Datos Categóricos Apareados

Lautaro Ochotorena

3 de octubre de 2023

Índice

1. Preliminares

Teorema 1.0.1 (Teorema de De Moivre-Laplace)**.** Si X es una variable aleatoria con distribución binomial, con $E(X) = np$ y $Var(X) = np(1-p)$, entonces

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

conforme $n \to \infty$, esta aproximación es buena si $np \geq 5$ y $n(1-p) \geq 5$.

Teorema 1.0.2. La propiedad de la invarianza del estimador de máxima verosimilitud se mantiene para el caso multivariado.
Si el estimador de máxima verosimilitud de $(\theta_1, ..., \theta_k)$ es $(\hat{\theta}_1, ..., \hat{\theta}_k)$ y $\tau(\theta_1, ..., \theta_k)$ es cualquier función dependiente de esos parámetros, entonces el estimador de máxima verosimilitud de $\tau(\theta_1, ..., \theta_k)$ es $\tau(\hat{\theta}_1, ..., \hat{\theta}_k)$

Definición 1.1. Definiremos dos tipos de situaciones en las que se muestran los datos para tablas de contingencias:

1) Hay I poblaciones de interés independientes, cada una correspondiente a una fila diferente de la tabla, y cada población está dividida en las mismas J categorías. Se toma una muestra de la i-ésima población ($i = 1, \ldots, I$) y las cantidades se introducen en las celdas de la i-ésima fila de la tabla.
   En este caso hay I distribuciones multinomiales (binomiales en caso de que J=2). Se puede aplicar un test de homogeneidad

2) Hay una sola población de interés, con cada individuo de la misma clasificado con respecto a dos factores diferentes. Hay I categorías asociadas con el primer factor, y J categorías asociadas con el segundo factor. Se toma una sola muestra, y el número de individuos pertenecientes tanto a la categoría i del factor 1 como a la categoría j del factor 2 se introduce en la celda de la fila i, columna j ($i = 1, \ldots, I$; $j = 1, \ldots, J$).
   En este caso se lo considera como una única distribución multinomial. Se puede aplicar test de independencia.

El método de propagación de erorres consiste en aproximar una variable aleatoria $Y = g(X)$ con g una función cualquiera y X una variable aleatoria con $E(X) = \mu_X$ y $Var(X) = \sigma_X^2$.
Mediante un desarrollo de Taylor se tiene que

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X)$$

y mediante las propiedades de la esperanza y la varianza se tiene que

$$\mu_Y \approx g(\mu_X)$$
$$\sigma_Y^2 \approx \sigma_X^2[g'(\mu_X)]^2$$

2. Datos Apareados

El diseño de datos apareados puede ser efectivo en experimentos que involucran a datos categóricos. Apropiadas técnicas deben ser utilizadas para estos datos. Veamos un par de ejemplos para saber de qué trata.

Ejemplo 2.0.1. Vianna et al. [1] recolectaron datos comparando los porcentajes de amigdalectomías (extirpación de las amígdalas) de un grupo de pacientes que sufren de la enfermedad de Hodgkin (un linfoma maligno) y lo compararon con un grupo de control. Los resultados fueron los siguientes

	Con amigdalectomía	Sin amigdalectomía	Total
Enfermedad de Hodgkin	67	34	101
Control	43	64	107
Total	110	98	208

La tabla muestra que 66 % de los que tienen la enfermedad de Hodgkin han tenido una amigdalectomía, comparado con el 40 % del grupo de control.
Esta tabla se obtuvo fijando el número total de personas con la enfermedad y fijando el número total de personas en el grupo de control, es una muestra de tipo 1.
Por lo cual tenemos dos distribuciones binomiales que podemos representar con la siguiente tabla

N_{11}	N_{12}	$n_1 = 101$
N_{22}	N_{22}	$n_2 = 107$
$N_{\cdot 1}$	$N_{\cdot 2}$	$n = 208$

donde N_{i1} y N_{i2} representan cantidad de éxitos y fracasos de una binomial $B(n_i, \pi_i)$, para $i = 1, 2$.
Se trató de ver si las personas sin amígdalas son más propensas a tener la enfermedad, para ello se somete a un test de homogeneidad que plantea

$$H_0 : \ \pi_1 = \pi_2$$

El test de homogeneidad utiliza como estadístico

$$T = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(N_{ij} - \frac{n_i}{n} N_{\cdot j})^2}{\frac{n_i}{n} N_{\cdot j}}$$

que tiene una distribución $\chi^2_{(2-1)(2-1)}$ bajo H_0 siempre que $\frac{n_i}{n} N_{\cdot j} \geq 5$ para $i, j = 1, 2$.
Evidentemente,

$$\frac{n_i}{n} N_{\cdot j} \geq \frac{101}{208} 98 \geq 5$$

por lo que cumple con lo pedido. Se obtiene finalmente que

$$t = 14{,}259$$

y un $p - valor$ de 0,0001, lo cual indica que para cualquier nivel de significancia aceptable se rechazaría la hipótesis nula.
Los investigadores conjeturaron que las amigdalas actúan como una "capa protectora" contra la enfermedad de Hodgkin.
Más tarde, se realizó otro experimento, esta vez por parte de Johnson and Johnson [2], en donde seleccionaron a 85 pacientes con la enfermedad de Hodgkin, los cuales tenían un hermano del mismo sexo pero que no poseía la enfermedad y cuya diferencia de edad no fuese mayor a 5 años. Los resultados obtenidos fueron los siguientes

	Con amigdalectomía	Sin amigdalectomía	Total
Enfermedad de Hodgkin	41	44	85
Hermano	33	52	85
Total	74	96	170

calcularon el estadístico del test y obtuvieron

$$t = 1{,}53$$

y un $p-valor$ de 0,2, lo cual indicaría que no se rechazaría la hipótesis nula para cualquier nivel de significancia aceptable. Esto iba en contraposición con la conclusión de los anteriores investigadores.

Varias cartas dedicadas al editor de la revista en donde había sido publicado el trabajo de Johnson and Johson hacían referencia a que se había cometido un error en el análisis por ignorar las paridades. Esto es porque el test de homogeneidad depende de una hipótesis de independencia multinomial de los datos, y los datos extraidos por Johnson and Johnson no eran independientes porque eran datos entre hermanos (datos apareados).
Un apropiado análisis de los datos de Johnson and Johnson surgen al análizar la siguiente tabla

		Hermano	
		Sin amigdalectomía	Con amigdalectomía
Paciente	Sin amigdalectomía	37	7
	Con amigdalectomía	15	26

Los datos vienen de una muestra de tamaño 85 de una distribución multinomial con cuatro celdas (tabla tipo 2). Se puede representar las probabilidades de la tabla como

π_{11}	π_{12}	$\pi_{1.}$
π_{21}	π_{22}	$\pi_{2.}$
$\pi_{.1}$	$\pi_{.2}$	1

La hipótesis nula apropiada establece que las probabilidades de personas con amigdalectomía y sin amigdalectomía son las mismas para pacientes y para hermanos, esto es que $\pi_{1.} = \pi_{.1}$ y $\pi_{2.} = \pi_{.2}$, o escrito de otra forma

$$\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21}$$
$$\pi_{12} + \pi_{22} = \pi_{21} + \pi_{22}$$

A lo que se puede simplificar con $\pi_{12} = \pi_{21}$.
Por lo que

$$H_0 : \pi_{12} = \pi_{21}$$

Cuando H_0 es verdadero, se esperan valores similares para n_{12} y n_{21}. Sea $n^* = n_{12} + n_{21}$ denotando la suma de estas dos celdas.
Bajo H_0, cada una de estas n^* observaciones tiene $1/2$ de chances de contribuir a n_{12} y $1/2$ de chances de contribuir a n_{21}. Por lo cual, N_{12} y N_{21} son número de éxito y fracaso de una distribución binomial con n^* sucesos y probabilidad de éxito de $\frac{1}{2}$.
Cuando $n^* > 10$, por el teorema 1,0,1 tenemos el estadístico

$$Z = \frac{N_{12} - (\frac{1}{2})n^*}{\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}} \sim N(0,1)$$

y el cuadrado de esto sigue una distribucion χ^2_1. El estadístico involucra dos parámetros que por la hipótesis nula deberían ser iguales y por lo cual el grado de libertad de la chi-cuadrado es 1.

Este es el llamado test de McNemar.
El valor del estadístico es

$$\frac{(n_{12} - (\frac{1}{2})n^*)^2}{(\sqrt{n^*(\frac{1}{2})(\frac{1}{2})})^2} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Usando este test en la tabla 2,0,1 se obtiene un $\chi_1^2 = \frac{(7-15)^2}{7+15} = 2{,}909$ con un $p - valor$ de 0.09, lo cual pone en duda a la hipótesis nula, contrario al análisis que obtuvo Johnson and Johnson.

Ejemplo 2.0.2. Celulares y conducir
¿El uso de celulares mientras se maneja causa accidentes? Esto es una pregunta difícil de estudiar. Un estudio de observación comparando porcentajes de accidentes de personas que usan el celular manejando y personas que no lo usan puede involucran variables como la edad, el género, el tiempo, lugar en donde se conduce, etc. Un experimento aleatorizado y controlado en donde de forma aleatoria se asigna a personas a usar o no usar el celular sería poco ético al exponer a gente a condiciones potencialmente peligrosas.
En 1997, Redelmeier y Tibshirani [3] hicieron un estudio en el que identificaron a 699 conductores que poseían celulares y habían estado involucrados en accidentes automovolísticos.
Luego, se usaron archivos para determinar el uso del celular durante los 10 minutos antes del accidente y durante el mismo tiempo en la semana anterior. Por lo cual cada persona actúo como su propio grupo de control, eliminando las variables dichas anteriormente. Estos resultados fueron los siguientes

Colisión	Antes de la colisión		
	En el celular	No en el celular	Total
En el celular	13	157	170
No en el celular	24	505	529
Total	37	662	699

El día de la colisión 24 % de los conductores habían estado utilizando el celular comparado con el 5 % de la semana previa a colisionar.
El test de McNemar se puede aplicar para ver la hipótesis nula de no asosiación, obteniendo así

$$\chi_1^2 = \frac{(157 - 24)^2}{157 + 24} = 97{,}7$$

y con un $p - valor$ muy chico (casi 0). Es por ello que se rechazaría la hipótesis nula, sin embargo, el autor apunta a que el resultado puede no indicar que el uso de celulares mientras se maneja causa más accidentes, esto es debido a que pueden intervenir otras variables en el momento del accidente como es el estrés emocional que puede generar que los conductores miren más el celular.

3. Razón de probabilidades (Odds Ratio)

Para un evento A con probabilidad $P(A)$, las odds de A se definen como

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

Y esto implica que

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

En el caso de las tablas de contingencia 2×2, para una probabilidad de éxito $P(A) = \pi$, las *odds* de A son

$$odds = \frac{\pi}{(1-\pi)}$$

Las *odds* son no negativas, con valor mayor a 1 cuando el éxito es más probable que el fracaso. Cuando $odds = 4$, el éxito es 4 veces más probable que el fracaso. Caso contrario es cuando $odds = 1/4$, esto dice que el fracaso es cuatro veces más probable que el éxito.

En una tabla 2×2 de tipo 1, la fila 1 tiene $odds_1 = \frac{\pi_1}{1-\pi_1}$ y la fila 2 tiene $odds_2 = \frac{\pi_2}{1-\pi_2}$.
La razón de probabilidad (u odds ratio) se define como

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

3.1. Propiedades de la Odds Ratio

Puede tomar cualquier valor no negativo. Cuando X y Y son independientes con $\pi_1 = \pi_2$, entonces $odds_1 = odds_2$ y por lo cual $\theta = 1$. Cuando $\theta > 1$, las *odds* de éxito con más altas en la fila 1 que en la fila 2, esto implica que $\pi_1 > \pi_2$. Por ejemplo, cuando $\theta = 4$, las *odds* de éxito en la fila 1 es cuatro veces las *odds* de éxito de la fila 2. Cuando $\theta < 1$, el éxito es menos probable en la fila 1 que en la fila 2, esto es $\pi_1 < \pi_2$.

Observación. Tener cuidado: $\theta = 4$ implica que $\pi_1 > \pi_2$ pero no implica que $\pi_1 > 4\pi_2$.

Valores de θ lejos del 1 en cierta dirección representan una fuerte asociación. Una odds ratio de 4 está más lejos de la independencia que una odds ratio de 2, así como una odds ratio de 0.25 está más lejos de la independencia que una odds ratio de 0.5.
Dos valores de θ representan la misma fuerza de asociación pero en diferentes direcciones cuando un valor es el inverso del otro. Por ejemplo, cuando $\theta = 0{,}25$, las odds de éxito en la fila 1 son 0.25 veces las odds de éxito de la fila 2, o equivalentemente las odds de la filas 2 son $1/0.25 = 4$ veces las odds de la fila 1.
Además cuando se cambian las filas por las columnas, la odds ratio no cambia.

3.2. Otra forma de ver las Odds Ratios

Sea X el evento de que un individuo es expuesto a una patógeno potencialmente peligroso y D al evento en el que el individuo se enferma de ese patógeno. Denotaremos a los eventos complementarios como $\bar{X}$ y $\bar{D}$.
Se define

$$\theta = \frac{odds(D|X)}{odds(D|\bar{X})}$$

Visto de esta forma, la odds ratio es una medida de la influencia de exposición sobre la enfermedad.
Consideremos que tenemos una distribución multinomial como la tabla de tipo 2, usando las probabilidades conjuntas tenemos que

	$\bar{D}$	D	
$\bar{X}$	π_{11}	π_{12}	$\pi_{1.}$
X	π_{21}	π_{22}	$\pi_{2.}$
	$\pi_{.1}$	$\pi_{.2}$	

por lo que

$$P(D|X) = \frac{\pi_{22}}{\pi_{21} + \pi_{22}}$$
$$P(D|\bar{X}) = \frac{\pi_{12}}{\pi_{11} + \pi_{12}}$$

y entonces

$$odds(D|X) = \frac{P(D|X)}{1 - P(D|X)} = \frac{\pi_{22}}{\pi_{21} + \pi_{22}} : \frac{\pi_{21}}{\pi_{21} + \pi_{22}} = \\ = \frac{\pi_{22}}{\pi_{21}}$$

$$odds(D|\bar{X}) = \frac{\pi_{12}}{\pi_{11}}$$

y por lo tanto

$$\theta = \frac{odds(D|X)}{odds(D|\bar{X})} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

De hecho, a la odds ratio también se la conoce como el *cross-product ratio* ya que es el cociente entre el producto de las diagonales de la tabla.
Por lo que tenemos dos formas de estimar la odds ratio, o bien estimando las odds o bien estimando las probabilidades conjuntas.

Observación. Si consideramos ahora

$$\theta = \frac{odds(X|D)}{odds(X|\bar{D})}$$

veamos que llegamos al mismo resultado:

$$odds(X|D) = \frac{P(X|D)}{1 - P(X|D)} = \frac{\pi_{22}}{\pi_{12} + \pi_{22}} : \frac{\pi_{12}}{\pi_{12} + \pi_{22}} = \\ = \frac{\pi_{22}}{\pi_{12}}$$

$$odds(X|\bar{D}) = \frac{\pi_{21}}{\pi_{11}}$$

y por lo cual

$$\theta = \frac{odds(X|D)}{odds(X|\bar{D})} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

Consideremos tres posibles maneras de tomar un muestreo para estudiar la relación entre enfermos y exposición en una población:
1) Podría considerarse hacer un muestreo aleatorio de toda la población, así podemos estimar todas las probabilidades conjuntas directamente. Sin embargo, si la enfermedad es rara, se requerirá un gran muestreo para garantizar que un número substancial de individuos enfermos estén incluidos en el muestro.

2) Un segundo método de muestreo se llama **prospective study**, un cierto número fijado de personas expuestas y no expuestas son muestreadas y las incidencias de ambos grupos son comparadas. En este caso se puede estimar y comparar $P(D|X)$ y $P(D|\bar{X})$ y por lo tanto se puede estimar la *odds ratio*. Sin embargo, no se pueden estimar las probabilidades conjuntas π_{ij} porque los conteos marginales de individuos expuestos y no expuestos han sido fijados arbitrariamente por el diseño del muestreo.
3) Un tercer método de muestreo se llama **retrospective study**, un cierto número fijado de personas con la enfermedad y sin la enfermedad son muestreadas y las incidencias de exposición y no exposición en los dos grupos son comparadas. Esto permite estimar $P(X|D)$ y $P(X|\bar{D})$ y por lo tanto se puede estimar la *odds ratio*. Sin embargo, tampoco se pueden estimar las probabilidades conjuntas.
En este último caso, teniendo el resultado del muestreo como

	$\bar{D}$	D	
$\bar{X}$	n_{11}	n_{12}	$n_{1.}$
X	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	

Se puede estimar

$$\hat{P}(X|D) = \frac{n_{22}}{n_{.2}}$$

$$1 - \hat{P}(X|D) = \frac{n_{12}}{n_{.2}}$$

$$\hat{odds}(X|D) = \frac{n_{22}}{n_{12}}$$

Similarmente,

$$\hat{odds}(X|\bar{D}) = \frac{n_{21}}{n_{11}}$$

por lo que, la estimación de la odds ratio es

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

Observación. En el resto de casos se llega al mismo estimador.

Es más, para una distribución multinomial con 4 celdas (tabal tipo 2) o para distribuciones binomiales independientes con dos filas (tabla tipo 1), este estimador es el estimador de máxima verosimilitud (EMV) de θ. Esto sale, en el caso de la tabla tipo 1, por usar los EMV de $P(X|D)$ y $P(X|\bar{D})$ (o $P(D|X)$ y $P(D|\bar{X})$ según sea el caso) junto al teorema 1,0,2 y para el caso multinomial sale usando el mismo teorema junto con $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$ (EMV para la multinomial).
Como ejemplo, la tabla de Vianna, Greenwald y Davies (tabla 2,0,1) se obtuvo usando el método de **retrospective study**, la estimación de la odds ratio es

$$\hat{\theta} = \frac{67 \times 64}{43 \times 34} = 2{,}93$$

Por lo que, las odds de contraer la enfermedad de Hodgkin aumentan por un factor de 3 al someterse a una amigdalectomía.

3.3. Log Odds Ratios e inferencias

La distribución de la odds ratio es altamente sesgada. Cuando $\theta = 1$, por ejemplo, $\hat{\theta}$ no puede ser mucho más pequeño que θ (porque $\hat{\theta} \geq 0$), pero sí puede ser tan grande como quiera ya que no hay una cota superior.
Debido a este sesgo, en la inferencia estadística se usa una alternativa de una medida equivalente, $\log(\theta)$. Debido a su propiedad de que $\log(\frac{1}{n}) = -\log n$ hace que el logaritmo de la odds ratio sea simétrico alrededor del 0 en el sentido de que si invierto filas o invierto columnas cambio su signo. Dos valores de $\log(\theta)$ que son lo mismo pero en signo contrario, como por ejemplo $\log(2) = 0{,}7$ y $\log(0{,}5) = -0{,}7$ tienen la misma fuerza de asociación.
Además, duplicar el logaritmo de la odds ratio corresponde a tener un odds ratio al cuadrado, $2\log(\theta) = \log(\theta^2)$.
Para tablas tipo 1 se tiene que $\log(\hat{\theta})$ tiene aproximadamente una distribución normal con media $\log(\theta)$ y un desvío estandar de

$$SE = \sqrt{\frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}}$$

La demostración de que sigue aproximadamente una distribución normal no es fácil pero usando el método de propagación de errores se puede llegar la media y el desvío estándar anteriores.
Veámoslo, tenemos que N_{i1} y N_{i2} representan la cantidad de éxitos y fracasos de una binomial $B(n_{i.}, \pi_i)$ para $i = 1, 2$, además $N_{i2} = n_{i.} - N_{i1}$. Usando $g_1(x) = \log(\frac{x}{n_{1.}-x})$ y $g_2(x) = \log(\frac{n_{2.}-x}{x})$ se obtiene

$$\begin{aligned}
E(log(\hat{\theta})) &= E\left(\log\left(\frac{N_{11}}{N_{12}}\right) + \log\left(\frac{N_{22}}{N_{21}}\right)\right) = \\
&= E\left(\log\left(\frac{N_{11}}{n_{1.} - N_{11}}\right)\right) + E\left(\log\left(\frac{n_{2.} - N_{21}}{N_{21}}\right)\right) \approx \\
&\approx \log\left(\frac{n_{1.}\pi_1}{n_{1.} - n_{1.}\pi_1}\right) + \log\left(\frac{n_{2.} - n_{2.}\pi_2}{n_{2.}\pi_2}\right) = \\
&= \log\left(\frac{\pi_1}{1-\pi_1}\right) + \log\left(\frac{1-\pi_2}{\pi_2}\right) = \\
&= \log(\theta)
\end{aligned}$$

Por otro lado, como $g_1'(x) = \frac{n_{1.}}{x(n_{1.}-x)}$, $g_2'(x) = \frac{-n_{2.}}{x(n_{2.}-x)}$ y las binomiales son independientes se tiene

$$\begin{aligned}
Var(log(\hat{\theta})) &= Var\left(\log\left(\frac{N_{11}}{N_{12}}\right) + \log\left(\frac{N_{22}}{N_{21}}\right)\right) = \\
&= Var\left(\log\left(\frac{N_{11}}{n_{1.} - N_{11}}\right)\right) + Var\left(\log\left(\frac{n_{2.} - N_{21}}{N_{21}}\right)\right) \approx \\
&\approx n_{1.}\pi_1(1-\pi_1)\left(\frac{n_{1.}}{n_{1.}\pi_1(n_{1.} - n_{1.}\pi_1)}\right)^2 + n_{2.}\pi_2(1-\pi_2)\left(\frac{-n_{2.}}{n_{2.}\pi_2(n_{2.} - n_{2.}\pi_2)}\right)^2 = \\
&= \frac{1}{n_{1.}\pi_1(1-\pi_1)} + \frac{1}{n_{2.}\pi_2(1-\pi_2)} \approx \\
&\approx \frac{n_{1.}}{N_{11}N_{12}} + \frac{n_{2.}}{N_{21}N_{22}} = \frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}
\end{aligned}$$

Observación. Para tablas de tipo 2 hay que usar el método de propagación de errores para multivariables.

Mediante esto se puede construir un intervalo de confianza para $\log(\theta)$ de la forma

$$IC_{1-\alpha}(\log(\theta)) = [\log(\hat{\theta}) \pm z_{\alpha/2}(SE)]$$

y luego obtener un intervalo de confianza para θ.
Veamos un ejemplo, usando nuevamente la tabla de Vianna, Greenwald y Davies. Se tiene que $\log(\hat{\theta}) = 1{,}075$ y el desvío estándar es

$$SE = \sqrt{\frac{1}{67} + \frac{1}{64} + \frac{1}{43} + \frac{1}{34}} = 0{,}288$$

Con un nivel de confianza de 0,95, se obtiene que el intervalo de confianza para $\log(\theta)$ es $1{,}075 \pm 1{,}96(0{,}288)$ o

$$IC_{0{,}95}(\log(\theta)) = [0{,}51, 1{,}64]$$

lo que corresponde a un intervalo de confianza de θ de la forma

$$IC_{0{,}95}(\theta) = [e^{0{,}51}, e^{1{,}64}] = [1{,}66, 5{,}15]$$

Notar que como la distribución de θ es sesgada, el estimador $\hat{\theta}$ no es el centro del intervalo.
Como el intervalo de confianza de θ no contiene al 1, entonces hay una diferencia entre las odds; las odds de contraer la enfermedad de Hodgkin es mayor si se somete a una amigdalectomía, es más, al menos aumentan un 66 %.

La odds ratio $\hat{\theta}$ equivale a 0 o a ∞ si algún $n_{ij} = 0$, es por ello que una estimación ligeramente enmendada

$$\tilde{\theta} = \frac{(n_{11} + 0{,}5)(n_{22} + 0{,}5)}{(n_{12} + 0{,}5)(n_{21} + 0{,}5)}$$

es preferible. Y en este caso,

$$SE = \sqrt{\frac{1}{n_{11} + 0{,}5} + \frac{1}{n_{12} + 0{,}5} + \frac{1}{n_{21} + 0{,}5} + \frac{1}{n_{22} + 0{,}5}}$$

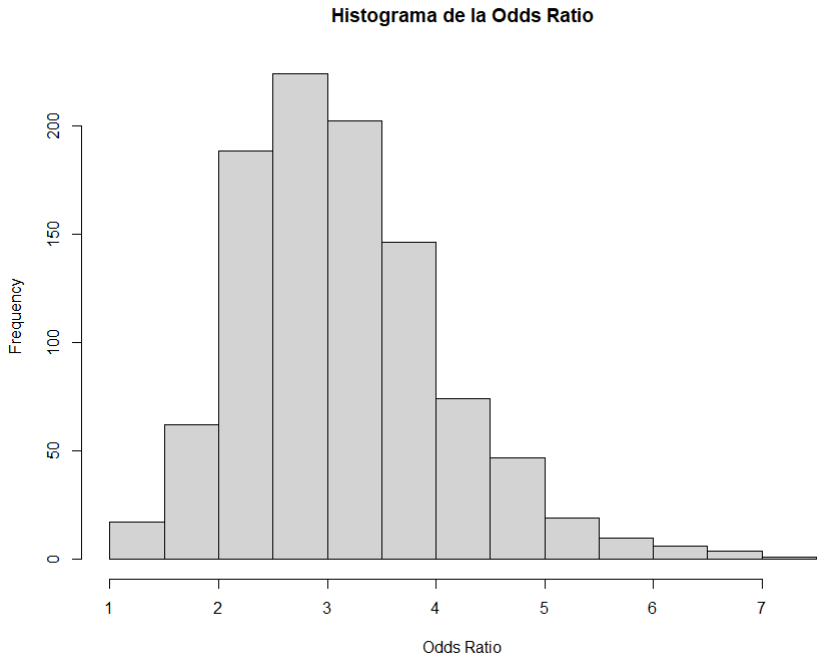
3.4. Desvío estandar usando Bootstrap

Otra forma de estimar el desvío estándar de $\hat{\theta}$ es mediante simulaciones de bootstrap.
Volviendo al ejemplo de Vianna, Greenwald y Davies (tabla 2,0,1):
El modelo indica que la cantidad N_{11} de la fila 1 y columna 1 tiene una distribución binomial con $n = 101$ ensayos y probabilidad π_1. La cantidad N_{22} de la fila 2 y columna 2 tiene una distribución independiente binomial con $n = 107$ ensayos y probabilidad π_2.
Por lo que la distribución de la variable aleatoria

$$\hat{\theta} = \frac{N_{11}N_{22}}{(101 - N_{11})(107 - N_{22})}$$

es determinado por dos distribuciones binomiales y podemos aproximarlo bien si se extrae un gran número de muestras de ellos.

Como las probablidades π_1 y π_2 son desconocidas, se estiman mediante los valores observados $\hat{\pi}_1 = \frac{67}{101} = 0{,}663$ y $\hat{\pi}_2 = \frac{64}{107} = 0{,}598$. Se generaron mediante computadora 1000 realizaciones de las variables binomiales N_{11} y N_{22}.
La siguiente figura muestra el histograma resultado de 1000 valores de $\hat{\theta}$.



con un desvío estandar muestral de 0.949.

3.5. Riesgo Relativo (Relative risk) y relación con el Odds ratio

El riesgo relativo (o relative risk) para tablas 2×2 de tipo 1 se define como

$$RR = \frac{\pi_1}{\pi_2}$$

No puede ser un número negativo. Si el riesgo relativo es igual a 1 entonces $\pi_1 = \pi_2$, esto dice quue la respuesta es independiente al grupo.
El riesgo relativo muestral se define como la proporción de las probabilidades estimadas, esto es

$$rr = \frac{\hat{\pi}_1}{\hat{\pi}_2}$$

donde $\hat{\pi}_i = \frac{N_i}{n_i}$ (EMV) donde $N_i \sim B(n_i, \pi_i)$ para $i = 1, 2$.

Observación. No confundir el riesgo relativo con la odds ratio. Por ejemplo, $\hat{\theta} = 1{,}83$ no significa que $\hat{\pi}_1$ es 1.83 veces $\hat{\pi}_2$, esa es la interpretación del riesgo relativo.

Se puede ver que $\log(rr) = \log(\frac{\hat{\pi}_1}{\hat{\pi}_2})$ se aproxima a una distribución normal con media $\log(\frac{\pi_1}{\pi_2})$ y

$$Var(\log(rr)) = \frac{1 - \hat{\pi}_1}{n_1 \hat{\pi}_1} + \frac{1 - \hat{\pi}_2}{n_2 \hat{\pi}_2}$$

Veamos que efectivamente esas son las aproximaciones de la media y la varianza:
Usando el método de propagación de errores y tomando $g_i(x) = \log(\frac{x}{n_i})$ para $i = 1, 2$ se tiene

$$\begin{aligned} E\left(\log\left(\frac{\hat{\pi_1}}{\hat{\pi_2}}\right)\right) &= E\left(\log\left(\hat{\pi_1}\right) - \log\left(\hat{\pi_2}\right)\right) = \\ &= E\left(\log(\hat{\pi_1})\right) - E\left(\log(\hat{\pi_2})\right) = \\ &= E\left(\log\left(\frac{N_1}{n_1}\right)\right) - E\left(\log\left(\frac{N_2}{n_2}\right)\right) \approx \\ &\approx \log\left(\frac{n_1\pi_1}{n_1}\right) - \log\left(\frac{n_2\pi_2}{n_2}\right) = \\ &= \log\left(\pi_1\right) - \log\left(\pi_2\right) = \log\left(\frac{\pi_1}{\pi_2}\right) \end{aligned}$$

Por otro lado, como $g_i'(x) = \frac{1}{x}$ y al ser distribuciones binomiales independientes se tiene

$$\begin{aligned} Var\left(\log\left(\frac{\hat{\pi_1}}{\hat{\pi_2}}\right)\right) &= Var\left(\log\left(\hat{\pi_1}\right) - \log\left(\hat{\pi_2}\right)\right) = \\ &= Var\left(\log(\hat{\pi_1})\right) + Var\left(\log(\hat{\pi_2})\right) = \\ &= Var\left(\log\left(\frac{N_1}{n_1}\right)\right) + Var\left(\log\left(\frac{N_2}{n_2}\right)\right) \approx \\ &\approx n_1\pi_1(1-\pi_1)\left(\frac{1}{n_1\pi_1}\right)^2 + n_2\pi_2(1-\pi_2)\left(\frac{1}{n_2\pi_2}\right)^2 = \\ &= \frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2} \approx \\ &\approx \frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2} \end{aligned}$$

Por ello se puede construir el intervalo de confianza de nivel $1-\alpha$ de la siguiente forma

$$IC_{1-\alpha}(\log(RR)) = [\log(\frac{\hat{\pi}_1}{\hat{\pi}_2}) \pm z_{\alpha/2}\sqrt{\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}}]$$

y por lo tanto

$$IC_{1-\alpha}(RR) = [e^{(\log(\frac{\hat{\pi}_1}{\hat{\pi}_2}) \pm z_{\alpha/2}\sqrt{\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}})}]$$

Veamos la relación entre el odds ratio y el riesgo relativo:

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = rr \times \frac{1-\hat{\pi}_2}{1-\hat{\pi}_1}$$

cuando $\hat{\pi}_1$ y $\hat{\pi}_2$ son cercanos a 0, $\hat{\theta}$ y rr toman valores similares.
En algunos casos, no es posible estimar el riesgo relativo, sin embargo sí es posible estimar la odds ratio y se puede usar esta para aproximar el riesgo relativo.

Referencias

[1] Vianna, N., Greenwald, P., and Davies, J. (1971). Tonsillectomy and Hodgkin's disease: The lymphoidtissue barrier. Lancet, 1, 431-432.

[2] Johnson, S., and Johnson, R. (1972). Tonsillectomy history in Hodgkin's disease. N. Eng. J. Med., 287, 1122-1125.

[3] Redelmeier, D. A., and Tibshirani, R. J. (1997). Association between cellular-telephone calls and motor vehicle collisions. N. Eng. J. Med., 336, 453-458.

[4] John A. Rice (2007). Mathematical Statistics and Data Analysis, Third Edition. 526-529.

[5] Alan Agresti (2007). An Introduction to Categorical Data Analysis, Second Edition. 28-32.