COMPARACION ENTRE DIFERENTES METODOS DE ANALISIS DE RESIDUOS EN TABLAS DE CONTINGENCIA Y UN NUEVO ENFOQUE

Adalberto González Debén, Instituto de Cibernética, Matemática y Física, CITMA, Cuba
Ignacio Méndez Ramírez, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, México

RESUMEN
Rudas, Clogg y Lindsay, en 1994, propusieron un nuevo índice para medir la falta de ajuste de un modelo cualquiera H en tablas de contingencia. Este índice está basado en un modelo de dos clases latentes, en la primera clase latente se cumple el modelo H y para la segunda clase latente no se hace ninguna suposición. El índice de falta de ajuste se interpreta como la proporción de individuos de la población que está intrínsecamente fuera del modelo H. En este trabajo se propone el análisis de la segunda clase latente para estudiar la falta de ajuste de un modelo y detectar posibles celdas atípicas. Por medio de tres ejemplos clásicos reportados en la literatura, se compara el análisis de la segunda clase latente con el análisis de residuos estandarizados, el análisis de residuos eliminados y la variación del estadístico de razón de verosimilitud. En ninguno de los tres ejemplos el análisis de la segunda clase latente fue peor que los otros criterios.

Palabras claves: tabla de contingencia, bondad de ajuste, análisis de residuos.

ABSTRACT
In 1994, Rudas, Clogg and Lindsay proposed a new index of lack of fit for a certain model H in contingency table analysis. The framework of the index is a two latent class model where the model H applies to the first latent class and the second latent class is unrestricted. The index of lack of fit is the proportion of individuals of the population intrinsically outside the model. In this work we propose the analysis of the second latent class to study the causes of lack of fit of a model and to detect outlying cells. By means of three classical examples from the literature Rudas' second latent class analysis is compared to standardized residual analysis, the deleted residual analysis, and the variations of likelihood ratio statistics. The second latent class was not worse than any other criterion in the three examples.

Key words: contingency table, goodness of fit, residual análisis.

MSC: 62H17

1. INTRODUCCION

Dada una tabla de contingencia de N celdas con el conjunto de probabilidades en cada una de ellas

$$P = \{P_h ; h = 1, \cdots N\}, \qquad \sum_{h=1}^{N} P_h = 1 ,$$

sean las frecuencias observadas

$$f = \{f_h ; h = 1, \cdots N\}$$

y el tamaño de la muestra

$$n = \sum_{h=1}^{N} f_h$$

Bajo el esquema de muestreo multinomial las frecuencias esperadas son

$$e = \{e_h ; h = 1, \cdots N\}$$

donde

$$e_h = nP_h$$

El procedimiento estándar para comprobar si un cierto modelo H es adecuado es el siguiente:

1. Se obtiene $\hat{e}^{(H)}$, que son las estimaciones máximo verosímiles de los valores esperados bajo el modelo H.

2. Se comparan f *y* $\hat{e}^{(H)}$ para decidir si el modelo H se ajusta a los datos.

Los estadígrafos más utilizados para cuantificar la discrepancia entre los valores observados y los valores esperados son el ji-cuadrado y el de razón de verosimilitud. El estadígrafo ji-cuadrado es

$$X^2 = \sum_{h=1}^{N} \frac{(f_h - \hat{e}_h)^2}{\hat{e}_h}$$

y el estadígrafo de razón de verosimilitud,

$$G^2 = 2\sum_{h=1}^{N} f_h \log(\frac{f_h}{\hat{e}_h})$$

Son muy conocidas las limitaciones prácticas de esta metodología, a saber: cuando n es pequeño o las tablas son ralas (no densas) no resulta apropiada la suposición de distribución asintótica de los estadígrafos de prueba y si, por el contrario, n es muy grande se rechaza cualquier modelo sencillo.

Las pruebas de bondad de ajuste global, como lo indica su nombre, proporcionan una respuesta general a la pregunta de si el modelo se ajusta a los datos o no. Sin embargo, no sirven para determinar las causas de falta de ajuste ni tampoco, en caso afirmativo, permiten saber si el modelo realmente representa a todos los datos o si por el contrario hay celdas que se desvían considerablemente del valor esperado. Esto sólo se logra observando celda por celda las discrepancias entre el valor observado y el esperado.

Los residuos estandarizados y eliminados son los más conocidos y utilizados en la literatura con estos fines. El residuo estandarizado se define como

$$r_h = \frac{f_h - e_h}{\sqrt{e_h}} .$$

Bajo el supuesto de que el modelo se cumple, r_h tiene distribución asintótica normal estandar.

El residuo eliminado Simonoff (1988), es menos conocido. Se trata de la comparación entre el valor observado y el valor perfecto de la celda. El valor perfecto es el que tendría que haber en esa celda para que coincida exactamente con el valor esperado según el modelo de interés y la información de las celdas restantes. El residuo eliminado se define como

$$r_h^* = \frac{f_h - x}{\sqrt{x}}$$

donde *x* es el valor perfecto de la celda h.

2. DETECCION DE CELDAS ATIPICAS (outliers)

Según Barnett y Lewis (1994), “En sentido general, un outlier en un conjunto de datos es una observación que parece inconsistente con el modelo supuesto”. En el caso de tablas de contingencia este concepto se refiere a las celdas y no a los individuos aislados.

2.1. Análisis usual

Haberman (1973) propuso inspeccionar gráficamente los residuos ajustados. Brown (1974) propuso declarar una celda como atípica si el modelo de independencia no se ajusta y el de cuasi-independencia, que resulta de considerar esa celda como cero estructural, sí se ajusta. Fuchs y Kenet (1980) propusieron un test con este mismo fin y Simonoff (1988) otro que lo supera en cuanto a la capacidad para detectar más de una celda.

Otros autores han explorado la conveniencia de aplicar conceptos ya establecidos en el contexto de análisis de regresión como son los de palanca Andersen (1992), Upton y Guillen (1995) e influencia Upton y Guillen (1995).

El consenso parece ser favorable al residuo eliminado r^* y a la diferencia del estadígrafo de razón de verosimilitud

$$\Delta G^2 = G^2 - G_h^2$$

donde G^2 es el estadígrafo de razón de verosimilitud para la tabla original y G_h^2 es el estadígrafo de razón de verosimilitud que resulta de sustituir f_h por x.

Cuando sólo se cuenta con la información de que un cierto modelo no se ajusta, se deben hacer tantas pruebas como celdas tiene la tabla. Para poder garantizar que la probabilidad de cometer el error de tipo I no exceda el nivel de significación prefijado α, se debe tomar un nivel tal que permita que esto no suceda. Una solución conservadora muy utilizada es la que se conoce como ajuste de Bonferroni, que consiste en adoptar un valor

$$\alpha^* = \alpha / N.$$

donde N es el número de celdas.

2.2. Ejemplos

Para ilustrar la utilización de r,r* y ΔG^2 se muestran tres ejemplos muy conocidos en la literatura acerca de esta temática. El primer ejemplo Wickens (1989), es una tabla de 3 x 4 con una celda atípica. El segundo ejemplo es el de la tabulación cruzada de los colores de los ojos y el pelo para una muestra de 592 individuos Rudas **et al**. (1994), donde parece haber dos celdas atípicas. El tercer ejemplo Simonoff (1988), Barnett y Lewis (1994), es una tabla de 5 x 5 con tres celdas atípicas. En todos los casos se trabajó con el modelo de independencia.

EJEMPLO 1

En la Tabla 2.1a se muestran los datos que analiza Wickens (1989). No se cumple el modelo de independencia entre filas y columnas (G^2 = 18.2 con 6 grados de libertad). En las Tablas 2.1b, 2.1c y 2.1d se muestran los resultados del análisis de residuos usual.

Tabla 2.1a. Tabla 10.4 página 252, Wickens (1989).

12	18	3	10
18	3	7	10
20	6	9	11

Tabla 2.1b. Residuos estandarizados

-1.2	**2.9**	-1.4	-0.2
0.8	-1.8	0.6	0.2
0.4	-1.2	0.8	-0.1

.

Tabla 2.1c. Residuos eliminados para cada una de las celdas de la Tabla 2.1a.

-2.68	**8.66**	-2.09	-0.3
2.02	-2.76	0.99	0.46
1.2	-2.14	1.66	-0.14

Tabla 2.1d. ΔG^2 para cada una de las celdas de la Tabla 2.1a.

3.67	**15.74**	3.63	0.047
1.44	6.66	0.49	0.11
0.51	3.08	1.17	0.0096

Para α^* = 0.05/12 = 0.00416 el valor crítico de la distribución ji-cuadrado con un grado de libertad es 8.21. Como conclusión resulta evidente que según todos los puntos de vista considerados se detecta la celda (1,2) como atípica.

EJEMPLO 2

En la Tabla 2.2 a se muestran los datos que analizan Rudas **et al**. (1994). No se cumple el modelo de independencia (X^2 =138.29 y G^2 =146.44, con 9 grados de libertad). En las Tablas 2.2b, 2.2c y 2.2d se muestran los resultados.

Tabla 2.2a. Tabla 1 página 625, Rudas **et al**. (1994).

OJOS / PELO	Negro	Castaño	Rojo	Rubio
Café	68	119	26	7
Azul	20	84	17	94
Gris	15	54	14	10
Verde	5	29	14	16

Tabla 2.2b. Residuos estandarizados.

OJOS / PELO	Negro	Castaño	Rojo	Rubio
Café	**4.4**	1.23	-0.08	**-5.85**
Azul	**-3.07**	-1.95	-1.73	**7.05**
Gris	-0.48	1.35	0.85	-2.23
Verde	-1.95	-0.35	2.28	0.61

Tabla 2.2c. Residuos eliminados para cada una de las celdas de la Tabla 2.2a.

OJOS / PELO	Negro	Castaño	Rojo	Rubio
Café	**11.6**	4.05	-0.13	**-9.38**
Azul	**-5.11**	**-5.47**	-2.8	**24.18**
Gris	-0.68	3.45	1.19	-3.06
Verde	-2.46	-0.73	3.23	0.9

Tabla 2.2d. ΔG^2 para cada una de las celdas de la Tabla 2.2a.

OJOS / PELO	Negro	Castaño	Rojo	Rubio
Café	**36.45**	4.69	0.01	**85.69**
Azul	**19.69**	**11.62**	5.65	**97.12**
Gris	0.34	4.21	0.93	8.54
Verde	6.24	0.26	5.66	0.52

Para $\alpha^* = 0.05/16 = 0.031$ el valor crítico de la distribución ji-cuadrado es 8.73. Según el criterio ΔG^2 se detectan 5 celdas atípicas (1,1), (1,4), (2,1), (2,2) y (2,4). Estas celdas son las que tienen valores mayores de r*.

EJEMPLO 3

En la Tabla 2.3a se muestran los datos que analizan Simonoff (1988), y Barnett y Lewis (1994), y en las Tablas 2.3b, 2.3c y 2.3d se muestran los resultados.

Tabla 2.3a. Tabla 10.4. página 252, Simonoff (1988), Barnett y Lewis (1994).

18	41	41	20	21
39	20	20	22	22
24	20	20	16	18
20	20	19	19	19
23	19	20	17	20

Tabla 2.3b. Residuos estandarizados.

-2.38	**1.94**	**1.94**	-0.77	-0.85
2.23	-1.25	-1.25	0.28	-0.009
4.8	-0.23	-0.23	-0.13	0.1
-0.34	-0.19	-0.41	0.66	0.39
0.21	-0.5	-0.28	0.08	0.54

Tabla 2.3c. Residuos eliminados para cada una de las celdas de la Tabla 2.3a.

-3.69	**3.65**	**3.64**	-1.19	-1.33
4.11	-1.94	-1.94	0.44	-0.014
0.76	-0.36	-0.36	-0.18	0.15
-0.51	-0.29	-0.62	0.99	0.58
0.34	-0.75	-0.43	0.12	0.81

Tabla 2.3d. ΔG^2 para cada una de las celdas de la Tabla 2.3a.

10.66	6.12	6.12	0.98	1.21
7.74	2.7	2.7	0.12	0.00
0.35	0.085	0.085	0.023	0.016
0.18	0.06	0.26	0.61	0.22
0.07	0.39	0.12	0.01	0.42

Los residuos de las celdas (1,1), (1,2), (1,3) y (2,1) son los más grandes. Para $\alpha^* = 0.05/25 = 0.02$ el valor crítico de la distribución ji-cuadrado es 9.55. Según el criterio ΔG^2 sólo se detecta la celda (1,1) como atípica.

Como conclusión se puede decir que:

1.Todos estos criterios son efectivos cuando hay una sola celda atípica.

2. Cuando hay más de una celda atípica puede que alguna quede oculta o que alguna, que en realidad no lo es, lo parezca.

3. INDICE DE FALTA DE AJUSTE DE RUDAS, CLOGG Y LINDSAY

Rudas **et al**. (1994), proponen un nuevo índice de falta de ajuste **(NIFA)** que sirve para cualquier tabla de contingencia y cualquier modelo. Dados una tabla de contingencia P, con N celdas, y un modelo H, se propone una familia de modelos $\{H_\pi\}$ de la forma:

$$P_h = (1-\pi)P_{1h} + \pi P_{2h},\ h = 1,...,N; \qquad 0 \le \pi \le 1 \tag{1}$$

Se trata de expresar la tabla P como una combinación convexa de dos subtablas P_1 y P_2, también llamadas primera cara y segunda cara respectivamente, de manera que el modelo H se cumple para la subtabla P_1 y no se hace ninguna suposición respecto a la subtabla P_2.

Se define el índice de falta de ajuste π^* como el menor valor de π para el que se cumple la ecuación (1). La interpretación de π^* es muy sencilla: es la proporción de individuos de la población que está intrínsecamente fuera del modelo H (medida con error o mal clasificada) y, consecuentemente, 1 $-\pi^*$ representa la proporción de la población intrínsecamente descrita por H.

La ecuación (1) es una generalización del modelo H, pues representa una familia que va desde el modelo H (π = 0) hasta el modelo completamente irrestricto (π =1). Por otro lado, al no incluir ninguna suposición para la segunda tabla P_2, constituye una generalización del modelo usual de clases latentes (que supone el modelo de independencia para ambas tablas).

Clogg **et al**. (1995) utilizan el índice de falta de ajuste para comparar los modelos de independencia, cuasiindependencia y asociación cuasiuniforme en tablas de mobilidad; y González (1998) lo utiliza para comparar el modelo de independencia con el modelo de asociación fila-columna en tablas de contingencia ordinales.

4. UNA ALTERNATIVA AL ANALISIS DE RESIDUOS USUAL

Como en el modelo (1) no se le impone ninguna restricción a la subtabla P_2, se supone que en ella se concentra la parte de la población que no es modelada por H. Si π^* resulta grande, y por lo tanto se requiere mejorar el modelo H, se puede inspeccionar esta subtabla y de esta manera obtener información útil para este propósito. Esto se considera una alternativa al análisis de residuos en el procedimiento usual. La diferencia radica en que los residuos, en el procedimiento usual, se basan en la suposición de que el modelo de interés se cumple para toda la población; y cuando esta suposición no es válida los residuos tienen muy poco sentido Rudas (1998).

Clogg **et al**. (1995) analizan la segunda clase latente con vistas a investigar la contribución de cada celda a la falta de ajuste de cada uno de los modelos utilizados; y en Clogg **et al**. (1997) se proponen algunos gráficos sencillos para facilitar el análisis de la segunda clase latente.

En este acápite se propone el análisis de la segunda clase latente como otra forma de detectar celdas atípicas (outliers). A modo de ilustración se consideran los tres ejemplos ya vistos en el epígrafe II.

En el primer ejemplo Wickens (1989), el 23 % de los individuos se aparta de la hipótesis de independencia y más de la mitad de ellos se concentra en la celda (1,2) (ver Tabla 4.1).

Para el segundo ejemplo Rudas **et al**. (1994) el valor del índice de falta de ajuste es NIFA =29.34 %.

En la Tabla 4.2a se muestran los valores brutos de la segunda cara y en las tablas 4.2b y 4.2c aparecen expresados en porcientos del total de individuos y del total de individuos en la segunda cara respectiva-mente. Por último en la tabla 4.2d se muestra el porciento de individuos de cada celda que no sigue el modelo de independencia entre filas y columnas.

Tabla 4.1. Análisis de la segunda cara de Rudas para la Tabla 2.1a.

Celda	Observado	cara 1	cara 2	cara 2 (%)
1,1	12	11.95	0.01	0.05
1,2	18	2.11	**15.75**	**53.92**
1,3	3	3.05	0.00	0.00
1,4	10	6.66	3.26	11.17
2,1	18	17.2	0.76	2.58
2,2	3	3.03	0.0	0.00
2,3	7	4.39	2.56	8.76
2,4	10	9.58	0.38	1.31
3,1	20	19.8	0.15	0.5
3,2	6	3.49	2.46	8.43
3,3	9	5.05	3.88	13.27
3,4	11	11.02	0.00	0.02
Total	127	95.25	25.75	100.0

Tabla 4.2a. Valores brutos de la segunda cara.

OJOS / PELO	Negro	Castaño	Rojo	Rubio
Café	**39.8**	2.1	2.2	0.0
Azul	0.1	0.6	0.1	**88.7**
Gris	2.3	0.9	3.2	6.8
Verde	0.0	7.8	9.7	14.7

Tabla 4.2b. % en cada celda de la segunda entre el total de individuos.

OJOS / PELO	Negro	Castaño	Rojo	Rubio
Café	**6.72**	0.35	0.37	0.0
Azul	0.02	0.1	0.02	**14.98**
Gris	0.39	0.15	0.54	1.15
Verde	0.0	1.32	1.64	2.48

Tabla 4.2c. % entre el total de individuos de la segunda cara.

OJOS/PELO	Negro	Castaño	Rojo	Rubio
Café	**22.3**	1.2	1.2	0.0
Azul	0.0	0.4	0.1	**49.6**
Gris	1.3	0.5	1.8	3.8
Verde	0.0	4.3	5.4	8.2

Tabla 4.2d. % de cada celda en la segunda cara.

OJOS/PELO	Negro	Castaño	Rojo	Rubio
Café	**58.5**	2.0	8.5	0.0
Azul	0.5	0.7	0.6	**94.4**
Gris	15.3	1.7	22.9	68
Verde	0.0	26.9	69.3	91.9

Como se puede apreciar, sólo se destacan las celdas (1,1) ojos café y pelo negro y (2,4) ojos azules y pelo rubio.

El tercer ejemplo es la tabla presentada por Simonoff (1988) y Barnett y Lewis (1994). Los resultados se muestran en las tablas 4.3a y 4.3b.

Tabla 4.3a. Método de Rudas para la Tabla 2.3a y el modelo de independencia (cara 1).

18.14	20.97	20.98	18.27	20.13
17.29	19.99	20	17.42	19.19
15.89	18.37	18.37	16	17.63
16.32	18.87	18.87	16.44	18.11
16.33	18.89	18.89	16.45	18.12

Tabla 4.3b. Método de Rudas para la tabla 2.3a y el modelo de independencia (cara 2).

0.0	19.92	19.91	1.67	0.84
21.61	0.05	0.04	4.53	2.77
8.05	1.56	1.56	0.1	0.4
3.63	1.09	0.2	2.52	0.9
6.6	0.18	1.06	0.54	1.85

Como se puede apreciar, en la segunda cara se destacan las celdas (1,2), (1,3) y (2,1).

En la Tabla 4.4 se muestra un resumen de los tres ejemplos considerados. En la misma aparecen las celdas detectadas como atípicas con cada uno de los métodos estudiados.

Tabla 4.4. Celdas detectadas como atípicas según cada método.

Ejemplo / criterio	**residuo estandarizado (r)**	**residuo eliminado (r*)**	**ΔG^2**	**Cara 2**
Wickens	1,2	1,2	1,2	1,2
Pelo / ojos	1,1 1,4 2,1 2,4	1,1 1,4 2,1 2,2 2,4	1,1 1,4 2,1 2,2 2,4	1,1 2,4
Barnett, Simonoff	1,1 1,2 1,3 2,1	1,1 1,2 1,3 2,1	1,1	1,2 1,3 2,1

En ninguno de los tres ejemplos el análisis de la segunda cara fue peor que los otros criterios. En el primer caso todos los criterios coincidieron. En el segundo ejemplo, con el análisis de la segunda cara sólo se detectaron dos celdas atípicas; mientras que por el contrario, el criterio r fue elevado para 4 celdas y los criterios r* y ΔG^2 detectaron esas mismas cuatro celdas y además la celda (2,2). En el tercer ejemplo, con el análisis de la segunda cara de Rudas se detectaron las tres celdas atípicas. Sin embargo, los criterios r y r* resaltaron una celda más y, por el contrario, el criterio ΔG^2 sólo permitió detectar una celda, que además no es ninguna de las tres señaladas.

5. CONCLUSIONES Y RECOMENDACIONES

Como la primera cara queda con "todas las celdas perfectas" y la segunda contiene a los individuos que "sobran", el análisis de la segunda cara sirve como una herramienta más para estudiar las causas de falta de ajuste de un modelo y detectar posibles celdas atípicas. Con este fin se puede utilizar como complemento del análisis de residuos estandarizados, el análisis de residuos eliminados y la variación del estadístico de razón de verosimilitud.

Se necesita seguir experimentando esta metodología con otros ejemplos de la literatura y problemas reales, que involucren más variables y modelos más complejos.

En el trabajo original de Rudas **et al.** (1994), se recomienda estudiar el caso de los ceros muestrales; esto debe extenderse al caso más general de celdas subfrecuentes.

REFERENCIAS

AGRESTI, A. (1990): **Categorical data analysis**. Wiley: New York.

ANDERSEN, E.B. (1992): "Diagnostic in categorical data analysis", **Journal of the Royal Statistical Society,** B54, 781- 791.

BARNETT, V.D. and T. LEWIS (1994): **Outliers in statistical data**. 3rd ed. Wiley: New York.

BROWN, M.B. (1974): "Identification of the sources of significance in two-way contingency tables", **Applied Statistics** 23, 405-413.

CLOGG, C.C.; T. RUDAS and L. XI (1995): "A new index of structure for the analysis of models for mobility tables and other cross-classifications", **Sociological Methodology**, 25, 197-222.

CLOGG, C.C.; T. RUDAS and S. MATTHEWS (1997): "Analysis of contingency tables using graphical displays based on the mixture index of fit", In Blasius, J. y Greenacre, M. (Eds). **Visualization of categorical data**, 425-439. A.P: New York.

FUCHS, C. and R. KENETT (1980): "A test for detecting outlying cells in the multinomial distribution and two-way contingency tables", **Journal of the American Statistical Association**, 75, 395-398.

GONZALEZ, D.A. (1998): Experiencias con un nuevo índice de falta de ajuste en el análisis de tablas de contingencia. Tesis de Maestría. Facultad de Matemática y Computación, Universidad de La Habana.

HABERMAN, S.J. (1973): "The analysis of residuals in cross-classified tables", **Biometrics**, 29, 205-220.

LITTLE, R.J.A. and B. RUBIN (1987): **Statistical analysis with missing data**, Wiley, New York.

RUDAS, T. (1998): "The mixture index of fit", **Advances in Methodology, Data Analysis, and Statistics**, Anuška Ferligoj (Editor), Metodološki zvezki, 14. Ljubljana: FDV.

RUDAS, T.; CLOGG, C.C. and B.G. LINDSAY (1994): "A new index of fit based on Mixture Methods for the analysis of contingency tables", **Journal of the Royal Statistical Society**. B56, 623-639.

SIMONOFF, J.S. (1988): "Detecting outlying cells in two-way contingency tables via backwards-stepping", **Technometrics**, 30, 339-345.

UPTON, G.J.G. and M. GUILLEN (1995): "Perfect cells, direct models and contingency table outliers", **Commun. Statist.- Theory Meth.**, 24, 1843-62.

WICKENS, T.D. (1989): **Multivariate contingency tables analysis for the behavioral sciences**, Lawrence Erlbaum Associates Inc: Hillsdale, New Jersey.