

Influencia de la cantidad de éxitos en la precisión de los intervalos de confianza de la razón de odds y en la potencia de la prueba, tomando en cuenta el tamaño de muestra.



Miguel Coto-García¹, Natalia Díaz-Ramírez¹, Andrés Flores-Cruz¹⁻²

¹Estudiante de Estadística, Universidad de Costa Rica, San Pedro, Costa Rica

²Estudiante de Microbiología y Química Clínica, Universidad de Costa Rica, San Pedro, Costa Rica

Resumen

Los modelos de regresión logística podrían presentar problemas con las estimaciones ya que estas se pueden indefinir. Esto sucede cuando hay una gran cantidad de éxitos. El problema de separación completa se debe a la estructura de los datos, la cual conduce a estimaciones de razón de odds infinitas.

El objetivo de la investigación es ilustrar la influencia en la precisión de los intervalos de la razón de odds y en la potencia de la prueba asociada a los coeficientes de la concentración de insecticida y especie, cuando hay un número alto de éxitos, tomando en cuenta el tamaño de muestra.

Se realizaron simulaciones tomando como base un experimento con un diseño de parcelas divididas y un modelo de regresión logística. La parcela corresponde a la concentración de insecticida (25 mg/L, 75 mg/L, 100 mg/L) y la subparcela es la especie (*Aedes aegypti*, *Aedes Albopictus*). El éxito corresponde a la muerte de un mosquito.

Como principal conclusión se obtuvo que cuando hay gran cantidad de éxitos en la variable respuesta, los porcentajes de intervalos de la razón de odds que se indefinen son altos y para tamaños de muestra pequeños el porcentaje de intervalos indefinidos es mayor que el porcentaje con un tamaño de muestra mayor.

Palabras clave

Simulación, parcelas divididas, regresión logística, virus



Introducción

Las enfermedades virales transmitidas por artrópodos (arbovirus) afectan a muchas personas a nivel mundial. En el continente americano, los arbovirus que generan más impacto son: el virus del dengue (DENV), virus del zika (ZIKV), ambos pertenecientes a la familia Flaviviridae especie *Flavivirus*, y virus chikungunya (CHIKV), el cual pertenece a la familia Togaviridae especie *Alphavirus*. [2]

Durante el 2016 en América, para DENV al 14 de octubre, se reportaron 2 048 182 casos, para CHIKV, en abril, se contaba con el reporte de 31 000 casos y en septiembre, solo Chile y Uruguay no reportaron casos de ZIKV. [4], [5], [6]

En Costa Rica, al 19 de octubre del 2016, existían 1,317 casos de ZIKV, 2 976 de CHIKV y 18 638 de DENV. [8]

En la actualidad la arbovirosis de mayor importancia mundial es la causada por DENV. Los mosquitos del género *Aedes*, son considerados los vectores naturales del DENV, siendo *Aedes aegypti* el principal vector en áreas urbanas y *Aedes albopictus* el segundo vector en importancia [7].

Se han desarrollado muchas técnicas para eliminar criaderos de estos vectores y a los adultos. Aun así, no se ha encontrado una alternativa que brinde resultados satisfactorios. A partir de esto, surge el interés por seguir analizando los efectos que nuevos tipos de insecticidas producen sobre los vectores.

A partir de un experimento, en el cual se pretendía analizar el efecto de la concentración de insecticida y especie en la probabilidad de muerte de un vector adulto (hembras), se va a simular el escenario donde se tiene gran cantidad de éxitos.

Un problema que aparece con frecuencia en los datos usados para un modelo de regresión logística es el de separación completa. La separación se puede definir como una división completa de los dos “grupos” de puntos asociados a los valores que toma la variable respuesta (en estos conjuntos de datos, la codificación general es 0 y 1). La principal consecuencia de la separación es la no existencia de los estimadores de máxima verosimilitud. [1] La separación conduce a estimaciones de razón de odds infinitas, las cuales rara vez se pueden asumir como ciertas en la práctica. [3]

La escasez de datos se relaciona con tamaños de muestra pequeños, lo cual es frecuente en muchos diseños de datos y si este tamaño de muestra es tan pequeño, que

conduce al problema de la separación, no es posible inferir a partir de este conjunto de datos. [1]

El presente trabajo tiene como objetivo ilustrar la influencia en la precisión de los intervalos de la razón de odds y en la potencia de la prueba asociada a los coeficientes de la concentración de insecticida y especie, cuando hay un número alto de éxitos en la variable respuesta, tomando en cuenta el tamaño de muestra.

Materiales y métodos

El diseño experimental corresponde a un diseño de parcelas divididas, sin bloques, con un modelo de regresión logística. Donde la unidad experimental es el mosquito, la variable respuesta es el número de muertes de mosquitos y la media de la variable respuesta es la probabilidad de muerte del mosquito.

Además, la parcela corresponde a la concentración de insecticida (25 mg/L, 75 mg/L, 100 mg/L) y la subparcela es la especie (*Aedes aegypti*, *Aedes Albopictus*).

Respecto al experimento, para cada cilindro de 15 cm, aproximadamente, se coloca una concentración de insecticida y se depositan los mosquitos correspondientes de las dos especies. Se realizó una corrida por lo que el total de cilindros a utilizar son 3, ya que en cada cilindro se coloca una sola concentración de insecticida.

Una vez depositados los mosquitos en el cilindro al pasar una hora se cambia la tapa donde se encuentra el insecticida y se coloca una nueva tapa sin insecticida. Luego a las 24 horas se verifican los mosquitos de cada especie que murieron.

El modelo que se va utilizar es el siguiente:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \alpha_2 I_2 + \alpha_3 I_3 + \gamma_2 E_2 \quad (1)$$

Donde:

$$\alpha_1 = -(\alpha_2 + \alpha_3) , \gamma_1 = -\gamma_2$$

$$I_2 = \begin{cases} 1 & \text{si insecticida es 75} \\ 0 & \text{si insecticida es 100} \\ -1 & \text{si insecticida es 25} \end{cases} , \quad I_3 = \begin{cases} 0 & \text{si insecticida es 75} \\ 1 & \text{si insecticida es 100} \\ -1 & \text{si insecticida es 25} \end{cases}$$

$$E_2 = \begin{cases} 1 & \text{si especie es aedes albopictus} \\ -1 & \text{si especie es aedes aegypti} \end{cases}$$

Los supuestos del modelo (1) son homoscedasticidad, independencia de los errores y Y/X se distribuye Binomial con parámetros n y p.

Para la elaboración de la simulación se asumió que no hay interacción entre la concentración de insecticida y la especie. Además, se consideró un nivel de significancia de 0,05.

Si las probabilidades de muerte de los mosquitos son altas debería de haber gran cantidad de éxitos (muerte de los mosquitos) en la variable respuesta. Para establecer las probabilidades se consideró que la razón de odds entre la concentración de insecticidas de 75 mg/L y la de 100 mg/L es de 1.5 y entre las especies *Aedes aegypti* y *Aedes albopictus* es de 2. Se tomaron en cuenta las siguientes probabilidades:

Cuadro 1. Probabilidades altas		
Concentración de insecticida	Especie	
	Aedes Aegypti	Aedes Albopictus
25 mg/L	0.975	0.95
75 mg/L	0.97	0.94
100 mg/L	0.98	0.96

Cuadro 2. Probabilidades moderadas		
Concentración de insecticida	Especie	
	Aedes Aegypti	Aedes Albopictus
25 mg/L	0.82	0.70
75 mg/L	0.79	0.65
100 mg/L	0.85	0.74

Para un tamaño de muestra fijo se tomaron en cuenta, probabilidades altas (ver cuadro 1) y probabilidades moderadas (ver cuadro 2). Para cada caso mencionado se realizaron mil simulaciones. Donde se obtuvo la amplitud de los intervalos de la razón de odds, correspondientes a cada comparación en la concentración de insecticida y en la especie, y se calculó el porcentaje de intervalos que se indefinen. También se tomó en cuenta el valor p, asociado a la concentración de insecticida y la especie, de la prueba de razón de verosimilitud, y se obtuvo la proporción promedio de rechazo. Lo anterior se hace primero para un tamaño de muestra pequeño de 4 réplicas por tratamiento y luego para un tamaño de muestra más grande, correspondiente a 12 réplicas por tratamiento.

Resultados

Para el análisis en la precisión de los intervalos de la razón de odds se obtuvo el porcentaje de intervalos que se indefinen, los resultados se presentan a continuación:

Cuadro 3. Porcentaje de intervalos de la razón de odds que se indefinen

Comparaciones	Tamaño de muestra pequeño		Tamaño de muestra grande	
	Probabilidades	Probabilidades	Probabilidades	Probabilidades
	altas	moderadas	altas	moderadas
Insecticida 25-75	97.9	27.4	79.5	0.3
Insecticida 25-100	97.9	27.4	79.5	0.3
Insecticida 75-100	92.7	17.8	64.7	0.2
A. Aegypti-A. Albopictus	86.7	5.5	46.5	0

Se puede observar que cuando las probabilidades de muerte son altas el porcentaje de intervalos que se indefinen es mayor que el porcentaje obtenido cuando las probabilidades son moderadas, es decir cuando hay gran cantidad de éxitos en la variable repuesta la precisión en los intervalos de la razón de odds se ve afectada ya que estos tienden a indefinirse.

Además, el tamaño de muestra también influye ya que cuando la muestra es pequeña el porcentaje de intervalos que se indefine es mayor que el porcentaje para un tamaño de muestra más grade.

Con los resultados obtenidos se podría considerar que cuando se tiene gran cantidad de éxitos en la variable respuesta las probabilidades se deben ajustar para que no sean tan altas, ya que con probabilidades altas los intervalos de la razón de odds tienden a indefinirse. Así mismo, se debería de tomar en cuenta el tamaño de muestra.

Por su parte, para el análisis de la potencia de la prueba asociada a los coeficientes de la especie y la concentración de insecticida en cada simulación se tomó en cuenta los valores p obtenidos de la prueba de la razón de verosimilitud y se obtuvo la proporción de rechazo.

Cuadro 4. Proporción de rechazo promedio

Variable	Tamaño de muestra pequeño		Tamaño de muestra grande	
	Probabilidades	Probabilidades	Probabilidades	Probabilidades
	altas	moderadas	altas	moderadas
Insecticida	0	0.08	0.04	0.07
Especie	0.02	0.08	0.08	0.06

Como se observa en el cuadro 4, en general la proporción promedio de rechazo de la prueba de razón de verosimilitud es muy baja.

En el caso de probabilidades moderadas, aunque un poco más altas, la proporción promedio de rechazo es muy baja respecto a lo que se podría esperar al usar razón de odds de 2 y de 1.5 entre las especies *Aedes aegypti* y *Aedes albopictus* y entre la concentración de insecticidas de 75 mg/L y la de 100 mg/L respectivamente; pues ambas diferencias deberían arrojar valores de rechazo promedio por arriba de un 90%. Esto quiere decir, que en una proporción menor al 10% de los casos simulados se detectó una diferencia significativa, vía la prueba de máxima verosimilitud.

El problema podría venir en que la diferencia (razón de odds) establecida es muy pequeña para ser detectada y la forma típica de solucionar este problema es aumentando el tamaño de muestra lo suficiente para que a esos niveles de diferencia antes mencionado se logre detectar diferencias en una proporción cercana a lo esperado. Con el inconveniente de que en este tipo de experimentos los tamaños de muestra suelen ser pequeños y los costos asociados a aumentarlos son bastante altos.


Además, para el caso de probabilidades altas podría estar influyendo el hecho de haber utilizado la prueba de razón de verosimilitudes, ya que los estimadores de máxima verosimilitud se pueden ver afectados cuando hay gran cantidad de éxitos.

Conclusiones

Cuando hay gran cantidad de éxitos en la variable respuesta, los intervalos de la razón de odds se indefinen. Además, el tamaño de muestra también afecta ya que se obtuvo que para tamaños de muestra pequeños el porcentaje de intervalos indefinidos es mayor que el porcentaje con un tamaño de muestra mayor.

Además, debido a la gran cantidad de éxitos de la variable respuesta y los datos usados en el análisis en la mayoría de los casos no fue posible detectar alguna diferencia significativa en los niveles de las variables de tratamiento. Por lo tanto, se recomienda utilizar tamaños de muestra mayores o bien razón de odds a detectar más amplias entre los distintos niveles.

Para el caso específico del experimento sobre el que se basó la simulación, esto implica que, si se utiliza un insecticida muy potente, es probable que los análisis a través de

un modelo de regresión logístico, muestren los problemas de estimaciones ya mencionados. Esto, sin embargo, no es del todo un problema, ya que se encontró una manera para tratar el problema. 

Referencias

- [1] Correa, J., Valencia, M. (2011). La separación en regresión logística, una solución y aplicación. *Rev. Fac Nac. Salud Pública*, 29(3), 281-288.
- [2] Fenner, F. (1976). The Classification and Nomenclature of Viruses. *J. gen. virol.*, 31, 463-470.
- [3] Heinze, G. (2006) A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statist. Med.*, 25, 4216–4226.
- [4] OMS (2016). *Chikungunya*. Recuperado de:
<http://www.who.int/mediacentre/factsheets/fs327/es/>
- [5] OMS (2016). *Informe sobre la situación virus de zika microcefalia síndrome de guillain-barré*. Recuperado de:
<http://apps.who.int/iris/bitstream/10665/250575/1/zikasitre13Oct16-spa.pdf?ua=1>
- [6] OMS. (sin fecha) Dengue. Recuperado de:
http://www.paho.org/hq/index.php?option=com_topics&view=article&id=1&Itemid=4073
4
- [7] Quintero, D. , Osorio, J., Martínez, M. (2010). Competencia vectorial: consideraciones entomológicas y su influencia sobre la epidemiología del Dengue. *IATREIA*, 23(2), 146-156.
- [8] Ministerio de Salud. (2016). Boletín epidemiológico Zika, Chikungunya y Dengue. Recuperado de:
<https://www.ministeriodesalud.go.cr/index.php/biblioteca-de-archivos/vigilancia-de-la-salud/analisis-de-situacion-de-salud/3130-boletin-epidemiologico-no-34-2016-zika-chikungunya-y-dengue/file>