Inferencia Estadística

Marisol García Peña

Departamento de Matemáticas
Pontificia Universidad Javeriana

Bogotá, 2022

Pruebas chi-cuadrado

Distribución asintótica de la razón de verosimilitud generalizada

Teorema

Sea $X_1, \ldots, X_n$ es una muestra aleatoria con función de densidad conjunta $f(\bullet, \ldots, \bullet; \theta)$, donde $\theta = (\theta_1, \ldots, \theta_k)$, que se asume que satisface las condiciones de regularidad. Suponga que el espacio del parámetro Θ es de dimensión k. Para probar la hipótesis $H_0 : \theta_1 = \theta_1^0, \ldots, \theta_r = \theta_r^0, \theta_{r+1}, \ldots, \theta_k$, donde $\theta_1^0, \ldots, \theta_r^0$ son conocidos y $\theta_{r+1}, \ldots, \theta_k$ no son especificados, $-2 \log \Lambda_n$ es aproximadamente distribuida como chi-cuadrado con r g.l. cuando H_0 es verdadera y la muestra de tamaño n es grande.

Pruebas chi-cuadrado (cont.)

- En el teorema se asume que $1 \leq r \leq k$.
- Si $r = k$ todos los parámetros son especificados.
- El espacio del parámetro Θ es de dimensión k y como H_0 especifica el valor de r de los componentes de $(\theta_1, \ldots, \theta_k)$, la dimensión de Θ_0 es $k - r$.
- Los grados de libertad de la distribución asintótica chi-cuadrado puede verse como: el número de parámetros especificados por H_0 o como la diferencia en dimensiones de Θ y Θ_0.

Pruebas chi-cuadrado (cont.)

- Λ_n es la variable aleatoria que tiene valores

$$\lambda_n = \frac{\sup_{\Theta_0} L(\theta_1, \ldots, \theta_k; x_1, \ldots, x_n)}{\sup_{\Theta} L(\theta_1, \ldots, \theta_k; x_1, \ldots, x_n)}$$

- Corresponde a la razón de verosimilitud generalizada para una muestra de tamaño n.

- El principio de la razón de verosimilitud generalizada indica: rechazar H_0 para λ_n pequeño, pero $-2\log\lambda_n$ aumenta cuando λ_n decrece $\Longrightarrow$ prueba equivalente, rechazar H_0 para $-2\log\lambda_n$ grande.

- Usando la distribución asintótica, la prueba es:
Rechazar H_0 si y sólo si $-2\log\lambda_n > \chi^2_{(1-\alpha,r)}$

Pruebas chi-cuadrado (cont.)

Prueba chi-cuadrado de bondad de ajuste

- Si la población tiene distribución multinomial $f(x_1, \ldots, x_n; p_1, \ldots, p_k) = \prod_{j=1}^{k+1} p_j^{x_j}$, con $x_j = 0$ o 1, $j = 1, \ldots, k+1$; $0 \leq p_j \leq 1$, $j = 1, \ldots, k+1$; $\sum_{j=1}^{k+1} x_j = 1$ y $\sum_{j=1}^{k+1} p_j = 1$.

- Esto si fuera muestreo con reemplazo de una población de individuos que fueron clasificados en $k+1$ clases/categorías.

- Un problema común es probar si las probabilidades p_j tienen valores numéricos específicos.

- Por ejemplo, el resultado de lanzar un dado puede ser clasificado en una de 6 clases y con la información de una muestra de observaciones, se puede probar si el dado es verdadero, es decir, si $p_j = \frac{1}{6}, j = 1, \ldots, 6$.

Pruebas chi-cuadrado (cont.)

- También se puede pensar en términos de independencia, ensayos repetidos en donde cada ensayo resulta en uno de los $k+1$ resultados, llamados clases/categorías.
- En este caso la densidad de la multionmial proporciona la densidad para los resultados de un ensayo.
- El resultado de un ensayo $\Longrightarrow$ v.a. multivariada $(X_1, \ldots, X_k)$ donde X_j es 1 si el ensayo tiene como resultado la categoría j y 0 sino.
- p_j es la probabilidad de que el ensayo tenga como resultado la categoría j.

Pruebas chi-cuadrado (cont.)

- Si se repite en ensayo n veces $\Longrightarrow$ n observaciones de la v.a. multivariada $(X_1, \ldots, X_k)$, es decir, $(X_{11}, \ldots, X_{1k}), (X_{21}, \ldots, X_{2k}), \ldots, (X_{n1}, \ldots, X_{nk})$.
- $N_j = \sum_{i=1}^{n} X_{ij}$, la v.a. N_j es el número de n ensayos que tienen como resultado la categoría $j \Longrightarrow (N_1, \ldots, N_k)$ tiene distr. multinomial.
- $H_0 : p_j = p_j^0, j = 1, \ldots, k+1$, donde p_j^0 son las probabildiades dadas que suman 1 $\Longrightarrow$ Usar principio de la razón de verosimilitud generalizada.
- $L = L(p_1, \ldots, p_k; x_{11}, \ldots, x_{1k}, \ldots, x_{n1}, \ldots, x_{nk}) = \prod_{i=1}^{n} \prod_{j=1}^{k+1} p_j^{x_{ij}}$.

Pruebas chi-cuadrado (cont.)

- Espacio del parámetro Θ tiene k dimensiones (k de los $k+1$ p_j, el restante es determinado por $\sum p_j = 1$) y Θ_0 es un punto.
- L se maximiza en Θ cuando $p_j = \sum_{i=1}^{n} \frac{x_{ij}}{n} = \frac{n_j}{n}$, con n_j un valor de la v.a N_j.
- Entonces $\sup_{\Theta} L = \frac{1}{n^n} \prod_{j=1}^{k+1} n_j^{n_j}$.
- El máximo de L sobre Θ_0 solo es un valor $\prod_{j=1}^{k+1} (p_j^0)^{n_j}$ y la razón de verosimilitud generalizada es $\lambda = n^n \prod_{j=1}^{k+1} \left(\frac{p_j^0}{n_j} \right)^{n_j}$.

Pruebas chi-cuadrado (cont.)

- Una prueba de razón de verosimilitud generalizada es: Rechazar H_0 si y sólo si $\lambda < \lambda_0$, λ_0 se escoge para obtener una probabilidad de error tipo I deseada (α, tamaño de la prueba).

- Para n pequeño, la distribución de la prueba de razón de verosimilitud generalizada puede tabularse para determinar λ_0, para tamaños grandes de n puede usarse el teorema que establece que $-2\log\Lambda$ tiene aproximadamente distribución chi-cuadrado con k g.l.

- La aproximación chi-cuadrado es buena incluso si n es pequeña proporcionando $k > 2$.

Pruebas chi-cuadrado (cont.)

- Otra prueba que se usa para H_0 fue propuesto por Karl Pearson.
- La prueba usa la estadística $Q_k^0 = \sum_{j=1}^{k+1} \frac{(N_j - np_j^0)^2}{np_j^0}$.
- Tiende a se pequeña cunado H_0 es verdadera y grande cuando H_0 es falsa.
- N_j es el número de ensayos observados que tienen como resultado la categoría j y np_j^0 es el número esperado cuando H_0 es verdadera.
- $E[Q_k^0] = \sum_{j=1}^{k+1} \frac{1}{np_j^0}[np_j(1-p_j) + n^2(p_j - p_j^0)^2]$, donde p_j son los verdaderos parámetros. Si H_0 es verdadera, entonces $E[Q_k^0] = \sum(1 - p_j^0) = k + 1 - 1 = k$

Pruebas chi-cuadrado (cont.)

Teorema

Si los posibles resultados de cierto experimento aleatorio se descomponen en $k+1$ conjuntos mutuamente excluyentes, $A_1, \ldots, A_{k+1}$. Se define $p_j = P[A_j], j = 1, \ldots, k+1$. En n repeticiones independientes de un experimento aleatorio, sea N_j que denota el número de rsultados que pertenecen al conjunto $A_j, j = 1, \ldots, k+1$, tal que $\sum_{j=1}^{k+1} N_j = n$. Entonces

$$Q_k = \sum_{j=1}^{k+1} \frac{(N_j - np_j)^2}{np_j}$$

tiene como distribución límite cuando n tiende a infinito, la distribución chi-cuadrado con k grados de libertad.

Pruebas chi-cuadrado (cont.)

- El teorema establece la distribución límite para la estadística

$$Q_k^0 = \sum_{j=1}^{k+1} \frac{(N_j - np_j^0)^2}{np_j^0}$$

cuando $H_0 : p_j = p_j^0, j = 1, \ldots, k+1$ es verdadera.

- Prueba con tamaño α: Rechazar H_0 si y sólo si $Q_k^0 > \chi^2_{(1-\alpha,k)}$.

Pruebas chi-cuadrado (cont.)

Ejemplo

La teoría mendeliana indica que la forma y color de una cierta variedad de arveja debe ser agrupada en 4 grupos, "redonda y amarilla", "redonda y verde", "angular y amarilla" y "angular y verde", de acuerdo a las razones 9/3/3/1. Para $n = 556$ arvejas, se observa lo siguiente

Categoría	V.Observado	V.Esperado
Redonda y amarilla	315	312.75
Redonda y verde	108	104.25
Angular y amarilla	101	104.25
Angular y verde	32	34.75

Valor esperado$= 556\left(\frac{9}{16}\right)$.

Pruebas chi-cuadrado (cont.)

- Una prueba de tamaño 0.05 para $H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$ es:
  Rechazar H_0 si y sólo si $Q_3^0 = \sum_{j=1}^{4} \frac{(N_j - np_j^0)^2}{np_j^0} > \chi^2_{(1-\alpha;k)}$.

- $\chi^2_{(0.95;3)} = 7.81$.

- $Q_3^0 = \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(101 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} = 0.470$

- No hay evidencia suficiente para rechazar H_0, es decir, hay un buen ajuste de los datos al modelo.

Pruebas chi-cuadrado (cont.)

Teorema

Si los posibles resultados de un cierto experimento aleatorio se descomponen en $k+1$ conjuntos mutuamente excluyentes, $A_1, \ldots, A_{k+1}$. Se define $p_j = P[A_j], j = 1, \ldots, k+1$ y se asume que p_j depende de r parámetros desconocidos $\theta_1, \ldots, \theta_r$, tal que, $p_j = p_j(\theta_1, \ldots, \theta_r), j = 1, \ldots, k+1$. En n repeticiones del experimento aleatorio, sea N_j el número de resultados que pertenecen al conjunto $A_j, j = 1, \ldots, k+1$ y $\sum_{j=1}^{k+1} N_j = n$. Sean $\widehat{\Theta}_1, \ldots, \widehat{\Theta}_r$ estimadores BAN - mejores estimadores asintóticamente normales (estimadores de máxima versimilitud) de $(\theta_1, \ldots, \theta_r)$ basados en $N_1, \ldots, N_k$. Bajo condiciones de regularidad sobre p_j's

$$Q_k' = \sum_{j=1}^{k+1} \frac{(N_j - n\widehat{P}_j)^2}{n\widehat{P}_j}$$

tiene distribución límite chi-cuadrado con $k - r$ g.l., donde $\widehat{P}_j = p_j(\widehat{\Theta}_1, \ldots, \widehat{\Theta}_r), j = 1, \ldots, k+1$

Pruebas chi-cuadrado (cont.)

- El teorema no menciona cuál es la hipótesis que se prueba.
- $H_0 : X_i$ tiene densidad $f(x; \theta_1, \ldots, \theta_r)$.
- Una prueba para H_0 puede ser: Rechazar H_0 si y sólo si Q'_k es grande, es decir, rechazar H_0 si y sólo si $Q'_k > \chi^2_{(1-\alpha, k-r)}$.
- Si las observaciones $X_1, \ldots, X_n$ están disponibles, puede estimarse θ_i, $i = 1, \ldots, r$ usando los EMV, la distribución límite de Q'_k está entre/limitada por una chi-cuadrado con $k - r$ g.l. y una chi-cuadrado con k g.l.
- Los r grados de libertad son recuperados al estimar eficientemente $\theta_1, \ldots, \theta_r$.

Pruebas chi-cuadrado (cont.)

Ejemplo

Suponga que se quiere probar la hipótesis nula de que una muestra aleatoria observada $x_1, \ldots, x_n$ ha sido seleccionada de una población normal.

Si los n valores $x_1, \ldots, x_n$ son agrupados en $k+1$ clases. Por ejemplo, la j-ésima clase puede tomarse como todas las observaciones que están en el intervalo $(z_{j-1}, z_j], j = 1, \ldots, k+1$, para algún $z_0 < z_1 < z_2 < \cdots < z_k < z_{k+1}$, donde $z_0 = -\infty$ y $z_{k+1} = +\infty$. Entonces

$$p_j = p_j(\mu, \sigma^2) = \int_{z_j - 1}^{z_j} \phi_{\mu,\sigma^2}(x)dx = \Phi\left(\frac{z_j - \mu}{\sigma}\right) - \Phi\left(\frac{z_{j-1} - \mu}{\sigma}\right)$$

$\widehat{\mu}$ y $\widehat{\sigma}$ son EMV basados en $n_1, \ldots, n_k$, donde n_j es el número de observaciones que están en el j-ésimo intervalo.

Pruebas chi-cuadrado (cont.)

Entonces,

$$\widehat{p}_j = \Phi\left(\frac{z_j - \widehat{\mu}}{\widehat{\sigma}}\right) - \Phi\left(\frac{z_{j-1} - \widehat{\mu}}{\widehat{\sigma}}\right)$$

se puede determinat a partir de la muestra y el valor

$$q'_k = \sum_{j=1}^{k+1} \frac{(n_j - n\widehat{p}_j)^2}{n\widehat{p}_j}$$

de Q'_k también se puede obtener a partir de la muestra. La prueba: Rechazar H_0 al nivel α si $q'_k > \chi^2_{(1-\alpha;k-2)}$.

Si μ y σ son los EMV basados en $X_1, \ldots, X_n$ entonces la distribución asintótica de Q'_k estará entre una $\chi^2_{(k-2)}$ y una $\chi^2_{(k)}$, luego la prueba será: rechazar H_0 si $q'_k > C$, donde C está entre $\chi^2_{(1-\alpha;k-2)}$ y una $\chi^2_{(1-\alpha;k)}$, si k es grande la diferencia entre los percentiles es pequeña.

Pruebas chi-cuadrado (cont.)

Prueba de igualdad de 2 distribuciones multinomiales y generalizaciones

- Prueba de hipótesis de que dos poblaciones multinomiales pueden ser consideradas la misma.
- Se supone que hay $k+1$ grupos asociados con cada una de las poblaciones multinomiales.
- Primera polación tiene prob. asociadas $p_{11}, p_{12}, \ldots, p_{1k+1}$ y la segunda $p_{21}, p_{22}, \ldots, p_{2k}, p_{2k+1}$.
- $H_0 : p_{1j} = p_{2j} (= p_j), j = 1, \ldots, k+1$

Pruebas chi-cuadrado (cont.)

- Muestra de tamaño n_1 de la primera población, N_{1j} número de resultados en el grupo $j, j = 1, \ldots, k1$.
- Muestra de tamaño n_2 de la primera población, N_{2j} número de resultados en el grupo $j, j = 1, \ldots, k1$.
- $\sum_{j=1}^{k+1} \frac{(N_{ij} - n_i p_{ij})^2}{n_i p_{ij}}$ tiene distribución límite $\chi^2_{(k)}$ para $i = 1, 2$.
- $\sum_{i=1}^{2} \sum_{j=1}^{k+1} \frac{(N_{ij} - n_i p_{ij})^2}{n_i p_{ij}}$ tiene distribución límite $\chi^2_{(2k)}$ si las dos muestras sin independientes.

Pruebas chi-cuadrado (cont.)

- Si H_0 es verdadera, entonces $Q_{2k} = \sum_{i=1}^{2} \sum_{j=1}^{k+1} \frac{(N_{ij} - n_i p_j)^2}{n_i p_j}$ tiene distribución límite $\chi^2_{(2k)}$.
- Si H_0 especifica los valores p_j, entonces Q_{2k} es la estadística y se puede usar para la prueba.
- Si p_j en H_0 son desconocidas, deben ser estimadas.
- Si H_0 es verdadera $\Longrightarrow$ dos muestras se pueden considerar como una muestra de tamaño $n_1 + n_2$ de una multinomial con probabilidades $p_1, \ldots, p_{k+1}$.

Pruebas chi-cuadrado (cont.)

- Estimadores de máxima verosimilitud de p_j son $\frac{N_{1j}+N_{2j}}{n_1+n_2}, j = 1, \ldots, k$.
- $Q'_{2k} = \sum_{i=1}^{2} \sum_{j=1}^{k+1} \frac{[N_{ij} - n_i(N_{1j} + N_{2j})/(n_1 + n_2)]^2}{n_i(N_{1j} + N_{2j})/(n_1 + n_2)} \sim \chi^2_{(2k-k=k)}$.
- Los g.l. se reducen en 1 por cada parámetro estimado.
- Otra prueba $\Longrightarrow$ razón de verosimilitud generalizada Λ y obtener la distribución de $-2 \log \Lambda$.

Pruebas chi-cuadrado (cont.)

Ejemplo

En una encuesta de opinión sobre un tema político, hay una pregunta sobre si los votantes menores de 25 años podrían ver el tema de manera diferente a los mayores de 25. Se entrevistaron 1500 personas mayores de 25 años y 1000 menores de 25 años con los siguientes resultados. Pruebe la hipótesis nula de que no hay evidencia de diferencia de opinión debido al grupo de edad, es decir, $H_0 : p_{1j} = p_{2j} = p_j, j = 1, 2, 3.$

	En contra	Indeciso	A favor	Total
<25	400	100	500	1000
>25	600	400	500	1500
Total	1000	500	1000	2500

Pruebas chi-cuadrado (cont.)

p_1 y p_2 deben ser estimados.

$$\begin{aligned} Q_2' =& \frac{(400 - 1000(1000)/2500)^2}{1000(1000)/2500} + \frac{(100 - 1000(500)/2500)^2}{1000(500)/2500} \\ &+ \frac{(500 - 1000(1000)/2500)^2}{1000(1000)/2500} + \frac{(600 - 1500(1000)/2500)^2}{1500(1000)/2500} \\ &+ \frac{(400 - 1500(500)/2500)^2}{1500(500)/2500} + \frac{(500 - 1500(1000)/2500)^2}{1500(1000)/2500} \\ =& 125 \end{aligned}$$

El percentil 99 de la distribución $\chi^2_{(0.99;2)} = 9.21$, $2 = 2k - k$ g.l., $k = 2, k+1$ categorías. Entonces, existe fuerte evidencia de que los dos grupos tienen diferentes opiniones sobre el tema político.

Pruebas chi-cuadrado (cont.)

Ejemplo

100 observaciones fueron seleccionadas de dos poblaciones Poisson. ¿Existe evidencia sólida en los datos para apoyar la afirmación de que las dos poblaciones de Poisson son diferentes?

	0	1	2	3	4	5	6	7	8	≥ 9	Total
Pop.1	11	25	28	20	9	3	3	0	1	0	100
Pop.2	13	27	28	17	11	1	2	1	0	0	100
Total	24	52	56	37	20			11			200

Pruebas chi-cuadrado (cont.)

- Se agrupan lo datos en 6 categorías, el último incluye todos los mayores a 4.
- Si las dos poblaciones son iguales $\Longrightarrow$ estimar un parámetro, la media de la distr. Poisson en común.
- El EMV es la media muestral.

$$\widehat{\lambda} = \frac{0(24) + 1(52) + 2(56) + 3(37) + 4(20) + 5(4) + 6(5) + 7(1) + 8(1)}{200}$$

$$= \frac{420}{200} = 2.1$$

- Número esperado en cada grupo de cada población

0	1	2	3	4	>5
12.25	25.72	27.00	18.90	9.92	6.21

- Como se trata de una Poisson, el número esperado es $100P[X = 0] = 100(0.1225)$.

Pruebas chi-cuadrado (cont.)

- $Q_{(2k-1)} = 1.68$, 1 parámetro estimado, $2k - 1 = 9$ g.l., $k = 5$.
- Para el cálculo de Q, $\frac{(11-12.25)^2}{12.25} + \frac{(25-25.72)^2}{25.72} + \cdots + \frac{(4-6.21)^2}{6.21}$.
- El percentil 99 de la distribución $\chi^2_{(0.99;9)} = 21.67$, luego no hay razón para sospechar que las dos poblaciones Poisson sean diferentes.

Pruebas chi-cuadrado (cont.)

- Tabla de contingencia $\Longrightarrow$ clasificación múltiple.
- Los individuos se clasifican por dos criterios/variables que tienen distinto número de categorías $\Longrightarrow$ tablas a dos vías.
- Las clasificaciones resultantes se conocen como celdas.
- Tabla a tres vías $\Longrightarrow$ los individuos son clasificados de acuerdo a tres criterios/variables.
- Las tablas de contingencia porporcionan una presentación conveniente de los datos y poder investigar la posible relación.

Pruebas chi-cuadrado (cont.)

- En tablas a dos vías $\Longrightarrow$ H_0 : las variables son independientes.
- En tablas a tres vías $\Longrightarrow$ H_0 : la variable de clasificación (1) es independiente de las variables (2) y (3).
- Es útil en cualquier campo de investigación.

Pruebas chi-cuadrado (cont.)

Tablas de contingencia a dos vías

- n individuos/objetos son clasificados de acuerdo a 2 criterios/variables A y B.
- $A_1, \ldots, A_r$ en A y $B_1, \ldots, B_s$ en B, el número de individuos/objetos que pertenecen a A_i y B_j es N_{ij}.
- Tabla de contingencia de $r \times s$, con frecuencias de celda N_{ij} y $\sum N_{ij} = n$.

	B_1	B_2	B_3	$\cdots$	B_s
A_1	N_{11}	N_{12}	N_{13}	$\cdots$	N_{1s}
A_2	N_{21}	N_{22}	N_{23}	$\cdots$	N_{2s}
A_3	N_{31}	N_{32}	N_{33}	$\cdots$	N_{3s}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	N_{r1}	N_{r2}	N_{r3}	$\cdots$	N_{rs}

Pruebas chi-cuadrado (cont.)

- Totales fila $N_{i\bullet} = \sum_j N_{ij}$ y totales columna $N_{\bullet j} = \sum_i N_{ij}$.
- $\sum_i N_{i\bullet} = \sum_j N_{\bullet j} = n$.
- Los n individuos son considerados como una muestra de tamaño n de una distribución multinomial con probabilidades $p_{ij} (i = 1, \ldots, r; j = 1, \ldots, s)$.
- La función de densidad de probabilidad para una observación es $f(x_{11}, \ldots, x_{rs}; p_{11}, \ldots, p_{rs}) = \prod_{i,j} p_{ij}^{x_{ij}}$, donde $x_{ij} = 0, 1$ y $\sum_{i,j} x_{ij} = 1$

Pruebas chi-cuadrado (cont.)

- H_0 : A y B son independientes, es decir, la prob. de que un individuo caiga en B_j no es afectada por la categoría A a la cual pertenece el individuo.
- $P[B_j|A_i] = P[B_j]$ y $P[A_i|B_j] = P[A_i]$ o $P[A_i \bigcap B_j] = P[A_i]P[B_j]$.
- Denotando las probabilidades marginales $P[A_i]$ como $p_{i\bullet}(i = 1, \ldots, r)$ y las probabilidades marginales $P[B_j]$ como $p_{\bullet j}(j = 1, \ldots, s) \Longrightarrow H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}, \sum p_{i\bullet} = 1, \sum p_{\bullet j} = 1$.
- Si la hipótesis nula es falsa $\Longrightarrow$ interacción/dependencia entre las variables de clasificación.

Pruebas chi-cuadrado (cont.)

- Espacio del parámetro Θ para la distribución de $N_{11}, \ldots, N_{rs}$ tiene $rs - 1$ dimensiones, especificando todas menos una de las p_{ij}, la restante se fija usando $\sum_{i,j} p_{ij} = 1$.
- Bajo H_0 el espacio del parámetro Θ_0 tiene $r - 1 + s - 1$ dimensiones, H_0 es especificada por $p_{i\bullet}, i = 1, \ldots, r$ y $p_{\bullet j}, j = 1, \ldots, s$ pero solo hay $r - 1 + s - 1$ dimensiones porque $\sum p_{i\bullet} = 1$ y $\sum p_{\bullet j} = 1$.
- La verosimilitud para una muestra de tamaño $n \Longrightarrow L = \prod_{i,j} p_{ij}^{n_{ij}}$.
- Máximo en Θ ocurre cuando $\widehat{p}_{ij} = \frac{n_{ij}}{n}$.
- En Θ_0, $L = \prod_{i,j} (p_{i\bullet} p_{\bullet j})^{n_{ij}} = \left(\prod_i p_{i\bullet}^{n_{i\bullet}} \right) \left(\prod_j p_{\bullet j}^{n_{\bullet j}} \right)$ y máximo en $\widehat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}$ y $\widehat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}$.

Pruebas chi-cuadrado (cont.)

- Razón de verosimilitud generalizada, $\lambda = \dfrac{\left(\prod_i n_{i\bullet}^{n_{i\bullet}}\right)\left(\prod_j n_{\bullet j}^{n_{\bullet j}}\right)}{n^n \prod_{i,j} n_{ij}^{n_{ij}}}$.

- La distribución de Λ bajo H_0 no es única porque la hipótesis es compuesta y la distribución exacta de Λ involucra los parámetros desconocidos $p_{i\bullet}$ y $p_{\bullet j}$.

- Es díficil resolver para λ_0 en $\sup_{\Theta_0} P[\Lambda \leq \lambda_0] = \alpha$.

- Para muestras grandes, se tiene la prueba porque $-2\log\Lambda$ es aproximadamente una $\chi^2_{(rs-1-(r+s-2)=(r-1)(s-1))}$.

Pruebas chi-cuadrado (cont.)

- Puede usarse otra prueba para $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}, \sum p_{i\bullet} = 1, \sum p_{\bullet j} = 1$.
- $Q = \sum_{i,j} \frac{[N_{ij} - n(N_{i\bullet}/n)(N_{\bullet j}/n)]^2}{n(N_{i\bullet}/n)(N_{\bullet j}/n)} \sim \chi^2_{(rs-1-(r-1+s-1)=(r-1)(s-1))}$.
- Rechazar H_0 para Q grande. Propuesto por Karl Pearson.
- Difiere de $-2\log\lambda$ en términos de orden $1/\sqrt{n}$, las dos pruebas son equivalentes a menos que n sea pequeño.
- N_{ij} es el número observado en la celda ij y $n(N_{i\bullet}/n)(N_{\bullet j}/n)$ es el estimador del número esperado en la celda ij cuando H_0 es verdadera.

Pruebas chi-cuadrado (cont.)

Ejemplo

Localización de política de privacidad y nacionalidad del sitio en la web. Política de privacidad. Reglas que definen los grandes distribuidores "online" relacionadas con la información recolectada de los usuarios (a través de programas).

	Nacionalidad del sitio web			
Localización política	Francia	Reino Unido	EUA	Total Fila
Pág.principal	56	68	35	159
Pág.pedido	19	19	28	66
Pág.cliente	6	10	16	32
Otra pág.	12	9	13	34
Total col.	93	106	92	291

Políticas de privacidad

- Paso 1. Formular hipótesis.
  H_0: La posición de la política de privacidad es independiente de la nacionalidad de sitio en la web.
  H_1: La posición de la política de privacidad depende de la nacionalidad del sitio en la web.

- Paso 2. Construir la regla de decisión.
  Se tienen $r = 4$ filas y $s = 3$ columnas $\Longrightarrow v = (r-1)(s-1) = 3(2) = 6$g.l.
  χ^2_{tabla} con 6g.l. y $\alpha = 0.05 \Longrightarrow \chi^2_{tabla} = 12.59$. Valor crítico.
  Rechazar H_0 si $Q > 12.59$, caso contrario no se rechaza.

Pruebas chi-cuadrado (cont.)

- Paso 3. Calcular las frecuencias esperadas. $n\frac{N_{i\bullet}N_{\bullet j}}{n^2}$.

Localización	Nacionalidad Francia	Reino Unido	EUA	Total
Principal	$\frac{159(93)}{291} = 50.81$	$\frac{159(106)}{291} = 57.92$	$\frac{159(92)}{291} = 50.27$	159
Pedido	$\frac{66(93)}{291} = 21.09$	$\frac{66(106)}{291} = 24.04$	$\frac{66(92)}{291} = 20.87$	66
Cliente	$\frac{32(93)}{291} = 10.23$	$\frac{32(106)}{291} = 11.66$	$\frac{32(92)}{291} = 10.12$	32
Otra	$\frac{34(93)}{291} = 10.87$	$\frac{34(106)}{291} = 12.38$	$\frac{34(92)}{291} = 10.75$	34
Total	93	106	92	291

- Paso 4. Calcular la estadística de prueba.

$$Q = \frac{(56 - 50.81)^2}{50.81} + \cdots + \frac{(13 - 10.75)^2}{10.75}$$
$$= 0.53 + \cdots + 0.74 = 17.54$$

Pruebas chi-cuadrado (cont.)

- Paso 5. Tomar la decisión.
  Dado que $Q = 17.54 > \chi^2_{tabla} = 12.59$, se rechaza H_0, la diferencia observada entre las frecuencias esperadas y observadas son significativas para $\alpha = 0.05$.
  La posición de la política de privacidad está asociada a la nacionalidad para $\alpha = 0.05$, con base en la muestra de 291 sitios en la web.

- Regla de Cochran $\Longrightarrow$ frecuencias esperadas> 5 para todas las celdas y realizar prueba chi-cuadrado.

Tablas de contingencia a tres vías

- n individuos/objetos son clasificados de acuerdo a 3 criterios/variables A, B y C.
- $A_i, (i = 1, \ldots, s_1)$, $B_j, (j = 1, \ldots, s_2)$ y $C_k, (k = 1, \ldots, s_3)$, el número de individuos/objetos que pertenecen a A_i, B_j y C_k es N_{ijk}.
- Tabla de contingencia de $s_1 \times s_2 \times s_3$, con frecuencias de celda N_{ijk} y $\sum N_{ijk} = n$.
- p_{ijk} representa las proabilidades asociadas con las celdas individuales y los totales marginales $N_{i\bullet k} = \sum_{j=1}^{s_2} N_{ijk}$ y $N_{\bullet\bullet k} = \sum_{i=1}^{s_1} \sum_{j=1}^{s_2} N_{ijk}$.

Pruebas chi-cuadrado (cont.)

- $H_0 : p_{ijk} = p_{i\bullet\bullet}p_{\bullet j\bullet}p_{\bullet\bullet k}$, las tres variables/criterios son mutuamente independientes, donde $p_{i\bullet\bullet} = \sum_j \sum_k p_{ijk}$, $p_{\bullet j\bullet} = \sum_i \sum_k p_{ijk}$ y $p_{\bullet\bullet k} = \sum_i \sum_j p_{ijk}$.

- También se puede probar si uno de los tres criterios es independiente de los otros dos. Por ejemplo, si B es independiente de A y C, $H_0 : p_{ijk} = p_{i\bullet k}p_{\bullet j\bullet}$, con $p_{i\bullet k} = \sum_j p_{ijk}$.

- El proceso para probar estas hipótesis es análogo al de las tablas a dos vías.

- Para probar $H_0 : A$ y C son independientes de B, se usa la estadística
$$Q = \sum_i \sum_j \sum_k \frac{[N_{ijk} - n(N_{i\bullet k}/n)(N_{\bullet j\bullet}/n)]^2}{n(N_{i\bullet k}/n)(N_{\bullet j\bullet}/n)}.$$

- $\chi^2_{(s_1 s_2 s_3 - 1 - (s_1 s_3 - 1) - (s_2 - 1) = (s_1 s_3 - 1)(s_2 - 1))}$.