VOLUMEN 3, NÚMERO 5

FOCUS on Field Epidemiology

Enfoque en Epidemiología de Campo

Aspectos Básicos del Análisis de Datos: Variables y Distribución

CONTRIBUYENTES

Autores:

Kim Brunette, MPH

Amy Nelson, PhD, MPH

Grupo de trabajo FOCUS*

Críticos:

Grupo de trabajo FOCUS*

Gloria C. Mejia, DDS, MPH, PhD

(Versión en español)

Editoras de Producción :

Tara P. Rybka, MPH

Lorraine Alexander, DrPH

Rachel A. Wilfert, MD, MPH

Gloria C. Mejia, 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 estan nombrados en la última página de la publicación.**

UNC SCHOOL OF PUBLIC HEALTH

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.

¿Cómo puedes determinar si un derrame químico en una fábrica causó enfermedad a los trabajadores? ¿Cómo puedes saber qué alimento causó un brote de salmonella en tu comunidad?

En una investigación de campo, muchas veces quieres saber si una exposición determinada (por ejemplo un derrame químico) está asociado a cualquier enfermedad posible, o cuáles de las muchas posibles exposiciones están asociadas a una enfermedad en particular (por ejemplo, ¿cuál es la causa potencial de un brote de salmonella? Comienzas el proceso de responder estas preguntas eligiendo un diseño de estudio, desarrollando un cuestionario y recolectando datos. Todo esto fue discutido en ediciones anteriores de FOCUS, Una vez que estos pasos se han completado y has recolectado tus datos, ¿qué sigue?

A diferencia de la caracterización de los epidemiólogos en algunos programas de televisión, después de recolectar los datos no tienes simplemente un destello de inspiración y resuelves el brote; de hecho tienes que sentarte y analizar los datos! No es la parte más glamorosa del trabajo del epidemiólogo, pero cuando los datos llevan al origen de un brote, el análisis es definitivamente estimulante. Esta edición de FOCUS te guiará en los pasos básicos del análisis descriptivo de datos, incluyendo los tipos de variables, principios básicos de codificación y análisis de datos univariados simples.

Tipos de Variables

Antes de continuar con el análisis, tomemos un momento para discutir las variables. Esto puede parecer trivial para aquellos que tienen experiencia en análisis, pero las variables no son un tema trivial. Al igual que las personas, las variables vienen en distintos tamaños y formas. Sin embargo, la mayor parte de la epidemiología de campo se centra en variables continuas y categóricas.

Las variables **continuas** son siempre numéricas, y teóricamente pueden ser cualquier número, positivo o negativo (en realidad, esto depende de la variable). Ejemplos de variables continuas son la edad en años, el peso, presión arterial, temperaturas interiores y exteriores, concentraciones de contaminantes en el aire o agua, y otras mediciones.

Las variables **categóricas** contienen información que puede organizarse en categorías, de manera similar a como se organizaría la información en canastas. Cada información pertenece a una, y sólo una, canasta. Existen diversos tipos de variables categóricas, ordinales, nominales y dicotómicas o binarias.

Una variable **ordinal** es cualquier variable categórica con algún orden intrínseco o valor numérico. Por ejemplo, podríamos categorizar información acerca del nivel educacional de un grupo de personas en una variable denominada EDUCACION. Una persona puede no haber terminado la educación secundaria, otra podría haberse graduado de secundaria pero no

haber recibido educación superior, una tercera podrá haber tenido alguna educación superior o haber recibido algún entrenamiento posterior, y otra podrá haberse graduado de la universidad. Los niveles de educación de todos los miembros del grupo se acomodan claramente dentro de estas categorías, y las categorías tienen un orden intrínseco. Un egresado de la universidad tiene más educación que un graduado de la escuela secundaria, y un graduado de la escuela secundaria tiene más educación que alguien que no terminó la escuela. De este modo, a medida que las categorías van de 1 a 5, aumenta el nivel de educación. Otros ejemplos de variables ordinales son:

- aceptación (por ejemplo, total desacuerdo, desacuerdo, neutro, acuerdo, totalmente de acuerdo)
- escalafón (por ejemplo excelente, bueno, aceptable, pobre)
- frecuencia (por ejemplo siempre, a menudo, a veces, nunca)
- o cualquier otra escala (por ejemplo, "en una escala del 1 al 5 ¿cuánto te gustan los cacahuates?")

Una variable **nominal** es una variable categórica sin orden intrínseco. Por ejemplo, digamos que tenemos una variable llamada LUGAR DE RESIDENCIA que caracteriza la parte del país en que vive una persona: el noreste, sur, medio-oeste, o noroeste. Las categorías de esta variable no tienen valor numérico ni orden. La residencia en el noroeste no tiene valor cuantitativo en comparación con el noreste. Otros ejemplos de variables nominales incluyen el sexo (masculino, femenino), nacionalidad (Estadounidense, Mexicano, Francés), raza-etnia (Afroamericano, hispano, blanco, asiático americano) o mascota favorita (perro, gato, pez, culebra).

Una **variable dicotómica** o **binaria** es una variable categórica que tiene sólo dos niveles o categorías. Muchas variables dicotómicas representan la respuesta a una pregunta cerrada (de respuesta sí o no). Por ejemplo, "¿asistió usted al picnic de la iglesia el 24 de mayo?" o "¿Consumió ensalada de papas en el picnic?" Una variable no requiere ser variable si/no para ser dicotómica, sólo necesita tener dos categorías, como por ejemplo, sexo (masculino/femenino).

Codificación

Una vez que has recopilado tus cuestionarios u otra información debes elegir los códigos para ingresarlos a una base de datos. La codificación es el proceso de traducir la información recolectada de los cuestionarios u otras investigaciones a algo que pueda ser analizado, por lo general utilizando un programa computacional. La codificación incluye el asignar un valor a la información entregada en el cuestionario, y muchas veces a ese valor se le asigna un nombre. Por ejemplo, si tienes la pregunta "Sexo?", podrías tener respuestas tales como "masculino", "femenino" o "M", "F", etc. La codificación evitará tales inconsistencias.

Un sistema común de codificación (codificación y nombre) para variables dicotómicas es el siguiente:

0 = No 1 = Si,

donde el número 1 es el valor asignado, y SI es la etiqueta o significado de dicho valor.

A algunos les gusta utilizar un sistema de 1 y 2, donde

1 = No 2 = Si.

Esto apunta a un aspecto importante en la codificación. Cuando asignas un valor a un pedazo de información, también debes dejar en claro lo que significa ese valor. En el primer ejemplo anterior, 1= sí, pero en el segundo ejemplo, 1= No. Cualquiera de los dos está bien, siempre y cuando quede claro cómo se ha codificado la información. Puedes aclarar esto creando un **diccionario de datos** como un archivo separado que acompañe la base de datos.

De manera similar, podríamos codificar la variable dicotómica para sexo:

0 = Femenino 1 = Masculino

Las variables dicotómicas también pueden ser **variables falseadas o ficticias (dummy, en inglés)**. Una variable "ficticia" es cualquier variable que se codifica para que tenga dos niveles, como las variables si/no y las variables femenino/masculino del ejemplo anterior. También pueden ser usadas para representar variables más complicadas. Esto es especialmente útil cuando tienes muchos valores que son más significativos cuando se analizan en términos de una respuesta sí o no.

Por ejemplo, puedes haber recopilado datos sobre el número de cigarrillos fumados por semana, con 75 respuestas que van de cero cigarrillos a 3 paquetes por semana, pero puedes volver a codificar esta información como variable ficticia: 1= fuma, 0=no fuma. También podrías hacer esto para la educación (1=cualquier educación posterior a la secundaria, 0= sin educación posterior a la secundaria), consumo de alimentos (1= comió el producto durante el período de tiempo, 0=no comió el producto) y muchas otras variables. Este tipo de codificación es útil en las etapas posteriores del análisis.

Muchos paquetes de software para análisis te permiten asignar un nombre a los valores de las variables. Luego el computador automáticamente nombra los 0 como masculinos y los 1 como femeninos, lo que facilita tu vida cuando observes el resultado, tal como lo muestra

el ejemplo siguiente:

Sin nombre:	Variable SEXO	Frecuencia	Porcentaje
	0	21	60%
	1	14	40%
Con nombre:	Variable SEXO	Frecuencia	Porcentaje
	Masculino	21	60%
	Femenino	14	40%

El proceso de codificación es similar con otras variables categóricas. Para la variable EDUCACION mencionada anteriormente, podríamos codificarla de la siguiente manera:

0 = No se graduó de la escuela secundaria
1 = Se graduó de la escuela secundaria
2 = Algún estudio universitario o superior
3 = Egresado de la universidad.

Observa que para esta variable ordinal categórica debemos ser consistentes con la enumeración, porque el valor del código asignado tiene significado. Mientras más alto el código, más educado es el que responde. También podríamos haber codificado esta variable en el orden inverso, de modo que 0= egresado de la universidad y 3= no se graduó de la escuela secundaria. En este caso, mientras más alto es el código, menor es el grado de educación del que responde. Cualquiera de los dos modos está bien, siempre y cuando recordemos el código al interpreta el análisis.

El siguiente es un ejemplo de mala codificación:

0 = Algún estudio universitario o superior
1 = Egresado de la escuela secundaria
2 = Egresado de la universidad
3 = no se graduó de la escuela secundaria.

Los datos que estamos tratando de codificar tienen un orden inherente, pero la codificación en este ejemplo no sigue ese orden. Esta *no* es una codificación apropiada para una variable ordinal categórica.

Para una variable nominal categórica, sin embargo, el orden no tiene ningún efecto. Aunque codifiquemos cada variable con un número, el número no representa un valor numérico. Por ejemplo, usando la variable LUGAR DE RESIDENCIA, mencionada anteriormente,

1 = Noreste
2 = Sur
3 = Noroeste
4 = Medio-oeste
5 = Suroeste.

No importa el orden que utilicemos para estas categorías. El medio-oeste pudo haberse codificado como 4, 2, o 5 porque no existe un valor ordenado asociado a cada respuesta.

La codificación de variables continuas es unidireccional. Si alguien dice que su edad en años es 37 años, la ingresas como 37 a tu base de datos. Pero ¿qué pasaría si usases categorías de edad en lugar de los datos en años que recopilaste?

Es común crear categorías a partir de variables continuas, y puede hacerse fácilmente usando un software de análisis. Con un paquete de software, puedes desglosar una variable continua como la edad en categorías, creando una variable ordinal categórica como la siguiente:

EDADCAT
1 = 0–9 años de edad
2 = 10–19 años de edad
3 = 20–39 años de edad
4 = 40–59 años de edad
5 = 60 años o más.

Es posible que también necesites codificar respuestas de preguntas de oraciones para completar y preguntas abiertas. Con una pregunta abierta como por ejemplo, ¿por qué decidió no consultar al doctor acerca de esta enfermedad?, los entrevistados responderán todos de manera distinta. También podrías dar a los entrevistados opciones de respuesta para una pregunta específica y ofrecer una opción "otro (especifique)", en la cual los entrevistados podrán escribir la respuesta que deseen. Este tipo de preguntas abiertas pueden ser difíciles de analizar. Una manera de analizar la información es agrupar las respuestas con temas similares. Para la pregunta anterior, quienes respondieron que "no se sentían lo suficientemente enfermos para ir al doctor", "los síntomas desaparecieron", y "la enfermedad no duró mucho tiempo" podrían agruparse como "la enfermedad no era grave."

También necesitarás codificar respuestas del tipo "no sé". Por lo general, la respuesta "no sé" se codifica como 9.

Dato de ayuda para la Codificación:

Pese a que no codificas hasta que se hayan recopilado los datos, debes pensar cómo vas a codificar cuando estés diseñando el cuestionario, antes de que recopiles datos. Esto te ayudará a recopilar los datos en un formato que puedas usar.

Limpieza de los Datos

Uno de los primeros pasos en el análisis de datos es mirar la información obtenida y "limpiarla" de cualquier error evidente, debido al ingreso incorrecto de datos.

Si existen valores extremos (números demasiado altos o demasiado bajos), ¿son correctos esos números? Un

valor de 110 en edad podría ser un error para quien en realidad tiene 10 o 11 (¡o 101!)

¿Se ingresó un valor que no existe para una variable? Por ejemplo, si 1=masculino y 0= femenino, si se ingresó "2", es claramente un error.

Si existen valores faltantes, ¿acaso la persona no respondió, o accidentalmente no se ingresó a la base de datos?

Algunos software de análisis permiten al usuario establecer límites definidos al ingresar datos. Esto evita que una persona ingrese un 2 cuando los valores aceptables son sólo 1 y 0. Los límites también pueden establecerse para variables continuas y nominales, por ejemplo permitiendo sólo 3 dígitos para la edad, o limitando la cantidad de palabras que se ingresan. También puedes asignar tipos de campos para la mayoría de los tipos de variables basados en el tipo de datos que el campo debiera contener (por ejemplo formatear las fechas como mm/dd/aaaa o valores numéricos o textos específicos). De manera similar, algunos protocolos de estudio permiten que se ingresen datos de otras fuentes. Por ejemplo, si una persona no respondió una pregunta sobre edad, esa información puede estar disponible a partir de un registro médico que se esté usando en el estudio.

A modo de verificación del ingreso de datos, algunos sistemas te permiten ingresar datos dos veces y luego compararlos para ver si existen discrepancias. Este proceso se llama "doble ingreso."

El análisis univariado de datos, que discutiremos a continuación es también una forma útil de revisar la calidad de los datos, incluyendo la revisión de los valores extremos.

Análisis univariado de datos

El análisis univariado de datos es importante por muchas razones. Al mirar a cada variable individualmente, aprendemos mucho acerca de la información recopilada. De manera similar, el análisis univariado es un buen método para verificar la calidad de los datos. Siempre deben investigarse las inconsistencias o los resultados inesperados, usando los datos originales como punto de referencia.

Las frecuencias simples te dirán si la mayoría de los casos son jóvenes, o si la mayoría de los casos son mujeres, o si muchos de los participantes del estudio comparten alguna otra característica que te interese. Primero, examina la distribución **univariada** (una variable). Los gráficos y tablas te pueden dar una buena idea de cómo se ven tus variables y pueden ayudar también a encontrar algunos de los problemas de limpieza de datos mencionados anteriormente.

Variables continuas

Una mirada inicial a las variables continuas te puede entregar varias piezas importantes de información:

- ¿Hay datos para todos los sujetos, o faltan valores?
- ¿Se encuentra la mayoría de los valores agrupados, o hay mucha variación? (figuras 1a y 1b)
- ¿Existen valores extremos?
- ¿Tienen sentido los valores mínimos y máximos, o podría haber errores en la codificación?

También podemos realizar un análisis univariado de variables continuas para obtener información valiosa. Entre las estadísticas comúnmente usadas se encuentran las siguientes:

Media- promedio de todos los valores de esta variable en la base de datos (fig.2).

Mediana- el medio de la distribución, el número en el cual la mitad de los valores están por encima y la otra mitad está por debajo (figura 2)

Moda- el valor que más se repite (figura 2)

Rango de valores- desde el valor mínimo hasta el máximo (figura 2)

Desviación estándar- una medida de qué tan confiables son los datos. En la figura 1a, la desviación estándar es de 20.4; en la figura 1b es de 7.6. Una desviación estándar grande en comparación a los valores de la variable indica que los datos se encuentran ampliamente distribuidos. Las desviaciones estándar son fáciles de calcular con software de análisis o pueden ser calculadas a mano.

Distribución- muestra si la mayoría de los valores se en-

Figura 1a: ejemplo de valores de edad ampliamente distribuidos

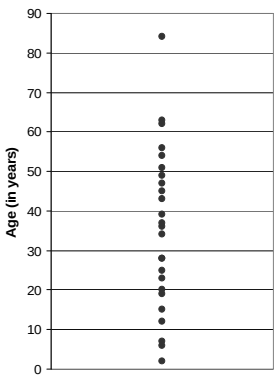
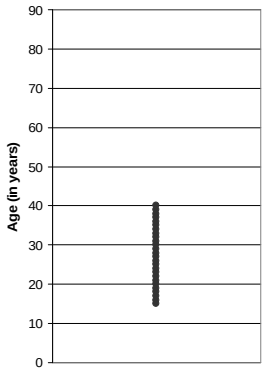


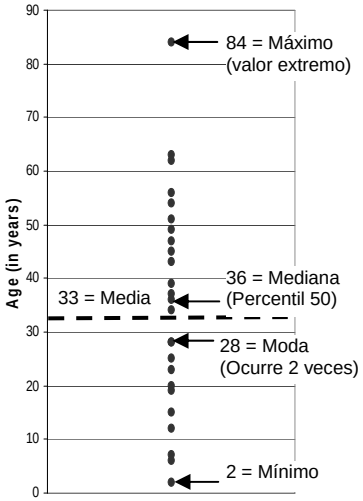
Figura 1b: ejemplo de valores de edad poco distribuidos.



cuentran en la parte baja del rango, o en la parte alta, o agrupados en el medio (figura 3)

Percentil- el porcentaje de la distribución que es igual a o menor que un determinado valor (figura 3). En la figura 3a, el percentil 25 ocurre en los 4 años pues el 25% del total de quienes responden tienen 4 o menos años.

Figura 2: estadísticas que describen una distribución de variable continua.



Datos categóricos

Figura 3a: "distribución de campana" para la variable EDAD

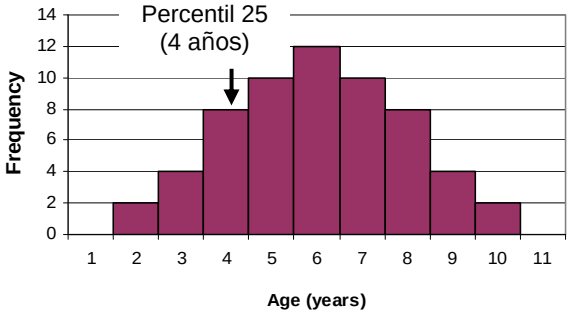
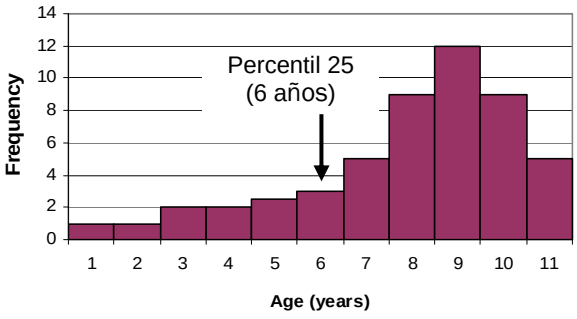


Figura 3b: distribución asimétrica para la variable EDAD



La distribución de datos categóricos también debe ser examinada antes de llevar a cabo análisis más profundos.

Por ejemplo, observemos la variable LUGAR DE RESIDENCIA, de la figura 4a. Vemos que más gente vive en el sur que en ningún otro lugar, mientras que menos gente vive en el suroeste que en cualquier otro lugar. El área geográfica de residencia puede ser importante en algunos estudios porque los estilos de vida y las influencias culturales tienden a ser distintas en diferentes partes del país. Una forma de poner estos resultados en contexto es comparar esta distribución a la distribución "esperada" de residencia. Si seleccionamos a nuestros participantes del estudio al azar a partir de la población de los Estados Unidos, podríamos esperar tener la misma distribución de personas en cada región, suponiendo que todas tienen el mismo tamaño de población, como lo muestra la figura 4b.

Otra forma de observar estos datos categóricos es

Figura 4a: número de personas que responden el cuestionario y que residen en 5 regiones de los Estados Unidos.

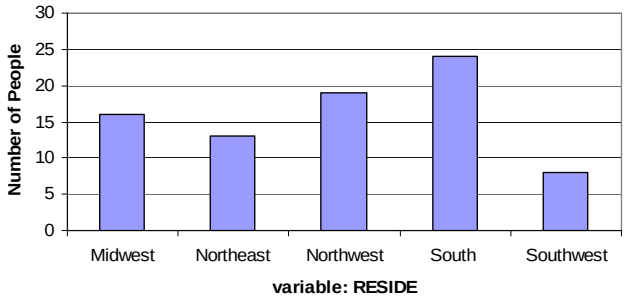
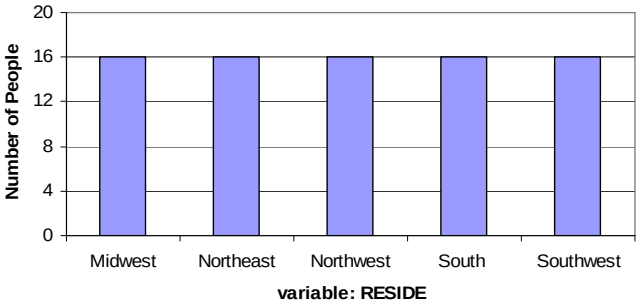


Figura 4b: número esperado de personas que residen en 5 regiones de los Estados Unidos.



haciendo listas de los datos en tablas. Por ejemplo, la tabla 1 nos da la misma información que la figura 4a, pero en un formato distinto. Muchas tablas muestran la frecuencia (número) así como también el porcentaje del total de los individuos en la tabla. Cuando los gráficos o tablas se usan para informar y

Tabla 1: Numero de personas que responden el cuestionario y que residen en 5 regiones de los Estados Unidos

	Frequencia	Porcentaje
Medio-oeste	16	20%
Noreste	13	16%
Noroeste	19	24%
Sur	24	30%
Suroeste	8	10%
Total	80	100%

describir datos, deben incluir toda la información relevante en los títulos, encabezados y nombres, tales como población que se estudia, fechas del estudio, y número de participantes. En el ejemplo anterior, un epidemiólogo querría comparar el número observado de personas en cada región (figura 4a) frente al número esperado de personas en cada región (figura 4b). Al comparar los datos observados con los datos esperados, los epidemiólogos pueden ver si hay algo extraordinario asociado a los datos observados. Podemos realizar el proceso con nuestra variable ordinal categórica EDUCACION. La figura 5a nos muestra la distribución observada de niveles de educación en una población determinada de adultos. Pese a que el gráfico está marcado con niveles de educación, esos niveles representan los números usados en la codificación (0 para menos que educación secundaria, 1 para egresados de educación secundaria, y así sucesivamente.)

Este gráfico contiene información descriptiva útil acerca de la población del estudio. También podemos compararla con la distribución esperada de educación entre nuestros participantes en el estudio. La información obtenida de la Oficina del Censo de los Estados Unidos respecto al nivel educacional de la población de los Estados Unidos de 20 años o más aparece en la figura 5b. (1). Esta es la distribución esperada de nivel educacional para la población del país. Al mirar los gráficos y comparar las categorías, vemos que la población de nuestro estudio parece ser más educada de lo que esperábamos.

Figura 5a. Datos observados respecto a nivel de educación a partir de un cuestionario hipotético.

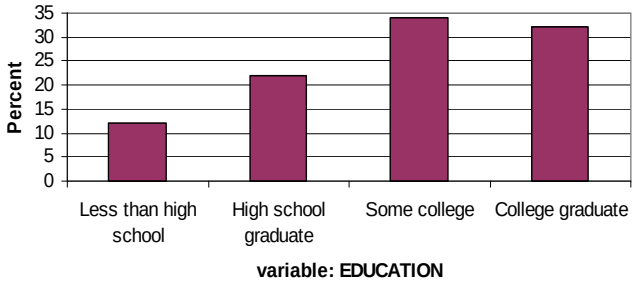
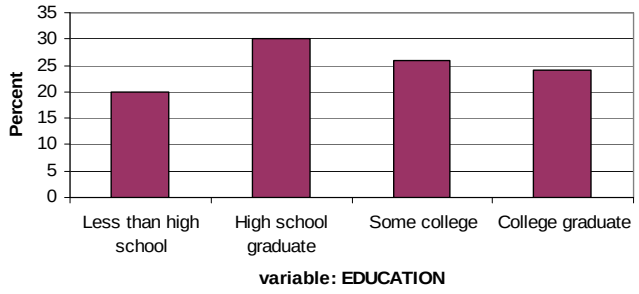


Figura 5b. Datos de nivel educacional de la población de los Estados Unidos de 20 o más años, Oficina del Censo de los Estados Unidos.



¿Son los datos observados tan distintos de los datos esperados? Este es el tipo de preguntas que un epidemiólogo querría explorar en profundidad. Una forma de comparar los datos categóricos observados con los datos categóricos esperados es usando una prueba estadística como el ji-cuadrado. Las próximas dos ediciones de FOCUS discutirán los ji-cuadrado y otros tipos de análisis de datos más extensos.

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

REFERENCIAS:

1. US Census Bureau. Educational Attainment in the United States: 2003---Detailed Tables for Current Population Report, P20-550 (All Races). Available at: http://www.census.gov/population/www/socdemo/education/cps2003.html. Accessed December 11, 2006.

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

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

- **Análisis de datos: pruebas estadísticas simples**
- **Análisis avanzado de datos: métodos para el control de confusión**
- Recolección de especimenes en investigaciones de brotes

¡Estamos en Internet!
http://www.sph.unc.edu/nccphp