

# Estimaciones usadas en diseños muestrales complejos: aplicaciones en la encuesta de salud cubana del año 2001

Mayilée Cañizares Pérez,<sup>1</sup> Isabel Barroso Utra,<sup>2</sup> Alina Alfonso León,<sup>3</sup>  
René García Roche,<sup>2</sup> Karen Alfonso Sagué,<sup>2</sup> Martha Chang de la Rosa,<sup>2</sup>  
Mariano Bonet Gorbea<sup>2</sup> y Esther M. León<sup>3</sup>

## Forma de citar

Cañizares Pérez M, Barroso Utra I, Alfonso León A, García Roche R, Alfonso Sagué K, Chang de la Rosa M, et al. Estimaciones usadas en diseños muestrales complejos: aplicaciones en la encuesta de salud cubana del año 2001. *Rev Panam Salud Publica* 2004;15(3):176-84.

## RESUMEN

**Objetivos.** En el presente trabajo se discuten las particularidades de tres métodos de estimación de parámetros simples —medias, totales y porcentajes, y sus respectivos errores estándar—, así como de los modelos de regresión logística, en el análisis de los datos de muestras complejas.

**Métodos.** Se tomaron los datos de la Segunda Encuesta Nacional de Factores de Riesgo y Afecciones Crónicas No Transmisibles, realizada en Cuba en el año 2001. Se usó un diseño muestral complejo estratificado y por conglomerados polietápicos. Los estratos estuvieron conformados por las 14 provincias de Cuba y el municipio especial Isla de la Juventud, mientras que los conglomerados fueron las áreas geográficas muestrales (AGEM), manzanas y secciones. Las muestras se ponderaron en proporción inversa a las probabilidades de selección y para la estimación se realizó una estratificación por sexo y grupo de edad (15-34, 35-54, 55-74 y 75 años o más). Las varianzas se estimaron mediante las aproximaciones de Taylor. Se compararon tres métodos estadísticos: el análisis convencional, que da por sentado que los datos provienen de un muestreo simple aleatorio; el análisis ponderado, que solo toma en consideración los pesos muestrales para obtener las estimaciones; y el análisis con ajustes, que toma en cuenta la desigualdad en las probabilidades de selección y el efecto de la conglomeración en los datos.

**Resultados.** Las estimaciones puntuales obtenidas mediante las tres estrategias de análisis fueron similares. Los errores estándar estimados para la prevalencia de sobrepeso obtenidos por el análisis convencional se subestimaron en 19,3% y para la prevalencia de hipertensión en más de 11,5% en relación con los otros métodos. El análisis ponderado generó valores de los errores estándar mucho menores que los calculados por otros métodos. Resultados similares se encontraron al estimar los valores de la razón de posibilidades.

**Conclusiones.** Los métodos de análisis que toman en cuenta la estructura de los datos y las características del diseño ofrecen una visión más realista del problema estudiado. Con ellos se obtienen estimaciones más precisas de los parámetros y de sus errores estándar. La frecuencia con que se encuentran datos procedentes de diseños muestrales complejos en estudios epidemiológicos y de salud pública exige una mayor utilización de estos métodos y de los paquetes estadísticos que los contemplan.

## Palabras clave

Diseño de investigaciones epidemiológicas, muestras, muestreo, técnicas de estimación.

<sup>1</sup> Arthritis Community Research & Evaluation Unit (ACREU), Toronto, Canadá. La correspondencia debe enviarse a: Mayilée Cañizares Pérez, Arthritis Community Research & Evaluation Unit (ACREU),

Division of Outcomes & Population Health, University Health Network, 399 Bathurst Street, Toronto, Canada, M5T 2S8. Correo electrónico: mcanizar@uhnres.utoronto.ca

<sup>2</sup> Instituto Nacional de Higiene, Epidemiología y Microbiología (INHEM), Ciudad de La Habana, Cuba.

<sup>3</sup> Oficina Nacional de Estadística (ONE), Ciudad de La Habana, Cuba.

En los estudios epidemiológicos a menudo se utilizan encuestas domiciliarias para estimar la frecuencia de determinadas características de salud o de factores de riesgo en la población. Debido a restricciones prácticas y económicas es frecuente que el muestreo simple aleatorio (MSA) no resulte suficientemente eficiente o factible y que sea necesario recurrir a estrategias alternativas de muestreo, como la estratificación, la selección en etapas, la formación de conglomerados o el empleo de probabilidades de selección desiguales. Los diseños muestrales que incorporan combinaciones de estas estrategias se denominan complejos, por contraste con el MSA, en el que las unidades muestrales se seleccionan independientemente unas de otras y todas tienen igual probabilidad de selección y distribución.

El análisis de los datos obtenidos mediante diseños muestrales complejos puede resultar más complicado por la posible existencia de una correlación entre las observaciones de un mismo conglomerado. Esta correlación se debe tomar en cuenta para obtener estimaciones apropiadas de los parámetros, sus varianzas y la correcta distribución de los estadígrafos (1, 2). Por su parte, se ha documentado que las varianzas de los datos obtenidos mediante MSA se subestiman, lo que lleva a obtener intervalos de confianza más estrechos y niveles de significación mayores de lo debido (2, 3–5).

Durante mucho tiempo no se han tomado en cuenta las particularidades de la estructura de este tipo de datos y se han aplicado métodos estadísticos estándares para analizarlos. Levy y Stolte (6) encontraron una tendencia cada vez mayor a usar métodos estadísticos más novedosos, como el análisis de muestras complejas, la regresión logística y la regresión proporcional de Cox, que fueron desarrollados en las décadas de 1960 y 1970 y que han tenido un impacto notable en el análisis de datos biomédicos. El porcentaje de artículos publicados en *American Journal of Public Health* y *American Journal of Epidemiology* en los que se empleó este tipo de análisis cuando se requería se incrementó de 0% en la década de 1970 a 27,8% en la

de 1980 y a más de 75,0% en la década siguiente. A pesar del notable aumento en el uso de estas técnicas estadísticas, muchos investigadores todavía no tienen en cuenta la estructura de los datos procedentes de muestras complejas durante el análisis de los resultados.

En el presente trabajo se discuten las particularidades de tres métodos de estimación de parámetros simples —medias, totales y porcentajes, y sus respectivos errores estándar—, así como de los modelos de regresión logística, en el análisis de los datos de muestras complejas, aplicado a los resultados de la Segunda Encuesta Nacional de Factores de Riesgo y Afecciones Crónicas No Transmisibles, que fue realizada en Cuba durante los años 2000 y 2001 por el Instituto Nacional de Higiene, Epidemiología y Microbiología (INHEM) y la Oficina Nacional de Estadísticas (ONE).

## MATERIALES Y MÉTODOS

### Diseño muestral de la encuesta

Se elaboró una encuesta dirigida a obtener la información necesaria para calcular las prevalencias nacionales y provinciales de las principales afecciones crónicas no transmisibles y sus factores de riesgo.

La muestra fue seleccionada a partir de la muestra general del sistema de encuestas domiciliarias elaborado por la ONE de Cuba en el año 1995 y actualizado en 1999 (7). Se usó un diseño muestral complejo estratificado y por conglomerados polietápicos. Los estratos estuvieron conformados por las 14 provincias de Cuba y el municipio especial Isla de la Juventud, mientras que los conglomerados fueron las áreas geográficas muestrales (AGEM), las manzanas y las secciones. Se calculó el tamaño muestral de cada estrato de manera que la muestra fuera representativa y permitiera obtener estimaciones provinciales. La muestra nacional se conformó mediante la agregación de las provinciales, por lo que no fue equiprobabilística.

El tamaño muestral calculado fue de 9 260 viviendas, distribuidas en 926 AGEM. Esta cifra se basó en la hipótesis de que en Cuba viven 2,6 personas por vivienda y de que debían encontrarse aproximadamente 23 743 personas mayores de 15 años residentes en viviendas particulares urbanas. En los cuadros 1 y 2 se presenta la distribución de los tamaños muestrales planificados para la encuesta.

Las provincias fueron divididas en AGEM, manzanas y secciones, conformadas por 180, 30 y 5 viviendas, respectivamente, es decir, cada AGEM estaba compuesta por 6 manzanas y

**CUADRO 1. Distribución por provincias de la muestra de viviendas planificada y estudiada. Cuba, 2001**

Provincia	Muestra planificada	Muestra estudiada	Porcentaje de respuesta
Pinar del Río	680	665	97,8
La Habana	637	623	97,8
Ciudad de La Habana	957	896	93,5
Matanzas	610	589	96,4
Villa Clara	635	613	96,4
Cienfuegos	596	568	95,1
Sancti Spiritus	622	603	96,8
Ciego de Ávila	584	574	98,1
Camagüey	625	616	98,6
Las Tunas	603	583	96,7
Holguín	608	584	96,1
Granma	600	574	95,7
Santiago de Cuba	648	623	96,1
Guantánamo	619	575	92,9
Isla de la Juventud	236	235	99,6
Total	9 260	8 921	96,3

**CUADRO 2. Distribución por provincias de la muestra de personas planificada y estudiada. Cuba, 2001**

Provincia	Muestra planificada	Muestra estudiada	Porcentaje de respuesta	Personas medidas <sup>a</sup> (%)	
				Medidas antropométricas	Tensión arterial
Pinar del Río	1 789	1 777	99,3	99,9	99,0
La Habana	1 653	1 625	98,3	98,7	98,5
Ciudad de La Habana	2 528	2 361	93,4	96,9	95,1
Matanzas	1 509	1 458	96,6	97,4	96,7
Villa Clara	1 511	1 497	99,1	99,3	99,1
Cienfuegos	1 510	1 492	98,8	91,8	91,5
Sancti Spiritus	1 565	1 497	95,7	97,9	94,9
Ciego de Ávila	1 453	1 398	96,2	96,3	95,6
Camagüey	1 685	1 531	90,9	90,5	90,5
Las Tunas	1 496	1 461	97,7	96,7	96,6
Holguín	1 563	1 544	98,8	98,7	98,1
Granma	1 545	1 452	94,0	89,1	87,3
Santiago de Cuba	1 717	1 686	98,2	99,9	96,6
Guantánamo	1 592	1 447	90,9	96,7	95,6
Isla de la Juventud	627	625	99,7	99,8	99,8
Total	23 743	22 851	96,2	96,6	95,5

<sup>a</sup> Porcentaje de la muestra estudiada.

cada manzana tenía 6 secciones (figura 1). Tanto las AGEM como las manzanas y las secciones tenían igual población. La selección de las AGEM se realizó mediante muestreo aleatorio sistemático, en cada AGEM elegida se seleccionó una manzana de manera aleatoria y de cada manzana se tomaron dos secciones al azar. En cada sección seleccionada se visitaron todas las viviendas para entrevistar a todos los residentes permanentes de 15 años de edad o más.

Las entrevistas fueron realizadas en los hogares por encuestadores profesionales. Las mediciones antropométricas (peso, talla, circunferencia de cintura y cadera) y de tensión arterial fueron tomadas por personal de enfermería en centros de mediciones habilitados para tales efectos. Los encuestadores y los enfermeros fueron previamente entrenados por el equipo de investigación, los primeros en la familiarización y estandarización de los cuestionarios, y los segundos, en la toma de la tensión arterial y de las mediciones antropométricas. Las entrevistas fueron estructuradas y abordaron los temas siguientes: características sociodemográficas; condiciones de vida (tipo y estado físico de las viviendas, combustible utilizado para cocinar, ingresos familiares y equi-

pos electrónicos en la vivienda); calidad de la vida relacionada con la salud, medida a través del cuestionario EuroQol-5 (8); percepción de riesgo de enfermedades crónicas; hábitos tóxicos (hábito de fumar y de consumir bebidas alcohólicas y drogas); actividad física, medida por la combinación de dos criterios: la actividad principal, según la clasificación de la Organización de las Naciones Unidas para la Agricultura y la Alimentación y la Organización Mundial de la Salud (9), y la actividad física adicional, según las recomendaciones del Colegio Estadounidense del Deporte (10); hábitos dietéticos; afecciones que padecía el encuestado (hipertensión arterial, enfermedades del corazón, enfermedades cerebrovasculares, diabetes mellitus, asma y bronquitis crónica); discapacidades y minusvalías; lesiones sufridas; comportamientos violentos (si era víctima de agresiones físicas o verbales); actividades de prevención de cáncer (bucal, de mama y cervicouterino, entre otros); y consumo de medicamentos.

### Ponderaciones

Las muestras se ponderaron con el fin de compensar las diferencias en las

probabilidades de selección de los diferentes subgrupos, reducir el sesgo debido a la disminución de la muestra (ya que no siempre se logra encuestar a todas las personas planificadas), estimar los totales poblacionales y compensar las deficiencias del marco muestral (por ejemplo, la omisión de algunas de las viviendas en las listas) y la reducción de las varianzas en el proceso de hacer estimaciones a partir de información auxiliar, como los tamaños poblacionales.

Las ponderaciones muestrales se obtuvieron en proporción inversa a las probabilidades de selección y para la estimación se realizó una estratificación por sexo y grupo de edad. Los grupos de edades usados para estos fines fueron de 15 a 34, 35 a 54, 55 a 74, y 75 años o más.

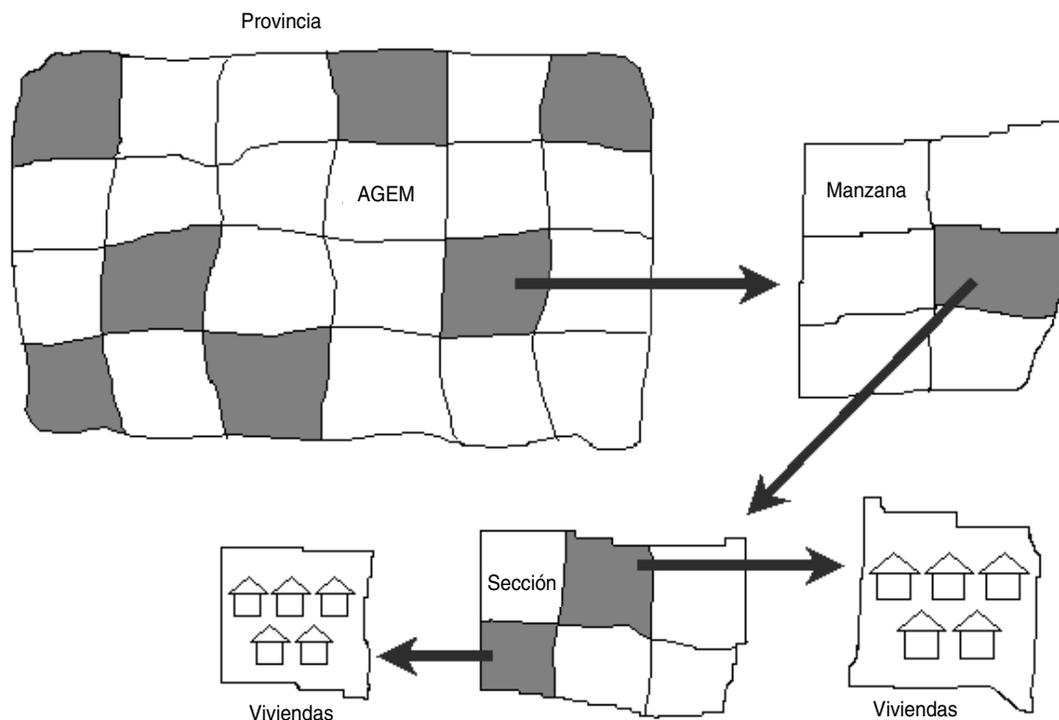
### Análisis de los datos incompletos

Los estudios poblacionales realizados por encuesta adolecen con frecuencia de datos incompletos, ya sea porque faltan unidades enteras (cuando las personas consideradas en el diseño del estudio no desean participar o cuando no es posible ponerse en contacto con ellas mediante el mecanismo de selección establecido) o porque faltan respuestas aisladas (cuando algunas personas no contestan determinadas preguntas del cuestionario) (11).

La falta de respuesta en unidades completas se compensó mediante los pesos muestrales, para los cuales se tomó en cuenta la disminución del tamaño de la muestra.

La falta de respuestas aisladas podía deberse a la falta de coherencia entre respuestas; a que el entrevistado no poseía la información solicitada o no deseaba proporcionarla; a que el entrevistado se negaba a ser medido, o a la presencia de errores en las mediciones. Este efecto se compensó mediante imputaciones, es decir, la asignación de un valor a cada dato ausente para completar el conjunto de datos y obtener estimaciones más coherentes. Los procedimientos que se usaron para imputar los valores ausentes en la encuesta fueron:

**FIGURA 1. Esquema que representa el diseño muestral de la Segunda Encuesta Nacional de Factores de Riesgo y Afecciones Crónicas No Transmisibles. Cuba, 2001**



**Nota:** Áreas geográficas muestrales (AGEM), 180 viviendas  
 Manzanas, 30 viviendas  
 Secciones, 5 viviendas

**Imputaciones lógicas:** cuando el valor ausente se podía inferir de las respuestas a otras preguntas, por ejemplo, cuando un encuestado no había respondido si era o no fumador, pero había indicado la cantidad de cigarrillos que fumaba diariamente. Este procedimiento se empleó para sustituir valores ausentes en menos de 10 observaciones.

**Imputaciones por modelos de regresión:** este procedimiento consiste en desarrollar un modelo de regresión en el que la variable dependiente es la que ha de imputarse. Una vez obtenido el modelo, se asigna el valor calculado a cada una de las observaciones que faltaban. Este procedimiento se usó para imputar los ingresos.

**Imputaciones múltiples:** en este enfoque, basado en simulaciones, a cada

valor ausente se le asignan  $m$  valores extraídos de una distribución de predicción (donde  $m > 1$ ), lo que produce  $m$  bases de datos. Después, en cada base de datos se realiza el análisis estadístico que responde al propósito del estudio y se obtienen tantas estimaciones como imputaciones se hayan hecho. Finalmente, se combinan estas estimaciones mediante la regla de Rubin para llegar al valor definitivo (12). Este procedimiento se usó para imputar las variables resultantes de las mediciones: peso, talla, circunferencia de la cintura y la cadera y tensión arterial diastólica y sistólica.

**Estimaciones en diseños muestrales complejos**

En el presente trabajo se estimaron las varianzas mediante las aproximaciones de Taylor (13). Los resultados

obtenidos por este método fueron muy similares a los obtenidos por el procedimiento *Jackknife* (14), que emplea un enfoque radicalmente diferente, lo cual confirma la confiabilidad de estos procedimientos (15).

**Efectos del diseño muestral**

Con el propósito de ilustrar el efecto de no tomar en cuenta el diseño muestral en el análisis estadístico se utilizó el procedimiento *Surveymeans* para calcular parámetros simples, como las medias y las proporciones y sus intervalos de confianza. Se utilizó también el procedimiento *Genmod* para estimar las razones de posibilidades mediante modelos de regresión logística. Las variables escogidas para este ejercicio fueron la tensión arterial diastólica y sistólica, el sobrepeso y la hipertensión arterial.

Se compararon tres métodos estadísticos: el análisis convencional, en el que se da por sentado que los datos provienen de un MSA; el análisis ponderado, en el que solo se toman en cuenta los pesos muestrales para obtener las estimaciones; y el análisis ajustado, en el que se toman en cuenta todos los aspectos del diseño muestral.

## RESULTADOS

### Disminución de la muestra y fuentes de información

Se encuestó a 22 851 personas mayores de 15 años durante el período comprendido entre noviembre de 2000 y marzo de 2001. La distribución por provincias de la muestra estudiada de viviendas y personas se presenta en los cuadros 1 y 2, respectivamente. Se visitó un total de 96,3% de las viviendas planificadas. Guantánamo resultó ser la provincia con el menor porcentaje de respuestas (92,9%) y la Isla de la Juventud fue la que tuvo el mayor (99,6%) (cuadro 1).

En cuanto a las personas encuestadas, la disminución de la muestra a escala nacional fue de 3,8%. Camagüey y Guantánamo tuvieron el menor porcentaje de respuesta (90,9%), mientras que el mayor porcentaje lo tuvo el municipio especial de Isla de la Juventud (99,7%). Para las medidas antropométricas y la tensión arterial, la disminución de la muestra fue de 3,4% y 4,5%, respectivamente (cuadro 2). La caída de la muestra se debió fundamentalmente a viviendas que se encontraban cerradas.

### Análisis de los datos incompletos

Las variables que mayor porcentaje de valores ausentes tuvieron fueron los ingresos económicos, las medidas antropométricas y la tensión arterial. Los hombres fueron los que omitieron los ingresos con mayor frecuencia (6,5%) y los que menos asistieron al centro de medición antropométrica (10,5%). Los jóvenes de 15 a 34 años y las personas de 75 años o más fueron

**CUADRO 3. Porcentaje de falta de respuesta a la pregunta sobre los ingresos económicos y de ausencia de las mediciones antropométricas, según características demográficas. Cuba, 2001**

Variable	<i>n</i>	Pregunta sobre ingresos	Mediciones antropométricas
Sexo			
Hombres	10 777	6,5	10,5
Mujeres	12 074	3,3	9,4
Edad (años)			
15 a 34	8 796	7,1	10,5
35 a 54	8 270	3,6	8,8
55 a 74	4 413	2,2	8,3
75 años y más	1 372	6,4	17,6
Educación			
Ningún nivel aprobado	2 595	7,5	12,1
Primaria	3 501	4,4	9,4
Nivel medio	7 346	5,2	9,2
Nivel medio superior	724	4,4	9,8
Universitario	2 167	2,7	10,4
Ocupación			
Ninguna	11 115	7,3	10,1
Trabajador estatal	10 162	1,9	9,9
Trabajador por cuenta propia	1 071	6,1	8,2
Empresa mixta	265	2,3	9,1
No informó	238	11,8	9,2
Región			
Oriente	7 590	5,3	11,8
Centro	7 415	4,6	10,7
Occidente	7 846	4,6	7,2

quienes dejaron más preguntas sin responder, al igual que las personas que no terminaron la escuela primaria. En relación con la ocupación, los que no identificaron el trabajo que realizaban omitieron con mayor frecuencia el monto de sus ingresos (11,8%). La falta de respuesta sobre los ingresos mostró pocas diferencias por regiones ( $P = 0,06$ ), pero en la región oriental se observó un mayor porcentaje de personas sin información antropométrica (11,8%) ( $P = 0,02$ ) (cuadro 3).

### Efectos del diseño muestral

En el cuadro 4 se presenta el efecto del diseño en la estimación de las variables escogidas. Los valores encontrados (de 1,23 en hipertensos a 1,59 en los encuestados con desnutrición) se encuentran dentro del espectro de valores aceptados (de 1,0 a 3,0), por lo que se puede considerar que el estudio tuvo un buen diseño.

Las prevalencias estimadas con los tres métodos empleados fueron muy

**CUADRO 4. Estimaciones, errores estándar (EE) y efectos del diseño para las variables seleccionadas. Cuba, 2001**

Variable	<i>n</i>	Estimación (%)	EE bajo el supuesto de MSA (%)	EE ajustado según el diseño (%)	Efecto del diseño
Desnutridos	2 946	12,7	0,22	0,35	1,59
Peso normal	11 303	49,5	0,33	0,43	1,30
Sobrepeso	8 602	37,8	0,32	0,44	1,38
Hipertensos	7 678	33,2	0,31	0,38	1,23

MSA: muestreo simple aleatorio.

**CUADRO 5. Estimaciones de porcentajes y medias por sexo y método de análisis. Cuba, 2001**

Indicadores	Análisis convencional		Análisis ponderado		Análisis ajustado	
	Hombres n = 10 777	Mujeres n = 12 074	Hombres n = 3 080 587	Mujeres n = 3 316 293	Hombres n = 3 080 587	Mujeres n = 3 316 293
Sobrepeso <sup>a</sup>	34,1 (0,46)	42,4 (0,45)	34,1 (0,026)	42,8 (0,026)	34,1 (0,57)	42,8 (0,57)
Hipertensos <sup>a</sup>	33,5 (0,46)	34,6 (0,43)	33,1 (0,026)	33,3 (0,025)	33,1 (0,52)	33,3 (0,53)
Tensión arterial diastólica <sup>b</sup>	80,4 (0,1052)	77,1 (0,1043)	80,4 (0,0006)	77,0 (0,0006)	80,4 (0,1352)	77,0 (0,1446)
Tensión arterial sistólica <sup>b</sup>	150,7 (0,1507)	121,1 (0,1571)	125,6 (0,0009)	121,1 (0,0010)	125,6 (0,1924)	121,1 (0,2192)

<sup>a</sup> Prevalencia estimada (error estándar entre paréntesis).

<sup>b</sup> Media estimada (error estándar entre paréntesis).

similares. Esto pudo deberse a la pequeña variabilidad de las ponderaciones muestrales y a que las provincias más pobladas —que están subrepresentadas en la muestra nacional— tenían las mayores prevalencias de los factores de riesgo estudiados. Sin embargo, la precisión de la estimación difiere según el método de cálculo utilizado. Los errores estándar son menores con el análisis convencional que con el análisis ajustado. Por ejemplo, los errores estándar de la prevalencia de sobrepeso calculados mediante el análisis convencional se subestimaron en 19,3% y los de la hipertensión, en más de 11,5%.

El análisis ponderado generó errores estándar mucho menores que los otros métodos. Esto se debe a que en el análisis convencional el tamaño de la muestra es igual a la suma de los pesos muestrales (por ejemplo, la muestra de los hombres sería de 3 080 587 en vez de 10 777, que es el verdadero tamaño muestral). Los errores estándar tan pequeños hacen que los intervalos de confianza resulten muy estrechos y, por lo tanto, algunas diferencias no significativas cobran significación. Por

ejemplo, con el análisis ponderado los intervalos de confianza para la prevalencia de hipertensos serían de 33,45 a 33,55% en los hombres y de 34,55% a 34,65% en mujeres. Estos intervalos son muy estrechos y no se traslapan, de lo que se podría inferir que existen diferencias significativas entre hombres y mujeres. Sin embargo, los intervalos de confianza calculados mediante el análisis ajustado serían de 33,08% a 34,12% en el caso de los hombres y de 32,26% a 34,34% en el de las mujeres, es decir, no indican que existan diferencias significativas entre los dos sexos.

Al estimar las medias de las tensiones arteriales por los tres métodos se encontraron resultados similares a los observados al estimar las prevalencias de sobrepeso e hipertensión. Los valores medios de la tensión arterial diastólica y sistólica obtenidos con los tres métodos fueron muy similares y, de igual manera, solo se observaron cambios en la precisión de las estimaciones (cuadro 5).

Al ajustar los modelos de regresión se observaron diferencias en las estimaciones de las razones de posibilidad-

des en todos los casos. Al igual que en la estimación de las prevalencias y las medias, la precisión de las estimaciones resultó la más afectada. El análisis ajustado produjo los intervalos de confianza más amplios para las razones de posibilidades (entre 0,43 y 1,75), mientras que el análisis ponderado dio los más estrechos (entre 0,01 y 0,08). Esto se debe a que los métodos que no incorporan la estructura de los datos subestiman el error estándar (cuadro 6). En este caso el efecto de diseño fue mayor en alrededor de 1,6 veces. De igual manera estos resultados indican que si no se toma en cuenta la estructura del diseño al calcular la varianza de los estimadores, se obtendrán intervalos de confianza más estrechos y, por ende, pruebas de hipótesis con resultados más significativos de los que se obtendrían si se tomara en cuenta el diseño en el análisis, lo que conduce a conclusiones inapropiadas.

## DISCUSIÓN

Con frecuencia se realizan estudios a gran escala para obtener estimaciones

**CUADRO 6. Razón de posibilidades e intervalos de confianza para el efecto del índice de masa corporal sobre la hipertensión arterial, según el sexo y el procedimiento de análisis. Cuba, 2001**

Relación	Análisis convencional		Análisis ponderado		Análisis ajustado	
	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres
Sobrepeso frente a peso normal	4,16 (3,65–4,74)	3,33 (3,30–3,36)	3,97 (3,93–4,01)	3,33 (3,30–3,36)	4,14 (3,36–5,11)	3,31 (2,71–4,06)
Sobrepeso frente a desnutridos	2,23 (2,13–2,34)	2,11 (2,10–2,12)	2,23 (2,22–2,24)	2,11 (2,10–2,12)	2,24 (2,01–2,50)	2,12 (1,91–2,34)

nacionales de problemas de salud. Uno de los inconvenientes de las muestras que no son equiprobabilísticas es que para ciertas subpoblaciones minoritarias no se obtienen estimaciones precisas. En estos casos, es necesario someter la muestra a un tratamiento especial mediante ponderaciones muestrales que permitan compensar las distorsiones de la representatividad de la muestra. Las ponderaciones toman en cuenta las probabilidades de selección de cada persona, la disminución de la muestra y la población. De esta manera, la ponderación puede interpretarse como la medida del número de personas de la población que representa cada individuo de la muestra (1, 2).

La importancia de tomar en cuenta los pesos muestrales al analizar este tipo de datos se hace más evidente si el valor de estos pesos tiene una correlación inversa con la variable analizada, ya que en ese caso la estimación muestral no ponderada será mayor que el valor real. Por ejemplo, si se toman dos muestras de igual tamaño en dos áreas geográficas con tamaños poblacionales diferentes y la estatura promedio en el área de mayor población es menor que en el área menos poblada, la media muestral global no ponderada será mayor que la media muestral ponderada debido a que los individuos del área menor tienen una mayor representación en la muestra y su contribución a la estimación global es mayor que su contribución en la población.

Aunque la evidencia empírica de este trabajo se basa en un tipo de diseño muestral complejo, los resultados coinciden con los de otros estudios similares (1, 16). El uso de paquetes estadísticos estándar para analizar datos provenientes de diseños muestrales complejos lleva generalmente a obtener estimaciones puntuales sesgadas de los parámetros poblacionales, los errores estándar se subestiman, se obtienen intervalos de confianza muy estrechos y las pruebas de significación llevan con mayor frecuencia a rechazar la hipótesis de nulidad, ya que se subestiman las medidas de variabilidad.

La magnitud del sesgo en las estimaciones convencionales depende de

los datos que se analizan, en particular de los métodos empleados para calcular los pesos muestrales y su variabilidad (2, 16, 17). Si los pesos muestrales tienen poca variabilidad en los datos, las estimaciones puntuales convencionales son similares a las estimaciones puntuales ponderadas (18). En la encuesta cubana, los pesos tuvieron una variabilidad pequeña, lo que contribuyó a que las estimaciones puntuales obtenidas por los tres métodos fueran similares. Sin embargo, los resultados de las regresiones logísticas tuvieron mayor variación. De ahí la importancia de tomar en cuenta el diseño en el análisis para llegar a conclusiones correctas.

En el análisis convencional, además del sesgo que se produce en las estimaciones puntuales, generalmente se subestiman los errores estándar y las medidas de variabilidad. Esto se debe a que este tipo de análisis no toma en cuenta la estructura de los datos y las diferencias de las probabilidades de selección. En este caso, las estimaciones de los errores estándar más eficientes fueron las que se obtuvieron mediante el análisis ajustado, ya que este método toma en cuenta la estructura de los datos para calcular las estimaciones de la varianza. Estos resultados son similares a los notificados por Brogan (1), Barroso y Cañizares (18) y Warszawski y col. (19).

Con el análisis convencional también se subestiman las medidas de variabilidad, y la magnitud de esta subestimación depende del tamaño del coeficiente de correlación de las variables que se analizan dentro del conglomerado. Cuando la correlación es grande, la variabilidad se subestima en mayor medida. El análisis ponderado es una estrategia de análisis razonable si el coeficiente de correlación es muy pequeño. En estos casos es preciso evitar el empleo de los totales poblacionales para estimar los errores estándar cuando se utilice la opción de análisis ponderado disponible en los paquetes estadísticos comerciales.

Bajo determinadas circunstancias, los paquetes estadísticos estándar pueden brindar resultados cercanos a los obtenidos mediante paquetes especia-

lizados. Sin embargo, es importante reconocer estas circunstancias y estar atentos a los errores cuando se utilicen estos métodos. No existen muchas dificultades para obtener indicadores ponderados, ya que estos paquetes contemplan esta opción. Sin embargo, este análisis ofrece estimaciones sin sesgo de los parámetros, pero no de sus errores estándar. De los tres procedimientos aplicados para estimar los parámetros poblacionales en el ejemplo, los que usan los pesos (análisis ponderado y análisis ajustado) proporcionan una estimación de ellos que no está sesgada. Esto se debe a que no todos los individuos tienen iguales probabilidades de ser incluidos en la muestra y los pesos muestrales incorporan estas diferencias, así como las propias de la cobertura del muestreo y las generadas por la falta de respuesta a algunas preguntas (2).

Las estimaciones ponderadas no estarían sesgadas si todas las personas en las viviendas seleccionadas respondieran a todas las preguntas del cuestionario y asistieran a los centros de medición. Sin embargo, con frecuencia esto no ocurre así, aun en países como Cuba donde el sistema de salud cubre a todo el país y la población está muy dispuesta a colaborar en estudios relacionados con la salud. La ausencia de datos, ya sea en unidades o en respuestas aisladas, puede crear sesgos en las estimaciones derivadas de diseños muestrales complejos. Si existen diferencias entre las personas que no responden y las que sí lo hacen y dichas diferencias no se tienen en cuenta en el análisis, perderá validez cualquier estimación o inferencia poblacional que se haga (8, 20).

La mayor parte de las veces, al hacer una inferencia estadística se parte del supuesto de que los datos provienen de un MSA y de una población infinita, es decir, que las observaciones son independientes y pertenecen a una distribución teórica dada. En la práctica, sin embargo, la muestra se obtiene mediante diseños muestrales complejos. El problema central radica en que las observaciones obtenidas mediante estos diseños casi siempre tienen cierto grado de dependencia mutua entre sí.

Consecuentemente, si esto no se toma en cuenta se producen sesgos en las estimaciones y ellos a su vez producen incrementos indebidos en la potencia de las pruebas de hipótesis.

Este es un tema de gran complejidad teórica (2). Las fórmulas aproximadas para realizar las estimaciones de parámetros sencillos (medias y totales) y sus varianzas bajo diseños muestrales complejos pueden deducirse de diversos modos. Para los primeros, generalmente se utilizan estimadores de razón que utilizan información auxiliar acerca de la población finita —como los tamaños de las unidades y las variables de estratificación— que los investigadores conocen desde antes de seleccionar la muestra. Esta información auxiliar influye en la elección del diseño muestral y en la estimación de los parámetros, ya que refleja la estructura de la población. Para estimar las varianzas se utilizan diferentes fórmulas que no son expresiones algebraicas exactas. Por lo general se utilizan dos aproximaciones para calcular las varianzas: linealizaciones a partir de las series de Taylor (13) y las técnicas de replicación (14). Para usar cualquiera de estos métodos es necesario conocer las características del diseño muestral, es decir, la variable que identifica las ponderaciones y las variables que identifican los estratos y los conglomerados.

Las fórmulas para resolver problemas inferenciales más complejos —por ejemplo, la estimación de coeficientes de regresión— parten de supuestos tanto o más avanzados. La ecuación de estimación generalizada se aplica a los modelos de regresión donde no se

cumple el supuesto de independencia entre las observaciones, pues se toma en cuenta la dependencia que pueden tener las observaciones dentro de los conglomerados (21). El método de la ecuación de estimación generalizada puede considerarse una extensión de los modelos lineales generalizados en muestras aleatorias procedentes de diseños complejos (22).

Para estimar las varianzas muestrales se requieren programas estadísticos desarrollados especialmente para acomodar el diseño muestral en las estimaciones que se vayan a realizar. Ninguno de los paquetes estadísticos estándares toma en cuenta adecuadamente los aspectos del diseño muestral en el análisis estadístico ni provee estimaciones apropiadas de los errores estándar, debido a que todos dan por sentado que los datos provienen de un MSA. No obstante, algunos ofrecen la opción de incluir los pesos muestrales para obtener los estimadores de los parámetros simples, como SUDAAN (23), STATA (24) y SAS (25). Los dos primeros proveen estimaciones precisas de una gran variedad de estadígrafos —como medias, porcentajes, estimados poblacionales—, así como de coeficientes de regresión y de razones de posibilidades.

El *Genmod* es un procedimiento de carácter general del SAS que permite ajustar una gran variedad de modelos y diseños, basado en la ecuación de estimación generalizada. En versiones más recientes del SAS se han incorporado otros procedimientos, como el *Surveymeans* y el *Surveyreg*, diseñados especialmente para el análisis de datos complejos. Con el procedimiento *Sur-*

*veymeans* se pueden estimar medias, totales y proporciones, así como sus intervalos de confianza. El *Surveyreg* permite ajustar modelos de regresión lineal múltiple. Ambos procedimientos utilizan la linealización de series de Taylor para estimar las varianzas.

La elección de cuál procedimiento usar para completar los datos incompletos resulta una tarea compleja, pues un mismo procedimiento en determinadas situaciones produce estimaciones precisas y en otras no. Cuando los investigadores se enfrentan con datos incompletos deben haber valorado previamente más de una alternativa de tratamiento de datos y haber realizado, además, un análisis de sensibilidad que les permita elegir el mejor procedimiento a utilizar.

Los métodos de análisis que toman en cuenta la estructura de los datos y las características del diseño ofrecen una versión más realista del problema estudiado. Con ellos se obtienen estimaciones más precisas de los parámetros y de sus errores estándar. La frecuencia con que se encuentran datos procedentes de diseños muestrales complejos en estudios epidemiológicos y de salud pública exige una mayor utilización de estos métodos y de los paquetes estadísticos que los contemplan.

**Agradecimientos.** Los autores agradecen la colaboración del colectivo de investigadores de la Segunda Encuesta Nacional de Factores de Riesgo y Afecciones No Transmisibles, así como los valiosos comentarios y recomendaciones de los revisores de la *Revista Panamericana de Salud Pública*.

## REFERENCIAS

1. Brogan D. Software for sample survey data: misuse of standard packages. En: Armitage P, Colton T, eds. *Encyclopedia of biostatistics*. New York: John Wiley; 1998. Pp. 4167–74.
2. Skinner CJ, Holt D, Smith TMF. *Analysis of complex surveys*. New York: John Wiley; 1989.
3. Hidiriglou MA, Rao JNK. Chi-squared tests with categorical data from complex surveys. Part I. *Journal of Official Statistics*. 1987;3(2): 117–32.
4. Hidiriglou MA, Rao JNK. Chi-squared tests with categorical data from complex surveys. Part II. *Journal of Official Statistics*. 1987;3(2): 133–40.
5. Binder DA. On the variances of asymptotically normal estimators from complex surveys. *Int Stat Rev*. 1983;51:279–92.
6. Levy PS, Stolte K. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Stat Methods Med Res*. 2000;9:41–55.
7. *Diseño muestral general del sistema de encuestas de hogares*. La Habana: Oficina Nacional de Estadísticas; 1999.
8. Badia X, Roset M, Montserrat S, Herdman M, Segura A. La versión española del EuroQol:

- descripción y aplicaciones. *Med Clin.* 1999; 112(Supl 1):79–86.
9. Organización de las Naciones Unidas para la Agricultura y la Alimentación, Organización Mundial de la Salud. Necesidades de energía y de proteínas. Informe de una Reunión Consultiva Conjunta FAO/OMS/UNU de expertos. Ginebra: FAO, OMS; 1985. (Serie Informes Técnicos No. 724).
  10. American College of Sports Medicine. Experts release new recommendation to fight America's epidemic of physical inactivity. *Am Coll Sports Med News Release*, July 29, 1993.
  11. Cañizares M, Barroso I, Alfonso K. Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. *Gac Sanit.* 2003 (en prensa).
  12. Shaffer JL. *Analysis of incomplete multivariate data.* London: Chapman and Hall; 1997.
  13. Binder DA. On the variances of asymptotically normal estimators from complex surveys. *Int Stat Rev.* 1983;51:279–92.
  14. Rust KF, Rao JNK. Variance estimation for complex surveys using replication techniques. *Stat Meth Med Res.* 1996;5:283–310.
  15. Silva LC, Barroso IM. *Regresión logística.* Madrid: La Muralla/Hespérides; 2003.
  16. Kneipp SM, Yarandi HN. Complex sampling designs and statistical issues in secondary analysis. *West J Nurs Res.* 2002;24:552–66.
  17. Design and estimation for the National Health Interview Survey, 1995–2004. *Vital Health Stat 2.* 2000;130:1–31.
  18. Barroso IM, Cañizares M, Lera L. Influencia de la estructura de los datos en la selección de los métodos de análisis estadísticos. *Rev Esp Salud Publica.* 2002;76(2):95–103.
  19. Warszawski J, Messiah A, Lellouch J, Meyer L, Deville JC. Estimating means and percentages in a complex sampling survey: application to a French national survey on sexual behavior (ACSF). *Stat Med.* 1997;16:397–423.
  20. Verbeke G, Molenberghs G. *Linear mixed models for longitudinal data.* New York: Springer Verlag; 2000.
  21. Preisser JS, Koch GG. Categorical data analysis in public health. *Ann Rev Pub Health.* 1997;18:51–82.
  22. McCullagh P, Nelder JA. *Generalised linear models.* London: Chapman and Hall; 1987.
  23. Shah BV, Barnwell BG, Bieler GS. *SUDAAN user's manual: Release 7.0.* Research Triangle Park, North Carolina: Research Triangle Institute; 1996.
  24. STATA. *Statistical software for Windows.* College Station, Texas: Stata Corporation; 1996.
  25. SAS. *Statistical software for Windows.* Cary, North Carolina: SAS Institute Inc.; 1996.

Manuscrito recibido el 20 de junio de 2003. Aceptado para publicación, tras revisión, el 30 de diciembre de 2003.

## ABSTRACT

### Estimate methods used with complex sampling designs: their application in the Cuban 2001 health survey

**Objectives.** To look at the individual features of three different methods used to estimate simple parameters—means, totals, and percentages, as well as their standard errors—and of logistic regression models, and to describe how such methods can be used for analyzing data obtained from complex samples.

**Methods.** Data from Cuba's Second National Survey of Risk Factors and Non-Communicable Chronic Ailments [*Segunda Encuesta Nacional de Factores de Riesgo y Afecciones Crónicas No Transmisibles*], which was conducted in 2001, were studied. A complex, stratified multi-stage cluster sampling design was used. Cuba's 14 provinces and the municipality of Isla de la Juventud served as the strata, while the clusters consisted of sampled geographic areas (SGA), blocks, and sectors. Samples were weighted in inverse proportion to their probability of being selected, and estimates were performed by sex and age group (15–34, 35–54, 55–74, and 75 or more years). Taylor approximations were used to estimate variances. Three statistical methods were compared: conventional analysis, which assumes all data were obtained through simple random sampling; weighted analysis, which only takes into account the weight of the samples when performing estimates; and adjusted analysis, which looks at all aspects of the sampling design (namely, the disparity in the probability of being included in the sample and the effect of clustering on the data).

**Results.** The point estimates obtained with the three different types of analytic methods were similar. Standard error (SE) estimates for the prevalence of overweight and of arterial hypertension that were obtained by conventional analysis were underestimated by 19.3% and by more than 11.5%, respectively, when such estimates were compared to those obtained with the other two analytic methods. On the other hand, weighted analysis generated SE values that were much smaller than those obtained with the other two types of analyses. The same pattern was noted when odds ratios were calculated using the different methods.

**Conclusions.** Analytic methods that take into account the way the data are structured as well as the study design give a more realistic picture of the problem under study and provide more exact estimates of the study parameters and their SE than conventional analytic methods. Because data from epidemiologic and public health research are often obtained through complex sampling designs, the methods described in this paper and the statistical packages that utilize them should be used more widely.