas de interpretación sobre las medidas del tamaño del efecto corregidas por el máximo, como $\hat{\eta}_Q^2$. El problema con $\hat{\eta}_Q^2$ radica en la forma en que se maximiza $\hat{\eta}_Q^2$. El denominador de $\hat{\eta}_Q^2$, $Q_{max} = N(g-1)$, estandariza el valor observado de Q para el tamaño de la muestra (N) y el número de tratamientos (g). Desafortunadamente, $N(g-1)$ no estandariza Q para los datos en los que se basa, sino que estandariza Q en otro conjunto hipotético de datos no observados.

$\mathfrak{R}$ es la alternativa preferida a $\hat{\eta}_Q^2$ como medida del tamaño del efecto por dos razones:

- Puede lograr un tamaño del efecto de 1 para los datos observados, mientras que esto es a menudo imposible para $\hat{\eta}_Q^2$.
- Es una medida del tamaño del efecto corregida por el azar, lo que significa que es 0 en condiciones de azar, 1 cuando la concordancia entre los N sujetos es perfecta, y negativo en condiciones de desacuerdo.

Por lo tanto, posee una interpretación clara que se corresponde con el coeficiente de concordancia entre evaluadores de Cohen y otras medidas corregidas por el azar que son familiares para la mayoría de los investigadores. Por otro lado otro lado, $\hat{\eta}_Q^2$ no posee una interpretación significativa, excepto para los valores límite de $Q = 0$ y $Q = 1$.

2.9. Medida de asociación d_N^c de Leik y Gove

En 1971, Robert Leik y Walter Gove propusieron una nueva medida de asociación nominal basada en comparaciones por pares de las diferencias entre las observaciones

[Leik and Gove, 1971]. Insatisfechos con las medidas existentes de asociación nominal, Leik y Gove sugirieron una medida de asociación de reducción proporcional en el error que fuese corregida para la cantidad máxima real de asociación, dadas las distribuciones de frecuencias marginales observadas. La nueva medida se denominó d_N^c, donde d indicaba el índice, siguiendo otros índices como d_{yx} y d_{xy} de Somers; el subíndice N indicaba la relevancia de d para una variable dependiente nominal; y el superíndice c indicaba que la medida estaba corregida por las restricciones impuestas por las distribuciones de frecuencia marginal.

Al igual que d_N^c, muchas medidas de asociación para dos variables se han basado en comparaciones por pares de las diferencias entre las observaciones. Consideremos dos variables de nivel nominal que se han clasificado de forma cruzada en una tabla de contingencia $r \times c$, donde r y c denotan el número de filas y columnas, respectivamente. Sean $n_{i.}$, $n_{.j}$, y n_{ij} la frecuencia marginal de la fila i, la frecuencia marginal de la columna j y el número de objetos en la celda ij, respectivamente, para $i = 1, ., r$ y $j = 1, ...c$, y sea N el número total de objetos en la tabla de contingencia $r \times c$. Si y y x representan las variables de fila y columna, respectivamente, hay $N(N-1)/2$ pares de objetos en la tabla que pueden dividirse en cinco tipos de pares exhaustivos y mutuamente excluyentes:

- pares concordantes
- pares discordantes
- pares empatados en la variable y pero que difieren en la variable x
- pares empatados en la variable x pero que difieren en la variable y
- pares empatados en ambas variables x e y

Para una tabla de contingencia $r \times c$, los pares concordantes (pares de objetos que están clasificados en el mismo orden tanto en la variable x como en la variable y) vienen dados por:

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \Big(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \Big) \ ,$$

los pares discordantes (pares de objetos que se clasifican en un orden en la variable x y en el orden inverso en la variable y) vienen dados por:

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \Big(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \Big) \ ,$$

los pares de objetos empatados en la variable x pero que difieren en la variable y vienen dados por:

$$T_x = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \Big(\sum_{k=j+1}^{c} n_{ik} \Big) \ ,$$

los pares de objetos empatados en la variable y pero que difieren en la variable x vienen dados por:

$$T_y = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} (\sum_{k=i+1}^{r} n_{kj}) \ ,$$

y los pares de objetos empatados en la variable x y en la variable y vienen dados por:

$$T_{xy} = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}(n_{ij} - 1) \ .$$

Entonces,

$$C + D + T_x + T_y + T_{xy} = \frac{N(N-1)}{2} \ .$$

2.9.1. Tabla de Contingencia Observada

A partir de la tabla de contingencia observada se obtienen los parámetros C, D, T_x, T_y y T_{xy}.

2.9.2. Tabla de Contingencia Esperada

Usando la tabla de contingencia observada, se contruye la tabla de contingencia esperada de la siguiente forma:

$$E_{ij} = \frac{n_{i.} n_{.j}}{N} \qquad con \quad i = 1, ..., r \quad y \quad j = 1, .., c \ .$$

A partir de la tabla de contingencia esperada se obtienen los parámetros C', D', T'_x, T'_y y T'_{xy}.

Afortunadamente, existe una forma más eficiente de calcular C', D', T'_x, T'_y y T'_{xy} sin calcular primero los valores esperados. En primer lugar, dadas las distribuciones de frecuencias marginales de filas y columnas observadas se calcula el número de pares de valores esperados de frecuencia de celdas esperadas empatados a las variables x e y,

$$T'_{xy} = \frac{1}{2N^2} (\sum_{i=1}^{r} n_{i.}^2)(\sum_{j=1}^{c} n_{.j}^2) - \frac{N}{2}$$

A continuación, se calcula el número de pares de valores de frecuencia de celda esperados empatados a la variable y,

$$T'_y = \frac{1}{2} \sum_{i=1}^{r} n_{i.}^2 - \frac{N}{2} - T'_{xy}$$

Del mismo modo, se calcula el número de pares de valores de frecuencia de celda esperados empatados a la variable x,

$$T'_x = \frac{1}{2} \sum_{j=1}^{c} n_{.j}^2 - \frac{N}{2} - T'_{xy}$$

Por último, se calcula el número de pares concordantes y discordantes de valores de frecuencia de celdas esperados,

$$C' = D' = \frac{1}{2}[\frac{N(N-1)}{2} - T'_x - T'_y - T'_{xy}]$$

Hay que tener en cuenta que C', D', T'_x, T'_y y T'_{xy} se calculan sobre los totales de frecuencia marginal de la tabla de contingencia observada, que son invariantes bajo permutaciones.

2.9.3. Tabla de Contingencia Maximal

El estadístico de prueba d_N^c se basa en tres tablas de contingencia: la tabla de valores observados, la tabla de valores esperados y una tabla de valores máximos que se describirá a continuación.

Un algoritmo para generar un ordenamiento de las frecuencias de las celdas en una tabla de contingencia $r \times c$ que proporciona el valor máximo de un conjunto estadístico es:

PASO 1: Mantener las frecuencias marginales observadas de una tabla de contingencia $r \times c$ y eliminar los valores de frecuencias de celdas (n_{ij}).

PASO 2: Si algún par de frecuencias marginales, uno de cada conjunto de marginales, son iguales entre sí, introducir ese valor en la tabla como n_{ij} y restar el valor de los dos totales de frecuencia marginal. Repetir el PASO 2 hasta que no haya dos totales de frecuencia marginal iguales. Si todos los totales de frecuencia marginal se han reducido a cero, ir al PASO 5; de lo contrario, ir al PASO 3.

PASO 3: Observar la frecuencia marginal más grande que queda en cada conjunto e introducir el menor de los dos valores en n_{ij} . A continuación, restar ese valor (más pequeño) de los dos totales de frecuencias marginales.

PASO 4: Si todos los totales de frecuencia marginal se han reducido a cero, ir al PASO 5; En caso contrario, ir al PASO 2.

PASO 5: Establecer los valores n_{ij} restantes como 0, $i = 1, ..., r$ y $j = 1, ..., c$.

Denotamos la doble comilla (' ') como una suma de pares calculada sobre los valores de frecuencias de las celdas maximizadas. Entonces, el número de pares concordantes de valores de frecuencia de celda maximizada es

$$C'' = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} n_{ij}(\sum_{k=i+1}^{r}\sum_{l=j+1}^{c} n_{kl}) \ ,$$

el número de pares discordantes de valores de frecuencia celular maximizada es

$$D'' = \sum_{i=1}^{r-1}\sum_{j=1}^{c-1} n_{i,c-j+1}(\sum_{k=i+1}^{r}\sum_{l=1}^{c-j} n_{kl}) \ ,$$

el número de pares de valores de frecuencia de celda maximizados empatados a la variable x es

$$T_x'' = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} (\sum_{k=j+1}^{c} n_{ik}) \ ,$$

el número de pares de valores de frecuencia celular maximizados empatados a la variable y es

$$T_y'' = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} (\sum_{k=i+1}^{r} n_{kj}) \ ,$$

y el número de pares de valores de frecuencia celular maximizada empatados en ambas variables x e y es

$$T_{xy} = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}(n_{ij} - 1) \ .$$

Entonces,

$$C'' + D'' + T_x'' + T_y'' + T_{xy}'' = \frac{N(N-1)}{2} \ .$$

2.9.4. Cálculo de la d_N^c de Leik y Gove

Dados los valores observados, esperados y maximizados de C, D, T_x , T_y y T_{xy}, los errores del primer tipo (E_1) -la variación entre la independencia y la máxima asociación- vienen dados por:

$$E_1 = T_y' - T_y''$$

y los errores del segundo tipo (E_2) -la variación entre la tabla observada y la tabla de máxima asociación- vienen dados por:

$$E_2 = T_y - T_y''$$

Entonces, a la manera de las medidas de asociación de reducción proporcional en el error,

$$d_N^c = \frac{E_1 - E_2}{E_1} = \frac{(T_y' - T_y'') - (T_y - T_y'')}{T_y' - T_y''} = \frac{T_y' - T_y}{T_y' - T_y''}$$

Dado que d_N^c es una medida simétrica, el número de valores empatados en la variable x puede utilizarse en lugar del número de valores empatados en la variable y. Así pues,

$$d_N^c = \frac{T_x' - T_x}{T_x' - T_x''}$$

Alternativamente, d_N^c puede definirse en términos del número de valores empatados tanto en x como en y. Así,

$$d_N^c = \frac{T'_{xy} - T_{xy}}{T'_{xy} - T''_{xy}}$$

Como los datos son categóricos, C y D pueden considerarse agrupados. Por lo tanto,

$$d_N^c = \frac{(C' + D') - (C + D)}{(C' - D') - (C'' + D'')}$$

Como señalan Leik y Gove, para ayudar a interpretar la relación entre las variables x e y, sería preferible determinar explícitamente el número de pares perdidos por los requisitos marginales de la tabla de contingencia. La asociación puede entonces ser definida dentro de esos límites, permitiendo que el índice alcance la unidad si las frecuencias de las celdas están tan cerca de un patrón perfecto como lo permitan las distribuciones marginales. Así pues, la proporción de casos que se considera es

$$1 - \frac{2(T''_x + T''_y)}{N(N-1)}$$

2.9.5. Un test de permutación para d_N^c

Leik y Gove no proporcionaron un error estándar para el estadístico de prueba d_N^c [Leik and Gove, 1969]. Por otro lado, las pruebas de permutación no suponen ni requieren el conocimiento de los errores estándar. Consideremos la expresión

$$d_N^c = \frac{T'_y - T_y}{T'_y - T''_y}$$

Es evidente que T'_y y T''_y son invariantes bajo permutación. Por lo tanto la probabilidad de d_N^c bajo la hipótesis nula puede determinarse por la distribución discreta de permutación de T_y sólo, que se obtiene fácilmente de la tabla de contingencia observada. Los métodos estadísticos de permutación exacta son muy eficaces cuando sólo se calcula la parte variable del estadístico de prueba definido en cada uno de los M ordenamientos posibles de los datos observados; en este caso, T_y.

2.10. Un problema de ocupación de la matriz

En muchas situaciones de investigación, es necesario examinar una secuencia de observaciones sobre un pequeño grupo de sujetos, donde cada observación se clasifica de una de dos maneras. Supongamos, por ejemplo, que se registra un Éxito (1) o un Fracaso (0) para cada uno de $N \geq 2$ sujetos en cada una de $k \geq 2$ tareas. La prueba estándar en estos casos es el Test Q de Cochran para el cambio.

Sin embargo, cuando el número de sujetos es pequeño, por ejemplo, $2 \leq N \leq 6$, y el número de tratamientos es grande, por ejemplo, $20 \leq k \leq 400$, puede ser preferible una prueba alternativa al test Q de Cochran. Estas condiciones de investigación surgen por varias razones. En primer lugar, se propone un estudio de grupo a largo plazo, pero pocos sujetos están dispuestos a comprometerse con la investigación debido al tiempo prolongado de la misma. En segundo lugar, un estudio longitudinal comienza con un número adecuado

de sujetos, pero hay una alta tasa de abandono y no se puede justificar el análisis de supervivencia. En tercer lugar, muy pocos sujetos satisfacen el protocolo de investigación. Cuarto, el coste de cada observación/tratamiento es caro para el investigador. Quinto, los sujetos también generan mucho gasto. Sexto, se puede realizar un estudio piloto con un número reducido de sujetos para establecer la validez de la investigación antes de solicitar financiación para un estudio más amplio.

Consideremos una matriz de ocupación de $N \times k$ con N sujetos (filas) y k condiciones de tratamiento (columnas). Sea x_{ij} la observación del i-ésimo sujeto ($i = 1, \ldots, N$) en la j-ésima condición de tratamiento ($j = 1, \ldots, k$), donde un éxito se codifica como 1 y un fracaso como 0. Para cualquier sujeto, un éxito puede ser resultado del tratamiento administrado o puede ser resultado de alguna otra causa o de una respuesta aleatoria, es decir, un falso positivo. Por lo tanto, una respuesta exitosa al tratamiento se cuenta sólo cuando todos los N sujetos obtienen un éxito, es decir, una columna completa de valores 1. Evidentemente, este enfoque no se generaliza bien a un gran número de sujetos, ya que no es realista que un gran número de sujetos responda de forma conjunta. El test Q de Cochran es preferible cuando N es grande.

En 1965, Mielke y Siddiqui presentaron un procedimiento exacto de permutación para el problema de ocupación de la matriz en el "Journal of the American Statistical Association" que es apropiado para muestras pequeñas (N) y un gran número de tratamientos (k) [Mielke and Siddiqui, 1965]. Sean

$$R_i = \sum_{j=1}^{k} x_{ij}$$

para $i = 1, ...N$, los totales de los sujetos (filas), sea

$$M = \prod_{i=1}^{N} \begin{pmatrix} k \\ R_i \end{pmatrix}$$

el número de matrices de ocupación $N \times k$ igualmente distinguibles en el conjunto de referencia, bajo la hipótesis nula, y sea $\nu = min(R_1, ..., R_N)$. La hipótesis nula estipula que cada una de las M configuraciones de 1s y 0s dentro de cada una de las N filas ocurre con igual probabilidad, dado que los valores $R_1, ..., R_N$ son fijos. Si U_g es el número de configuraciones distintas en las que exactamente k condiciones de tratamiento (columnas) se llenan de aciertos (1s), entonces

$$U_\nu = \begin{pmatrix} k \\ \nu \end{pmatrix} \prod_{i=1}^{N} \begin{pmatrix} k-\nu \\ R_i - \nu \end{pmatrix}$$

es el valor inicial de la relación recursiva

$$U_g = \begin{pmatrix} k \\ g \end{pmatrix} [\prod_{i=1}^{N} \begin{pmatrix} k-g \\ R_i - g \end{pmatrix} - \sum_{j=g+1}^{\nu} \begin{pmatrix} k-g \\ j-g \end{pmatrix} \frac{U_j}{\begin{pmatrix} k \\ j \end{pmatrix}}] \,,$$

donde $0 \leq g \leq \nu - 1$. Si $g = 0$, entonces

$$M = \sum_{g=0}^{\nu} U_g$$

y la probabilidad exacta de observar s o más condiciones de tratamiento (columnas) completamente llenas de éxitos (1s) viene dada por:

$$P = \frac{1}{M} \sum_{g=s}^{\nu} U_g \ ,$$

donde $0 \leq s \leq v$.

2.11. Test exacto de Fisher

Aunque el test de probabilidad exacta de Fisher no es, estrictamente hablando, una medida de asociación entre dos variables de nivel nominal, ha adquirido tal importancia en el análisis de las tablas de contingencia 2×2, que excluir la prueba exacta de Fisher de su consideración sería una grave omisión. Dicho esto, sin embargo, la prueba de probabilidad exacta de Fisher proporciona la probabilidad de asociación más que una medida de la fuerza de la asociación. La prueba de probabilidad exacta de Fisher fue desarrollada de forma independiente por R.A. Fisher [Fisher, 1935], Frank Yates [Yates, 1934] y Joseph Irwin [Irwin, 1935] a principios de la década de 1930. En consecuencia, la prueba se denomina a menudo test exacto de Fisher-Yates o test de probabilidad exacta de Fisher-Irwin.

Aunque la prueba de probabilidad exacta de Fisher se diseñó originalmente para tablas de contingencia 2×2 y se utiliza casi exclusivamente para este fin, en esta sección la prueba se amplía para aplicarla a otras tablas de contingencia más grandes. Para facilitar el cálculo y evitar expresiones factoriales grandes, un procedimiento de recursión con un valor inicial arbitrario proporciona un método eficaz para obtener los valores exactos de la probabilidad.

2.11.1. Análisis exacto de Fisher con una tabla 2×2

Considere una tabla de contingencia 2×2 con N casos, donde x_o denota la frecuencia observada de cualquier celda y r y c representan las frecuencias marginales de fila y columna, respectivamente, correspondientes a x_o. La siguiente tabla ilustra la notación de una tabla de contingencia 2×2:

Tabla 2.7: Notación de una tabla de contingencia 2x2

	A_1	A_2	Total
B_1	x	$r - x$	r
B_2	$c - x$	$N - r - c + x$	$N - r$
Total	c	$N - c$	N

Si $H(x|r, c, N)$ es una función positiva definida recursivamente en la que

$$H(x|r,c,N) = D\times\begin{pmatrix} r \\ x \end{pmatrix}\begin{pmatrix} N-r \\ c-x \end{pmatrix}\begin{pmatrix} N \\ c \end{pmatrix}^{-1} = D\times\frac{r!c!(N-r)!(N-c)!}{N!x!(r-x)!(c-x)!(N-r-c+x)!}$$

donde $D > 0$ es una constante desconocida, entonces resolviendo la relación recursiva

$$H(x+1|r,c,N) = H(x|r,c,N) \times g(x)$$

se obtiene

$$g(x) = \frac{(r-x)(c-x)}{(x+1)(N-r-c+x+1)} \ .$$

El algoritmo puede emplearse entonces para enumerar todos los valores de $H(x|r,c,N)$, donde $a \leq x \leq b$, $a = max(0, r+c-N)$, $b = min(r,c)$, y $H(a|N,r,c)$ es inicialmente tomado como un valor positivo pequeño. El total sobre toda la distribución puede hallarse mediante:

$$T = \sum_{k=a}^{b} H(x|r,c,N) \ .$$

Para calcular el valor de la probabilidad de x_o, dadas las distribuciones de frecuencias marginales observadas, se debe determinar la probabilidad puntual de la tabla observada. Este valor, designado por $U_2 = H(x|r,c,N)$, se encuentra recursivamente. A continuación, hay que identificar la cola de la distribución de probabilidad asociada a U_2. Sea

$$U_1 = \begin{cases} H(x_o - 1|r,c,N) & si \ \ x_0 > a \\ 0 & si \ \ x_o = a \end{cases}$$

y

$$U_3 = \begin{cases} H(x_o + 1|r,c,N) & si \ \ x_0 < b \\ 0 & si \ \ x_o = b \end{cases} \ .$$

Si $U_1 > U_3$, U_2 se encuentra en la cola derecha de la distribución; en caso contrario, se define que U_2 está en la cola izquierda de la distribución, y los subtotales de una cola (S_1) y de dos colas (S_2) pueden hallarse mediante:

$$S_1(x_o|r,c,N) = \sum_{k=a}^{b} K_k H(k|r,c,N)$$

y

$$S_2(x_o|r,c,N) = \sum_{k=a}^{b} L_k H(k|r,c,N) \ ,$$

respectivamente, donde

$$K_k = \begin{cases} 1 & si \;\; U_1 \leq U_3 \;\; y \;\; k \leq x_o \;\; \text{ó} \;\; si \;\; U_1 > U_2 \;\; y \;\; k \geq x_o \\ 0 & c.c \end{cases}$$

y

$$L_k = \begin{cases} 1 & si \;\; H(k|r, c, N) \leq U_2 \\ 0 & c.c \end{cases}$$

para $k = a, ..., b$. Los valores de la probabilidad exacta de una y dos colas vienen dados entonces por:

$$P_1 = \frac{S_1}{T} \qquad y \qquad P_2 = \frac{S_2}{T} \;,$$

respectivamente.

2.11.2. Análisis exacto de Fisher con otras tablas de contigencia

Aunque el test exacto de Fisher se ha limitado en gran medida al análisis de tablas de contingencia 2×2, no es difícil extender el test exacto de Fisher a tablas de contingencia más grandes, aunque tales extensiones pueden generar un gran esfuerzo computacional. Consideremos un ejemplo de tabla de contingencia 2×3 con N casos, donde x_o denota la frecuencia observada de la celda en la primera fila y la primera columna, y_o denota la frecuencia observada de la celda en la segunda fila y la primera columna, y r_1, r_2 y c_1 son las frecuencias marginales observadas en la primera fila, la segunda fila y la primera columna, respectivamente. Si $H(x, y)$, dados N, r_1,r_2 y c_1, es una función positiva definida recursivamente, la resolviendo de la relación recursiva

$$H(x, y + 1) = H(x, y) \times g_1(x, y)$$

se obtiene

$$g_1(x, y) = \frac{(c_1 - x - y)(r_2 - y)}{(1 + y)(N - r_1 - r_2 - c_1 + 1 + x + y)} \;. \tag{2.2}$$

Si $y = min(r_2, c_1 - x)$, entonces $H(x + 1, y) = H(x, y) \times g_2(x, y)$ donde

$$g_2(x, y) = \frac{(c_1 - x - y)(r_1 - x)}{(1 + x)(N - r_1 - r_2 - c_1 + 1 + x + y)} \;, \tag{2.3}$$

dado que $max(0, r_1 + r_2 + c_1 - N - x) = 0$. Sin embargo, si $y = min(r_2, c_1 - x)$ y $max(0, r_1 + r_2 + c_1 - N - x) > 0$, entonces $H(x + 1, y - 1) = H(x, y) \times g_3(x, y)$, donde

$$g_3(x, y) = \frac{y(r_1 - x)}{(1 + x)(r_2 + 1 - y)} \;. \tag{2.4}$$

Las tres expresiones recursivas dadas en las ecuaciones (2.2), (2.3) y (2.4) pueden emplearse para enumerar completamente la distribución de $H(x, y)$, donde $a \leq x \leq b$, $a = max(0, r_1+c_1-N)$, $b = min(r_1, c_1)$, $c(x) \leq y \leq d(x)$, $c(x) = max(0, r_1+r_2+c_1-N+x)$, $d(x) = min(r_2, c_1 - x)$ y $H[a, c(x)]$ se fija inicialmente en algún valor positivo pequeño. El total sobre la distribución completamente enumerada se puede calcular mediante:

$$T = \sum_{x=a}^{b} \sum_{y=c(x)}^{d(x)} H(x, y) \ .$$

Para calcular el valor de la probabilidad de (x_o, y_o), dadas las distribuciones marginales de frecuencia observadas, debe ser obtenido el valor de la probabilidad puntual hipergeométrica de la tabla de contingencia observada 2×3; este valor también puede hallarse recursivamente. A continuación, hay que hallar la probabilidad de un resultado tan o más extremo. El subtotal viene dado por:

$$S = \sum_{x=a}^{b} \sum_{y=c(x)}^{d(x)} J_{x,y} H(x, y) \ ,$$

donde

$$J_{x,y} = \begin{cases} 1 & si \quad H(x, y) \leq H(x_o, y_o) \\ 0 & c.c \end{cases}$$

para $x = a, ..., b$ e $y = c(x), ..., d(x)$. El valor exacto de probabilidad para la independencia asociada a las frecuencias de celdas observadas, x_o e y_o, viene dado por $P = \frac{S}{T}$.

2.11.3. Análisis de tablas 2×2×2

La prueba de probabilidad exacta de Fisher no se limita a las tablas de contingencia de dos dimensiones. Considere una tabla de contingencia de $2 \times 2 \times 2$, como la representada en la siguiente figura:

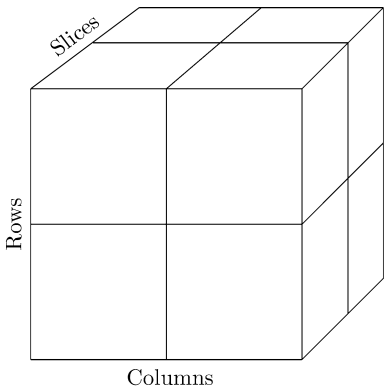

Figura 2.1: Representación gráfica de una tabla de contingencia 2x2x2

donde n_{ijk} denota la frecuencia de la celda de la i-ésima fila, j-ésima columna y k-ésimo ¿corte? para $i, j, k = 1, 2$. Sea $n_{i..}$ la frecuencia marginal de la i-ésima fila, $i = 1, ..., r$; $n_{.j.}$ la frecuencia marginal de la j-ésima columna, $j = 1, ..., c$, y $n_{..k}$ la frecuencia marginal del ¿corte?-ésimo, $k = 1, ...s$. Por lo tanto, $A = n_{1..}$, $B = n_{.1.}$, $C = n_{..1}$, y $N = n_{...}$ denotan los totales de frecuencia marginal observados de la primera fila, la primera columna, el primer ¿corte? y toda la tabla, respectivamente, de manera que $1 \leq A \leq B \leq C \leq N/2$. Además, sea $w = n_{111}$, $x = n_{112}$, $y = n_{121}$ y $z = n_{211}$ las frecuencias de las celdas de la tabla de contingencia $2 \times 2 \times 2$. Entonces, la probabilidad para cualquier w, x, y, y z está dada por:

$$P(w, x, y, z|A, B, C, N) = [A!(N-A)!B!(N-B)!C!(N-C)!]\times$$

$$\times[(N!)^2 w!x!y!z!(A-w-x-y)!(B-w-x-z)!(C-w-y-z)!(N-A-B-C+2w+x+y+z)!]^{-1}$$

[Mielke et al., 1994]. Un algoritmo para calcular la probabilidad exacta de Fisher implica una estructura de bucle anidado y requiere dos etapas. En la primera se obtiene la probabilidad exacta, U, de la tabla de contingencia $2 \times 2 \times 2$ observada. En la segunda etapa se obtiene el valor de la probabilidad exacta de todas las tablas con valores de probabilidad puntual hipergeométrica iguales o inferiores a la probabilidad puntual de la tabla de contingencia observada. Los cuatro bucles anidados dentro de cada etapa son sobre los índices de frecuencia de las celdas w, x, y y z, respectivamente. Los límites de w, x, y y z son

$$0 \leq w \leq M_w, \quad 0 \leq x \leq M_x, \quad 0 \leq y \leq M_y \quad y \quad L_x \leq z \leq M_z,$$

respectivamente, donde $M_w = A$, $M_x = A - w$, $M_y = A - w - x$, $M_z = min(B - w - x, C - w - y)$, y $L_z = max(0, A + B + C - N - 2w - x - y)$.

El método de recursión puede ilustrarse con el cuarto bucle (interno) sobre z, dado w, x, y, A, B, C, y N porque el bucle interno produce tanto U en la primera etapa como el valor exacto de la probabilidad en la segunda etapa. Sea $H(w, x, y, z)$ una función recursiva positiva definida dados A, B, C y N, que satisface

$$H(w, x, y, z+1) = H(w, x, y, z) \times g(w, x, y, z) \ ,$$

donde

$$g(w, x, y, z) = \frac{(B - w - x - z)(C - w - z)}{(z+1)(N - A - B - C + 2w + x + y + z + 1)} \ .$$

Los tres bucles restantes de cada etapa inicializan $H(w, x, y, z)$. Sea $I_x = max(0, A + B + C - N)$ y se establece como valor inicial de $H(0, 0, 0, I_z)$ a una constante positiva arbitraria pequeña. Entonces, el total sobre la distribución completamente enumerada sería:

$$T = \sum_{w=0}^{M_w} \sum_{x=0}^{M_x} \sum_{y=0}^{M_y} \sum_{z=L_x}^{M_x} H(w, x, y, z) \ .$$

Si w_o, x_o, y_o y z_o son los valores de w, x, y y z en la tabla de contingencia observada $2 \times 2 \times 2$, entonces U y el valor exacto de la probabilidad (P) vienen dados por:

$$U = H(w_o, x_o, y_o, z_o)/T$$

y

$$P = \sum_{w=0}^{M_w} \sum_{x=0}^{M_x} \sum_{y=0}^{M_y} \sum_{z=L_x}^{M_x} H(w, x, y, z)\psi(w, x, y, z)/T \ ,$$

respectivamente, donde

$$\psi(w, x, y, z) = \begin{cases} 1 & si \ \ H(w, x, y, z) \leq H(w_o, x_o, y_o, z_o) \\ 0 & si \ \ c.c. \end{cases} \ .$$

Capítulo 3

Medidas para Variables Ordinales (Parte 1)

Las medidas de las relaciones entre dos variables de nivel ordinal suelen ser más informativas que las medidas entre simples variables de nivel nominal (categóricas), ya que las categorías disjuntas y ordenadas suelen contener más información que las categorías disjuntas y desordenadas. Las medidas de asociación para dos variables de nivel ordinal suelen ser de dos tipos: las que se basan en las diferencias entre pares, como las medidas τ_a y τ_b de Kendall y la medida γ de Goodman y Kruskal, y las que se basan otros criterios, como la medida kappa ponderada de Cohen de concordancia entre evaluadores y el análisis ridit de Bross.

En este capítulo se aplican los métodos estadísticos de permutación a una variedad de medidas de asociación diseñadas para variables de nivel ordinal que se basan en las comparaciones por pares. Se incluyen las medidas de asociación ordinal τ_a y τ_b de Kendall, τ_c de Stuart, las medidas asimétricas d_{yx} y d_{xy} de Somers, las medidas $d_{y.x}$ y $d_{x.y}$ de Kim, la medida e de Wilson o el coeficiente de correlación rango-biserial de Cureton.

3.1. Principales medidas de Asociación Ordinal por pares

Una serie de medidas de asociación para dos variables de nivel ordinal se basan en comparaciones por pares de las diferencias entre los rangos. El estadístico de prueba S, definido por Maurice Kendall en 1938 [Kendall, 1938], desempeña un papel importante; este estadístico se expresa a menudo como $S = C - D$, donde C y D indican el número de pares concordantes y discordantes, respectivamente. Sean dos variables ordinales clasificadas de forma cruzada en una tabla de contingencia $r \times c$, donde r y c denotan el número de filas y columnas, respectivamente. Sean $n_{i.}$, $n_{.j}$, y n_{ij} los totales de frecuencia marginal de las filas, los totales de las frecuencias marginales de las columnas y el número de indivisuos en la celda ij, respectivamente, para $i = 1, ..., r$ y $j = 1, ...c$, y sea N el número total de individuos en la tabla de contingencia $r \times c$, es decir

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \ , \qquad n_{.j} = \sum_{i=1}^{r} n_{ij} \qquad y \qquad N = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \ .$$

La siguiente tabla muestra una notación convencional para una tabla de contingencia $r \times c$ para dos variables categóricas, X_i para $i = 1, ..., r$ e Y_j para $j = 1, ..., c$:

Tabla 3.1: Notación para la clasificación cruzada de dos variables categóricas, X e Y.

X\Y	Y_1	Y_2	...	Y_c	**Total**
X_1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
X_2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	N

Si X e Y representan las variables de fila y columna, respectivamente, hay $N(N-1)/2$ pares de individuos en la tabla que pueden dividirse en cinco tipos de pares mutuamente exhaustivos y exclusivos: pares concordantes, pares discordantes, pares empatados en la variable X pero no en la variable Y, pares empatados en la variable Y pero no en la variable X, y pares empatados en ambas variables.

Para una tabla de contingencia $r \times c$, los pares concordantes (pares de objetos que están clasificados en el mismo orden tanto en la variable X como en la variable Y) vienen dados por:

$$C = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \Big(\sum_{k=i+1}^{r} \sum_{l=j+1}^{c} n_{kl} \Big) \ ,$$

los pares discordantes (pares de objetos que se clasifican en un orden en la variable X y en el orden inverso en la variable Y) vienen dados por:

$$D = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{i,c-j+1} \Big(\sum_{k=i+1}^{r} \sum_{l=1}^{c-j} n_{kl} \Big) \ ,$$

los pares de objetos empatados en la variable X pero que difieren en la variable Y vienen dados por:

$$T_x = \sum_{i=1}^{r} \sum_{j=1}^{c-1} n_{ij} \Big(\sum_{k=j+1}^{c} n_{ik} \Big) \ ,$$

los pares de objetos empatados en la variable Y pero que difieren en la variable X vienen dados por:

$$T_y = \sum_{j=1}^{c} \sum_{i=1}^{r-1} n_{ij} \Big(\sum_{k=i+1}^{r} n_{kj} \Big) \ ,$$

y los pares de objetos empatados en la variable X y en la variable Y vienen dados por:

$$T_{xy} = \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}(n_{ij} - 1) \ .$$

Entonces,

$$C + D + T_x + T_y + T_{xy} = \frac{N(N-1)}{2} \ .$$

Dados C, D, T_x, T_y y N, se suelen definir seis medidas de asociación ordinal, cada una de las cuales tiene el mismo numerador, $S = C - D$, pero diferentes denominadores[1][2].

3.1.1. Medida τ_a de Kendall de Asociación Ordinal

La primera de estas medidas de asociación por pares fue la τ_a de Kendall [Kendall, 1948], que es una medida simétrica de asociación ordinal y se diseñó originalmente para medir la asociación entre dos conjuntos de valores de rangos no empatados, donde los dos conjuntos de puntuaciones de rango se etiquetan habitualmente como X e Y, aunque los rangos también pueden representarse en una tabla de contingencia $r \times c$ donde $n_{i.} = n_{.j} = 1$ para $i = 1, ..., r$ y $j = 1, ..., c$.

Se define simplemente como la diferencia entre las proporciones de pares concordantes y discordantes, está dada por

$$\tau_a = \frac{C}{\frac{N(N-1)}{2}} - \frac{D}{\frac{N(N-1)}{2}} = \frac{C-D}{\frac{N(N-1)}{2}} = \frac{2S}{N(N-1)} \ .$$

La τ_a de Kendall se presenta a veces como una alternativa al coeficiente de correlación de rango de Spearman [Kraft and van Eeden, 1968].

3.1.2. Medida τ_b de Kendall de Asociación Ordinal

Cuando existen valores empatados, la medida de asociación ordinal τ_a de Kendall no es la mejor opción, ya que ignora los dos conjuntos de valores empatados, T_x y T_y. Por esta razón, Kendall desarrolló τ_b [Kendall, 1948], una alternativa a τ_a, dada por

$$\tau_b = \frac{S}{\sqrt{(C + D + T_x)(C + D + T_y)}} \ .$$

La τ_b de Kendall es una medida estrictamente monótona de asociación ordinal, es decir, para cada aumento de categoría en la variable X, se espera que haya un aumento de categoría en la variable Y. Por consiguiente, τ_b sólo puede alcanzar límites de ± 1 para tablas de contingencia en las que $r = c$ y las distribuciones de frecuencias marginales de filas y columnas sean idénticas. Más concretamente, τ_b no puede alcanzar generalmente valores de ± 1 debido a la desigualdad de Cauchy:

El cuadrado de la suma de los productos de dos conjuntos será igual o menor que el producto de las sumas al cuadrado de dos conjuntos. Formalmente, para las variables X e Y,

[1] El número de pares empatados en ambas variables X e Y (T_{xy}) no se utiliza en ninguna de las seis medidas.

[2] En realidad hay más de seis medidas de asociación ordinal basadas en comparaciones por pares; aquí sólo se tratan las seis medidas más comunes.

$$(\sum_{i=1}^{N} x_i y_i))^2 \leq \sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2 \ .$$

En consecuencia, el numerador de τ_b será igual o menor que el denominador, permitiendo que τ_b alcance ± 1 sólo cuando todas las observaciones se concentren en una de las dos diagonales principales de la tabla de contingencia. Si ninguna frecuencia marginal es cero, esto significa que τ_b sólo puede alcanzar ± 1 para una tabla de contingencia cuadrada con distribuciones de frecuencias marginales idénticas. Es importante señalar que, dado que las categorías están ordenadas, las distribuciones de frecuencias marginales deben ser idénticas, no simplemente equivalentes. Así, distribuciones de frecuencias marginales para filas y columnas de, por ejemplo, $50, 30, 20$ y $50, 30, 20$, respectivamente, son idénticas, lo que proporciona la posibilidad que τ_b sea igual a $+1$, y distribuciones de frecuencias marginales para filas y columnas de $50, 30, 20$ y $20, 30, 50$, respectivamente, son idénticas, por lo que existe la posibilidad de que posibilidad de que τ_b sea igual a -1, pero las distribuciones de frecuencia marginal de filas y columnas de fila y columna de $50, 30, 20$ y $30, 20, 50$, respectivamente, son equivalentes pero no idénticas, y por lo tanto obligan a que la τ_b de Kendall sea menor que $+1$ o mayor que -1.

Por lo tanto, la τ_b de Kendall no es la medida más apropiada de asociación ordinal para la tabla de contingencia $r \times c$, con $r \neq c$.

3.1.3. Medida τ_c de Stuart de Asiciación Ordinal

Alan Stuart propuso la τ_c [Stuart, 1953], que modifica la τ_b de Kendall para las tablas de contingencia en las que $r \neq c$, la medida viene dada por

$$\tau_c = \frac{2mS}{N^2(m-1} \ ,$$

donde $m = min(r, c)$.

Stuart demostró que si N es un múltiplo de m y $r = c$ con distribuciones marginales de frecuencia idénticas, de modo que todas las observaciones caen en la diagonal de la tabla de contingencia y todas las frecuencias de las celdas son iguales, el valor máximo de la S de Kendall viene dado por

$$S_{max} = \frac{N^2(m-1)}{2m} \ .$$

Entonces, si $N = m$,

$$\frac{N^2(m-1)}{2m} = \frac{N^2(N-1)}{2N} = \frac{N(N-1)}{2} \ .$$

Sin embargo, si N no es un múltiplo de m, la expresión de S_{max} sigue siendo un límite superior que no se puede alcanzar. De ello se concluye que la τ_c de Stuart puede alcanzar a veces (y para N grande, puede alcanzar casi siempre) ± 1.

3.1.4. Medida γ de Goodman y Kruskal

En 1954 Goodman y Kruskal desarrollaron una nueva medida de asociación simétrica para dos variables de nivel ordinal que denominaron gamma, γ [Goodman and Kruskal, 1954]. Gamma es una medida de asociación ordinal de reducción proporcional al error que se basa únicamente en los pares no empatados, C y D, y viene dada por

$$\gamma = \frac{S}{C+D} = \frac{C-D}{C+D} = \frac{C}{C+D} - \frac{D}{C+D} \ .$$

Por tanto, de la expresión se deduce que γ es simplemente la diferencia entre las proporciones de los pares iguales y no iguales, ignorando todos los pares empatados, es decir, T_x, T_y y T_{xy}.

Hay un problema potencial con γ que fue reconocido inmediatamente por Goodman y Kruskal: Gamma es inestable en varios "puntos de corte", es decir, γ tiende a aumentar a medida que las categorías de una tabla de contingencia se colapsan porque no tiene en cuenta los pares empatados (y el número de pares empatados enumenta a medida que la tabla se colapsa). Además, γ es una medida de asociación ordinal (no estrictamente) monótona, es decir, para cada aumento (disminución) de la categoría ordenada en la variable x, la variable y aumenta (disminuye) o permanece igual.

3.1.5. Medidas d_{yx} y d_{xy} de Somers

En 1962 el sociólogo Robert Somers [Somers, 1962] se opuso a la medida simétrica de asociación ordinal de Goodman y Kruskal, γ, y propuso dos alternativas asimétricas dadas por

$$d_{yx} = \frac{C-D}{C+D+T_y} = \frac{S}{C+D+T_y} \ ,$$

donde T_y denota el número de pares empatados en la variable Y pero no empatados en la variable X, y

$$d_{xy} = \frac{C-D}{C+D+T_x} = \frac{S}{C+D+T_x} \ ,$$

donde T_x denota el número de pares empatados en la variable X pero no empatados en la variable Y.

A diferencia de las cuatro medidas simétricas, τ_a, τ_b, τ_c y γ, las medidas d_{yx} y d_{xy} de Somers dependen de qué variable, Y o X, se considera que es la dependiente. Observando las ecuaciones anteriores, se ve que Somers incluyó en los denominadores de d_{yx} y d_{xy} el número de valores empatados en la variable dependiente: T_y para d_{yx} y T_x para d_{xy}. La razón para incluir los valores empatados es simplemente que cuando la variable Y es la variable dependiente (d_{yx}), si dos valores de la variable independiente, X, difieren pero los dos valores correspondientes de la variable dependiente, Y, no difieren (están empatados), hay evidencia de una falta de asociación y los empates en la variable Y (T_y) deben ser incluidos en el denominador donde actúan para disminuir el valor de d_{yx}. El mismo razonamiento es válido para la medida d_{xy} de Somers.

Por último, es evidente que que la medida τ_b de Kendall de asociación ordinal es simplemente la media geométrica de las medidas d_{yx} y d_{xy} de Somers, dadas por

$$\tau_b = \sqrt{d_{yx} d_{xy}} \ .$$

3.2. Métodos estadísticos de permutación

Para el análisis de permutación exacto de una tabla de contingencia $r \times c$, es necesario calcular la medida de asociación ordinal seleccionada para las frecuencias de celdas observadas y enumerar exhaustivamente los M posibles ordenamientos igualmente probables de los N individuos en las rcceldas, dadas las distribuciones de frecuencias marginales observadas. Para cada ordenamiento en el conjunto de referencia de todas las permutaciones, se calculan una medida de asociación ordinal, digamos T , y el valor exacto de probabilidad hipergeométrica puntual bajo la hipótesis nula, $p(n_{ij}|n_{i.}, n_{.j}, N)$, donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^{r} n_{i.}!)(\prod_{j=1}^{c} n_{.j}!)}{N! \prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!} \ ,$$

n_{ij} es la frecuencia de celdas observada para $i = 1, ..., r$ y $j = 1, ..., c$, $n_{i.}$ es la i-ésimo frecuencia marginal, $n_{.j}$ es la j-ésima frecuencia marginal, y N es el total de todos los valores de n_{ij} para $i = 1, ..., r$ y $j = 1, ..., c$ [Burr, 1960]. Si T_o denota el valor del estadístico de prueba observado, los valores exactos de probabilidad (P) a una cola de T_o son las sumas de los valores $p(n_{ij}|n_{i.}, n_{.j}, N)$ asociados a los valores T calculados en los posibles ordenamientos igualmente probables de las frecuencias de las celdas que son iguales o mayores que T_o cuando T_o es positivo e iguales o menores que T_o cuando T_o es negativo. Así, el valor exacto de la probabilidad hipergeométrica de T_o cuando T es positivo viene dado por

$$P = \sum_{k=1}^{M} \Psi(T_k) p(n_{ij}|n_{i.}, n_{.j}, N) \ ,$$

donde

$$\Psi(T_k) = \begin{cases} 1 & T_k \geq T_o \\ 0 & c.c. \end{cases}$$

y el valor exacto de la probabilidad hipergeométrica de T_o cuando T es negativo viene dado por

$$P = \sum_{k=1}^{M} \Psi(T_k) p(n_{ij}|n_{i.}, n_{.j}, N) \ ,$$

donde

$$\Psi(T_k) = \begin{cases} 1 & T_k \leq T_o \\ 0 & c.c. \end{cases} .$$

Cuando el número de posibles ordenamientos de las frecuencias de las celdas es muy grande, las pruebas exactas son poco prácticas y se hacen necesarios los métodos de Monte Carlo. Los métodos estadísticos de permutación de Monte Carlo generan una muestra aleatoria de todos los posibles ordenamientos de las frecuencias de las celdas, extraídos con reemplazo, dadas las distribuciones de frecuencias marginales observadas. Los valores de probabilidad (de cola superior e inferior) del estadístico T son simplemente las proporciones de los valores T calculados en los ordenamientos de frecuencias de celdas seleccionadas aleatoriamente que son iguales o mayores que T_o cuando T_o es positivo e iguales o menores que T_o cuando T_o es negativo. Así, el valor de la probabilidad de remuestreo de Monte Carlo de T_o cuando T es positivo viene dado por

$$P(T \geq T_o|H_0) = \frac{número \;\; de \;\; veces \;\; que \;\; T \geq T_o}{L} ,$$

donde L denota el número de ordenamientos aleatorios de los datos observados.

3.3. Distribuciones de las Frecuencias Marginales

Sea C el número de pares concordantes, D el número de pares discordantes, T_x el número de pares empatados en la variable X pero no en la variable Y, T_y el número de pares empatados en la variable Y pero no en la variable X y T_{xy} denota el número de pares empatados tanto en variable X como en la variable Y. Entonces, el número total de pares puede dividirse como

$$\left(\begin{array}{c} N \\ 2 \end{array} \right) = \frac{N(N-1)}{2} = C + D + T_x + T_y + T_{xy} .$$

Obsérvese que

$$\frac{1}{2}(N^2 - \sum_{j=1}^{c} n_{.j}^2) = C + D + T_x$$

y

$$\frac{1}{2}[\sum_{j=1}^{c} n_{.j}(n_{.j} - 1)] = T_y + Y_{xy} ,$$

donde $n_{.j}$ indica la frecuencia marginal total de la j-ésima columna, $j = 1, ..., c$.

Entonces, todos los pares posibles se pueden dividir en términos de los totales de frecuencia marginal como

$$\left(\begin{array}{c} N \\ 2 \end{array} \right) = \frac{1}{2}(N^2 - \sum_{j=1}^{c} n_{.j}^2) + \frac{1}{2}[\sum_{j=1}^{c} n_{.j}(n_{.j} - 1)] =$$

$$= \frac{1}{2}[N^2 - \sum_{j=1}^{c} n_{.j}^2 + \sum_{j=1}^{c} n_{.j}(n_{.j} - 1)] = \frac{1}{2}(N^2 - \sum_{j=1}^{c} n_{.j}) = \frac{N(N-1)}{2} \ .$$

Mientras que la relación dada en la anterior ecuación es en términos de los totales de frecuencia marginal de columna, los mismos resultados pueden obtenerse de los totales de frecuencia marginal de fila, es decir

$$\begin{pmatrix} N \\ 2 \end{pmatrix} = \frac{1}{2}[N^2 - \sum_{i=1}^{c} n_{i.}^2 + \sum_{i=1}^{c} n_{i.}(n_{i.} - 1)] \ ,$$

donde $n_{i.}$ indica el total de frecuencia marginal de la fila i, $i = 1, ..., r$.

Por lo tanto, como las distribuciones de frecuencias marginales se fijan bajo permutación, los valores de probabilidad exactos de τ_a de Kendall, τ_b de Kendall, d_{yx} de Somers y d_{xy} de Somers se basan totalmente en la distribución de permutación del numerador común, S [Burr, 1960]. En el caso de la medida de asociación ordinal de Stuart, la fórmula para τ_c no incluye ni $C + D + T_x$ ni $C + D + T_y$, sino que utiliza $m = min(r, c)$, que se basa en el número de filas o columnas que se fijan bajo permutación. En consecuencia, el valor de la probabilidad de τ_c de Stuart también se basa únicamente en la distribución de permutación del estadístico S. En el caso de la medida de asociación ordinal de Goodman y Kruskal, γ, no considera que T_x ni T_y proporcionen ninguna información utilizable; por lo tanto, su valor de probabilidad difiere ligeramente del valor de valor de probabilidad común para τ_a y τ_b de Kendall, τ_c de Stuart y d_{yx} y d_{xy} de Somers.

3.4. Otras medidas de Asociación Ordinal por pares

3.4.1. Medidas $d_{y.x}$ y $d_{x.y}$ de Kim

En 1971, Jae-On Kim propuso medidas asimétricas proporcionales de reducción del error de asociación ordinal dadas por

$$d_{y.x} = \frac{C - D}{C + D + T_x} \qquad y \qquad d_{x.y} = \frac{C - D}{C + D + T_y}$$

[Kim, 1971]. A diferencia de las medidas d_{yx} y d_{xy} de Somers de asociación ordinal, que ajustan los empates en la variable dependiente; las medidas $d_{y.x}$ y $d_{x.y}$ de Kim ajustan los empates en la variable independiente. Es evidente que las medidas $d_{y.x}$ y $d_{x.y}$ de Kim son equivalentes a las medidas d_{xy} y d_{yx} de Somers, respectivamente.

3.4.2. Medida e de Asociación Ordinal de Wilson

En 1974, Thomas Wilson propuso otra medida de asociación ordinal que denominó e [Wilson, 1974]. Argumentando que una medida de asociación debería ajustarse para los valores empatados tanto en la variable X como en la variable Y, Wilson sugirió una medida simétrica de asociación ordinal dada por

$$e = \frac{C - D}{C + D + T_x + T_y} = \frac{S}{C + D + T_x + T_y} \ .$$

Como observó Wilson, e toma los valores de ± 1 si y sólo si los datos son estríctamente monótonos. Es obvio, a partir de la ecuación anterior, que la e de Wilson es equivalente a la d_{yx} de Somers cuando $T_x = 0$ y es equivalente a la d_{xy} de Somers cuando $T_y = 0$. Además, si $T_x = 0$ como $T_y = 0$, entonces $e = d_{yx} = d_{xy} = \gamma = \tau_a = \tau_b$.

3.4.3. Medida S de Whitfield para una Variable Binaria y una Variable Ordinal

En 1947 John Whitfield, un psicólogo experimental de la Universidad de Cambridge propuso una medida de correlación entre dos variables en la que una variable estaba compuesta por N rangos y la otra variable era dicotómica [Whitfield, 1947]. Un ejemplo de análisis servirá para ilustrar el procedimiento de Whitfield:

Considere las puntuaciones de rango que figuran en la siguiente tabla, donde las categorías de la variable dicotómica vienen dadas por "0" y "1" y las puntuaciones de rango van de 1 a 6:

Tabla 3.2: Ejemplo para la medida S de Whitfield

Ordinal	1	2	3	4	5	6
Binaria	0	1	0	0	0	1

Sea $n_0 = 4$ el número de puntuaciones de rango en la categoría "0", sea $n_1 = 2$ el número de puntuaciones de rango en la categoría "1" y sea $N = n_0 + n_1$.

Whitfield diseñó un procedimiento para calcular un estadístico que denominó S, siguiendo notación de Kendall en un artículo de *Biometrika* de 1945 sobre "The treatment of ties in ranking problems" [Kendall, 1945]. Teniendo en cuenta las $N = 6$ puntuaciones de la tabla, se consideran los $n_0 = 4$ rangos en la categoría identificada por "0": 1, 3, 4 y 5. Empezando por el rango 1 con la categoría "0", no hay puntuaciones de rango con la categoría "1" a la izquierda de "0" y hay dos puntuaciones de rango con la categoría "1" a la derecha de "0" (rangos 2 y 6), por lo que Whitfield calculó $0 - 2 = -2$. Para el rango 3 con la categoría "0", hay una puntuación de rango a la izquierda de "0" con la categoría "1" (rango 2) y una puntuación de rango a la derecha de "0" con la categoría "1" (rango 6); por tanto, $1 - 1 = 0$. Para el rango 4 con la categoría "0", hay una puntuación de rango a la izquierda de "0" con la categoría "1" (rango 2) y una puntuación de rango a la derecha de "0" = 4 con la categoría "1" (rango 6); por tanto, $1 - 1 = 0$. Por último, para el rango 5 con la categoría "0", hay una puntuación de rango a la izquierda de "0" con la categoría "1" (rango 2) y una puntuación de rango a la derecha de "0" con la categoría "1" (rango 6); por tanto, $1 - 1 = 0$. La suma de las diferencias entre las variables "0" y "1" es $S = -2 + 0 + 0 + 0 = -2$. De este modo, el enfoque de Whitfield se adapta a muestras con $n_0 = n_1$, así como a cualquier número de puntuaciones de rango empatadas.

Como el número de pares posibles de N enteros consecutivos viene dado por

$$\frac{N(N-1)}{2}$$

Whitfield definió y calculó la medida de asociación de rangos entre las variables "0" y "1" como

$$\tau = \frac{2S}{N(N-1)} \ .$$

La S de Whitfield es idéntica a la S de Kendall y está directamente relacionada con el estadístico U de suma de rangos de dos muestras de Mann y Whitney y con el estadístico W de suma de rangos de dos muestras de Wilcoxon. Las relaciones entre los estadísticos S de Whitfield y U de Mann y Whitney [Mann and Whitney, 1947] vienen dadas por

$$S = 2U - n_0 n_1 \qquad y \qquad U = \frac{S + n_0 n_1}{2}$$

y las relaciones entre la S de Whitfield y la W de Wilcoxon [Wilcoxon, 1945] vienen dadas por

$$S = n_1(N+1) - 2W \qquad y \qquad W = \frac{n_1(N+1) - S}{2} \ .$$

3.4.4. Coeficiente de correlación rango-biserial del Cureton

Sean dos variables correlacionadas, una representada por rangos y la otra por una dicotomía. En 1956, el psicólogo Edward Cureton propuso una nueva medida de correlación para una variable de rangos y una variable dicotómica denominada r_{rb} para la correlación rango-biserial [Cureton, 1956].

Cureton afirmó que el coeficiente de correlación debería tipificarse adecuadamente entre ± 1 y debería ser estrictamente no paramétrico, definido únicamente en términos de inversiones y concordancias entre pares de rangos, sin el uso de medias, varianzas, covarianzas o regresión. El coeficiente de correlación rango-biserial se define como:

$$r_{rb} = \frac{S}{S_{max}} \ ,$$

donde $S = C - D$ es el estadístico del test de Kendall [Kendall, 1938] y Whitfield [Whitfield, 1947], con C el número de pares concordantes y D el número de pares discordantes; y $S_{max} = n_0 n_1$, con n_0 el número de puntuaciones de rango en la categoría "0" y n_1 el número de puntuaciones de rango en la categoría "1".

En 1966, Glass [Glass, 1966] dedujo una fórmula simplificada para la r_{rb}, suponiendo que no hay puntuaciones de rango empatadas, dada por

$$r_{rb} = \frac{2}{N}(\bar{y}_1 - \bar{y}_0) \ ,$$

donde $\bar{y}_0$ y $\bar{y}_1$ son las medias aritméticas de los valores de la variable dicotómica codificados como "0" y "1", respectivamente.

Glass proporcionó dos fórmulas de cálculo alternativas dadas por

$$r_{rb} = \frac{2}{n_0}(\bar{y}_1 - \frac{N+1}{2}) \qquad \text{ó} \qquad r_{rb} = \frac{2}{n_1}(\frac{N+1}{2} - \bar{y}_0) \ .$$

3.4.4.1. Relaciones entre las diferentes medidas

A veces es interesante examinar las relaciones entre tests y medidas estadísticas aparentemente no relacionadas. Cureton propuso originalmente la r_{rb} como una medida del efecto del tamaño para la prueba de suma de rangos de dos muestras de Wilcoxon-Mann-Whitney, por tanto, se espera que la r_{rb} de Cureton y la prueba de Wilcoxon-Mann-Whitney estén relacionadas.

Además, dado que la medida rango-biserial de Cureton se basa en la S de Kendall, es de esperar que la r_{rb} de Cureton y la τ_a de Kendall estén relacionadas. Finalmente, en 2008 Roger Newson estableció la identidad entre el estadístico r_{rb} de Cureton y el estadístico d_{yx} de Somers [Newson, 2008].

Wilcoxon y Cureton

El test de Wilcoxon de suma de rangos de dos muestras, W, es simplemente la menor de las sumas de las puntuaciones de rango de las dos muestras, es decir,

$$W = min\{\sum_{i=1}^{n_0} rango_i, \sum_{j=1}^{n_1} rango_j\}$$

Cuando no hay valores de rango empatados, las relaciones entre W de Wilcoxon y r_{rb} de Cureton vienen dadas por

$$W = \frac{n_0(N+1) - n_0 n_1 r_{rb}}{2} \qquad y \qquad r_{rb} = \frac{n_0(N+1) - 2W}{n_0 n_1} \ ,$$

donde n_0 es el número de objetos del grupo con la menor de las dos sumas.

Mann-Whitney y Cureton

La prueba de dos muestras de Mann y Whitney de Mann y Whitney, U, es la suma del número de valores en un grupo, precedido por el número de valores del otro grupo. Alternativamente,

$$U = n_0 n_1 + \frac{n_0(n_0+1)}{2} - W$$

Cuando no hay valores de rango empatados, las relaciones entre la U de Mann y Whitney y la r_{rb} de Cureton vienen dadas por

$$U = \frac{n_0 n_1 (1 + r_{rb})}{2} \qquad y \qquad r_{rb} = \frac{2U}{n_0 n_1} - 1 \ .$$

Kendall y Cureton

El estadístico de la prueba τ_a de Kendall es

$$\tau_a = \frac{2S}{N(N-1)}$$

y relaciones entre la τ_a de Kendall y la r_{rb} de Cureton vienen dadas por

$$\tau_a = \frac{2n_0 n_1 r_{rb}}{N(N-1)} \qquad y \qquad r_{rb} = \frac{\tau_a N(N-1)}{2n_0 n_1} \ .$$

Capítulo 4

Medidas para Variables Ordinales (Parte 2)

Este capítulo continúa la descripción de las medidas de asociación para dos variables de nivel ordinal iniciada en el capítulo 3, pero se centra las medidas que se basan en criterios distintos a las comparaciones por pares entre las puntuaciones de los rangos, aunque es inevitable cierto solapamiento.

Se incluyen métodos estadísticos de permutación exactos y de Monte Carlo para el coeficiente de correlación de Spearman, la medida de concordancia de la regla del pie de Spearman, el coeficiente de concordancia de Kendall y Babington Smith, la medida kappa ponderada de Cohen para la concordancia, y el análisis ridit de Bross.

4.1. Coeficiente de correlación de rango de Spearman

Se consideran dos clasificaciones de N objetos consistentes en los N primeros enteros y sean X_i e Y_i para $i = 1, ..., N$ la primera y la segunda clasificación, respectivamente. La correlación de rangos no está exenta de críticas, ya que la derivación de la fórmula para esta correlación implica la suposición de que las diferencias entre los rangos sucesivos son iguales.

Una medida popular de la correlación entre las dos clasificaciones es el coeficiente de correlación de orden de rango de Spearman, dado por

$$\rho = 1 - \frac{\sum_{i=1}^{N} d_i^2}{\frac{N(N^2-1)}{6}} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \ , \qquad (4.1)$$

donde $d_i = X_i - Y_i$ para $i = 1, ..., N$. Charles Spearman desarrolló ρ en el primero de los dos artículos sobre la medición de la asociación y la correlación en 1904 y 1906 que aparecieron en el *American Journal of Psychology* [Spearman, 1904] y en el *British Journal of Psychology* [Spearman, 1906], respectivamente. Con dos conjuntos de puntuaciones de rango sin valores empatados, X_i e Y_i para $i = 1, ..., N$, se tiene que

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i = \frac{N(N+1)}{2}$$

y

$$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i^2 = \frac{N(N+1)(2N+1)}{6} \ ,$$

Spearman simplemente sustituyó en la fórmula de Pearson del coeficiente de correlación producto-momento dado por

$$r_{xy} = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}\sqrt{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}$$

y simplificó la ecuación, dando lugar a la Ecuación (4.1).

Obsérvese que el denominador del coeficiente de correlación de orden de rango de Spearman, $N(N^2-1)/6$, como se indica en la Ec. (4.1), representa la mitad del valor máximo de $\sum_{i=1}^{N} d_i^2$ cuando X_i e Y_i , $i = 1, ..., N$, son puntuaciones de rango sin valores empatados y las puntuaciones de rango de Y_i son la inversa exacta de las puntuaciones de rango de X_i, es decir, $Y_i = N - X_i + 1$ para $i = 1, ..., N$. Por lo tanto, la ρ de Spearman es una medida de correlación de rango corregida al máximo y toma valores entre ± 1, donde $+1$ indica una asociación positiva perfecta y -1 indica una asociación negativa perfecta.

Se demuestra fácilmente que el denominador de la ecuación (4.1), $N(N^2-1)/6$, es la mitad del valor máximo de $\sum_{i=1}^{N} d_i^2$ cuando X_i e Y_i para $i = 1, ..., N$ son ambas son puntuaciones de rango no empatadas y las puntuaciones de rango Y_i son la inversa de las puntuaciones de rango x_i. Para el valor máximo de $\sum_{i=1}^{N} d_i^2$, se define:

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (X_i - Y_i)^2 = \sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N} Y_i^2 - 2\sum_{i=1}^{N} X_i Y_i \ .$$

Ya que, para N valores de rango no empatados,

$$\sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N} Y_i^2 = \frac{N(N+1)(2N+1)}{6}$$

y, para $X_i = 1, ..., N$ e $Y_i = N - x_i + 1$, $i = 1, ..., N$,

$$\sum_{i=1}^{N} X_i Y_i = \frac{N(N+1)(N+2)}{6} \ ,$$

entonces, al sustituir en la Ec. (4.1) se obtiene:

$$\sum_{i=1}^{N} d_i^2 = \frac{2N(N+1)(2N+1)}{6} - \frac{2N(N+1)(N+2)}{6} =$$

$$= \frac{2N(N+1)(N-1)}{6} = \frac{N(N^2-1)}{3} \ ,$$

que es el doble del valor de $N(N^2-1)/6$.

Kendall y Babington Smith observaron que para interpretar la significación de un valor de ρ, es necesario considerar sólo la distribución de valores obtenida a partir de las clasificaciones observadas con todas las demás permutaciones posibles de los números enteros de 1 a N, y señalaron además que en la práctica suele ser más conveniente considerar sólo la distribución de $\sum_{i=1}^{N} d_i^2$ ya que $N(N^2-1)/6$ es invariable bajo permutación [Kendall et al., 1939].

4.2. Medida de concordancia de la regla del pie de Spearman

Los artículos de 1904 [Spearman, 1904] y 1906 [Spearman, 1906] de Charles Spearman contenían dos nuevas medidas de correlación de rango: el conocido coeficiente de correlación de rango de Spearman, ρ, y un segundo coeficiente de correlación menos conocido que Spearman denominó "la regla del pie". Se consideran dos clasificaciones de N objetos que consisten en los primeros N enteros y sean X_i e Y_i para $i = 1, ..., N$ la primera y la segunda clasificación, respectivamente. Entonces, la regla de Spearman viene dada por:

$$\mathcal{R} = 1 - \frac{\sum_{i=1}^{N} |X_i - Y_i|}{\frac{N^2-1}{3}} = \frac{3\sum_{i=1}^{N} |X_i - Y_i|}{N^2-1} \ . \qquad (4.2)$$

A diferencia del coeficiente de correlación de orden de rango de Spearman, el denominador del coeficiente de la regla del pie de Spearman, $\sim(N^2-1)/3$, como se da en la Ec. (4.2), no representa la mitad del valor máximo de $\sum_{i=1}^{N} |X_i - Y_i|$ cuando X_i e Y_i para $i = 1, ..., N$ son ambas puntuaciones de rango no empatadad y las puntuaciones de rango Y_i son la inversa exacta de las puntuaciones de rango X_i, es decir, $Y_i = N - X_i + 1$ para $i = 1, ..., N$. Por lo tanto, la $\mathcal{R}$ de Spearman no es una medida máximo-corregida de correlación de rangos y es, en cambio, una medida de concordancia corregida por el azar.

Se puede demostrar fácilmente que la $\mathcal{R}$ de Spearman es una medida de concordancia corregida por el azar y no es, de hecho, una medida convencional de correlación, lo que explica por qué $\mathcal{R}$ puede, en ocasiones, dar valores negativos y sólo puede alcanzar un valor de -1 cuando $N = 2$. Para demostrar que el valor esperado de $\sum_{i=1}^{N} |d_i|$ viene dado por $(N^2-1)/3$, se considera que:

$$\sum_{i=1}^{N} |d_i| = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} |i-j| = \frac{2}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (j-i) =$$

$$= \frac{1}{N} \sum_{i=1}^{N-1} [N(N+1) + i^2 - i(2N+1)] =$$

$$= \frac{N(N+1)}{6N} [6(N+1) + (2N-1) - 3(2N+1)] =$$

$$= \frac{N^2 - 1}{3} \ .$$

Por lo tanto, el coeficiente de la regla del pie de Spearman, dado por

$$\mathcal{R} = a - \frac{\sum_{i=1}^{N} |d_i|}{\frac{N^2-1}{3}}$$

es una medida de concordancia corregida por el azar cuando el valor esperado de $\sum_{i=1}^{N} |d_i|$ viene dado por $(N^2 - 1)/3$, ya que adopta la forma clásica de las medidas de concordancia corregidas por el azar dada por

$$concordancia = \frac{discordancia \ \ observada}{discordancia \ \ esperada}$$

[Krippendorff, 1970].

Existen tres limitaciones de la regla del pie de Spearman que contribuyen a su falta de uso en la investigación contemporánea [Stuart, 1977]: En primer lugar, a diferencia de otras medidas de correlación de rangos, $\mathcal{R}$ no norma adecuadamente entre los límites de ± 1; en segundo lugar, al igual que la ρ de Spearman, $\mathcal{R}$ se limita a los datos totalmente clasificados y no se adapta a las puntuaciones de rango empatadas; y en tercer lugar, debido a la suma de las diferencias absolutas entre las puntuaciones de rango, tradicionalmente ha sido algo engorroso establecer el valor de probabilidad de un valor observado de $\mathcal{R}$, especialmente cuando N es pequeño.

La $\mathcal{R}$ de Spearman alcanza un valor máximo de $+1$ cuando X_i es idéntica a Y_i para $i = 1, ..., N$ y no hay valores empatados. Sin embargo, si $Y_i = N - x_i + 1$ para $i = 1, ..., N$, entonces $\mathcal{R} = -0,5$ cuando N es impar y

$$\mathcal{R} = -0.5(1 + \frac{3}{N^2 - 1})$$

cuando N es par [Kendall, 1962]. En consecuencia, $\mathcal{R}$ no puede alcanzar un valor mínimo de -1, excepto cuando $N = 2$. Spearman, aparentemente sin saber que R era una medida corregida por el azar y reconociendo que los valores negativos de $\mathcal{R}$ no representaban una correlación inversa, sugirió ingenuamente que "es mejor tratar toda correlación como positiva" [Spearman, 1904]. Maurice Kendall señaló explícitamente esta aparente falta tipificación como un defecto de la regla del pie y sugirió una corrección dada por

$$\mathcal{R}' = 1 - \frac{\sum_{i=1}^{N} |X_i - Y_i|}{N^2} \ ,$$

que aseguraba un límite adecuado de $+1$ cuando las dos clasificaciones estaban en completa concordancia y -1 cuando las dos clasificaciones eran inversas entre sí [Kendall, 1962]. Sin embargo la corrección, aunque bien intencionada, destruyó por completo la interpretación de la regla del pie de Spearman.

4.2.1. Probabilidad de la regla del pie de Spearman

Cuando las variables X e Y consisten en su totalidad en puntuaciones de rango no empatadas de 1 a N y la variable Y es una permutación de las observaciones de rango en la variable X, existen métodos para determinar la probabilidad de una $\mathcal{R}$ observada bajo la hipótesis nula de que cualquiera de los $N!$ ordenamientos de los valores de X o Y es igualmente probable. Si

$$D = \sum_{i=1}^{N} |X_i - Y_i|$$

entonces, dado que $\mathcal{R}$ es simplemente una transformación lineal de D, la probabilidad de un valor observado de D es la probabilidad de un valor observado de $\mathcal{R}$. Las tablas de la función de distribución acumulativa exacta de D para $2 \leq N \leq 10$ y los valores de probabilidad aproximados basados en métodos deMonte Carlo para $11 \leq N \leq 15$ fueron publicados por Ury y Kleinecke en 1979 [Ury and Kleinecke, 1979]. En 1988 Franklin amplió el trabajo de Ury y Kleinecke, informando de la función de distribución acumulativa exacta de D para $11 \leq N \leq 18$, y discutió la tasa de convergencia a una distribución normal aproximada [Franklin, 1988]. En 1990 Salama y Quade utilizaron las propiedades de la cadena de Markov para obtener la exacta función de distribución acumulativa de D para $4 \leq N \leq 40$ y siguieron investigando aproximaciones a la distribución discreta de D [Salama and Quade, 1990]. Si la variable X o la variable Y contiene valores empatados, entonces el cálculo de un valor de probabilidad exacto es más complejo.

4.2.2. Rangos Múltiples

La regla del pie de Spearman, tal como se presentó originalmente en sus artículos de 1904 y 1906 en el *American Journal of Psychology* [Spearman, 1904] y el *British Journal of Psychology* [Spearman, 1906], respectivamente, se limitaba a $N \geq 2$ puntuaciones de rango no empatadas y $b = 2$ jueces. Sin embargo, como Berry y Mielke demostraron en 1998, la regla de Spearman puede generalizarse para incluir tanto puntuaciones de rango empatadas como no empatadas y $b \geq 2$ conjuntos de clasificaciones [Berry and Mielke, 1998]. Sea

$$\delta = [N \begin{pmatrix} b \\ 2 \end{pmatrix}]^{-1} \sum_{i=1}^{N} \sum_{r<s} |X_{ri} - X_{si}|$$

denota una función de distancia media basada en todas las $\begin{pmatrix} b \\ 2 \end{pmatrix}$ posibles diferencias absolutas emparejadas entre los valores de las clasificaciones por b jueces y sea

$$\mu_\delta = [N^2 \begin{pmatrix} b \\ 2 \end{pmatrix}]^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{r<s} |X_{ri} - X_{sj}|$$

denota el valor esperado de δ donde b es el número de jueces, N es el número de objetos, y $\sum_{r<s}$ es la suma sobre todos los r y s tales que $1 \leq r < s \leq N$. Entonces, la generalización de la medida de la regla del pie de Spearman viene dada por

$$\mathfrak{R} = 1 - \frac{\delta}{\mu_\delta} \ ,$$

donde $\mathfrak{R}$ es una medida de concordancia corregida por el azar entre los b jueces que no se limita a las puntuaciones de rango no empatadas. Nótese que en el caso de $b = 2$ jueces, la ecuación anterior se reduce a la regla del pie de Spearman de 1906 para $b = 2$ jueces.

4.3. Coeficiente de Concordancia

Mientras que el coeficiente de correlación de rango de Spearman y las medidas τ_a y τ_b de Kendall expresan el grado de asociación entre dos variables medidas en puntuaciones de rango o transformadas en ellas, el coeficiente de concordancia expresa el grado de asociación entre múltiples conjuntos de puntuaciones de rango.

En 1939, Maurice Kendall y Bernard Babington Smith publicaron un artículo en *The Annals of Mathematical Statistics* sobre "The problem of m rankings" en el que desarrollaron el conocido coeficiente de concordancia [Kendall and Smith, 1939]. Sean N y m el número de puntuaciones de rango y el número de jueces, respectivamente, entonces Kendall y Babington Smith definieron el coeficiente de concordancia como

$$W = \frac{12S}{m^2(N^3 - N)} \ ,$$

donde S es la suma observada de los cuadrados de las desviaciones de las sumas de los rangos respecto al valor medio $m(N + 1)/2$.

Dado que $m^2(N^3 - N)$ en el denominador de la ecuación anterior es invariable en todas las permutaciones de los datos observados, Kendall y Babington Smith demostraron que para para comprobar si un valor observado de W es estadísticamente significativo sólo es necesario considerar la distribución de S permutando los N rangos de todas las formas posibles e igualmente posibles. Si uno de los rangos es fijo, hay $(N!)^{m-1}$ valores posibles de S. Basándose en este procedimiento de permutación, Kendall y Babington Smith crearon cuatro tablas que proporcionan valores de probabilidad exactos para $N = 3$ y $m = 2, ..., 10$, $N = 4$ y $m = 2, ..., 6$, y $N = 5$ y $m = 3$.

W también puede definirse como

$$W = \frac{12\sum_{i=1}^{N} R_i^2 - 3m^2N(N + 1)}{m^2N(N^2 - 1)} \ ,$$

donde R_i para $i = 1, ..., N$ es la suma de las puntuaciones de rango para el i-ésimo de los N objetos y no hay puntuaciones de rango empatadas. También se sabe que W puede definirse como una función del valor medio de todos los coeficientes de correlación de orden de rango de Spearman por pares, dado por

$$\bar{\rho} = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij} \ .$$

A este respecto, Kendall y Babington Smith demostraron que $\bar{\rho}$ es simplemente el coeficiente intraclase, r_I, para los m conjuntos de clasificaciones. Las relaciones entre W y $\bar{\rho}$ vienen dadas por

$$\bar{\rho} = \frac{mW - 1}{m - 1} \qquad y \qquad W = \frac{\bar{\rho}(m-1) + 1}{m} \ .$$

Si todos los ordenamientos de los conjuntos de puntuaciones de rango observadas ocurren con igual probabilidad, el valor exacto de la probabilidad del valor observado de W calculado sobre M posibles ordenamientos igualmente probables de las puntuaciones de rango observadas bajo la hipótesis nula es

$$P(W \geq W_o | H_0) = \frac{número \ \ de \ \ valores \ \ W \geq W_o}{M} \ ,$$

donde W_o indica el valor observado de W.

4.3.1. Procedimientos relacionados

Desde hace tiempo se conoce que la estructura de datos para el coeficiente de concordancia de Kendall y Babington Smith [Kendall and Smith, 1939] es la misma que la del análisis de varianza de dos vías de Friedman para rangos [Friedman, 1937] y la misma que la de la relación de correlación de Wallis para datos de puntuación de rangos [Wallis, 1939]. Mientras que la prueba de Friedman, por ejemplo, proporciona un valor de probabilidad global de las diferencias globales entre rangos, existe un procedimiento relacionado que proporciona un valor de probabilidad exacto para la suma de rangos de un solo objeto, respondiendo a la pregunta: ¿cuándo el total de un solo objeto no se debe al azar bajo la hipótesis nula de asignación aleatoria?

Supongamos que cada uno de los N jueces asigna de forma independiente K rangos distintos no empatados a $K \geq 2$ objetos. Si S denota la suma de los N rangos para un objeto determinado bajo la hipótesis nula de que cada uno de los N jueces asigna los K rangos a los K objetos al azar, es decir, que cada objeto ocurre con probabilidad $1/K$, entonces la probabilidad puntual exacta de S viene dada por

$$p_S = K^{-N} C_{S-N}$$

para $S = N, N+1, ..., NK$. Sea $m = S - N$, entonces

$$C_m = \sum_{j=0}^{v} (-1)^j \begin{pmatrix} N \\ j \end{pmatrix} \begin{pmatrix} m - jK + N - 1 \\ N - 1 \end{pmatrix}$$

y $v = min(N, m/K)$, es decir, el mayor número entero no negativo menor o igual que N o m/K. El valor exacto de la probabilidad unilateral de S viene dado por

$$P_1 = \sum_{j=N}^{w} p_j \ ,$$

donde $w = min(S, NK + N - S)$ y el valor exacto de la probabilidad bilateral de S es dado por

$$P_2 = min(2P_1, 1) \ ,$$

ya que la distribución de S es simétrica respecto a $N(K+1)/2$ bajo la hipótesis nula.

4.4. Medida de acuerdo u de Kendall

A veces, en lugar de pedir a un grupo de jueces que clasifiquen un conjunto de objetos, se les presenta una serie de pares de objetos y se les pide que indiquen su preferencia por uno de los dos objetos de cada par. Este procedimiento se denomina comparación por pares. Cuando se recogen datos mediante el método de las comparaciones por pares, es posible calcular el grado de concordancia entre los jueces. En 1940, Kendall y Babington Smith [Kendall and Smith, 1940] propusieron un coeficiente de concordancia para evaluar las comparaciones pareadas entre k jueces para N clasificaciones, dado por

$$u = \frac{2S}{\binom{k}{2}\binom{N}{2}} - 1$$

donde

$$S = \sum_{i=1}^{N}\sum_{j=1}^{N}\binom{a_{ij}}{2} = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^2 - k\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} + \binom{k}{2}\binom{N}{2} \ ,$$

a_{ij} es el número de veces que un objeto asociado a la fila i de una matriz de preferencias es preferido al objeto asociado a la fila j, y $a_{ij} \geq 2$ para $i, j = 1, ..., N$. El número máximo de concordancias, que se produce cuando $\binom{N}{2}$ celdas de la matriz de preferencias contienen k cada una, es $\binom{N}{2}\binom{k}{2}$ y, por tanto, sólo en el caso de concordancia completa, $u = +1$ [Kendall and Smith, 1940].

Mientras que el valor máximo de u es $+1$ cuando hay concordancia completa entre los k jueces, el número mínimo de concordancias se produce cuando cada celda de la matriz de preferencias contiene $k/2$ si k es par o $(k \pm 1)/2$ si k es impar. Así, cuando k es par, el valor mínimo de u es $-1/(k-1)$ y cuando k es impar, el valor mínimo de u es $-1/k$. Como el valor esperado de u es cero [@44] y los valores mínimos de u son $-1/(k-1)$ cuando k es par y $-1/k$ cuando k es impar, u es claramente una medida de concordancia corregida por el azar, aunque aparentemente esto no fue reconocido por Kendall y Babington Smith cuando desarrollaron u en 1940.

4.5. Medida Kappa de Cohen

En 1960, Jacob Cohen desarrolló el estadístico kappa, una medida corregida por el azar de la concordancia entre dos jueces para un conjunto de c categorías disjuntas y no ordenadas [Cohen, 1960]. En 1968, Cohen amplió el kappa para medir la concordancia entre dos jueces para un conjunto de c categorías ordenadas y disjuntas [Cohen, 1968]. El kappa original para c categorías disjuntas y no ordenadas se conoció como kappa "no ponderado", o κ, y el kappa para c categorías disjuntas y ordenadas se conoció como kappa "ponderado", o κ_w. Mientras que la kappa no ponderada no distinguía entre magnitudes de

desacuerdo, la kappa ponderada incorporaba la magnitud de cada desacuerdo y otorgaba un crédito parcial a las discrepancias cuando la concordancia no era completa [Maclure and Willett, 1987]. El enfoque habitual consiste en asignar pesos a cada par de desacuerdos (pesos mayores indican un mayor desacuerdo). El kappa no ponderado para c categorías disjuntas y no ordenadas se trata en el capítulo 2. A continuación se presenta el kappa ponderado para c categorías disjuntas y ordenadas:

La medición de la concordancia es un caso especial de medición de la asociación entre dos variables de nivel ordinal. Los índices de concordancia miden el grado en que un conjunto de medidas de respuesta son idénticas a otro conjunto, es decir, concuerdan, en lugar del grado en que un conjunto de medidas de respuesta es una función lineal de otro conjunto de medidas de respuesta, es decir, correlacionan. Al igual que la regla de Spearman, la medida kappa ponderada de Cohen de la concordancia es una medida corregida por el azar, que refleja la cantidad de concordancia por encima de lo que cabría esperar por azar. Así, la kappa ponderada es igual a uno cuando se produce una concordancia perfecta, es igual a cero en caso de independencia y puede ser ligeramente negativo cuando la concordancia es inferior a la esperada por el azar [Fleiss et al., 2003].

Para simplificar, consideremos $N \geq 2$ objetos clasificados de forma cruzada por $b = 2$ jueces independientes en una tabla de contingencia $c \times c$ con c categorías disjuntas y ordenadas denotadas por $a_1, ..., a_c$, como en la siguiente tabla:

Tabla 4.1: Notación para una tabla de validación cruzada de N objetos por 2 jueces en c categorías disjuntas y ordenadas

Juez 1 \ Juez 2	a_1	a_2	...	a_c	Total
a_1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
a_2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_c	n_{c1}	n_{c2}	...	n_{cc}	$n_{c.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	N

Donde $n_{i.}$ es el total de la frecuencia marginal de la i-ésima fila, $i = 1, ..., c$, $n_{.j}$ es la frecuencia marginal total de la j-ésima columna, $j = 1, ..., c$, n_{ij} son las frecuencias de las celdas y w_{ij} denotan los pesos de las celdas. Cuando las c categorías para los $b = 2$ jueces están ordenadas de forma similar, entonces n_{ii} , $i = 1, \ldots, c$, y n_{ij} , $i \neq j$, denotan las frecuencias de las celdas de concordancia y discordancia, respectivamente.

Aunque se han propuesto diversos esquemas de ponderación para la kappa ponderada de Cohen, el más popular es la ponderación cuadrática dada por $w_{ij} = (i - j)^2$ para $i, j = 1, ..., c$, donde las ponderaciones de discordancia de la categoría progresan geométricamente hacia fuera desde la diagonal de concordancia, es decir, 0^2, 1^2, 2^2, 3^2, etc. Sin embargo, la ponderación lineal en la que $w_{ij} = |i - j|$ para $i, j = 1, ..., c$, donde las ponderaciones de la categoría de desacuerdo progresan linealmente hacia fuera desde la diagonal de concordancia, es decir, 0, 1, 2, 3, y así sucesivamente.

Una fórmula sencilla para el cálculo del estadístico de la prueba kappa ponderada de Cohen con $b = 2$ jueces viene dada por

$$\kappa_w = 1 - \frac{\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}n_{ij}}{\frac{1}{N^2}\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}n_{i.}n_{.j}} \ .$$

Dada una tabla de concordancia $c \times c$ con N objetos clasificados por las valoraciones de dos jueces independientes en c categorías ordenadas y disjuntas, una prueba de permutación exacta genera un conjunto de referencia de todos los M posibles ordenamientos igualmente probables de los N objetos en las c^2 celdas, preservando el número total de objetos en cada categoría, es decir, las distribuciones de frecuencia marginal. Para cada ordenamiento de frecuencias de celdas con distribuciones de frecuencias marginales fijas, se calculan el estadístico kappa ponderado, κ_w, y el valor de probabilidad exacto, $p(n_{ij}|n_{i.}, n_{.j}, N)$, donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^{c} n_{i.}!)(\prod_{j=1}^{c} n_{.j}!)}{N! \prod_{j=1}^{c}\prod_{j=1}^{c} n_{ij}!}$$

es el valor de probabilidad hipergeométrico convencional de una tabla de contingencia $c \times c$.

Sea κ_o el valor del estadístico kappa ponderado observado y M el número total de ordenamientos distintos de la frecuencia de las celdas de los N objetos en la tabla de clasificación $c \times c$, dados los totales de frecuencia marginales fijos. Entonces el valor exacto de la probabilidad de κ_o bajo la hipótesis nula viene dado por

$$P(\kappa_o|H_0) = \sum_{k=1}^{M} \Psi(\kappa_k) p(n_{ij}|n_{i.}, n_{.j}, N) \ ,$$

donde

$$\Psi(\kappa_k) = \begin{cases} 1 & \kappa_k \geq \kappa_o \\ 0 & c.c. \end{cases} \ .$$

Cuando el conjunto de referencia de todos los M ordenamientos posibles es muy grande, los análisis exactos de permutación son poco prácticos y se hacen necesarias las aproximaciones de remuestreo de Monte Carlo. Sea L una muestra aleatoria de todos los M valores posibles de κ_w. Entonces, bajo la hipótesis nula, el valor de la probabilidad aproximada de remuestreo para el valor observado de κ_w, κ_o, viene dado por

$$P(\kappa_o) = \frac{1}{L}\sum_{l=1}^{L} \Psi_l(\kappa_w) \ .$$

4.5.1. Comparación de la ponderación lineal y cuadrática

Existe una considerable controversia sobre qué pesos deben utilizarse con el estadístico kappa ponderado de Cohen, κ_w. La elección de los pesos es completamente arbitraria y se puede utilizar cualquier peso de celda de desacuerdo. La ponderación lineal es quizás la

más útil, porque como las ponderaciones de las celdas de desacuerdo progresan linealmente hacia fuera desde la diagonal de concordancia, su interpretación es más facil.

Es evidente que la ponderación lineal y la ponderación cuadrática producen los mismos resultados para las tablas de contingencia 2×2. También está muy claro que los valores kappa ponderados linealmente y los ponderados cuadráticamente suelen diferir muy poco para las tablas de contingencia 3×3. La ponderación lineal y cuadrática generalmente produce mayores diferencias con tablas de contingencia más grandes. Brenner y Kliebsch demostraron que la forma lineal del coeficiente kappa ponderado es menos sensible al número de categorías que la forma cuadrática; en consecuencia, recomendaron usar la forma lineal siempre que el número de categorías de la escala ordinal sea grande [Brenner and Kliebsch, 1996].

4.5.2. Kappa ponderado con múltiples jueces

Aunque la kappa ponderada de Cohen fue diseñada originalmente para $b = 2$ jueces independientes y se limita a ellos, la kappa ponderada puede generalizarse y ampliarse para medir la concordancia entre múltiples jueces. La generalización de la kappa de Cohen a múltiples jueces ha sido controvertida durante mucho tiempo, con muchos pasos en falso y callejones sin salida en el camino. En 1988 Berry y Mielke generalizaron la medida de concordancia kappa de Cohen para dar cabida a múltiples jueces [Berry and Mielke, 1988] y en 2008 Mielke, Berry y Johnston proporcionaron un eficiente algoritmo de remuestreo de Monte Carlo para analizar los datos de concordancia con múltiples jueces [Mielke et al., 2008].

En esta sección se presenta un procedimiento algorítmico para calcular el kappa ponderado y no ponderado con múltiples calificadores. Aunque el procedimiento es apropiado para cualquier número de $c \geq 2$ categorías disjuntas y ordenadas y $b \geq 2$ jueces, la descripción del procedimiento se limitan a $b = 3$ jueces independientes para simplificar la presentación, sin pérdida de generalidad.

Consideremos $b = 3$ jueces que clasifican independientemente N objetos en c categorías ordenadas y disjuntas. La clasificación puede conceptualizarse como una tabla de contingencia $c \times c \times c$ con c filas, c columnas y c cortes. Sean n_{ijk}, R_i, C_j y S_k las frecuencias de las celdas y los totales de las frecuencias marginales de las filas, columnas y cortes para $i, j, k = 1, ..., c$ y sea el total de frecuencias dado por

$$N = \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} n_{ijk} \ .$$

La estadística de la prueba kappa ponderada de Cohen para una tabla de contingencia de tres vías viene dada por

$$\kappa_w = \frac{N^2 \sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} n_{ijk}}{\sum_{i=1}^{c} \sum_{j=1}^{c} \sum_{k=1}^{c} w_{ijk} R_i C_j S_k} \ ,$$

donde w_{ijk} son los pesos de desacuerdo asignados a cada celda para $i, j, k = 1, ..., c$. Bajo la hipótesis nula de que los jueces clasifican los N objetos de forma independiente con totales de frecuencia marginal fijos, $E[\kappa_w] = 0$.

Como se ha comentado anteriormente, se han propuesto diversas funciones de ponderación para el kappa ponderado para dos jueces, donde las ponderaciones de celdas arbitrarias se denotan como w_{ij} e i y j designan las c categorías para cada juez, $i, j = 1, ..., c$. Normalmente, los pesos de las celdas se definen de forma que $w_{ii} = 0$ para $i = 1, ..., c$ y los pesos son simétricos, es decir, $w_{ij} = w_{ji}$ para $i, j = 1, ..., c$. Algunos ejemplos de sistemas de ponderación para dos jueces son la ponderación lineal, en la que $w_{ij} = |i - j|$, ponderación cuadrática donde $w_{ij} = (i - j)^2$, y kappa no ponderado donde

$$w_{ij} = \begin{cases} 0 & i = j \\ 1 & c.c. \end{cases} .$$

Para tres jueces, las ponderaciones de desacuerdo de las celdas vienen dadas por w_{ijk}, donde i, j, y k designan las c categorías de cada juez. De forma análoga a w_{ij}, w_{ijk} puede ser definida de forma que $w_{iii} = 0$ para $i = 1, ..., c$ y las ponderaciones son simétricas, es decir, $w_{ijk} = w_{ikj} = w_{jik} = w_{jki} = w_{kij} = w_{kji}$ para i, j, $k = 1, ..., c$. Ejemplos de sistemas de ponderación para tres jueces incluyen la ponderación lineal donde

$$w_{ijk} = |i - j| + |i - k| + |j - k|$$

y la ponderación cuadrática donde

$$w_{ijk} = (i - j)^2 + (i - k)^2 + (j - k)^2$$

para i, j, $k = 1, ..., c$.

El kappa ponderado para tres jueces se reduce al kappa no ponderado cuando

$$w_{ijk} = \begin{cases} 0 & i = j = k \\ 1 & c.c. \end{cases} .$$

Dada una tabla de contingencia $c \times c \times c$ con N objetos clasificados de forma cruzada por tres jueces independientes, una prueba de permutación exacta implica generar todos los ordenamientos posibles, igualmente probables, de los N objetos en las c^3 celdas, preservando las distribuciones de frecuencia marginal observadas. Para cada ordenamiento en el conjunto de referencia de todas las permutaciones de las frecuencias de las celdas, se calcula el estadístico kappa ponderado, κ_w, y el valor de la probabilidad puntual hipergeométrica exacta bajo la hipótesis nula, $p(n_{ijk}|R_i, C_j, S_k, N)$, donde

$$p(n_{ij}|n_{i.}, n_{.j}, N) = \frac{(\prod_{i=1}^{c} R_i!)(\prod_{j=1}^{c} C_j!)(\prod_{j=1}^{c} S_k!)}{(N!)^2 \prod_{j=1}^{c}\prod_{j=1}^{c}\prod_{k=1}^{c} n_{ijk}!}$$

[Mielke and Berry, 1988b].

Si κ_o denota el valor del estadístico kappa ponderado observado, el valor de probabilidad exacto de κ_o bajo la hipótesis nula viene dado por

$$P(\kappa_o|H_0) = \sum_{l=1}^{M} \Psi_l(n_{ijk}|R_i, C_j, S_k, N) \ ,$$

donde

$$\Psi_l(n_{ijk}|R_i, C_j, S_k, N) = \begin{cases} p(n_{ijk}|R_i, C_j, S_k, N) & \kappa_w \geq \kappa_o \\ 0 & c.c. \end{cases}$$

y M denota el número total de posibles ordenamientos de frecuencia de celdas igualmente probables dados los totales de frecuencia marginal observados. Cuando el conjunto de referencia de M ordenamientos posibles es muy grande, como es típico en las tablas de contingencia multidireccionales, las pruebas exactas son poco prácticas y se hacen necesarios los procedimientos de remuestreo de Monte Carlo. En el remuestreo, una muestra aleatoria de tamaño L extraída de los M posibles ordenamientos de frecuencias de celdas permite comparar los estadísticos de prueba κ_w calculados en las L tablas aleatorias con el estadístico de prueba κ_w calculado en la tabla observada.

4.6. Análisis Ridit

En 1958, I.D.J. Bross introdujo la puntuación ridit para el análisis de datos categóricos ordenados, donde "ridit" es un acrónimo de *Relative to an I dentified Distribution* (relativo a una distribución identificada) y la "it" representa un tipo de transformación similar a logit y probit [Bross, 1958]. Son comunes dos aplicaciones del análisis ridit:

- La primera compara los grupos de tratamiento y de control, donde el grupo de control observado sirve como grupo de referencia y los ridits se calculan para las c categorías disjuntas y ordenadas del grupo de control y se aplican a las c categorías disjuntas y ordenadas del grupo de tratamiento.

  En la primera aplicación, el grupo de control y los ridits correspondientes se tratan como una población infinita y parámetros de población, respectivamente.

- La segunda aplicación compara dos grupos de tratamiento independientes en los que ninguno de los grupos de tratamiento se considera un grupo de referencia y se calculan los ridits para las c frecuencias de categorías ordenadas y disjuntas de cada grupo de tratamiento; y se aplican a las c categorías ordenadas y disjuntas del otro grupo de tratamiento. En esta aplicación, los $k = 2$ grupos de tratamiento se consideran muestras finitas independientes, sin que ninguno se identifique como grupo de referencia. El supuesto de la segunda aplicación de que los grupos son finitos es, en realidad, más realista. En 2009 Mielke, Long, Berry y Johnston generalizaron el análisis ridit para $k \geq 2$ grupos de tratamiento independientes [Mielke et al., 2009].

Se considera una tabla de contingencia de clasificación cruzada $c \times k$ con c categorías de respuesta disjuntas y ordenadas y k grupos de tratamiento desordenados. Siguiendo la notación de Bross, m_{ij} denota la frecuencia de celda observada de la i-ésima fila y j-ésima columna para $i = 1, ..., c$ y $j = 1, ..., k$, sean

$$M_j = \sum_{i=1}^{c} m_{ij}$$

los totales de frecuencia de tratamiento no ordenados para $j = 1, ..., k$, y sea

$$N = \sum_{i=1}^{c} \sum_{j=1}^{k} m_{ij}$$

el total de la frecuencia de la tabla para todas las ck celdas. Las puntuaciones ridit para el j-ésimo tratamiento observado, $j = 1, ..., k$, vienen dadas por

$$R_{1j} = \frac{m_{1j}}{2M_j} \ , \qquad R_{2j} = \frac{m_{1j} + \frac{m_{2j}}{2}}{M_j} \ , \quad ... \quad , \qquad R_{cj} = \frac{m_{1j} + ... + m_{c-1,j} + \frac{m_{cj}}{2}}{M_j} \ .$$

Así, la puntuación ridit R_{ij} para la i-ésima de las c categorías en el j-ésimo de los k tratamientos es la proporción de observaciones en las categorías inferiores a la i-ésima categoría en el j-ésimo tratamiento mas la mitad de la proporción de observaciones en la i-ésima categoría del j-ésimo tratamiento.

4.6.1. Cálculo

Se defina el estadístico de prueba T como

$$T = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} |x_{ij} - x_{ji}| \ ,$$

donde

$$x_{ij} = \sum_{k=1}^{c} \frac{R_{ki} m_{kj}}{M_j}$$

para $i, j = 1, ..., k$.

En el contexto de un análisis ridit de k tratamientos, los procedimientos de permutación exacta examinan todas las posibles asignaciones igualmente probables de los N sujetos a las c categorías ordenadas y disjuntas. Alternativamente, los procedimientos de permutación de remuestreo de Monte Carlo examinan un subconjunto aleatorio seleccionado de entre todas las posibles asignaciones de los N sujetos a las c categorías ordenadas y separadas. La hipótesis nula de una prueba de permutación especifica que todos los resultados posibles del análisis ridit son igualmente probables.

4.6.1.1. Procedimientos de permutación exacta

Los M_j sujetos del j-ésimo grupo de tratamiento, $j = 1, ..., k$, se clasifican en c categorías disjuntas y ordenadas. Entre las c^N configuraciones de asignación igualmente probables bajo la hipótesis nula, hay

$$W = \prod_{j=1}^{k} \begin{pmatrix} M_j + c - 1 \\ c - 1 \end{pmatrix}$$

particiones distintas de las c^N configuraciones de asignación de los k grupos de tratamiento. En una aplicación típica, W y c^N suelen ser muy grandes.

Por lo tanto, un análisis de permutación exacto no suele ser práctico para los análisis ridit con $k > 2$ tratamientos y se recomiendan los procedimientos de permutación de Monte Carlo.

4.6.1.2. Procedimientos de permutación de remuestreo

Un procedimiento de permutación de remuestreo de Monte Carlo genera L conjuntos de \$N\$ asignaciones aleatorias seleccionadas con reemplazo de las c^N configuraciones de asignación igualmente probables de los k grupos de tratamiento. En general, $L = 1.000.000$

es suficiente para garantizar una precisión de tres decimales [Johnston et al., 2007]. Para cada uno de los L conjuntos, los contadores de las c categorías disjuntas y ordenadas indexadas por $i = 1, ..., c$ se ponen a cero y se generan j variables aleatorias uniformes independientes, U_j, sobre $[0, 1)$, para $j = 1, ..., N$. Si U_j pertenece a $[\frac{i-1}{c}, \frac{i}{c})$, el i-ésimo de los contadores c se incrementa en 1. A continuación se calcula el estadístico T de la prueba ridit para cada uno de los L conjuntos de N asignaciones aleatorias de las frecuencias de las categorías ordenadas. Sea T_o el valor observado de T. Entonces, dados los estadísticos ridit del remuestreo $T_1, ..., T_L$, el valor de la probabilidad de la cola superior del remuestreo de T_o bajo la hipótesis nula viene dado por

$$P = \frac{1}{L} \sum_{i=1}^{L} \Psi(T_i) \ ,$$

donde

$$\Psi(T_i) = \begin{cases} 1 & T_i \geq T_o \\ 0 & c.c. \end{cases} \ .$$

Apéndice A

Apéndice: Título del Apéndice

A.1. Primera sección

Apéndice B

Apéndice: Título del Apéndice

B.1. Primera sección

Bibliografía

A.C. Acock and G.R. Stavig. A measure of association for nonparametric statistics. *Social Forces*, 57, 1979.

A. Agresti and B. Finlay. *Statistical Methods for the Social Sciences*. Prentice–Hall, 1997.

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2021. URL https://CRAN.R-project.org/package=rmarkdown. R package version 2.11.

P. Armitage, L.M. Blendis, and H.C. Smyllie. The measurement of observer disagreement in the recording of signs. *J. R. Stat. Soc. A Gen.*, 129, 1966.

J. Berkson. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Stat. Assoc.*, 33, 1938.

Kenneth J. Berry, Janis Johnston, and Paul W. Jr. Mielke. *The Measurement of Association: A Permutation Statistical Approach*. 2010.

K.J. Berry and P.W Mielke. Goodman and kruskal's tau-b statistic: A nonasymptotic test of significance. *Sociol Method Res.*, 13, 1985.

K.J. Berry and P.W. Mielke. A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educ. Psychol. Meas.*, 48, 1988.

K.J. Berry and P.W. Mielke. Extension of spearman's footrule to multiple rankings. *Psychol. Rep.*, 82, 1998.

D. Böhning and H. Holling. A monte carlo study on minimizing chi-square distances under the hypothesis of homogeneity or independence for a two-way contingency table. *Statistics*, 20:55–70, 1989.

H.M. Blalock. Probabilistic interpretations for the mean square contingency. *J. Am. Stat. Assoc.*, 53, 1958.

R.L. Brennan and D.J. Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educ. Psychol. Meas.*, 41, 1981.

H. Brenner and U. Kliebsch. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 7, 1996.

I.D.J. Bross. How to use ridit analysis. 14, 1958.

E.J. Burr. The distribution of kendall's score s for a pair of tied rankings. *Biometrika*, 47, 1960.

D.V. Cicchetti, D. Showalter, and P.J. Tyrer. The effect of number of rating scale categories on levels of interrater reliability. *Appl. Psychol. Meas.*, 9, 1985.

W.G. Cochran. The comparison of percentages in matched samples. *Biometrika*, 37, 1950.

J. Cohen. *A coefficient of agreement for nominal scales*, volume 20. 1960.

J. Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, 70, 1968.

A.J. Conger. Kappa reliabilities for continuous behaviors and events. *Educ. Psychol. Meas.*, 45, 1985.

Herbert L. Costner. *Criteria for Measures of Association.* American Sociological Review, 1965.

M. Cowles. *Statistics in Psychology: An Historical Perspective.* 2 edition, 2001.

E.E. Cureton. Rank-biserial correlation. *Psychometrika*, 21, 1956.

G.A. Ferguson. *Statistical Analysis in Psychology and Education.* McGraw–Hill, 1981.

R.A. Fisher. *Statistical Methods for Research Workers.* Oliver and Boyd, 5th ed edition, 1934.

R.A. Fisher. *The logic of inductive inference (with discussion)*, volume 98. J. R. Stat. Soc., 1935.

J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psych. Bull.*, 76, 1971.

J.L. Fleiss. The equivalence of weighted kappa and the intraclass coefficient as measures of reliability. *Educ. Psychol. Meas.*, 33, 1973.

J.L. Fleiss, B. Levin, and M.C. Paik. *Statistical Methods for Rates and Proportions.* 5 edition, 2003.

L.A. Franklin. Exact tables of spearman's footrule for $n = 11(1)18$ with estimate of convergence and errors for the normal approximation. *Stat. Probab. Lett.*, 6, 1988.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, 32, 1937.

G.V Glass. Note on rank-biserial correlation. *Educ. Psychol. Meas.*, 26, 1966.

L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications. *J. Am. Stat. Assoc.*, 49, 1954.

L.A. Goodman and W.H. Kruskal. Measures of association for cross classifications, iii: Approximate sampling theory. *J. Am. Stat. Assoc.*, 58, 1963.

P. Graham and R. Jackson. The analysis of ordinal agreement data: Beyond weighted kappa. *J. Clin. Epidemiol*, 46, 1993.

J.P. Guilford. *Fundamental Statistics in Psychology and Education*. McGraw–Hill, 1950.

A.A. Hunter. *On the validity of measures of association: The nominal-nominal two-by-two case*. 1973.

R. Iachan. Measures of agreement for incompletely ranked data. *Educ. Psychol. Meas.*, 44, 1984.

J.O. Irwin. Tests of significance for differences between percentages based on small numbers. *Metron*, 12, 1935.

J.E. Johnston, K.J. Berry, and P.W. Mielke. Permutation tests: Precision in estimating probability values. *Percept. Motor Skill*, 105, 2007.

M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30, 1938.

M.G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33, 1945.

M.G. Kendall. *Rank Correlation Methods*. 1948.

M.G. Kendall. *Rank Correlation Methods, 3rd edn*. 1962.

M.G. Kendall and B. Babington Smith. The problem of m rankings. *Ann.Math. Stat.*, 10, 1939.

M.G. Kendall and B. Babington Smith. On the method of paired comparisons. *Biometrika*, 31, 1940.

M.G. Kendall, S.F.H. Kendall, and B. Babington Smith. The distribution of spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, 30, 1939.

J.-O. Kim. Predictive measures of ordinal association. *Am. J. Soc.*, 76, 1971.

C.A. Kraft and C. van Eeden. *A Nonparametric Introduction to Statistics*. Macmillan, 1968.

K. Krippendorff. *Bivariate agreement coefficients for reliability of data*. Borgatta, E.F. (ed.) Sociological Methodology, 1970.

J.R. Landis and G.G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33, 1977.

R.K. Leik and W.R. Gove. The conception and measurement of asymmetric monotonic relationships in sociology. *Am. J. Sociol.*, 74, 1969.

R.K. Leik and W.R. Gove. *Integrated approach to measuring association*. Costner, H.L. (ed.) Sociological Methodology, 1971.

R.J. Light. *Measures of response agreement for qualitative data: Some generalizations and alternatives*, volume 76. Psychol. Bull, 1971.

R.J. Light and B.H. Margolin. An analysis of variance for categorical data. *J. Am. Stat. Assoc.*, 66, 1971.

M. Maclure and W.C. Willett. Misinterpretation and misuse of the kappa statistic. *Am. J. Epidemiol.*, 126, 1987.

H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, 18, 1947.

L.A. Marascuilo and McSweeney. *Nonparametric and Distribution-free methods in the Social Sciences*. Brooks–cole edition, 1977.

B.H. Margolin and R.J. Light. An analysis of variance for categorical data, ii: Small sample comparisons with chi square and other competitors. *J. Am. Stat. Assoc*, 69, 1974.

Q. McNemar. Note on the sampling error of the differences between correlated proportions and percentages. *Psychometrika*, 12, 1947.

P.W. Mielke and K.J. Berry. Cumulant methods for analyzing independence ofr-way contingency tables and goodness-of-fit frequency data. *Biometrika*, 1988a.

P.W. Mielke and K.J. Berry. Cumulant methods for analyzing independence of r-way contingency tables and goodness-of-fit frequency data. *Biometrika*, 75, 1988b.

P.W. Mielke and M.M. Siddiqui. *A combinatorial test for independence of dichotomous responses*, volume 60. J. Am. Stat. Assoc., 1965.

P.W. Mielke, K.J. Berry, and D. Zelterman. Fisher's exact test of mutual independence for 2×2×2 cross-classification tables. *Educ. Psychol. Meas.*, 54, 1994.

P.W. Mielke, K.J. Berry, and J.E. Johnston. Resampling programs for multiway contingency tables with fixed marginal frequency totals. *Psychol. Rep.*, 101, 2007.

P.W. Mielke, K.J. Berry, and J.E. Johnston. Resampling probability values for weighted kappa with multiple raters. *Psychol. Rep.*, 102, 2008.

P.W. Mielke, M.A. Long, K.J. Berry, and J.E. Johnston. g-treatment ridit analysis: Resampling permutation methods. *Stat. Methodol.*, 6, 2009.

R. Newson. *Identity of Somers' D and the rank biserial correlation coefficient.* 2008.

W.S. Robinson. The statistical measurement of agreement. *Am. Sociol. Rev.*, 22, 1957.

I.A. Salama and D. Quade. A note on spearman's footrule. *Commun. Stat. Simul. C*, 19, 1990.

W.A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Quart.*, 19, 1955.

R.C. Serlin, J. Carr, and L.A. Marascuilo. A measure of association for selected nonparametric procedures. *Psychol. Bull.*, 92, 1982.

R.H. Somers. A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.*, 27, 1962.

C.E. Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 15, 1904.

C.E. Spearman. 'footrule' for measuring correlation. *Brit. J. Psychol.*, 2, 1906.

C.E. Särndal. A comparative study of association measures. *Psychometrika*, 39, 1974.

A. Stuart. The estimation and comparison of strengths of association in contingency tables. *Biometrika*, 40, 1953.

A. Stuart. Spearman-like computation of kendall's tau. *Brit. J. Math. Stat. Psy.*, 30, 1977.

H.K. Ury and D.C. Kleinecke. Tables of the distribution of spearman's footrule. *J. R. Stat. Soc. C Appl.*, 28, 1979.

W.A. Wallis. The correlation ratio for ranked data. *J. Am. Stat. Assoc.*, 34, 1939.

R.S. Weiss. *Statistics in Social Research: An Introduction*. Wiley, 1968.

J.W. Whitfield. Rank correlation between two variables, one of which is ranked, the other dichotomous. *Biometrika*, 34, 1947.

T.D. Wickens. *Multiway Contingency Tables Analysis for the Social Sciences*. 1989.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2021a. URL https://CRAN.R-project.org/package=ggplot2. R package version 3.3.5.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021b. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.7.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bull.*, 1, 1945.

T.P. Wilson. *Measures of association for bivariate ordinal hypotheses*. Blalock, H.M., 1974.

Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2021. URL https://yihui.org/knitr/. R package version 1.34.

F. Yates. Contingency tables involving small numbers and the χ^2 test. *Suppl. J. R. Stat. Soc.*, 1, 1934.