Página www

Página de Abertura

Contenido

Página 1 de 28

Regresar

Full Screen

Cerrar

Abandonar

Análisis de Datos Categóricos
Clase 2

Juan Carlos Correa
e-mail: jccorrea@unal.edu.co

1 de mayo de 2019

Prueba de la Razón de Verosimilitud (LRT)

Sea $X_1, \cdots, X_n$ una m.a. con verosimilitud $L(\theta)$, $\theta \in \Omega$. Considere la hipótesis

$$\begin{aligned} \mathrm{H}_0 &: \theta \in \Omega_0 \\ \mathrm{H}_1 &: \theta \in \Omega - \Omega_0 \end{aligned}$$

La LRT está definida por

$$\lambda = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})}$$

donde $\hat{\theta}$ es el e.m.v. sobre el espacio completo (estimador irrestricto) de θ y $\hat{\theta}_0$ es el e.m.v. bajo la restricción que H_0 sea cierta.

Resultado MUY Importante

Bajo condiciones de regularidad y cuando $n \longrightarrow \infty$ entonces

$$-2\log(\lambda) = -2\log\left(\frac{\text{máx}_{\theta\in\Omega_0} L(\theta)}{\text{máx}_{\theta\in\Omega} L(\theta)}\right) = -2\log\left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right) \sim \chi^2_\nu$$

donde

$$\nu = \dim(\Omega) - \dim(\Omega_0)$$

Pruebas de hipótesis con respecto a π

Asumamos que deseamos verificar $H_o : \pi = \pi_o$. La función de verosimilitud para π está dado por

$$L(\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

donde $x = \sum_{i=1}^n x_i$.
La razón de verosimilitudes está dada por

$$\lambda(\pi) = \frac{L(\omega)}{L(\Omega)} = \frac{\binom{n}{x} \pi_o^x (1-\pi_o)^{n-x}}{\binom{n}{x} \pi^x (1-\pi)^{n-x}} = \left(\frac{\pi_o}{\pi}\right)^x \left(\frac{1-\pi_o}{1-\pi}\right)^{n-x}$$

Tenemos que $-2\ln\left(\lambda(\hat{\pi})\right) \sim \chi^2_{(\nu)}$, con $\nu = \dim(\Omega) - \dim(\omega)$.
Por lo tanto

$$-2\ln\left(\lambda(\hat{\pi})\right) = -2\ln\left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right) = -2\ln\left(\left(\frac{\pi_o}{\hat{\pi}}\right)^x \left(\frac{1-\pi_o}{1-\hat{\pi}}\right)^{n-x}\right)$$

Ejemplo

En experimento en el cual a un sujeto se le pedía que usara toda su fuerza de voluntad para tratar de afectar el resultado de una moneda al ser lanzada al aire para que se lograra obtener cara, en 20 ensayos se obtuvieron 12 caras. Se podrá afirmar que el sujeto tuvo un efecto?

- Queremos verificar $H_0 : \pi = 0{,}5$ vs. $H_1 : \pi \neq 0{,}5$
- El estimador de máxima verosimilitud es $\hat{\pi} = 12/20 = 0{,}6$
- $-2\ln\left(\lambda(\hat{\pi})\right) = -2\ln\left(\left(\frac{0{,}5}{0{,}6}\right)^{12}\left(\frac{1-0{,}5}{1-0{,}6}\right)^{20-12}\right) = 0{,}6685058$
- $\chi^2_{0{,}95,1} = 3{,}841459.$
- Ya que el valor observado del estadístico, 0.6685058, es menor que el valor crítico no podemos rechazar H_0

Modelo Multinomial

El modelo multinomial es uno de los más comunes en el trabajo estadístico aplicado. Surge naturalmente cuando se reponden preguntas de selección múltiple, etc.

Estimación

Asumamos que $X_1, X_2, \cdots, X_n$ es una muestra aleatoria de una multinomial

$$M\left(1, (\pi_1, \pi_2, \cdots, \pi_k)'\right)$$

donde $\sum_{i=1}^{k} \pi_i = 1$. Cada X_i es un vector con ceros y con un único uno en la posición correspondiente a la categoría que pertenece la observación.

$$E(X_i) = \begin{bmatrix} \pi_1 \\ \pi_2 \\ . \\ . \\ . \\ \pi_k \end{bmatrix}$$

$$var(X_i) = \Sigma = \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_k \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \cdots & -\pi_2\widehat{\pi} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_k\pi_1 & -\pi_k\pi_2 & \cdots & \pi_k(1-\pi_k) \end{bmatrix}$$

La función de verosimilitud será:

$$L(\pi_1, \pi_2, \cdots, \pi_k) = \frac{n!}{n_1!n_2!\cdots n_k!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_k^{n_k}$$

donde n_i es el número de observaciones que pertenecen a la i-ésima categoría y $n = \sum_{i=1}^{n} n_i$.
El log de la verosimilitud será

$$l = \log(L(\pi_1, \pi_2, \cdots, \pi_k)) = \log\left(\frac{n!}{n_1!n_2!\cdots n_k!}\right) + \sum_{i=1}^{k} n_i \log(\pi_i)$$

Para hallar los estimadores de máxima verosimilitud derivamos la función anterior con respecto a cada uno de los parámetros (aquí abusamos tanto de notación como de lenguaje) teniendo en cuenta la restricción $\sum_{i=1}^{k} \pi_i = 1$, utilizando el multiplicador de Lagrange, $l^* = l(\pi_1, \pi_2, \cdots, \pi_k) - \lambda(\sum_{i=1}^{k} \pi_i - 1)$. Igualamos a cero y resolvemos el sistema de ecuaciones resultante.

$$\begin{aligned}
\frac{\partial l^*}{\partial \pi_1} &= \frac{n_1}{\pi_1} + \lambda \\
\frac{\partial l^*}{\partial \pi_2} &= \frac{n_2}{\pi_2} + \lambda \\
\vdots \; &\vdots \; \vdots \\
\frac{\partial l^*}{\partial \pi_k} &= \frac{n_1}{\pi_k} + \lambda \\
\frac{\partial l^*}{\partial \lambda} &= \sum_{i=1}^{k} \pi_i - 1
\end{aligned}$$

Igualando a cero y resolviendo, obtenemos

$$\hat{\pi}_i = \frac{n_i}{n} \text{ para todo } i = 1, \cdots, k.$$

Pruebas de hipótesis

Asumamos que deseamos verificar la hipótesis

$$H_0 : \pi_1 = \pi_1^*, \cdots, \pi_k = \pi_k^*$$

contra la alternativa $H_A : \pi_j \neq \pi_j^*$, para algún π_j. La razón de verosimilitud es

$$\lambda(\pi_1, \cdots, \pi_k) = \frac{L(\pi_1^*, \cdots, \pi_k^*)}{L(\pi_1, \cdots, \pi_k)} = \frac{\frac{n!}{n_1! n_2! \cdots n_k!} \pi_1^{*n_1} \pi_2^{*n_2} \cdots \pi_k^{*n_k}}{\frac{n!}{n_1! n_2! \cdots n_k!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_k^{n_k}}$$

Lo cual se reduce a

$$\lambda(\pi_1, \cdots, \pi_k) = \left(\frac{\pi_1^*}{\pi_1}\right)^{n_1} \left(\frac{\pi_2^*}{\pi_2}\right)^{n_2} \cdots \left(\frac{\pi_k^*}{\pi_k}\right)^{n_k}$$

Sabemos que $-2\log\left(\lambda(\hat{\pi}_1, \cdots, \hat{\pi}_k)\right) \sim \chi^2_\nu$, donde $\nu = \dim(\Omega) - \dim(\omega) = (k-1) - 0 = k-1$, tenemos entonces que

$$-2\log\left(R(\hat{\pi}_1, \cdots, \hat{\pi}_k)\right) = -2\sum_{i=1}^{k} n_i \log\left(\frac{\pi_i^*}{\hat{\pi}_i}\right) \sim \chi^2_{(k-1)}$$

El Estadístico G^2

El estadístico G^2 está basado en la razón de verosimilitud, y es tal vez la medida de ajuste que más sirve en el análisis de datos categóricos, dadas sus propiedades.

$$G^2 = 2\sum_i n_i\left[\log(n_i) - \log(e_i)\right]$$

El periódico El Tiempo (Abril 2 del 2000) presentó una tabla con los porcentajes de los diferentes tipos de sangre en la población.

$$H_o : \pi_O = 0{,}577, \pi_A = 0{,}292, \pi_{AB} = 0{,}091, \pi_B = 0{,}021$$

La siguiente tabla presenta los datos sobre el tipo de sangre en una muestra de personas de la región central y oriental de Antioquia

	Tipo de Sangre			
	O	A	AB	B
Frecuencia	474	246	11	59
$\hat{\pi}_i$	0.60000000	0.31139241	0.01392405	0.07468354

La siguiente función en R nos permite realizar los cálculos:

```
prueba.multinomial<-function(observado,prob.teoricas,nivel=0.05){

if(length(observado)!=length(prob.teoricas))stop('Longitudes
diferentes!')
observado<-ifelse(observado==0,0.5,observado)
G2<--2*sum(observado*log(prob.teoricas/(observado/sum(observado))))
gl<-length(observado)-1
valor.critico<-qchisq(1-nivel,gl)
list(G2=G2,valor.critico=valor.critico)
}
```

```
prueba.multinomial(c(474,246,11,59),c(0.577,0.292,0.091,0.021))
$G2
[1] 177.1022

$valor.critico
[1] 7.814728
```

Como un ejemplo consideremos la siguiente tabla que presenta el mes de nacimiento de los estudiantes de pregrado de la Universidad Nacional-Sede Medellín, también aparece la probabilidad estimada de nacer en cada mes y la probabilidad teórica asumiendo uniformidad, o sea asumiendo que una persona extraída al azar tiene igual probabilidad de haber nacido en cualquier día del año. Observe que los meses tienen diferentes probabilidades teóricas debido a que los números de días por mes no es constante.

Mes	Nacimientos	Frecuencia Relativa	Probabilidad Teórica
Enero	856	0.09069	0.08493
Febrero	716	0.07586	0.07671
Marzo	740	0.07840	0.08493
Abril	721	0.07639	0.08219
Mayo	803	0.08507	0.08493
Junio	751	0.07956	0.08219
Julio	790	0.08370	0.08493
Agosto	830	0.08793	0.08493
Septiembre	830	0.08793	0.08219
Octubre	801	0.08486	0.08493
Noviembre	789	0.08359	0.08219
Diciembre	812	0.08603	0.08493

```
$G2
[1] 18.55022

$valor.p
[1] 0.06966068
```

No parece existir evidencia que nos indique que los niños prefieran unos meses u otros para nacer.

Sobre los resultados del juego de dados

En un juego de parqués se registraron los resultados del lanzamiento de un par de dados 130 veces. A partir de estos resultados quiere uno ver si los dados son conjuntamente buenos.

Resultado	2	3	4	5	6	7	8	9	10	11	12
Frecuencia	4	8	10	11	22	14	22	18	10	5	6

La hipótesis a verificar es la que la suma de los dos dados tiene una distribución producida por un par de dados justos:

Resultado	2	3	4	5	6	7	8	9	10	11	12
Probabilidad esperada	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

```
> prueba.multinomial(c(4,8,10,11,22,14,22,18,10,5,6),
+ c(1,2,3,4,5,6,5,4,3,2,1)/36)
$G2
[1] 8.75751

$valor.p
[1] 0.555261
```

Los resultados no nos permiten rechazar la hipótesis nula sobre la distribución de la suma de los dos dados.

Igualdad de Dos Proporciones: $H_0 : \pi_1 = \pi_2 (= \pi)$

- Dos muestras independientes de poblaciones Bernoulli

	Resultado	
	Éxito	Fracaso
Muestra I	x_1	$n_1 - x_1$
Muestra II	x_2	$n_2 - x_2$

Prueba LRT

$$\lambda = \frac{\hat{\pi}^{x_1+x_2}\left(1-\hat{\pi}\right)^{n_1+n_2-(x_1+x_2)}}{\hat{\pi}_1^{x_1}\left(1-\hat{\pi}_1\right)^{n_1-x_1}\hat{\pi}_2^{x_2}\left(1-\hat{\pi}_2\right)^{n_2-x_2}}$$

donde

$$\begin{aligned}\hat{\pi} &= \frac{x_1+x_2}{n_1+n_2}\\ \hat{\pi}_1 &= \frac{x_1}{n_1}\\ \hat{\pi}_2 &= \frac{x_2}{n_2}\end{aligned}$$

Re-escribiendo λ

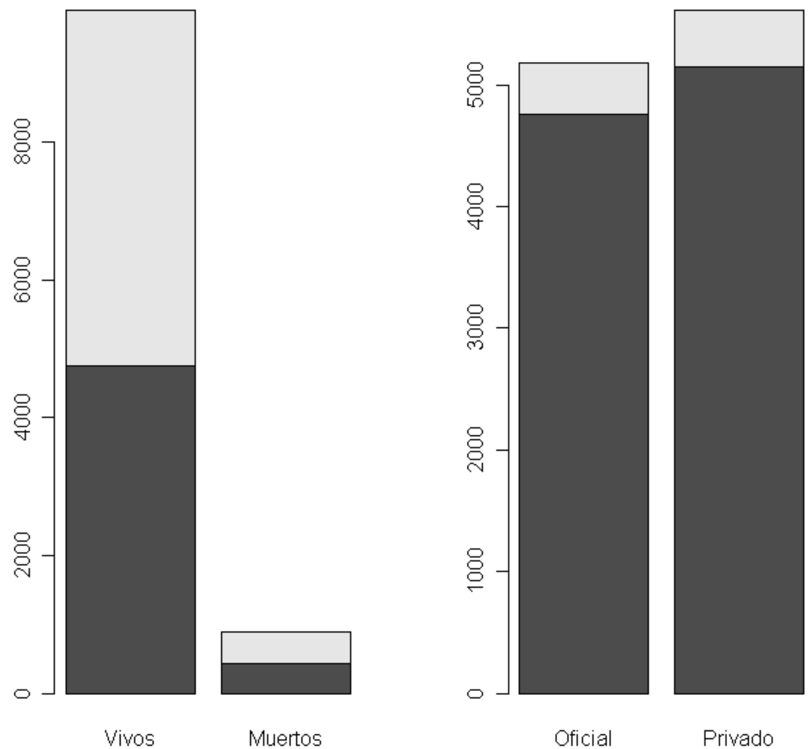
$$\lambda = \left(\frac{\hat{\pi}}{\hat{\pi}_1}\right)^{x_1} \left(\frac{1-\hat{\pi}}{1-\hat{\pi}_1}\right)^{n_1-x_1} \left(\frac{\hat{\pi}}{\hat{\pi}_2}\right)^{x_2} \left(\frac{1-\hat{\pi}}{1-\hat{\pi}_2}\right)^{n_2-x_2}$$

$$-2\log(\lambda) \sim \chi^2_1$$

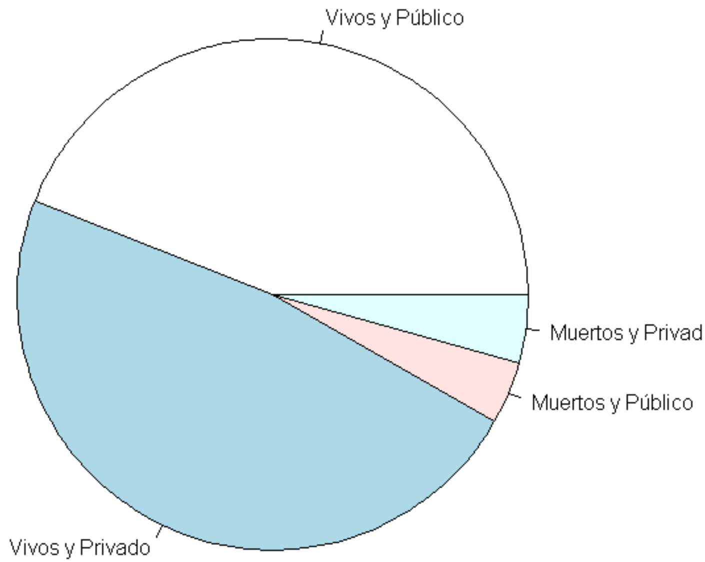
Mortalidad en Instituciones Públicas o Privadas

	Resultado	
	Vivo	Muerto
Oficial	4757	430
Privado	5148	464

```
> partos.dat<-array(c(4757,5148,430,464),c(2,2))
> partos.dat
     [,1] [,2]
[1,] 4757  430
[2,] 5148  464

> rownames(partos.dat)<-c('Oficial','Privado')
> colnames(partos.dat)<-c('Vivos','Muertos')
> partos.dat
        Vivos Muertos
Oficial  4757     430
Privado  5148     464

> par(mfrow=c(1,2))
> barplot(partos.dat)
> barplot(t(partos.dat))

> pie(partos.dat,labels=c('Vivos y Público','Vivos y Privado',
   'Muertos y Público','Muertos y Privado'))
```



Página www
Página de Abertura
Contenido
Página 25 de 28
Regresar
Full Screen
Cerrar
Abandonar


Vivos y Público
Muertos y Privad
Muertos y Público
Vivos y Privado

```
> prop.table(partos.dat)
            Vivos    Muertos
Oficial 0.4405038 0.03981850
Privado 0.4767108 0.04296694
> sum(prop.table(partos.dat))
[1] 1
>

> prop.table(partos.dat,1)
            Vivos    Muertos
Oficial 0.9171004 0.08289956
Privado 0.9173200 0.08267997


> prop.table(partos.dat,2)
            Vivos   Muertos
Oficial 0.4802625 0.4809843
Privado 0.5197375 0.5190157
>
```

```
Igualdad.2.prop<-function(N){
# N: tabla 2x2
n1<-N[1,1]+N[1,2]
pi1<-N[1,1]/n1
n2<-N[2,1]+N[2,2]
pi2<-N[2,1]/n2
pi0<-(N[1,1]+N[2,1])/(n1+n2)
lambda<-(pi0/pi1)^N[1,1]*((1-pi0)/(1-pi1))^(n1-
N[1,1])*(pi0/pi2)^N[2,1]*((1-pi0)/(1-pi2))^(n2-N[2,1])
G2<--2*log(lambda)
valor.p<-pchisq(G2,1,lower.tail=F)
list(G2=G2,valor.p=valor.p)
} # Fin Igualdad.2.prop

partos.dat<-array(c(4757,5148,430,464),c(2,2))

Igualdad.2.prop(partos.dat)

$G2
[1] 0.00171166

$valor.p
[1] 0.9669992
```