Análisis de datos Categóricos

Pruebas Chi-cuadrado para tablas de dos vías

Ms Carlos López de Castilla Vásquez

Universidad Nacional Agraria La Molina

2017-1

Independencia y Homogeneidad

- Una prueba de *independencia* se aplica sobre una muestra multinomial con probabilidades $\{\pi_{ij}\}$ para una tabla de contingencia $I \times J$.
- La hipótesis nula es:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad i = 1, \cdots, I \quad j = 1, \cdots, J$$

- La prueba de *homogeneidad* se aplica a las muestras multinomiales en las I filas de una tabla de contingencia.
- En los capítulos anteriores se introdujo el estadístico chi-cuadrado de Pearson para probabilidades multinomiales.

Estadístico de prueba de Pearson

- Los estimadores de máxima verosimilitud para $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$:

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n} \qquad \hat{\pi}_{+j} = \frac{n_{+j}}{n}$$

- Las frecuencias esperadas:

$$\left\{\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}\right\}$$

- El estadístico de prueba de Pearson es:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

Estadístico de prueba de Pearson

- Pearson (1900, 1904, 1922) consideró que la distribución de $X^2 \to \chi^2$ con $I \times J - 1$ grados de libertad.
- Fisher (1922) corrige el error de Pearson (sección 16.2) introduciendo en su artículo la noción de *grados de libertad*.
- Como se requiere estimar $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$ los grados de libertad son:

$$(I \times J - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1)$$

- Las dimensiones de $\{\pi_{i+}\}$ y $\{\pi_{+j}\}$ reflejan la restricción $\sum_i \pi_{i+} = \sum_j \pi_{+j} = 1$.

Prueba de razón de verosimilitud

- Para un muestreo multinomial, el núcleo de la verosimilitud es:

$$\prod_i \prod_j \pi_{ij}^{n_{ij}}$$

- Bajo la hipótesis de independencia:

$$\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n^2}$$

- La razón de verosimilitud es:

$$\Lambda = \frac{\prod_i \prod_j (n_{i+}n_{+j})^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}}$$

Prueba de razón de verosimilitud

- El estadístico de prueba G^2:

$$G^2 = -2 \log \Lambda = 2 \sum_i \sum_j n_{ij} \log \left(\frac{n_{ij}}{\hat{\mu}_{ij}} \right)$$

  donde $\{\hat{\mu}_{ij} = n_{i+} n_{+j} / n\}$.
- Cuanto mayores son los valores de G^2 y X^2 mayor evidencia existe en contra de la independencia.
- Para muestras grandes la distribución aproximada de $G^2 \sim \chi^2$ con $(I-1)(J-1)$ grados de libertad.

Ejemplo: Creencias religiosas y educación

- La siguiente tabla clasifica una muestra de 2726 personas de acuerdo a su creencia religiosa y al nivel de educación alcanzado.

Tabla 1: *Creencias religiosas y educación*

	Creencia religiosa			
Educación	Extrema	Moderada	Liberal	Total
Primaria	178	138	108	424
Secundaria	570	648	442	1660
Bachiller	138	252	252	642
Total	886	1038	802	2726

Ejemplo: Creencias religiosas y educación

- Los estadísticos $X^2 = 69{,}2$ y $G^2 = 69{,}8$ con 4 grados de libertad proporcionan una fuerte evidencia de asociación entre las variables.

Tabla 2: *Frecuencias esperadas estimadas*

	Creencia religiosa			
Educación	Extrema	Moderada	Liberal	Total
Primaria	137.8	161.5	124.7	424
Secundaria	539.5	632.1	488.4	1660
Bachiller	208.7	244.5	188.9	642
Total	886	1038	802	2726

Introducción

- Una prueba de independencia chi-cuadrado tiene utilidad limitada.
- Un valor-p pequeño indica una fuerte evidencia de asociación pero proporciona poca información acerca de su intensidad.
- Se ha advertido sobre los peligros de confiar únicamente en los resultados de las pruebas chi-cuadrado en lugar de estudiar la naturaleza de la asociación.
- En esta sección se discute una segunda etapa en el análisis para descubrir más acerca de la asociación entre las variables.

Pearson y residuales estandarizados

- Una comparación celda a celda entre la frecuencia observada y la frecuencia esperada estimada puede ayudar a descubrir la naturaleza de la dependencia.
- El *residual de Pearson* se define por:

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

- El estadístico de prueba de Pearson es $X^2 = \sum_i \sum_j e_{ij}^2$.
- Bajo la hipótesis nula $\{e_{ij}\}$ son asintóticamente normales con media cero y variancia menor que uno.

Pearson y residuales estandarizados

- Comparando los residuales de Pearson con los percentiles de la distribución normal estándar es posible identificar, de manera conservadora, las celdas que presentan falta de ajuste.
- El *residual estandarizado de Pearson* se define por:

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}\left(1 - p_{i+}\right)\left(1 - p_{+j}\right)}}$$

- Un residual estandarizado de Pearson mayor a 2 o 3, en valor absoluto, indica falta de ajuste al supuesto de independencia para dicha celda.

Ejemplo: Creencias religiosas y educación

- Los residuales estandarizados de Pearson se muestran a continuación:

Tabla 3: *Residuales estandarizados de Pearson*

	Creencia religiosa		
Educación	Extrema	Moderada	Liberal
Primaria	4.5	-2.6	-1.9
Secundaria	2.6	1.3	-4.0
Bachiller	-6.8	0.7	6.3

Ejemplo: Creencias religiosas y educación

- La tabla anterior muestra un residual positivo grande para los sujetos con educación primaria y con creencia religiosa extrema. Lo mismo se observa para los sujetos con bachillerato y creencia religiosa liberal.
- Lo anterior significa que hay más sujetos en esas combinaciones de lo que predice el supuesto de independencia.
- Similarmente, existen menos estudiantes con bachillerato y creencia religiosa extrema y menos estudiantes con secundaria y creencia religiosa liberal de lo que predice el supuesto de independencia.

Particionado el estadístico chi-cuadrado

- Sea $Z \sim \mathcal{N}(0,1)$ entonces $Z^2 \sim \chi^2_1$.
- Una variable aleatoria con distribución chi-cuadrado con ν grados de libertad tiene representación $Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$.
- Luego, el estadístico chi-cuadrado podría particionarse en ν componentes chi-cuadrado con 1 grado de libertad.
- Ademas, si X_1^2 y X_2^2 son variables aleatorias independientes con distribuciones chi-cuadrado con ν_1 y ν_2 grados de libertad respectivamente, entonces $X^2 = X_1^2 + X_2^2 \sim \chi^2_{\nu_1+\nu_2}$.
- Un particionamiento puede mostrar que una asociación refleja básicamente diferencias entre ciertas categorías o grupos de categorías.

Ejemplo: Origen de la esquizofrenia

- La siguiente tabla clasifica una muestra de psiquiatras de acuerdo al pensamiento de la escuela a la que pertenecen y su opinión sobre el origen de la esquizofrenía.

Tabla 4: *Origen de la esquizofrenia*

	Opinión sobre el origen		
Escuela	Biogen	Amb	Comb
Ecléctica	90	12	78
Médica	13	1	6
Psico	19	13	50

- En la tabla anterior se tiene que $G^2 = 23{,}04$ con 4 grados de libertad.

Ejemplo: Origen de la esquizofrenia

- Para comprender mejor esta asociación se particiona G^2 en cuatro componentes independientes.
- Para la primera sub-tabla $G_1^2 = 0{,}29$ con 1 grado de libertad.
- Para la segunda sub-tabla $G_2^2 = 1{,}36$ con 1 grado de libertad.

Tabla 5: *Sub-tablas 1 y 2*

	Bio	Amb		Bio+Amb	Comb
Ecl	90	12	Ecl	102	78
Med	13	1	Med	14	6

Ejemplo: Origen de la esquizofrenia

- Existe poca evidencia de tener diferencias entre el pensamiento de las escuelas Ecléctica y Médica sobre el origen de la esquizofrenia.
- Para la tercera sub-tabla $G_3^2 = 12{,}95$ con 1 grado de libertad.
- Para la cuarta sub-tabla $G_4^2 = 8{,}43$ con 1 grado de libertad.

Tabla 6: *Sub-tablas 3 y 4*

	Bio	Amb		Bio+Amb	Comb
Ecl+Med	103	13	Ecl+Med	116	84
Psic	19	13	Psic	32	50

Ejemplo: Origen de la esquizofrenia

- Los resultados obtenidos en el análisis de las sub-tablas 3 y 4 indican que existen diferencias entre el pensamiento de la escuela Psicoanalítica y las otras escuelas (Ecl + Med).
- Los miembros de la escuela Psicoanálitica son mucho más contundentes que las otras escuelas en atribuir el origen de la esquizofrenia al *Ambiente* (sub-tabla 3) y a una *Combinación* (sub-tabla 4).
- La suma de los cuatro componentes es $G^2 = 23{,}04$ que es el valor obtenido en la prueba de independencia de la tabla 5.

Reglas para particionar

- La suma de los grados de libertad de las subtablas deber ser igual a los grados de libertad de la tabla completa.
- Cada frecuencia observada en la tabla completa debe aparecer en solamente una de las sub-tablas.
- Cada total marginal en la tabla completa debe ser un total marginal en solamente una de las sub-tablas.
- La suma del G^2 en cada sub-tabla es igual al que se obtiene con la tabla completa. Sin embargo no siempre ocurre lo mismo cuando se utiliza el estadístico X^2.

Limitaciones de las pruebas chi-cuadrado

- Las pruebas de independencia chi-cuadrado solo indican el grado de evidencia que tiene la hipótesis alterna de asociación entre las variables.
- Requieren un tamaño de muestra grande.
- Los estadísticos X^2 y G^2 no cambian de valor cuando se reordenan las filas y columnas, lo cual indica que las variables son tratadas como nominales.
- El valor del X^2 depende del tamaño de muestra ya que:

$$X^2 = n \sum_i \sum_j \frac{(\hat{\pi}_{ij} - \hat{\pi}_{i+}\hat{\pi}_{+j})^2}{\hat{\pi}_{i+}\hat{\pi}_{+j}}$$

Prueba exacta de Fisher para tablas 2×2

- Las pruebas de independencia chi-cuadrado se basan en distribuciones asintóticas.
- Considere una tabla 2×2 donde se tiene una hipótesis nula de independencia, es decir $\theta = 1$.
- Suponga que los $\{n_{ij}\}$ provienen de dos muestras aleatorias independientes o de una única distribución multinomial definida sobre las 4 casillas en la tabla.
- Si se consideran todas las posibles tablas que tienen como totales por filas y columnas los valores observados entonces la distribución de las frecuencias es *hipergeométrica*.

Prueba exacta de Fisher para tablas 2×2

- Si condicionamos los totales por fila y columna solamente queda libre una frecuencia, por ejemplo n_{11}, de modo que ésta determina las otras tres frecuencias.
- Dados los totales marginales las tablas que tienen mayores valores de n_{11} también tienen mayores valores en su odds ratio:

$$\theta = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

- La prueba de independencia se puede plantear como:

$$\begin{aligned} H_0 &: \theta \leq 1 \\ H_1 &: \theta > 1 \end{aligned}$$

Ejemplo

- Considere un ejemplo clásico de Fisher. Un colega Bristol suyo afirmaba que era capaz de distinguir en una taza de té con leche, qué se había echado primero.
- Para comprobarlo diseñó un experimento donde se probaban 8 tazas de té. De ellas, en 4 se había echado primero la leche y en las otras 4 primero el té.
- Se trataba de adivinar en que orden se había echado la leche y el té.
- Las tasas se presentaron de manera aleatoria obteniendo los resultados que se muestran en la siguiente tabla.

Ejemplo

Tabla 7: *Experimento de Fisher*

	Predicción		
Primero	Leche	Té	Total
Leche	3	1	4
Té	1	3	4
Total	4	4	

- En base a los resultados obtenidos en este experimento Bristol convenció a Fisher sobre su habilidad en distinguir si se había echado primero el té o la leche.

Medidas basadas en X^2

- El *coeficiente* ϕ es:

$$\phi = \sqrt{\frac{X^2}{n}}$$

- El *coeficiente de contingencia* C es:

$$C = \sqrt{\frac{X^2}{X^2 + n}}$$

- El *coeficiente V de Crámer* es:

$$V = \sqrt{\frac{X^2}{n(k-1)}} \qquad k = \min(I, J)$$

Medidas basadas en la reducción proporcional del error

- Son apropiadas para estudios realizados bajo el modelo de homogeneidad.
- El *coeficiente lambda* de Goodman y Kruskal se define por:

$$\lambda_1 = \frac{\sum_j \max\{n_{ij}\} - \max\{n_{i+}\}}{n - \max\{n_{i+}\}}$$

si la variable respuesta se encuentra en fila, caso contrario:

$$\lambda_2 = \frac{\sum_i \max\{n_{ij}\} - \max\{n_{+j}\}}{n - \max\{n_{+j}\}}$$

Medidas basadas en la reducción proporcional del error

- Una versión simétrica del coeficiente, apropiada para el modelo de independencia, ya que no distingue ninguna variable respuesta es:

$$\lambda = \frac{\sum_j \max\{n_{ij}\} + \sum_i \max\{n_{ij}\} - \max\{n_{i+}\} - \max\{n_{+j}\}}{2n - \max\{n_{i+}\} - \max\{n_{+j}\}}$$

- El *coeficiente tau* de Goodman y Kruskal se define por:

$$\tau_1 = \frac{n \sum_i n_{i+}^{-1} \sum_j n_{ij}^2 - \sum_j n_{+j}^2}{n^2 - \sum_j n_{+j}^2}$$

si la variable respuesta se encuentra en fila.

Medidas de transmisión de información

- El índice de concentración de *Ghini* es:

$$\eta = 1 - \sum_j p_{+j}$$

- La medida de *entropía* es:

$$\eta_H = -\sum_j p_{+j} \log p_{+j}$$

- El *coeficiente de incertidumbre* es:

$$U = -\frac{\sum_i \sum_j p_{ij} \log(p_{ij}/p_{i+}p_{+j})}{\sum_j p_{+j} \log p_{+j}}$$

Introducción

- El objetivo en una prueba de asociación con variables ordinales es si existe un incremento en una variable como consecuencia de un incremento en la otra.
- Puesto que el concepto de linealidad de Pearson no es apropiado con variables ordinales el concepto que lo sustituye es *monoticidad*.
- Con variables cualitativas ordinales es posible observar una *tendencia monótona* en su comportamiento.
- Cuando se clasifican los individuos usando dos variables ordinales es posible determinar comparativamente si un par es *concordante*, *discordante* o *empatado*.

Ejemplo: Ingreso y nivel de satisfacción

Tabla 8: *Ingreso y satisfacción en el trabajo*

Ingreso (miles)	Satisfacción en el trabajo			
	Muy insatisfecho	Algo insatisfecho	Algo satisfecho	Muy satisfecho
< 15	1	3	10	6
15 - 25	2	3	10	7
25 - 40	1	6	14	12
≥ 40	0	1	9	11

Concordancias y discordancias

- Es de esperar que la satisfacción en el trabajo tienda a aumentar cuando el ingreso lo haga.
- Un par es *concordante* cuando uno de ellos se encuentra en un nivel más alto para X y Y en comparación al otro.
- Un par es *discordante* cuando uno de ellos se encuentra en un nivel más alto para X y en un nivel más bajo para Y.
- Un par esta *empatado* cuando ambos se encuentran en el mismo nivel para X y Y .
- El número total de pares concordantes es $C = 1331$.
- El número total de pares discordantes es $D = 849$.

Coeficiente Gamma

- El coeficiente *gamma* (Goodman y Kruskal 1954) es:

$$\hat{\gamma} = \frac{C - D}{C + D}$$

- Al igual que el coeficiente de correlación: $-1 \leq \gamma \leq 1$.
- El coeficiente gamma trata las variables de forma simétrica, es decir no es necesario identificar la variable respuesta.
- La condición de independencia implica que $\gamma = 0$, sin embargo el resultado anterior no implica la independencia.
- Para el ejemplo de la tabla 8:

$$\hat{\gamma} = \frac{1331 - 849}{1331 + 849} = 0{,}2211$$

τ-b de Kendall

- Sea T_X y T_Y el número de pares empatados en X y Y respectivamente.
- El τ-b de Kendall se define por:

$$\tau_b = \frac{C - D}{\left\{\left[\binom{n}{2} - T_X\right]\left[\binom{n}{2} - T_Y\right]\right\}^{1/2}}$$

- Para el ejemplo de la tabla 8:

$$\tau_b = \frac{1331 - 849}{\{[4560 - 1159]\,[4560 - 1617]\}^{1/2}} = 0{,}1524$$

τ-c de Kendall

- El τ-c de Kendall se define por:

$$\tau_c = \frac{2k(C-D)}{n^2(k-1)}$$

donde k es el menor número de casos no empatados que hay en X y Y .

- Para el ejemplo de la tabla 8 se tiene $k = \min\{3401, 2942\}$:

$$\tau_c = \frac{2(2942)(1331-849)}{96^2(2942-1)} = 0{,}1046$$

Prueba de Mantel-Haenszel

- Cuando las variables X y Y son ordinales es común observar una tendencia positiva o negativa en la asociación.
- Se asignan scores a las categorías para resumir la tendencia lineal.
- Una prueba estadística que es sensible a las tendencias lineales positivas o negativas utilizan información sobre la *correlación*.
- Sean $u_1 \leq u_2 \leq \cdots \leq u_I$ y $v_1 \leq v_2 \leq \cdots \leq v_J$ los scores para filas y columnas respectivamente.
- Los scores presentan mayores diferencias entre las categorías más separadas.

Prueba de Mantel-Haenszel

- Una vez elegidos los scores se calcula el coeficiente de correlación r entre X y Y:

$$r = \frac{\sum_i \sum_j n_{ij} u_i v_j - n\bar{u}\,\bar{v}}{S_u S_v}$$

- Una estadística para probar la independencia versus $r \neq 0$ es:

$$M^2 = (n-1)\, r^2$$

- Si n es grande M^2 tiene una distribución chi-cuadrado con 1 grado de libertad.

Ejemplo: Satisfacción en el trabajo

- Las pruebas chi-cuadrado de independencia $X^2 = 6{,}0$ con $p = 0,74$ y $G^2 = 6{,}8$ con $p = 0{,}66$ presentan poca evidencia de asociación pero ignoran la naturaleza ordinal de las variables.
- Suponga que los scores para la satisfacción en el trabajo y el ingreso son (1, 2, 3, 4) y (7.5, 20, 32.5, 60) respectivamente.
- Usando los scores anteriores la correlación es $r = 0{,}20$.
- La prueba estadística de tendencia lineal $M^2 = 3{,}81$ con $p = 0{,}051$ y muestra cierta evidencia de asociación.

Ejemplo: Satisfacción en el trabajo

- La evidencia es mucho más fuerte para la hipótesis alterna de una cola (tendencia positiva) usando $M = \sqrt{n-1}r = 1{,}95$ con $p = 0{,}026$.
- La evidencia no trivial de asociación positiva podria sorprender ya que X^2 y G^2 no tienen valores de probabilidad significativos.
- Cuando existe una tendencia lineal, el análisis usado en su detección puede proporcionar valores de probabilidad p mucho más pequeños que los obtenidos de un análisis que ignora dicha tendencia.