**Un marco para el muestreo límite en las empresas**

**Diseño de encuesta**

En la teoría del muestreo, las distribuciones sesgadas de muchas de las variables de la encuesta en una población dificultan el uso de herramientas clásicas. Una posible solución es el muestreo de corte, que descarta una parte de la población del marco muestral. Aunque el muestreo de corte es común entre los profesionales, sus fundamentos teóricos son débiles porque las probabilidades de inclusión de algunas de las unidades son cero. En este artículo proponemos un marco que justifica el muestreo de corte y proporciona un medio para determinar el censo y los umbrales de corte. Utilizamos un modelo de estimación que supone que se conocen los tamaños de las unidades descartadas para cada variable. Calculamos la varianza del estimador resultante y su sesgo. Desarrollamos un algoritmo de minimización del error cuadrático medio en función de información auxiliar multivariada a nivel de población. Debido a la naturaleza multivariada del modelo, empleamos la teoría de relajación estocástica y utilizamos el algoritmo de recocido simulado.

**Introducción**

El muestreo de corte es un procedimiento comúnmente utilizado por los institutos nacionales de estadística para seleccionar muestras, pero no es fácil dar una definición única y clara de la metodología.

En términos generales, la población se divide en dos o tres estratos, de modo que las unidades de cada estrato reciben un trato diferente. En particular, parte de la población objetivo suele quedar excluida a priori de la selección de la muestra.

En Knaub (2008b) se ofrece una breve introducción al muestreo de corte. La formulación básica (Hansen et al. 1953, pp. 486-490; Sa¨rndal et al. 1992, pp. 531-533), frecuentemente empleada en el campo de la recopilación de precios, se caracteriza por un umbral tal que las unidades por encima este umbral se incluyen en la muestra con una probabilidad positiva.

Las unidades por debajo de este umbral se descartan, siendo nula su probabilidad de ser incluidas en la muestra. En este caso, como señalaron de Haan et al. (1999), la varianza muestral es cero por definición. Sin embargo, esto no implica la solución de todos los problemas de precisión. Es bien sabido (ver, por ejemplo, Sarndal et al. 1992, p. 531) que el muestreo de corte produce estimadores sesgados. Por lo tanto, la medida de error que normalmente se utiliza es el error cuadrático medio (es decir, la suma de la varianza y el sesgo al cuadrado). De ello se deduce que el muestreo de corte podría ser una buena opción cuando la reducción de la varianza compensa con creces la introducción de un pequeño sesgo (Knaub 2007).

Hidiroglou (1986) propone una interpretación alternativa, que considera dos estratos. En el primero se incluyen todas las observaciones en la muestra, mientras que en el segundo no se descartan las unidades sino que se muestrean. En este contexto, el algoritmo propuesto por Lavalle´e e Hidiroglou (1988) se utiliza a menudo para determinar los límites de los estratos y los tamaños de muestra de los estratos. Para el escenario donde difieren la variable de encuesta y la variable de estratificación, Rivest (2002) propuso una generalización del algoritmo de Lavalle´e e Hidiroglou.

El algoritmo Rivest incluye un modelo que tiene en cuenta las diferencias entre la encuesta y la variable de estratificación y permite encontrar el tamaño de muestra óptimo y los límites óptimos del estrato para un diseño de tomar todo/tomar algo.

Finalmente, el enfoque más general (el adoptado en este artículo) considera tres estratos cuyas unidades se enumeran completamente, se muestrean y se descartan, respectivamente. Como señalaron Sigman y Monsour (1995), este tipo de estratificación es particularmente apropiado en las encuestas empresariales, porque las empresas tienden a tener distribuciones sesgadas. Por lo tanto, el tamaño tiene un impacto considerable en la precisión de las estimaciones de las encuestas, y no darse cuenta de que tales poblaciones deben estratificarse de la manera antes mencionada puede causar una subestimación de las características de la población. Aunque esta variable se concentra en unos pocos establecimientos grandes, esta metodología proporciona al investigador una muestra cuyo tamaño es bastante pequeño pero cuyo grado de cobertura es alto.

El problema tratado en este artículo es una generalización del muestreo de corte estándar. Como es habitual en las encuestas de empresas, asumimos que la población de interés tiene un sesgo positivo, debido a la presencia de pocas unidades “grandes” y muchas unidades “pequeñas”. Si uno está interesado en estimar el total de la población, un porcentaje considerable de las observaciones hace una contribución insignificante al total; por otro lado, la inclusión en la muestra de las observaciones más grandes es esencialmente obligatoria.

En tales situaciones, los profesionales suelen utilizar particiones de la población en tres conjuntos: un estrato de toma todo cuyas unidades se encuestan en su totalidad (UC), un estrato de toma algo del cual se extrae una muestra aleatoria simple (EE. UU.) y un estrato de toma nada del cual se extrae una muestra aleatoria simple (EE. UU.) estrato cuyas unidades se descartan (UE). En otras palabras, los encuestadores deciden a priori excluir del análisis a parte de la población (por ejemplo, empresas con menos de cinco empleados). Sin embargo, esta elección suele estar motivada por el deseo de armonizar las normas administrativas: en este caso, la división de las empresas en pequeñas, medianas y grandes. Esta estrategia se emplea con tanta frecuencia en las encuestas empresariales que su uso es “implícito” y “acrítico”, de modo que las consecuencias inferenciales de las restricciones causadas al archivo por este procedimiento se ignoran en su mayor parte.

El problema de determinar el umbral óptimo, es decir, la división de la población en estratos UC y US, es relativamente sencillo tanto desde el punto de vista técnico como metodológico (Hidiroglou 1986). Por otro lado, encontrar un criterio que asigne cada unidad exactamente a uno de los tres estratos tiende a considerarse una alternativa no viable, principalmente porque algunas probabilidades de inclusión se igualan a cero. De ello se deduce que el muestreo de corte se encuentra, en cierto sentido, en una posición intermedia entre los esquemas de muestreo probabilístico y no probabilístico, una característica que no es apreciada por los expertos en este campo. Como resultado, en la literatura existen muy pocos artículos sobre sus fundamentos metodológicos.

No obstante, en aplicaciones se utiliza con frecuencia. Es el caso, por ejemplo, de la encuesta mensual sobre el sector manufacturero realizada por Statistics Canada (véase, por ejemplo, Statistics Canada 2001), que implícitamente utiliza un muestreo de corte, sin prestar demasiada atención a las implicaciones metodológicas: “ El marco de muestreo para la Encuesta Mensual Canadiense de Manufactura (MSM) se determina a partir de la población objetivo después de restar los establecimientos que representan el 2% inferior de los envíos totales de manufactura estimados para cada provincia. Estos establecimientos fueron excluidos del marco para poder reducir el tamaño de la muestra sin afectar significativamente la calidad”. Se emplean procedimientos similares en encuestas realizadas por otros institutos nacionales de estadística (para una revisión exhaustiva, ver Knaub 2007, Sección II): el muestreo de corte se utiliza ampliamente pero los aspectos metodológicos no están documentados. Dos excepciones son el libro de Sarndal et al. (1992, págs. 531 –533), y el artículo de de Haan et al. (1999): este último presenta aplicaciones exitosas del muestreo de corte en el campo de los índices de precios al consumidor. Como señala Knaub (2007, p. 2), el muestreo de corte para estimar los precios unitarios puede ser útil: “Si se utiliza una muestra de corte para los ingresos y otra para el volumen de ventas, entonces la relación tenderá ser más exacto que el numerador o el denominador”.

Finalmente, Elisson y Elvers (2001) realizaron un análisis univariado que compara el muestreo de corte con el muestreo estratificado simple. Concluyen que el muestreo de corte merece más consideración y sugieren su uso en aplicaciones; sin embargo, encuentran que la variable dimensional que determina el umbral de corte tiene un impacto relevante en los resultados, por lo que destacan que se debe tener mucho cuidado en la elección de esta variable. Además, señalan la necesidad de un modelo apropiado para la estimación de la fracción de población excluida de la muestra.

En cualquier caso, cabe mencionar las ventajas prácticas del muestreo por corte en cuanto a los costes de una encuesta:

1. construir y actualizar un marco muestral para unidades de pequeñas empresas podría ser demasiado costoso, considerando que la ganancia en eficiencia de los estimadores probablemente sería pequeña;
2. excluir las unidades de la población que contribuyen poco a los agregados a estimar normalmente implica una gran disminución en el número de unidades que deben ser encuestadas para obtener un nivel de precisión predefinido de las estimaciones;
3. poner una restricción a la población del marco y, en consecuencia, a la muestra, permite reducir el problema de los estratos vacíos que afecta principalmente a las empresas más pequeñas. Al respecto, varios análisis empíricos mostraron que algunas dificultades, como la tasa de no respuesta, la tasa de rotación de las unidades económicas y los errores de sub o sobre cobertura del marco, se vuelven más relevantes a medida que el tamaño de las unidades se reduce;
4. se puede demostrar que el muestreo de corte es más práctico en términos de precisión cuando se tiene en cuenta el error total de la encuesta. Knaub (2004) muestra una manera de medir el error total de la encuesta en el contexto de errores ajenos al muestreo, como el error de medición.

Dado que los profesionales están a favor de tales particiones de la población y existen razones técnicas que justifican su uso, la pregunta básica es: ¿es posible considerar el muestreo de corte como un esquema de muestreo válido? Si la respuesta es positiva, la cuestión es definir un marco estadístico para el muestreo de corte.

En este sentido, intentamos desarrollar una solución fácilmente implementable al problema de la construcción de los tres estratos UC, US y UE en una configuración multipropósito y multivariada. En otras palabras, de manera similar a lo que sucede en aplicaciones prácticas, asumimos un interés en encuestas con más de una variable objetivo, utilizando información auxiliar contenida en múltiples variables.

Sin embargo, el caso en el que se dispone de una única medida de tamaño no es tan infrecuente. Por ejemplo, en las encuestas empresariales, la única información auxiliar suele ser el número de empleados. Además, cuando las encuestas son voluntarias, la tasa de participación de las pequeñas empresas es en su mayoría muy baja. En este caso, sin duda es conveniente un procedimiento de muestreo de corte basado en una variable dimensional (Bailar et al. 1983, Sección 5.1). La situación debería simplificarse sustancialmente en el sentido de que la toma –todas las categorías serían las unidades con las medidas de mayor tamaño y la toma –ninguna serían las unidades con las medidas más pequeñas. Quedaría por determinar los puntos límite. Para este problema probablemente haya que encontrar una solución analítica, en lugar de algorítmica. Sin embargo, el problema actual es multivariado, por lo que dejamos para futuras investigaciones el estudio del problema univariado.

La estructura del artículo es la siguiente. En la Sección 2 definiremos un modelo de estimación que supone, para cada variable, que el peso de las unidades excluidas del análisis es conocido y constante. Sin embargo, esta hipótesis, en general, no está bajo el control del investigador, por lo que este estimador está sesgado, y tendremos que encontrar su sesgo y su error cuadrático medio (MSE). El modelo será desarrollado para la estimación de un total. La sección 3 se dedicará a la derivación del tamaño de muestra para el esquema de corte, centrándose en su optimización y, en consecuencia, en la construcción del diseño óptimo. El problema se abordará definiendo el tamaño de la muestra como una función de la partición UC, US y UE determinada sobre la base de información auxiliar multivariada que se supondrá conocida para toda la población. En vista de la naturaleza combinatoria de este problema, utilizaremos la teoría de relajación estocástica y, en particular, el algoritmo de recocido simulado (SA). En la Sección 4 mostraremos alguna evidencia empírica sobre el sesgo del estimador cuando utilizamos datos de encuestas sobre empresas mataderos en Italia. En la misma sección presentaremos los principales resultados de la aplicación del esquema de muestreo a este conjunto de datos. Finalmente, la Sección 5 concluirá el artículo y señalará algunos problemas abiertos.

4. CASO DE ESTUDIO

En esta sección aplicaremos el diseño de corte óptimo propuesto anteriormente a la encuesta mensual de sacrificio de carnes rojas realizada por ISTAT (Instituto Nacional de Estadística de Italia). El objetivo consiste en obtener información sobre el número y el peso de los animales sacrificados mensualmente en Italia. Esta encuesta se basa en un muestreo estratificado, con estratificación por tipos de matadero y división geográfica, para un total de 5 estratos, dos de ellos con referencia geográfica. Las divisiones geográficas son Noroeste (1), Noreste (2), Centro (3), Sur (4) e Islas (5). Los estratos son los siguientes:

* Estrato 1 (siempre totalmente observado): mataderos privados con sello de la Comunidad Económica Europea (CEE) en la división geográfica 1 o 2;
* Estrato 2: mataderos privados con sello CEE en la división geográfica 3, 4 ó 5;
* Estrato 3: mataderos privados con baja capacidad (independientemente de la división geográfica);
* Estrato 4: mataderos privados en abandono, públicos con sello CEE y públicos en derogación (aparte de división geográfica);
* Estrato 5: mataderos públicos con baja capacidad.

La estratificación también actúa como un criterio dimensional que asigna al Estrato 1 las empresas con más de 10.000 sacrificios de ovinos y caprinos o más de 50.000 porcinos.

En promedio, la muestra es de aproximadamente 460 unidades para una población de 2211 unidades con el nivel deseado de precisión c establecido en 5% (ISTAT 2007). A continuación compararemos la estratificación ISTAT, preparada por expertos en base a su conocimiento de las variables investigadas, con nuestra estratificación óptima.

Además de la encuesta mensual, Istat realiza anualmente el censo de los mataderos. Por lo tanto, nuestro marco contiene N = 2.211 mataderos para los cuales conocemos cuatro variables enumeradas completamente en 1999, 2000 y 2001. Son, respectivamente, el número total de (i) bovinos sacrificados, (ii) porcinos, (iii) ovinos y caprinos y (iv) equinos. Primero consideraremos el conjunto de datos completo (para cada uno de los tres años) para evaluar el comportamiento del sesgo b(tE) y buscar posibles regularidades. Recordando que d se define como la relación entre el número de unidades de población descartadas y el número de unidades de población muestreadas y enumeradas completamente, es necesario conocer la lista completa de las variables auxiliares rezagadas.

Luego se implementará el diseño de corte propuesto en este artículo, con el objetivo de establecer una encuesta mensual sobre sacrificio para el año 2002, utilizando como variables auxiliares sólo los datos enumerados íntegramente en 2001. Nuestro ejercicio consiste en estimar los mismos totales estimado por el ISTAT en su encuesta mensual. Comenzamos con una breve descripción del archivo que nos ocupa. Los diagramas de dispersión de todos los pares de las cuatro variables en 2001 se muestran en la Figura 1; Los mismos gráficos para los años 1999 y 2000 son casi idénticos y por lo tanto no se presentan aquí. La principal evidencia es que las variables están esencialmente no correlacionadas, como lo confirma el coeficiente de correlación lineal, que oscila en el intervalo ½20:0096; 0:0566. Además, los mataderos están muy especializados y la mayoría de las empresas son pequeñas. En particular, el 38,9% de las empresas sacrifican un solo tipo de animales, el 24,06% dos tipos, el 21,85% tres tipos y sólo el 15,2% los cuatro tipos de animales.

**4.2 Diseño Muestral**

Pasemos ahora finalmente a los resultados de la implementación del diseño de corte para la estimación de los mismos totales de la encuesta mensual de sacrificio de carnes rojas del ISTAT. Por lo tanto, en la función objetivo (14) los tamaños de muestra nj(phi) ( j ¼ 1; : : : ; 4) están dados por (12).

En el algoritmo utilizamos un nivel deseado de precisión c = 1%. La razón por la que empleamos un nivel deseado de precisión del 1% en lugar del 5% como lo hace el ISTAT es que el muestreo de corte es considerablemente más eficiente que el muestreo estratificado estándar. Por lo tanto, es prácticamente imposible alcanzar un nivel de precisión del 1% mediante el enfoque estratificado estándar utilizado por el ISTAT, a menos que el tamaño de la muestra sea irrealmente grande. En esta sección analizamos las consecuencias del uso de diferentes valores numéricos de c.

La Figura 3 muestra el tamaño de muestra óptimo total en función del número de iteraciones del recocido simulado. Es inmediatamente evidente que la “mayor disminución” en el tamaño de la muestra tiene lugar en las primeras iteraciones; las iteraciones restantes parecen proporcionar sólo un ajuste hacia el óptimo global. Más precisamente (ver Figura 4), aproximadamente después de las primeras 100 iteraciones, el algoritmo simplemente mueve algunas observaciones de UE a EE. UU.; A estas operaciones de cambio de etiquetas corresponden disminuciones muy pequeñas en el tamaño total de la muestra.

La Tabla 3 brinda algunos detalles sobre la implementación del algoritmo. La cantidad NS es el tamaño del estrato US; el número de unidades realmente muestreadas de este estrato se puede calcular como n - NC; por ejemplo, en la iteración 298 (es decir, cuando el algoritmo converge) tomamos una muestra de n - NC < 314 - 280 = 34 unidades. Vale la pena añadir que el algoritmo es bastante lento: una iteración dura casi cinco segundos, por lo que la convergencia se alcanza en aproximadamente 22 minutos en un Pentium 4 a 3,00 GHz. Según nuestra experiencia computacional, el tiempo de convergencia aumenta muy rápidamente a medida que N aumenta, por lo que la aplicación del método en poblaciones grandes puede resultar difícil.

El esquema de muestreo desarrollado en este artículo produce la partición de la población que se muestra en la Tabla 4 y la Figura 5. Sólo se muestra la partición relativa al ganado vacuno y porcino. Las gráficas de los pares restantes son muy similares y por eso las omitimos. Esta figura muestra el diagrama de dispersión de las raíces cuartas de las dos variables auxiliares. Se excluyen las observaciones en la parte gris claro del gráfico (no se toma nada), las de la parte blanca la porción se enumera completamente (se lleva todo), y aquellos en la porción gris oscuro se enumeran por completo muestreado (tomar algo).

La estratificación es muy clara, con dos estratos (UC y UE) cuyos tamaños son mayores que los de Estados Unidos. El estrato de tomar algo está anidado en el estrato de no tomar nada, con un muestreo fracción igual al 16%. Esto significa que en nuestra aplicación el esquema de muestreo es bastante similar a un diseño de toma todo/no toma nada. De hecho, en el estudio de caso, 1.723 unidades están cortadas, 280 unidades están completamente enumeradas y sólo 208 pertenecen al estrato muestral genuino (Cuadro 3). Puede parecer sorprendente que aproximadamente 3/4 de la población puedan quedar fuera sin perjudicar la encuesta. Además, más del 50% de las unidades con probabilidad de inclusión positiva pertenecen al estrato completamente censado. Sin embargo, esto no es inesperado, porque normalmente sucede si una variable de estratificación está altamente correlacionada con la variable de estudio, como es el caso aquí dado el corto lapso entre censos.

Además, según los resultados teóricos derivados de la sección anterior, se esperaba una fracción de muestreo tan pequeña. Teniendo en cuenta la gran concentración de la población, el estrato estadounidense contiene principalmente empresas cuyos valores de las cuatro variables auxiliares son diferentes de cero, es decir, las menos especializadas.

Para completar esta discusión, considere las Tablas 5 y 6. La Tabla 5 muestra una comparación de los resultados que se acaban de presentar y algunos resultados obtenidos mediante muestreo de corte a medida que cambia el sesgo bð^tEÞ. La posición de nuestro caso de estudio en la tabla se explica por el hecho de que, aunque hemos utilizado cuatro valores numéricos para el sesgo (tantos como variables encuestadas), el valor medio del sesgo es aproximadamente igual a 0,1.