Modelo Lineal Generalizado

2do. Cuatrimestre de 2015

Nuestro objetivo será estudiar la relación entre dos o más variables, que podrán ser tanto continuas como categóricas.

En algunos casos diferenciaremos entre variable de respuesta y variables explicativas, mientras que en otros, simplemente, nos interesará estudiar la asociación entre las variables presentes sin hacer esta distinción.

A diferencia de lo que se trata habitualmente en Modelo Lineal, la variable de respuesta podrá ser categórica.

Abordaremos tres grandes temas:

- Tablas de Contingencia
- Modelo Lineal Generalizado
- Modelos Log-lineales

Veremos algunos ejemplos que introduzcan estos temas.

Consideremos el caso en el una muestra de 980 norteamericanos fue clasificada de acuerdo con el género y su identificación político-partidaria. En esta situación nos **interesa estudiar si hay asociación o no entre las variables categóricas**

G: Género

y

C:Identificación partidaria.

Esta es una tablas de contingencia bastante sencilla.

	C: Identificación partidaria			
G: Género	Demócrata	Independiente	Republicano	Total
Femenino	279	73	225	577
Masculino	165	47	191	403
Total	444	120	416	980

Cuadro 1: General Social Survey, 1991

Para responder a esto, en primera instancia, veremos los test de independencia o de homogeneidad basados en la distribución χ^2 que fueron introducidos por Pearson.

Sin embargo, estos tests, como muchos otros, tienen algunas limitaciones. Una de ellas es que si bien nos indican cuanta evidencia de asociación entre las variables de interés existe, no nos dicen nada sobre la naturaleza de esta relación.

Para comprender más profundamente la asociación entre variables nos ayudarán los modelos log-lineales y los modelos lineales generalizados, siendo estos últimos una generalización del modelo lineal habitual.

Ejemplo: Datos del Titanic

Estos datos tienen como fuente

http://www.encyclopedia-titanica.org/

Las variables consideradas son

Name: Nombre del Pasajero

PClass: Clase del Pasajero

Age: Edad del Pasajero

Sex: Género del Pasajero

Survived: Survived=1 el Pasajero sobrevivió

Survived=0 el Pasajero no sobrevivió

```
   titanic <- read.table("c:\\users\\ana\\glm\\titanic.txt", header = T)
attach(titanic)
names(titanic)
"Name"     "PClass"   "Age"      "Sex"      "Survived"

length(Age)
[1] 1313

table(PClass)
PClass
1st 2nd 3rd
322 280 711

table(Sex)
Sex
female   male
   462    851

 table(Survived)
Survived
  0   1
863 450
```

```
table(Survived,PClass)
        PClass
Survived 1st 2nd 3rd
       0 129 161 573
       1 193 119 138

table(Survived,Sex)
        Sex
Survived female male
       0    154  709
       1    308  142
```

El objetivo es describir la asociación entre las variables presentes. Por ejemplo, de estas dos últimas tablas podríamos decir que los pasajeros hombres y los de tercera clase sobrevivieron menos.

Aquí consideramos a *Survival* como variable de respuesta y a las demás como predictoras. En este caso, las variables predictoras pueden ser variables categóricas, ordinales o no, o bien cuantitativas, ya sea continuas o discretas.

Survival	Sex	
	Female	Male
0	154	709
1	308	142

Cuadro 2: Tabla de 2×2

En forma genérica

Y	X	
	0	1
0	$1-\pi(0)$	$1-\pi(1)$
1	$\pi(0)$	$\pi(1)$

Cuadro 3: Tabla de 2×2

En esta situación podríamos intentar modelar la variable de respuesta en función de las explicativas.

Recordemos que en el modelo lineal simple habitual, si Y es nuestra variable de respuesta y x una variable explicativa podemos formular el modelo como:

$$E(Y) = \beta_0 + \beta_1 x \tag{1}$$

Si, como en el ejemplo, nuestra variable de respuesta es binomial, entonces $E(Y) = \pi$, por lo tanto la generalización inmediata de (1) sería:

$$E(Y) = \pi = \pi(x) = \beta_0 + \beta_1 x \tag{2}$$

Sin embargo, (2) no parece ser un modelo adecuado, pues $\beta_0 + \beta_1 x$ podría tomar valores fuera del intervalo $(0, 1)$.

Un problema evidente de este modelo es que la probabilidad π es acotada, mientras que las $\mathbf{x}'\boldsymbol{\beta}$ (en el caso de un vector de covariables) pueden tomar cualquier valor real. Si bien esto podría controlarse imponiendo complicadas restricciones a los coeficientes, esta solución no resulta muy natural.

Una solución sencilla es *transformar* la probabilidad mediante una función que mapee el intervalo $(0, 1)$ sobre la recta real y luego modelar esta transformación como una función lineal de las variables independientes.

Una elección muy frecuente es:

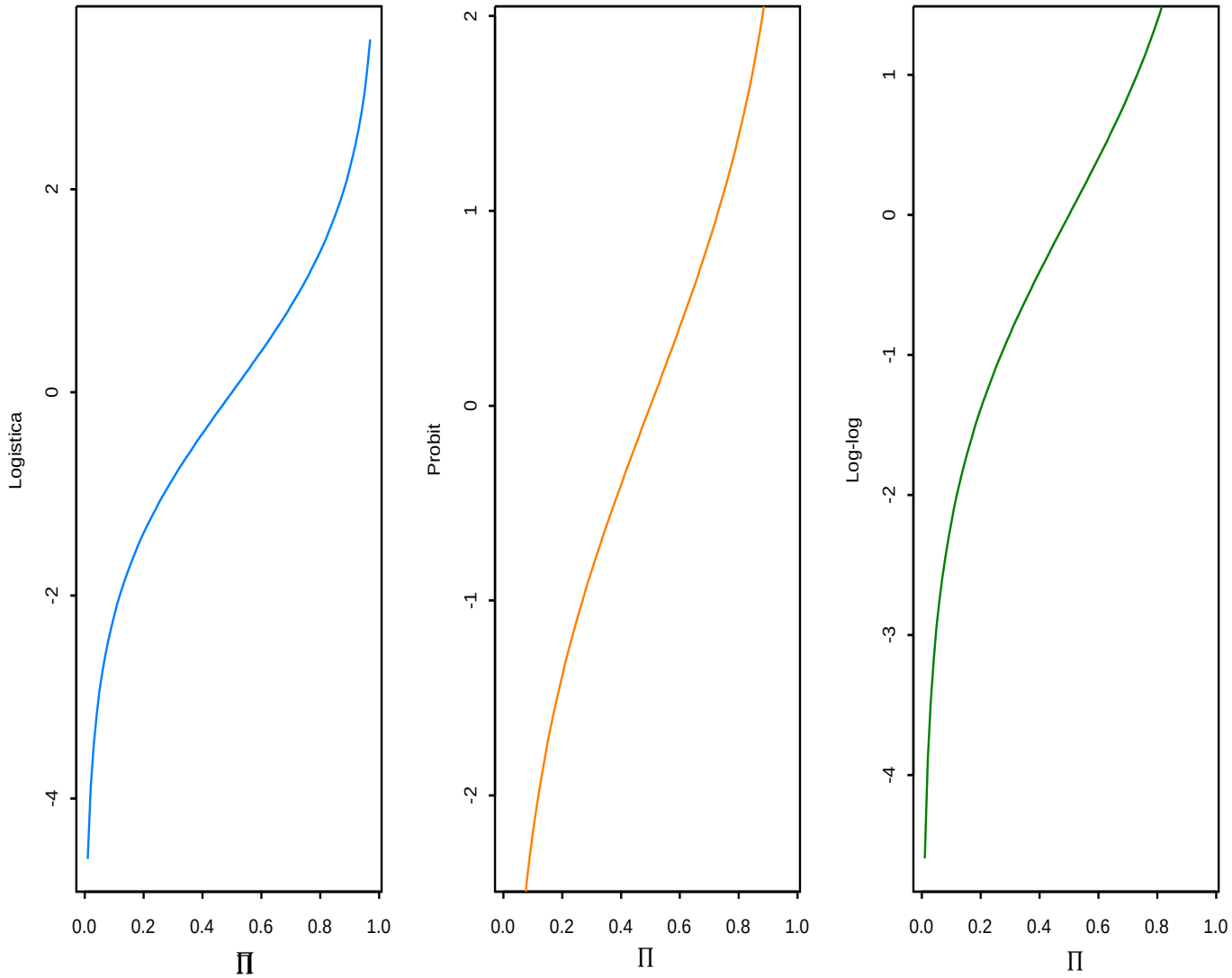
$$\text{logit}(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = \beta_o + \beta_1 x$$

que da origen al modelo de regresión logística. Otra forma de escribirlo es

$$\pi(x) = \frac{e^{\beta_o+\beta_1 x}}{1 + e^{\beta_o+\beta_1 x}}$$

Esta es sólo una elección posible y veremos más adelante porque es una elección razonable.

Un punto a destacar es que en este modelo es natural la heteroscedasticidad, pues $V(Y) = \pi(1-\pi)$, que será función de x.


Logistica
Probit
Log-log
Π
Π
Π
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
-2
-1
0
1
2
-4
-3
-2
-1
0
1

El modelo definido se conoce como modelo logístico, es un caso del **modelo lineal generalizado** con respuesta binomial y función de enlace logit.

Si bien los coeficientes $\boldsymbol{\beta}$ tienen una interpretación similar a la que tienen en el modelo lineal, debemos tener en cuenta que el miembro de la derecha es un logit y no una media, por lo que deberemos precisar cuál es su significado en este caso.

Estos temas los desarrollaremos en el contexto más general del **modelo lineal generalizado**. Este modelo es una extensión que comprende al modelo lineal que aplicamos cuando el supuesto de normalidad es razonable y que abarca también el caso de una respuesta Poisson, Binomial Negativa, Gamma, Exponencial, etc.

Una vez establecido el modelo que queremos ajustar deberemos estimar los parámetros, hallar intervalos de confianza para los mismos, evaluar la bondad del ajuste y es probable que nos interese realizar algún test que involucre a los parámetros. También tendremos que evaluar la influencia de las observaciones en la determinación de los valores estimados.

En nuestro último ejemplo consideramos de nuevo el caso de una tabla de contingencia. Supongamos que consideramos dos variables categóricas F y C. Sean F la variable que identificamos con las filas y C con las columnas. Supongamos nos interesa estudiar la asociación de las variables categóricas.

	C						
F	1	2	.	.	j	.	. J
1	.	.			. . .		.
2	.	.			. . .		.
.	.	.			. . .		.
i	.	.			π_{ij}		.
.	.	.			. . .		.
I	.	.			. . .		.

Cuadro 4: Tabla de distribución conjunta

Sabemos que si F y C son independientes, las π_{ij} se pueden escribir en términos

de las marginales como

$$\pi_{ij} = \pi_i^F \, \pi_j^C \ .$$

¿Qué ocurre en el caso general cuando no suponemos independencia?

Si pensamos en los valores esperados, $m_{ij} = n\pi_{ij}$ también podremos expresar a m_{ij} usando un modelo multiplicativo:

$$m_{ij} = \tau \tau_i^F \, \tau_j^C \tau_{ij}^{FC} \ , \tag{3}$$

donde, como en ANOVA, los τ deberán satisfacer ciertas restricciones.

Si tomamos logaritmo en (3) queda:

$$\begin{aligned} \log m_{ij} &= \log \tau + \log \tau_i^F + \log \tau_j^C + \log \tau_{ij}^{FC} \\ \log m_{ij} &= \mu + \mu_i^F + \mu_j^C + \mu_{ij}^{FC} \end{aligned}$$

que resulta un modelo aditivo, al que estamos más acostumbrados.

Este tipo de modelos recibe el nombre de log-lineal. Una diferencia con el modelo lineal habitual es que aquí las dos variables tienen un rol simétrico.

El investigador deducirá una asociación entre las variables interpretando los parámetros. Esta tarea puede ser más compleja si la cantidad de parámetros es muy elevada, como ocurre cuando aumenta el número de variables en el problema.

Bibliografía:

- Agresti, A. (1990). Categorical Data Analysis. Wiley, New York.
- Christensen, R. (1997). Log-linear Models and Logistic Regression. 2da. Edición. New York: Springer Verlag.
- Bishop, l., Fienberg, S. y Holland, P. (1976). Discrete Multivariater Analysis: Theory and Practice.
- Dobson, A. (2001). An Introduction to Generalized Linear Models. 2da. Edición. Londres: Chapman and Hall.
- Lindsey, J. (1997). Applying Generalized Linear Models. New York: Springer Verlag .
- Mc. Cullagh y Nelder, J. A. (1989). Generalized Linear Models. 2da. Edición. Londres: Chapman and Hall.
- Santner, T. y Duffy, D. (1989). The Statistical Analysis of Discrete Data. New York: Springer Verlag.

- Rao, C. R. (1965). Linear Statistical Inference and Its Applications. New York: Wiley.

Comentar: Régimen de Aprobación

Tablas de Contingencia

En la primera parte del curso estudiaremos la relación entre 2 ó 3 variables categóricas. Introduciremos parámetros que describan la asociación entre variables categóricas y luego haremos inferencia sobre estos parámetros.

Sean X e Y dos variables categóricas, donde X tiene I niveles e Y tiene J niveles posibles. Cuando clasificamos sujetos de acuerdo a las dos variables tenemos IJ combinaciones posibles.

Las casillas de la tabla representan los IJ resultados posibles. La probabilidad de que (X, Y) tome el valor (i, j) será π_{ij}. Cuando las celdas contienen la frecuencia de cada resultado ij tenemos una **tabla de contingencia**, término que introdujo Pearson en 1904. También suele llamársela **tabla de clasificación cruzada**. Una tabla de contingencia con I filas y J columnas se dice una tabla de $I \times J$.

			Y				
X	1	2	. .	j		. .	J
1	.	.		. . .			.
2	.	.		. . .			.
.	.	.		. . .			.
.	.	.		. . .			.
i	.	.		π_{ij}			.
.		.		. . .			.
.	.	.		. . .			.
I	.	.		. . .			.

Cuadro 5: Distribución Conjunta

La distribución de probabilidad π_{ij} es la distribución conjunta de X e Y, mientras que las marginales de ambas variables las obtendremos sumando columnas y filas, respectivamente: π_{i+} y π_{+j}, donde

$$\pi_{i+} = \sum_{j=1}^{J} \pi_{ij} \qquad \pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$$

En una tabla de 2×2 tendríamos:

	Columnas		
Filas	1	2	Total
1	π_{11}	π_{12}	π_{1+}
	$(\pi_{1\|1})$	$(\pi_{2\|1})$	(1)
2	π_{21}	π_{22}	π_{2+}
	$(\pi_{1\|2})$	$(\pi_{2\|2})$	(1)
Total	π_{+1}	π_{+2}	1

Cuadro 6: Distribución 2×2

En muchos casos una de las variables, digamos Y, es una variable de respuesta y la otra, X, es una variable explicativa.

En general, es de interés estudiar cómo cambia la distribución de Y cuando pasamos de un nivel de X a otro.

Dado que un sujeto está clasificado en la fila i de X, $\pi_{j|i}$ es la probabilidad de que clasifique en la columna j de Y, es decir $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$ es la **probabilidad**

condicional *de Y dado que $X = i$.*

En términos de las probabilidades definidas, tenemos que

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \quad \forall\, i, j\,.$$

Diremos que X e Y son **independientes** si

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \forall\, i, j\,,$$

y cuando vale la independencia

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j} \quad \forall\, i, j\,.$$

Supongamos que en n individuos observamos (X, Y) y volcamos esta información en una tabla de contingencia: n_{ij} el número de individuos que tienen $X = i$ e $Y = j$, de manera que

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \, .$$

En el **caso muestral** la información también suele disponerse sobre una tabla como la que sigue:

	$Y=1$	$Y=2$	$\cdot$	$Y=j$	$\cdot$	$Y=J$
$X=1$	n_{11}	n_{12}				n_{1J}
$X=2$	n_{21}	n_{22}				n_{2J}
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$X=i$	$\cdot$	$\cdot$	$\cdot$	n_{ij}	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$X=I$	n_{I1}	n_{I2}				n_{IJ}

Cuadro 7: Tabla de Contingencia Genérica de $I \times J$

Ejemplo de 2×2

Comencemos por considerar un ejemplo de los más sencillos de tablas de contingencia en el que tenemos dos variables.

G:Género	**C**: Cree en la vida postmortem		
	Si	No	Total
F	435	147	582
M	375	134	509
Total	810	281	1091

Cuadro 8: General Social Survey, 1991

En forma genérica, tendríamos

	C		
G	Si	No	Total
F	n_{11}	n_{12}	n_{1+}
M	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	$n_{++} = n$

Cuadro 9: Tabla de 2×2

Podría interesarnos testear la hipótesis de independencia entre las variables.

Podemos escribir esta hipótesis como

$$H_o : \pi_{ij} = \pi_{i+}\,\pi_{+j} \quad \forall i\,\forall j$$

Para resolver este problema necesitamos estimar las probabilidades.

Podríamos estimar las probabilidades bajo el supuesto de independencia y comparar los valores observados con los valores esperados bajo independencia. Esto

se puede realizar mediante un estadístico, conocido como χ^2 de Pearson, cuya distribución asintótica necesitaremos estudiar.

Podríamos estimar por máxima verosimilitud las probabilidades y realizar un test de cociente de verosimilitud. Para para esto necesitamos asumir una distribución subyacente.

Si cada una de las n observaciones es clasificada en forma independiente en una de las $I \times J$ celdas de la tabla con probabilidad π_{ij}, entonces el vector aleatorio que representa el número de individuos clasificados en la celda (i, j), $\mathbf{n}$, tiene distribución multinomial. La frecuencias esperadas en cada casilla son $\mu_{ij} = n\pi_{ij}$. Para $I = J = 2$ sería

$$P(\mathbf{n} = (n_{11}^*, n_{12}^*, , n_{21}^*, n_{22}^*)) = \frac{n!}{n_{11}^*!\, n_{12}^*!\, n_{21}^*!\, n_{22}^*!}\, \pi_{11}^{n_{11}^*}\, \pi_{12}^{n_{12}^*}\, \pi_{21}^{n_{21}^*}\, \pi_{22}^{n_2^*}\,.$$

Maximimizar el log–likelihood en el caso general resulta equivalente a maximizar:

$$\ell = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}^{*} \log \pi_{ij}$$

La maximización de ℓ debe contemplar que $\Sigma_{i=1}^{I} \Sigma_{j=1}^{J} \pi_{ij} = 1$.

Sin embargo, en principio esto es válido si la distribución subyacente es multinomial.

La pregunta aquí es cómo llegaron los datos a la tabla.

Tipos de Muestreo

Dada una tabla de contingencia hay varios esquemas de muestreo que pueden conducir a los datos tal como los hemos observado y que podrían influir en el modelo de probabilidad a utilizar. En este caso tenemos dos factores F y C cada uno con dos niveles. En general, tendremos un factor *fila* con I niveles y un factor *columna* con J niveles que corresponde a una tabla de $I \times J$.

El número total de celdas es $N = I \times J$. Los totales marginales muestrales son

$$
\begin{aligned}
n_{i+} &= \sum_{j=1}^{J} n_{ij} \qquad \text{total fila} \\
n_{+j} &= \sum_{i=1}^{I} n_{ij} \qquad \text{total columna} \\
n_{++} &= n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \qquad \text{gran total}
\end{aligned}
$$

Usaremos una notación similar a la anterior, por ejemplo, p_{ij} será proporción

muestral de la casilla (i, j) definida por

$$p_{ij} = \frac{n_{ij}}{n} .$$

Las condicionales muestrales quedarán definidas análogamente, por ejemplo,

$$p_{j|i} = \frac{p_{ij}}{p_{i+}} = \frac{n_{ij}}{n_{i+}} \quad \forall\, i, j .$$

En el ejemplo anterior asumimos que todos los datos han sido recolectados muestreando 1091 individuos que fueron clasificados de acuerdo con el **sexo** y la **creencia en la vida postmortem**. Vemos las dos variables como respuesta y nos interesa su distribución conjunta.

En los experimentos que responden a este esquema de muestreo, seleccionamos una muestra de n individuos de una población y registramos los valores (X, Y) para cada individuo. La distribución conjunta del vector $\mathbf{n}$ con componentes $\{n_{ij}\}$ es multinomial de parámetros n y $\boldsymbol{\pi} = \{\pi_{ij}\}$: $M(n, \boldsymbol{\pi})$, donde

$$\pi_{ij} = P(X = i, Y = j).$$

En este caso el gran total n es conocido y fijo.

A veces, se expresa los parámetros como medias de las celdas:

$$\mu_{ij} = E(n_{ij}) = n\pi_{ij} \, .$$

Este tipo de muestreo se conoce como **muestreo multinomial**.

Dado que la distribucón multinomial aparecerá con frecuencia en el análisis de datos categóricos, repasaremos algunas de sus propiedades.

Distribución Multinomial

Supongamos que realizamos n ensayos independientes, de manera que cada ensayo puede resultar en uno de los eventos $E_1, \ldots, E_K$ (los E_j's son mutuamente excluyentes y exhaustivos). En cada ensayo, el evento E_j puede ocurrir con probabilidad π_j y por lo tanto $\pi_1 + \ldots + \pi_K = 1$.

Si llamamos

$$W_j = \text{ número de veces que el evento } E_j \text{ ocurre },$$

entonces

$$W_1 + \ldots + W_K = n$$

y la distribución de $\mathbf{W} = (W_1, \ldots, W_K)'$ es multinomial de parámetros n y $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K)'$, es decir

$$\begin{aligned} \mathbf{W} &= (W_1, \ldots, W_K)' \sim M(n, \pi_1, \ldots, \pi_K) \\ \mathbf{W} &= (W_1, \ldots, W_K)' \sim M(n, \Pi)\,. \end{aligned}$$

De manera que:

$$\begin{cases} P(W_1 = w_1, \ldots, W_K = w_K) = \dfrac{n!}{w_1! \ldots w_K!} \pi_1^{w_1} \ldots \pi_n^{w_K} & \text{si } \Sigma_{i=1}^K w_i = n \\ \\ 0 & \text{caso contrario (c.c.)} \end{cases}$$

Como en el caso de la distribución binomial, algunas propiedades más elementales de esta distribución, como el cálculo de esperanza o matriz de covarianza, se deducen fácilmente pensando a $\mathbf{W}$ como una suma de vectores de 0's y 1's. Más precisamente, podemos escribir

$$\mathbf{W} = \mathbf{Z}_1 + \ldots + \mathbf{Z}_n$$

donde las $\mathbf{Z}_i$ son independientes y cada una es $M(1, \pi_1, \ldots, \pi_K) = M(1, \Pi)$.

$$\mathbf{Z}_i' = (0, \ldots, \underset{\substack{\downarrow \\ j}}{1}, \ldots, 0)$$

$\mathbf{Z}_i$ es un vector con un 1 en la coordenada j si E_j ocurrió en el i-ésimo ensayo y en el resto de las posiciones 0's.

Los elementos de $\mathbf{Z}_i$ son Bernoulli correlacionadas.

Esperanza y Varianza

Por ejemplo, supongamos que $K = 2$ y que $\mathbf{Z} \sim M(1, \pi_1, \pi_2)$. Los resultados posibles son son

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{con probabilidad } \pi_1$$

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{con probabilidad } \pi_2 = 1 - \pi_1$$

La media de $\mathbf{Z} = (Z_1, Z_2)'$ es

$$E(\mathbf{Z}) = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}$$

El segundo momento de $\mathbf{Z}$ es:

$$E(\mathbf{Z}\mathbf{Z}') = E\begin{pmatrix} z_1^2 & z_1 z_2 \\ \\ z_1 z_2 & z_2^2 \end{pmatrix}$$

$$= \begin{pmatrix} \pi_1 & 0 \\ \\ 0 & \pi_2 \end{pmatrix}$$

Por lo tanto, la matriz de covarianza de $\mathbf{Y}$ es:

$$\Sigma_{\mathbf{Z}} = E(\mathbf{Z}\mathbf{Z}') - E(\mathbf{Z})E(\mathbf{Z}')$$
$$= \begin{pmatrix} \pi_1 & 0 \\ \\ 0 & \pi_2 \end{pmatrix} - \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} (\pi_1 \quad , \quad \pi_2)$$

$$= \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}$$

Vamos a demostrar que si hay K resultados posibles y $\mathbf{Z} \sim M(1, \pi_1, \ldots, \pi_K)$

$$\begin{aligned} E(\mathbf{Z}) &= \Pi = (\pi_1, \ldots, \pi_K)' \\ \Sigma_{\mathbf{Z}} &= \Delta(\Pi) - \Pi\Pi' \end{aligned}$$

Probaremos que si $\pi_j \neq 0, \forall j$

$$rg(\Sigma_{\mathbf{Z}}) = K - 1\,.$$

Luego, si $\mathbf{W} \sim M(n, \pi_1, \ldots, \pi_K)$

$$\begin{aligned} E(\mathbf{W}) &= n\Pi \\ \Sigma_{\mathbf{W}} &= \Delta(n\Pi) - n\Pi\Pi' \end{aligned}$$

Otras Propiedades

Enumeraremos algunas propiedades de la distribución multinomial que nos resultarán útiles, algunas serán ejercicio de la práctica.

1. El espacio paramétrico natural de la distribución multinomial es el simplex, en R^K definido por
$$\mathcal{S} = \{\boldsymbol{\pi} : \pi_j > 0, \pi_1 + \ldots + \pi_K = 1\}.$$
2. Si $\mathbf{W} = (W_1, \ldots, W_K)' \sim M(n, \pi_1, \ldots, \pi_K)$, entonces
$$\begin{aligned} W_j &\sim Bi(n, \pi_j) \\ Cov(W_i, W_j) &= -n\,\pi_i\,\pi_j \quad i \neq j\,, \end{aligned}$$
es decir las W_i están negativamente correlacionadas.
3. Si $\mathbf{W} = (W_1, \ldots, W_K)' \sim M(n, \pi_1, \ldots, \pi_K)$, entonces
$$W^* = (W_1 + W_2, W_3, \ldots, W_K)' \sim M(n, \pi_1 + \pi_2, \pi_3, \ldots, \pi_K)$$

Es decir, si se colapsa una multinomial sumando celdas la distribución sigue siendo multinomial.

4. Sea $\mathbf{W} = (W_1, \ldots, W_K)' \sim M(n, \pi_1, \ldots, \pi_K)$. Consideremos la distribución condicional de

$$\mathbf{W} \Big| \begin{array}{l} W_1 + W_2 = z \\ W_3 + \ldots + W_K = n - z \end{array}$$

Los vectores $(W_1, W_2)'$ y $(W_3, \ldots, W_K)'$ son condicionalmente independientes y multinomiales:

$$(W_1, W_2)' \sim M(z, \frac{\pi_1}{\pi_1 + \pi_2}, \frac{\pi_2}{\pi_1 + \pi_2})$$

$$(W_3, \ldots, W_K)' \sim M(n - z, \frac{\pi_3}{\pi_3 + \ldots + \pi_K}, \ldots, \frac{\pi_K}{\pi_3 + \ldots + \pi_K})$$

5. Si $W_1, \ldots, W_K$ son variables independientes tales que $W_j \sim \mathcal{P}(\lambda_j)$, entonces

$$(W_1, \ldots, W_K)'|_{\sum_{j=1}^{K} W_j = n} \sim M(n, \pi_1, \ldots, \pi_K)$$

donde

$$\pi_j = \frac{\lambda_j}{\lambda_1 + \ldots + \lambda_K}$$

Por lo tanto, la distribución de $W_1, \ldots, W_K$ puede ser factorizada en el producto de

$$n = \sum_{j=1}^{K} W_j \sim \mathcal{P}(\sum_{j=1}^{K} \lambda_j)$$

y

$$(W_1, \ldots, W_K)'|_{n=n^*} \sim M(n^*, \pi_1, \ldots, \pi_K)$$

Esto será especialmente útil a la hora de calcular la función de verosimilitud bajo ciertas condiciones.

En nuestro ejemplo hemos hablado de la función de verosimilitud, recordemos algunas propiedades.

Propiedades de los Estimadores de Máxima Verosimilitud

Recordemos que si la variable aleatoria Y tiene función de densidad (f.d.)o probabilidad puntual (f.p.p.)$f(y, \boldsymbol{\theta})$, la verosimiltud $L(\boldsymbol{\theta}, y)$ es simplemente $f(y, \boldsymbol{\theta})$ mirada como función de $\boldsymbol{\theta}$ con y fijo.

La función de probabilidad puntual o densidad es definida sobre el soporte $y \in \mathcal{Y}$, mientras que la verosimilitud es definida sobre un espacio paramétrico Θ.

En muchos casos es conveniente trabajar con el logaritmo de la verosimilitud (log-likelihood)

$$l(\boldsymbol{\theta}, y) = \log L(\boldsymbol{\theta}, y) \ .$$

En general, tendremos una muestra aleatoria $Y_1, \ldots, Y_n$ con f.d. o f.p.p. $f(y, \boldsymbol{\theta})$, de manera que la verosimilitud será:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \boldsymbol{\theta})$$

y la log-versosimilitud

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i, \boldsymbol{\theta}) \ .$$

Una propiedad útil de los EMV es la de **invariancia** que dice que si g es una función con inversa g^{-1}, de manera que $\phi = g(\boldsymbol{\theta})$ implica que $\boldsymbol{\theta} = g^{-1}(\phi)$, entonces el EMV de ϕ, $\widehat{\phi}$, se calcula como

$$\widehat{\phi} = g(\widehat{\boldsymbol{\theta}}) \, ,$$

siendo $\widehat{\boldsymbol{\theta}}$ el EMV de $\boldsymbol{\theta}$.

Como ya sabemos, podemos maximizar $L(\boldsymbol{\theta})$ o bien maximizar $l(\boldsymbol{\theta})$. En problemas regulares, el EMV puede hallarse igualando a 0 las derivadas pimeras de $l(\boldsymbol{\theta})$ respecto de $\boldsymbol{\theta}$.

La derivada primera de $l(\boldsymbol{\theta})$ respecto de $\boldsymbol{\theta}$ se llama score. En el caso univariado tenemos:

$$l'(\theta) = \sum_{i=1}^{n} u_i(\theta) \ ,$$

donde

$$u_i(\theta) = \frac{\partial}{\partial \theta} \log f(y_i, \theta) \ .$$

Si tenemos q parámetros, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, el vector de scores es

$$l'(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial l}{\partial \theta_1} \\ \frac{\partial l}{\partial \theta_2} \\ . \\ . \\ \frac{\partial l}{\partial \theta_q} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_1} \log f(y_i, \theta) \\ \sum_{i=1}^{n} \frac{\partial}{\partial \theta_2} \log f(y_i, \theta) \\ . \\ . \\ \sum_{i=1}^{n} \frac{\partial}{\partial \theta_q} \log f(y_i, \theta) \end{bmatrix}$$

Una propiedad bien conocida del score es que su esperanza es nula:

$$E\left(l'(\boldsymbol{\theta})\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \int l'((\boldsymbol{\theta}_0))f(y,\boldsymbol{\theta}_0)dy = \int \frac{f'(y,\boldsymbol{\theta}_0)}{f(y,\boldsymbol{\theta}_0)}f(y,\boldsymbol{\theta}_0)dy = 0$$

La varianza de los score es conocida como la **información de Fisher**. En el caso univariado, la información de Fisher es:

$$i(\theta) \;=\; V(u(\theta)) = V(l'(\theta)) = E\left[(l'(\theta))^2\right]$$

Recordemos que

$$i(\theta) = E(-l''(\theta)) = -E\left(\frac{\partial^2}{\partial\theta^2}\log f(y,\theta)\right)$$

En el caso multivariado, tendremos $\mathbf{I}(\theta)$ es una matriz de $q \times q$ tal que:

$$\{\mathbf{I}(\theta)\}_{ij} \;=\; -E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(y,\theta)\right)$$

En Estadística se probó que, bajo condiciones de regularidad, los EMV son asintóticamente normales, de manera que

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, \mathbf{I}^{-1}(\theta))$$

¿Cómo sería en el caso general de una multinomial cualquiera?

Para simplificar la notación, indicaremos $\{n_1, \ldots, n_K\}$ las observaciones de cada casilla, con $n = \Sigma_{i=1}^{K} n_i$ y siendo $\{\pi_1, \ldots, \pi_K\}$ las probabilidades de cada celda.

Luego, la función de verosimilitud será:

$$L = L(\pi_1, \ldots, \pi_K) = \frac{n!}{\prod_{i=1}^{K} n_i!} \prod_{i=1}^{K} \pi_i^{n_i} \qquad \text{donde } \sum_{i=1}^{K} \pi_i = 1 \ .$$

Como el ln es una función estrictamente creciente, hallar el máximo de L equivale a hallar el máximo de

$$l = \ln L = \ln \left(\frac{n!}{\prod_{i=1}^{K} n_i!} \right) + \sum_{i=1}^{K} n_i \ln \pi_i \qquad \text{donde } \sum_{i=1}^{K} \pi_i = 1 \ .$$

Como $\Sigma_{i=1}^{K} \pi_i = 1$, entonces $\pi_K = 1 - \Sigma_{i=1}^{K-1} \pi_i$ y $n!/\prod_{i=1}^{K} n_i!$ es constante, buscamos el máximo de

$$l = \ln L = cte + \sum_{i=1}^{K-1} n_i \ln \pi_i + n_K \ln(1 - \sum_{i=1}^{K-1} \pi_i) \ .$$

Para buscar los puntos críticos planteamos:

$$\frac{\partial l}{\partial \pi_i} = \frac{n_i}{\pi_i} - \frac{n_K}{1 - \sum_{i=1}^{K-1} \pi_i} = \frac{n_i}{\pi_i} - \frac{n_K}{\pi_K} = 0$$

Esta igualdad es cierta $\forall i = 1, \ldots, K$ (para K se cumple trivialmente).

Luego,

$$\frac{n_i}{n_K} = \frac{\pi_i}{\pi_K} \Rightarrow \frac{n}{n_K} = \frac{1}{\pi_K} \Rightarrow \widehat{\pi}_K = \frac{n_K}{n}$$

$$\Rightarrow \widehat{\pi}_i = \frac{n_i \widehat{\pi}_K}{n_K} = \frac{n_i}{n}$$

Por lo tanto, tal como es de esperar

$$\widehat{\pi}_i = \frac{n_i}{n} = p_i \quad i = 1, \ldots, K$$

También podríamos usar multiplicadores de Lagrange. En ese caso, consideraríamos

$$\Lambda = \ln\left(\frac{n!}{\prod_{i=1}^{K} n_i!}\right) + \sum_{i=1}^{K} n_i \ln \pi_i + \lambda(1 - \sum_{i=1}^{K} \pi_i) \ .$$

Para buscar los puntos críticos planteamos:

$$\begin{aligned} \frac{\partial \Lambda}{\partial \pi_i} &= \frac{n_i}{\pi_i} - \lambda = 0 \quad i = 1, \ldots, K \\ \frac{\partial \Lambda}{\partial \lambda} &= 1 - \sum_{i=1}^{K} \pi_i = 0 \end{aligned}$$

Por lo tanto,

$$\lambda \pi_i = n_i \Rightarrow \lambda \sum_{i=1}^{K} \pi_i = n \Rightarrow \lambda = n \, ,$$

luego,

$$\widehat{\pi}_i = \frac{n_i}{n} = p_i \quad i = 1, \ldots, K$$

Muestreo Poisson

Otra posibilidad es que los datos de la tabla sean realizaciones independientes de v.a. con distribución Poison.

Un proceso que sustentaría este modelo es aquel en que cada celda los individuos llegan aleatoriamente al lugar donde se los clasifica. En este caso n no está prefijado y todos los valores de la tabla son aleatorios.

En el **muestreo de Poisson** tenemos que

$$n_{ij} \sim \mathcal{P}(\lambda_{ij}) \quad \begin{array}{l} i = 1, \ldots, I \\ j = 1, \ldots, J \end{array}$$

independientes. En este esquema el gran total n_{++} no está fijo, sino que es aleatorio.

Ejemplo:

Supongamos que se realiza en un control de velocidad durante una hora. Para ello se cuenta con un radar que registra la velocidad de cada auto que pasa por

el puesto de observación. Supongamos que de cada auto que pasa se registra la velocidad y la marca del auto. Así se obtienen

$$\begin{aligned} X &= \text{marca del auto (1 = Ford, 2 = Fiat, 3 = Chevrolet, 4 = Otros)} \\ Y &= \text{si el auto excede el límite de velocidad (1 = Si, 0 = No).} \end{aligned}$$

Es claro que, bajo independencia, $n \sim \mathcal{P}(\lambda_{++})$.

Si tenemos un muestreo de Poisson, la distribución de los n_{ij} condicional a que n está fijo en un valor, digamos m, ya no es más Poisson, más aún ya no son más independientes.

La distribución condicional de los n_{ij} dado que $n = m$ es multinomial. Para simplificar la notación olvidaremos el doble subíndice, indicaremos $\{n_1, \ldots, n_K\}$, entonces si $\Sigma_{i=1}^{K} n_i = m$

$$P(n_1 = m_1, \ldots, n_K = m_K \mid \sum_{i=1}^{K} n_i = m)$$

$$
\begin{aligned}
&= \frac{P\left(n_1 = m_1, \ldots, n_K = m_K \cap \Sigma_{i=1}^{K} n_i = m\right)}{P(n = m)} \\
&= \prod_{i=1}^{K} \frac{e^{-\lambda_i} \lambda_i^{m_i}}{m_i!} \quad \frac{m!}{e^{-\sum_{i=1}^{K} \lambda_i} \left\{\sum_{i=1}^{K} \lambda_i\right\}^m} \\
&= \left\{\prod_{i=1}^{K} \lambda_i^{m_i}\right\} \frac{m!}{\prod_{i=1}^{K} m_i!} \frac{1}{\left\{\sum_{i=1}^{K} \lambda_i\right\}^m} \\
&= \frac{m!}{\prod_{i=1}^{K} m_i!} \quad \prod_{i=1}^{K} \pi_i^{m_i}
\end{aligned}
$$

donde $\pi_i = \frac{\lambda_i}{\Sigma_{j=1}^{K} \lambda_j}$ o volviendo a la notación original:

$$\pi_{ij} = \frac{\lambda_{ij}}{\lambda_{++}}$$

Entonces:

$$(n_1, \ldots, n_K)|_{n=m} \sim M(m, \pi_1, \ldots, \pi_k)\,.$$

Esto nos servirá a la hora de plantear la verosimilitud cuando deseemos estimar.

En efecto, podemos factorizar la verosimilitud como el producto del likelihood de la Poisson n ($n \sim \mathcal{P}(\lambda_{++})$) y el likelihood de una multinomial correspondiente a $\{n_{ij}\}$ dado n, con parámetros

$$\pi_{ij} = \frac{\lambda_{ij}}{\lambda_{++}}$$

El total n no da información acerca de las π_{ij}.

Es interesante observar, que desde el punto de vista de la verosimilitud, la inferencia sobre $\boldsymbol{\pi}$ es la misma si n es considerado fijo o aleatorio.

Muestreo Multinomial Independiente

Volviendo al ejemplo de **creencia en la vida postmortem** vs. **género**, otra alternativa podría ser en realidad se haya tomado una muestra de 582 mujeres y otra muestra independiente de 509 hombres a los que se clasificó según su creencia. En este caso nos centramos en la distribución condicional de la creencia dado cada nivel de género.

En este esquema los totales por fila están fijos y tenemos n_1 (n_{1+}) individuos de género femenino y n_2 (n_{2+}) individuos de género masculino.

Si π_i es la probabilidad de que el individuo crea en la vida postmortem para el nivel i de género tendremos:

$$\pi_i = P(C = 1|G = i) = \frac{\pi_{i1}}{\pi_{i+}}$$

Las variables de interés serán: $Y_{i1} \sim Bi(n_i, \pi_i)$, $i = 1, 2$, que cuentan el número de individuos que sí creen en cada género.

La distribución conjunta de (Y_{11}, Y_{21}) es

$$P\left((Y_{11}, Y_{21}) = (y_{11}, y_{21})\right) = \frac{n_1!}{y_{11}!y_{12}!}\pi_1^{y_{11}}(1-\pi_1)^{y_{12}}\frac{n_2!}{y_{21}!y_{22}!}\pi_2^{y_{21}}(1-\pi_2)^{y_{22}}$$

La hipótesis de interés es la de **homogeneidad**, es decir que la probabilidad de creencia es la misma en ambos género:

$$H_0 : \pi_1 = \pi_2$$

Notemos que bajo independencia $\pi_{ij} = \pi_{i+}\pi_{+j}$, entonces la probabilidad condicional

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} = \pi_{+j}$$

es decir no depende de la fila i, con lo cual homogeneidad e independencia son equivalentes.

Supongamos que decidimos de antemano que vamos a muestrear n_{i+} individuos con $X = i\,(i = 1, \ldots, I)$ y que para cada uno de ellos registramos el valor de Y.

En este esquema cada fila de tabla $(n_{i1}, n_{i1}, \ldots, n_{iJ})'$ es multinomial con probabilidades

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$$

y las filas son muestreadas en forma independiente.

Este tipo de muestreo es razonable de aplicar cuando los datos provienen de un muestreo aleatorio estratificado (estratos definidos por X) o en un experimento en el X =grupo de tratamiento.

También es adecuado cuando no tenemos totales por filas fijos, pero estamos interesados en $P(Y|X)$ y no en $P(X)$, lo que corresponde a que Y es el resultado de interés y no deseamos modelar a X.

Por lo que vimos, en estos tres tipos de muestreo el núcleo de la verosimilitud es el mismo.

La importancia de estos resultados es que el análisis que hagamos es independiente del esquema de muestreo y depende de los parámetros de interés.

Estimación y Tests de Bondad de Ajuste

Supongamos que tenemos un muestreo multinomial y obtenemos la tabla (X, Y) en n individuos.

	$Y=1$	$Y=2$		$Y=J$
$X=1$	n_{11}	n_{12}		n_{1J}
$X=2$	n_{21}	n_{22}		n_{2J}
.	.	.	. .	.
.	.	.	. .	.
.	.	.	. .	.
$X=I$	n_{I1}	n_{I2}		n_{IJ}

Cuadro 10: Tabla de $I \times J$

Sea n_{ij} el número de individuos que tienen $P(X=i, Y=j)$, de manera que

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

Por lo que ya vimos, los estimadores de máxima verosimilitud de π_{ij} son

$$\widehat{\pi}_{ij} = \frac{n_{ij}}{n} \quad \forall i, j \, .$$

Si las dos variables categóricas fueran independientes, tendríamos

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \forall i, j \, ,$$

luego por la propiedad de invariancia de los estimadores de máxima verosimilitud, bajo independencia el estimador de máxima verosimilitud de π_{ij} sería:

$$\widehat{\pi}_{ij} = \widehat{\pi}_{i+}\widehat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n^2} \quad \forall i, j \, .$$

Dado que $n_{ij} \sim Bi(n, \pi_{ij})$,

$$m_{ij} = E(n_{ij}) = n\pi_{ij} \, .$$

Bajo el supuesto de independencia, el EMV es

$$\widehat{m}_{ij} = n\widehat{\pi}_{ij} = \frac{n_{i+}n_{+j}}{n}$$

Estos estimadores tienen la propiedad de tener las mismas marginales que la tabla:

$$\begin{aligned} \widehat{m}_{i+} &= \sum_{j=1}^{J} \frac{n_{i+}n_{+j}}{n} = n_{i+} \\ \widehat{m}_{+j} &= \sum_{i=1}^{I} \frac{n_{i+}n_{+j}}{n} = n_{+j} \end{aligned}$$

Test de Bondad de Ajuste

Pearson (1900) presentó un test que sirve para evaluar si una distribución multinomial tiene ciertas probabilidades π_{ij0} propuestas.

Por ejemplo, consideremos el caso de las leyes de herencia de la teoría de Mendel. Mendel cruzó arvejas de cepa amarilla con arvejas de cepa verde puras y predijo que la segunda generación de híbridos serían un 75 % amarillas y un 25 % verdes, siendo las amarillas las de carácter dominante.

En un experimento de $n = 8023$ semillas, resultaron $n_1 = 6022$ amarillas y $n_2 = 2001$ verdes. Las proporciones esperadas eran $\pi_1 = 0.75$ y $\pi_2 = 0.25$, por lo tanto $m_1 = 6017.25$ y $m_2 = 2005.75$. La pregunta es: ¿se verifican las leyes de Mendel en este caso?

Como antes, *vectorizaremos* la tabla notando $\{n_1, \ldots, n_K\}$ a las casillas y $\{\pi_1, \ldots, \pi_K\}$ las probabilidades correspondiente y $n = \sum_{i=1}^{K} n_i$.

Supongamos que las hipótesis a testear son

$$H_0 : \pi_i = \pi_{i0}, \sum_{j=1}^{K} \pi_{i0} = \sum_{j=1}^{K} \pi_i = 1 \quad H_1 : \exists i : \pi_i \neq \pi_{i0}$$

Pearson propuso el siguiente estadístico:

$$\chi^2 = \sum_{j=1}^{K} \frac{(n_i - m_{i0})^2}{m_{i0}} \quad \text{donde } m_{i0} = n\pi_{i0}$$

La idea intuitiva es que comparamos el valor **observado** (n_i) con el valor **esperado** (m_{i0}) bajo H_0, suele decirse :

$$\frac{(observado - esperado)^2}{esperado}.$$

Intuitivamente rechazaremos H_0 cuando esto sea muy grande. ¿Cuán grande?

El argumento heurístico que dio Pearson es el siguiente: si $n_1, \ldots, n_K$ fueran v.a. independientes tales que $n_i \sim \mathcal{P}(m_i)$, bajo ciertas condiciones de regularidad

$$\frac{n_i - m_i}{\sqrt{m_i}} \overset{aprox.}{\sim} N(0,1)$$

entonces

$$\sum_{i=1}^{K} \left[\frac{n_i - m_i}{\sqrt{m_i}}\right]^2 \overset{aprox.}{\sim} \chi^2_K$$

Si además agregásemos la restricción $\sum_{i=1}^{K} n_i = n$, sería natural perder un grado de libertad y que la distribución asintótica del estadístico resultase χ^2_{K-1}.Por todo esto, la regla de decisión sería

$$\text{Rechazamos } H_0 \text{ si } \chi^2 > \chi^2_{K-1,\alpha}$$

En el caso en que $N = 2$, el estadístico queda

$$n\frac{(\widehat{p}-\pi_0)^2}{\pi_0}+n\frac{(\widehat{p}-\pi_0)^2}{1-\pi_0}=n\frac{(\widehat{p}-\pi_0)^2}{\pi_0(1-\pi_0)}=\left[\frac{\widehat{p}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\right]^2$$

que es el cuadrado del test habitual para testear

$$H_0:\pi=\pi_0 \quad H_1:\pi\neq\pi_0$$

que tiene distribución asintótica normal y en consecuencia, su cuadrado lo compararíamos con una χ^2_1.

Veamos la justificación teórica de este test. Comenzaremos por presentar algunos resultados teóricos.

Proposición 1:

Sean $\mathbf{X}_n = (X_{1n}, \ldots, X_{Kn})'$ una sucesión de v.a. y $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)' \in \Re^K$.
Si $\forall \boldsymbol{\lambda} \in \Re^K$

$$\boldsymbol{\lambda}'\mathbf{X}_n = \lambda_1 X_{1n} + \ldots + \lambda_K X_{Kn} \xrightarrow{\mathcal{D}} \lambda_1 X_1 + \ldots + \lambda_K X_K \,,$$

donde $\mathbf{X} = (X_1, \ldots, X_K)' \sim \mathcal{F}$, entonces la distribución límite de $\mathbf{X}_n = (X_{1n}, \ldots, X_{Kn})'$ existe y es $\mathcal{F}$.

Teorema Central del Límite Multivariado (TCLM)

Sea $\mathbf{U}_n = (U_{1n}, \ldots, U_{Kn})'$ una sucesión de vectores aleatorios independientes tales que $E(\mathbf{U}_n) = \boldsymbol{\mu}$ y $\Sigma_{\mathbf{U}_n} = \Sigma$, $n = 1, 2, \ldots$

Si $\bar{\mathbf{U}}_n = (\bar{U}_{1n}, \ldots, \bar{U}_{Kn})'$ es el vector de promedios, donde para cada $1 \leq i \leq K$ $\bar{U}_{in} = \frac{1}{n} \sum_{j=1}^{n} U_{ij}$, entonces

$$\sqrt{n}(\bar{\mathbf{U}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} N_K(\mathbf{0}, \Sigma)\,.$$

Según la Proposición 1 debemos estudiar la distribución de $\boldsymbol{\lambda}'\bar{\mathbf{U}}_n$.

$$\begin{aligned} \boldsymbol{\lambda}'\bar{\mathbf{U}}_n &= \lambda_1\bar{U}_{1n} + \ldots + \lambda_K\bar{U}_{Kn} \\ &= \lambda_1\frac{\sum_{j=1}^{n} U_{1j}}{n} + \ldots + \lambda_K\frac{\sum_{j=1}^{n} U_{Kj}}{n} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{\lambda}' \mathbf{U}_j \\
&= \frac{1}{n} \sum_{j=1}^{n} W_j \\
&= \bar{W}_n
\end{aligned}
$$

donde $E(W_i) = \boldsymbol{\lambda}' \boldsymbol{\mu}$, $Var(W_i) = \boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}$.

Por el TCL univariado, tenemos que

$$\sqrt{n}(\bar{W}_n - \boldsymbol{\lambda}' \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} N_K(0, \boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}) \,,$$

es decir,

$$\sqrt{n}(\boldsymbol{\lambda}' \bar{\mathbf{U}}_n - \boldsymbol{\lambda}' \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} N_K(0, \boldsymbol{\lambda}' \Sigma \boldsymbol{\lambda}) \,,$$

que corresponde a la distribución de $\boldsymbol{\lambda}' \mathbf{U}$, con $\mathbf{U} \sim N_K(\mathbf{0}, \Sigma)$

por lo que

$$\sqrt{n}(\bar{\mathbf{U}}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} N_K(\mathbf{0}, \Sigma)\,.$$

Proposición 2:

Si $\mathbf{Z}_n \xrightarrow{\mathcal{D}} \mathbf{Z}$ y g es una función continua, entonces

$$g(\mathbf{Z}_n) \xrightarrow{\mathcal{D}} g(\mathbf{Z})\,.$$

En IR sabemos que si

$$\sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{D}} N(0, \sigma^2)\,,$$

entonces bajo condiciones de suavidad de la función g,

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{\mathcal{D}} N(0, \sigma^2(g'(\mu))^2)$$

El siguiente lema, conocido como Método Δ, generaliza este resultado para una función de vector aleatorio.

Lema 2: Método Δ una función de vector aleatorio.

Supongamos que $\mathbf{T}_n = (T_{1n}, \ldots, T_{Kn})$ es una sucesión de vectores aleatorios tal que

$$\sqrt{n}((T_{1n}, \ldots, T_{Kn}) - (\theta_1, \ldots, \theta_K)) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma)$$

Sea g una función tal que $g : IR^K \longrightarrow IR$ diferenciable. Luego, si

$$\phi = (\phi_i) = \frac{\partial g}{\partial t_i}|_{t=\theta} \neq \mathbf{0}\,,$$

entonces

$$\sqrt{n}(g(T_{1n}, \ldots, T_{Kn}) - g(\theta_1, \ldots, \theta_K)) \xrightarrow{\mathcal{D}} N(0, \phi'\Sigma\phi)$$

Análogamente, si en lugar de un campo escalar tenemos un campo vectorial, es decir $g : IR^K \longrightarrow IR^q$, donde cada componente g_i es diferenciable como en el lema anterior, obtenemos

$$\sqrt{n}(g(T_{1n}, \ldots, T_{Kn}) - g(\theta_1, \ldots, \theta_K)) \xrightarrow{\mathcal{D}} N_q(\mathbf{0}, \mathbf{G}\Sigma\mathbf{G}')$$

donde

$$G_{ij} = \frac{\partial g_i}{\partial t_j}|_{\mathbf{t}=\boldsymbol{\theta}} \cdot$$

Ahora estudiaremos la distribución asintótica de $(\frac{n_1}{n}, \ldots, \frac{n_{K-1}}{n})$ cuando

$$(n_1, \ldots, n_K)' \sim M(n, \pi_1, \ldots, \pi_K) \qquad \sum_{i=1}^{K} \pi_i = 1.$$

Llamemos $\mathbf{p} = (p_1, \ldots, p_K)'$, $p_i = \dfrac{n_i}{n}$.

Consideremos el vector $\mathbf{Y}_i \sim M(1, \pi_1, \ldots, \pi_K)$ que ya definimos con todas sus componentes iguales a 0 y un único 1 en la coordenada j-ésima si en la i-ésima observación ocurrió la categoría j:

$$\mathbf{Y}_i = (0, \ldots, \underset{\substack{\downarrow \\ j}}{1}, \ldots, 0)' \qquad 1 \leq i \leq n$$

Recordemos que si $\mathbf{Y}_i \sim M(1, \pi_1, \ldots, \pi_K)$

$$\begin{aligned} E(\mathbf{Y}_i) &= \boldsymbol{\pi} \\ \Sigma_{\mathbf{Y}_i} &= \Delta(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' \end{aligned}$$

Podemos escribir al vector $\mathbf{p}$ en términos de los $\mathbf{Y}_i$:

$$\mathbf{p} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i = (\bar{Y}_1, \ldots, \bar{Y}_K)$$

entonces por el T.C.L multivariado sabemos que

$$\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{\mathcal{D}} N_K(\mathbf{0}, \boldsymbol{\Delta}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}') \,.$$

Ya hemos visto que como los π_i's están relacionados, $\Sigma = \Delta(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ no es invertible.

Definamos $\tilde{\mathbf{p}} = (p_1, \ldots, p_{K-1})'$ y $\tilde{\boldsymbol{\pi}} = (\pi_1, \ldots, \pi_{K-1})'$.

Notemos que $\tilde{\mathbf{p}} = \mathbf{T}\mathbf{p}$, siendo $\mathbf{T}$ es una transformación lineal, tenemos que

$$\sqrt{n}(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}) \xrightarrow{\mathcal{D}} N_{K-1}(\mathbf{0}, \boldsymbol{\Delta}(\tilde{\boldsymbol{\pi}}) - \tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}')\,,$$

donde $\Delta(\tilde{\boldsymbol{\pi}}) - \tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}'$ sí es invertible.

Esto quiere decir que bajo H_0, si $\tilde{\Sigma}_0 = \Delta(\tilde{\boldsymbol{\pi}}_0) - \tilde{\boldsymbol{\pi}}_0\tilde{\boldsymbol{\pi}}_0'$, entonces

$$\sqrt{n}(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}_0) \xrightarrow{\mathcal{D}} N_{K-1}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{\mathbf{0}})\,.$$

Por lo tanto, como $\mathbf{x}'\mathbf{D}^{-1}\mathbf{x}$ es una función continua de $\mathbf{x}$ por la Proposición 2 tenemos que

$$n(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}_0')\,\tilde{\Sigma}_0^{-1}\,(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}_0) \xrightarrow{\mathcal{D}} \chi^2_{K-1}$$

Calculando efectivamente la forma cuadrática que estamos considerando, veremos que

$$n(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}_0')\,\tilde{\Sigma}_0^{-1}\,(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}_0) = n\sum_{j=1}^{K}\frac{(p_i - \pi_{i0})^2}{\pi_{i0}}$$

Ejemplo: Leyes de Mendel

El test de Pearson fue usado para testear las leyes de herencia de la teoría de Mendel. Mendel cruzó arvejas de cepa amarilla con arvejas de cepa verde puras y predijo que la segunda generación de híbridos serían un 75 % amarillas y un 25 % verdes, siendo las amarillas las de carácter dominante.

En un experimento de $n = 8023$ semillas, resultaron $n_1 = 6022$ amarillas y $n_2 = 2001$ verdes. Las frecuencias relativas esperadas eran $\pi_1 = 0.75$ y $\pi_2 = 0.25$, por lo tanto $m_1 = 6017.25$ y $m_2 = 2005.75$.

Luego, si queremos testear la hipótesis nula

$$H_0 : \pi_1 = 0.75,\ \pi_2 = 0.25$$

el estadístico χ^2 es:

$$\chi^2 = \frac{(n_1 - 6017.25)^2}{6017.25} + \frac{(n_2 - 2005.75)^2}{2005.75} = 0.015$$

con un p-valor=0.88, lo que no contradice la teoría de Mendel.

Otro ejemplo

Los consultores para estudiantes de un centro de cómputos se encuentran con preguntas acerca de programas escritos en FORTRAN(1), en BASIC(2), PASCAL(3) y ADA(4). Los consultores han sido contratados basándose en el supuesto de que el 40 % de todas las preguntas se refieren a programas FORTRAN, 25 % a Basic, 25 % a Pascal y 10 % a ADA. Durante el primer mes se registran las consultas realizadas y se observa que las mismas fueron las dadas en la siguiente tabla. ¿Sostienen los datos el supuesto realizado al 5 %?

	FORTRAN	BASIC	PASCAL	ADA	TOTAL
Observ.	52	38	21	9	120

Sea $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$.

$$H_o : \boldsymbol{\pi} = (0.4, 0.25, 0.25, 0.1) \quad \text{vs.} \quad H_1 : \boldsymbol{\pi} \neq (0.4, 0.25, 0.25, 0.1)$$

A mano podríamos hacer:

```
Calculo los m_i:
> 120*0.40
  48
> 120*0.25
  30
> 120*0.10
  12
Caculo el estadístico de Pearson:
> pearson=((52-48)**2/48)+((38-30)**2/30)+((21-30)**2/30)+((9-12)**2/12)
> pearson
  5.916667
Calculo el p-valor:
> 1-pchisq(5.916667,3)
 0.1157357
```

O bien usar el chisq.test

```
> obs=c(52,38,21,9)
> null.probs <- c(0.40,0.25,0.25,0.10)
> chisq.test(obs, p=null.probs)
        Chi-squared test for given probabilities
data:  obs
X-squared = 5.9167, df = 3, p-value = 0.1157
```

Cuando $\boldsymbol{\pi}$ puede yacer en cualquier lugar de $\mathcal{S}$ decimos que el modelo es saturado. Este modelo tiene $K-1$ parámetros. Sin embargo, con frecuencia suponemos que $\boldsymbol{\pi}$ yace en un subconjunto de menor dimensión de $\mathcal{S}$.

Por jemplo, los elementos de $\boldsymbol{\pi}$ podrían estar determinados por $q \leq K-1$ parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, como muestran los siguientes ejemplos.

Test de Independencia

Independencia en una tabla de 2×2

Supongamos que $\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22})'$ es el vector de frecuencias de una tabla de 2×2.

Luego, X_{ij} es el número de individuos para los cuales $(A = i, B = j)$. Si A y B no están relacionados, entonces en todas las casillas valdrá:

	$B=1$	$B=2$	
$A=1$	X_{11}	X_{12}	α
$A=2$	X_{21}	X_{22}	$1-\alpha$
	β	$1-\beta$	

Cuadro 11:

$$\pi_{ij} = P(A=i, B=j) = P(A=i)P(B=j)$$

Llamemos $\alpha = P(A=i)$ y $\beta = P(B=j)$, luego

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} \alpha\beta \\ \alpha(1-\beta) \\ (1-\alpha)\beta \\ (1-\alpha)(1-\beta) \end{bmatrix}$$

Este es un modelo restringido que depende del parámetro

$$\boldsymbol{\theta} = (\alpha, \beta)' ,$$

donde $0 \leq \alpha \leq 1, \quad 0 \leq \beta \leq 1$.

Para hallar los estimadores de máxima verosimilitud de α y β tenemos que maximizar:

$$\begin{aligned} L &= L(X_{11}, X_{12}, X_{21}, X_{22}, \alpha, \beta) = \\ &= \frac{n!}{X_{11}!X_{12}!X_{21}!X_{22}!}(\alpha\beta)^{X_{11}}(\alpha(1-\beta))^{X_{12}}((1-\alpha)\beta)^{X_{21}}((1-\alpha)(1-\beta))^{X_{22}} \end{aligned}$$

Después de tomar logaritmo, obtenemos:

$$\begin{aligned} l = \ln(L) = cte &+ X_{11}\ln(\alpha\beta) + X_{12}\ln(\alpha(1-\beta)) \\ &+ X_{21}\ln((1-\alpha)\beta) + X_{22}\ln((1-\alpha)(1-\beta)) \end{aligned}$$

Después de derivar e igualar a 0, queda:

$$(1) \ : \ \frac{\partial l}{\partial \alpha} = \frac{X_{11} + X_{12}}{\alpha} - \frac{X_{21} + X_{22}}{1 - \alpha} = 0$$
$$(2) \ : \ \frac{\partial l}{\partial \beta} = \frac{X_{11} + X_{21}}{\beta} - \frac{X_{12} + X_{22}}{1 - \beta} = 0$$

por lo tanto

$$\widehat{\alpha} = \frac{X_{11} + X_{12}}{n} .$$

$$\widehat{\beta} = \frac{X_{11} + X_{21}}{n} .$$

En el caso general, es decir en las tablas de contingencia de $I \times J$ con muestreo multinomial puede ser de interés testear la hipótesis de independencia, es decir:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad \forall i, j$$

es decir, la hipótesis nula depende de ciertos parámetros.

Por esto si bien para testear esta hipótesis usaremos un test de tipo Pearson, antes será necesario estudiar la distribución asintótica de dicho estadístico bajo estas condiciones.

Otro ejemplo es el de las tablas simétricas.

Ejemplo: Tabla de 2×2 con simetría

Consideremos $\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22})'$ como en el ejemplo anterior, pero supongamos) que ahora A y B representan dos características medidas en dos oportunidades distintas. Por ejemplo, A podría ser la respuesta a

$$A : \textbf{¿Apoya usted la gestión de gobierno?}$$

medida en el mes de enero (1=Si, 0=No) y B la misma pregunta hecha tres

meses después.

	Abril	
Enero	1	0
1	π_{11}	π_{12}
0	π_{21}	π_{22}

Cuadro 12: Ejemplo Simetría

En este tipo de esquema el interés del investigador es detectar un cambio en el tiempo. Si no hubiera ningún cambio, la probabilidad de "Si en enero"

$$P(A = 1) = \pi_{11} + \pi_{12}$$

sería igual a la probabilidad de "Si" tres meses después

$$P(B = 1) = \pi_{11} + \pi_{21} \, .$$

Observemos $P(A = 1) = P(B = 1)$ si y sólo si $\pi_{12} = \pi_{21}$, que se conoce como la condición de **simetría**. Bajo simetría, $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$ podría expresarse como:

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \\ \beta \\ 1-\alpha-2\beta \end{bmatrix},$$

con $\boldsymbol{\theta} = (\alpha, \beta)'$.

Será un ejercicio de la práctica probar que los EMV bajo este modelo son:

$$\widehat{\alpha} = \frac{X_{11}}{n}\,.$$

$$\widehat{\beta} = \frac{X_{12} + X_{21}}{2n}\,.$$

En algunos casos más complicados, los estimadoresde máxima verosimilitud no tiene una expresión cerrada y deben ser computados a través de un método iterativo, por ejemplo Newton–Raphson.

Aun cuando el método para calcular $\widehat{\boldsymbol{\theta}}$ va a cambiar de modelo en modelo, bajo condiciones de regularidad estos estimadores tendrán algunas propiedades comunes.

En los problemas regulares, $\widehat{\boldsymbol{\theta}}$ es solución de las ecuaciones en *scores*

$$\frac{\partial l(\boldsymbol{\pi}(\boldsymbol{\theta}), \mathbf{X})}{\partial \theta_j} = 0\,, \quad j = 1, \ldots, q\,. \tag{4}$$

y $\widehat{\boldsymbol{\theta}}$ es una función suave de las proporciones muestrales $\mathbf{X}/n$.

Supongamos que las probabilidades de las casillas son $\pi_1(\boldsymbol{\theta}), \ldots, \pi_K(\boldsymbol{\theta})$ que involucran a q parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ y que:

- a) θ_0, el verdadero valor del parámetro, es un punto interior del espacio paramétrico.
- b) $\pi_i(\theta_0) > 0 \; \forall i$.
- c) Cada $\boldsymbol{\pi}(\boldsymbol{\theta})$ admite derivadas parciales continuas de 1^{er} orden en un entorno de θ_0.
- d) La matriz $\mathbf{M} = \{\dfrac{\partial \pi_r(\boldsymbol{\theta})}{\partial \theta_s}\}$ de $K \times q$ evaluada en $\boldsymbol{\theta}_0$ tiene rango q, donde $q \leq K - 1$.

Notemos que $\mathbf{M}$ es

$$\begin{bmatrix} \frac{\partial \pi_1}{\partial \theta_1} & \frac{\partial \pi_1}{\partial \theta_2} & \ldots & \frac{\partial \pi_1}{\partial \theta_q} \\ \frac{\partial \pi_2}{\partial \theta_1} & \frac{\partial \pi_2}{\partial \theta_2} & \ldots & \frac{\partial \pi_2}{\partial \theta_q} \\ \vdots & \vdots & \ddots & \\ \frac{\partial \pi_K}{\partial \theta_1} & \frac{\partial \pi_K}{\partial \theta_2} & \ldots & \frac{\partial \pi_K}{\partial \theta_q} \end{bmatrix}$$

El requerimiento de que el jacobiano $\mathbf{M} = \{\frac{\partial \pi_r(\boldsymbol{\theta})}{\partial \theta_s}\}$ tenga rango completo asegura que el problema es identificable. Por ejemplo, esto se violaría si diferentes $\boldsymbol{\theta}$ dan lugar a iguales vectores $\boldsymbol{\pi}(\boldsymbol{\theta})$. En este caso, $\boldsymbol{\theta}$ no podria ser estimado consistentemente a partir de la muestra $\mathbf{X}$ y sería necesaria información adicional.

Bajo la condición fuerte de identificabilidad y continuidad de las funciones $\pi_i(\boldsymbol{\theta})$, se puede demostrar que el EMV de $\boldsymbol{\theta}$ existe y que converge a $\boldsymbol{\theta}_0$ con probabilidad 1.

Un resultado importante bajo estas condiciones de regularidad el EMV $\widehat{\boldsymbol{\theta}}$ es asintóticamente normal, es decir

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_q(0, (\mathbf{A}'\mathbf{A})^{-1})$$

donde

$$\mathbf{A} = \Delta(\boldsymbol{\pi}(\boldsymbol{\theta}_0))^{-1/2} \left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} .$$

En la práctica, este resultado es importante ya que a partir de él se pueden deducir intervalos de confianza y tests de hipótesis para las componentes de $\boldsymbol{\theta}$ o funciones de ellas si se reemplaza a $\mathbf{A}$ por

$$\widehat{\mathbf{A}} = \Delta(\widehat{\boldsymbol{\pi}})^{-1/2} \left(\frac{\partial \boldsymbol{\pi}}{\partial \widehat{\boldsymbol{\theta}}}\right) .$$

Aplicando todos los resultados anteriores obtenemos que:

$$\sqrt{n}(\boldsymbol{\pi}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\pi}(\boldsymbol{\theta}_0)) \xrightarrow{D} N(0, \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} (\mathbf{A}'\mathbf{A})^{-1} \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}_0)'}{\partial \boldsymbol{\theta}})$$

La estimación de la matriz de covarianza de $\boldsymbol{\pi}(\widehat{\boldsymbol{\theta}})$ es

$$\frac{1}{n} \frac{\partial \boldsymbol{\pi}(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} (\widehat{\mathbf{A}}'\widehat{\mathbf{A}})^{-1} \frac{\partial \boldsymbol{\pi}(\widehat{\boldsymbol{\theta}})'}{\partial \boldsymbol{\theta}} \quad \text{o bien} \quad \frac{1}{n} \frac{\partial \widehat{\boldsymbol{\pi}}}{\partial \boldsymbol{\theta}} (\widehat{\mathbf{A}}'\widehat{\mathbf{A}})^{-1} \frac{\partial \widehat{\boldsymbol{\pi}}'}{\partial \boldsymbol{\theta}}$$

En este punto, cabe observar que habría dos estimaciones posibles de Π. Una, dada por la relación entre cada π_i y $\boldsymbol{\theta}$ y otra dada por las proporciones muestrales.

Parece razonable comparar estas dos estimaciones con el estadístico χ^2:

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - \widehat{m}_i)^2}{\widehat{m}_i} = \sum_{i=1}^{K} \frac{(n_i - n\widehat{\pi}_i)^2}{n\widehat{\pi}_i} \; .$$

El estadístico χ^2 puede escribirse como

$$\chi^2 = n\|\mathbf{e}\|^2$$

donde el vector de residuos $\mathbf{e}$ se define como

$$\mathbf{e}' = \left(\frac{p_1 - \pi_1(\widehat{\boldsymbol{\theta}})}{\sqrt{\pi_1(\widehat{\boldsymbol{\theta}})}}, \ldots, \frac{p_K - \pi_K(\widehat{\boldsymbol{\theta}})}{\sqrt{\pi_K(\widehat{\boldsymbol{\theta}})}} \right).$$

Para derivar la distribución asintótica de χ^2 se necesita la conjunta de $(\mathbf{p}, \boldsymbol{\pi}(\widehat{\theta}))$.

Se puede ver que

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{(a)}{=} (\mathbf{A}'\mathbf{A})^{-1}\,\mathbf{A}'\;\Delta(\boldsymbol{\pi}_0^{-1/2})\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0)$$

$$(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(n^{-1/2})$$

Por lo tanto

$$\sqrt{n}\begin{pmatrix} \mathbf{p} - \boldsymbol{\pi}_0 \\ \widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{I} \\ \mathbf{D} \end{pmatrix}\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}_0) + o_p(1)$$

de donde

$$\sqrt{n}\begin{pmatrix} \mathbf{p} - \boldsymbol{\pi}_0 \\ \widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0 \end{pmatrix} \xrightarrow{D} N(0, \Sigma^*)$$

donde

$$\Sigma^* = \begin{pmatrix} \Delta(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0' & (\Delta(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')\mathbf{D}' \\ \mathbf{D}(\Delta(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0) & \mathbf{D}(\Delta(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0\boldsymbol{\pi}_0')\mathbf{D}' \end{pmatrix}$$

Luego, aplicando el método Δ resulta que

$$\sqrt{n}\,\mathbf{e} \xrightarrow{\mathcal{D}} N(0, \mathbf{I} - \boldsymbol{\pi}(\theta_0)^{1/2}\boldsymbol{\pi}'(\theta_0)^{1/2} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')$$

Teorema: Sea $\mathbf{Y}$ un vector con distribución $N(\nu, \Sigma)$. Una condición necesaria y suficiente para que $(\mathbf{Y}-\nu)'\mathbf{C}(\mathbf{Y}-\nu)$ tenga distribución χ^2 es que $\Sigma\mathbf{C}\Sigma\mathbf{C}\Sigma = \Sigma\mathbf{C}\Sigma$, donde los grados de libertad serán el rango de $\mathbf{C}\Sigma$ (si Σ es no singular la condición se simplifica a $\mathbf{C}\Sigma\mathbf{C} = \mathbf{C}$). (Rao, 1965, p. 150)

Dado que $\chi^2 = \sqrt{n}\mathbf{e}'\sqrt{n}\mathbf{e}$, luego aplicaremos el resultado de Rao con $\nu = 0$, $\mathbf{C} = I$, $\Sigma = \mathbf{I} - \boldsymbol{\pi}(\boldsymbol{\theta}_0)^{1/2}\boldsymbol{\pi}'(\boldsymbol{\theta}_0)^{1/2} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$). Como Σ aquí es una matriz idempotente, resulta que su rango coincide con su traza y en consecuencia

$$\chi^2 = \sqrt{n}\mathbf{e}'\sqrt{n}\mathbf{e} \xrightarrow{\mathcal{D}} \chi^2_{K-1-q}$$

Ver Rao, Capítulo 5e y Agresti, Capítulo 14.

Volviendo al Test de Independencia:

En una tabla de $I \times J$ con muestreo multinomial, la hipótesis nula de independencia equivale a

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad \forall i, j$$

Usando el estadístico de Pearson tendríamos

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \widehat{m}_{ij})^2}{\widehat{m}_{ij}},$$

donde $\widehat{m}_{ij} = n_{i+}n_{+j}/n$.

Respecto a los grados de libertad, estos están determinados por la cantidad de casillas y de parámetros, que en este caso serán

$$I * J - ((I-1) + (J-1)) - 1 = (I-1) * (J-1)\,.$$

Retomemos el ejemplo en que se clasifica a los individuos según su **Identificación Partidaria** y el **Género**. En la siguiente tabla tenemos los valores observados y en rojo los valores predichos bajo el modelo de independencia

	C: Identificación partidaria			
S:Género	Demócrata	Independiente	Republicano	Total
F	279	73	225	577
	(261.4)	(70.7)	(244.9)	
M	165	47	191	403
	(182.6)	(49.3)	(171.1)	
Total	444	120	416	980

Cuadro 13: Datos de la General Social Survey, 1991.

El valor del estadístico test de χ^2 es 7.01, con

$IJ-1-(I-1)-(J-1) = (I-1)(J-1) = (2-1)(3-1) = 2$ grados de libertad .

El p–valor correspondiente es 0.03, de manera que a los valores habituales se rechazaría la hipótesis de independencia, indicando que el género y la identificación partidaria estarían asociados.

Otro Ejemplo

Este es otro ejemplo en que las probabilidades dependen de una cantidad menor de parámetros desconocidos, θ.

Una muestra de 156 terneros nacidos en Florida fueron clasificados de acuerdo a que hayan contraído neumonía dentro de los 60 días de haber nacido. Los terneros que contrajeron neumonía fueron a su vez clasificados según se hayan infectado o no a los 15 días de haberse curado. La Tabla muestra los datos recolectados:

	Segunda Infección	
	Si	No
Primera Infección		
Si	30	63
No	0	63

Cuadro 14: Terneros de Florida

Es claro que los terneros que no tuvieron una primera infección no pudieron

reinfectarse, es por ello que ninguna observación puede verse en la casilla 21 y en consecuencia en la tabla $n_{21} = 0$. Esto es lo que se conoce como un **cero estructural**. El objetivo en este estudio era testear si la probabilidad de una primera infección era igual que la probabilidad de una segunda infección, dado que el ternero había contraído una primera infección.

Es decir, la hipótesis a testear es

$$H_0 : \pi_{11} + \pi_{12} = \frac{\pi_{11}}{\pi_{11} + \pi_{12}}$$

o equivalentemente $\pi_{11} = (\pi_{11} + \pi_{12})^2$. De manera que si llamamos $\pi = \pi_{11} + \pi_{12}$ a la probabilidad de infección primaria, el modelo bajo H_0 corresponde a una **trinomial** como muestra la siguiente tabla:

En este caso el likelihood resulta proporcional a

$$(\pi^2)^{n_{11}} (\pi(1-\pi))^{n_{12}} (1-\pi)^{n_{22}} ,$$

Primera Infección	**Segunda Infección** Si	No	Total
Si	π^2	$\pi(1-\pi)$	π
No	–	$1-\pi$	$1-\pi$

Cuadro 15: Distribución Conjunta

el log–likelihood queda

$$n_{11}\log(\pi^2) + n_{12}\log(\pi - \pi^2) + n_{22}\log(1-\pi)\,,$$

Derivando e igualando a 0 resulta

$$\widehat{\pi} = \frac{2n_{11} + n_{12}}{2n_{11} + 2n_{12} + n_{22}}$$

En la siguiente tabla se muestran en rojo (2do. renglón) los valores esperados bajo H_0

	Segunda Infección	
Primera Infección	Si	No
Si	30	63
	(38.1)	(39)
No	0	63
	(–)	(78.9)

Cuadro 16:

El estadístico de Pearson da χ^2 =19.7 con un total de (3-1)-1=1 grados de libertad. Dado que el p–valor es 0.00001 hay una fuerte evidencia contra H_0. Si miramos la tabla encontramos que muchos más terneros contraen una primera infección y no la segunda de lo que el modelo bajo H_0 predice. Con esto los investigadores concluyeron que la primera infección tiene un efecto inmunizador.

Estadístico G^2

Otra medida alternativa para la distancia entre $\widehat{\boldsymbol{\pi}}$ y $\mathbf{p}$ muy usada es la **deviance** G^2, que es un estadístico basado en el cociente de verosimilitud.

Si queremos testear

$$\begin{aligned} H_0 &: \text{Modelo restringido } \omega \\ H_1 &: \text{Modelo Saturado } \Omega\,, \end{aligned}$$

el cociente estaría dado por

$$\Lambda = \frac{\max\limits_{\omega} L}{\max\limits_{\Omega} L}$$

Si consideramos $G^2 = -2\log\Lambda$ queda definido el estadístico como

$$\begin{aligned} G^2 &= -2\log\Lambda = 2[l(\mathbf{p},\mathbf{X}) - l(\widehat{\boldsymbol{\pi}},\mathbf{X})] \\ &= 2\left[\sum_{i=1}^{K} X_i \log p_i - \sum_{i=1}^{K} X_i \log \widehat{\pi}_i\right] \end{aligned}$$

$$= 2 \sum_{i=1}^{K} X_i \log \frac{p_i}{\widehat{\pi}_i} = 2 \sum_{i=1}^{K} X_i \log \frac{X_i}{n\widehat{\pi}_i}$$

Por lo tanto,

$$G^2 = 2 \sum_{i=1}^{K} X_i \log \frac{X_i}{\widehat{\mu}_i} .$$

Probaremos que bajo H_0 la distribución límite de G^2 es también χ^2 con $K - 1 - \#\{$ parámetros bajo $\omega\}$, es decir la misma distribución límite que la del estadístico de Pearson.

Para derivar la distribución asintótica bajo H_0, se prueba que $G^2 - \chi^2 \xrightarrow{p} 0$ cuando H_O es cierta.

Una ventaja de G^2 es que tiene sentido en modelos más generales, como ya veremos.

En el ejemplo de **Identificación Partidaria vs. Sexo**, $G^2 = 7$, que da también un p–valor de 0.03.

Efecto de observar ceros

Si en alguna celda se observa un 0, el estadístico χ^2 puede calcularse sin problemas, siempre que las $\widehat{\pi}$'s sean todas positivas. Sin embargo, el estadístico G^2 tiene problemas, pues si $X_i = 0$, entonces $X_i \log \frac{X_i}{n\widehat{\pi}_i}$ no está definido. Si reescribimos a G^2 como

$$\begin{aligned} G^2 &= -2 \log \Lambda = 2 \log \frac{L(\mathbf{p}, \mathbf{X})}{L(\widehat{\boldsymbol{\pi}}, \mathbf{X})} \\ &= 2 \log \prod_{i=1}^{K} \left(\frac{X_i/n}{\widehat{\pi}_i} \right)^{X_i} \end{aligned}$$

una celda con un 0 aportaría un 1 al producto y por lo tanto, puede ser ignorada. Luego podemos calcular a G^2 como

$$2 \sum_{i:\, x_i > 0} X_i \log \frac{X_i}{n\widehat{\pi}_i}$$

Si alguna $\widehat{\pi}_i$ es 0, los dos estadísticos se *rompen*.

¿Cuán grande debe ser n para tener una buena aproximación?

Sabemos que a medida que n se hace más grande la distribución de χ^2 y de G^2 se aproximan a una distribución límite χ^2, sin embargo nos preguntamos cuán grande es grande.

• Una vieja regla conocida para las binomiales dice que la aproximación χ^2 es buena si $n\widehat{\pi}_i \geq 5$, $i = 1, \ldots, K$.

• Otra regla más permisiva establece que la aproximación χ^2 es buena si a lo sumo el 20 % de las casillas tienen $n\widehat{\pi}_i < 5$, $i = 1, \ldots, K$ y ninguna casilla tiene $n\widehat{\pi}_i < 1$.

• En tablas *sparse* (por ejemplo, $n/K < 5$) la aproximación χ^2 es pobre. En realidad, si los datos están distribuidos en la tabla de foma muy desigual, en el sentido de que hay zonas de la tabla que son *sparse*, la aproximación χ^2 también puede ser pobre, aún cuando el n total sea grande.

Hemos probado que los dos estadísticos se aproximan a 0, si el modelo es cierto. Si el modelo no es cierto, ambos crecen, pero la diferencia entre ambos también puede crecer. De manera, que si el modelo tiene un ajuste pobre los

dos estadísticos pueden ser grandes y estar lejos uno de otro, i.e., $|\chi^2 - G^2|$ no necesariamente tiende a 0 con n. Aún en esa situación, los correspondientes p–valores pueden estar cerca de 0 y podemos llegar a la misma conclusión a partir de ellos.

Para ser más precisos, consideremos una sucesión de situaciones $\boldsymbol{\pi}_n$ para las cuales la falta de ajuste disminuye con n, es decir alternativas contiguas. Supongamos que el modelo bajo la hipótesis nula es $\boldsymbol{\pi} = \mathbf{f}(\theta)$, pero en realidad

$$\boldsymbol{\pi}_n = \mathbf{f}(\theta) + \boldsymbol{\delta}/\sqrt{n}\,,$$

con $\sum_{i=1}^{K} \delta_i = 0$. Luego, si $\boldsymbol{\delta} = 0$, el modelo es cierto.

Para estas alternativas contiguas, Mitra (1958) demostró que el estadístico de Pearson tiene distribución asintótica χ^2 no central, con $N - 1 - q$ grados de libertad, con parámetro de no centralidad dado por

$$\lambda = n \sum_{i=1}^{N} \frac{(\pi_{ni} - f_i(\theta))^2}{f_i(\theta)}$$

Notemos que λ tiene la forma del estadístico χ^2 en el que se reemplazó a $\mathbf{p}$ por $\boldsymbol{\pi}_n$ y a $\widehat{\boldsymbol{\pi}}$ por $\mathbf{f}(\boldsymbol{\theta})$. Análogamente, utilizando los mismos reemplazos obtenemos el parámetro de no centralidad de G^2. Haberman (1974) demostró que bajo ciertas condiciones χ^2 y G^2 tienen el mismo parámetro de no centralidad, pero éste no es siempre el caso, (ver Agresti,2002, pag 590).

Residuos de Pearson y deviance

Como ya hemos visto podemos escribir al estadístico de Pearson como

$$\chi^2 = n \sum_{i=1}^{N} e_i^2 \, .$$

A

$$\epsilon_i = \sqrt{n}\, \frac{p_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i}} = \frac{n_i - \widehat{m}_i}{\sqrt{\widehat{m}_i}}$$

se lo conoce como el i–ésimo residuo de Pearson.

Estos residuos se comportan de alguna manera como los residuos estandarizados que conocimos en regresión lineal.

Teniendo en cuenta la distribución asintótica de los residuos podemos considerar una versión standarizada que es asintóticamente normal standard (Haberman, 1975). Para H_o : Independencia serían

$$\epsilon_i = \frac{n_i - \widehat{m}_i}{[\widehat{m}_i(1 - p_{i+})(1 - p_{+j})]^{1/2}}$$

Es común que se compare a $|\epsilon_i|$ con 2 (algunos autores comparan con 3),

indicándose falta de ajuste en la i–ésima casilla si $|\epsilon_i| > 2\,(3)$. El análisis de estos residuos puede sugerirnos en que sentido los datos se apartan del modelo ajustado.

De la misma forma, la deviance puede interpretarse como una suma de cuadrados de residuos

$$G^2 = \sum_{i=1}^{N} r_i^2 sgn(X_i - n\widehat{\pi}_i)$$

donde

$$r_i = \sqrt{\left|2X_i \log \frac{X_i}{n\widehat{\pi}_i}\right|} \times sgn(X_i - n\widehat{\pi}_i)$$

que se conocen como componentes de la deviance.

Veamos un ejemplo. En la siguiente tabla se muestran los resultados de un estudio en el que se clasifican los individuos de la muestra de acuerdo a su grado de fundamentalismo religioso y el nivel de educación alcanzado.

	Creencia Religiosa			
Nivel Educación	Fundamentalista	Moderado	Liberal	Total
Menos que High School	178	138	108	424
	(137.8)	(161.5)	(124.7)	
	(4.5)	(-2.6)	(-1.9)	
High School o junior College	570	648	442	1660
	(539.5)	(632.1)	(488.4)	
	(2.6)	(1.3)	(-4.0)	
Bachelor o Graduado	138	252	252	642
	(208.7)	(244.5)	(188.9)	
	(-6.8)	(0.7)	(6.3)	
Total	886	1038	802	2726

Cuadro 17: General Social Survey, National Opinion Research Center, 1996

Si se testea la hipótesis H_o : independencia, el estadístico χ^2=69.2 y el G^2=69.8 con g.l.=(3-1)(3-1)=4. Los p–valores son <0.0001.

En paréntesis se muestran los valores esperados ajustados bajo el modelo y los residuos standard.

El primero, por ejemplo, se calculó como

$$\frac{178 - 137.8}{[137.8(1 - 0.156)(1 - 0.325)]^{1/2}} = 4.5$$

Esta celda muestra una discrepancia importante entre n_{11} y su valor esperado ajusrado bajo independencia $\widehat{\mu}_{11}$.

Así vemos que los mayores desajustes se producen en los niveles de educación más alto, para las categorías Fundamentalista y Liberal.

Medidas de Asociación

A fin de describir el grado de asociación entre las variables de una tabla de contingencia es frecuente que se usen distintas medidas.

Comenzaremos con tablas de 2×2, como las que siguen

	Y		
X	1	2	Total
1	π_{11}	π_{12}	π_{1+}
2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

	Y		
X	1	2	Total
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Consideremos la siguiente tabla que corresponde a un informe sobre la relación entre el uso de aspirina y el infarto de miocardio realizado por el Physicians Health Study Research Group de Harvard Medical School:

	Infarto de Miocardio		
	si	no	Total
Aspirina	104	10933	11037
Placebo	189	10845	11034

Diferencia de Proporciones o Riesgo Atribuible

Miremos a Y como variable de respuesta y a X como variable explicativa, tal como seríia natural en un muestreo de producto multinomial en que

$$n_{11} \sim Bi(n_{1+}, \frac{\pi_{11}}{\pi_{1+}}) \text{ y } n_{21} \sim Bi(n_{2+}, \frac{\pi_{21}}{\pi_{2+}})$$

independientes.

La diferencia de proporciones se define como

$$\begin{aligned} \delta &= P(Y=1|X=1) - P(Y=1|X=2) \\ &= \frac{\pi_{11}}{\pi_{1+}} - \frac{\pi_{21}}{\pi_{2+}} \\ &= \pi_{1|1} - \pi_{1|2} \end{aligned}$$

Podemos estimar a δ como

$$\begin{aligned} d &= \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} \\ &= p_{1|1} - p_{1|2} \end{aligned}$$

En el ejemplo de Infarto de Miocardio tenemos

$$d = 104/11037 - 189/11034 = \text{ 0.0094 - 0.0171 = -0.0077}$$

Observemos que

$$\begin{aligned} E(d) &= E(p_{1|1} - p_{1|2}) = \pi_{1|1} - \pi_{1|2} \\ V(d) &= V(p_{1|1} - p_{1|2}) = \frac{\pi_{1|1}(1-\pi_{1|1})}{n_{1+}} + \frac{\pi_{1|2}(1-\pi_{1|2})}{n_{2+}} \end{aligned}$$

siendo la última igualdad cierta por la independencia entre las filas.

Si n_{1+} y n_{2+} son grandes, d es aproximadamente normal, es decir

$$\frac{(p_{1|1} - p_{1|2}) - (\pi_{1|1} - \pi_{1|2})}{\sqrt{\frac{\pi_{1|1}(1-\pi_{1|1})}{n_{1+}} + \frac{\pi_{1|2}(1-\pi_{1|2})}{n_{2+}}}}$$

es aproximadamente $N(0,1)$. Por lo tanto haciendo un plug–in para estimar la varianza podemos obtener un intervalo asintótitco para δ de nivel $1-\alpha$ como

$$d \pm z_{\alpha/2}\sqrt{\frac{p_{1|1}(1-p_{1|1})}{n_{1+}}+\frac{p_{1|2}(1-p_{1|2})}{n_{2+}}}$$

$$p_{1|1}-p_{1|2} \pm z_{\alpha/2}\sqrt{\frac{p_{1|1}(1-p_{1|1})}{n_{1+}}+\frac{p_{1|2}(1-p_{1|2})}{n_{2+}}}$$

Riesgo Relativo

Notemos que que la diferencia entre 41 % y 40.1 % es la misma que entre 1 % y 0.1 %. Sin embargo, 1 % es diez veces 0.1 %. Este es un problema de la diferencia de proporciones como medida. Si estamos trabajanos con eventos poco frecuentes $\pi_{1|1}$ y $\pi_{1|2}$ serán muy pequeñas y δ será cercano a 0, aún cuando el efecto sea importante, como en el ejemplo anterior.

Esto es frecuente en epidemiología en donde la prevalencia de ciertas enfermedades es muy baja.

Esto sugiere la conveniencia de considerar una medida relativa como el **riesgo relativo**

$$RR = \frac{P(Y=1|X=1)}{P(Y=1|X=2)} = \frac{\pi_{11}/\pi_{1+}}{\pi_{21}/\pi_{2+}}$$

El riesgo relativo es una medida no negativa y un riego relativo igual a 1 corresponde a independencia.

El EMV de RR es

$$rr = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

En el ejemplo quedaría:

$$rr = \frac{0.0094}{0.0171} = 0.55\,,$$

esto significa que el riesgo de infarto de miocardio en el grupo tratado con aspirina es aproximadamente la mitad que en grupo que recibió placebo.

Dado que podemos aproximar mediante una normal a su logaritmo suele usarse

como medida $\log(RR)$, que se estima por $\log(rr) = \log p_{1|1} - \log p_{1|2}$.

Sabemos que

$$\sqrt{n_{i+}}\,(p_{1|i} - \pi_{1|i}) \xrightarrow{D} N(0, \pi_{1|i}(1 - \pi_{1|i}))\,,$$

luego usando el método Δ obtenemos que

$$\sqrt{n_{i+}}\,(\log p_{1|i} - \log \pi_{1|i}) \xrightarrow{D} N\left(0, \frac{(1 - \pi_{1|i})}{\pi_{1|i}}\right)\,.$$

Por la independencia entre las filas, obtenemos que la varianza asintótica de $\log(rr)$ es

$$V(\log(rr)) \simeq \frac{(1 - \pi_{1|1})}{n_{1+}\pi_{1|1}} + \frac{(1 - \pi_{1|2})}{n_{2+}\pi_{1|2}}$$

y se puede estimar haciendo un plug–in por

$$\begin{aligned}\widehat{V}(\log(rr)) &\simeq \frac{(1 - p_{1|1})}{n_{1+}\,p_{1|1}} + \frac{(1 - p_{1|2})}{n_{2+}\,p_{1|2}} \\ &\simeq \frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}\end{aligned}$$

Un intervalo de nivel asintótico $1-\alpha$ para $\log(RR)$ es

$$\log(rr) \pm z_{\alpha/2}\sqrt{\widehat{V}(\log(rr))}$$

Como $\log(rr)$ no existe si algún $p_{1|i} = 0$ suele usarse

$$\log(\tilde{rr}) = \log\left(\frac{n_{11}+1/2}{n_{1+}+1/2}\right) - \log\left(\frac{n_{21}+1/2}{n_{2+}+1/2}\right)$$

RR es sólo función de $P(Y|X)$, por lo tanto la inferencia que hagamos sobre RR será la misma para los tres muestreos que hemos considerado. La comparación en la otra respuesta da otro riesgo relativo.

Odds Ratio (Producto Cruzado)

El riesgo relativo es el cociente de dos probabilidades. Podríamos comparar la probabilidad de **si** y de **no** en un mismo estrato. Eso nos lleva a la definición de **odss** o **chance**. El odds de un suceso A es

$$odds = \frac{P(A)}{1 - P(A)}$$

y toma cualquier valor mayor o igual a 0.

En el ejemplo, tenemos que para el grupo tratado el odds estimado resulta

$$0.0094/(1 - 0.0094) = 0.0094/0.9906 = 0.0095\,,$$

mientras que para el grupo placebo el odds estimado es

$$0.0171/(1 - 0.0171) = 0.0171/0.9829 = 0.0174\,.$$

En el grupo que recibió placebo la chance de tener infarto es 0.0174 la de no

tener infarto, mientras que en el grupo tratado la chance de infarto es 0.0095 la de no tener infarto.

Dicho de otra manera, la chance de tener infarto respecto de la de no tenerlo en el grupo placebo es aproximadamente el doble que la obtenida en en el grupo tratado.

Podríamos comparar los dos odds, por ejemplo considerando su cociente, esto da origen a

$$\theta = \text{odds ratio} = \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}$$

$$= \frac{\left[\frac{\pi_{11}}{\pi_{1+}}\right] / \left[\frac{\pi_{12}}{\pi_{1+}}\right]}{\left[\frac{\pi_{21}}{\pi_{2+}}\right] / \left[\frac{\pi_{22}}{\pi_{2+}}\right]}$$

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

Esta medida es función de $P(Y|X)$, la inferencia es válida para los tres muestreos vistos.

El EMV es

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

Las propiedades de $\widehat{\theta}$ son fáciles de deducir bajo muestreo multinomial, pero también son válidas con muestreo Poisson o Producto Multinomial en el que los totales marginales por filas o bien por columnas están fijos.

Como con el riesgo relativo podemos deducir un intervalo de nivel asintótico $1-\alpha$ para $\log(\widehat{\theta})$

$$\log(\widehat{\theta}) \pm z_{\alpha/2}\sqrt{\widehat{V}(\log\widehat{\theta})}$$

donde

$$\widehat{V}(\log(\widehat{\theta})) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Notemos además que si intercambiamos los roles de X e Y, obtenemos

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

por lo que también puede ser visto como función de $P(X|Y)$, que correspondería a tener n_{+j} fijos. El hecho de que los roles de X e Y puedan ser intercambiados es una propiedad interesante, pues puede ser de gran utilidad pues permite usar estudios restropectivos.

Estimación por Intervalos

Podría ser de interés obtener intervalos de confianza para $(\pi_1, \ldots, \pi_K)$ o para $\{\pi_i - \pi_j : 1 \leq i < j \leq K\}$

Recordemos distribución asintótica de $(\frac{n_1}{n}, \ldots, \frac{n_{K-1}}{n})$ cuando

$$(n_1, \ldots, n_K)' \sim M(n, \pi_1, \ldots, \pi_K) \qquad \sum_{i=1}^{K} \pi_i = 1.$$

Como antes llamemos $\mathbf{p} = (p_1, \ldots, p_K)'$, $p_i = \dfrac{n_i}{n}$.

Vimos que

$$\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{\mathcal{D}} N_K(0, \Delta(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}')\,.$$

Ya hemos visto que como los π_i's están relacionados, $\Sigma = \Delta(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$ no es invertible.

Si como antes definimos $\tilde{\mathbf{p}} = (p_1, \ldots, p_{K-1})'$ y $\tilde{\boldsymbol{\pi}} = (\pi_1, \ldots, \pi_{K-1})'$, tenemos que

$$\sqrt{n}(\tilde{\mathbf{p}} - \tilde{\boldsymbol{\pi}}) \xrightarrow{\mathcal{D}} N_{K-1}(0, \Delta(\tilde{\boldsymbol{\pi}}) - \tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}') \,,$$

donde $\Delta(\tilde{\boldsymbol{\pi}}) - \tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{\pi}}'$ sí es invertible.

Gold (1963) define la región de confianza de nivel asintótico $1 - \alpha$

$$R^G = \left\{ \mathbf{w} \in I\!R^{K-1} : (\tilde{\mathbf{p}} - \mathbf{w})' \widehat{\tilde{\Sigma}}^{-1} (\tilde{\mathbf{p}} - \mathbf{w}) \leq \frac{\chi^2_{\alpha,K-1}}{n} \right\}$$

donde $\widehat{\tilde{\Sigma}} = \Delta(\tilde{\mathbf{p}}) - \tilde{\mathbf{p}}\tilde{\mathbf{p}}'$.

Aplicando el método de proyecciones de Scheffé tenemos que:

$$\left\{ \tilde{\mathbf{p}} \in R^G \right\} = \cap_{\mathbf{l} \in I\!R^{K-1}} \left\{ \mathbf{l}'\tilde{\mathbf{p}} \in \mathbf{l}'\tilde{\mathbf{p}} \pm \left(\frac{\chi^2_{\alpha,K-1}}{n} \right)^{\frac{1}{2}} \left(\mathbf{l}' \widehat{\tilde{\Sigma}}^{-1} \mathbf{l} \right)^{\frac{1}{2}} \right\}$$

Como caso especial la subfamilia de direcciones $(1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 1), \ldots (-1, -1, \ldots, -1)$ da por resultado los intervalos de confianza de nivel asintótico para $\pi_1, \ldots, \pi_K$

$$I_i^G = p_i \pm \left(\frac{p_i(1-p_i)}{n}\right)^{\frac{1}{2}} \{\chi^2_{\alpha,K-1}\}^{\frac{1}{2}}$$

Quesenberry y Hurst (1964) propusieron intervalos de confianza análogos a los propuestos por Gold, pero basados en la matriz $\tilde{\Sigma}(\mathbf{w}) = \Delta(\tilde{\mathbf{w}}) - \tilde{\mathbf{w}}\tilde{\mathbf{w}}'$ en lugar de $\widehat{\tilde{\Sigma}}$. De esta forma los intervalos de confianza simultáneos resultan:

$$I_i^{QH} = \left\{\mathbf{w} \in I\!R^{K-1} : \frac{(p_i - w_i)^2}{w_i(1-w_i)} \leq \frac{\chi^2_{\alpha,K-1}}{n}\right\}$$

De todas formas, ninguna de estas dos soluciones es completamente satisfactoria en tanto según los estudios de Ghosh (1979) las aproximaciones no son del todo buenas y además los IC son conservativos en tanto se comienza con una elipse y se termina extrayendo un subconjunto finito de proyecciones.

Esto llevó a Goodman (1965) a considerar intervalos simultáneos basados en el criterio de Bonferroni. Recordemos que si los eventos E_j son tales que $P(E_j) \geq 1 - \frac{\alpha}{K}, 1 \leq j \leq K$, entonces la desigualdad de Bonferroni establece que

$$\begin{aligned} P\left[\bigcap_{j=1}^{K} E_j\right] &= 1 - P\left[\bigcup_{j=1}^{K} E_j^c\right] \\ &\geq 1 - \sum_{j=1}^{K} P\left[E_j^c\right] \geq 1 - \sum_{j=1}^{K} (\alpha/K) = 1 - \alpha \end{aligned}$$

La cota inferior es bastante fina para valores de α pequeños.

Goodman aplicó el criterio de Bonferroni a los eventos

$$E_j = \left\{ \frac{n(p_j - \pi_j)^2}{\pi_j(1 - \pi_j)} \leq \chi^2_{\alpha/K,1} \right\}$$

para los cuales $P(E_j) \to 1 - \alpha/K$ cuando $n \to \infty$, dando lugar a los intervalos de confianza asintóticos I_i^{GM}. Estos intervalos tienen la misma forma que los I_i^{QH} salvo por el percentil χ^2 utilizado.

En general, Goodman demostró que

$$\frac{\text{long. de GM interval}}{\text{long. de QH interval}} \rightarrow \left\{ \frac{\chi^2_{\alpha/K,1}}{\chi^2_{\alpha,K-1}} \right\}^2 \text{ en prob.}$$

Los intervalos de Bonferroni son superiores cuando el límite es inferior a 1. Goodman encontró que es menor que 1 para $\alpha \leq 0{,}10$ y para un amplio rango de valores de K. Por ejemplo, para α =0.10 y $K = 10$ el límite es 0.67.

El método de Gold de la elipse puede se usado para obtener intervalos de confianza silmutáneos para $\pi_i - \pi_j$.

Los intervalos resultantes son:

$$p_i - p_j \pm \{\chi^2_{\alpha,K-1}\}^{\frac{1}{2}} \left(\frac{p_i + p_j - (p_i - p_j)^2}{n} \right)^{\frac{1}{2}},$$

teniendo en cuenta que:

$$\begin{aligned} Var(p_i - p_j) &= Var(p_i) + Var(p_j) - 2Cov(p_i, p_j) \\ &= \frac{\pi_i(1-\pi_i)}{n} + \frac{\pi_j(1-\pi_j)}{n} + 2\frac{\pi_i\pi_j}{n} \\ &= \frac{\pi_i + \pi_j - (\pi_i - \pi_j)^2}{n} \end{aligned}$$