



# FOCUS on Field Epidemiology

Enfoque en Epidemiología de Campo

## CONTRIBUYENTES

### Autores:

David Bergmire-Sweat, MPH  
Amy Nelson, PhD, MPH  
Grupo de trabajo FOCUS\*

### Críticos:

Grupo de trabajo FOCUS\*  
Gloria C. Mejía, DDS, MPH, PhD  
(Versión en español)

### Editoras de Producción :

Tara P. Rybka, MPH  
Lorraine Alexander, DrPH  
Rachel A. Wilfert, MD, MPH  
Gloria C. Mejía, DDS, MPH, PhD  
(Versión en español)

### Jefe de Edición:

Pia D.M. MacDonald, PhD, MPH

### Traducción al español por:

Pelusa Orellana

\* Todos los miembros del Grupo de Trabajo FOCUS están nombrados en la última página de la publicación.

## Análisis avanzado de datos: Métodos para el control de variables de confusión (apareamiento y regresión logística)

*No es sabio estar demasiado seguro de la propia sabiduría. Es sano recordar que el más fuerte puede debilitarse y que el más sabio puede errar.*

—Mahatma Gandhi (1869 - 1948)

### Introducción

Sorprendente. Desconcertante. Impresionante. Complejo. Místico. Perplejo. Asombroso. Confuso.

¿Qué piensas cuando lees estas palabras? En tus intentos por alcanzar un plano más alto en la práctica de la epidemiología puedes pensar en un sabio brillante estudiando un problema muy antiguo. O puedes pensar en las palabras de un viejo texto o en la historia de un pueblo olvidado. Tal vez tienes una visión más trivial y piensas en un mago asombroso.

En esta edición de FOCUS explorarás la capacidad mágica de la epidemiología para reducir los efectos de confusión, otorgándote las herramientas para volverte más sabio. Asegúrate, eso sí, de no volverte muy altivo por ello.

La epidemiología es la ciencia fundamental en la salud pública, por que usa el poder del análisis científico para evidenciar que una exposición puede estar asociada al desarrollo o la prevención de una enfermedad. A veces, sin embargo, los métodos más básicos de la epidemiología no son suficientes para determinar “las causas de las ocurrencias” o si una exposición está verdaderamente

asociada a una enfermedad. Otras exposiciones o características en la población pueden confundir la relación exposición-enfermedad, situación conocida como **efecto de confusión**.

Por ejemplo, podría parecer que las personas afectadas por un brote gastrointestinal son principalmente miembros de un mismo club gastronómico, pero muchos miembros también asistieron a un festival gastronómico de la ciudad. Se podría haber culpado a las prácticas de manipulación de alimentos al interior del club, cuando en efecto la causa haya sido la comida consumida en el festival, y habría un número mucho mayor de gente en riesgo de enfermar. La membresía al club gastronómico puede ser un factor que confunde la relación entre el asistir al festival de comida y la enfermedad. Si pudiéramos analizar los datos de manera de explicar al mismo tiempo tanto la membresía al club gastronómico como la asistencia al festival, podríamos determinar cuál de estos eventos estaba realmente asociado al brote.

En una edición anterior de FOCUS tratamos este tipo de situación a través de métodos de estratificación. Podríamos calcular el riesgo de enfermedad debido a la comida del festival entre quienes estaban en el club gastronómico, y luego calcular el riesgo de enfermedad por separado, entre quienes no estaban en el club. Si ambos grupos, o ambos estratos, muestran que la asistencia al festival de comida es un riesgo significativo para la enfermedad, luego el festival de comida puede estar implicado,



NORTH CAROLINA  
CENTER FOR PUBLIC  
HEALTH PREPAREDNESS

The North Carolina Center for Public Health Preparedness is funded by Grant/Cooperative Agreement Number U90/CCU424255 from the Centers for Disease Control and Prevention. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the views of the CDC.

porque la enfermedad ocurrió independientemente de si las personas eran o no miembros del club.

Sin embargo, ¿qué ocurre si hay múltiples factores que pudieran estar confundiendo la relación entre exposición y enfermedad? En el ejemplo anterior, ¿qué pasaría si hubiéramos tenido que estratificar por membresía al club y por estado de salud (ya que las personas con peor estado de salud son más susceptibles a la enfermedad)? Puede haber otros factores que también podrían ser considerados como variables de confusión potenciales, como por ejemplo edad, ocupación, ingreso, ejercicio y otros factores de estilo de vida. Estos otros factores pueden disfrazar asociaciones relevantes o efectos protectores. El tratar de estratificar según todos estos factores distintos se vuelve difícil de calcular, y el número de tablas de 2x2 que se necesitan crece enormemente.

Se necesitan métodos más avanzados para tratar este tipo de situación. En esta edición de FOCUS exploramos la regresión logística y explicamos datos apareados. La regresión logística es una manera eficiente de controlar al mismo tiempo muchas variables de confusión potenciales. El apareamiento, si se hace correctamente en la etapa del diseño de estudio de la investigación, reduce la confusión incluso antes de que comience el análisis.

### Confundiendo las variables que confunden.

Comúnmente, en la epidemiología de campo, comparamos dos grupos usando medidas de asociación tales como la razón de riesgos (RR) en estudios de cohorte, y la oportunidad relativa o razón de posibilidades (OR) en estudios de casos y controles. Por favor revisa las ediciones de FOCUS que cubren estos temas (*volumen 3, edición 1 y volumen 3 edición 2*) en detalle antes de continuar.

En un mundo perfecto, tú podrías recopilar datos, analizarlos usando una tabla de 2x2 para obtener una medida de asociación e intervalo de confianza (IC), o un ji-cuadrado ( $X^2$ ) y un valor p, y el análisis estaría completo. Podrías saber, sin embargo, que rara vez la epidemiología de campo es tan simple.

En algunas situaciones, podrás tener múltiples exposiciones que estén estadísticamente asociadas a la enfermedad... entonces, ¿cómo puedes saber cuál es el verdadero culpable? En otras situaciones, puede que *no* tengas exposiciones asociadas a la enfermedad cuando recopilaste los datos para evaluar cada exposición posible. En cualquiera de estas situaciones, necesitas ahondar en los datos. Puede haber una variable de confusión dando vueltas por ahí, haciendo que parezca que las exposiciones están asociadas con la enfermedad, cuando en realidad no lo están, o bien, que no exista asociación entre una exposición y la enfermedad cuando en realidad si la hay.

Una variable de confusión es una tercera variable que

distorsiona la razón de riesgo o la oportunidad relativa entre una exposición y un resultado. El fenómeno de confusión es una forma de sesgo que puede causar una distorsión en la medida de asociación entre una exposición y enfermedad y debe ser eliminada. (1)

El fenómeno de confusión puede ocurrir en un estudio epidemiológico observacional cada vez que se comparan dos grupos entre sí. Estos grupos pueden ser un grupo expuesto y un grupo no expuesto, como en un estudio de cohorte, o un grupo de enfermos y un grupo de no enfermos, como en el estudio de casos y controles. El fenómeno de confusión es una “mezcla de efectos” al comparar estos grupos entre sí. En cualquier relación exposición-enfermedad examinada, otros factores pueden afectar tanto la exposición como la enfermedad. Estos terceros factores pueden confundir la relación entre la exposición y la enfermedad que tratas de estudiar.

Entre algunas variables de confusión comunes que deben ser observadas se incluye la edad, grupo socioeconómico y sexo, entre otros. He aquí algunos ejemplos:

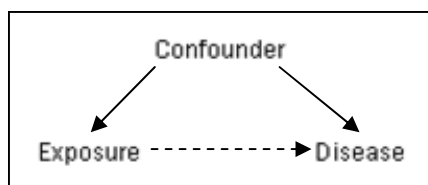
- Los niños nacidos mas tarde en el orden de nacimiento (que sean primogénitos, segundos, o menores, etc.) tienen más posibilidades de tener síndrome de Down. ¿Quiere decir que podemos concluir que el orden de nacimiento causa síndrome de Down? No, la relación entre el orden de nacimiento y el síndrome de Down está siendo confundida por la edad de la madre. Las mujeres mayores tienen más posibilidades de tener hijos con síndrome de Down. Las mujeres mayores también tienen más posibilidades de tener niños que sean los menores en el orden de los nacimientos (es mucho más común que una mujer de 35 años tenga 6 niños que una de 21 años de edad). La edad de la madre está confundiendo la asociación entre el orden de nacimiento y el síndrome de Down: pareciera que hubiera una asociación, cuando en realidad no la hay. (2)
- Algunos estudios del uso de terapia de reemplazo hormonal (TRH) muestran que no hay una asociación entre el uso de TRH y la enfermedad cardiovascular, sin embargo, otros sugieren que hay un aumento en el riesgo. Las mujeres de estratos socioeconómicos (ESE) altos, tienen más posibilidades de comprar TRH, y las mujeres de ESE bajo tienen, por lo general más riesgo de enfermedad cardiovascular. Las diferencias de ESE pueden causar confusión en la relación entre TRH y aumento del riesgo de enfermedad cardiovascular, que sólo se hace aparente cuando se controla el ESE entre las pacientes del estudio. (3)
- Un brote hipotético de gastroenteritis en un restaurante descubre que las mujeres tienen un riesgo más alto de enfermedad que los hombres. Sin embargo, esta

asociación está confundida por el consumo de ensaladas. Las mujeres tienden más a pedir ensaladas que los hombres, y la ensalada estaría contaminada con el agente que causó la enfermedad. La relación entre sexo y enfermedad está confundida por el consumo de ensalada (que fue, en realidad, la causa del brote).

Si examinas estos ejemplos verás que las variables de confusión tienen dos características fundamentales:

- Una variable de confusión debe estar asociada con la *enfermedad* que se estudia y
- Una variable de confusión debe estar asociada con la *exposición* que se estudia.

En otras palabras, el “triángulo” a la derecha debe darse para que se presente el efecto de confusión.



Ahora que sabes cómo reconocer el efecto de confusión, ¿cómo haces que desaparezca? Para controlarlo, debes eliminar del cuadro a la variable de confusión.

Existen tres formas de hacerlo:

1. **Restringe el análisis.** Solamente analiza la relación exposición-enfermedad entre aquellos en un nivel de la variable de confusión. (Por ejemplo, examina la relación de HRT y enfermedad cardiovascular sólo entre mujeres de estrato socioeconómico alto). El problema con este enfoque es que acabas por desear una gran parte de tus datos (en este ejemplo, todas las mujeres de estrato socioeconómico bajo).
2. **Estratifica.** Analiza la relación de la exposición y la enfermedad por separado en todos los niveles de la variable de confusión. (Por ejemplo, examina la relación entre HRT y enfermedad cardiovascular entre mujeres de nivel socioeconómico alto, y separadamente, entre mujeres de nivel socioeconómico bajo. El estrato socioeconómico incluso puede dividirse en múltiples niveles tales como bajo, medio, y alto).
3. **Realiza regresiones logísticas.** La regresión es una manera de poner todas las variables dentro de un modelo matemático. La regresión es muy útil cuando hay múltiples variables de confusión que requieren ser controladas (tal vez nivel socioeconómico y sexo y edad).

*Estratificación para controlar variables de confusión*

La estratificación puede ser usada para separar los efectos de las exposiciones y de las variables de confusión en un resultado. Por ejemplo, en un brote de tuberculosis entre hombres vagabundos, se ha implicado como el lugar

de transmisión tanto un hogar para vagabundos como un comedor de asistencia pública (olla común u olla popular). Los hombres solían pasar tiempo en ambos lugares. Para determinar cuál sitio debiera tener prioridad para la intervención, podríamos examinar la asociación entre el hogar para vagabundos y la tuberculosis entre hombres que *no* fueron al comedor público, y luego examinar la misma relación por separado entre hombres que *sí* fueron al comedor público.

Tomemos un ejemplo y hagamos los cálculos. Un brote ha ocurrido en una recepción y se ha implicado tanto a las galletas como el ponche en los análisis iniciales. Sospechamos que uno de estos artículos de comida está confundiendo al otro, pero no podemos eliminar los efectos sin estratificar ya que muchas personas consumen tanto galletas como ponche.

Después de realizar un estudio de casos y controles, los datos globales se ven de la siguiente manera:

**Exposición a las galletas**

	Casos	Controles	Total
Galletas	37	21	58
No galletas	13	29	42
Total			100

$$OR = (37 \times 29) / (21 \times 13) = 3.93; 95\% CI, 1.69 - 9.15$$

$$p = 0.001^*$$

**Exposición al Ponche**

	Casos	Controles	Total
Ponche	40	20	60
No Ponche	10	30	40
Total			100

$$OR = (40 \times 30) / (20 \times 10) = 6.00; 95\% CI, 2.83 - 12.71$$

$$p = 0.0004^*$$

Tanto las galletas como el ponche tienen una alta oportunidad relativa (OR) para la enfermedad e intervalos de confianza que no incluyen 1. Podemos estratificar para tratar de eliminar estos efectos.

*\*Nota: los intervalos de confianza (CI) y los valores p exactos fueron calculados en tablas con tamaños de celda menores a 5 usando SAS 9.13. La prueba estadística X<sup>2</sup> (ji-cuadrado) fue utilizada para detectar diferencias entre grupos expuestos y no expuestos y entre enfermos y no enfermos, de acuerdo a la definición de caso.*

Para estratificar según la exposición al ponche, queremos saber:

- *Entre quienes no tomaron ponche, ¿cuál es la medida de efecto (OR) para la asociación entre galletas y enfermedad?*
- *Entre quienes sí tomaron ponche, ¿cuál es la medida de efecto para la asociación entre galletas y enfermedad?*

Si las galletas son las culpables, tendría que haber una asociación entre galletas y enfermedad, independientemente de si alguien tomó o no ponche.

Estratificación de la asociación de galletas según exposición al ponche:

**Tomaron Ponche**

	Casos	Controles	Total
Galletas	35	17	52
No galletas	5	3	8
Total			60

$$OR = (35 \times 3) / (17 \times 5) = 1.3; 95\% \text{ CI, } 0.17 - 7.22$$

**No tomaron ponche**

	Casos	Controles	Total
Galletas	2	4	6
No galletas	8	26	34
Total			40

$$p = 1.0^*$$

$$OR = (2 \times 26) / (4 \times 8) = 1.63; 95\% \text{ CI, } 0.12 - 13.86$$

$$p = 0.63^*$$

Para estratificar según exposición a las galletas, queremos saber:

- *Entre quienes no comieron galletas, ¿cuál es la medida de efecto para la asociación entre el ponche y la enfermedad?*
- *Entre quienes sí comieron galletas, ¿cuál es la medida de efecto para la asociación entre ponche y enfermedad?*

Si el ponche es el culpable, tendría que haber una asociación entre el ponche y la enfermedad, independientemente de si alguien consumió o no galletas.

Estratificación de la asociación del ponche según exposición a las galletas:

**Comieron galletas**

	Casos	Controles	Total
Ponche	35	17	52
No Ponche	2	4	6
Total			58

$$OR = (35 \times 4) / (17 \times 2) = 4.12; 95\% \text{ CI, } 0.52 - 48.47$$

$$p = 0.18^*$$

**No comieron galletas**

	Casos	Controles	Total
Ponche	5	3	8
No Ponche	8	26	34
Total			42

$$OR = (5 \times 26) / (3 \times 8) = 5.42; 95\% \text{ CI, } < 0.80 - 40.95$$

$$p = 0.08^*$$

La estratificación nos permite examinar dos factores de riesgo independientemente el uno del otro. En este ejemplo, podemos ver que la galletas no fueron realmente un factor de riesgo independiente del ponche (ORs estratificados  $\approx 1$ ), mientras que el ponche se mantuvo como un factor potencial de riesgo independiente de las galletas (las tasas de posibilidad son grandes y los valores p cercanos a significativos). El paso siguiente sería averiguar qué pasó en la compra, preparación y almacenamiento de ese ponche, que hubiera hecho enfermar a las personas (es decir, una investigación de salud ambiental).

¿Recuerdas a los estadísticos Mantel y Haenszel, de la última edición de FOCUS? Bien, ellos diseñaron una manera inteligente para controlar el efecto de confusión usando análisis estratificado. En el análisis estratificado anterior nos quedamos con dos valores de la asociación entre el ponche y la enfermedad por estratos específicos: 4.1 y 5.4. Pero sería mucho más conveniente tener un solo valor; después de todo las dos oportunidades relativas son cálculos de la misma oportunidad relativa “real” que mide la asociación entre el ponche y la enfermedad. La oportunidad relativa de Mantel-Haenszel toma una asociación (como la nuestra entre ponche y enfermedad), la estratifica según una variable de confusión potencial (como nuestras galletas), y luego las recombina promediándolas en un valor en el que está “controlada” la variable estratificadora. Esta oportunidad relativa recombina es la **oportunidad relativa común**. Esta es una técnica muy útil de usar cuando sabes que tienes una variable de confusión pero quieres presentar sólo un cálculo. También funciona bien cuando estratificas según variables

que tienen más de dos valores potenciales, tales como la cantidad consumida de un artículo de comida (dos pedazos, o un mordisco, o nada), o el número de veces que una persona fue a nadar. Pero probablemente no estratificarías por nivel en una variable continua ya que los números serían demasiado pequeños. Algunas exposiciones están asociadas a enfermedad, pero sólo una vez que se alcanza una cantidad crítica de exposición. En esas instancias, las personas que estuvieron expuestas a un bajo nivel podrían estar confundiendo la capacidad de identificar la asociación debido a que no enfermaron pese a estar expuestos.

#### *Estratificación y Modificadores de Medidas de Efecto*

Hemos visto una estratificación que no produce asociación (exposición a la galleta estratificada por consumo de ponche), y una estratificación que produce una asociación consistente y fuerte (exposición al ponche estratificada por consumo de galleta). Pero existe un tercer resultado posible. Podríamos descubrir que un estrato mostraba que no había asociación con una oportunidad relativa cercana a 1.0, mientras que el otro estrato mostraba que sí había asociación. Esta ocurrencia se conoce como **modificación de la medida de efecto**. En este caso, la tercera variable no está confundiendo la asociación entre exposición y resultado y no es un sesgo que quisiéramos eliminar. En lugar de ello, queremos identificar y presentar por separado cálculos para cada nivel. Así que si el sexo es un modificador de la medida de efecto, deberías entregar dos oportunidades relativas o razones de riesgo: una para los hombres y una para las mujeres. Identificamos la modificación de medidas de efecto por estratificación, la misma técnica que usamos para identificar variables de confusión, pero estamos buscando que la medida de efecto sea bien diferente (heterogénea) entre los 2 o más estratos. Por ejemplo:

- Entre los adultos mayores, el sexo es un modificador de efecto de la asociación entre consumo nutricional y osteoporosis. El consumo nutricional (especialmente calcio) se asocia con resultado de osteoporosis en las mujeres. Entre los hombres no es tan fuerte porque el contenido mineral de los huesos de los hombres es distinto al de las mujeres y no está tan afectado por el consumo nutricional.
- En los países en vías de desarrollo, la higiene es un modificador de efecto en la asociación entre bebés alimentados por leche materna y mortalidad infantil. En condiciones que son muy poco sanitarias, la alimentación por leche materna tiene un fuerte efecto en la reducción de la mortalidad infantil. En condiciones mucho más limpias, sin embargo, la mortalidad infantil no difiere mucho entre niños alimentados con leche materna o con biberón. (En este ejemplo la higiene se podría considerar como una variable de

aproximación para estrato socioeconómico).

#### **Apareamiento**

En una edición anterior de FOCUS mencionamos la técnica de apareamiento de características de casos y controles para reducir el fenómeno de confusión. Si necesitas un repaso en el tema de apareamiento, revisa la edición de FOCUS sobre estudios de casos y controles (*Volumen 3 edición 2*). En estudios de cohorte, las personas no expuestas son apareadas a personas expuestas en relación a las características deseadas. En estudios de casos y controles, los controles son apareados a casos en relación a las características deseadas. Generalmente tienes que explicar el porqué del apareamiento cuando realizas el análisis de datos apareados, de lo contrario terminas introduciendo sesgo. Piensa esto por un momento: si eliges controles que tengan la misma edad, sexo y ocupación que los casos mediante apareamiento, ¡estás introduciendo a propósito un sesgo de selección en tu estudio! Eso está bien si es que el apareamiento reduce el fenómeno de confusión, siempre y cuando expliquemos el apareamiento en el análisis y las variables apareadas no sean exposiciones de interés.

Digamos que estamos realizando un estudio en una escuela en la que un número de estudiantes han informado de un extraño olor y enfermedad repentina. Evaluamos la asociación entre el oler algo inusual en la escuela y un conjunto de síntomas (náusea, vómitos, desmayos), y apareamos casos y controles sobre la base de sexo, curso, y pasillo. ¿Por qué aparear sobre la base de estos factores? Hay un precedente de un “brote” de enfermedad con olores poco comunes en edificios, por lo general sin una causa definida incluso después de mucha investigación ambiental y epidemiológica. Algunos hacen hipótesis de que este tipo de enfermedad es psicogénica, lo que significa que una vez que unas pocas personas sienten el olor y se sienten enfermos, otros en el edificio reaccionan ante esto con cierto pánico, causando una enfermedad contagiosa que se esparce más bien por histeria que por una causa real. Los factores sobre los cuales apareamos son variables de confusión potenciales. El sexo es una variable de confusión porque las mujeres tienden a ser más reactivas en este tipo de situaciones; el curso es una manera de controlar la edad, ya que estudiantes mayores o más jóvenes pueden reaccionar de manera diferente; y el apareamiento según pasillo controla el olor percibido, ya que los alumnos en un pasillo en el que hay una sala de química podrían percibir un olor inusual diferente a como lo percibirían alumnos en un pasillo cercano a la cafetería.

Debido a que hemos apareado casos y controles, armamos una tabla de 2x2 distinta: examinamos *parejas* en la tabla, por lo que tenemos casos en uno de los lados y controles en el otro, y las celdas de la tabla contienen parejas.

**Tabla 1: Análisis apareado para un estudio de casos y controles**

		Controles		Total
		Expuestos	No expuestos	
Casos	Expuestos	e	f	e + f
	No expuestos	g	h	g + h
	Total	e + g	f + h	

La celda “e” contiene el número de parejas caso-control apareadas en las que tanto el caso como el control habían estado expuestos. Esta es una celda concordante (y también lo es la celda “h”) porque el caso y el control tienen el mismo estado de exposición. La celda “f” contiene el número de pares caso-control apareados en los que los casos estuvieron expuestos pero los controles de cada par no habían estado expuestos. Esta es una celda discordante (y también lo es la celda “g”) porque el caso y el control tienen un estado de exposición distinto. Ya que queremos contrastar la exposición entre casos y controles, sólo las celdas discordantes (f y g) nos otorgan información útil.

Un ji-cuadrado para datos apareados – de otro estadístico llamado McNemar - puede ser fácilmente calculado a partir de estos datos usando un programa computacional estadístico. El cálculo examina parejas discordantes y luego otorga un valor McNemar ji-cuadrado y un valor p. Si el valor p es <0.05 (conoces este ejercicio) puedes concluir que hay una diferencia estadísticamente significativa en la exposición entre casos y controles.

También podemos usar la tabla de parejas discordantes para calcular una medida de asociación. La tabla 2 usa datos de ejemplo del brote de enfermedad repentina mencionado anteriormente.

**Tabla 2: Datos de ejemplo para enfermedad repentina en una escuela.** Los controles apareados con los casos según sexo, curso, y pasillo al interior de la escuela.

		Controles		Total
		Olor	No olor	
Casos	Olor	6	12	18
	No olor	4	5	9
	Total	10	17	

Podemos calcular la tasa de posibilidades directamente a partir de estos datos.

$$OR = \frac{\text{\# de pares con casos expuestos y no expuestos}}{\text{\# de pares con casos no expuestos y controles expuestos}} = f / g = 12 / 4 = 3.0$$

Interpretación: la oportunidad o posibilidad de presentar repentinamente náuseas, vómitos y desmayos si los alumnos sintieron un olor inusual en la escuela es 3.0 veces la oportunidad de tener repentinamente estos síntomas si los alumnos no sintieron el olor extraño en la escuela, controlando el sexo, el curso y la ubicación física en la escuela.

Una nota importante acerca del apareamiento: una vez que has apareado sobre la base de una variable, ¡no puedes usar esa variable como factor de riesgo en tu análisis! Los casos y controles tendrán exactamente el mismo nivel de las variables apareadas (porque las elegiste conscientemente para que fuesen así), de modo que son inútiles como factores de riesgo. La moraleja del cuento: ¡no apares sobre la base de ninguna variable de la que sospeches pueda ser un factor de riesgo! Si las apareas, asegúrate de tener una buena razón para hacerlo.

**Introducción a la Regresión Logística.**

En la cumbre de la epidemiología de campo, en relación al análisis de datos, se encuentra la regresión logística. Al igual que con muchos términos estadísticos, las palabras suenan complicadas pero el proceso es relativamente sencillo. La regresión logística es simplemente un proceso matemático que resulta en una oportunidad relativa, lo cual está disponible para los epidemiólogos comunes y corrientes gracias al computador. Lo especial acerca de la regresión logística es que puedes controlar las variables de confusión. No estamos hablando de una sola variable de confusión, como en los métodos de estratificación, nos referimos a controlar tantas variables como desees (siempre y cuando tengas un tamaño de muestra lo suficientemente grande, pero discutiremos eso más tarde). La regresión logística es un modelo matemático que puede entregar una oportunidad relativa conocido como oportunidad relativa “ajustada”, porque su valor ha sido ajustado en razón a las variables de confusión. Algunos epidemiólogos usan la regresión logística todo el tiempo y no pueden vivir sin ella, mientras que otros piensan que las tablas de 2x2 son, en realidad, todo lo que necesitan.

Al igual que en la tabla de 2x2, la variable resultante (enfermo o no enfermo) y la variable de exposición (expuesta o no expuesta) deben ser ambas dicotómicas. Otras variables (las variables de confusión que quieres ajustar) pueden ser dicotómicas, categóricas, con distintos niveles, o continuas.

Tenemos que tomar en cuenta el hecho de que la regresión implica matemáticas avanzadas, pero no te preocupes, esta discusión no se volverá más complicada. El meollo del asunto es que la regresión logística usa una ecuación llamada *función logit* para calcular la oportunidad relativa. Revisemos el ejemplo anterior en el que tanto las galletas como el ponche han sido implicados en el análisis inicial como la fuente de un brote en una recep-

ción. Sospechamos que uno de estos artículos de comida esté confundiendo el otro, pero no podemos eliminar los efectos sin estratificar porque muchas personas consumen tanto galletas como ponche. Las variables son:

- ENFERMO (donde el valor es 1, si está enfermo, ó 0 si no está enfermo)
- PONCHE (1 si tomó ponche, 0 si no tomó ponche)
- GALLETAS (1 si comió galletas, 0 si no comió galletas)

En general, la ecuación se vería de la siguiente manera:

$$\text{Logit (RESULTADO)} = \text{EXPOSICION} + \text{VARIABLE DE CONFUSIÓN 1} + \text{VARIABLE DE CONFUSIÓN 2} + \text{VARIABLE DE CONFUSIÓN 3} + \dots \text{ (etc.)}$$

En nuestro caso, el resultado es la variable ENFERMO. La exposición es la variable PONCHE, y la variable de confusión es GALLETAS. La ecuación se ve de la siguiente manera:

$$\text{Logit (ENFERMO)} = \text{PONCHE} + \text{GALLETAS}$$

La computadora usa la matemática de fondo de la regresión logística (no cubierto aquí, pero de fácil acceso si te interesa) para entregar los resultados como oportunidades relativas.

Cada variable en el lado derecho de la ecuación (estas son variables dependientes) tendrán su propia oportunidad relativa. En nuestro ejemplo, la oportunidad relativa para el PONCHE será la oportunidad o posibilidad de enfermarse si se consumió ponche, en comparación a la oportunidad de enfermarse si no se consumió el ponche, controlando por las GALLETAS. Si tuviéramos otras variables de confusión en la ecuación (como variables dependientes), la oportunidad relativa para el PONCHE también estaría controlando esas variables.

Más aún, hay una oportunidad relativa para las GALLETAS. Esta oportunidad relativa es la oportunidad de enfermarse si se consumieron galletas en comparación a la oportunidad de enfermarse si no se consumieron galletas, controlando el PONCHE.

**Ecuación logística y oportunidad relativa:**

$$\text{logit (ENFERMO)} = \text{PONCHE} + \text{GALLETAS}$$

**OR ponche:**

La oportunidad de enfermar entre quienes consumieron ponche comparados con quienes no consumieron ponche (controlando por galletas)

**OR galletas:**

La oportunidad de enfermar entre quienes consumieron galletas comparados con quienes no consumieron galletas (controlando por ponche)

Cada variable en el lado derecho de la ecuación está controlando todas las otras variables en el lado derecho de la ecuación. Esto hace de la regresión logística algo muy útil para los epidemiólogos. Si no estás seguro si una de varias variables es una variable de confusión, puedes examinarlas todas al mismo tiempo. Sin embargo, hay dos advertencias importantes:

- No ingreses demasiadas variables en la ecuación. Si tienes una muestra de 30 o menos (por ejemplo, 30 observaciones o menos), debieras sólo poner una variable dependiente en la ecuación. En este caso, también podrías usar una tabla de 2x2. Una tabla de 2x2 te dará la misma oportunidad relativa que la regresión logística con sólo una variable dependiente. Una regla general es que puedes agregar a la ecuación una variable por cada 25 observaciones. Por lo tanto, si estás analizando 50 sujetos de estudio, puedes incluir 2 variables dependientes en el modelo de regresión. En el ejemplo de las galletas y el ponche, tienes 200 casos y controles, lo que significa que puedes tener hasta 8 variables dependientes en el modelo de regresión para explicar el resultado u observaciones.
- No puedes controlar variables de confusión que no has medido. Si la asistencia de un niño en un centro infantil determinado fuera una variable de confusión en la relación ENFERMEDAD-PONCHE, pero no tienes datos sobre la asistencia del niño al centro infantil, no podrías controlarlo. De manera similar, podría haber una variable de confusión de la cual no se te ocurrió recopilar información así que no puedes controlarla. Esto siempre es posible, y es parte de la razón por la que nunca podemos estar 100% seguros de nuestros cálculos oportunidades relativas o razones de riesgo.

La regresión logística también puede fácilmente tener en cuenta el apareamiento en el análisis de los datos, usando un método especial llamado *regresión logística condicional*. El computador calcula las oportunidades relativas de manera similar a la prueba de McNemar, pero los resultados están “condicionados” sobre las variables de apareamiento. Esto también puede hacerse en Epi Info. La interpretación de las oportunidades relativas apareadas usando regresión logística condicional es la misma de las oportunidades relativas calculadas a partir de tablas.

Muchos epidemiólogos de campo jamás tendrán que utilizar regresión logística al analizar datos de brotes u otras investigaciones. Sin embargo, la regresión logística es extremadamente útil en el manejo de variables de confusión o de aquellas que potencialmente pueden causar confusión, y es particularmente útil cuando se analizan bases de datos grandes y estudios diseñados para establecer factores de riesgo para condiciones crónicas, investigaciones de conglomerados de cáncer, evaluaciones de

exposiciones ambientales, y otras situaciones en las que numerosos factores de confusión podrían oscurecer las relaciones entre factores de riesgo asociados con resultados de enfermedad. Muchos paquetes de programas informáticos pueden simplificar el análisis de datos usando la regresión logística, incluyendo el SAS, SPSS, STATA, y EpiInfo.\*

### Ejemplos: Regresión Logística

En 1997, los asistentes a una cena de boda en una residencia privada se quejaron de enfermedad diarreica diagnosticada como cyclosporiasis. En el análisis univariable (en la que cada exposición se analiza por separado en una tabla de 2x2), la exposición al consumo de frambuesas fue la exposición más fuertemente asociada con el riesgo de enfermar; más aun, fue la única exposición significativamente asociada al riesgo de enfermedad en un análisis de regresión logística multivariable. Al final, los investigadores determinaron que las frambuesas no habían sido lavadas. (4)

Para evaluar la relación entre la obesidad y la preocupación sobre seguridad alimentaria, el Departamento Estatal de Salud de Washington analizó datos a nivel estatal del sistema de vigilancia de factores conductuales de riesgo. Una variable que indicaba preocupación acerca de seguridad alimentaria se analizó usando un modelo de regresión logística, incluyendo el ingreso económico y la educación como posibles variables de confusión. Las personas

que informaron haber estado preocupados de la seguridad alimentaria eran más propensos a ser obesos que aquellos que no informaron tales preocupaciones (ORajustado = 1.29, 95% CI: 1.04-1.83). (5)

### Ejemplos: Apareamiento y regresión logística condicional

En 2002, un brote alimentario de *Salmonella* Newport se detectó entre 47 personas de 5 estados distintos. Se aparearon controles obtenidos a través de discado telefónico al azar según grupo etario con casos de Nueva York, Michigan, y Pennsylvania. Se realizó regresión logística para calcular oportunidades relativas controlando las variables de confusión (estas no fueron especificadas, pero pudieron haber sido sexo y estado). La investigación reveló que los casos eran más propensos a haber comido carne molida, en comparación a los controles (ORapareado=2.3, 95% CI: 0.9-5.7), y tuvieron más posibilidades de haber comido carne cruda o poco cocida en comparación a los controles (ORapareado=50.9, 95% CI: 5.3-489.0). No se identificó ningún evento específico de contaminación, y se emitió una alerta de salud pública para recordar a los consumidores acerca de las prácticas seguras de manipulación de alimentos. (6)

Un estudio de casos y controles se llevó a cabo para determinar factores de riesgo para desarrollar fiebre tifoidea (*Salmonella* serotipo Typhi) en un brote en Tajikistan afectando a más de 10,000 personas en 1996-1997. Los casos presentaban cultivos positivos para el organismo y se apareó a controles y casos por edad y vecindario. Usando una tabla de 2x2, la enfermedad se asoció al consumo de agua no hervida en los 30 días previos a su inicio (ORapareado= 6.5, 95%CI: 3.0-24.0), la obtención de agua potable de un grifo exterior (ORapareado= 9.1, 95% CI: 1.6-82.0), y el consumo de comida de un vendedor callejero (ORapareado= 2.9, 95% CI: 1.4-7.2). Una vez que todas las variables fueron incluidas en un análisis de regresión logística condicional para eliminar los efectos de cada factor manteniendo fijos todos los otros datos, sólo el consumo de agua no hervida (ORapareado = 9.6, 95%CI: 2.7-34.0) y la obtención de agua a partir de un grifo exterior (ORapareado= 16.7, 95% CI:2.0-138.0) estaban significativamente asociadas a la enfermedad. El hecho de hervir el agua en casa resultó ser protectorio (ORapareado= 0.2, 95% CI: 0.05-0.5). (7)

### Conclusión

En este número de FOCUS hemos discutido el complejo tema de control del fenómeno de confusión. El objetivo de este número era el comprender el uso del diseño de estudio apareado y la regresión logística con múltiples variables para controlar el efecto de confusión. Con un poco de práctica, estos métodos se te harán tan fáciles como la creación de una tabla de 2x2.

\*Los siguientes paquetes de programas informáticos son utilizados frecuentemente por epidemiólogos para el análisis de datos, incluyendo la regresión logística. Para mas información sobre programas específicos visite sus sitios Web en Internet. Esta no es una lista comprensiva y UNC no respalda específicamente a ningún paquete particular.

SAS – Cary, North Carolina

<http://www.sas.com/index.html>

SPSS – Chicago, Illinois

<http://www.spss.com/>

STATA –College Station, Texas

<http://www.stata.com>

Epi Info –Atlanta, Georgia

<http://www.cdc.gov/EpiInfo/>

Episheet – Boston, Massachusetts

<http://members.aol.com/krothman/modepi.htm>

(Episheet no puede hacer regresión logística pero es útil para análisis más simples, por ejemplo, tablas de 2x2 y análisis estratificado.)

## CONTACTO:

The North Carolina Center for Public Health Preparedness  
The University of North Carolina at Chapel Hill  
Campus Box 8165  
Chapel Hill, NC 27599-8165

Phone: 919-843-5561  
Fax: 919-843-5563  
Email: [nccphp@unc.edu](mailto:nccphp@unc.edu)

### Equipo de trabajo FOCUS:

- Lorraine Alexander, DrPH
- Meredith Anderson, MPH
- David Bergmire-Sweat, MPH
- Kim Brunette, MPH
- Anjum Hajat, MPH
- Pia D.M. MacDonald, PhD, MPH
- Gloria C. Mejia, DDS, MPH
- Amy Nelson, PhD, MPH
- Tara P. Rybka, MPH
- Rachel A. Wilfert, MD, MPH

## REFERENCIAS:

1. Gregg MB. *Field Epidemiology*. 2<sup>nd</sup> ed. New York, NY: Oxford University Press; 2002.
2. Hecht CA, Hook EB. Rates of Down syndrome at livebirth by one-year maternal age intervals in studies with apparent close to complete ascertainment in populations of European origin: a proposed revised rate schedule for use in genetic and prenatal screening. *Am J Med Genet*. 1996;62:376-385.
3. Humphrey LL, Nelson HD, Chan BKS, Nygren P, Allan J, Teutsch S. Relationship between hormone replacement therapy, socioeconomic status, and coronary heart disease. *JAMA*. 2003;289:45..
4. Centers for Disease Control and Prevention. Update: Outbreaks of Cyclosporiasis – United States, 1997. *MMWR Morb Mort Wkly Rep*. 1997;46:461-462. Available at: <http://www.cdc.gov/mmwr/PDF/wk/mm4621.pdf>. Accessed December 12, 2006.
5. Centers for Disease Control and Prevention. Self-reported concern about food security associated with obesity – Washington, 1995–1999. *MMWR Morb Mort Wkly Rep*. 2003;52:840-842. Available at: <http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5235a3.htm>. Accessed December 12, 2006.

Si le gustaría recibir copias electrónicas del periódico FOCUS on Field Epidemiology por favor llene la siguiente forma:

- NOMBRE: \_\_\_\_\_
- TÍTULO (S): \_\_\_\_\_
- AFILIACIÓN: \_\_\_\_\_
- CORREO ELECTRÓNICO: \_\_\_\_\_
- ¿Podemos contactar por correo electrónico a sus colegas?: Si es así, por favor incluya su correo electrónico a continuación  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Por favor enviar por fax a: (919) 919-843-5563

O por correo a: North Carolina Center for Public Health Preparedness  
The University of North Carolina at Chapel Hill  
Campus Box 8165  
Chapel Hill, NC 27599-8165

O en línea en: <http://www.sph.unc.edu/nccphp/focus/>

## PRÓXIMOS TEMAS

- Recolección de especímenes en investigaciones de brotes
- Diagnóstico de laboratorio: un repaso general
- Diagnóstico de laboratorio: técnicas moleculares

¡Estamos en Internet!

<http://www.sph.unc.edu/nccphp>