Revista Española de Nutrición Comunitaria

Rev Esp Nutr Comunitaria 2015;21(Supl. 1):188-194
ISSN 1135-3074

DOI: 10.14642/RENC.2015.21.sup1.5065

Identificación, impacto y tratamiento de datos perdidos y atípicos en epidemiología nutricional

Rosa Abellana Sangra[1], Andreu Farran Codina[2]

[1]Departamento de Salud Pública. Facultad de Medicina. Universidad de Barcelona. [2]Departamento de Nutrición y Bromatología. Facultad de Farmacia. Universidad de Barcelona.

Resumen

Cuando se realiza un estudio epidemiológico nutricional, es inevitable que aparezcan valores perdidos y atípicos. Los datos perdidos aparecen, por ejemplo, por la dificultad de recoger los datos en las encuestas dietéticas que conducen a una falta de información sobre la cantidad de alimentos consumidos y una pobre descripción de ellos. Un inadecuado tratamiento durante el proceso de recolección nos conduce a sesgos y pérdida de precisión y consecuentemente una incorrecta interpretación de los resultados. El objetivo de este artículo es proporcionar recomendaciones sobre el tratamiento de datos perdidos y atípicos, y algunas orientaciones sobre el software existente para calcular el tamaño de muestra y realizar el análisis estadístico. También se realizan recomendaciones sobre la recolección de datos que es un paso importante en la investigación nutricional. Se comentan los métodos que se usan para hacer frente a los datos perdidos, específicamente, el método eliminación de casos, imputación simple o múltiple con indicaciones y ejemplos. También se relata cómo se identifican datos atípicos, el impacto que tienen en el análisis estadístico, las opciones para un adecuado tratamiento y se ilustra mediante un ejemplo. Finalmente, se menciona el software existente que aborda total o parcialmente las cuestiones tratadas, específicamente el software de libre distribución.

Palabras clave: *Valores perdidos. Valore atípicos. Recogida de datos. Epidemiología nutricional.*

THE IDENTIFICATION, IMPACT AND MANAGEMENT OF MISSING VALUES AND OUTLIER DATA IN NUTRITIONAL EPIDEMIOLOGY

Abstract

When performing nutritional epidemiology studies, missing values and outliers inevitably appear. Missing values appear, for example, because of the difficulty in collecting data in dietary surveys, leading to a lack of data on the amounts of foods consumed or a poor description of these foods. Inadequate treatment during the data processing stage can create biases and loss of accuracy and, consequently, misinterpretation of the results. The objective of this article is to provide some recommendations about the treatment of missing and outlier data, and orientation regarding existing software for the determination of sample sizes and for performing statistical analysis. Some recommendations about data collection are provided as an important previous step in any nutritional research. We discuss methods used for dealing with missing values, especially the case deletion method, simple imputation and multiple imputation, with indications and examples. Identification, impact on statistical analysis and options available for adequate treatment of outlier values are explained, including some illustrative examples. Finally, the current software that totally or partially addresses the questions treated is mentioned, especially the free software available.

Key words: *Missing data. Outliers. Data collection. Epidemiology nutritional.*

Introducción

De acuerdo con la Organización Mundial de la Salud (OMS), la epidemiologia es el "estudio de la distribución de los determinantes de estados o eventos (incluyendo las enfermedades) relacionados con la salud, y la aplicación de este estudio al control de enfermedades y otros problemas de salud". La epidemiologia nutricional se orienta a aspectos de la dieta que pueden influir en la aparición de enfermedades en humanos.

La dieta es un complejo repertorio de exposiciones que están fuertemente correlacionadas. Los individuos se exponen a la dieta en diferentes grados, con pocos cambios claros en la dieta que se produzcan en momentos fácilmente identificables. La evaluación de la ingesta de alimentos es difícil y está sujeta a múltiples sesgos.

Correspondencia: Rosa Abellana Sangra.
Faculta de Medicina. Universidad de Barcelona.
C/ Casanova, 143.
08036 Barcelona. España.
E-mail:rabellana@ub.edu

Además, el consumo de nutrientes se determina normalmente de manera indirecta, a partir del consumo de alimentos reportado o a partir de los niveles de determinados parámetros bioquímicos. Por consiguiente, la limitación más seria a la investigación en epidemiologia nutricional es la medida de la exposición a factores dietético-nutricionales.

Entre otros problemas, los valores desconocidos en los estudios de la dieta pueden aparecer a causa de la falta de registro de consumo de alimentos en determinados días u ocasiones de ingesta, información descriptiva inadecuada para una correcta codificación a nivel individual, la ausencia de alimentos consumidos o de nutrientes de interés en las tablas de composición de alimentos y, en un sentido amplio, los no participantes en un estudio que incluya una muestra aleatoria representativa. Si la imputación de los valores perdidos no se realiza, el efecto de esta falta de información debe tenerse en cuenta en la interpretación de los resultados obtenidos en los estudios de investigación.

El proceso de investigación científica puede dividirse en diferentes etapas. En primer lugar, es importante revisar la literatura científica y formular apropiadamente un objetivo de investigación y una hipótesis. Luego, es preciso elaborar un buen diseño de investigación que sea capaz de responder a la pregunta formulada. Los procedimientos de muestreo y de determinación del tamaño de la muestra son partes importantes del diseño. Toda esta información debe explicitarse en un protocolo de investigación, en el cual deben figurar detalles instrumentales y procedimentales del estudio. Por ejemplo, deben incluirse los cuestionarios, los análisis bioquímicos u otros procedimientos de obtención de datos. La validación de los cuestionarios de alimentos es un punto importante para evitar sesgos en los datos. Una vez el protocolo se ha elaborado y ha sido revisado, puede empezar el trabajo de campo y la recopilación de datos. Los datos obtenidos tienen que ser codificados y procesados, y este proceso de datos es una parte importante en los estudios de evaluación de la dieta, especialmente si los datos de consumo de alimentos se tienen que utilizar para estimar la ingesta de nutrientes. Los programas informáticos adecuados permiten introducir, gestionar y procesar grandes volúmenes de datos con el propósito de preparar una matriz numérica lista para su análisis estadístico. En este paso, la identificación y extracción de valores atípicos (*outliers*) y el estudio de los valores desconocidos son esenciales para evitar problemas en el análisis estadístico. Se han desarrollado diferentes métodos para tratar los valores extremos y los valores perdidos, y la elección correcta del método es un punto crítico.

El propósito de este artículo es facilitar algunas recomendaciones sobre el tratamiento de los valores perdidos y los valores atípicos, y algunas orientaciones con respecto a los programas informáticos existentes para el cálculo de tamaños muestrales y para la realización de análisis estadístico. Estos programas pueden ayudar a prevenir posibles resultados incorrectos en los análisis estadísticos y la mala interpretación de las observaciones realizadas.

Recolección de datos. Recomendaciones

La recolección de los datos es una parte importante de la investigación. La información recogida y cómo se miden las variables condiciona el análisis estadístico posterior y la validez del estudio. Por eso, se recomienda que se registre la información de los sujetos de forma original en lugar de variables calculadas o categorizadas. Por ejemplo, en lugar de registrarse la edad del sujeto se recomienda recoger el año de nacimiento o el estado nutricional no debería de ser registrado en categorías como "bajo peso", "normal", "sobrepeso" y "obesidad". Es mejor pedir el peso y la altura y luego calcular el índice de masa corporal y de ahí generar las categorías de peso. Lo mismo ocurre con la frecuencia de la ingesta de alimentos. Si la variable frecuencia de consumo se registra utilizando las categorías "menos de 2 veces por semana", "entre 2 y 5 veces" y "5 o más veces a la semana", posteriormente es imposible saber el número de ingestas que ha realizado un sujeto de un alimento y, por lo tanto, la variable no podrá modificarse.

Datos perdidos y atípicos

Datos perdidos

Una vez se han registrado los datos es importante tener en cuenta la información no proporcionada por los sujetos, es decir, los datos faltantes o perdidos. Rubin (1976)[4] clasifica los datos perdidos en tres tipos: datos perdidos completamente al azar (MCAR = *missing completely at random*), datos perdidos al azar (MAR = *missing at random*) y datos perdidos no debidos a azar (NMAR = *not missing at random*). Se considera que los datos perdidos son MCAR cuando la probabilidad de qué un sujeto presente un valor ausente en una variable no depende ni de la propia variable ni de ninguna otra variable recogida. En cambio, los datos perdidos se consideran MAR cuando la probabilidad de no observar un dato depende de otras variables pero no de los valores de la variable con valores perdidos. Finalmente, los datos perdidos se consideran de tipo NMAR cuando la probabilidad de que un sujeto presente un valor faltante depende de dicha variable con valores perdidos.

Por ejemplo, cuando se registra el índice de masa corporal según el sexo de los sujetos, si no existe ninguna razón en particular de porque un sujeto no informa de su peso, entonces los datos faltantes se consideran MCAR. Sin embargo, si es más probable que las mujeres no nos revelen su peso, estos datos perdidos dependen del sexo y se consideran datos MAR. Y finalmente, en el caso de que los sujetos obesos sean más propensos a no revelar su peso, la probabilidad de que el índice de masa corporal presente datos perdidos depende de la propia varia-

Tabla I
Datos de 12 estudiantes. Los valores perdidos de la variable IMC se muestran mediante casillas vacías

Sexo	*Índice de obesidad*	*IMC*	*Energía*	*Prot.*	*Imputación media IMC*	*Imputación regresión IMC*
2	80		901,8	53,1	21,33	27,15
2	78		3.197,2	177,6	21,33	24,30
2	72		2.295,5	96,0	21,33	19,89
2	65		2.229,8	113,6	21,33	19,55
2	62		2.131,1	79,0	21,33	21,18
2	82		2.137,9	125,4	21,33	27,17
2	80		1.453,3	69,7	21,33	22,25
2	63		2.927,2	124,4	21,33	19,55
2	69	20,10	2.684,6	104,8	20,10	20,10
1	58	23,78	2.681,9	144,5	23,78	23,78
2	78	22,12	2.677,3	136,4	22,12	22,12
1	67	21,29	2.674,6	127,5	21,29	21,29

IMC: Índice de masa corporal; Energía: Ingesta total de energía; Prot.: Ingesta de proteína.

ble, estos datos perdidos no son debidos al azar sino a la propia variable que se pide información; son por tanto NMAR.

Ejemplo ilustrativo

Se ha seleccionado una muestra aleatoria de 58 estudiantes de los grados de Nutrición Humana y Dietética y Ciencia y Tecnología de los Alimentos de la Universidad de Barcelona para evaluar su estado nutricional. La ingesta de alimentos de los estudiantes se ha recogido mediante un recordatorio de 24h y un registro de 3 días y un cuestionario de frecuencia de alimentos. Los estudiantes también han cumplimentado el cuestionario de hábitos y estilos de vida y obesidad[5]. El objetivo principal es estudiar la relación entre la ingesta de proteínas con el género, el índice de masa corporal (IMC), la ingesta de energía total y el índice de sobrepeso y obesidad. Ocho mujeres no han informado de su peso y altura, por tanto el IMC presenta ocho valores perdidos (tabla I).

Eliminación de los casos

Hay dos formas de eliminar los datos perdidos: eliminación de los casos (*listwise*) o eliminación por pares (*pairwise*). En la eliminación de los casos el sujeto con datos perdidos se eliminan del análisis. Si los datos son MCAR, este tipo de eliminación no presenta sesgo, pero el tamaño de la muestra se reduce y por tanto puede afectar a la potencia de los contrastes de hipótesis (disminuyendo) o al error estándar de la estimación (incrementando). Además, este método descarta la otra información proporcionada por el sujeto.

En la eliminación por pares (o análisis de los casos disponibles) se elimina el sujeto del análisis cuando los datos son perdidos en la variable que se precisa para el análisis, pero se incluye el sujeto en los análisis en los que se disponga información. Cuando se utiliza la eliminación por pares, el tamaño de la muestra a analizar no es consistente en todas las estimaciones realizadas.

En la tabla II, se muestra las estimaciones de los coeficientes del modelo lineal del logaritmo de las proteínas

Tabla II
Estimación del modelo de regresión lineal entre el logaritmo de las proteínas y el sexo, IMC, índice de obesidad e ingesta total de energía usando método eliminación de casos (panel izquierdo) e imputación múltiple (panel derecho)

	Eliminación de los casos			*Imputación múltiple*			
Variable	*Beta*	*Error Std.*	*P valor*	*Beta*	*Error Std.*	*P valor*	*TIF*
Mujeres	-0,245	0,089	< 0,001	-0,240	0,085	< 0,001	0.003
IMC	-0,003	0,092	0,009	-0,020	0,013	0,008	0,109
Índice Obesidad	-0,002	0,003	0,001	0,006	0,003	0.001	0,036
Energía total	0,0003	0,00005	< 0,001	0,0003	0,0004	< 0,001	0,017
8 observaciones con datos perdidos eliminadas				Sin datos perdidos			

Error Std.: Error estándar; TIF: Tasa de información faltante.

consumidas en función del sexo, IMC, del índice de obesidad y la ingesta total de energía. La información de los estudiantes con datos faltantes en IMC se ha eliminado (eliminación de los casos).

Imputación simple o múltiple

La imputación es un proceso de reemplazar los datos perdidos por estimaciones. Existen varios métodos: imputación mediante la media, imputación mediante regresión, imputación mediante el algoritmo de esperanza-maximización e imputación múltiple.

El método de imputación mediante la media consiste en reemplazar los datos perdidos por la media de los datos no perdidos. Si aplicamos este método a nuestros datos, todos los datos perdidos son reemplazados por la media del IMC (21,32). El problema de este tipo de imputación es que puede atenuar cualquier correlación entre las variables que se han imputado valores. En la tabla 1 se muestra los valores del IMC utilizando este método.

En la imputación mediante regresión, los datos perdidos son reemplazados por el valor predicho de la regresión que se deriva de los datos. En contraste con la imputación de la media, el valor imputado está condicionado a la información que se dispone de los sujetos. Teniendo en cuenta el ejemplo de los datos perdidos en la variable IMC, con la imputación de la media todos los datos perdidos son reemplazados por el mismo valor (la media del IMC). Sin embargo, con la imputación mediante regresión se pueden reemplazar los datos perdidos por los valores predichos del IMC según el sexo, el sobrepeso y la obesidad total e ingesta de energía de los estudiantes. En la tabla 1 se muestran los valores de IMC reemplazados por imputación mediante regresión. Cada estudiante tiene un valor predicho de IMC diferente según su índice y el consumo total de energía. Por lo tanto, hay una mejoría al comparar la imputación por regresión con la imputación por la media, pero el valor predicho con la regresión tiene un error que no es considerado al realizar la imputación. Sin embargo, esta dificultad puede superarse mediante la imputación de regresión estocástica. Este enfoque añade un término aleatorio residual de la distribución normal (u otra) para cada valor imputado

Otra manera de tratar con datos perdidos es la técnica llamada el algoritmo de expectación-maximización[6] (*EM algorithm*). Este método asume una distribución de los datos perdidos parcialmente y la inferencia se basa en la verosimilitud bajo esta distribución. Es un proceso iterativo, en el cual se repiten los dos pasos siguientes hasta convergencia. En el paso *E* se calcula la expectativa condicional de los datos perdidos, condicionado a los valores observados y las estimaciones actuales de los parámetros. Entonces estas expectativas se imputan a los datos perdidos. En el paso *M*, se calculan las estimaciones máximo-verosímiles de los parámetros. No obstante, este método no considera a la incertidumbre de los datos perdidos.

En la imputación múltiple[7] en lugar de imputar un valor único para cada dato perdido, cada uno de ellos se sustituye por *m* datos simulados que representa la incertidumbre del valor a imputar. Entonces cada imputación genera un conjunto de datos diferentes los cuales se analizan por separado, obteniéndose *m* estimaciones y sus errores estándar. La estimación global es el promedio de todas las estimaciones. El error estándar de la estimación se realiza calculando la varianza intra-imputaciones, promedio de los errores estándar *m*, así como la varianza entre los imputaciones, varianza muestral de las *m* estimaciones. Se suman estas dos varianzas y su raíz cuadrada determina el error estándar de la estimación. Mediante este método se introduce la incertidumbre de los datos perdidos en el error estándar de la estimación. La varianza entre las *m* estimaciones también refleja incertidumbre estadística debido a los datos perdidos.

En nuestros datos se realizaron 15 imputaciones para cada valor faltante. Así, tenemos 15 conjuntos de datos según los valores de la imputación. La tabla II muestra la estimación de los coeficientes de la regresión del logaritmo de las proteínas según sexo, índice de masa corporal. En este caso todos los estudiantes se han utilizado porque el IMC se ha imputado. La tasa de información faltante cuantifica el aumento relativo de varianza debido a los datos perdidos del IMC. El índice de masa corporal tiene una tasa de 0,109 y las variables restantes tienen una tasa muy baja porque no se ha realizado ninguna imputación.

Si tenemos datos perdidos del tipo MCAR, entonces no hay ningún sesgo en los datos y si además son unos pocos casos entonces una buena opción es elegir el método de la eliminación *listwise*. Si los datos son MAR, la mejor solución es la imputación múltiple. La imputación por máxima verosimilitud y la imputación por regresión estocástica también son adecuadas, pero se recomienda la imputación múltiple. Si los datos son NMAR entonces estos métodos a menudo están sesgados y existen métodos específicos para esta tipo de datos[7].

Valor atípico

Un valor atípico es una observación claramente diferente del resto de datos, es una observación extrema. Hay varios métodos para detectar valores atípicos: gráficos como los gráficos de normalidad, diagrama de cajas o métodos basados en distribuciones

Los métodos basados en distribuciones se asume que los datos provienen de una distribución Normal. Existen varios test como la prueba de *Grubbs* para valores atípicos [8], el criterio de *Pierce*[9], or la prueba Q de *Dixon*[10]. Un método común para la detección de valores atípicos es mediante el rango intercuartílico. Una observación se considera atípica si está fuera de los limites y ; k se fija normalmente a 1,5 o 3.

Es importante estudiar los datos atípicos porque la mayoría de los procedimientos estadísticos están influenciados por estos datos y no son robustos. Por

ejemplo, la media es sensible a las observaciones extremas, y la mediana no. Supongamos que tenemos 10 estudiantes que tienen un consumo de proteínas entre 50 y 160 g/día pero hay uno que tiene un consumo de 250 g/día. La media es de 166 g/día en cambio la mediana es de 81 g/día.

En el análisis de regresión los valores atípicos también pueden influir en los resultados. En la regresión se diferencia entre valores atípicos y observaciones que tienen una alta influencia (*leverage*). Concretamente, un valor atípico es una observación extrema en la variable respuesta. Sin embargo una observación que tiene un valor de X muy lejos de su media puede ser un punto altamente influyente.

El *leverage* o influencia mide la distancia del punto a la media de la distribución de la X. Cuando el *leverage* es dos o tres veces superior que la media del *leverage*, (p+1)/n, se considera que el punto tiene un alto *leverage*, siendo p el número de parámetros de regresión y n el tamaño de la muestra.

Datos con alta influencia y valores atípicos pueden tener una influencia potencial en la regresión, generando un impacto negativo porque pueden sesgar las estimaciones. Por otra parte, no todos los puntos con una alta influencia o valores atípicos influencian en la estimación de los coeficientes. Es posible por ejemplo tener una observación con alto *leverage*, pero estar alineada con el patrón del resto de los datos y por tanto no generar un impacto negativo en los resultados.

En la figura 1, se muestra el peso y la altura de 60 sujetos. Las variables presentan una relación lineal. Hay tres puntos que se han añadido al gráfico (A, B y C). A es un valor atípico respecto la altura pero no respecto el peso. Su *leverage* es bajo (0,016) porque es inferior a 2*(2/61) = 0,06. B es un valor atípico respecto el peso y tiene un *leverage* alto, y C no es un valor atípico respecto la altura pero tiene un alto *leverage*.

Se puede realizar un análisis preliminar para detectar valores extremos mediante los residuos del modelo). Un problema con los residuos es que sus valores dependen de las unidades de medida utilizados. Puesto que los residuos están en las unidades de la variable dependiente, Y, no disponemos de unos puntos de corte para definir un residuo grande. Este problema se puede solucionar mediante el uso de residuos estandarizados, que se calculan dividiendo el residuo por su error estándar.

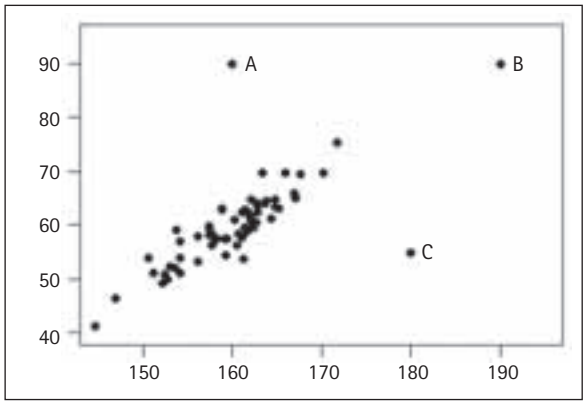

Fig. 1.—Diagrama de puntos de peso y altura con tres potenciales observaciones influyentes (A, B y C).

Las observaciones con residuos estandarizados en valor absoluto superiores a 3 deben ser considerados como potenciales valores atípicos. Los puntos A, B y C en la figura 1 tienen un residual estandarizado igual a 6,39, -0,26 y -5,76 respectivamente. Los puntos A y C tienen un valor residual alto, sin embargo B tiene un residuo pequeño pero en cambio es un valor atípico con un alto *leverage*.

La distancia de Cook[11] es una medida para detectar observaciones potencialmente influyentes. La distancia mide el efecto de la eliminación de una observación. Los puntos de datos con un valor residual alto (aislados) y / o un *leverage* alto pueden distorsionar la estimación y la precisión de las estimaciones del modelo de regresión. En el caso de existir puntos con una distancia de un Cook grande, (distancias superiores a 1) se recomienda el estudio de su influencia. Otra regla común es considera el umbral el percentil 1-alfa de la distribución de Fisher Snedecor (F (p, n-p, 1-alfa)).

Los puntos A, B y C tienen una distancia de Cook de 0,34, 0,02 y 4,25, respectivamente. Aunque A es un valor atípico, no es una observación con una influencia potencial alta. Sin embargo, como se muestra en la tabla III, el error estándar de los coeficientes aumentó y la bondad de ajuste (coeficiente de determinación) disminuyó de 0,82 a 0,58 (tabla III). Los puntos B y C tienen un alto *leverage* pero solamente C es un potencial valor influyente. En la tabla III muestra que las estimaciones, los errores estándar y los coeficientes de determinación cuando el punto C se agrega a los datos.

Tabla III
Estimación de los coeficientes de regresión, el error estándar y el coeficiente de determinación de la regresión entre el peso y la altura en función de las observaciones A, B o C incluidas

	Constante	*Error Std.*	*Beta*	*Error Std.*	R^2
Sin A, B y C	-106,4	10,13	1,04	0,06	0,82
Observación A	-106,8	18,10	1,04	0,11	0,58
Observación B	-104,9	8,17	1,03	0,05	0,87
Observación C	-68,6	13,69	0,80	0,08	0,60

Regresión lineal: peso = constante + beta*altura; R^2: coeficiente de determinación.

Existen otros métodos de diagnóstico para detectar observaciones potencialmente influyentes que son: los estadísticos DFBETAS, DFFITS y COVRATIO[12]. Todos miden el impacto al eliminar una observación del análisis. Concretamente DFBETAS mide el efecto sobre la estimación de los coeficientes, DFFITD sobre el valor predicho y COVRATIO sobre las varianzas (error estándar) de los coeficientes de regresión y sus covarianzas.

Cuando se detecta un valor atípico, primero se debería evaluar su procedencia. Si el valor procede de un error humano o del instrumento de medida entonces el error debe ser corregido. Sin embargo, pueden surgir datos atípicos por diferentes causas tales como la variabilidad inherente de la variable o si la distribución subyacente tiene una distribución asimétrica o porque es un dato que proviene de otra población. Alternativamente, valores atípicos pueden sugerir que deben ser incluidos en el análisis de regresión variables explicativas adicionales. La eliminación de datos atípicos es una práctica controvertida y en lugar de omitirlos se recomienda el uso de métodos estadísticos robustos los cuales no están excesivamente afectados por valores atípicos.

Software estadístico

Paquetes para el cálculo tamaño de muestra

Una vez definido el objetivo y el tipo de diseño, es importante calcular el número de sujetos a estudiar. La muestra ha de ser representativa de la población estudiada. En función del objetivo del estudio y de la estructura de la población, existen varios tipos de muestreos: muestreo aleatorio simple, muestreo sistemático, muestreo estratificado o muestreo por conglomerados. Además, el cálculo del tamaño de la muestra depende del objetivo principal y de si se requiere trabajar con una precisión mínima de las estimaciones o con una potencia prefijada. Es importante también considerar un porcentaje extra de sujetos porque podemos tener valores perdidos. Es difícil recomendar una cantidad de porcentaje de individuos que no responderán y básicamente depende del área de estudio. También es conveniente diseñar estrategias para garantizar o controlar que los sujetos responden a toda la información del cuestionario.

Para calcular el tamaño de las muestras están disponibles muchos programas comerciales o libres. En relación con el software libre, EPIDAT 4.0 y GRANMO permiten calcular el tamaño de la muestra según la metodología estadística que se va a usar.

El software EPIDAT 4.0 fue creado por *Servizo de Epidemioloxía de la Dirección Xeral de Innovación e Xestión da Saúde Pública de la Consellería de Sanidade* (Xunta de Galicia) con el apoyo de la Organización Panamericana de la Salud y la Universidad CES de Columbia. Puede descargarse desde la página http://www.sergas.es/ en la sección de investigación e innovación sanitaria/datos/ Software.

El software GRANMO fue desarrollado por el *Program of Research in Inflammatory and Cardiovascular Disorders y el Institut Municipal d'Investigació Mèdica*, Barcelona, España. Puede descargarse desde la página web http://www.imim.cat.

Software de análisis estadístico

Hay una gran variedad de programas estadísticos disponibles. En la actualidad, el software libre más comúnmente utilizado es el *R-project*[13]. Es un proyecto GNU que fue desarrollado en Bell por John Chambers y sus colegas. Se compila y ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS. R es un lenguaje interpretado. R tiene un intérprete de línea de comandos y dispone de varios paquetes que los usuarios pueden descargar de la página web. El proyecto R no tiene una interfaz gráfica de usuario (GUI) amigable pero algunos paquetes como *R commander*[14] o *Deducer*[15] proporcionan una GUI basada en menús.

De software tipo comercial también existe una gran variedad como: *S-plus 16* (versión comercial del proyecto R), *SPSS*[17] (Statistical Package for the Social Sciences), *SAS institute*[18] (sistema de análisis estadístico), *STATA* [19] (*Statistics and Data*) o *Minitab*[20].

La imputación múltiple se ha vuelto cada vez más popular, y todos estos softwares permiten aplicar esta técnica. Yucel (2011)[21] proporciona una descripción de la metodología de imputación implementada por varios softwares.

Finalmente, todos estos programas realizan una amplia variedad de análisis estadísticos y tienen gran poder para generar gráficos de los resultados. La opción más adecuada del software por lo tanto depende de los costos y las preferencias de cada usuario.

Conflictos de intereses

Los autores declaran que no hay ningún conflicto de intereses con respecto a la publicación de este documento.

Referencias

1. Willet W. Nutritional epidemiology. 2nd ed. Oxford: Oxford University Press; 1998.
2. Arab L. Analyses, presentation, and interpretation of results. In: Cameron ME, Van Staveren W, editors. Manual on methodology for food consumption studies. Oxford: Oxford University Press; 1988.
3. Polgar S, Thomas SA. Introducción a la investigación en ciencias de las salud. Madrid: Churchill Livingstone; 1993.
4. Rubin, D.B. Inference and missing data. *Biometrika* 1976; 63: 581-92.
5. Pardo A, Ruiz M, Jódar E, Garrido J, De Rosendo J M, Usán L A. Development of a questionnaire for the assessment and quantification of overweight and obesity related lifestyles. *Nutrición Hospitalaria* 2004; XIX (2): 99-109.

6. Dempster, AP, Laird NM, Rubin DB. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion) 1977. *Journal of the Royal Statistical Association* 1977; B39: 1-38.
7. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 1987.
8. Grubbs FE. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics* 1950; 21 (1): 27-58.
9. Pierce B.Criterion for the Rejection of Doubtful Observations. *Astronomical Journal II* 1852; 45: 161-310.
10. Dean RB and Dixon WJ. Simplified Statistics for Small Numbers of Observations. *Anal Chem*1951; 23 (4): 636-8.
11. Cook R D Influential Observations in Linear Regression. *Journal of the American Statistical Association* 1979; 74 (365): 169-74.
12. Belsley DA, Kuh E, Welsh RE. Regression diagnostics: identifying influential data and sources of collinearity. Wiley series in probability and mathematical statistics. New York 1980.
13. R-project version 3.1.2. Dowload from: http://www.r-project.org/
14. Fox J, Bouchet-Valat M *et a*l. A platform-independent basic-statistics GUI (graphical user interface) for R, based on the tcltk package. Version 2.1-5.2014.
15. Fellows I *et al.* Deducer: A data analysis GUI for R. Version 0.7-7. 2014.
16. S-PLUS. TIBCO Software Inc.2014.
17. IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.
18. Statistical Analysis System (SAS) Institute Inc. 2013. Unitate States. Version 9.4.
19. StataCorp. 2013. *Stata Statistical Software: Release 13.* College Station, TX: StataCorp LP.
20. Minitab 17 Statistical Software (2010). [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com)
21. Yucel RM. State of the Multiple Imputation Software. *Journal of statistical software* 2011;45(1).