Datos Categóricos

Clase 8

Juan Carlos Correa

30 de marzo de 2022

Resultado Importante I *Suponga que* X_n *es* $AN\left(\mu, \sigma_n^2\right)$ *con* $\sigma_n \to 0$. *Sea* g *una función de valor real diferenciable en* $X = \mu$ *con* $g'(\mu) \neq 0$. *Entonces*

$$g(X_n) \sim AN\left(g(\mu), \left[g'(\mu)\right]^2 \sigma_n^2\right)$$

Resultado Importante II *Sea* $\mathbf{X}_n = (X_{n1}, X_{n2}, \cdots, X_{nk})'$ *y además asuma que*

$$\mathbf{X}_n \sim AN\left(\mu, b_n^2 \mathbf{\Sigma}\right)$$

con $\mathbf{\Sigma}$ *matriz de covarianzas y* $b_n \to 0$*.*

Sea $g(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \cdots, g_m(\mathbf{x}),)'$*, donde* $\mathbf{x} = (x_1, x_2, \cdots, x_k)'$*, una función con argumento un vector y donde cada componente es una función de valor real y tiene un diferencial no cero* $g_i(\mu; \mathbf{t})$*,* $\mathbf{x} = (t_1, t_2, \cdots, t_k)'$*, en* $\mathbf{x} = \mu$*. Haga*

$$\mathbf{D} = \left[\frac{\partial g_i}{\partial x_j} \mid_{\mathbf{x}=\mu}\right]_{m \times k}$$

Entonces $g\left(\mathbf{X}_n\right) \sim AN\left(g(\mu), b_n^2 \mathbf{D}\mathbf{\Sigma}\mathbf{D}'\right)$

La Razón de Odds

La siguiente tabla presenta el modelo poblacional para una tabla 2×2, donde cada celda presenta la probabilidad de ella.

	A	A^c
B	$P(A \cap B)$	$P(A^c \cap B)$
B^c	$P(A \cap B^c)$	$P(A^c \cap B^c)$

Los odds* de que el evento B ocurra relativo al evento A se define como la razón de las probabilidades

$$\frac{P\left[B \mid A\right]}{P\left[B^c \mid A\right]}$$

*La palabra *odds* no tiene una única y precisa traducción, algunos la traducen como disparidad y otros como apuestas.

La interpretación de la razón anteriror es directa: Asumiendo que el evento A ha ocurrido, esta razón nos dice cúantas veces ocurre el evento B por cada aparición del evento B^c.

Los odds de B relativo a A^c son

$$\frac{P\left[B \mid A^c\right]}{P\left[B^c \mid A^c\right]}$$

Cornfield (1951) definió la razón de odds como

$$\psi = \frac{\frac{P[B|A]}{P[B^c|A]}}{\frac{P[B|A^c]}{P[B^c|A^c]}}$$

Fisher (1962) la llamó *Razón del Producto Cruzado.*

El estimador muestral de ψ será

$$r = \frac{\left(\frac{\frac{n_{11}}{n_{+1}}}{\frac{n_{21}}{n_{+1}}}\right)}{\left(\frac{\frac{n_{12}}{n_{+2}}}{\frac{n_{22}}{n_{+2}}}\right)} = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

para lo anterior, se presupone una tabla conteos de como la que aparece a continuación

	A	A^c	
B	n_{11}	n_{12}	n_{1+}
B^c	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{n+2}	

	Relaciones?	
Género	Sí	No
Masculino	368	284
Femenino	148	218

```
> datos<-matrix(c(368 , 284,148 , 218),ncol=2,byrow=T)
> proba.cond<-prop.table(datos,1)
> proba.cond
          [,1]      [,2]
[1,] 0.5644172 0.4355828
[2,] 0.4043716 0.5956284
> odds<-proba.cond[,1]/proba.cond[,2]
> odds
[1] 1.2957746 0.6788991
> odds[1]/odds[2]
[1] 1.908641

# Prob(Sí|Hombre)/Prob(Sí|Mujer)
> proba.cond[1,1]/proba.cond[2,1]
[1] 1.395788
```

Problema con celdas con ceros Un problema con este estimador r es la presencia de ceros en las celdas, ya que puede convertirse en una forma indeterminada.

Varios estimadores adicionales han sido propuestos para la razón odds y para el logarítmo de la razón de odds. Entre ellos tenemos:

- El de Haldane:
$$\widehat{\psi}_H = \frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(c+\frac{1}{2})(b+\frac{1}{2})}$$

- El de Jewell:
$$\widehat{\psi}_J = \frac{ad}{(b+1)(c+1)}$$

- Estimador de máxima verosimilitud condicional: Este estimador es la solución a un polinomio de alto grado de la forma:

$$\sum_{j=s}^{\delta} \left(\begin{array}{c} N_1 \\ j \end{array} \right) \left(\begin{array}{c} N_2 \\ k_1 - j \end{array} \right) (a-j)\rho^j$$

donde $s = \text{máx}(0,\, k_1 - N_2)$ y $\delta = \text{mín}(k_1, N_1)$

Propiedades de la razón de odds Algunas propiedades de la razón de odds son las siguientes:

- Es un número nonegativo.
- Cuando todas las celdas tienen probabilidades positivas, la independencia entre las dos variables es equivalente a $\psi = 0$.
- Es invariante bajo el intercambio de filas o columnas.
- Es invariante bajo multiplicaciones de filas y columnas.

- La interpretación es clara. Valores de ψ que se alejen de 1.0 en una dirección particular representa una asocición fuerte. Dos valores de ψ pueden representar un mismo nivel de asociación (un valor y su inverso) pero en direcciones opuestas. Para simetrizar esta medida se trabaja con el $log(\psi)$. Valores menores que uno indican una asociación negativa, mientras valores mayores que 1 indican una asociación positiva.

- Puede usarse en tablas $I \times J$ (y tablas multidimensionales) mirando series de particiones 2×2 o mirando subtablas 2×2.

Distribución asintótica de la Razón de Odds:
Esquema de muestreo *multinomial*

Sean

$$(n_1, ..., n_k) \sim Multinomial(\pi, n)$$

$$\pi = (\pi_1, \pi_2, ..., \pi_k)^T$$

$$n = n_1 + \cdots + n_k$$

Una estimación para el vector π es el vector

$$\widehat{\pi} = (\widehat{\pi}_1, \widehat{\pi}_2, ..., \widehat{\pi}_k)^T.$$

La i-ésima observación es

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{ik})'$$

donde

$$Y_{ij} = \begin{cases} 1 & \text{si cae en la celda } j \\ 0 & \text{en otro caso} \end{cases}$$

y además

$$\sum_j Y_{ij} = 1$$

Ahora

$$
\begin{aligned}
E[\mathbf{Y}_i] &= \pi \\
cov(\mathbf{Y}_i) &= \mathbf{\Sigma} \quad i = 1, ..., n \\
\sigma_{jj} &= var(Y_{ij}) = \pi_j(1 - \pi_j) \\
\sigma_{jk} &= cov(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) \\
&= -\pi_j\pi_k \qquad j \neq k \\
\mathbf{\Sigma} &= Diag(\pi) - \pi\pi^T \\
\widehat{\pi} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i \\
cov(\widehat{\pi}) &= \frac{(Diag(\pi) - \ \underline{\pi}\pi^T)}{n} \rightarrow \text{Matriz singular}
\end{aligned}
$$

Teorema central del límite multivariable Bajo el supuesto que $\mathbf{Y}_i, i = 1, \cdots, n$ sea una muestra aleatoria de una distribución $Multinomial(\pi, 1)$, entonces

$$\sqrt{n}(\widehat{\pi} - \pi) \xrightarrow{a} N(\mathbf{0}, Diag(\pi) - \pi\pi^T)$$

cuando $n \to \infty$.

Ahora

$$\begin{aligned} g(\pi) &= \mathsf{log}(\pi) \\ \frac{\partial g}{\partial \pi} &= Diag(\pi)^{-1} \end{aligned}$$

La covarianza de la matriz asintótica de

$$\sqrt{n}\left[\mathsf{log}(\widehat{\pi}) - \mathsf{log}(\pi)\right]$$

es

$$Diag(\pi)^{-1}\left[Diag(\pi) - \pi\pi^T\right]Diag(\pi)^{-1} = Diag(\pi)^{-1} - \mathbf{1}\mathbf{1}^T$$

Para una matriz C de constantes

$$\sqrt{n}C\left[\mathsf{log}(\widehat{\pi}) - \mathsf{log}(\pi)\right] \stackrel{a}{\rightarrow} N\left(\mathbf{0}, CDiag(\pi)^{-1}C^T - C\mathbf{1}\mathbf{1}^TC^T\right)$$

Con base en el anterior resultado, consideremos el siguiente vector

$$\begin{pmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{pmatrix}$$

El Odds ratio será

$$OR = \psi = \frac{\frac{\pi_{11}}{\pi_{21}}}{\frac{\pi_{12}}{\pi_{22}}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

Ahora

$$\log(\psi) = C(\log(\pi)) = [1\ \ -1\ \ -1\ \ 1]\begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix}$$

entonces

$$\begin{aligned}
nVar\left(\log\left(\widehat{\psi}\right)\right) &= CDiag(\pi)^{-1}C^T - C\mathbf{1}\mathbf{1}^T C^T \\
&= [1\ \ -1\ \ -1\ \ 1]\begin{bmatrix} \frac{1}{\pi_{11}} & 0 & 0 & 0 \\ 0 & \frac{1}{\pi_{12}} & 0 & 0 \\ 0 & 0 & \frac{1}{\pi_{21}} & 0 \\ 0 & 0 & 0 & \frac{1}{\pi_{22}} \end{bmatrix}\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \\
&\quad -[1\ \ -1\ \ -1\ \ 1]\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}[1\ 1\ 1\ 1]\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix} \\
&= \left[\frac{1}{\pi_{11}}\quad -\frac{1}{\pi_{12}}\quad -\frac{1}{\pi_{21}}\quad \frac{1}{\pi_{22}}\right]\begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}
\end{aligned}$$

$$= \quad \frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}$$

Distribución Asintótica de $\log\left(\widehat{\psi}\right)$

$$log\left(\widehat{\psi}\right) \sim AN\left(log\left(\psi\right), \frac{1}{n}\left(\frac{1}{\pi_{11}}+\frac{1}{\pi_{12}}+\frac{1}{\pi_{21}}+\frac{1}{\pi_{22}}\right)\right)$$

Un intervalo de confianza para $log\left(\psi\right)$ del 95 % es

$$\left(log\left(\widehat{\psi}\right) \mp 1{,}96\sqrt{\frac{1}{n}\left(\frac{1}{\widehat{\pi}_{11}}+\frac{1}{\widehat{\pi}_{12}}+\frac{1}{\widehat{\pi}_{21}}+\frac{1}{\widehat{\pi}_{22}}\right)}\right)$$

o

$$\left(\log\left(\widehat{\psi}\right) \mp 1{,}96\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}}\right)$$

Una forma muy común de hallar un intervalo de confianza para ψ se calcula invirtiendo el intervalo anterior

$$LI = \exp\left(\log\left(\widehat{\psi}\right) - 1{,}96\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\right)$$

y

$$LS = \exp\left(\log\left(\widehat{\psi}\right) + 1{,}96\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}\right)$$

```
> odds<-odds[1]/odds[2]
> LI.log<-log(odds)-1.96*sqrt(sum(1/datos))
> LI.log
[1] 0.3864979
> LS.log<-log(odds)+1.96*sqrt(sum(1/datos))
> LS.log
[1] 0.906285
> exp(LI.log)
[1] 1.471817
> exp(LS.log)
[1] 2.475111
```

Distribución Asintótica de $\widehat{\psi}$ Tenemos

$$Y = log\left(\widehat{\psi}\right) \sim AN\left(log\left(\psi\right),\, \frac{1}{n}\left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right)\right)$$

y que

$$\widehat{\psi} = \exp\left(\log\left(\widehat{\psi}\right)\right) = e^{Y}$$

Entonces

$$Y \sim AN\left(\exp\left(log\left(\psi\right)\right),\, \frac{1}{n}\left(\exp\left(log\left(\psi\right)\right)\right)^{2}\left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right)\right)$$

$$Y \sim AN\left(\psi,\, \psi^{2}\left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right)\right)$$

Un intervalo de confianza para ψ del 95 % basado en la distribución anterior es

$$LI = \widehat{\psi} - 1{,}96\widehat{\psi}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

y

$$LS = \widehat{\psi} + 1{,}96\widehat{\psi}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

```
> odds-1.96*odds*sqrt(sum(1/datos))
[1] 1.412598
> odds+1.96*odds*sqrt(sum(1/datos))
[1] 2.404685
```

```
> library(Epi)

> twoby2(datos)
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Col 1
Comparing : Row 1 vs. Row 2

      Col 1 Col 2    P(Col 1) 95% conf. interval
Row 1   368   284      0.5644    0.5261   0.6020
Row 2   148   218      0.4044    0.3553   0.4555

                                   95% conf. interval
             Relative Risk: 1.3958    1.2117   1.6079
         Sample Odds Ratio: 1.9086    1.4718   2.4751
Conditional MLE Odds Ratio: 1.9074    1.4597   2.4970
    Probability difference: 0.1600    0.0962   0.2218

             Exact P-value: 0
        Asymptotic P-value: 0
------------------------------------------------------
```