Datos Categóricos: Clase 3

Juan Carlos Correa

16 de marzo de 2022

Teorema central del límite multivariable para la Multinomial

Bajo el supuesto que $\mathbf{Y}_i, i = 1, \cdots, n$ sea una muestra aleatoria de una distribución $Multinomial(1, \pi_{k\times 1})$, entonces

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

cuando $n \to \infty$.

Esquemas de Muestreo para tablas $I \times J$

Las tablas bidimensionales son importantes por varios aspectos:

- Permiten el cruce de dos variables, lo que es manejable de una forma sencilla por parte del investigador y usualmente los tamaños muestrales no permiten elaborar tablas más complejas, ya que las tablas comienzan a presentar muchas celdas con muy pocas observaciones, esto se puede ver si tenemos 5 variables hay 32 celdas posibles.
- El usuario las entiende y las visualiza sin mayores dificultades.

Un caso de importancia es el de las Tablas 2×2, el cual desarrollaremos en detalle, ya que permite la introducción de conceptos importantes de una forma simple.
Hay un ejemplo clásico narrado por Fisher (1951) en su libro "El Planeo de Experimentos" (la traducción es argentina) y que presentamos a continuación:

> "Una dama declara que catando una taza de té con leche, puede distinguir si la leche o la infusión de té fué vertida primero en la taza. Consideremos el problema de diseñar un experimento por medio del cual este aserto puede ser testado. Con este propósito permítasenos primero formular un experimento de forma simple con miras a estudiar sus limitaciones y sus características: aquellas que aparecen como fundamentales para el método experimental, cuando está bien desarrollado y las que no son esenciales sino auxiliares.
>
> Nuestro experimento consiste en mezclar ocho tazas de té cuatro en una forma y cuatro en la otra, y presentarlas ordenadas al azar al sujeto que debe juzgarlas. El sujeto ha sido informado de antemano en qué consistirá el test, a saber: que se le pedirá que cate ocho tazas, que éstas serán cuatro de cada clase, y que le serán presentadas ordenadas al azar, que es un orden no determinado arbitrariamente por elección humana, sino por la manipulación actual de los aparatos físicos usados en juegos de azar, cartas, dados, ruletas, etc., o, más expeditivamente a partir de una tabla de números para muestras al azar, destinada a dar el resultado actual de tal manipulación. Su tarea es separar las ocho tazas en dos grupos de 4, estipulando, si es posible, los tratamientos recibidos."

Esquemas de Muestreo para Tablas 2×2

Las tablas 2×2 son del estilo de la que aparece a continuación:

		Clasificación II		
		1	2	Total
Clasificación I	1	a	b	$k_1 = a + b$
	2	c	d	$k_2 = c + d$
Total		$n_1 = a + c$	$n_2 = b + d$	N

Muestreando con ambos conjuntos de marginales fijos

Sean A, B, C y D variables aleatorias con valores observados a, b, c y d. Bajo este esquema de muestreo sólo una es independiente, digamos A.

Prueba Exacta de Fisher

Suponga que en un proceso de selección de personal para cierta labor de promoción se decide entrevistar k_1 hombres y k_2 mujeres. De antemano se sabe que n_1 personas serán seleccionadas.

		Seleccionado?		
		Sí	No	Total
Sexo	Hombre	a	b	$k_1 = a + b$
	Mujer	c	d	$k_2 = c + d$
Total		$n_1 = a + c$	$n_2 = b + d$	N

- Una pregunta que podría ser de interés es la siguiente:

  Exite sesgo a favor de la selección de mujeres (u hombres)?

- Si $\frac{a}{k_1}$ y $\frac{c}{k_2}$ son muy diferentes, uno puede sospechar un sesgo. La hipótesis nula será en este caso:

  H_0: La selección es estrictamente aleatoria.

Bajo la hipótesis nula la distribución de A será :

$$P_{H_0}(A = a) = \frac{\begin{pmatrix} n_1 \\ a \end{pmatrix}\begin{pmatrix} n_2 \\ b \end{pmatrix}}{\begin{pmatrix} N \\ k_1 \end{pmatrix}} = \frac{\begin{pmatrix} k_1 \\ a \end{pmatrix}\begin{pmatrix} k_2 \\ c \end{pmatrix}}{\begin{pmatrix} N \\ n_1 \end{pmatrix}}$$

para $a = 0, 1, ..., \text{mín}(k_1, n_1)$ y $\text{máx}(0, k_1 + n_1 - N) \leq a \leq \text{mín}(k_1, n_1)$.
Es fácil ver que $a \geq 0$. Se deja como ejercicio verificar que $a \geq k_1 + n_1 - N$.

$$E[A] = k_1 \frac{n_1}{N} = k_1 p \text{ , donde } p = \frac{n_1}{N}$$

Si el número observado es mucho mayor que el valor esperado, digamos $\geq a_\alpha$, esto indicará un sesgo a favor de los hombres, aquí a_α es el entero más pequeño tal que

$$P(A \geq a_\alpha) \leq \alpha$$

α es el nivel de significancia deseado para probar H_0, donde la alternativa sería

H_a: Sesgo a favor de los hombres

El *valor-p* se calcula como

$$valor - p = \sum_{j=a}^{min(n_1,k_1)} P[A = j] = \sum_{j=a}^{n_1} \frac{\binom{n_1}{j}\binom{n_2}{k_1 - j}}{\binom{N}{k_1}}$$

Se rechaza H_0 a un nivel α si $p \leq \alpha$. Esta prueba es conocida como la *prueba exacta de Fisher-Irwin*[a].

La prueba anterior es de una cola. La prueba de dos colas, esto es: Sesgo hacia alguno de los sexos, puede construirse de muchas formas. Una es: escoja α_1 y α_2 tal que $\alpha_1 + \alpha_2 = \alpha$ con a_{α_1} tal que

$$P(A \leq a_{\alpha_1}) \leq \alpha_1$$

y $a^*_{\alpha_2}$tal que.

$$P(A \leq a^*_{\alpha_2}) \leq \alpha_2$$

Rechace H_0 si $a \leq a_{\alpha_1}$ ó $a \leq a^*_{\alpha_2}$.

[a]Debemos anotar que varios autores hacen comentarios sobre la falsa idea que produce la palabra *exacta* cuando se habla de la Prueba Exacta de Fisher. Como D'Agostino et al. (1988) notan, esta prueba es muy conservadora y tiene una potencia muy pobre comparada con la chi-cuadrada.

Prueba exacta de Fisher en R La función `fisher.test()` permite realizar la prueba exacta de Irwin-Fisher.

```
fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE,
             or = 1, alternative = "two.sided", conf.level = 0.95)
```

x Es una matriz $I \times J$ de enteros no negativos o un objeto tipo factor.

y Es un objeto tipo factor y solo es considerardo si el argumento anterior no es una matriz.

workspace Un entero que especifica el espacio de trabajo en R.

hybrid Un valor lógico que indica si se calculan las probabilidades exactas o un híbrido basado en una aproximación chi-cuadrada.

or Valor hipotético de la razón de odds.

alternative Solo se utiliza en matrices 2×2 y debe especificar el tipo de hipótesis a ser verificada: "two-sided", "greater" o "less"

conf.level Solo se utiliza en matrices 2×2 y especifica el nivel de confianza.

Consideremos el famoso ejemplo de la dama que declara que conoce si en una taza de té con leche fue colocado primero el té o la leche descrito por Fisher.

		Decisión	
		Té	Leche
Lo que primero se colocó	Té	3	1
	Leche	1	3

```
> data.te <- matrix(c(3,1,1,3),ncol=2,byrow=T)
> fisher.test(data.te)

        Fisher's exact test

data: data.te
p-value = 0.4857
alternative hypothesis: two.sided
```

Muestrando con un conjunto de marginales fijo

Suponga que una muestra aleatoria de tamaño n_1 es sacada de la población I con probabilidad de éxito π_1 y a es el número de éxitos observados. Suponga también que otra muestra aleatoria de tamaño n_2 es sacada de la población II con probabilidad de éxito π_2 y b es el número de éxitos observados. El modelo de probabilidad postulado se conoce como *Producto-Binomial*

$$P(A = a, B = b) = \binom{n_1}{a} \pi_1^a (1-\pi_1)^{n_1-a} \binom{n_2}{b} \pi_2^b (1-\pi_2)^{n_2-b}$$

Aquí el problema de interés será verificar la siguiente hipótesis:

$$H_0 : \pi_1 = \pi_2 = (\pi)$$

Bajo H_0 tenemos,

$$P(A = a, B = b) = \binom{n_1}{a}\binom{n_2}{b}\pi^{a+b}(1-\pi)^{N-(a+b)}$$

Nota: $a + b$ es un estadístico suficiente para el parámetro π (de perturbación o molestia), bajo H_0, pero por sí mismo no proporciona información alguna acerca de H_0. Bajo H_0 tenemos que

$$P(A + B = a + b) = \binom{N}{a+b}\pi^{a+b}(1-\pi)^{N-a-b}$$

Para probar $H_0 : \pi_1 = \pi_2$ vs. $H_1 : \pi_1 > \pi_2$ rechazamos H_0 si a es suficientemente grande con respecto a b, dado $a + b$. Esto es, rechazamos H_0 si $a \geq a_\alpha$, donde a_α es el entero más pequeño tal que

$$P(A \geq a_\alpha \mid A + B = a + b) \leq \alpha$$

Esta es una prueba condicional con nivel α.
Coincide con la prueba de una cola de *Fisher-Irwin*, tomando $a+b = k_1$. Tal prueba condicional con nivel α para todo posible valor de $a + b$ se puede aceptar como una prueba incondicional de nivel α.

Prueba LRT para $H_0 : \pi_1 = \pi_2$ vs. $H_1 : \pi_1 \neq \pi_2$ La prueba de la razón de verosimilitud para este problema la construimos así:

$$L(\pi_1, \pi_2) = \pi_1^a (1-\pi_1)^b \pi_2^c (1-\pi_2)^d$$

Bajo H_0 tenemos

$$L(\pi_1 = \pi_2 = \pi) = \pi^{a+c}(1-\pi)^{b+d}$$

El e.m.v. bajo el modelo restricto es $\hat{\pi} = (a+c)/N$. los e.m.v. bajo el modelo irrestricto son $\hat{\pi}_1 = a/(a+b)$ y $\hat{\pi}_2 = c/(c+d)$. Por lo tanto

$$G^2 = -2\log(\lambda) = -2\log\left(\frac{\hat{\pi}_1^a(1-\hat{\pi}_1)^b\hat{\pi}_2^c(1-\hat{\pi}_2)^d}{\hat{\pi}^{a+c}(1-\hat{\pi})^{b+d}}\right) \sim \chi^2_{dim(\Omega)-di}$$

donde $dim(\Omega) - dim(\omega) = 2 - 1 = 1$.

Muestreando sólo con N fijo

El modelo de probabilidad observado es

$$P(A=a, B=b, C=c) = \frac{N!}{a!b!c!d!}\pi_{11}^{a}\pi_{12}^{b}\pi_{21}^{c}\pi_{22}^{d}$$

donde $d = N - a - b - c$ y π_{ij} es la probabilidad de la (i, j)-ésima celda, $i, j = 1, 2$.
La hipótesis de interés corriente es

$$\begin{aligned} H_0 &: \text{ Independencia de las dos respuestas o} \\ H_0 &: \quad \pi_{ij} = \pi_{i+}\pi_{+j}, \qquad i, j = 1, 2. \end{aligned}$$

Considere la prueba de una cola

$$H_1 : \pi_{11} > \pi_{1+}\pi_{+1},$$

esto es, asociación positiva.
Bajo H_0 el modelo se convierte en

$$P(A = a, B = b, C = c) = \frac{N!}{a!b!c!d!}\pi_{1+}^{a+b}\left(1-\pi_{1+}\right)^{c+d}\pi_{+1}^{a+c}\left(1-\pi_{+1}\right)^{b+d}$$

$(a+b, a+c)$ son estadísticos suficientes para los parámetros de molestia (π_{1+}, π_{+1}). La distribución condicional de A dado $a+b=k_1$ (y $a+c=n_1$) es multinomial bajo H_0. La prueba condicional de *Fisher-Irwin* rechaza H_0 si $a \geq a_\alpha$, esta es una prueba unilateral.

Siguiendo los argumentos previos se puede generar una prueba para
$H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$

$$P\left(A = a, B = b \Big/ \begin{array}{c} A + C = n_1 \\ B + D = n_2 \end{array}\right) = \binom{n_1}{a} \pi_1^{*a}(1-\pi_1^*)^{n_1-a} \binom{n_2}{b} \pi_2^{*b}(1-\pi_2^*)^{n_2-b}$$

donde $\pi_1^* = \frac{\pi_{11}}{\pi_{+1}}$ y $\pi_2^* = \frac{\pi_{12}}{\pi_{+2}}$.
Así en el marco condicional $a + c = n_1$ y $b + d = n_2$ reduce el modelo de probabilidad a la hipótesis nula y a la alternativa del caso anterior.

Prueba LRT para $H_0 : \pi_1 = \pi_2$ vs. $H_1 : \pi_1 \neq \pi_2$ La prueba de la razón de verosimilitud para este problema la construimos así:

$$L(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) = \frac{N!}{a!b!c!d!}\pi_{11}^a \pi_{12}^b \pi_{21}^c \pi_{22}^d$$

Bajo H_0 tenemos

$$L(\pi_{11} = \pi_{21} = \pi, \pi_{12}, \pi_{22}) = \frac{N!}{a!b!c!d!}\pi^{a+c} \pi_{12}^b \pi_{22}^d$$

Los e.m.v. bajo el modelo restricto son $\hat{\pi} = (a+c)/(2N)$, $\hat{\pi}_{12} = b/N$, $\hat{\pi}_{22} = d/N$. Los e.m.v. bajo el modelo irrestricto son $\hat{\pi}_{11} = a/N$, $\hat{\pi}_{12} = b/N$,$\hat{\pi}_{21} = c/N$, $\hat{\pi}_{22} = d/N$. Por lo tanto

$$G^2 = -2\log(\lambda) = -2\log\left(\frac{\frac{N!}{a!b!c!d!}\hat{\pi}^{a+c}\hat{\pi}_{12}^b\hat{\pi}_{22}^d}{\frac{N!}{a!b!c!d!}\hat{\pi}_{11}^a\hat{\pi}_{12}^b\hat{\pi}_{21}^c\hat{\pi}_{22}^d}\right) \sim \chi^2_{dim(\Omega)-dim(\omega)}$$

donde $dim(\Omega) - dim(\omega) = 3 - 2 = 1.$

Muestreo bajo el esquema Poisson

Suponga que A, B, C y D son Poisson independientes con medias λ_{ij}, $i, j = 1, 2$.
El modelo de probabilidad es

$$P(A = a, B = b, C = c, D = d) = \exp(-\lambda_{++})\frac{\lambda_{11}^{a}\lambda_{12}^{b}\lambda_{21}^{c}\lambda_{22}^{d}}{a!b!c!d!}$$

donde $\lambda_{++} = \lambda_{11} + \lambda_{12} + \lambda_{21} + \lambda_{22}$. Como $N = A + B + C + D$, entonces $N \sim poisson(\lambda_{++})$. Si condicionamos en N reducimos este modelo al caso anterior con
$\pi_{ij} = \frac{\lambda_{ij}}{\lambda_{++}}$, $i, j = 1, 2$.