Comptant poblacions amagades: el mètode de captura-recaptura

Pere Puig
Departament de Matemàtiques
Universitat Autònoma de Barcelona

Resum

No sempre comptar és fàcil, perquè algunes vegades allò que volem comptar és poc accessible, o, fins i tot, pot estar amagat. En aquest article farem una petita introducció als mètodes de captura-recaptura, que són unes eines estadístiques molt útils per a estimar el nombre d'individus de les anomenades *poblacions amagades*.

Abstract

Counting is not always a simple task, as sometimes what we want to count is not accessible or, indeed, can be hidden. This article presents a short introduction to capture-recapture methods, which are very useful statistical tools for estimating the number of individuals in so-called «hidden populations».

L'esser humà, d'ençà que existeix, ha volgut controlar el seu entorn. Una manera bàsica de fer-ho és comptant. Segurament va començar comptant coses quotidianes, com ara dits, aliments, animals o persones, desenvolupant pel camí sistemes de numeració i algorismes. Però comptar no sempre és fàcil, perquè algunes vegades allò que volem comptar és poc accessible, o, fins i tot, pot estar amagat. Com comptar el nombre de peixos d'una certa espècie que hi ha dins d'un llac? Com saber el nombre de consumidors d'un cert tipus de droga en una ciutat? En aquest article farem una petita introducció dels anomenats mètodes de captura-recaptura, que són unes eines estadístiques molt útils per a estimar el nombre d'individus de les anomenades *poblacions amagades*.

Per a il·lustrar la metodologia més bàsica, començarem amb un experiment molt senzill que pot dur-se a terme en una aula de classe.

Quants cigrons hi ha a la tassa? L'experiment

Tenim una tassa plena de cigrons i preguntem a la classe com saber (estimar) el nombre de cigrons sense haver-los de comptar un per un. Es pot explicar, per exemple, que la tassa pot

representar un llac i que cada cigró seria un peix d'una determinada espècie. Algun estudiant podria suggerir pesar un cigró (o potser vint, segons la precisió que tingui la balança) i després pesar la tassa plena, i a partir d'aquí deduir el nombre de cigrons (descomptant el pes de la tassa buida). Tot i que aquest mètode podria ser molt bo per al nostre experiment, cal explicar als estudiants que en el cas del llac i els peixos seria irrealitzable i no tindria sentit.

Comencem per descriure el nostre mètode. Agafem amb la mà un grapat de cigrons, els comptem i els fem una marca amb un retolador vermell . El nombre de cigrons que hem agafat l'anomenem n_1. Vegeu la figura 1. Continuant el símil del llac, n_1 seria el nombre de peixos que hauríem capturat utilitzant una xarxa, parant trampes o amb algun mètode no lesiu, de manera que els animals estiguin vius i no pateixin. La marca, també no lesiva, podria ser algun tipus d'anella o un punt de pintura resistent a l'aigua. D'altra banda, hi ha peixos que individualment es poden identificar per la diferent estructura de taques dels seus cossos (els neros, per exemple). En aquests casos, la manera de marcar-los seria simplement fer-los una fotografia.

Posem els n_1 cigrons marcats dins de la tassa amb els altres i els barregem ben barrejats. És molt important que la barreja estigui ben feta. En el cas dels peixos, el que es faria en aquest punt del procés seria alliberar al llac els n_1 peixos marcats prèviament i esperar un cert temps, de manera que els peixos es barregin els uns amb els altres.

Tornem a agafar un grapat de cigrons de la bossa i els comptem. El nombre de cigrons que hem agafat ara l'anomenem n_2. En principi, no importa si el valor de n_2 és diferent o igual al de n_1. Amb aquesta segona captura trobarem que alguns cigrons ja estan marcats (recapturats), i d'altres, no. El nombre de cigrons que ja estan marcats l'anomenarem m. Vegeu la figura 2. En el cas dels peixos, es faria una segona captura no lesiva, comptant-los (n_2) i veient quants n'hi havia que ja estaven marcats (m).

Figura 1. Agafem de la tassa $n_1 = 83$ cigrons, i els marquem amb un retolador.

Figura 2. Barregem bé el cigrons marcats amb els altres i agafem una nova mostra de $n_2 = 62$ cigrons; en trobem $m = 8$ que ja estaven marcats i 54 que no.

Quants cigrons hi ha a la tassa? Els càlculs

Resumint els resultats de l'experiment, tenim una primera captura de cigrons que marquem amb un retolador de mida n_1 i una segona captura de mida n_2, dels quals trobem que m ja estaven marcats. Anomenem N el nombre total de cigrons que hi ha a la tassa, que és justament la quantitat que volem estimar. Observem que la proporció de cigrons marcats que hi ha a la tassa és n_1/N. D'altra banda, la proporció de cigrons marcats que trobem en la segona captura és m/n_2. Si la barreja dels cigrons marcats amb els altres ha estat ben feta, aquestes dues proporcions haurien de ser semblants, és a dir,

$$\frac{n_1}{N} \sim \frac{m}{n_2}$$

D'aquí, aïllant N, trobem una manera aproximada de calcular (això ho anomenem *estimar*) el nombre de cigrons:

$$N \sim \frac{n_1\, n_2}{m} \qquad (1)$$

Aquesta expressió és coneguda com l'estimador de Lincoln-Petersen. El nom prové de dos dels seus precursors, Frederick C. Lincoln i Carl G. J. Petersen. El primer va néixer a Denver, Colorado, el 1892, i va ser un dels principals desenvolupadors dels mètodes de registre, anellatge i seguiment d'ocells migratoris als EUA. Petersen va néixer a Dinamarca el 1860 i va estudiar l'abundància i les migracions de poblacions de peixos (en particular, palaies). Tot i que aquest estimador ha passat a la historia de l'estadística amb el nom d'aquests dos científics, de fet molt abans, al voltant del 1662, John Graunt va estimar la població de Londres utilitzant mètodes molt semblants. De la mateixa manera, Laplace, a finals del segle xviii, va estimar també el nombre d'habitants de França. En ambdós casos, en lloc de fer una captura física dels individus es van fer servir llistes de registres públics. Aquesta és una manera molt freqüent d'aplicar els mètodes de captura i recaptura en medicina i en ciències socials, com veurem posteriorment.

Condicions de validesa del model

La realitat experimental no sempre permet que l'estimador de Lincoln-Petersen (1) ens proporcioni una bona estimació de N. Per a això es necessari assegurar-nos que certes condicions bàsiques experimentals es compleixen. Aquestes condicions són les següents:

1. La població ha de ser *tancada*. Això vol dir que N es considera constant durant el període d'estudi. No poden haver-hi naixements, morts ni migracions durant el període de captura-recaptura.
2. Tots els individus han de tenir la mateixa probabilitat de ser capturats, tant en la primera com en la segona captura. No hi pot haver diferencies degudes, per exemple, a la mida de l'animal, a l'edat o al sexe.
3. El marcatge no ha d'afectar la probabilitat de captura de l'individu. Per exemple, si la primera captura és traumàtica, és molt possible que sigui més difícil capturar l'animal la segona vegada i aleshores el model no seria vàlid. En el cas de poblacions humanes en què s'utilitzen llistes, aquest condicionament exigeix que les dues llistes siguin independents.
4. Els individus no han de perdre les marques entre la primera captura i la segona. Cada individu ha de ser perfectament identificable i distingible.

De totes aquestes condicions, la que falla més sovint és la segona. Quan passa això, una manera d'arreglar-ho és estratificant. Per exemple, en una espècie animal en què els individus joves són més fàcils de capturar que els vells, hauríem de fer a la pràctica dos experiments, comptant els joves i els vells per separat.

Funciona sempre aquest estimador? Parlem de probabilitats

Notem en (1) que si m és zero, és a dir, si no trobem cigrons o peixos marcats en la segona captura, l'estimador no tindria sentit perquè estaríem dividint per zero. Intuïtivament ja es veu que això ho podem evitar si agafem n_1 i n_2 prou grans; però, com de grans? Per estudiar aquest problema, cal observar que la quantitat m és una *variable aleatòria*. És a dir, si tornem els cigrons novament a la tassa i repetim la segona captura (després de barrejar-los bé), agafant el mateix nombre de peixos o de cigrons n_2, el nombre d'individus marcats m serà segurament diferent, ja que aquest nombre varia a l'atzar, és fruit d'un experiment aleatori. Aleshores ens preguntem: quina és la probabilitat que m sigui zero? És freqüent o no que pugui passar això? Aquesta probabilitat es pot calcular amb l'anomenada regla de Laplace, segons la qual n'hi ha prou amb comptar i fer el quocient entre el nombre d'esdeveniments elementals que componen l'esdeveniment $m = 0$ i el nombre total d'*esdeveniments elementals*, que és el nombre total de maneres en què podem agafar n_2 peixos d'un total de N, és a dir, les combinacions sense repetició de N elements presos de n_2 en n_2, que s'expressen com a C_{N,n_2} i que es calculen amb la coneguda fórmula:

$$C_{N,\,n_2} = \frac{N!}{(N - n_2)!\,n_2!}$$

Recordem que el signe d'admiració indica el factorial d'un número de manera que $k! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (k-1) \cdot k$ sent per definició $0! = 1$. D'altra banda, el nombre d'esdeveniments elementals que componen l'esdeveniment $m = 0$ és justament el nombre de maneres en què podem agafar n_2 peixos d'un total de $N - n_1$ que no estan marcats. Això novament ens porta al càlcul del nombre de combinacions sense repetició, en aquest cas de

$$C_{N-n_1,\,n_2} = \frac{(N-n_1)!}{(N-n_1-n_2)!\,n_2!}$$

Naturalment, aquesta expressió només té sentit sempre que $N \geq n_1 + n_2$ que és la situació habitual. Notem que si $N < n_1 + n_2$ mai podríem observar a la segona captura $m = 0$. Seguint ara la regla de Laplace, la probabilitat que m sigui zero, que denotarem com a $P(m = 0)$, vindrà donada per l'expressió:

$$P(m=0) = \frac{C_{N-n_1,\,n_2}}{C_{N,\,n_2}} = \frac{(N-n_1)!(N-n_2)!}{(N-n_1-n_2)!\,N!} \tag{2}$$

Per exemple, si el nombre de peixos en ell llac fos de $N = 500$ i en la primera captura n'agaféssim $n_1 = 20$ i en la segona $n_2 = 25$, aleshores tindríem:

$$P(m=0) = \frac{480!\,475!}{455!\,500!}$$

Atès que aquests factorials donen valors molt grossos, és millor simplificar primer aquesta darrera expressió, obtenint:

$$P(m=0) = \frac{480!\,475!}{455!\,500!} = \frac{475 \cdot 474 \cdots 457 \cdot 456}{500 \cdot 499 \cdots 482 \cdot 481} = \frac{475}{500} \cdot \frac{474}{499} \cdots \frac{457}{482} \cdot \frac{456}{481} = 0{,}3512$$

En el pla numèric, per tal d'evitar *overflows* en fer operacions amb números molt grossos, és molt millor avaluar les 20 fraccions que hi ha entre 475/500 i 456/481 i després multiplicar els resultats. Això ho podem fer fàcilment amb una calculadora o amb un full d'Excel. Procedint així, finalment la probabilitat calculada ens està indicant que un 35,12% de les vegades que féssim una segona captura d'aquesta mida obtindríem un valor $m = 0$ i, per tant, l'estimador de Lincoln-Petersen no seria útil. Aquesta probabilitat la trobem massa alta! Això simplement ens està informant del fet que uns valors de $n_1 = 20$ i de $n_2 = 25$ no serien gaire apropiats per a mesurar una població de l'ordre de $N = 500$. A continuació plantegem com un exercici d'ampliació el càlcul de la probabilitat de trobar en la segona captura exactament k peixos marcats, on k és evident que pot variar entre 0 i n_1:

Exercici. Demostreu que

$$P(m=k) = C_{n_1,\,k}\,\frac{C_{N-n_1,\,n_2-k}}{C_{N,\,n_2}} \tag{3}$$

Aquesta distribució de probabilitats de m, expressada a l'exercici, és coneguda en la literatura com a *distribució hipergeomètrica*. Podeu trobar molta informació sobre aquesta distribució en llibres d'introducció a la teoria de probabilitats i també en enllaços d'internet. El que té de bo de

ser una distribució coneguda és que molts programes i aplicacions informàtiques la tenen implementada. Per exemple, per a calcular $P(m = k)$ amb Excel només cal escriure =DISTR.HIPERGEOM(k; n_2; n_1;N) i automàticament obtindrem el resultat indicat en (3). D'aquesta manera, podem fer moltes exploracions i càlculs interessants. Per exemple, explorant la situació experimental en què $n_1 = n_2 = n$, podem fer un gràfic posant en l'eix de les abscisses els valors de la grandària poblacional N i en l'eix de les ordenades el valor de n que fa que $P(m = 0)$ sigui petita, posem pel cas de 0,1 (10%). Per exemple, per a $N = 500$ i $n = 20$ trobem que =DISTR.HIPERGEOM(0; 20; 20; 500) ens dona 0,4349; per a $n = 35$ tenim que =DISTR.HIPERGEOM(0; 35; 35; 500) ens dona 0,0718, i així per tempteig trobem que el valor de n que proporciona la probabilitat més pròxima a 0,1 és $n = 33$ (sempre aproximem per defecte). La figura 3 ens mostra aquests valors de n en funció de la grandària poblacional N. Observem que per a una grandària $N = 1.000$ recomanaríem utilitzar una mida de la captura de $n = 47$, que representa un 4,7% de la població. En canvi, per a $N = 100$ resultaria $n = 15$, que representa un 15% de la població. Això ens indica una cosa que és molt lògica: per a poblacions petites el mètode de captura-recaptura no és gaire eficient.

La correcció de Chapman

L'estimador de Lincoln-Petersen (1), tot i que és molt senzill i intuïtiu, no funciona gaire bé quan m és petit. És clar que en el cas extrem en què $m = 0$ el valor de l'estimador seria molt i molt gran: infinit! Però aquest estimador també tendeix a donar valors més grans del compte quan m és petit, és a dir, tendeix a sobreestimar el valor de la grandària poblacional per als valors petits de m, produint el que anomenem un *biaix*. Per aquesta raó, Douglas G. Chapman va introduir el 1951 una correcció de l'estimador de Lincoln-Petersen (1), que té millors propietats perquè redueix aquest biaix o sobreestimació de la grandària poblacional per a valors de m petits. L'estimador o correcció de Chapman és el següent:

$$N_C \sim \frac{(n_1 + 1)(n_2 + 1)}{m + 1} - 1 \qquad (4)$$

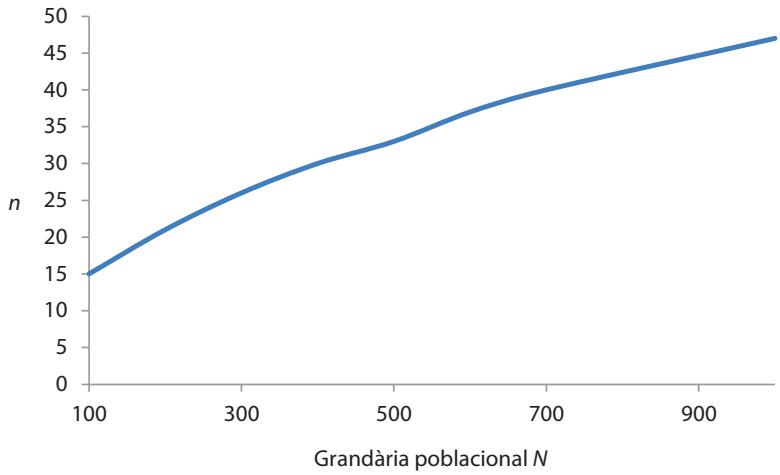

Figura 3. Mida recomanada de la captura *n* en funció de la grandària poblacional *N*.

Tot i que la correcció de Chapman elimina el problema que teníem quan $m = 0$, encara convé tractar d'evitar els valors petits de m agafant valors prou alts de n_1 i n_2.

Aplicacions diverses

A continuació descriurem alguns exemples d'aplicació, fora de l'àmbit de l'ecologia (abundància d'animals), per posar de manifest la utilitat d'aquesta metodologia en diverses àrees.

Càlcul del nombre de consumidors d'heroïna: Bangkok 1994

Un clàssic en l'àmbit de les aplicacions dels mètodes de captura-recaptura en la salut pública és l'article de Mastro *et al.* (1994). En aquest article s'estima d'una manera certament enginyosa el nombre de consumidors d'opiacis (essencialment heroïna) per via injectada. Es considera com a primera captura el conjunt de tots els individus que es troben sota tractament de metadona per tal de superar la seva addicció, en 18 centres d'atenció a la ciutat. Aquest valor és de $n_1 = 4.064$. Com a segona captura, es consideren els registres policials d'individus sota detenció que donen positiu d'opiacis en una anàlisi d'orina. Això es fa en 72 comissaries i s'obté $n_2 = 1.540$. El nombre d'individus coincidents en les dues llistes va ser de $m = 171$. Partint d'aquestes dades, l'estimador de Lincoln-Petersen (1) dona un nombre de consumidors de $N \sim \frac{4064 \cdot 1540}{171} = 36600$.

En aquest cas, la correcció de Chapman (3) ens dona un valor molt semblant:

$$N_C \sim \frac{4065 \cdot 1541}{172} - 1 = 36419$$

Malgrat que aquest exemple ha tingut una gran influència i ha servit com a punt de partida per a altres estudis que s'han fet en diversos indrets del món, cal assenyalar que algunes de les condicions de validesa del model són qüestionables. Si bé les condicions 1 i 4 les podem admetre, això no passa amb les condicions 2 i 3. És creïble que tots els consumidors d'heroïna tinguin la mateixa probabilitat d'aparèixer en una d'aquestes dues llistes? Creiem que no! Segurament existeix un perfil de consumidor que mai se sotmetrà a un tractament de metadona. Tampoc no és clar que tots els consumidors hagin de delinquir forçosament i hagin d'aparèixer tard o d'hora en un registre policial. Això no obstant, la xifra obtinguda sembla bastant raonable i proporcionalment similar a altres recollides en diferents ciutats en aquella època.

Càlcul del nombre de fumadors: Universitat de Barcelona, curs 1997-1998

Freixa *et al.* (2000) van portar a terme un interessant treball sobre el nombre de fumadors de deu o més cigarrets diaris entre els estudiants de les facultats d'Econòmiques i de Psicologia i l'Escola Universitària de Professorat de la Universitat de Barcelona. Les «captures» van consistir a passar una enquesta als estudiants que utilitzaven les màquines expenedores de tabac que hi havia en els serveis de bar-restaurant de la Facultat d'Econòmiques i del campus de la Vall d'Hebron. En la primera captura, obtinguda després d'una observació de vuit hores al llarg de dos dies, es va registrar un total de $n_1 = 111$ estudiants. En la segona captura,

també obtinguda després d'una observació de vuit hores al llarg de dos dies, diferents dels de la primera captura, però en la mateixa franja horària, el resultat va ser de $n_2 = 97$ estudiants. El nombre d'estudiants que apareix en les dues captures és de $m = 5$. En aquest cas, l'estimador de Lincoln-Petersen (1) i la correcció de Chapman (3) ens donen uns valors de $N \sim 2153$ i $N_C \sim 1828$. Els autors utilitzen aquesta estimació del nombre de fumadors, juntament amb la informació del nombre d'estudiants matriculats, per a estimar la prevalença dels fumadors.

Conductors «capturats» en controls d'alcoholèmia: UK 2011-2015

El diari del Regne Unit *The Guardian* publicava el 30 de desembre de 2016 una relació del nombre de conductors que havien donat positiu en controls d'alcoholèmia entre 2011 i 2015. En concret, la notícia es referia als conductors que havien rebut la sanció coneguda com a *DR10 endorsement*. La notícia especificava que 219.008 conductors havien donat positiu un cop, 8.068 havien donat positiu dos cops, 449 tres cops, 46 quatre, 5 cinc i 2 sis cops. Notem que aquest és un patró de captura-recaptura molt més complex que el que hem descrit anteriorment, però el podem simplificar si considerem la primera captura com aquells individus que durant aquest període han donat positiu només en un control d'alcoholèmia, és a dir, $n_1 = 219.008$. La grandària de la segona captura no la coneixem, perquè no tenim informació de quants conductors no han donat positiu en els controls d'alcoholèmia. Però el nombre de conductors recapturats sí que el tenim, i correspondria a aquells que han donat positiu més d'un cop, és a dir, $m = 8068 + 449 + 46 + 5 + 2 = 8.570$. Aleshores, pensant que l'esforç en la segona captura al llarg dels cinc anys d'estudi seria el mateix que en la primera, suposaríem que $n_2 = n_1 = 219.008$. Això, ho repetim, seria una simplificació del problema, que de fet és molt més complex. Partint d'aquí, l'estimació del nombre total de conductors que han conduït sota els efectes d'una ingestió no autoritzada d'alcohol seria de $N \sim 5.586.792$ i $N_C \sim 5.596.189$. Cal dir que aquests valors són molt semblants als obtinguts utilitzant mètodes més sofisticats, tenint en compte el patró de captura-recaptura real d'aquest problema. Com a punt de reflexió comentarem que, segons *The Guardian*, el nombre total de conductors que han donat positiu en els controls ha estat de 227.578. Tenint en compte l'estimador de Chapman, aquesta xifra només representaria el 4,1% del total de conductors que donarien positiu en el cas que els aturessin per a fer un control. Aquest tant per cent és molt baix, però sincerament no ens sorprèn!

Les prostitutes de Costa d'Ivori i de Kisumu, Kenya

Els mètodes de captura-recaptura també s'han utilitzat per a mesurar grups de població socialment marginats, com és el cas de les prostitutes. Vuylsteke *et al.* (2010) estimen la població de prostitutes en tres ciutats de Costa d'Ivori i a Kisumu (Kenya). A les quatre ciutats la primera captura es va fer un dissabte. Els equips d'investigadors assignats per a aquesta tasca van visitar els punts clau de cada ciutat en hores punta de l'activitat. Van identificar les prostitutes i van obtenir el seu consentiment per a participar en l'estudi contestant a un petit qüestionari. A cadascuna se li va donar una targeta, que va ser el mecanisme de «marcatge». Per exemple, amb aquest procediment experimental, a la ciutat de San Pedro es va obtenir $n_1 = 949$ i a Kisumu el valor va ser de $n_1 = 651$. La segona captura es va fer sis dies més tard, un divendres, en els mateixos llocs i hores que en la primera, utilitzant els mateixos equips d'investigadors i els mateixos procediments operatius. Aquesta vegada es van obtenir uns

valors de $n_2 = 862$ a San Pedro i de $n_2 = 680$ a Kisumu. A les prostitutes de la segona captura se'ls va preguntar si ja havien rebut anteriorment una targeta de reconeixement de la seva col·laboració. Aquelles que van contestar que sí van ser les «recapturades», i es van obtenir unes xifres de $m = 427$ a San Pedro i de $m = 328$ a Kisumu. Els valors de les grandàries de les poblacions en les dues ciutats estimades usant l'estimador de Lincoln-Petersen (1) van ser de $N \sim 1.916$ i $N \sim 1.350$, respectivament. Els autors de l'article fan, de fet, unes estimacions una mica més elaborades tenint en compte també el percentatge de prostitutes que van refusar de participar en l'estudi.

Com a cloenda: captura-recaptura, un ventall de possibilitats

Els mètodes de captura-recaptura constitueixen una línia de recerca actual en estadística matemàtica, disciplina en què treballen molts investigadors. El llibre recent de Böhning *et al.* (2017) és un compendi de les aportacions més recents a la medicina i les ciències socials dins d'aquest àmbit. Si fem als nostres estudiants una petita introducció als mètodes de captura-recaptura, en termes similars al que hem tractat en aquest article, podem produir un *feedback* juntament amb uns punts de reflexió comuns que poden ser molt interessants. Preguntem als estudiants, com un repte, on aplicarien aquesta metodologia. Quines poblacions amagades podrien ser del seu interès? És possible que tinguem sorpreses. Al llarg dels anys he sentit moltes propostes diferents, algunes assequibles, d'altres no, però moltes han estat singularment creatives. Des de propostes d'experiments per a mesurar quanta gent copia en un examen, o entra al metro sense pagar, fins a experiments per a saber quants cotxes Volkswagen escarabat de color groc hi ha a Barcelona. En qualsevol cas, pot ser un mitjà útil per a fomentar la seva imaginació i la seva creativitat.

Referències

Bohning, D., Van der Heijden, P.G.M., Bunge, J. (2017). Capture-Recapture Methods for the Social and Medical Sciences. Series: Chapman & Hall/CRC Interdisciplinary Statistics, Chapman and Hall/CRC [ISBN 9781498745314]

Freixa, M., Guàrdia, J., Luisa Honrubia, M., Peró, M. (2000). Estimación de la prevalencia a partir de los métodos de captura-recaptura. *Psicothema*, 12, Supl. 2, 231-235.

Mastro, T. D., Kitayaporn, D., Weniger, B.G. *et al.* (1994). Estimating the number of HIV-infected injection drug users in Bangkok: a capture-recapture method. *American Journal of Public Health*, 84(7), 1094-1099.

Vuylsteke, B., Vandenhoudt, H., Langat, L., Semde, G., Menten, J., Odongo, F., Anapapa, A., Sika, L., Buve, A., Laga, M. (2010). Capture-recapture for estimating the size of the female sex worker population in three cities in Côte d'Ivoire and in Kisumu, western Kenya. *Tropical Medicine & International Health*, 15, 1537-1543[DOI 10.1111/j.1365-3156.2010.02654.X]