Introducción

La expansión de datos y la ponderación de datos a menudo se consideran parte del mismo proceso, pero en realidad deberían distinguirse entre sí. La expansión de datos es simplemente el procedimiento de multiplicar cada observación en los datos por un factor que representa cuántos miembros de la población están representados por esa observación. La ponderación de datos es el procedimiento de desarrollar factores de multiplicación que intentan corregir los sesgos en el diseño de la muestra que se han introducido, ya sea intencionalmente o no, en el proceso de muestreo y encuesta. Tanto la expansión como la ponderación de datos tienen como objetivo principal proporcionar una estimación lo más precisa posible de las estadísticas de población a partir de una muestra.

Hay muchos casos en la recopilación de datos en los que no hay interés en producir estadísticas de población a partir de los datos. En muchos casos de investigación y práctica que tienen como objetivo desarrollar modelos estadísticos de un proceso, utilizando datos recopilados de una muestra, puede que no haya necesidad o interés de ampliar los datos a toda la población. De hecho, es una propiedad de varios procedimientos de modelización que, siempre que se utilice una especificación apropiada del modelo, los sesgos en los datos no afectarán a la mayoría de los parámetros estimados del modelo, de modo que ni la ponderación ni la expansión de los datos son necesarias. necesario. Esto es cierto, por ejemplo, para el modelado de elección utilizando el método llamado análisis logit multinomial (Manski y Lerman, 1977; Stopher y Meyburg, 1979), en el cual, siempre que se especifique un conjunto completo de constantes alternativas específicas, los coeficientes de Todas las variables del modelo serán estimadores insesgados, incluso si los datos de entrada están significativamente sesgados. Además, en este caso particular, las constantes alternativas específicas que se estiman incorrectamente pueden corregirse mediante un proceso simple (Manski y Lerman, 1977).

Sin embargo, hay una serie de situaciones en las que la ampliación de los datos y la corrección mediante ponderación son deseables, si no necesarias. Los propósitos principales incluyen:

(1) comprobar la validez de los datos de la muestra;

(2) proporcionar comparaciones con el censo y otros datos;

(3) proporcionar comparaciones con encuestas anteriores;

(4) eliminar preocupaciones sobre posibles sesgos o proporcionar detalles sobre posibles sesgos y cómo superarlos;

(5) proporcionar datos descriptivos sobre la población; y

(6) proporcionar perfiles de grupos y subgrupos de población.

Ampliación de datos

Para poder ampliar una encuesta por muestreo a la población de la que se extrajo la muestra, existen dos requisitos principales. Primero, la población debe ser finita y debe conocerse su tamaño. En segundo lugar, la muestra debe haber sido extraída mediante un procedimiento de muestreo aleatorio o una aproximación cercana a un procedimiento de muestreo aleatorio. Si no se cumple una o ambas condiciones, no es posible ampliar los datos.

Muestreo aleatorio simple

El muestreo aleatorio simple es el procedimiento más sencillo para ilustrar la expansión de datos. El MAS sólo se puede llevar a cabo cuando la población es finita y conocida. Además, el propio MAS cumple el segundo criterio para la ampliación de datos. Si los datos se han muestreado a una tasa de muestreo de $f(= n/N)$, entonces el factor de expansión para cada observación en la muestra es simplemente el inverso de la tasa de muestreo, generalmente escrito como $g(= 1/f = N/n)$. Por tanto, supongamos que se ha muestreado una población de 1 millón de hogares con una tasa de muestreo del 1 por ciento, entonces la muestra consta de 10.000 hogares. Por lo tanto, cada hogar representa 100 hogares en la población total, y 100 es el factor de expansión que se aplicaría a todas y cada una de las observaciones de la muestra.

Muestreo estratificado

En el muestreo proporcional (muestreo estratificado con una fracción de muestreo uniforme), el factor de expansión es el mismo que en el muestreo aleatorio simple, porque cada observación de la muestra se muestrea al mismo ritmo que cualquier otra observación y solo hay una fracción de muestreo y, por lo tanto, una factor de expansión. En el muestreo desproporcionado (muestreo estratificado con una fracción de muestreo variable), cada estrato se muestrea a una frecuencia de muestreo diferente. Por lo tanto, habrá una tasa de muestreo dentro de cada estrato y cada estrato tendrá su propio factor

de expansión distintivo. Por tanto, cada observación del conjunto de datos recibirá un factor de expansión que está determinado simplemente por su pertenencia a un estrato.

Por ejemplo, supongamos que una población se estratifica en cinco estratos, con tasas de muestreo del 1 por ciento, 2 por ciento, 0,5 por ciento, 1,5 por ciento y 4 por ciento. Entonces, los factores de expansión de los cinco estratos serán $100;\ 50;\ 200; 66.6667$ y 25, respectivamente. En otras palabras, cada observación de muestra en el primer estrato representará a 100 miembros de la población del estrato 1, cada observación de muestra en el segundo estrato representará a cincuenta miembros de la población del estrato 2, y así sucesivamente. Cuando los factores de expansión se aplican correctamente, el número total de miembros de la muestra ampliada será igual al número total de miembros de la población, aparte de cualquier error de redondeo en el cálculo de los factores de expansión.

Muestreo multietapa

En el muestreo multietapa, habrá diferentes fracciones de muestreo en las diferentes etapas de la muestra, y también se pueden utilizar diferentes métodos de muestreo, de modo que algunas etapas pueden usar un muestreo desproporcionado, mientras que otras pueden usar un muestreo proporcional o un muestreo aleatorio simple. En este caso, es posible calcular una fracción de muestreo en cada etapa del muestreo y, por tanto, también un factor de expansión. El factor de expansión final para cada observación será el producto de los factores de expansión para cada etapa del muestreo.

Por ejemplo, supongamos que en la primera etapa se extrae una muestra desproporcionada, utilizando cinco estratos con tasas de muestreo del 2 por ciento, 5 por ciento, 10 por ciento, 20 por ciento y 25 por ciento, entonces los factores de expansión para esta etapa serán $50; 20; 10; 5$ y 4, respectivamente. Supongamos que en la segunda etapa se extrae una muestra aleatoria simple a una tasa del 0,1 por ciento. En esta etapa, el factor de expansión es, por tanto, 1.000 para todas las observaciones. Esto significa que el factor de expansión para las observaciones que están en el estrato 1 de la muestra de primera etapa es $50.000 (= 50 \times 1.000)$, mientras que el de las unidades muestrales del estrato 2 de la primera etapa será 20.000 , y así sucesivamente.

Luego se pueden realizar cálculos similares para más etapas en una muestra de varias etapas. Es una simple cuestión de multiplicar los factores de expansión de etapas

sucesivas, asegurándose de mantener los factores de expansión correctos para cada etapa, especialmente cuando se utiliza la estratificación.

Muestras de conglomerados

Los factores de expansión para las muestras por conglomerados son similares a los del muestreo multietápico, con excepción de la etapa final. Debido a que la etapa final en una muestra de conglomerados es un censo de las unidades en los conglomerados, no se produce ninguna expansión adicional en esta etapa. Supongamos que se va a extraer una muestra nacional de hogares de una nación que no tiene una lista de direcciones de hogares que puedan usarse para el muestreo, utilizando una técnica de muestreo por conglomerados. Se decide tomar una muestra primero de los condados de la nación, utilizando una muestra aleatoria simple de condados. Dentro de los condados se extraerá una muestra estratificada de municipios, y dentro de los municipios se extraerá una muestra aleatoria simple de bloques residenciales. Se encuestarán todos los hogares de cada bloque muestreado. Supongamos que la tasa de muestreo para los condados es una décima parte y que las tasas de muestreo para los estratos de municipios son 2 por ciento, 3,3333 por ciento, 1 por ciento y 0,5 por ciento para los cuatro estratos. La tasa de muestreo para bloques es uno en 100 o 1 por ciento. La tasa de muestreo para los hogares dentro de los bloques es, por supuesto, del 100 por ciento. Para el primer estrato de corregimientos, el factor de expansión será $10 \times 50 \times 100(\times 1) = 50{,}000$. Para el segundo estrato, es$10 \times 30 \times 100(\times 1) = 30.000$. Se aplicarían cálculos similares a los otros dos estratos, utilizando factores de expansión de 100 y 200 para las segundas etapas de cada uno de esos dos estratos, respectivamente.

Tenga en cuenta que el factor de expansión de la etapa final es 1 en cada caso, porque se muestra todo el conglomerado.

Otros métodos de muestreo

Otros métodos de muestreo siguen el mismo principio que se describe aquí para los principales métodos de muestreo probabilístico. Por ejemplo, si se extrae una muestra sistemática, a razón de uno entre cincuenta, entonces el factor de expansión es simplemente 50 para cada observación. Este será el caso incluso si se utilizan diferentes puntos de partida a lo largo del muestreo, para introducir cierta apariencia de aleatoriedad en el muestreo. En el caso de una muestra sistemática de una población que aún no está

enumerada, como el muestreo de compradores que llegan a un centro comercial, en el que, por ejemplo, se contacta a uno de cada veinte compradores para la encuesta, puede ser necesario realizar un recuento del total. población de la que se extrae la muestra para que la ampliación se realice correctamente.

Sin embargo, es importante señalar que la expansión es sólo una parte de la historia en las encuestas de población humana, así como en algunas otras encuestas, en las que hay cualquier forma de incumplimiento del diseño de la muestra, como no completar una encuesta con un grupo seleccionado. unidad de muestreo. Tal incumplimiento del diseño de muestreo requiere ponderar los datos. También es importante señalar que la expansión de datos no es posible si el muestreo proviene de una población indefinida, es decir, una población que tiene un tamaño desconocido.

Ponderación de datos

Como se señaló previamente, la expansión de datos se refiere correctamente simplemente al proceso de estimar estadísticas de población a partir de estadísticas muestrales, con base en el diseño muestral. La ponderación es el proceso necesario para intentar corregir los sesgos en los datos. El hecho más común que requiere ponderación es la falta de respuesta, que es, con diferencia, la causa más común de incumplimiento del diseño muestral en las encuestas con sujetos humanos.

Cuando se va a realizar la ponderación de una muestra, debe realizarse para características específicas de la población que sean características clave o variables criterio para las cuales el sesgo se considera inaceptable, o que se consideren las características más útiles frente a las cuales se puede ponderar una muestra. para comprobar si hay sesgos. Por ejemplo, un hecho común en las encuestas entre personas es que las personas de hogares de diferentes tamaños tendrán diferente disposición a responder una encuesta. A menudo, quienes viven en hogares pequeños (principalmente unipersonales) y quienes viven en hogares grandes (con cinco o más miembros) tienen menos probabilidades de responder; el primero debido a la preocupación por la invasión de la privacidad y la vulnerabilidad creada por la información recopilada por la encuesta, y el segundo porque la encuesta puede representar una carga mayor para los hogares más grandes. Por lo tanto, el primer paso en la ponderación es decidir qué características de la población se ponderarán. para ser llevado a cabo. No es necesario restringir la ponderación a una sola característica, aunque los procedimientos para ponderar la muestra se vuelven cada vez más complejos

a medida que se utilizan más características. Lo más común es que las características que se eligen sean aquellas que también están disponibles en alguna fuente de datos secundaria, como un censo de población. Esto se debe a que la comparación entre la muestra y la fuente secundaria proporciona una indicación de la posibilidad de sesgo proporciona el medio más sencillo para ponderar la muestra. Sin embargo, no es cierto que las características sean siempre las disponibles en fuentes de datos secundarias. Es posible que las características más importantes a utilizar para pesar una muestra en particular no estén disponibles en otras fuentes. Por lo tanto, pueden existir dos situaciones alternativas en las que es necesario ponderar la falta de respuesta: primero, los totales reales de población para las características seleccionadas no se conocen a partir de otras fuentes; y, en segundo lugar, los totales reales de población para las características seleccionadas se conocen a partir de otras fuentes de datos.

Ponderación con totales de población desconocidos

Cuando no se conocen los totales de la población sobre las características clave para la ponderación, es necesario mantener un registro durante toda la encuesta de todos los que no respondieron. Los que no respondieron se dividen en dos categorías principales. En primer lugar, están los no encuestados que rechazan la encuesta antes de que se responda cualquier pregunta y, por lo tanto, no proporcionan información alguna (normalmente denominados “rechazos”). Normalmente se requiere un recuento de los que no respondieron. En segundo lugar, están los que no responden y responden algunas preguntas, pero no responden suficientes preguntas para ser considerados respuestas completas (normalmente denominados “terminaciones”). No sólo es necesario un recuento de los que no respondieron, sino que las respuestas a las preguntas a las que sí responden deben registrarse para su uso en la ponderación. De hecho, en el caso en que los totales de población no se conocen por las características en cuestión, estas terminaciones son una fuente muy valiosa de información importante sobre los sesgos que pueden existir en la encuesta, y generalmente son la única fuente de ese tipo.

Diferentes métodos de encuesta surgirán con respecto a quienes no respondieron. En el caso de las encuestas autoadministradas, como las postales o por Internet, lo más frecuente es que solo haya rechazos, aunque puede haber un pequeño número de casos en los que los encuestados completaron parcialmente la encuesta y aun así la enviaron por

correo o lo envió en línea. En las encuestas a los entrevistadores, ya sea por teléfono o cara a cara, generalmente habrá una combinación de negativas y despidos. Considerando este aspecto de la metodología de la encuesta, conviene considerar dos casos. En el primer caso, sólo hay negativas y no hay despidos. En tal caso, cuando no existen datos complementarios sobre las características para la ponderación, en realidad no hay ponderación posible. El único supuesto que se puede hacer es que los que no respondieron se distribuyen de manera idéntica a los encuestados y, por lo tanto, no es posible realizar ninguna ponderación.

Si hay terminaciones en el conjunto de datos, entonces la posible suposición es que los rechazos y las terminaciones se distribuyen de manera idéntica con respecto a las características de la población que se utilizarán para la evaluación del sesgo y la posible ponderación de los datos.

Ahora la prueba adecuada es determinar si las terminaciones tienen una distribución diferente a la de los datos completados sobre las características clave. Si las distribuciones son las mismas, entonces no hay sesgo detectado por las variables clave. Si las distribuciones son diferentes, entonces hay evidencia prima facie de sesgo y se debe realizar una ponderación.

Suponiendo que se encuentra que las terminaciones tienen una distribución diferente en las características clave y que, por lo tanto, la ponderación es necesaria, entonces el procedimiento para ponderar los datos es el siguiente. Para cada característica que se haya seleccionado para evaluar el sesgo, determine la distribución de las terminaciones y de las encuestas completadas. Combine estos dos para formar una distribución compuesta y estime las proporciones de las terminaciones combinadas y completas para cada combinación de categorías de las características clave. Calcule la relación entre la proporción combinada y la proporción solo en las encuestas completas. Esta relación es el peso apropiado por el cual multiplicar las observaciones dentro de la encuesta completa. Un ejemplo es útil para ilustrar el procedimiento.

Un ejemplo

Supongamos que se ha realizado una encuesta de hogares mediante entrevistas telefónicas. Supongamos que la característica clave que se utilizará para evaluar el sesgo es la del tamaño del hogar. Se supone aquí que la pregunta sobre el tamaño del hogar se formuló al principio de la encuesta, de modo que la mayoría de los hogares que terminaron respondieron a esta pregunta. Por supuesto, las negativas rotundas no proporcionan ningún dato sobre el tamaño del hogar. Se supone que los resultados de la encuesta están distribuidos como se muestra en la Tabla 1.

Del examen de este cuadro se desprende claramente que las encuestas completadas se distribuyen de manera diferente a las encuestas terminadas, con una proporción mucho mayor de terminaciones en hogares unipersonales y una tasa de terminaciones mucho menor en hogares de tres personas. Como resultado, parece ser necesaria una corrección del sesgo. Esto se hace calculando los pesos como se muestra en la Tabla 2.

En la Tabla 2, la fila superior es la suma de las filas primera y tercera de la Tabla 1 y la segunda fila es el porcentaje que estas representan del total de los terminados y terminados. La tercera fila es una repetición de la segunda fila de la Tabla 1 y la cuarta fila es la relación entre la segunda fila y la tercera fila y representa el peso. Supongamos que el factor de expansión para los hogares de tamaño 1 fuera 100, entonces el factor de expansión ponderado sería el producto de 100 y 1.22, o 122. Esto explica la tasa de respuesta más baja de los hogares unipersonales.

Tabla 1 Resultados de una encuesta de hogares hipotética

	Tamaño del hogar					
Estado	1	2	3	4	5 +	Total
Terminado	85	268	293	212	142	1000
Porcentaje completado	8.5	26,8	29.3	21.2	14.2	100.0
Terminado	39	57	30	40	34	200
Porcentaje rescindido	19.5	28,5	14.8	20.1	17.1	100.0

Tabla 2 Cálculo de ponderaciones para la encuesta de hogares hipotética

	Tamaño del hogar					
Fuente	1	2	3	4	5 +	Total
Completado y terminado	124	325	323	252	176	1.200
Porcentaje completado + terminado	10.33	27.08	26,91	21.00	14.67	100.0
Porcentaje completado	8.5	26,8	29.3	21.2	14.2	100.0
Pesos (relación de dos proporciones)	1.22	1.01	0,92	0,99	1.03	-