Estadística Española • Volumen 60, número 196 / 2018, pp. 185-217

Utilidad de los descriptores aplicativos y su estimación en la estadística oficial

Wenceslao González Manteiga
Departamento de Estadística, Análisis Matemático y Optimización
Universidad de Santiago de Compostela

Carlos L. Iglesias Patiño
Instituto Galego de Estatística

Resumen

En un universo, se estudian dos caracteres, X e Y, Y cuantitativo y X, fundamentalmente, nominal. Interesa describir la variable Y basándose en la relación entre Y y X. Para ello se proponen una serie de aplicaciones, que se denominan descriptores aplicativos. También se definen estimadores de estos parámetros generalizados de la población finita. Se calculan su esperanza y varianza en dos diseños muestrales: muestreo aleatorio simple sin reemplazamiento y estratificado. Después se obtienen expresiones prácticas para sus sesgos absoluto y relativo y también para su varianza, para ésta se presentan estimadores. Esas expresiones son exactas o aproximadas en muestras "finitas". El regresograma resulta un caso particular de descriptor aplicativo. También se presenta otra aplicación a la representación mediante pictogramas.

Palabras clave: descriptor aplicativo, regresograma, muestreo aleatorio simple sin reemplazamiento, muestreo estratificado

Clasificación AMS: 62D, 62G

Usefulness of the mapping descriptive statistics and their estimation in official statistics

Abstract

Let be a finite population with two variates, X and Y, Y quantitative and X, namely, nominal. It is interesting to describe variate Y from relationship between Y and X. A sort of mappings will be proposed; they will be called mapping descriptive statistics. Also estimators for these generalized parameters of the finite population are defined. Their expectancy and variance are calculated based on two sampling designs: simple random sampling without replacement and stratified sampling. Then practical expressions for absolute and relative bias and also for their variance

are obtained and estimators for the latter are introduced. Those expressions are exact or approximate in "finite" samples. Regressogram is a particular case of mapping descriptive statistics. Also, another application is introduced to graphic representation, plotting pictograms.

Keywords: mapping descriptive statistics, regressogram, simple random sampling without replacement, stratified sampling

AMS:*classification:* 62D, 62G

1. Introducción

Tres décadas atrás, se produjo la revolución en el tratamiento de datos: la informática personal. Se consideraba como la panacea de la estadística exhaustiva. La realidad es que ha resuelto numerosos problemas en la fase de transformación de los microdatos a macrodatos pero, aunque ha facilitado la captación, subsiste el problema de su calidad inicial así como su integración en sistemas de cuentas u otras relaciones o en otro tipo de modelos.

Antiguamente existía una diferencia más clara entre datos estadísticos y datos de origen administrativo o de otra índole jurídica, entre estadísticas exhaustivas y por muestreo o entre información estadística y cartográfica. Actualmente los datos de origen no estadístico poseen un interés cada vez mayor y más, si cabe, con los escenarios presupuestarios restrictivos. En la actualidad es más frecuente que se mezclen este tipo de datos con los de carácter estadístico para elaborar, a su vez, una nueva estadística (Moral-Arce y Martín, 2009).

Otro ejemplo paradigmático son los censos demográficos de 2011 donde se empleó, además de la información estadística, otra de tipo administrativo, fiscal, etc.. Los datos no disponibles sobre población se captaron mediante una gran encuesta por muestreo. A todo lo anterior hay que añadir la recogida de datos multicanal: Internet, correo o por visita de un agente censal (Teijeiro y Vega, 2014 e INE, 2011).

Todos estos cambios recomiendan variaciones en los métodos estadísticos empleados. La multiplicidad de fuentes constituye un nuevo generador de variabilidad que puede aconsejar el empleo de la suavización. Su necesidad ya se había manifestado en el uso de la información estadística en contextos como los análisis demográfico y de la coyuntura o en las estadísticas de síntesis (cuentas, balances), ahora se manifiesta en operaciones estadísticas básicas (censos, encuestas…).

En este tipo de operaciones o por necesidades de la difusión, conviven abundantes caracteres nominales con variables. Por ello, se introducirán unas aplicaciones como forma de expresar la relación entre una cualidad y una magnitud que resultarán interesantes para describir la variabilidad a través de la población finita (p.f.) por eso se denominarán descriptores aplicativos o, por brevedad, descriptores. Serán parámetros de la p.f. que generalicen conceptos habituales en estadística descriptiva, incluso los que no han cabido en un tratamiento unitario como la tabulación y la representación gráfica.

A continuación se presenta como se estructura el resto del artículo. En la siguiente sección se definen los descriptores aplicativos. La tercera sección se dedica a introducir estimadores de estos descriptores en dos diseños muestrales, muestreo aleatorio simple sin reemplazamiento y muestreo estratificado y a estudiar su sesgo y varianza así como sus estimadores. En la cuarta, se presentan aplicaciones de estos descriptores y de sus estimadores entre las que destaca el regresograma como caso particular. Por último, se presentan las conclusiones y un anexo con las demostraciones.

2. Descriptores aplicativos

Sea U un universo de tamaño N que se identificará con sus etiquetas, $\mathrm{U} = \{1,2,\cdots,N\}$. En cada elemento del universo se estudian una variable Y y un carácter X que induce una partición $\Pi = \{B_c\}_{c=1}^{C}$ de U, siendo C el número de clases en que se divide el universo. Sea $\mathcal{P} = \{(i, x_i, Y_i), i = 1,\cdots,N\}$ la p.f., donde i es la etiqueta, x_i el carácter de clasificación e Y_i la variable de interés o estudio correspondientes al elemento etiquetado con i.

Obsérvese que no es restrictivo la consideración de un carácter nominal ya que si se quieren estudiar varios, k, simultáneamente basta considerar como X el producto cartesiano de los k caracteres. En el caso de $k = 2$, $C = IJ$ siendo I el número de clases del primer carácter y J el del segundo.

Ejemplos paradigmáticos de X son un carácter nominal en el que se emplea una clasificación estándar (CNAE, CNO, NUTS…) o el de un carácter cuantitativo, variable X, que toma valores en un intervalo y se discretiza mediante subintervalos. La partición correspondiente a este carácter de clasificación puede estar determinada por estándares, por tradición o venir condicionada por el secreto estadístico.

Se pueden definir aplicaciones con origen en la partición generada por X que describirán la p.f. $\mathcal{P}$, en concreto, las que asignan a cada clase su frecuencia absoluta o relativa, su media o su total de Y. Se podría pensar en emplear otros estadísticos descriptivos como, por ejemplo, la mediana aunque esto rebasa el objetivo del artículo.

Se denotará por $N(B_c)$ la frecuencia absoluta, en lo sucesivo recuento, reservando el término frecuencia para la relativa. $N(B_c) \geq 2$ de facto $N(B_c) \geq 3$ por secreto estadístico ya que «mientras para el matemático un total empieza con dos números, para el estadístico empieza con 3» (Piatier, 1967). En numerosos casos, se considera 4 como umbral para evitar problemas de concentración excesiva. Sea $P_{N,c} = N^{-1}N(B_c)$ la frecuencia, se verifica que $P_{N,c} > 0$.

Se denominará subtotalizador a la aplicación $Y_+(\cdot)\colon \Pi \to \mathbb{R}$, $Y_+(B_c) = \sum_{i\in B_c} Y_i$ y, del mismo modo, promediador a la que asigna $\bar{Y}(B_c) = N^{-1}(B_c)\sum_{i\in B_c} Y_i$ abusando de la notación $N^{-1}(B_c) \coloneqq \big(N(B_c)\big)^{-1}$. También se podría utilizar el subíndice, $N_{B_c}^{-1}$, de ahí el abuso, pero se reservará esta posibilidad para poder tratar la estratificación en secciones posteriores. Otras expresiones son:

$$Y_+(B_c) = \sum_{i=1}^{N} Y_i 1_{B_c}(i)$$

para el subtotalizador, y para el promediador

$$\bar{Y}(B_c) = \frac{\sum_{i=1}^{N} Y_i 1_{B_c}(i)}{\sum_{i=1}^{N} 1_{B_c}(i)} = \frac{Y_+(B_c)}{N(B_c)}.$$

Este descriptor se puede concebir como el resultado de un ajuste por mínimos cuadrados ponderados

$$\bar{Y}(B_c) = \arg\min_{\alpha} \sum_{i=1}^{N} (Y_i - \alpha)^2 1_{B_c}(i)$$

En particular, si se considera la variable degenerada Y igual a la constante unidad, el subtotalizador coincide con la aplicación $N(\cdot)$, que se podría denominar recontador, $N(B_c) = \sum_{i=1}^{N} 1_{B_c}(i)$. Si se divide por N, el tamaño del universo, sería la aplicación que asigna frecuencias, mide la abundancia relativa de cada clase.

El recontador constituye la expresión matemática de lo que se conoce habitualmente en estadística pública como cuadro o tabla de recuentos de ahí la denominación que se propone, mientras que el promediador y el subtotalizador, el cuadro o tabla de magnitudes (INE, 2004) cuando se consignan medias aritméticas o totales, respectivamente.

El problema del promediador es la constancia dentro de la clase. Una manera de solventarlo es dar un resumen de como fluctúa la variable en ella, esto es, se puede complementar con una medida de dispersión de los valores de Y correspondientes a la clase respectiva. Sea

$$S^2(B_c) = \frac{1}{\sum_{i=1}^{N} 1_{B_c}(i) - 1} \sum_{i=1}^{N} \left(Y_i - \bar{Y}(B_c)\right)^2 1_{B_c}(i)$$

esa medida. Con ella se ha introducido un nuevo descriptor aplicativo el variador cuya raíz cuadrada $S(\cdot)$ se denominará, en lo sucesivo, desviador.

3. Estimación de los descriptores aplicativos

Resulta habitual en las administraciones públicas pensar que, por disponer de un procedimiento administrativo amparado en una legislación sectorial y de un sistema informático que lo soporta, se posee un sistema de información. Esta identificación automática está lastrando el aprovechamiento estadístico de las fuentes administrativas.

Aunque la implantación haya sido adecuada, la dinámica de la gestión diaria conduce a una pérdida progresiva de calidad en los datos si no existe un control que, a su vez, implica dedicación extra.

Que se solicite un dato en un procedimiento administrativo o de gestión y que se cubra no quiere decir que su contenido sea válido. Si no es relevante para el procedimiento, no se suele dedicar esfuerzos para que se pueda procesar posteriormente. Es decir, el campo acaba por ser, de facto, de cumplimentación voluntaria (Durán, 2007).

Cuando, en un campo de un formulario o de una tabla relacional, no existan garantías de que se haya recogido con suficiente rigor desde el punto de vista estadístico, por no ser importante para el procedimiento administrativo, puede ser inabordable revisar todos los elementos y, por tanto, una solución sería extraer una muestra y operar sobre ella.

Si los campos con este problema son numerosos o bien existe incertidumbre sobre su número –como es habitual– conviene elegir una muestra aleatoria simple sin reemplazamiento que se abreviará MAS. Mientras que si se sabe con seguridad que es reducido (por ejemplo, de 1 a 3) se podría considerar un diseño muestral mejor, por ejemplo estratificando.

Cuando se quiere obtener una muestra que sea equilibrada para muchas variables, una estrategia aceptable consiste en extraer una MAS (Royall&Herson, 1973) ya que se podría decir que el muestreo aleatorio simple sin reemplazamiento es equilibrado "en promedio".

Ese mismo proceder sería recomendable, entre otros, en un procedimiento de actualización de marcos y directorios, siempre que no implique salir a campo; en estudios piloto que permitan explorar la viabilidad de una nueva metodología y compararla con la tradicional. O cuando el sistema de información contiene una enorme cantidad de elementos como es el caso del sistema de la Seguridad Social, cuyo fin principal es agilizar su gestión.

En resumen, se ha mencionado una forma de simbiosis entre estadística exhaustiva y muestral, el censo de 2011 y la Encuesta anual de estructura salarial. Otra podría ser la explotación de registros administrativos y sistemas de información por muestreo que ya se empleó en la explotación de censos antes de la generalización de la informática (Arribas y Almazán, 2006) o, incluso, después de ella (Pérez, 2011) y, tras la revolución ofimática, en la construcción de la muestra continua de vidas laborales (MCVL) en 2004, donde se extrajo una MAS, dado que se pretendía que la muestra fuera multipropósito.

De hecho para mantener la MCVL a lo largo de los años, y así constituir una muestra continua, el procedimiento de actualización intenta que la muestra resultante referida a cada año se asemeje a una MAS (Durán, 2007).

3.1 Muestreo aleatorio simple sin reemplazamiento

Llegado a este punto, interesará estimar estas aplicaciones descriptivas. Para ello se deberá proceder con el total, la media o la cuasivarianza de una clase cualquiera. Se emplearán diferentes estrategias muestrales: dos diseños, MAS y muestreo estratificado con MAS en cada estrato (MASE), con sus estimadores respectivos.

3.1.1 Promediador

Se denotará por MAS(N,n) el muestreo aleatorio simple sin reemplazamiento de un universo de tamaño N del que se extrae una MAS s de tamaño n, $\#s = n$. Se tratará la estimación del promediador cuando se emplea la estrategia constituida por el diseño MAS(N,n) y el siguiente estimador

$$\widehat{\bar{Y}}(B_c) = \frac{\sum_{i=1}^{N} Y_i 1_{B_c}(i) I_i}{\left(\sum_{i=1}^{N} 1_{B_c}(i) I_i\right) \vee 1} = \frac{\sum_{i=1}^{N} Y_i 1_{B_c}(i) I_i}{n(B_c) \vee 1}$$

siendo I_i, la variable indicatriz de pertenencia a la muestra del elemento i, $a \vee b = \max\{a,b\}$ y $n(B_c) = \#(s \cap B_c)$ con s MAS.

3.1.1.1 Estudio del sesgo

Teorema 1. En MAS(N,n), se verifica que $\mathrm{E}_p\left[\widehat{\bar{Y}}(B_c)\right] = \bar{Y}(B_c)[1 - p(n(B_c) = 0)]$

Corolario 1. En MAS(N,n), se verifica que $\mathrm{B}_p\left[\widehat{\bar{Y}}(B_c)\right] = -\bar{Y}(B_c)p(n(B_c) = 0)$

Si $\bar{Y}(B_c) > 0$, el sesgo es negativo y cuando $\bar{Y}(B_c) < 0$, es positivo. En el caso de que sea nulo también es nulo el sesgo. Es decir, el sesgo es de signo contrario al promediador. Además, en caso de que Y sea no negativa, situación habitual en estadística pública, si $\bar{Y}(B_c) = 0$, $\widehat{\bar{Y}}(B_c) = 0$ trivialmente.

Corolario 2. En MAS(N,n), si $\bar{Y}(B_c) \neq 0$ se verifica que el sesgo relativo es $\left|\mathrm{b}_p\left[\widehat{\bar{Y}}(B_c)\right]\right| = p(n(B_c) = 0)$

Lema 1. En MAS(N,n), se verifica que $p(n(B_c) = 0) < e^{-nP_{N,c}}$.

Obsérvese que $\mathrm{E}_p(n(B_c)) = nP_{N,c}$ es decir, el número promedio de elementos de la MAS en la clase. Por tanto, siempre que esta esperanza sea grande, el sesgo relativo es pequeño. En la siguiente tabla se presenta el valor de la cota para diferentes n y $P_{N,c}$.

Tabla 1

Cota del sesgo relativo del promediador en muestreo aleatorio simple sin reemplazamiento

	$P_{N,c}$ =	0,0001	0,001	0,01	0,1	0,111	0,125	0,143	0,167	0,2	0,25	0,333	0,5	1
n =	10	0,999	0,990	0,905	0,368	0,329	0,287	0,240	0,189	0,135	0,082	0,036	0,007	0,000
	20	0,998	0,980	0,819	0,135	0,108	0,082	0,057	0,036	0,018	0,007	0,001	0,000	0,000
	40	0,996	0,961	0,670	0,018	0,012	0,007	0,003	0,001	0,000	0,000	0,000	0,000	0,000
	80	0,992	0,923	0,449	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	160	0,984	0,852	0,202	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	320	0,969	0,726	0,041	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Corolario 3. En MAS(N,n), se verifica que

$$\left|\mathrm{B}_p\left[\widehat{\bar{Y}}(B_c)\right]\right| \leq \max_{B_c}|Y_i|\, e^{-nP_{N,c}} \leq \max_{N}|Y_i|\, e^{-nP_{N,c}}$$

Por tanto, esta estrategia muestral resulta aproximadamente insesgada si el máximo se comporta adecuadamente.

Si se conoce $N(B_c)$, se puede dar un estimador insesgado $[1-p(n(B_c)=0)]^{-1}\widehat{\overline{Y}}(B_c)$ o estimar el sesgo mediante $\widehat{\mathrm{B}}_p\left[\widehat{\overline{Y}}(B_c)\right] = -\widehat{\overline{Y}}(B_c)p(n(B_c)=0)$ donde $p(n(B_c)=0)$ se calcula o se aproxima mediante $\left(1-P_{N,c}\right)^n$.

3.1.1.2 Estudio de la varianza

Del mismo modo que se ha introducido el desviador, sea D_p la desviación típica según el diseño muestral, entonces la varianza admite la siguiente expresión

Teorema 2. En MAS(N,n), se verifica que

$$\mathrm{D}_p^2\left[\widehat{\overline{Y}}(B_c)\right] = S^2(B_c)\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m) +$$

$$+\overline{Y}^2(B_c)[1-p(n(B_c)=0)]p(n(B_c)=0)$$

Por ejemplo, si $p(n(B_c)=0)=0'1$, el segundo sumando de la varianza es el 9% del cuadrado de la media pudiendo resultar en algunos casos un sumando apreciable. En general

$$\mathrm{D}_p^2\left[\widehat{\overline{Y}}(B_c)\right] \le S^2(B_c)\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m)+\frac{1}{4}\,\overline{Y}^2(B_c)$$

por tanto, la rel-varianza o varianza relativa $\mathrm{d}_p^2\left[\widehat{\overline{Y}}(B_c)\right]$ no rebasa

$$\frac{S^2(B_c)}{\overline{Y}^2(B_c)}\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m)+\frac{1}{4}$$

Considérese $\overline{Y}^{-1}(B_c)S(B_c)$ como el coeficiente de variación de Y en la clase c que daría lugar a un nuevo descriptor aplicativo, $(S/\overline{Y})^2(\cdot)$. Muchos de estos descriptores resultan de operar con dos de estas aplicaciones.

Si se considera la variable degenerada Y igual a la constante unidad, $\overline{Y}(B_c)=1, S^2(B_c)=0$ y el estimador se convierte en una Bernoulli, $\widehat{\overline{Y}}(B_c)=0$ si $n(B_c)=0$ y $\widehat{\overline{Y}}(B_c)=1$ si $n(B_c)\neq 0$, por tanto, $\mathrm{D}_p^2\left[\widehat{\overline{Y}}(B_c)\right]=[1-p(n(B_c)=0)]p(n(B_c)=0)$ que corresponde con lo que se obtendría tras aplicar el teorema.

Por otra parte, de la demostración del teorema, se obtiene el sesgo de $\widehat{\overline{Y}}^2(B_c)$ como estimador de $\overline{Y}^2(B_c)$.

Corolario 4. En MAS(N,n), se verifica que

$$\mathrm{B}_p\left[\widehat{\bar{Y}}^2(B_c)\right] = S^2(B_c)\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m) - \bar{Y}^2(B_c)p(n(B_c)=0)$$

Por este corolario, la siguiente expresión resulta una aproximación por exceso de este sesgo

$$\mathrm{B}_p\left[\widehat{\bar{Y}}^2(B_c)\right] \cong S^2(B_c)\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m)$$

Lema 2. En MAS(N,n), se verifica que

$$\bar{Y}^2(B_c)p(n(B_c)=0) \le \max_{B_c} Y_i^2\, e^{-nP_{N,c}} \le \max_{N} Y_i^2\, e^{-nP_{N,c}}$$

Corolario 5. En MAS(N,n), se verifica que

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right] = S^2(B_c)\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m) + R$$

con $R \le \max_{N} Y_i^2\, e^{-nP_{N,c}}$.

La igualdad del teorema también se puede escribir como

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right] = \left[\frac{S^2(B_c)}{1-p(n(B_c)=0)}\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m) + \bar{Y}^2(B_c)p(n(B_c)=0)\right]$$

$$[1-p(n(B_c)=0)] =$$

$$= \left[S^2(B_c)\,\overline{\left(\frac{1}{n^+(B_c)}-\frac{1}{N(B_c)}\right)} + \bar{Y}^2(B_c)p(n(B_c)=0)\right][1-p(n(B_c)=0)]$$

donde aparece un factor idéntico al del sesgo. En numerosos casos, se podría despreciar $\bar{Y}^2(B_c)p(n(B_c)=0)[1-p(n(B_c)=0)]$ porque el tamaño esperado de la muestra en la clase sea grande. En esta situación $1-p(n(B_c)=0)$ sería aproximadamente igual a la unidad y el primer sumando resulta igual al cuadrado del desviador por otro factor que admite la interpretación de un coeficiente de exhaustividad promedio, $\overline{\left(\frac{1}{n^+(B_c)}-\frac{1}{N(B_c)}\right)}$, ya que se puede ver como el valor medio, $\overline{(\cdot)}$, de la variable $\frac{1}{n^+(B_c)}-\frac{1}{N(B_c)}$, donde $n^+(B_c)$ denota la que toma los valores de $n(B_c)$ salvo el 0.

La combinación lineal que aparece en el teorema se puede acotar del siguiente modo

Lema 3. Se verifica que

$$\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m)$$

$$<\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1-p(n(B_c)=0)}{P_{N,c}}+\frac{1-p(n(B_c)=0)}{\left(nP_{N,c}\right)^2}+\left(\frac{2}{nP_{N,c}}\right)^3$$

Por consiguiente, la varianza se puede mayorar como se presenta a continuación

Teorema 3. Se verifica que

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right]<S^2(B_c)\left\{\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1-p(n(B_c)=0)}{P_{N,c}}+\frac{1-p(n(B_c)=0)}{\left(nP_{N,c}\right)^2}+\left(\frac{2}{nP_{N,c}}\right)^3\right\}+R$$

con $R\leq\max_{B_c}Y_i^2\,e^{-nP_{N,c}}$.

Para la interpretación del primer sumando, conviene percatarse de que $f=\left(NP_{N,c}\right)^{-1}nP_{N,c}$ si $P_{N,c}>0$ como se está suponiendo por hipótesis. Si $nP_{N,c}$ es grande, el segundo sumando de orden $\left(nP_{N,c}\right)^{-2}$ y, aún más, los siguientes resultan pequeños si el máximo no es muy grande, por tanto la expresión

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right]\approxeq(1-f)\frac{S^2(B_c)}{nP_{N,c}}$$

constituye una aproximación aceptable que coincide con la habitual fórmula para la media en $\mathrm{MAS}\left(NP_{N,c},nP_{N,c}\right)$, abusando de la notación en el tamaño muestral.

Pero, si el tamaño medio de la muestra en esa clase, $nP_{N,c}$, es pequeño, podrían ser apreciables, por ejemplo, $n=1000$ y $P_{N,c}=1\%$ entonces $nP_{N,c}=10$ y la varianza se puede ver incrementada en casi un 10%.

A continuación, se intentará dar una aproximación más operativa si B_c es una clase propia de U

Lema 4. Si $N(B_c)<N$, se verifica que

$$\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m)\approxeq\frac{1-f}{nP_{N,c}}\left(1+\frac{1-P_{N,c}}{nP_{N,c}}\right)$$

Por tanto, si se pueden despreciar los términos de orden 3 y exponenciales, una expresión aproximada de la varianza del estimador, que generaliza la de Mirás(1985), es

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right]\approxeq(1-f)\frac{S^2(B_c)}{nP_{N,c}}\left(1+\frac{1-P_{N,c}}{nP_{N,c}}\right)$$

donde aparecen dos multiplicadores, uno menor que la unidad y otro, mayor. El menor está relacionado con la corrección por poblaciones finitas. En el mayor que la unidad, el segundo sumando es igual a n^{-1} multiplicado por el predominio en términos de tamaño del resto de clases de la partición frente a la clase en cuestión.

3.1.2 Estimación del subtotalizador en muestreo aleatorio simple sin reemplazamiento

Si se conoce el recontador, $N(B_c)$, se dispone del siguiente estimador del subtotalizador $\hat{Y}_{+,e}(B_c) = N(B_c)\hat{\bar{Y}}(B_c)$ que se obtiene expandiendo el estimador del promediador mediante el recontador, de ahí la notación. Su esperanza es igual a

$$\begin{aligned}\mathrm{E}_p\left[\hat{Y}_{+,e}(B_c)\right] &= N(B_c)\mathrm{E}_p\left[\hat{\bar{Y}}(B_c)\right] = N(B_c)\bar{Y}(B_c)[1-p(n(B_c)=0)] \\ &= Y_+(B_c)[1-p(n(B_c)=0)]\end{aligned}$$

Por tanto, $\mathrm{B}_p\left[\hat{Y}_{+,e}(B_c)\right] = -Y_+(B_c)p(n(B_c)=0)$, es decir, su sesgo absoluto puede llegar a ser apreciable aunque su sesgo relativo sea el mismo que para el promediador, $p(n(B_c)=0)$. Con respecto a la varianza, se verifica la siguiente desigualdad

$$\mathrm{D}_p^2\left[\hat{Y}_{+,e}(B_c)\right] < N^2(B_c)S^2(B_c)\left\{\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{P_{N,c}}+\frac{1}{\left(nP_{N,c}\right)^2}\right\}[1-p(n(B_c)=0)]+R$$

con $R \le N^2(B_c)\left(\frac{1}{2}\max_{B_c}\left|Y_i-Y_j\right|^2\left(nP_{N,c}/2\right)^{-3}+\max_{B_c}Y_i^2\,e^{-nP_{N,c}}\right)$.

La varianza relativa está mayorada por

$$\mathrm{d}_p^2\left[\hat{Y}_{+,e}(B_c)\right] = \frac{\mathrm{D}_p^2\left[\hat{Y}_{+,e}(B_c)\right]}{N^2(B_c)\bar{Y}^2(B_c)} < \frac{S^2(B_c)}{\bar{Y}^2(B_c)}\left\{\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{P_{N,c}}+\frac{1}{\left(nP_{N,c}\right)^2}\right\}[1-p(n(B_c)=0)]+R$$

con

$$R \le \max_{B_c}\left(\frac{Y_i}{\bar{Y}(B_c)}\right)^2\left(2\left(\frac{nP_{N,c}}{2}\right)^{-3}+e^{-nP_{N,c}}\right).$$

Una vez acotada, se puede pasar a ver aproximaciones

$$\mathrm{D}_p^2\left[\hat{Y}_{+,e}(B_c)\right] \cong N^2(B_c)S^2(B_c)\frac{1-f}{nP_{N,c}}\left[1+\frac{1-P_{N,c}}{nP_{N,c}}\right]$$

que también se puede expresar como

$$(N-n)N(B_c)\frac{S^2(B_c)}{n}\left[1+\frac{1-P_{N,c}}{nP_{N,c}}\right] = \frac{N-n}{n}NP_{N,c}S^2(B_c)\left[1+\frac{1-P_{N,c}}{nP_{N,c}}\right],$$

donde $n^{-1}(N-n)$ se interpreta como el predominio de la comuestra frente a la muestra.

Se puede definir otro estimador del subtotalizador $\hat{Y}_{+,HT}(B_c) = \sum_{i=1}^{N} Y_i 1_{B_c}(i) I_i^{(a}$ donde $I_i^{(a} = \pi_i^{-1} I_i$ si el diseño es probabilístico. Este estimador no precisa conocer el recontador. Corresponde al estimador de Horvitz-Thompson de la variable $Y1_{B_c}$, es decir, en la p.f. $\{(i, x_i, Y_i 1_{B_c}(i)), i = 1, \cdots, N\}$, y por tanto es insesgado para el total de esa variable que coincide con $Y_+(B_c)$.

En MAS(N, n), coincide con $\hat{Y}_{+,se}(B_c) = Nn^{-1} \sum_{i=1}^{N} Y_i 1_{B_c}(i) I_i$, el estimador por simple expansión de esta variable transformada, de ahí la notación. Por otra parte,

$$\mathrm{E}_p \left[n^{-1} \sum_{i=1}^{N} Y_i 1_{B_c}(i) I_i \right] = N^{-1} Y_+(B_c) = P_{N,c} \bar{Y}(B_c)$$

por tanto $\mathrm{E}_p \left[\hat{Y}_{+,se}(B_c) \right] = N(B_c) \bar{Y}(B_c) = Y_+(B_c)$, otra forma de ver su insesgadez. Si se considera la variable $Y_i = 1$, se obtiene también $\mathrm{E}_p \left[n^{-1} \sum_{i=1}^{N} 1_{B_c}(i) I_i \right] = N^{-1} N(B_c) = P_{N,c}$, es decir $\mathrm{E}_p \left[\hat{P}_{N,c} \right] = N^{-1} N(B_c) = P_{N,c}$ donde el estimador es la proporción muestral de elementos de la clase.

Para conocer la varianza de este segundo estimador del subtotalizador, se tiene esta cadena de igualdades inspirada en Hedayat&Sinha(1991)

$$\mathrm{D}_p^2 \left[\hat{Y}_{+,se}(B_c) \right] = (N - n) N \frac{1}{n} \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i 1_{B_c}(i) - \overline{Y 1_{B_c}} \right)^2 =$$

$$= \frac{N-n}{n} \frac{N}{N-1} \left[\sum_{i \in B_c} \left(Y_i - P_{N,c} \bar{Y}(B_c) \right)^2 + \sum_{i \notin B_c} \left(0 - P_{N,c} \bar{Y}(B_c) \right)^2 \right] =$$

$$= \frac{N-n}{n} \frac{N}{N-1} \left[\sum_{i \in B_c} \left(Y_i - \bar{Y}(B_c) \right)^2 + N(B_c) \left(\left(1 - P_{N,c} \right) \bar{Y}(B_c) \right)^2 + \left(N - N(B_c) \right) \left(P_{N,c} \bar{Y}(B_c) \right)^2 \right]$$

$$=$$

$$= \frac{N-n}{n} \left[N \frac{N(B_c) - 1}{N-1} S^2(B_c) + \frac{N}{N-1} N P_{N,c} \left(\left(1 - P_{N,c} \right) \bar{Y}(B_c) \right)^2 \right.$$
$$\left. + \frac{N}{N-1} N \left(1 - P_{N,c} \right) \left(P_{N,c} \bar{Y}(B_c) \right)^2 \right]$$

Un caso particular interesante consiste en estimar el recontador cuando los tamaños de clase son desconocidos mediante $\hat{N}(B_c) = \sum_{i=1}^{N} 1_{B_c}(i) I_i^{(a}$. Sea $Y_i = 1, \forall i$, en este caso la p.f. es $\{(i, x_i, 1_{B_c}(i)), i = 1, \cdots, N\}$ e $\left(\mathrm{Y}\, 1_{B_c} \right)_+ = N(B_c)$. En MAS, se puede construir un estimador $\hat{N}(B_c) = {}^{N}/_{n} \sum_{i=1}^{N} 1_{B_c}(i) I_i = N\, n(B_c)/n = N \hat{P}_{N,c}$ que es insesgado con varianza

$$\mathrm{D}_p^2\left[\widehat{N}(B_c)\right] = \frac{N-n}{n}\left[\frac{N}{N-1}NP_{N,c}\left(1-P_{N,c}\right)^2 + \frac{N}{N-1}N\left(1-P_{N,c}\right)\left(P_{N,c}\right)^2\right] =$$

$$= \frac{N-n}{n}\frac{N}{N-1}NP_{N,c}\left(1-P_{N,c}\right) = N^2\frac{N-n}{N-1}\frac{P_{N,c}\left(1-P_{N,c}\right)}{n}$$

porque $S^2(B_c) = 0$ e $\bar{Y}^2(B_c) = 1$ y del mismo modo

$$\mathrm{D}_p^2\left[\hat{P}_{N,c}\right] = \frac{N-n}{N-1}\frac{P_{N,c}\left(1-P_{N,c}\right)}{n}$$

corresponden a las que se obtendrían teniendo en cuenta la distribución hipergeométrica.

En el caso general, la igualdad obtenida permite mayorar la varianza del siguiente modo

$$\mathrm{D}_p^2\left[\hat{Y}_{+,se}(B_c)\right] < \frac{N-n}{n}\left[NP_{N,c}S^2(B_c) + \frac{N}{N-1}NP_{N,c}\left(1-P_{N,c}\right)\bar{Y}^2(B_c)\right]$$

debido a que $(N-1)^{-1}(N(B_c)-1) < N^{-1}N(B_c) < 1$.

Para dar una interpretación más fluida, también se puede aproximar mediante

$$\mathrm{D}_p^2\left[\hat{Y}_{+,se}(B_c)\right] \approxeq \frac{N-n}{n}\left[NP_{N,c}S^2(B_c) + NP_{N,c}\left(1-P_{N,c}\right)\bar{Y}^2(B_c)\right] =$$

$$= \frac{N-n}{n}NP_{N,c}\left[S^2(B_c) + \left(1-P_{N,c}\right)\bar{Y}^2(B_c)\right]$$

porque la aproximación del primer sumando es por exceso y la segunda, por defecto, pero adecuada si N es grande.

Si se aplica esta igualdad aproximada al caso particular del recontador se verifica que

$$\mathrm{D}_p^2\left[\widehat{N}(B_c)\right] \approxeq \frac{N-n}{n}NP_{N,c}\left(1-P_{N,c}\right) = \frac{1-f}{n}N^2P_{N,c}\left(1-P_{N,c}\right)$$

porque $S^2(B_c) = 0$ e $\bar{Y}^2(B_c) = 1$ y del mismo modo

$$\mathrm{D}_p^2\left[\hat{P}_{N,c}\right] \approxeq \frac{1-f}{n}P_{N,c}\left(1-P_{N,c}\right)$$

ambas se asemejan a las habituales de la proporción y el total de clase y resulta una aproximación por defecto.

Por otra parte, se ha visto que si $nP_{N,c}$ es grande de tal modo que se puedan ignorar los términos cúbicos y exponenciales, se tiene que

$$\mathrm{D}_p^2\left[\hat{Y}_{+,e}(B_c)\right] \approxeq \frac{N-n}{n}NP_{N,c}S^2(B_c)\left[1 + \frac{1-P_{N,c}}{nP_{N,c}}\right]$$

con lo cual para que

$$\mathrm{D}_p^2\left[\hat{Y}_{+,e}(B_c)\right] \leq \mathrm{D}_p^2\left[\hat{Y}_{+,se}(B_c)\right]$$

se debe verificar

$$\frac{N-n}{n} N P_{N,c} S^2(B_c) \frac{1-P_{N,c}}{n P_{N,c}} \leq \frac{N-n}{n} N P_{N,c}\left(1-P_{N,c}\right) \bar{Y}^2(B_c)$$

$$\frac{S^2(B_c)}{n P_{N,c}} \leq \bar{Y}^2(B_c) \Rightarrow \frac{S^2(B_c)}{\bar{Y}^2(B_c)} \leq n P_{N,c}$$

es decir $nP_{N,c}$ debe ser mayor que el cuadrado del coeficiente de variación de la variable en la celdilla. Por consiguiente, $\hat{Y}_{+,se}(B_c)$ tiene una varianza mayor cuando el número esperado de elementos en la clase, $nP_{N,c}$, sea grande bien porque el tamaño de la muestra lo sea, bien porque la clase sea relativamente abundante.

Además, si B_c = U, de ambas fórmulas aproximadas resulta la expresión habitual de la varianza para el estimador del total en MAS

$$\frac{N-n}{n} N S^2 = \frac{1-f}{n} N^2 S^2$$

3.1.3 Estimación del variador

Sea I_{ij} la variable indicatriz de pertenencia simultánea a la muestra de los elementos i y j. Se puede construir un estimador del variador del siguiente modo

$$\hat{S}^2(B_c) = \frac{1}{2} \frac{\sum\sum_{i \neq j} \left(Y_i - Y_j\right)^2 1_{B_c}(i) 1_{B_c}(j) I_{ij}}{\left(\sum\sum_{i \neq j} 1_{B_c}(i) 1_{B_c}(j) I_{ij}\right) \vee 2}$$

Teorema 4. En MAS(N,n), se verifica que

$$\mathrm{E}_p\left[\hat{S}^2(B_c)\right] = S^2(B_c)[1 - p(n(B_c)=0) - p(n(B_c)=1)]$$

Corolario 6. En MAS(N,n), si $S^2(B_c) \neq 0$ se verifica que el sesgo relativo es $\left|\mathrm{b}_p\left[\hat{S}^2(B_c)\right]\right| = p(n(B_c)=0) + p(n(B_c)=1)$

Además, se verifican las siguientes acotaciones

$$\left(1 - P_{N,c} - f + {}^1/_N\right)^n \leq p(n(B_c)=0) \leq \left(1-P_{N,c}\right)^n$$

$$n P_{N,c}\left(1-P_{N,c}\right)\left(1 - P_{N,c} - f + {}^2/_N\right)^{n-2} \leq p(n(B_c)=1) \leq n P_{N,c}\left(1 - P_{N,c} + {}^1/_N\right)^{n-1}$$

con las desigualdades estrictas para $n > 1$.

Corolario 7. En MAS(N,n), se verifica que

$$\mathrm{E}_p\left[\hat{S}^2(B_c)\right] = S^2(B_c)[1-R]$$

con $0 \leq R < e^{-nP_{N,c}}(1+nP_{N,c}e^{P_{N,c}})$.

Si $nP_{N,c}$ es grande, un estimador aproximado de la varianza del estimador del promediador es

$$\widehat{\mathrm{D}}_p^2\left[\widehat{\overline{Y}}(B_c)\right] \coloneqq \hat{S}^2(B_c)\frac{1-f}{nP_{N,c}}\left[1+\frac{1-P_{N,c}}{nP_{N,c}}\right]$$

En las clases donde no caigan elementos del universo, $P_{N,c}=0$, carecería de sentido estimar la varianza. En las que $P_{N,c}$ sea pequeño, puede ofrecer una estima grande e inestable.

Para el sesgo del cuadrado del estimador, se puede emplear

$$\widehat{\mathrm{B}}_p\left[\widehat{\overline{Y}}^2(B_c)\right] \cong \hat{S}^2(B_c)\sum_{m=1}^{n}\left(\frac{1}{m}-\frac{1}{N(B_c)}\right)p(n(B_c)=m)$$

3.2 Muestreo estratificado

3.2.1 Promediador

Se estudiarán estrategias basadas en el muestreo estratificado con MAS en cada estrato. Considérese una partición, a priori, del universo $\mathrm{U} = \mathrm{U}_1 \sqcup \cdots \sqcup \mathrm{U}_l \sqcup \cdots \sqcup \mathrm{U}_L$. Sea $N_l = \#\mathrm{U}_l$ de tal modo que $\sum_{l=1}^{L} N_l = N$. Este diseño que consta de MAS(N_l,n_l) en cada estrato U_l, se denotará MASE$\big((N_l,n_l)\ l=1,\cdots,L\big)$.

Se define un nuevo descriptor aplicativo, el ponderador $W_l(B_c) = N^{-1}(B_c)N_l(B_c)$, donde $N_l(B_c) = \#(B_c \cap \mathrm{U}_l)$, que se interpreta como el peso de la subclase del estrato dentro de toda la clase, de aquí la denominación del descriptor. $W_l(\cdot)$ se puede ver, a su vez, como cociente de dos descriptores aplicativos. Sea $W_l = W_l(\mathrm{U}) = N^{-1}N_l$.

Como $\overline{Y}(B_c) = \sum_{l=1}^{L} W_l(B_c)\overline{Y}_l(B_c)$ donde $\overline{Y}_l(B_c) = N_l^{-1}(B_c)Y_{l,+}(B_c)$ se puede definir el siguiente estimador del promediador en MASE$\big((N_l,n_l)\ l=1,\cdots,L\big)$.

$$\widehat{\overline{Y}_{st}}(B_c) \coloneqq \sum_{l=1}^{L} W_l(B_c)\,\widehat{\overline{Y}}_l(B_c) = \sum_{l=1}^{L}\frac{N_l(B_c)}{N(B_c)}\,\widehat{\overline{Y}}_l(B_c) = \frac{1}{N(B_c)}\sum_{l=1}^{L} N_l(B_c)\,\widehat{\overline{Y}}_l(B_c)$$

Teorema 5. En MASE$\big((N_l,n_l)\ l=1,\cdots,L\big)$, $\mathrm{E}_p\left[\widehat{\overline{Y}_{st}}(B_c)\right] = \overline{Y}(B_c) - R$ con R verificando que $\min_l \overline{Y}_l(B_c)p_{U_l}(n_l(B_c)=0) \leq R \leq \max_l \overline{Y}_l(B_c)p_{U_l}(n_l(B_c)=0)$, donde $p_{U_l}(\cdot)$ corresponde a MAS(N_l,n_l).

Para variables positivas, una situación poco deseable consiste en que algún estrato con una media alta y una probabilidad apreciable de que no haya elementos en la submuestra correspondiente pese dentro del universo. Como consecuencia del teorema, se puede acotar el sesgo para este tipo de variables del siguiente modo

Corolario 8. Si $Y > 0$, se verifica que

$$-\max_l \bar{Y}_l(B_c) p_{U_l}(n_l(B_c) = 0) \leq \mathrm{B}_p\left[\widehat{\bar{Y}_{st}}(B_c)\right] \leq -\min_l \bar{Y}_l(B_c)\, p_{U_l}(n_l(B_c) = 0).$$

La cota $\max_{B_c} |Y_i|\, e^{-nP_{N,c}}$ puede ser fina y sin embargo no serlo $\max_{B_c \cap U_l} |Y_i|\, e^{-n_l P_{N_l,c}}$ en algunos de los estratos, por tanto, hasta podría convenir no estratificar. Los sesgos que resultan de esta definición rigurosa podrían acabar por acumularse, no se compensan los de un estrato con los de otro si la variable es positiva.

Con respecto a la varianza, se verifica que

Teorema 6. En MASE$\big((N_l, n_l)\ l = 1, \cdots, L\big)$,

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}_{st}}(B_c)\right] = \sum_{l=1}^{L} W_l^2(B_c)\left[S_l^2(B_c)\overline{\left(\frac{1}{n_l^+(B_c)} - \frac{1}{N_l(B_c)}\right)} + \bar{Y}_l^2(B_c) p_{U_l}(n_l(B_c) = 0)\right]$$

$$\left[1 - p_{U_l}(n_l(B_c) = 0)\right]$$

Una posible cota es

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}_{st}}(B_c)\right] \leq \sum_{l=1}^{L} W_l^2(B_c)\, S_l^2(B_c)\overline{\left(\frac{1}{n_l^+(B_c)} - \frac{1}{N_l(B_c)}\right)} + \frac{1}{4}\sum_{l=1}^{L} W_l^2(B_c)\, \bar{Y}_l^2(B_c)$$

Este segundo sumando se puede acotar, a su vez, de los dos modos siguientes

$$\min_l W_l(B_c)\left(\frac{\bar{Y}_l(B_c)}{2}\right)^2 \leq \frac{1}{4}\sum_{l=1}^{L} W_l^2(B_c)\, \bar{Y}_l^2(B_c) \leq \max_l W_l\,(B_c)\left(\frac{\bar{Y}_l(B_c)}{2}\right)^2$$

o

$$\min_l \left(\frac{\bar{Y}_l(B_c)}{2}\right)^2 \sum_{l=1}^{L} W_l^2(B_c) \leq \frac{1}{4}\sum_{l=1}^{L} W_l^2(B_c)\, \bar{Y}_l^2(B_c) \leq \max_l \left(\frac{\bar{Y}_l(B_c)}{2}\right)^2 \sum_{l=1}^{L} W_l^2(B_c)$$

donde $\sum_{l=1}^{L} W_l^2(B_c)$ se puede interpretar como una medida de concentración de los elementos de la clase entre los L estratos. Además $\min_l W_l(B_c) \leq \sum_{l=1}^{L} W_l^2(B_c) \leq \max_l W_l\,(B_c)$. Por otra parte, también se puede considerar

$$\sum_{l=1}^{L} W_l^2(B_c)\,\bar{Y}_l^2(B_c) = \bar{Y}^2(B_c)\frac{\sum_{l=1}^{L} W_l^2(B_c)\,\bar{Y}_l^2(B_c)}{\left(\sum_{l=1}^{L} W_l(B_c)\,\bar{Y}_l(B_c)\right)^2} = \bar{Y}^2(B_c)\sum_{l=1}^{L}\left(\frac{W_l(B_c)\bar{Y}_l(B_c)}{\sum_{l=1}^{L} W_l(B_c)\,\bar{Y}_l(B_c)}\right)^2$$

con $L^{-1} \leq \sum_{l=1}^{L}\left(\frac{W_l(B_c)\bar{Y}_l(B_c)}{\sum_{l=1}^{L} W_l(B_c)\bar{Y}_l(B_c)}\right)^2 \leq 1$, por tanto se llega a que el segundo sumando de esa cota no rebasa al obtenido mediante el correspondiente MAS$(\sum_{l=1}^{L} N_l, \sum_{l=1}^{L} n_l)$.

En ocasiones, se podrá utilizar la siguiente aproximación

$$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right] \cong \sum_{l=1}^{L} W_l^2(B_c)\left[(1-f_l)\frac{S_l^2(B_c)}{n_l P_{N_l,c}}\left(1+\frac{1-P_{N_l,c}}{n_l P_{N_l,c}}\right)\right]$$

aunque se pueden acumular errores porque el segundo sumando depende de cuánto valgan $\bar{Y}_l(B_c)$. Para hacerse una idea de la magnitud de este resto se puede estimar $0'25\bar{Y}_l^2(B_c)$ mediante $0'25\,\widehat{\bar{Y}_l^2}(B_c)$. Incluso se ha visto un estimador de su sesgo. Si no se puede despreciar el segundo sumando, se puede aproximar o acotar.

3.2.2 Estimación de la varianza

Se puede estimar la varianza mediante

$$\widehat{\mathrm{D}}_p^2\left[\widehat{\bar{Y}_{st}}(B_c)\right] \coloneqq \sum_{l=1}^{L} W_l^2(B_c)\left[\hat{S}_l^2(B_c)\overline{\left(\frac{1}{n_l^+(B_c)}-\frac{1}{N_l(B_c)}\right)} + \widehat{\bar{Y}}_l^2(B_c)p_{U_l}(n_l(B_c)=0)\right]$$

$$\left[1-p_{U_l}(n_l(B_c)=0)\right]$$

Prescindiendo de estos segundos sumandos, se puede emplear un estimador aproximado

$$\widehat{\mathrm{D}}_p^2\left[\widehat{\bar{Y}}(B_c)\right] \coloneqq \sum_{l=1}^{L} W_l^2(B_c)\left[(1-f_l)\frac{\hat{S}_l^2(B_c)}{n_l P_{N_l,c}}\left(1+\frac{1-P_{N_l,c}}{n_l P_{N_l,c}}\right)\right]$$

A continuación, a modo de resumen, se presentan unas tablas con la esperanza y la varianza, con el sesgo y la varianza y con algunos de los estimadores de la varianza propuestos.

Tabla 2

Expresiones de la esperanza y varianza de los estimadores del promediador

Diseño	*Esperanza*	*Varianza*
Muestreo aleatorio simple sin reemplazamiento (MAS)	$\mathrm{E}_p\left[\widehat{\bar{Y}}(B_c)\right] = [1 - p(n(B_c) = 0)]\bar{Y}(B_c)$	$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right] = [1 - p(n(B_c) = 0)]\left[S^2(B_c)\overline{\left(\frac{1}{n^+(B_c)} - \frac{1}{N(B_c)}\right)}\right] + \bar{Y}^2(B_c)p(n(B_c) = 0)$
Muestreo estratificado (MASE)	$\mathrm{E}_p\left[\widehat{\bar{Y}_{st}}(B_c)\right] = \sum_{l=1}^{L}\{1 - p_{U_l}(n_l(B_c) = 0) W_l(B_c)\bar{Y}_l(B_c)\}$	$\mathrm{D}_p^2\left[\widehat{\bar{Y}_{st}}(B_c)\right] = \sum_{l=1}^{L}\{W_l^2(B_c)\left[1 - p_{U_l}(n_l(B_c) = 0)\right]\left[S_l^2(B_c)\overline{\left(\frac{1}{n_l^+(B_c)} - \frac{1}{N_l(B_c)}\right)} + \bar{Y}_l^2(B_c)p_{U_l}(n_l(B_c) = 0)\right]\}$

Tabla 3

Expresiones del sesgo y la varianza de los estimadores

Estrategia	*Sesgo*	*Varianza*
MAS Promediador	$\mathrm{B}_p\left[\widehat{\bar{Y}}(B_c)\right] = -\bar{Y}(B_c)p(n(B_c) = 0)$	$\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right] \cong (1 - f)\frac{S^2(B_c)}{nP_{N,c}}\left(1 + \frac{1 - P_{N,c}}{nP_{N,c}}\right)$
Subtotalizador expansión	$\mathrm{B}_p\left[\hat{Y}_{+,e}(B_c)\right] = -Y_+(B_c)p(n(B_c) = 0)$	$\mathrm{D}_p^2\left[\hat{Y}_{+,e}(B_c)\right] \cong \left(NP_{N,c}\right)^2(1 - f)\frac{S^2(B_c)}{nP_{N,c}}\left(1 + \frac{1 - P_{N,c}}{nP_{N,c}}\right)$
Subtotalizador simple expansión	$\mathrm{B}_p\left[\hat{Y}_{+,se}(B_c)\right] = 0$	$\mathrm{D}_p^2\left[\hat{Y}_{+,se}(B_c)\right] \cong \left(NP_{N,c}\right)^2(1 - f)\frac{S^2(B_c)}{nP_{N,c}}\left(1 + \frac{1 - P_{N,c}}{S^2(B_c)/\bar{Y}^2(B_c)}\right)$
Recontador	$\mathrm{B}_p\left[\hat{N}(B_c)\right] = 0$	$\mathrm{D}_p^2\left[\hat{N}(B_c)\right] \cong \left(NP_{N,c}\right)^2(1 - f)\frac{(1 - P_{N,c})}{nP_{N,c}}$
MASE Promediador	$\mathrm{B}_p\left[\widehat{\bar{Y}_{st}}(B_c)\right] = \sum_{l=1}^{L}\{W_l(B_c)\bar{Y}_l(B_c) p_{U_l}(n_l(B_c) = 0)\}$	$\mathrm{D}_p^2\left[\widehat{\bar{Y}_{st}}(B_c)\right] \cong \sum_{l=1}^{L}\{W_l^2(B_c)(1 - f_l)\frac{S_l^2(B_c)}{n_l P_{N_l,c}}\left(1 + \frac{1 - P_{N_l,c}}{n_l P_{N_l,c}}\right)\}$

Tabla 4

Expresiones de los estimadores de la varianza

Estrategia	*Estimador*
MAS Variador	$\hat{S}^2(B_c) = \frac{1}{2}\frac{\sum\sum_{i\neq j} \left(Y_i - Y_j\right)^2 1_{B_c}(i)1_{B_c}(j)I_{ij}}{\left(\sum\sum_{i\neq j} 1_{B_c}(i)1_{B_c}(j)I_{ij}\right) \vee 2}$
MAS Promediador	$\widehat{\mathrm{D}}_p^2\left[\widehat{\overline{Y}}(B_c)\right] \approxeq (1-f)\frac{\hat{S}^2(B_c)}{nP_{N,c}}\left(1+\frac{1-P_{N,c}}{nP_{N,c}}\right)$
Subtotalizador expansión	$\widehat{\mathrm{D}}_p^2\left[\hat{Y}_{+,e}(B_c)\right] \approxeq \left(NP_{N,c}\right)^2(1-f)\frac{\hat{S}^2(B_c)}{nP_{N,c}}\left(1+\frac{1-P_{N,c}}{nP_{N,c}}\right)$
Subtotalizador simple expansión	$\widehat{\mathrm{D}}_p^2\left[\hat{Y}_{+,se}(B_c)\right] \approxeq \left(N\hat{P}_{N,c}\right)^2(1-f)\frac{\hat{S}^2(B_c)}{n\hat{P}_{N,c}}\left(1+\frac{1-\hat{P}_{N,c}}{\hat{S}^2(B_c)/\widehat{\overline{Y}}^2(B_c)}\right)$
MASE Promediador	$\widehat{\mathrm{D}}_p^2\left[\widehat{\overline{Y}_{st}}(B_c)\right] \approxeq \sum_{l=1}^{L} W_l^2(B_c)\,(1-f_l)\frac{\hat{S}_l^2(B_c)}{n_l P_{N_l,c}}\left(1+\frac{1-P_{N_l,c}}{n_l P_{N_l,c}}\right)$

4. Aplicaciones de estos descriptores

En la siguiente subsección se estudia el regresograma como descriptor aplicativo al igual que su estimador. En la segunda se estudia una situación especial, el estimador regresograma aplicado a la descripción de la propia variable. En la última se muestra como traducir geométricamente estos descriptores para construir un pictograma.

4.1 De su empleo en el estudio del regresograma

Si se representa un promediador con el propósito de aproximar la regresión de Y sobre X siendo ambos caracteres cuantitativos, se está en presencia del regresograma, con una partición de toda la recta real o, más habitual, de un intervalo, por ejemplo, de la semirrecta $[0,\infty)$ o del semisegmento $(a,b]$.

Se dispone de unos números reales $t_0 < t_1 < \cdots < t_c < \cdots < t_C$, que dividen el conjunto donde toma valores X, de tal forma que la partición está formada por clases centrales que son semisegmentos y la última y la primera, las clases cola, admiten diferentes soluciones dependiendo del problema tratado: una semirrecta o un intervalo abierto o cerrado.

4.1.1 Regresograma en la población finita o descriptor regresograma

Se presenta, a continuación, un ejemplo extraído de una operación básica exhaustiva, la explotación estadística del Rexistro de buques pesqueiros de la Comunidad Autónoma, con el arqueo en GT de la flota pesquera con puerto base en Galicia según tramos de eslora en metros, como ejemplo de subtotalizador. Del mismo modo, la línea de número constituye un recontador.

Tabla 5

Buques según eslora en Galicia. Año 2011

Eslora (m)	*Menos de 12*	*De 12 a 18*	*De 18 a 24*	*24 y más*
Número	3982	296	107	349
Arqueo (GT)	6218	5948	8346	142700

Fuente: Consellería do Medio Rural e do Mar. *Rexistro de buques pesqueiros de Galicia*, en www.pescadegalicia.com

En este ejemplo, la variable de clasificación es la eslora discretizada que corresponde al carácter X y la variable de estudio, Y, es el arqueo. Obsérvese la partición de la variable de clasificación con los dos intervalos centrales de igual longitud, pero con el de la izquierda y el de la derecha de diferente amplitud, el segundo incluso sin acotar. Considerando las filas de arqueo y recuentos, se construye el siguiente regresograma, ejemplo de promediador.

Tabla 6

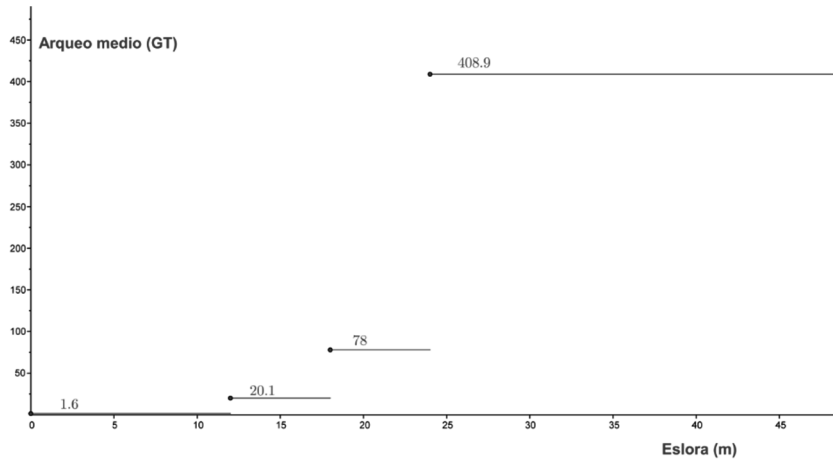
Arqueo medio en GT clasificado por tramos de eslora

Eslora (m)	*[0,12)*	*[12,18)*	*[18,24)*	*[24, ∞)*
Arqueo medio (GT)	1,6	20,1	78,0	408,9

Con él se pretende ver la relación entre el arqueo y la eslora de un buque, es decir presentar este en función de aquella. Gráficamente está constituido por una línea fragmentariamente constante. En abscisas, se coloca la variable de clasificación de la tabla y en ordenadas la variable de estudio.

Figura 1

Regresograma del arqueo en función de la eslora



Por brevedad, se supondrá que $(-\infty, t_0], (t_0, t_1], \cdots, (t_{c-1}, t_c], \cdots (t_{C-1}, t_C], (t_C, \infty)$. En el caso de intervalos, $(t_{c-1}, t_c]$ con $t_0 = a, t_C = b$. Sea $B_x = (l_x, L_x]$ la celdilla 'bin' de la partición que contiene a x y $1_{B_x}(\cdot)$ la variable indicatriz de pertenencia a ella. El regresograma de la p.f. es

$$\alpha_H(x) = \frac{\sum_{i=1}^{N} Y_i 1_{B_x}(x_i)}{\sum_{i=1}^{N} 1_{B_x}(x_i) \vee 1},$$

La precisión del denominador puede resultar redundante en estadística pública donde el recontador, que se puede abreviar $N_x = \sum_{i=1}^{N} 1_{B_x}(x_i)$, es siempre mayor que 1. Si se divide el recontador por N, se obtiene $P_{N,x} = N^{-1} N_x$.

El variador en este caso concreto es

$$S_x^2 = \frac{1}{\sum_{i=1}^{N} 1_{B_x}(x_i) - 1} \sum_{i=1}^{N} \left(Y_i - \alpha_H(x)\right)^2 1_{B_x}(x_i) = \frac{1}{N_x - 1} \sum_{j=1}^{N_x} \left(Y_j - \alpha_H(x)\right)^2$$

siempre que $N_x \geq 2$, como se está suponiendo. Su raíz se podría bautizar, como función de x, con la denominación de desviograma $S(x) = \left|\sqrt{S_x^2}\right|$ puesto que, en el caso de su cuadrado, el término variograma ya se ha empleado en otro contexto.

Una expresión alternativa es

$$S_x^2 = \frac{1}{N_x - 1} \left[\sum_{i=1}^{N} Y_i^2 \, 1_{B_x}(x_i) - \frac{\left(\sum_{i=1}^{N} Y_i 1_{B_x}(x_i)\right)^2}{\sum_{i=1}^{N} 1_{B_x}(x_i)} \right],$$

que permite interpretarlo como $\frac{N_x}{N_x - 1}$ multiplicado por el regresograma del cuadrado de Y menos el cuadrado del regresograma de Y.

4.1.2 Regresograma muestral o estimador regresograma

El estimador del regresograma cuando se realiza un MAS sería

$$\hat{\alpha}_H(x) = \frac{\sum_{i=1}^{N} Y_i 1_{B_x}(x_i) I_i}{\left(\sum_{i=1}^{N} 1_{B_x}(x_i) I_i\right) \vee 1} = \frac{\sum_{i=1}^{N} Y_i 1_{B_x}(x_i) I_i}{n_x \vee 1}$$

siendo $n_x = \sum_{i=1}^{N} 1_{B_x}(x_i) I_i$. Se verifica que $\mathrm{E}_p\left[\hat{\alpha}_H(x)\right] = \alpha_H(x)[1 - p(n_x = 0)]$ con $p(n_x = 0) < e^{-nP_{N,x}}$ y, por tanto, $\left|\mathrm{B}_p\left[\hat{\alpha}_H(x)\right]\right| \leq \max_{B_x} |Y_i| \, e^{-nP_{N,x}}$. Como ya se ha mencionado es frecuente en estadística pública que Y sea positiva, por tanto, $\alpha_H(x) > 0$, el sesgo es negativo $\mathrm{B}_p[\hat{\alpha}_H(x)] < 0$ es decir el estimador subestima el regresograma.

Por consiguiente, cuando $nP_{N,x}$ es grande, el sesgo resulta despreciable a efectos prácticos siempre que Y esté acotada o su máximo crezca según un orden polinómico en $nP_{N,x}$.

La varianza del estimador del regresograma verifica que

$$\mathrm{D}_p^2\,[\hat{\alpha}_H(x)] = \left[S_x^2\overline{\left(\frac{1}{n_x^+}-\frac{1}{N_x}\right)} + \alpha_H^2(x)p(n_x=0)\right][1-p(n_x=0)] \le S_x^2\overline{\left(\frac{1}{n_x^+}-\frac{1}{N_x}\right)} + R$$

con $0 < R < \max_{B_x} Y_i^2\, e^{-nP_{N,x}}$, o de modo aproximado cuando $nP_{N,x}$ sea grande

$$\mathrm{D}_p^2\,[\hat{\alpha}_H(x)] \cong (1-f)\frac{S_x^2}{nP_{N,x}}\left[1+\frac{1-P_{N,x}}{nP_{N,x}}\right]$$

Esta expresión también constituye una aproximación del sesgo del cuadrado del regresograma

$$\mathrm{B}_p\,[\hat{\alpha}_H^2(x)] \le S_x^2\overline{\left(\frac{1}{n_x^+}-\frac{1}{N_x}\right)} \cong (1-f)\frac{S_x^2}{nP_{N,x}}\left[1+\frac{1-P_{N,x}}{nP_{N,x}}\right]$$

Se construye un estimador del cuadrado del desviograma del siguiente modo

$$\widehat{S_x^2} = \frac{1}{2}\frac{\sum\sum_{i\neq j}\left(Y_i-Y_j\right)^2 1_{B_x}(x_i)1_{B_x}(x_j)I_{ij}}{\left(\sum\sum_{i\neq j}1_{B_x}(x_i)1_{B_x}(x_j)I_{ij}\right)\vee 2}$$

En MAS, se verifica que $\mathrm{E}_p\left[\widehat{S_x^2}\right] = S_x^2[1-p(n_x=0)-p(n_x=1)]$ por tanto $[1-p(n_x=0)-p(n_x=1)]^{-1}\widehat{S_x^2}$ sería un estimador insesgado.

4.2 De su empleo en la descripción de la propia variable

Se podría emplear el promediador de la p.f. con la propia variable X, es decir que coincida la variable de estudio y de clasificación, por ejemplo, cuando se disponga de una partición definida por un estándar y se quiera comparar la misma variable en dos situaciones diferentes, temporales o espaciales. Si se crea un factor a partir de ella, discretizándola, el promediador puede ser interesante para describirla como complemento del histograma sobre todo para el estudio de las clases inicial y final, especialmente esta última cuando $X > 0$.

La tabla de frecuencias, y el histograma asociado, transmite una idea clara de las características globales de la distribución o de lo que acontece en el término central de ella. Sin embargo, no ocurre lo mismo en las clases cola donde es mucho más difícil percibir lo que sucede.

El promediador también sirve como chequeo de la hipótesis de uniformidad aproximada dentro de las celdillas. Si se sostiene, la tabulación es más interesante ya que admite las aproximaciones típicas mediante las marcas de clase. O simplemente, por cuestiones de difusión, para evitar el problema de las clases cola donde muchas veces resulta inadecuado o imposible utilizar la marca de clase.

Si no hay razones para emplear una partición determinada, una buena práctica puede ser emplear inicialmente una partición de igual amplitud y luego agregar celdillas, lo que desde luego producirá un sobresuavizado de las celdillas unificadas ‘colapsadas’. Por

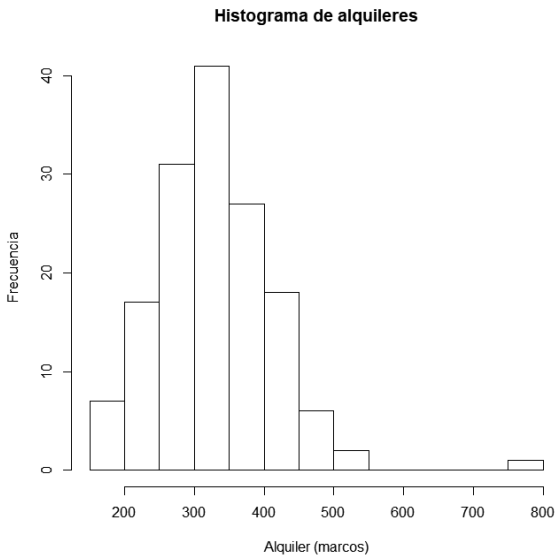
ejemplo, se puede emplear igual amplitud para la macrodepuración y amplitud variable para la difusión. También hay que tener en cuenta que, en algunas situaciones, una transformación previa puede convertir la amplitud variable en fija.

Conviene percatarse de que, en esta situación concreta, la mayoración de R en la expresión de la varianza bajo el diseño, en la que intervienen $\max_{B_c}\left|Y_i - Y_j\right|^2$ y $\max_{B_c} Y_i^2$, se puede refinar notablemente dentro de cada intervalo central en función de sus extremos inferior y superior; y también en las clases colas en función de su extremo inferior o superior, respectivamente.

Se estudiará a continuación un ejemplo extraído de un libro antiguo (Schott, 1928) que corresponde a una muestra de 150 viviendas del mismo tipo en la que se analiza su alquiler anual en marcos. El total muestral asciende a 49.871 marcos. Entre las razones para emplear este libro, existe una histórica. En su sección correspondiente a correlación, se menciona el empleo del regresograma aunque no se utiliza este vocablo ni se presenta un ejemplo numérico.

Figura 2

Distribución de frecuencias del alquiler anual en marcos



A la vista del histograma, obtenido mediante el paquete R, se unificarán los últimos intervalos englobándolos en un intervalo abierto y se tabularán los valores de la muestra. En la primera fila de la tabla, se presentan los intervalos que se han empleado para discretizar la variable, B_c. La segunda es el número de viviendas de la muestra en cada uno de ellos, $n(B_c)$. Y la tercera es el valor total de los alquileres de la muestra en el correspondiente intervalo

Tabla 7

Muestra de alquileres anuales de viviendas en marcos

Alquiler (marcos)	*(150,200]*	*(200,250]*	*(250,300]*	*(300,350]*	*(350,400]*	*(400,450]*	*(450,500]*	*(500,∞)*
$\boldsymbol{n}(\cdot)$	7	17	31	41	27	18	6	3
Total de alquileres	1326	3872	8808	13341	10300	7656	2760	1808

A continuación, se presentan en otra tabla los datos y cálculos efectuados con ellos que se emplearían en el proceso de estimación. Obsérvese que, en las clases segunda, cuarta y sexta, el alquiler medio está muy próximo a la marca de clase. El penúltimo intervalo está muy concentrado debido a cinco valores iguales.

Tabla 8

Estimadores del promediador y de su varianza

$\boldsymbol{B_c}$	*(150,200]*	*(200,250]*	*(250,300]*	*(300,350]*	*(350,400]*	*(400,450]*	*(450,500]*	*(500,∞)*
$\boldsymbol{n}(\cdot)$	7	17	31	41	27	18	6	3
$\widehat{\boldsymbol{P}}_{\boldsymbol{N,c}}$	0’047	0’113	0’207	0’273	0’180	0’120	0’040	0’020
$\widehat{\overline{\boldsymbol{Y}}}(\cdot)$	189’4	227’8	284’1	325’4	381’5	425’3	460’0	602’7
$\widehat{\boldsymbol{S}}(\cdot)$	5’968	14’317	14’660	14’431	12’801	11’061	9’798	170’896
$\left(\widehat{\boldsymbol{S^2}}/\boldsymbol{n}\right)(\cdot)$	5’088	12’055	6’933	5’079	6’069	6’797	16’000	9735’111
$(\boldsymbol{1}+\boldsymbol{n^{-1}})(\cdot)$	1’143	1’059	1’032	1’024	1’037	1’056	1’167	1’333
$\left[\left(\widehat{\boldsymbol{S^2}}/\boldsymbol{n}\right)(\boldsymbol{1}+\boldsymbol{n^{-1}})\right](\cdot)$	5’815	12’765	7’156	5’203	6’294	7’175	18’667	12980’148

En las primeras filas, se presentan los estimadores aplicativos $\widehat{P}_{N,x}$, promediador y raíz del variador. Como no se dispone del recontador, $N(B_c)$, ni del tamaño poblacional, N, no se puede emplear directamente la fórmula introducida en la sección 3.1.3 pero se puede utilizar otros estimadores de la varianza del promediador construidos por el método de sustitución ‘plug-in’, la generalización del habitual $\widehat{S^2}(B_c)/\left(n\widehat{P}_{N,c}\right) = \widehat{S^2}(B_c)/n(B_c)$ o corrigiéndolo

$$\frac{\widehat{S^2}(B_c)}{n(B_c)}\left(1+\frac{1}{n(B_c)}\right),$$

cuyos resultados se muestran en las últimas filas de la tabla. En las clases centrales, es posible que funcionen adecuadamente estos estimadores, sobre todo, el último. En las otras, resulta más dudoso[1].

[1] En la clase cola derecha, resulta evidente observando la tabla. Si se estimase el siguiente término de la cota, el cúbico $(2/n(B_c))^3 \cong 0'3$, daría un incremento del 30% que habría que añadir al 33% de la corrección del estimador habitual. A la vista de los resultados, sería recomendable unificar las dos últimas clases e, incluso, las dos primeras no obstante este ejemplo sólo pretende ilustrar los resultados y comentarios que se han ido desarrollando en las primeras secciones del artículo.

4.3 De su empleo en la construcción de pictogramas

También se podrían emplear estos descriptores aplicativos para realizar otro tipo de diagramas o representaciones gráficas, por ejemplo, el subtotalizador para construir un pictograma en el que la superficie de los rectángulos fuera igual al total en la clase si no existen problemas de revelación con el extremo final de la partición, porque hay que fijarlo previamente. Precisamente para sobrellevar esta última situación se introdujo en la subsección anterior el promediador aplicado a la propia variable X.

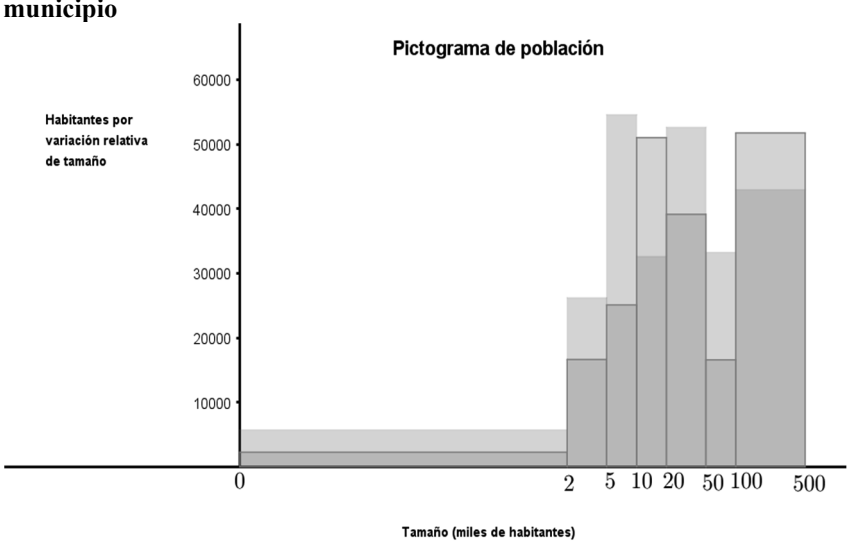
Como ejemplo concreto de este diagrama se puede representar la tabla «Concellos e a súa poboación clasificados polo número dos seus habitantes. Padrón municipal de habitantes. Ano 2014» (IGE, 2015). El carácter de clasificación X es la propia variable discretizada. Antes de su conversión en un pseudofactor, se transformará calculando el logaritmo decimal del número de habitantes del municipio para permitir una representación más clara. Es decir, se emplea una especie de papel semilogarítmico.

El tratamiento de las clases-cola no ofrece ningún problema de definición porque no existen municipios sin habitantes en Galicia y se conviene en considerar como primera clase el segmento [1, 2000] en lugar de la equivalente (0,2000] en esta situación concreta. Tampoco existen municipios con más de medio millón de habitantes.

Para la construcción del pictograma se operará con el subtotalizador. A partir de él, se define un nuevo descriptor aplicativo $Y_+(B_x)/(\log L_x - \log l_x)$ que determina la altura de los rectángulos que se emplearán en el pictograma.

Figura 3

Población de las provincias de A Coruña y Pontevedra según el tamaño del municipio



Como resultado de la transformación, se percibe cierta asimetría a la izquierda. Aunque los intervalos centrales resultantes no sean de igual amplitud, esta resulta más homogénea tras la transformación.

Con borde obscuro se representan los rectángulos correspondientes a la provincia de Pontevedra y con el claro, los de la provincia de A Coruña. No se representa la última clase (más de 500.000 habitantes) porque el subtotalizador es nulo en ambas provincias. Se observa que A Coruña tiene más población que Pontevedra en general y en la mayoría de los intervalos excepto en la última clase central (de 100.000 hasta 500.000) y en el intervalo (10.000, 20.000].

5. Conclusiones

El concepto de descriptor aplicativo introducido en este artículo permite extender la suavización a caracteres no cuantitativos. El proceso de codificación y presentación de valores medios a diferentes niveles de una clasificación jerárquica, con resultados cada vez más agregados, se puede ver como una suavización progresiva pudiendo llegar a una sobresuavización.

Se ha tratado la estimación de estos descriptores bajo dos diseños estándar. Se han calculado expresiones exactas para su sesgo y varianza empleando un método diferente a la linealización que han permitido ofrecer cotas y expresiones aproximadas, que resultan operativas en estadística pública.

En concreto, se ofrecen dos estimadores del subtotalizador en MAS, uno que se emplea cuando se conoce el recontador, la distribución de efectivos por clase, y otro cuando no se conoce. Se han comparado sus varianzas. Aunque el segundo presenta una varianza mayor si el número esperado de elementos en la clase es grande, puede resultar interesante como se ha mostrado en el caso particular de estimar el propio recontador.

Las expresiones de las varianzas obtenidas dependen de otro descriptor aplicativo, el variador. Se ha introducido un estimador para él en MAS del que se ha estudiado su sesgo. Con ello se puede estimar las varianzas del promediador y del subtotalizador en MAS y MASE.

Todas estas expresiones han permitido generalizar los resultados a MASE donde también se ha estudiado el sesgo y la varianza. Estas expresiones, además, permiten el estudio de la conveniencia de la estratificación ya que se puede analizar el proceso de acumulación de sesgos si se dispone de información previa como sucede habitualmente en estadística pública.

Como caso particular se estudia el regresograma con celdas de longitud variable, incluso no acotadas, que debería de tener un papel más activo en estadística pública, debido a su carácter marcadamente intuitivo dentro de los suavizadores locales. No parece casualidad que E. Engel, autor considerado por parte de la literatura como pionero o precursor en su utilización (Härdle&Linton, 1994), desempeñara su labor dentro de este ramo, llegó a ser director de órgano central de estadística. Se complementa con el desviograma como aplicación que mide la variabilidad.

Se obtienen, particularizando, estimadores del regresograma y del cuadrado del desviograma, así como de la varianza del primero. Cuando se emplean con la propia variable, sirven como complemento de la tabla de recuentos, con mayor interés en las clases-cola sobre todo en la superior de las pp.ff. con variables positivas de distribución asimétrica.

Por último, en la difusión, los resultados suelen presentarse relativizados para facilitar la comparación pero, en algunos casos, conviene mostrar una idea del volumen del que se derivan. Una posibilidad es emplear el pictograma que se ha construido basándose en el subtotalizador para que se aprecie el nivel de algunas de las variables fundamentales.

Los descriptores aplicativos y sus estimadores favorecerían la macrodepuración y explotación de los registros y sistemas de información de las administraciones tanto de forma directa, por muestreo o, incluso, con una combinación de explotación exhaustiva con una muestra ligera para obtener nuevos datos o bien confirmar o corregir los existentes en las bases de datos. El tratamiento por muestreo también facilitaría la imputación y el casamiento 'matching' entre sistemas de información para fines estadísticos puesto que permitiría controlar mejor los errores ajenos al muestreo y sería más proporcionado en lo que respecta a confidencialidad.

Estos objetos acelerarían la automatización dada la reiteración de su utilización en procesos de tabulación, representación gráfica, cálculo de errores de muestreo… Permiten un tratamiento conjunto más claro. Por ejemplo, las variables sintéticas (Silva, 1997) habituales suelen ser operaciones aritméticas sobre estos descriptores definidos en el mismo conjunto origen o las expresiones de las varianzas de los estimadores resultan de operaciones realizadas con esta tipo de aplicaciones salvo constantes que dependen del diseño. El propio cuadrado del desviograma se puede expresar como la diferencia de regresogramas salvo un coeficiente próximo a la unidad.

En resumen, permitirían sistematizar la explotación de microdatos. Suponen una matematización del proceso estadístico con una formulación más acorde con las corrientes actuales que reduciría la brecha entre estadística teórica y estadística pública.

Anexo

Demostración del teorema 1

$$\mathrm{E}_p\left[\widehat{\bar{Y}}(B_c)\right] = \sum_{m=0}^{n} \mathrm{E}_p\left[\widehat{\bar{Y}}(B_c)|n(B_c)=m\right] p(n(B_c)=m) = \bar{Y}(B_c)[1-p(n(B_c)=0)]$$

Demostración del lema 1. Si $n \leq N - N(B_c)$, $p(n(B_c)=0) < \left(1-P_{N,c}\right)^n < e^{-nP_{N,c}}$ por la desigualdad $1-u \leq e^{-u}$ que es estricta salvo en cero y se está suponiendo $P_{N,c} > 0$. Si $n > N - N(B_c)$, la probabilidad sería nula, verificándose la desigualdad de modo trivial.

Demostración del teorema 2. Como, en general, se verifica que $\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)\right] = \mathrm{E}_p\left[\widehat{\bar{Y}}^2(B_c)\right] - \mathrm{E}_p^2\left[\widehat{\bar{Y}}(B_c)\right]$ habrá que calcular $\mathrm{E}_p\left[\widehat{\bar{Y}}^2(B_c)\right]$ porque el segundo sumando es conocido, $\mathrm{E}_p^2\left[\widehat{\bar{Y}}(B_c)\right] = \bar{Y}^2(B_c)[1 - p(n(B_c) = 0)]^2$. El primero admite la siguiente expresión

$$\mathrm{E}_p\left[\widehat{\bar{Y}}^2(B_c)\right] = \sum_{m=0}^{n}\left\{\mathrm{D}_p^2\left[\widehat{\bar{Y}}(B_c)|n(B_c)=m\right] + \mathrm{E}_p^2\left[\widehat{\bar{Y}}(B_c)|n(B_c)=m\right]\right\}p(n(B_c)=m) =$$

$$= S^2(B_c)\sum_{m=1}^{n}\left(\frac{1}{m} - \frac{1}{N(B_c)}\right)p(n(B_c)=m) + \bar{Y}^2(B_c)[1 - p(n(B_c)=0)]$$

y de ambas igualdades se deduce el teorema.

Demostración del lema 3

$$\sum_{m=1}^{n}\left(\frac{1}{m} - \frac{1}{N(B_c)}\right)p(n(B_c)=m) = A - B$$

donde

$$B = \sum_{m=1}^{n}\frac{1}{N(B_c)}p(n(B_c)=m) = \frac{1 - p(n(B_c)=0)}{N(B_c)}$$

y

$$A = \sum_{m=1}^{n}\frac{1}{m}p(n(B_c)=m) = \sum_{m=1}^{n}\frac{1}{m}\frac{\binom{N(B_c)}{m}\binom{N-N(B_c)}{n-m}}{\binom{N}{n}}$$
$$= \sum_{m=1}^{n}\left(\frac{1}{m+1} + \frac{1}{(m+1)(m+2)} + \frac{2}{(m+1)(m+2)(m+3)}\right.$$
$$\left. + \frac{6}{m(m+1)(m+2)(m+3)}\right)\frac{\binom{N(B_c)}{m}\binom{N-N(B_c)}{n-m}}{\binom{N}{n}} = A_1 + A_2 + A_3 + A_4$$

Para estudiar estos cuatro sumandos positivos, considérese

$$A_1 = \frac{1}{(n+1)\frac{N(B_c)+1}{N+1}}\sum_{m=1}^{n}\frac{\binom{N(B_c)+1}{m+1}\binom{N-N(B_c)}{n-m}}{\binom{N+1}{n+1}} =$$

$$= \frac{1}{(n+1)\dfrac{N(B_c)+1}{N+1}} \sum_{m'=2}^{n'} \frac{\binom{N(B_c)+1}{m'}\binom{N-N(B_c)}{n'-m'}}{\binom{N+1}{n'}} =$$

$$= \frac{1}{(n+1)\dfrac{N(B_c)+1}{N+1}} [1 - p(m'=0) - p(m'=1)] <$$

$$< \frac{1}{n\dfrac{N(B_c)}{N}} [1 - p(m'=0) - p(m'=1)] < \frac{1}{nP_{N,c}} [1 - p(m=0)]$$

y, también,

$$A_2 = \sum_{m=1}^{n} \frac{1}{(m+1)(m+2)} \frac{\binom{N(B_c)}{m}\binom{N-N(B_c)}{n-m}}{\binom{N}{n}} =$$

$$= \frac{1}{(n+1)(n+2)\dfrac{N(B_c)+1}{N+1}\dfrac{N(B_c)+2}{N+2}} \cdot [1 - p(m''=0) - p(m''=1) - p(m''=2)]$$

$$< \frac{1}{n^2\dfrac{N(B_c)}{N}\dfrac{N(B_c)}{N}} [1 - p(m''=0) - p(m''=1) - p(m''=2)] < \frac{1}{\left(nP_{N,c}\right)^2} [1 - p(m=0)]$$

De igual modo, se mayoran los restantes sumandos, $A_3 + A_4 < \left(2/(nP_{N,c})\right)^3$. Entonces, reuniéndolos y teniendo en cuenta que $N(B_c) = NP_{N,c}$, se obtiene la desigualdad

$$\sum_{m=1}^{n} \left(\frac{1}{m} - \frac{1}{N(B_c)}\right) p(n(B_c)=m) = A - \frac{1 - p(n(B_c)=0)}{NP_{N,c}}$$
$$< \frac{1 - p(n(B_c)=0)}{nP_{N,c}} - \frac{1 - p(n(B_c)=0)}{NP_{N,c}} + \frac{1 - p(n(B_c)=0)}{\left(nP_{N,c}\right)^2} + \left(\frac{2}{nP_{N,c}}\right)^3$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1 - p(n(B_c)=0)}{P_{N,c}} + \frac{1 - p(n(B_c)=0)}{\left(nP_{N,c}\right)^2} + \left(\frac{2}{nP_{N,c}}\right)^3$$

Demostración del teorema 3

$$\mathrm{D}_p^2 \left[\widehat{\bar{Y}}(B_c)\right] < S^2(B_c) \left\{ \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1 - p(n(B_c)=0)}{P_{N,c}} + \frac{1 - p(n(B_c)=0)}{\left(nP_{N,c}\right)^2} + \left(\frac{2}{nP_{N,c}}\right)^3 \right\}$$
$$+ \bar{Y}^2(B_c) p(n(B_c)=0)[1 - p(n_c=0)]$$

Demostración del lema 4

Se consideran hasta los sumandos de orden cuadrático (A_1, A_2 y B), obviando los de tercer orden

$$A_1 = \frac{1}{(n+1)\frac{N(B_c)+1}{N+1}} \sum_{m'=2}^{n'} \frac{\binom{N(B_c)+1}{m'}\binom{N-N(B_c)}{n'-m'}}{\binom{N+1}{n'}}$$

$$= \frac{1}{n\left(1+\frac{1}{n}\right)\frac{N(B_c)}{N}\frac{1+1/N(B_c)}{1+1/N}}[1-p(m'=0)-p(m'=1)]$$

$$= K\left(1+\frac{1}{N}\right)\left(1-\frac{1}{n}+\frac{1}{n^2}-\cdots\right)\left(1-\frac{1}{N(B_c)}+\frac{1}{N(B_c)^2}-\cdots\right)[1$$
$$-p(m'=0)-p(m'=1)]$$

$$\approxeq K\left(1+\frac{1}{N}\right)\left(1-\frac{1}{n}\right)\left(1-\frac{1}{N(B_c)}\right)[1-p(m'=0)-p(m'=1)]$$

$$\approxeq K\left(1-\left(\frac{1}{n}-\frac{1}{N}\right)-\frac{1}{N(B_c)}\right)[1-p(m'=0)-p(m'=1)]$$

donde $K = \left(nP_{N,c}\right)^{-1}$, por tanto, el primer sumando es aproximadamente igual a

$$A_1 \approxeq K\left(1-\left(\frac{1}{n}-\frac{1}{N}\right)-\frac{1}{N(B_c)}\right)[1-p(m'=0)-p(m'=1)] \approxeq$$

$$\approxeq \frac{1}{nP_{N,c}}\left(1-\frac{1-f}{n}-\frac{1}{NP_{N,c}}\right) = \frac{1}{nP_{N,c}}-\frac{1-f}{n^2P_{N,c}}-\frac{1}{nNP_{N,c}^2}$$

El segundo se puede aproximar del siguiente modo

$$A_2 = \frac{1}{(n+1)(n+2)\frac{N(B_c)+1}{N+1}\frac{N(B_c)+2}{N+2}}[1-p(m''=0)-p(m''=1)-p(m''=2)] =$$

$$= \frac{[1-p(m''=0)-p(m''=1)-p(m''=2)]}{n^2\left(1+\frac{1}{n}\right)\left(1+\frac{2}{n}\right)\left(\frac{N(B_c)}{N}\right)^2\frac{1+1/N(B_c)}{1+1/N}\frac{1+2/N(B_c)}{1+2/N}}$$

$$= K^2\left(1+\frac{1}{N}\right)\left(1+\frac{2}{N}\right)\left(1-\frac{1}{n}+\frac{1}{n^2}-\cdots\right)\left(1-\frac{2}{n}+\left(\frac{2}{n}\right)^2-\cdots\right)\left(1\right.$$
$$\left.-\frac{1}{N(B_c)}+\frac{1}{N(B_c)^2}-\cdots\right)\left(1-\frac{2}{N(B_c)}+\left(\frac{2}{N(B_c)}\right)^2-\cdots\right)[1-p(m''=0)$$
$$-p(m''=1)-p(m''=2)] \approxeq$$

$$\approxeq K^2\left(1+\frac{1}{N}+\frac{2}{N}-\frac{1}{n}-\frac{2}{n}-\frac{1}{N(B_c)}-\frac{2}{N(B_c)}\right)[1-p(m''=0)-p(m''=1)-p(m''=2)]$$

$$= K^2\left(1-3\left(\frac{1}{n}-\frac{1}{N}+\frac{1}{N(B_c)}\right)[1-p(m''=0)-p(m''=1)-p(m''=2)]\right)$$

$$\approxeq \frac{1}{\left(nP_{N,c}\right)^2}\left(1-3\left(\frac{1-f}{n}+\frac{1}{NP_{N,c}}\right)\right)$$

teniendo en cuenta que

$$\frac{1}{\left(nP_{N,c}\right)^2}-\frac{1}{nNP_{N,c}^2}=\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{nP_{N,c}^2}=\frac{1-f}{\left(nP_{N,c}\right)^2}$$

y reuniendo los sumandos

$$A_1+A_2-B \approxeq \frac{1}{nP_{N,c}}-\frac{1}{NP_{N,c}}-\frac{1-f}{n^2P_{N,c}}+\frac{1-f}{\left(nP_{N,c}\right)^2}=\frac{1-f}{nP_{N,c}}-\frac{1-f}{n^2P_{N,c}}+\frac{1-f}{\left(nP_{N,c}\right)^2}=$$

$$=\frac{1-f}{nP_{N,c}}\left(1-\frac{1}{n}+\frac{1}{nP_{N,c}}\right)=\frac{1-f}{nP_{N,c}}\left(1+\frac{1-P_{N,c}}{nP_{N,c}}\right)$$

Demostración del teorema 4

$$\mathrm{E}_p\left[\hat{S}^2(B_c)\right]=\sum_{m=2}^{n}\mathrm{E}_p\left[\hat{S}^2(B_c)|n(B_c)=m\right]p(n(B_c)=m)=$$

$$=S^2(B_c)[1-p(n(B_c)=0)-p(n(B_c)=1)]$$

porque, si $m=2$,

$$\hat{S}^2(B_c)=\frac{1}{2}\frac{2(Y_k-Y_l)^2}{2}=\frac{(Y_k-Y_l)^2}{2}$$

y si $m \geq 3$,

$$\hat{S}^2(B_c)=\frac{1}{2}\frac{\sum\sum_{k\neq l}^{m}(Y_k-Y_l)^2}{m(m-1)}$$

Demostración del corolario 6

$$\left|\,\mathrm{b}_p\left[\hat{S}^2(B_c)\right]\right|=\left|\frac{\mathrm{B}_p\left[\hat{S}^2(B_c)\right]}{S^2(B_c)}\right|=p(n(B_c)=0)+\,p(n(B_c)=1)$$

Demostración del corolario 7

$$\frac{p(n(B_c)=1)}{p(n(B_c)=0)}=n\frac{N(B_c)}{N-N(B_c)-(n-1)}\geq n\frac{N(B_c)}{N-N(B_c)}$$

$$p(n(B_c)=1) \geq p(n(B_c)=0)\frac{nP_{N,c}}{1-P_{N,c}}$$

$$1-p(n(B_c)=0)-p(n(B_c)=1) \leq 1-p(n(B_c)=0)\left(1+\frac{nP_{N,c}}{1-P_{N,c}}\right)=1-R$$

con R verificando

$$0 \leq R < \left(1-P_{N,c}\right)^n\left(1+\frac{nP_{N,c}}{1-P_{N,c}}\right)=\left(1-P_{N,c}\right)^n+nP_{N,c}\left(1-P_{N,c}\right)^{n-1}<$$

$$< e^{-nP_{N,c}}+nP_{N,c}e^{-(n-1)P_{N,c}}=e^{-nP_{N,c}}\left(1+nP_{N,c}e^{P_{N,c}}\right)$$

Demostración del teorema 5

$$\mathrm{E}_p\left[\widehat{\overline{Y}_{st}}(B_c)\right]=\sum_{l=1}^{L}W_l(B_c)\,\mathrm{E}_p\left[\widehat{\overline{Y}}_l(B_c)\right]=\sum_{l=1}^{L}W_l(B_c)\,\overline{Y}_l(B_c)\left[1-p_{U_l}(n_l(B_c)=0)\right]=$$

$$=\sum_{l=1}^{L}W_l(B_c)\,\overline{Y}_l(B_c)-\sum_{l=1}^{L}W_l(B_c)\overline{Y}_l(B_c)\,p_{U_l}(n_l(B_c)=0)=\overline{Y}(B_c)-R$$

con $R=\sum_{l=1}^{L}W_l(B_c)\overline{Y}_l(B_c)\,p_{U_l}(n_l(B_c)=0)$ verificando que $\min_l \overline{Y}_l(B_c)\,p_{U_l}(n_l(B_c)=0) \leq R \leq \max_l \overline{Y}_l(B_c)\,p_{U_l}(n_l(B_c)=0)$.

Demostración del teorema 6

$$\mathrm{D}_p^2\left[\widehat{\overline{Y}_{st}}(B_c)\right]=\sum_{l=1}^{L}W_l^2(B_c)\,\mathrm{D}_p^2\left[\widehat{\overline{Y}}_l(B_c)\right]$$

$$=\sum_{l=1}^{L}W_l^2(B_c)\left[S_l^2(B_c)\overline{\left(\frac{1}{n_l^+(B_c)}-\frac{1}{N_l(B_c)}\right)}+\overline{Y}_l^2(B_c)p_{U_l}(n_l(B_c)=0)\right]\left[1-p_{U_l}(n_l(B_c)=0)\right]$$

Agradecimientos

El primer autor reconoce con gratitud el apoyo económico de los proyectos MTM2013—41383—P (Ministerio de Economía, Industria y Competitividad de España) y MTM2016—76969—P (Agencia Estatal de Investigación de España), ambos cofinanciados por el Fondo Europeo de Desarrollo Regional (FEDER) y la red IAP de la Política de Ciencia de Bélgica. Las representaciones del regresograma y del pictograma han sido realizadas por María Jesús Villaverde Mosquera.

Referencias

ARRIBAS, J.M. Y A. ALMAZÁN (2006) «La estadística española de posguerra (1939-1958)» en *Historia de la probabilidad y la estadística (III)* A.H.E.P.E. Delta, Publicaciones Universitarias, Madrid

COCHRAN, W. G. (1977) *Sampling Techniques*. 3rd ed. John Wiley& Sons, New York

DURÁN HERAS, A. (2007) «La Muestra Continua de Vidas Laborales de la Seguridad Social». *Revista del Ministerio de Trabajo y Asuntos Sociales*, 1, 231-240

FULLER, W. A. (2009) *Sampling Statistics*, John Wiley & Sons, Hoboken, New Jersey

HÄRDLE, W. AND O. LINTON. (1994) «Applied nonparametric methods» In Engle R.F., McFadden D.L. eds. *Handbook of econometrics*, Vol. IV. North Holland, Amsterdam

HEDAYAT, A.S. AND B.K. SINHA (1991) *Design and Inference in Finite Population Sampling.* John Wiley & Sons, New York

IGE (2015) Nomenclátor e Explotación do Padrón Municipal de Habitantes en www.ige.eu (consulta del 21 de septiembre de 2015)

INE (2004) *Buenas prácticas en la elaboración de Estadísticas Oficiales*. Madrid

INE (2011) *Proyecto de los Censos Demográficos 2011*. Subdirección General de Estadísticas de la Población. Madrid

KRISHNAIAH, P.R. AND C.R. RAO (1988) *Sampling Handbook of statistics 6*. North-Holland, Amsterdam

LOHR, S. (2000) *Muestreo: Diseño y Análisis*. Ed. Internacional Thomson, México

MIRÁS AMOR, J. (1985) *Elementos de muestreo en poblaciones finitas.* INE, Madrid

MORAL-ARCE, I. Y E. MARTÍN (2009) «Integración de información administrativa y muestral en estadísticas económicas estructurales. La Encuesta Anual de Estructura Salarial» *Estadística Española,* 51, 172, 487-504

PÉREZ LÓPEZ, C. (1999) *Técnicas de muestreo estadístico: teoría, prácticas y aplicaciones informáticas*. Ed. Rama, Madrid

PEREZ LOPEZ, C. (2011) «Fiscal panels data. Application to income TAX panel data (IRPF) of Spanish Institute for Fiscal Studies. Methodology, estimators and errors». *BEIO*, 27, 3, 204-220

PIATIER, A. (1967) *Estadística y Observación Económica.* Ediciones Ariel, Barcelona

R CORE TEAM (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

ROYALL, R.M. AND J. HERSON (1973) «Robust Estimation in Finite Populations I». *Journal of the American Statistical Association*, 68, 344, 890-893

SÁNCHEZ CRESPO, J. L. (1984) *Curso intensivo de Muestreo en poblaciones finitas*. INE, Madrid

SÄRNDAL, C.-E. et al. (1992) *Model assisted survey sampling*. Springer, New York

SCHOTT, S. (1928) *Estadística*. Editorial Labor, Barcelona

SILVA AYÇAGUER, L.C. (1997) *Cultura estadística e investigación científica en el campo de la salud*. Ed. Díaz de Santos, Madrid

TEIJEIRO, C. Y J. VEGA (2014) ¿Cómo se hizo el Censo 2011? *Índice*, núm. 60, pp. 7-9. INE

XUNTA DE GALICIA (2012) Rexistro de Buques Pesqueiros da Comunidade Autónoma de Galicia. Consellería de Pesca e Asuntos Marítimos, Santiago de Compostela en www.pescadegalicia.com (consulta del 7 de diciembre de 2012)