Categorización de variables en el análisis estadístico de datos: consecuencias sobre la interpretación de resultados

Francisco Cumsille[1] *y Shrikant I. Bangdiwala*[2]

RESUMEN *Es bastante frecuente que en estudios epidemiológicos, durante el proceso de análisis de datos, una o más variables continuas sean cambiadas de escala. El objetivo de este trabajo consistió en evaluar las consecuencias de la categorización de variables en el análisis de datos. Se estudian tres situaciones bajo diferentes escenarios de análisis estadístico en modelos de regresión. Los resultados muestran que la dicotomización de variables continuas puede modificar sustancialmente las relaciones entre variables dependientes e independientes. Así por ejemplo, en estudios epidemiológicos en los que se pretende evaluar el efecto de una exposición sobre una respuesta, la magnitud o la dirección de dicho efecto pueden estar sesgadas como consecuencia de la dicotomización de alguna variable. Se recomienda, por lo tanto, evitar, en la medida de lo posible, la categorización de variables en el análisis.*

El uso de procedimientos estadísticos en el análisis de los datos de las investigaciones en salud es cada vez más frecuente, debido a que constituyen la base de las decisiones que de esos estudios se puedan derivar. Ya sea en estudios por observación o en ensayos clínicos, la utilización de métodos estadísticos es un potente aporte, desde la planificación del estudio hasta su análisis e interpretación. Aceptado esto, es válido preguntarse si el procesamiento de los datos y los procedimientos estadísticos utilizados en un estudio son o no los adecuados. Una de las prácticas habituales en el procesamiento de los datos es la categorización de algunas variables continuas antes del análisis de los datos. Por ejemplo, es bastante frecuente que variables continuas como la tensión arterial o el índice de masa corporal se transformen en variables dicotómicas: presencia o ausencia de hipertensión arterial o de obesidad.

Es importante evaluar las razones por las cuales dichas variables son categorizadas y las consecuencias de esa categorización. Desde una perspectiva estadística, existen abundantes publicaciones acerca de la categorización de variables continuas, unas justificando su uso y otras poniéndolo en duda. Una razón estadística que puede justificar la categorización está relacionada con los errores de medición. A veces es bastante difícil obtener mediciones exactas de una variable continua, o simplemente se sospecha que pueden contener errores. En este aspecto, Flegal et al. (1) estudiaron el efecto de los errores de medición sobre los riesgos relativos (RR), y Reade-Christopher y Kupper (2) propusieron un modelo para evaluar el efecto de la mala clasificación de la exposición en estudios epidemiológicos de seguimiento. En estudios de casos y controles, Fung y Howe (3) estudiaron el efecto de la mala clasificación de la exposición y de las covariables del estudio sobre las estimaciones de las medidas de asociación y sobre la potencia de las pruebas estadísticas.

Otro argumento estadístico encontrado en la literatura para justificar la categorización de variables continuas está relacionado con los modelos estadísticos aplicados en algunos casos. Algunos investigadores prefieren dicotomizar la variable de respuesta o la variable de exposición, o ambas, porque las variables dicotomizadas representan mejor el estado de enfermo/no enfermo y pueden facilitar la interpretación de los coeficientes de un modelo estadístico, o porque el recorrido de una variable continua tiene distintos significados clínicos (4). Otra

[1] División de Bioestadística, Escuela de Salud Pública y Centro de Epidemiología Clínica, Universidad de Chile. Dirección: Independencia 939, Santiago, Chile.
[2] Departamento de Bioestadística, Universidad de Carolina del Norte, Chapel Hill.

importante razón tiene que ver con los supuestos de un modelo estadístico. Altman et al. (5), señalan que la categorización de variables continuas permite a los investigadores evitar los fuertes supuestos requeridos por ciertos modelos acerca de la relación entre las variables y la medición del riesgo, pero a expensas de perder información. En los modelos de regresión logística, por ejemplo, uno presupone la existencia de una relación lineal entre las variables independientes y el logaritmo de la razón de momios *(odds)*, mientras que si una variable independiente continua es dicotomizada, dicho supuesto se cumple automáticamente. Kahn y Sempos (6) estudiaron la razón de productos cruzados (RPC; *odds ratio*) asociada con la presencia o ausencia de enfermedad cardiovascular y diferentes variables continuas, y ajustaron los modelos usando las variables originales y también dos tipos de categorización. Las estimaciones de algunas RPC fueron bastante similares y otras bastante diferentes. Debido a que el supuesto de linealidad no es un requisito al usar variables categóricas, esos autores prefieren usar métodos que requieran el mínimo de supuestos.

Por otra parte, está la posición de quienes advierten de las consecuencias negativas de la categorización de variables continuas. La primera es la pérdida de información y de potencia estadística de los análisis (7). Por ejemplo, con respecto a la concordancia entre observadores, Donner y Eliasziw (8) estudiaron las consecuencias de la dicotomización de variables continuas con respecto a la pérdida de potencia estadística y concluyeron que en estudios de confiabilidad tal pérdida puede ser considerable.

Uno de los aspectos más controvertidos de la categorización de variables es la elección de los puntos de corte utilizados. En muchas ocasiones, la elección del valor de dicotomización es arbitraria. Amenta et al. (9) compararon diferentes métodos estadísticos para evaluar una prueba de efectividad y discriminación diagnóstica después de dicotomizar pruebas cuantitativas y concluyeron que la dicotomización, usando algún valor arbitrario, es inadecuada en la evaluación de la efectividad. Desde otra perspectiva, en estudios de casos y controles, Wartenberg y Northridge (10) presentaron un nuevo enfoque para definir el "mejor" punto de corte. Básicamente la idea es computar la asociación entre la respuesta de interés y la exposición para cada posible dicotomización de la exposición, y entonces escoger el valor que maximiza el riesgo. Aunque esta estrategia parece atractiva, Altman (11) señaló que el valor *P* obtenido por múltiples pruebas estadísticas no es válido, y que el error de tipo I asociado con dicho enfoque es muy alto. Ragland (4) consideró una situación bastante simple, en la cual el interés era evaluar la asociación entre la edad y la tensión arterial sistólica (TS). Usando diferentes puntos de corte para los valores de TS, la magnitud de la asociación (RPC y RR) y la potencia de las pruebas estadísticas cambian a medida que lo hace el punto de corte, de tal forma que la decisión final sobre la asociación entre las variables se vuelve arbitraria.

En un ejemplo hipotético, Maxwell y Delaney (12) exploraron las consecuencias de la dicotomización de dos variables independientes en un modelo de regresión y compararon los resultados obtenidos en el análisis de un diseño factorial 2×2 (usando las medianas como los puntos de corte para las dos variables independientes continuas) con los resultados de un modelo de regresión lineal múltiple. Los resultados revelaron que la dicotomización de una sola variable continua casi siempre produce un sesgo conservador, es decir, una subestimación de la asociación. En cambio, cuando se dicotomizan dos o más variables independientes, el efecto puede ser el contrario, es decir, una sobreestimación de la asociación.

Cumsille y Bangdiwala (13) estudiaron el efecto de dicotomizar una variable independiente continua en modelos de regresión logística, usando diferentes distribuciones de la misma. Las diferencias entre las medidas de asociación obtenidas con y sin dicotomización, y su significación estadística, pueden ser bastante considerables en algunos casos. En modelos de regresión lineal múltiple con una variable continua de exposición y una variable continua de control, Cumsille (14) ha demostrado que la dicotomización de la variable de control implica el cambio de un modelo de regresión lineal a un modelo no lineal. Así, si una variable independiente es dicotomizada, utilizar un modelo de regresión lineal tras la dicotomización implica simplemente una mala especificación del modelo, de tal forma que las conclusiones derivadas de él son dudosas.

Con estos antecedentes, podemos concluir que la dicotomización de variables para su posterior análisis estadístico no es una cuestión menor. Los métodos estadísticos son utilizados para derivar conclusiones a partir de ellos, pero, como hemos visto, si hay un procesamiento inadecuado de los datos, las conclusiones serán dudosas y de poco valor científico. El objetivo de este estudio consistió en ilustrar la afirmación precedente con tres situaciones en las que se dicotomizaron variables continuas, y describir cómo las conclusiones de un estudio se pueden ver afectadas por este proceder.

MATERIALES Y MÉTODOS

De las tres situaciones analizadas, dos provienen de resultados publicados en la literatura (15, 16), y la tercera corresponde a los resultados de una investigación propia. En cada caso se describen el escenario del estudio, sus conclusiones y las consecuencias de la dicotomización sobre las conclusiones.

RESULTADOS

Situación 1

Chang et al. (15) realizaron un estudio con el propósito de evaluar la asociación entre una respuesta binaria (enfermedad coronaria) y un conjunto de covariables: hábito tabáquico, TS,

CUADRO 1. Algunos conjuntos de riesgo y sus correspondientes riesgos relativos (RR)

Conjunto de riesgo	Peso relativo	Colesterol	TS	Hábito tabáquico	Edad	RR
1	0	0	0	0	0	1,00
2	1	0	0	0	0	1,56
3	0	1	0	0	0	1,55
4	0	0	1	0	0	1,64
5	0	0	0	1	0	2,06
6	0	0	0	0	1	2,76
7	1	1	0	0	0	2,42
8	1	0	1	0	0	2,56
…	…	…	…	…	…	…
31	0	1	1	1	1	6,79
32	1	1	1	1	1	10,60

TS: tensión sistólica

CUADRO 2. Perfiles hipotéticos de cuatro personas

Persona	Peso relativo	Colesterol total	TS	Cigarrillos/día	Edad
1	80	120	110	0	20
2	115	240	135	18	54
3	125	255	145	21	56
4	180	450	220	40	70

TS: tensión sistólica.

colesterol total, peso relativo y edad. Las covariables fueron dicotomizadas, asignándole el valor cero al consumo de menos de un paquete de cigarrillos por día, a la TS < 140 mm Hg, al peso relativo < 120%, al colesterol total < 250 mg% y a la edad < 55 años, y el valor 1 a la alternativa en cada caso.

Usando las cinco variables dicotómicas, es posible clasificar a los pacientes en 32 conjuntos diferentes de riesgo. Los autores analizaron los datos usando, entre otros, un modelo de regresión de Cox dependiente del tiempo. Algunos de los 32 conjuntos y sus correspondientes RR, obtenidos con el modelo de Cox, se muestran en el cuadro 1, del que se desprende que si el riesgo basal para un individuo que no tiene ningún factor de riesgo es 1,00, entonces para quien tiene alterado el peso relativo, el riesgo de contraer enfermedad coronaria es 1,56 veces mayor, y para quien presenta todos los factores de riesgo, 10,6 veces mayor.

Supongamos ahora que deseamos clasificar a cuatro personas cuyos perfiles hipotéticos para los cinco factores de riesgo se muestran en el cuadro 2. De acuerdo con la regla de clasificación y los resultados encontrados por los autores, en el cuadro 3 se presentan los valores de las variables dicotomizadas y el correspondiente RR en cada caso. Las primeras dos personas están en la categoría basal (RR = 1) de acuerdo con la dicotomización, y las otras dos en la peor situación (RR = 10,6).

De los resultados anteriores surgen algunas interrogantes: ¿es el riesgo de enfermedad coronaria el mismo en las personas 1 y 2?, ¿es el riesgo de enfermedad coronaria el mismo en las personas 3 y 4?, ¿es el RR de la persona 3 respecto de la persona 2, el mismo que el RR de la persona 4 respecto de la persona 1?, ¿son los valores originales equivalentes entre las personas 1 y 2?.

Al observar los valores originales de las cinco variables, se podría considerar que, desde una perspectiva clínica, las diferencias entre las personas 2 y 3 son poco relevantes; sin embargo, de acuerdo con la clasificación, aparecen en situaciones extremas. Probablemente sean mucho más relevantes las diferencias observadas entre las personas 1 y 2 o entre las personas 3 y 4, que, sin embargo, están clasificadas en los mismos grupos. Aunque es cierto que el procedimiento seguido por los autores parece muy atractivo como guía metodológica para evaluar el mayor riesgo de un perfil frente a otro, parece inadecuado clasificar individuos clínicamente homogéneos en clases distintas (personas 2 y 3), o clasificar en la misma clase a personas clínicamente diferentes (personas 3 y 4). Este es uno de los problemas que genera la dicotomización de variables continuas en el análisis estadístico de datos, y que nos debe llevar a cuestionar su práctica.

Situación 2

En un texto clásico sobre regresión logística, Hosmer y Lemeshow (16) presentan varios ejemplos para ilus-

CUADRO 3. Valores de las variables después de la dicotomización de los cinco factores, y riesgo relativo (RR)

Persona	Peso relativo	Colesterol total	TS	Hábito tabáquico	Edad	RR
1	0	0	0	0	0	1,00
2	0	0	0	0	0	1,00
3	1	1	1	1	1	10,60
4	1	1	1	1	1	10,60

TS: tensión sistólica.

trar la teoría de dichos modelos. En uno de ellos el objetivo es identificar factores de riesgo asociados con el bajo peso al nacer (BPN). Los potenciales factores de riesgo considerados son: edad de la madre (años), peso de la madre en el último período menstrual (libras), raza (blanca, negra u otra), hábito de fumar durante el embarazo (sí, no), antecedentes de parto prematuro (ninguno, uno, etc.), antecedentes de hipertensión arterial (sí, no), presencia de irritabilidad uterina (sí, no) y número de controles médicos (visitas) durante el primer trimestre del embarazo (ninguna, una, etc.). La variable de estudio original fue el peso al nacer, en gramos, que se transformó en una variable binaria (presencia o ausencia de BPN) de acuerdo con el siguiente criterio: BPN ausente si el peso al nacer fue > 2 500 g, y BPN presente si fue ≤ 2 500 g. En el cuadro 4 se muestran los resultados usando un modelo de regresión logística con el BPN como variable dependiente y la edad, peso, raza y visitas como variables independientes. La categoría "raza blanca" se tomó como referencia con respecto a las otras categorías de dicha variable.

Se observa que la asociación entre BPN y raza no es clara. La diferencia de BPN entre la raza blanca y la negra es ligeramente significativa, pero no lo es entre la raza blanca y otras razas. Supongamos ahora que en lugar de la variable BPN se usa la variable original, peso al nacer, en gramos, y que, como esta es continua, se usa un modelo de regresión lineal múltiple en lugar de uno de regresión logística. Usando las mismas variables independientes, los resultados se presentan en el cuadro 5. Como se puede observar, al usar la variable dependiente en su estado original, la fuerza de la asociación de la respuesta con la raza es mucho más clara, y difiere de la conclusión obtenida mediante regresión logística usando la variable dicotomizada.

CUADRO 5. Resultados del ajuste del modelo de regresión lineal múltiple

Variable	Estimación	Error estándar	*P*
Edad	0,779	10,308	0,9396
Peso	4,580	1,799	0,0117
Raza (negra)	–448,634	161,706	0,0061
Raza (otra)	–240,593	115,574	0,0388
Visitas	10,989	50,152	0,8268

En la discusión anterior, se usó el valor de 2 500 g para dicotomizar la variable. Sin embargo, se podría haber utilizado otro. ¿Qué ocurre con la conclusión si el punto de corte hubiera sido otro? En el cuadro 6 se presentan los resultados de aplicar el modelo de regresión logística usando diferentes puntos de corte, en términos de RPC y valores *P* para el mismo conjunto de variables independientes. De estos resultados podemos observar varios hechos interesantes. En primer lugar, para la variable peso de la madre en el último período menstrual (peso), observamos que si bien las RPC son similares para los diferentes punto de corte del peso al nacer, sus significaciones son bastante diferentes. Si el punto de corte fuera 2 200 g, no se observaría una asociación significativa entre el peso de la madre y el BPN, pero, en cambio, si el punto de corte fuera superior a 2 200 g, la conclusión sería la opuesta. Algo similar ocurre al comparar las madres de raza negra y de raza blanca. Para la variable número de visitas, si bien la asociación con el BPN no es significativa con ningún punto de corte, las RPC cambian de sentido. Hasta los 2 500 g, la RPC es superior a 1, sugiriendo ser un factor de riesgo, y a partir de allí es inferior a 1, sugiriendo ser un factor de protección.

Estos resultados nos llevan a cuestionar la validez de la dicotomización de la variable de respuesta, ya que la fuerza de las asociaciones y su dirección dependen del punto de corte elegido y, por lo tanto, las conclusiones que de allí se derivan parecen ser bastante arbitrarias.

CUADRO 4. Resultados del ajuste del modelo de regresión logística

Variable	Estimación	Error estándar	*P*
Edad	–0,0238	0,0337	0,4800
Peso	–0,0142	0,0065	0,0298
Raza (negra)	1,0039	0,4979	0,0438
Raza (otra)	0,4331	0,3622	0,2318
Visitas	–0,0493	0,1672	0,7681

Situación 3

El tercer ejemplo se refiere a un estudio basado en simulaciones (14) cuyo objetivo principal consistió en evaluar las consecuencias de la dicotomización de una variable de control continua sobre la medida de asociación entre una respuesta continua y una exposición también continua, basado en un modelo de regresión lineal múltiple. Las variables utilizadas para la simulación fueron la TS como variable dependiente, el colesterol como variable de exposición y la edad como variable de control. Las medias (± desviaciones estándares) usadas en las simulaciones fueron: 150 (± 30) mm Hg para la TS, 200 (± 40) mg/dL para el colesterol y 40 (± 10) años para la edad.

Si bien es cierto que la dicotomización de la variable edad (control) im-

plica que el modelo ya no es lineal al utilizar la variable dicotomizada (14), en la práctica lo que se hace es precisamente eso, ajustar un modelo de regresión lineal con la variable de exposición continua y la variable de control transformada, como variables independientes. La simulación está basada en conocer *a priori* la magnitud de la asociación entre el colesterol y la TS (dada por el coeficiente de la variable de exposición en el modelo). Después de generar los datos de las tres variables en un proceso de simulación, se procedió a dicotomizar la edad usando 5 puntos de corte diferentes: deciles 1, 3, 5, 7 y 9. Además, se procedió a controlar la asociación de la variable de control (edad) con la variable de exposición (colesterol), y también con la variable de respuesta (TS), a través de los correspondientes coeficientes de correlación: 0,0, 0,1, 0,4 y 0,7. Asimismo, se controló la asociación entre la TS y el colesterol, mediante coeficientes de correlación de 0,1, 0,4 y 0,7. Por otra parte se consideraron dos tamaños muestrales: 100 y 1 000. Combinando todos los aspectos anteriores, se generaron 96 escenarios distintos. En cada uno de los casos se conocía anticipadamente el grado de asociación entre las variables de respuesta (TS) y exposición (colesterol). Para cada uno de esos escenarios se procedió a simular datos con 500 réplicas en cada caso. Una de las formas de evaluar las consecuencias de la dicotomización fue usar la diferencia relativa (parámetro del coeficiente de regresión de la variable de exposición menos la estimación del parámetro, dividido por el parámetro) media porcentual de las 500 réplicas. En el cuadro 7 se presentan los resultados para algunos escenarios, fijando la correlación entre TS y colesterol en 0,4. Observamos que cuando la variable de control (edad) no está correlacionada con la de respuesta (TS) ni con la de exposición (colesterol), la dicotomización de la edad en el decil 1 implica que la diferencia relativa media de las 500 réplicas es de 1,86%, y para el decil 5 de 1,81%; es decir, hay poco impacto en este caso. Sin embargo, esta situación va cambiando a medida que la variable de control (edad) aumenta su correlación con las otras dos variables, llegándose a una diferencia relativa media de 306,2% cuando la correlación de la edad con el colesterol y la TS es de 0,70, y la edad ha sido dicotomizada en el decil 1. Para la mediana (decil 5), la diferencia relativa media es de 222,6% en este caso. Es importante destacar que, en general, las diferencias son mayores cuando la dicotomización se realiza en el decil 1, comparado con el decil 5.

En el cuadro 8 se presentan las medias de las diferencias relativas combinando los coeficientes de correlación

CUADRO 6. Resultados del modelo de regresión logística para diferentes puntos de corte

Variable		Puntos de corte del peso al nacer (en gramos)						
		2 000	2 200	2 400	2 500	2 600	2 800	3 000
Edad	RPC	0,929	0,962	0,989	1,024	1,026	1,043	0,991
	P	0,123	0,339	0,767	0,480	0,439	0,196	0,767
Peso	RPC	1,002	1,007	1,015	1,014	1,013	1,020	1,021
	P	0,853	0,359	0,047	0,030	0,038	0,002	0,001
Raza	RPC	0,512	0,506	0,303	0,366	0,355	0,401	0,228
(negra)	*P*	0,381	0,266	0,025	0,044	0,035	0,066	0,006
Raza	RPC	0,482	0,707	0,583	0,648	0,662	0,710	0,849
(otra)	*P*	0,195	0,454	0,177	0,232	0,245	0,322	0,629
Visitas	RPC	1,450	1,299	1,379	1,015	0,978	0,986	0,956
	P	0,215	0,265	0,119	0,768	0,887	0,931	0,768

RPC: razón de productos cruzados.

CUADRO 7. Comparación entre el coeficiente de regresión para el colesterol, y el promedio de las estimaciones basadas en las 500 réplicas, para algunos escenarios, con tamaño de la muestra = 100 y dicotomización de la edad en deciles 1 y 5

Coeficiente de correlación entre			Verdadero coeficiente de regresión	Promedio estimación decil 1	Diferencia relativa (%) decil 1	Promedio estimación decil 5	Diferencia relativa (%) decil 5
C y E	TS y E	TS y C					
0,0	0,0	0,4	0,300	0,3056	1,86	0,3054	1,81
0,0	0,4	0,4	0,300	0,2710	9,66	0,3214	7,14
0,0	0,7	0,4	0,300	0,2663	11,23	0,2619	12,69
0,4	0,0	0,4	0,3571	0,3044	14,78	0,3343	6,40
0,4	0,4	0,4	0,2143	0,2770	29,28	0,2415	12,68
0,4	0,7	0,4	0,1071	0,2025	89,00	0,2195	104,90
0,7	0,0	0,4	0,5882	0,3381	42,52	0,4740	19,42
0,7	0,4	0,4	0,1765	0,2620	48,47	0,2404	36,25
0,7	0,7	0,4	–0,1324	0,2729	306,20	0,1622	222,60

C: colesterol. E: edad. TS: tensión sistólica.

CUADRO 8. Medias de la diferencia relativa, para combinaciones de coeficientes de correlación de la variable de control con la exposición y la respuesta

Correlación entre variables de exposición y de control	Correlación entre variables de respuesta y de control				
	0,00	0,10	0,40	0,70	General
0,00	1,93	1,09	6,69	14,43	6,03
0,10	1,35	2,13	16,54	72,93	23,24
0,40	8,53	10,71	45,40	62,92	31,89
0,70	30,34	37,62	47,75	141,66	62,50
General	8,74	10,64	29,10	66,74	28,81

entre las variables de respuesta y de control, y las variables de exposición y de control. Por término medio, se encontró una diferencia relativa de 28,81% entre el valor estimado y el valor verdadero. Es importante destacar cómo la magnitud de la diferencia relativa va aumentando a medida que los coeficientes de correlación crecen, llegándose a una diferencia media de 141,66% en el caso de existir una fuerte asociación de la variable de control con la de respuesta y la de exposición, tal como se observaba en las situaciones puntuales analizadas en el cuadro 7. Cuando la variable de control tiene una débil asociación con las otras variables del modelo, como era de esperar, su dicotomización apenas afecta a la relación de la variable de respuesta con la de exposición.

DISCUSIÓN

Como se desprende tanto de las referencias bibliográficas como de los ejemplos presentados, las conclusiones de un estudio están estrechamente vinculadas con las estrategias de análisis y con el procesamiento de los datos anterior a dicho análisis. En muchos casos, puede parecer bastante difícil considerar como válidas las conclusiones de una investigación como las discutidas aquí, en virtud del procesamiento que hayan sufrido los datos. Como queda demostrado, si bien es cierto que la dicotomización puede ser un elemento útil para facilitar la interpretación de resultados y, en algunos casos, evitar algunos supuestos de los modelos, eso puede ser muy peligroso ya que se haría a expensas de la validez de las conclusiones. La disyuntiva no es trivial y requiere una atenta consideración. A la luz de los resultados mostrados, no es deseable dicotomizar variables continuas sin una buena justificación para ello y teniendo siempre consciencia de los sesgos que puede introducir en las estimaciones de las magnitudes de la asociación.

Si los motivos para realizar la dicotomización son tan poderosos que no se pueda evitar, es probable que sea preferible una categorización en más grupos, en vez de una simple dicotomización. Sin embargo, en cualquier caso, es recomendable un análisis inicial de los datos que sugiera algún tipo particular de categorización.

En resumen, hablando en términos muy generales, creemos que se debe evitar este tipo de prácticas, particularmente debido a los sesgos que pueden implicar en la evaluación de las asociaciones en estudios observacionales o experimentales.

REFERENCIAS

1. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. Am J Epidemiol 1991;134:1233–1244.
2. Reade-Cristopher S, Kupper LL. On the effects of predictor misclassification in multiple linear regression analysis. Comm Statistics: Theory and Methods 1995;24:13–37.
3. Fung KY, Howe GR. Methodological issues in case-control studies III: The effect of joint misclassification of risk factors and confounding factors upon estimation and power. Intl J Epidemiol 1984;13:366–370.
4. Ragland DR. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. Epidemiology 1992;3:434–440.
5. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors [commentary]. J Natl Cancer Inst 1994;86: 829–835.
6. Kahn H, Sempos C. Statistical methods in epidemiology. New York: Oxford University Press; 1989.
7. Zhao LP, Kolonel LN. Efficiency loss from categorizing quantitative exposures into qualitative exposure in case-control studies. Am J Epidemiol 1992;136:464–474.
8. Donner A, Eliasziw M. Statistical implications on the choice between a dichotomous or continuous trait in studies of interobserver agreement. Biometrics 1994;50: 550–555.
9. Amenta JS, Brocher SC, Serenko-Aber AL. Comparing different statistical methods for evaluating diagnostic effectiveness of clinical tests: respiratory distress syndrome as a model. Clin Chem 1988;34:273–280.
10. Wartenberg D, Northridge M. Defining exposure in case-control studies: a new approach. Am J Epidemiol 1991;133:1058–1071.
11. Altman D. Problems in dichotomizing continuous variables [letter]. Am J Epidemiol 1994; 139:442.
12. Maxwell S, Delaney H. Bivariate median splits and spurious statistical significance. Psychological Bull 1993;113:181–190.
13. Cumsille F, Bangdiwala SI. Dicotomización de variables continuas en modelos de regresión logística. Rev Med Chile 1996;124: 836–842.

14. Cumsille F. Effect of dichotomizing continuous variables in regression model [Dr.P.H. dissertation]. Department of Biostatistics, University of North Carolina, Chapel Hill, 1997.
15. Chang HG, Lininger LL, Doyle JT, Maccubbin PA, Rothenberg RB. Application of the Cox model as a predictor of relative risk of coronary heart disease in the Albany study. Stat Med 1990;9:287–292.
16. Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc.; 1989.

Manuscrito recibido el 15 de febrero de 1999 y aceptado para publicación, tras revisión, el 29 de agosto del 2000.

ABSTRACT

Categorizing variables in the statistical analysis of data: consequences for interpreting the results

Frequently during the process of data analysis in epidemiological studies, the scale of one or more continuous variables is changed. The objective of this paper was to assess the consequences of categorizing variables during data analysis. We studied three situations with different scenarios for statistical analysis with regression models. The results show that dichotomizing continuous variables can substantially modify the relationships between dependent and independent variables. Thus, for example, in epidemiological studies trying to evaluate the effect of an exposure on a response, the magnitude and/or the direction of this effect can be biased by dichotomizing a variable. We therefore recommend avoiding, as much as possible, the categorization of variables when doing analyses.

To understand fully all the variations which disease may show, it is necessary to draw on the experience of the world and not to reason too broadly from results obtained in a small section of a single country. Disease shows many peculiarities under the multiple influences of varying environment and that applies not alone to the communicable diseases. Diphtheria and cancer are almost universally described in terms of the clinical and epidemiological behavior they evidence in north temperate climates. What occurs in the tropics is often widely different. An international viewpoint becomes increasingly necessary for a full and clearer comprehension of disease.

[Para comprender completamente todas las variaciones que puede mostrar la enfermedad, es necesario aprovechar la experiencia del mundo y no razonar de forma demasiado general sobre los resultados obtenidos en una pequeña sección de un solo país. La enfermedad muestra muchas particularidades bajo las múltiples influencias de un ambiente variable, y esto se aplica no solamente a las transmisibles. La difteria y el cáncer se describen casi universalmente en términos del comportamiento clínico y epidemiológico que manifiestan en los climas templados del norte. Lo que ocurre en los trópicos es con frecuencia muy diferente. Un punto de vista internacional se hace cada vez más necesario para obtener una comprensión completa y más clara de la enfermedad.]

John E. Gordon
"Epidemiology—old and new"
Journal of the Michigan State Medical Society, 1950;49