Como loitar contra o nesgo

Ricardo Cao
Departamento de Matemáticas
Universidade da Coruña
e-mail: rcao@udc.es
http://www.udc.es/dep/mate/ricardo/homepage.htm

Resumo

Neste artigo faise un percorrido ao longo dalgunhas situacións reais que poden levar consigo a presenza de nesgo na mostraxe. Preséntase matematicamente o modelo de nesgo por lonxitude e trátase o problema da estimación da media poboacional nese contexto. Formúlase superficialmente tamén o problema de nesgo mediante unha función de peso xeral e mais as situacións con datos censurados e truncados.

Abstract

Several real situations that exhibit biased sampling are presented in this paper. The length bias model is mathematically introduced and the problem of estimating the mean of a population is presented in this context. Biased sampling with a general weight distribution is briefly introduced. The censored and truncated data setups are mentioned as well.

1. Bibliotecas, usuarios e nesgos

Hai uns anos a bibliotecaria da Facultade de Informática formuloume a seguinte pregunta: poderías axudarme a estimar o número medio de veces que un usuario entra e sae da biblioteca ao longo dun mesmo día? Para decatármonos do motivo da súa pregunta situémola no seu contorno.

Todas as bibliotecas públicas teñen a obriga de realizar un informe anual, que é enviado ao organismo estatal competente, en que se recolle, entre outra información, o número de usuarios anuais da biblioteca en cuestión. Moitas bibliotecas dispoñen dun arco na entrada en que, ademais de controlar que ninguén saque un libro da biblioteca sen autorización, se leva conta do número de persoas que entran e saen da biblioteca. O número de usuarios anuais defínese, neste ámbito, como a suma, ao longo de todos os días do ano, do número diario de usuarios distintos que fixeron uso da biblioteca. Deste xeito, se unha persoa sae da biblioteca para ir xantar, logo para facer un descanso e, finalmente, para tomar café e volve a ela tras cada interrupción, é considerado como un único usuario (non como catro). Ao contrario, se esa mesma persoa estivese eses mesmos catro períodos en catro días diferentes sería considerado como catro usuarios, para os efectos do número anual de usuarios.

A bibliotecaria da Facultade pode coñecer perfectamente o número de entradas-saídas que se produciron na biblioteca ao longo dun ano: a diferenza entre o rexistro do contador do arco ao final e ao comezo do ano. Dado que o número de entradas-saídas anuais é igual ao número de usuarios anuais multiplicado polo número medio de entradas-saídas diarias por persoa, resulta doado decatarse de que todo o que ten que

facer a bibliotecaria para subministrar a información requirida é dividir o contador de entradas-saídas entre o número medio de entradas-saídas por usuario. De aí a súa pregunta.

Despois de falar un anaco do tema, propúxenlle á bibliotecaria facer unha mostraxe con que estimar o número medio que ela necesitaba. Dado que era practicamente imposible coñecer a poboación de usuarios da biblioteca (mesmo a de usuarios dun día concreto) acordamos que a mostraxe se fixese seleccionando aleatoriamente instantes de tempo ao longo do ano. Con tal fin subministreille unha listaxe de 100 rexistros en que cada un indicaba o mes, o día, a hora e o minuto en que se tomaría un dato. Esta listaxe foi xerada aleatoriamente usando o método de aceptación/rexeitamento para evitar instantes fóra do calendario de apertura do local. O procedemento de recollida do dato consistía en que o axudante da biblioteca, encargado do mostrador de préstamo (moi preto do arco da entrada), solicitase á primeira persoa que entrase ou saíse da biblioteca, a partir do instante fixado na listaxe, que lle comunicase, na derradeira vez que saíse da biblioteca ao longo dese día, cal fora o número de entradas-saídas que fixera no día. Deste xeito disporiamos dunha mostra de 100 valores da variable "número de entradas-saídas diarias dun usuario".

A pouco que un pense no asunto decátase de que o procedemento anterior ten un serio inconveniente: non escolle cada un dos usuarios da biblioteca coa mesma probabilidade. De feito, teñen máis probabilidade de seren elixidos os usuarios que máis veces entran e saen na biblioteca que os que o fan menos veces. Este defecto de escoller con probabilidades diferentes os usuarios "cu inquedos" e os usuarios "pousóns" vén do feito de ter escollido aleatoriamente e con distribución uniforme os instantes de tempo (e non os usuarios mesmos).

Dado que outro tipo de mostraxe non semellaba posible non tiñamos máis remedio que analizar a situación con detemento e tratar de corrixir o nesgo introducido na mostraxe.

2. Nesgos, nesgos e máis nesgos

Problemas como o anterior aparecen decotío nas nosas vidas. Por exemplo, cando un ecólogo fai unha mostraxe, mediante fotografía aérea, co fin de estudar o número de individuos por manda de certa especie animal, é evidente que os grupos de animais que teñen máis probabilidade de seren elixidos son precisamente os de maior número. Algo semellante acontece se desexamos estimar o tamaño medio dos nódulos tumorais nun doente cando o rastrexo que realizan os aparatos de medida leva a cabo unha mostraxe espacial uniforme. Nesa situación tamén ocorre que os nódulos máis grandes teñen maior probabilidade de seren elixidos na mostra. Por último, nun contexto totalmente distinto, un problema semellante pode darse ao estimar a estadía media dos turistas nun país. Analicemos un pouco máis este último caso.

Un exemplo ben coñecido de problemas con nesgo na mostraxe foi o acontecido nun estudo da estadía media de turistas en Marrocos (véxase Patil 1984). Existían dúas mostras tomadas de xeito independente e con metodoloxías ben distintas. A primeira foi a enquisa levada a cabo entrevistando directamente, nos seus hoteis, os turistas que se atopaban no país. Sobra dicir que tanto a selección da mostra dos hoteis (de acordo co seu tamaño) como a dos hóspedes destes foron feitas aleatoriamente. O outro estudo foi feito polo Instituto de Estatística marroquí, nas aduanas de saída. Nos dous casos

preguntóuselles aos turistas, entre outros aspectos, a duración da súa estadía en Marrocos.

Un resultado quizais sorprendente foi que a estadía media estimada polo primeiro método foi aproximadamente o dobre que polo segundo. De feito, algo así era previsible porque, a pesar das precaucións de aleatorización que se tomaron, o primeiro método de mostraxe ten a peculiaridade de elixir con moita máis probabilidade os turistas que estiveron máis tempo no país que os que estiveron menos. Isto non acontece, evidentemente, ao facer a mostraxe nas aduanas.

Os problemas de nesgo por lonxitude foron descritos no ano 1963 polo egrexio estatístico hindú C. R. Rao no transcurso do *First International Symposium on Classical and Contagious Discrete Distribution,* que se levou a cabo en Montreal (véxase tamén Patil e Rao 1978). De todos os xeitos, as primeiras ideas sobre o tema xa se atopan nun clásico artigo de Sir Ronald A. Fisher (véxase Fisher 1934). En Cristóbal e Alcalá (2001) pódense consultar diversas referencias dos primeiros artigos en relación co tema.

Un dos problemas prácticos que Rao analizou foi o da estimación do número medio de membros por familia, tomando unha mostra de entre os rapaces e as raparigas nas escolas. Vexamos, a continuación, como formular matematicamente o problema en xeral.

3. O nesgo por lonxitude

Consideremos unha poboación, modelizada matematicamente mediante unha variable aleatoria, X, que suporemos discreta, que pode tomar k posibles valores: $x_1, x_2, \ldots, x_k$, con probabilidades $p_1, p_2, \ldots, p_k$, respectivamente. O nesgo por lonxitude (ou por tamaño) da variable supón que as probabilidades de observación da variable que realmente observamos (chamémoslle Y) se ven afectadas proporcionalmente polo propio valor da variable orixinal. É dicir, realmente observamos a variable aleatoria Y, que toma os mesmos valores que a variable orixinal: $x_1, x_2, \ldots, x_k$, mais con probabilidades $q_1, q_2, \ldots, B_k$, ao satisfacer a relación $o_i = C \cdot x_i \cdot p_i$, para todo i=1, 2, …, k, para certa constante C a determinar. Evidentemente, o valor de C vén dado polo feito de que os q_i tamén teñen que ser unha masa de probabilidade. En particular ten que verificarse que

$$1 = q_1 + q_2 + \cdots + q_k = C\,(x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_k \cdot p_k),$$

do que se deduce que $C = 1 / (x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_k \cdot p_k) = 1 / \mathrm{E}(X)$, é dicir, a constante C é a inversa da media da variable inobservable, X.

Dado que a variable X non pode observarse (soamente a Y), se desexamos estimar algunha característica dela, como a súa media poboacional, podemos tratar de relacionala con características da variable observable nesgada Y. Así, no caso da media demóstrase facilmente que

$$\mathrm{E}(1/Y) = (1/x_1) \cdot q_1 + (1/x_2) \cdot q_2 + \cdots + (1/x_k) \cdot q_k = C\,(p_1 + p_2 + \cdots + p_k) = C = 1 / \mathrm{E}(X).$$

Disto dedúcese que $\mathrm{E}(X) = 1 / \mathrm{E}(1/Y)$, é dicir, a media poboacional da variable de interese é a media harmónica da variable nesgada por lonxitude. A relación anterior dá pé a construír un estimador da media poboacional de X.

Efectivamente, supóñase que observamos unha mostra de tamaño n da variable nesgada por lonxitude: $Y_1, Y_2, \ldots, Y_n$. Isto non é máis que n variables aleatorias, independentes e coa mesma distribución de probabilidade que Y. Se o noso interese é estimar a media da

variable orixinal, $\theta = E(X)$, semella lóxico facelo usando a media harmónica da mostra da variable Y:

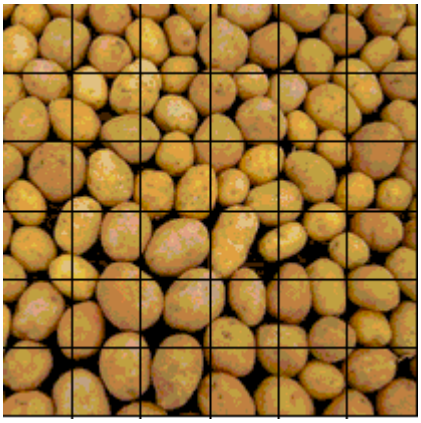
$$\theta_n = n / (1/Y_1 + 1/Y_2 + \cdots + 1/Y_n).$$

Pódese probar que θ_n é un estimador asintoticamente non esguellado de θ, e que, baixo hipóteses axeitadas, a distribución límite de θ_n é normal. A partir deste último feito é doado achar intervalos de confianza para o parámetro θ, neste contexto de nesgo por lonxitude.

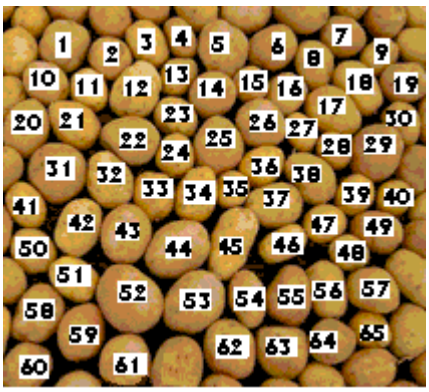
4. Patacas nesgadas e patacas aleatorias

Para ilustrar o comportamento do estimador θ_n na práctica e a diferenza entre unha mostraxe aleatoria simple clásica e unha mostraxe nesgada por lonxitude, imos crear unha situación ideal baixo o noso control. Temos unha poboación de 65 patacas da cal queremos estimar o seu peso medio. Obsérvese que, se se coñece nunha situación práctica, o número total de patacas, sabido o peso medio podemos calcular o peso total do lote. Como non queremos pesar todas as patacas, imos seleccionar unha mostra de 10 patacas, para soamente pesar esas poucas. Para seleccionar as 10 patacas da mostra procedemos de dous xeitos distintos:

a) Poñemos as patacas no chan, o máis xuntas que poidamos, formando unha figura cadrada de dimensións 60 cm. × 60 cm. Logo obtemos ao chou (por exemplo cun xerador de números aleatorios) e con distribución uniforme cada unha das dúas coordinadas dun punto no cadrado e tomamos a pataca que estea precisamente nesas coordinadas ou a máis próxima ao punto en cuestión. Repetimos o proceso un total de 10 veces, pesando de cada vez a pataca elixida.



b) Numeramos as patacas do 1 ao 65 e eliximos ao chou, con distribución uniforme, un número nese rango, que determina o número da pataca elixida. Procedemos 10 veces do mesmo xeito, pesando cada pataca seleccionada.



Obviamente, o procedemento a) proporciona unha mostra nesgada pois as patacas máis grandes son elixidas con maior probabilidade que as máis pequenas. Ao contrario, o método b) consiste nunca mostraxe aleatoria simple, pois cada pataca ten a mesma probabilidade de ser seleccionada na mostra, independentemente do seu peso.

Os pesos (en gramos) das dez patacas elixidas polo procedemento a) foron: 220, 140, 230, 265, 130, 295, 200, 110, 95, 175, mentres que os pesos das patacas elixidas segundo o método b) foron: 85, 165, 115, 165, 275, 95, 90, 105, 90, 195. As medias

mostrais resultaron 186 para o método a) e 138 para o b). Efectivamente, vese que os resultados son ben dispares, resultando a media da mostra nesgada moito maior que a media da mostra aleatoria simple.

Caso de non poder levar a cabo o procedemento b) (que sería o ideal), o nesgo introducido na mostraxe segundo o método a) debe corrixirse cunha estimación axeitada. Isto significa que non debemos utilizar a media mostral de 186 gramos obtida anteriormente, senón o valor dado polo estimador θ_n, media harmónica, introducido na sección anterior. Neste caso a estimación resultante para a mostra segundo a) é 163.34 gramos, que, como se ve, xa é bastante máis próxima á estimación mediante mostraxe aleatoria simple (138 gramos). De feito, neste caso (con soamente 65 patacas), podemos pesar todas as patacas e dividir a suma de todos os pesos entre 65 para coñecer a verdadeira media poboacional, e obteremos o valor de 156.39 gramos. Esta cantidade é próxima tanto á estimación con media harmónica da mostra nesgada como a estimación con media aritmética da mostra aleatoria simple.

Folga dicir que no noso problema inicial, en que desexabamos estimar o número medio de entradas-saídas diarias por usuario da biblioteca, o estimador axeitado é tamén a media harmónica dos datos obtidos, ao se presentar, igualmente, un nesgo por lonxitude. Esa foi a cantidade subministrada á bibliotecaria da Facultade de Informática.

5. Nesgos non lineares, truncamento e censura

Nalgunhas ocasións a función que altera as probabilidades de observación dos valores da variable non é linear nos devanditos valores. Nese caso tense unha función, w, coñecida, de tal xeito que as probabilidades dos valores que toma a variable aleatoria Y, son da forma $q_i = C \cdot w(x_i) \cdot p_i$. Na nosa exposición previa da sección 3, a función w era a identidade, mais existen situacións prácticas en que non ten porque selo.

Analizando máis polo miúdo o exemplo das patacas, decatámonos de que non é totalmente certa a suposición de que co método a) facemos que a probabilidade de observación dunha pataca se altera proporcionalmente ao seu peso. É certo que o seu peso pode suporse proporcional ao seu volume (xa que a densidade das patacas será case idéntica dunha a outra), mais a mostraxe uniforme sobre o chan, deseñada no método a), non provoca unha elección proporcional ao volume senón, posiblemente, á súa sección máxima.

Para simplificar os nosos argumentos, supoñamos por un momento que as patacas fosen esferas perfectas. Nese caso o seu volume é $V = 4/3\ \pi\ r^3$, mentres que a súa sección máxima é a superficie dun círculo máximo da esfera, é dicir, $S = \pi\ r^2$. Como consecuencia, $S = \pi^{1/3}\ (3/4)^{2/3}\ V^{2/3}$, e así teriamos que a sección máxima é proporcional ao volume elevado a 2/3, ou equivalentemente, ao peso elevado a 2/3. Noutros termos, teriamos $w(x)=x^{2/3}$.

Facendo cálculos semellantes aos da sección 3 é doado chegar ao valor da constante C na expresión $q_i = C \cdot w(x_i) \cdot p_i$, no contexto dunha función de nesgo xeral, resultando $E(1/w(Y)) = C = 1 / E(w(X))$. Tamén é moi sinxelo deducir $E(X) = E(Y/w(Y)) / C$, que resulta moi útil para encontrar un estimador axeitado de $\theta = E(X)$ neste contexto:

$$\theta_n = (Y_1/w(Y_1) + Y_2/w(Y_2) + \cdots + Y_n/w(Y_n)) / (1/w(Y_1) + 1/w(Y_2) + \cdots + 1/w(Y_n)).$$

No exemplo das patacas, dado que $w(x)=x^{2/3}$, a expresión redúcese a

$$\theta_n = (Y_1^{1/3} + Y_2^{1/3} + \cdots + Y_n^{1/3}) / (1/Y_1^{2/3} + 1/Y_2^{2/3} + \cdots + 1/Y_n^{2/3}).$$

Para a mostra obtida polo procedemento a) a estimación resulta 170.74 gramos.

Existen outros moitos contextos en que o procedemento de recollida de información mostral inclúe un nesgo de distinto tipo ao nesgo por lonxitude. Dous dos casos máis cotiáns son o truncamento e a censura.

O truncamento aparece en situacións onde a variable de interese só se pode observar cando o seu valor é suficientemente grande (ou pequeno) en termos doutra variable accesoria. Este pode ser o caso de estudos sobre o período de incubación dunha enfermidade. Nestas situacións é moi habitual que a mostra sexa recollida segundo un método de sección cruzada, en que se escollen individuos aleatoriamente, na poboación de referencia, nun instante dado. Se o individuo aínda non amosou síntomas da enfermidade no instante de mostraxe, non formará parte da mostra e, por tanto, o seu período de incubación non será observado. Deste xeito tenderase a non observar períodos de incubación grandes, e será improbable a non observación de períodos pequenos. De ignorar este nesgo na recollida dos datos estaríanse infraestimando características como a media do período de incubación.

O fenómeno de censura tamén aparece decotío en problemas biomédicos, dentro da análise de supervivencia. Tamén é frecuente no estudo de fiabilidade industrial. Neste caso o problema xorde ao non poder observar sempre o valor preciso da variable de interese (tipicamente un tempo), senón ter soamente unha cota para esta. Existen diversos tipos de censura: fixa, aleatoria, por intervalo, pola esquerda, pola dereita, … Vexamos un exemplo de censura aleatoria pola dereita.

Nun estudo médico de sección cruzada deséxanse estimar características da distribución do tempo de supervivencia ao cancro de mama. Nun instante concreto, en que se fai o estudo, recóllese o estado de cada unha das persoas doentes baixo tratamento por esta enfermidade nos últimos anos. Algunhas delas xa terán morto por causa do cancro de mama e, por tanto, coñeceremos xa o valor preciso da variable tempo de supervivencia á enfermidade. Outras poden ter falecido por unha causa distinta e, entón, tan só coñeceremos que o tempo de supervivencia ao cancro sería superior ao tempo de vida observado. Aínda que a causa sería distinta, o tratamento matemático sería semellante se non dispoñemos do verdadeiro tempo de supervivencia no caso de que a persoa doente saíse do programa de seguimento (por se ter mudado de cidade) ou se se atopa felizmente viva no momento do estudo.

Nos artigos de Kaplan e Meier (1958) e Lynden-Bell (1971) propuxéronse, por primeira vez, métodos para estimar a distribución de probabilidade de tempos de vida en presenza de censura e truncamento, respectivamente. No segundo deles a proposta veu motivada pola necesidade de estimar a distribución do cociente entre a potencia de radio a potencia óptica, en astronomía.

Nestes últimos cincuenta anos foron xurdindo moitas outras situacións da vida real en que, como no nesgo por lonxitude, na censura ou no truncamento, a variable de interese non se pode observar completamente. Aínda que o aparato matemático-estatístico necesario se vai complicando a medida que a situación real é máis complexa, a idea subxacente para abordar a estimación de características da distribución é a mesma: tratar de expresar as cantidades de interese da variable (inobservable) orixinal en termos

doutras que dependan da variable observable, que si se poderán entón estimar. Poderíase dicir que esta é unha receita universal para afrontar o problema de como loitar contra o nesgo.

Referencias

Cristóbal, J.A. e Alcalá, J.T. (2001). An overview of nonparametric contributions to the problem of functional estimation from biased data. *Test*, 10, 309-332.

Fisher, R.A. (1934). The effects of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6, 13-25.

Kaplan, E. L. e Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-48.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notes of the Royal Astronomical Society*, 155, 95.

Patil, G. P. (1984). Studies in statistical ecology involving weighted distributions. *Statistics: Applications and New Directions*, 478-503. Indian Statistical Institute.

Patil, G. P. e Rao, C. R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, 34, 179-189.