Mostreig

Àngel J. Gil Estallo

P08/05057/02303

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índex

Sessió 1
Mostreig 5
1. Introducció 5
2. Mostreig: població i mostra 5
3. Mostreig aleatori simple 7
   3.1. Tria d'una mostra aleatòria: ús de taules de dígits aleatoris 8
4. Mostreig sistemàtic 9
5. Mostreig estratificat 10
6. Mostreig per conglomerats 13
7. Mostreig polietàpic 14
8. Mostreig per quotes 16
9. Resum 17
Exercicis 19
Annex 22

Mostreig

1. Introducció

En aquesta sessió introduirem els aspectes més rellevants que cal tenir en compte a l'hora de tractar d'obtenir una mostra a partir d'una població. L'estudi de les tècniques de mostreig és molt complicat i molt important, ja que la majoria dels resultats teòrics es basen en la suposició que disposem d'una "bona" mostra, representativa de les característiques globals de la població. Si la mostra no és representativa, les conclusions que se'n puguin extreure seran poc correctes o simplement ens induiran a error.

Començarem per recordar la distinció entre **població** i **mostra**. Després tractarem diferents tipus de tècniques de mostreig, començant per la més important de totes: el **mostreig aleatori** simple. Aquesta tècnica assegura que tots els individus de la població tenen la mateixa probabilitat de ser escollits i que els individus se seleccionen de manera independent els uns dels altres. També tractarem detalladament un mètode molt important per a obtenir mostres aleatòries simples: el basat en **taules de dígits aleatoris**.

Una única mostra

De totes les possibles mostres de la població treballarem normalment amb una **única mostra**. A partir d'aquesta, hem de deduir tanta informació sobre la població global com sigui possible.

A continuació aprendrem a obtenir mostres **per mostreig sistemàtic**, mètode més simple que el mostreig aleatori simple.

Un cop fixats aquests dos mètodes, introduirem alguns refinaments que permeten d'incloure dins el procés de selecció de la mostra algunes de les característiques conegudes de la població; en concret estudiarem les mostres **estratificades** i les mostres per **conglomerats**:

1) En el primer cas (estratificació), es divideix la població en grups de manera que els elements de cada grup mostren un comportament similar, mentre que individus de diferents grups mostren comportaments diferents.

2) En el segon cas, tots els conglomerats (que poden ser una agrupació física o geogràfica) són similars els uns amb els altres, mentre que dins de cada conglomerat els individus mostren tanta heterogeneïtat com en la població total.

Per acabar, comentarem un tipus de mostreig, anomenat **mostreig per quotes**, en què es presenta una forma molt pragmàtica d'obtenir informació.

2. Mostreig: població i mostra

El cens

En cas que disposem d'un llistat de tots els individus de la població, direm que tenim un **cens** de la població.

En l'estudi de molts fets reals és convenient considerar i definir amb precisió el conjunt d'individus (siguin persones, màquines, motos o el que sigui) rellevants en la nostra investigació.

El conjunt dels individus objecte del nostre interès és el que s'anomena la **població**.

Normalment, accedir a cadascun dels individus de la població és impossible bé perquè la població és massa gran i resulta inviable econòmicament o bé perquè el temps necessari per a recollir totes les dades fa que l'estudi sigui inútil.

Motivacions per a l'ús de mostres

a) Si estudiem la duració d'un cert tipus de disc dur, no té sentit esperar que s'espatllin tots per a estudiar la mitjana de la duració dels discos.

b) D'altra banda, en el cas de les investigacions sanitàries, per exemple, no podem pretendre subministrar un nou fàrmac a tots els individus de Catalunya per saber si té les propietats requerides.

c) Una situació similar es dóna en l'anomenat *control de qualitat*: si es vol controlar la qualitat de la producció d'un cert producte, per exemple, dels iogurts produïts en una fàbrica, no podem obrir tots i cadascun dels iogurts (ja que destruiríem el producte i n'impossibilitaríem la venda). Fins i tot en el cas que vulguem controlar només el pes dels iogurts no els podem pesar tots individualment, ja que resultaria molt car i lent.

En general, s'acostuma a seleccionar una mostra i estudiar sobre aquesta la característica que ens interessa (la duració dels discos durs, l'efecte d'un tractament, el pes dels iogurts...).

Una **mostra** és qualsevol subconjunt de la població objecte del nostre estudi.

***Mostra* en anglès és *sample*, i *mostreig* és *sampling*.**

De mostres n'hi ha moltes. És fàcil veure que si la població total està formada per N individus i volem mostres de k individus, en podem formar $\binom{N}{k}$ de diferents.

En general, les mostres s'utilitzen per a obtenir informació numèrica sobre certes quantitats relacionades amb la població (i, per tant, desconegudes *a priori*). Establir procediments que garanteixin que la mostra sigui el més representativa possible de la població és, doncs, crucial.

Exemple d'utilització d'una mostra

Podem estar interessats en la mitjana de l'alçada de tots els catalans. Resulta evident que és impossible mesurar l'alçada de tots ells: haurem de recollir una mostra i intentar de deduir el valor de la mitjana poblacional a partir de la mitjana obtinguda amb els individus de la mostra.

En cas que la mostra sigui poc representativa de la població global direm que la mostra és **esbiaixada** (o que té **biaix**).

Aquest biaix acostuma a donar-se quan alguns sectors de la població estan més representats dins la mostra que d'altres.

Exemples d'errors que cal evitar

És molt fàcil veure que certes situacions acadèmiques produiran clarament mostres esbiaixades; però en les situacions de la vida real el biaix pot no ser tan evident. A continuació presentem alguns errors que cal evitar:

1) Imaginem que volem obtenir informació sobre el temps que els estudiants de la UOC dediquen a veure la televisió i que volem obtenir dades de 250 individus. Una opció seria preguntar als 250 primers alumnes que arribessin a una de les trobades presencials. Aquesta

mostra estaria molt probablement esbiaixada, ja que: 1) només respondran alguns dels alumnes que efectivament assisteixen a la trobada i, per tant, el col·lectiu dels alumnes que no hi assisteixen no estarà representat; 2) els alumnes que assisteixen a la trobada, però arriben tard, tampoc no estaran representats (és possible que arribin tard precisament perquè miren la TV a la nit, amb la qual cosa es podria quedar fora de la mostra un col·lectiu potencialment molt interessant en el nostre estudi). Els especialistes en mostreig han de tenir en compte aquestes qüestions.

2) En el cas anterior, ens podríem limitar a enviar un missatge a tots els estudiants de la UOC i estudiar totes les respostes que rebem. Aquest seria un cas de l'anomenada *resposta voluntària*. Les mostres obtingudes d'aquesta manera acostumen a tenir biaix a causa de les característiques dels individus que contesten, ja que normalment ho fan els individus que estan més posicionats (a favor o en contra) sobre el tema que es pregunta. En el cas del temps que es mira la TV, és possible que, en el conjunt d'individus que contesten estiguin sobrerepresentats els que miren molt la TV i subrepresentats els que la miren poc, perquè no els agrada (és possible que no responguin perquè no donen importància al tema, o perquè no volen perdre el temps en un tema que no els interessa). Aquests casos de resposta voluntària es donen molt en les enquestes telefòniques (o virtuals) de la televisió o la ràdio; com que s'ignoren els individus que no escolten el programa i els que responen acostumen a ser gent molt interessada en el tema, les conclusions que es poden extreure no són massa fiables, en general.

3) Un de clàssic i ben real. L'any 1936 als Estats Units es va obtenir una mostra de milions de votants amb la qual es va pronosticar la derrota de Roosvelt; una altra mostra, molt més modesta, de només milers d'electors, va servir per a pronosticar la victòria de Roosvelt. Finalment, va guanyar Roosvelt per majoria aclaparadora i amb uns resultats similars als predits per la segona mostra! El problema residia en el fet que la primera mostra es va obtenir telefònicament, en una època en què disposar de telèfon era sinònim d'un estatus social que afavoria un tendència especifica de vot. Avui en dia, en canvi, les enquestes telefòniques poden arribar a ser molt precises...

Resulta, doncs, que no totes les mostres que es poden extreure d'una població són útils des del punt de vista estadístic. Ara començarem l'estudi, molt descriptiu i resumit, de diferents tècniques de mostreig que permeten d'evitar (o almenys de reduir) l'impacte sobre el resultat final dels errors que acabem de mencionar.

3. Mostreig aleatori simple

Per la seva importància, començarem per introduir el concepte de mostreig aleatori simple.

Es diu que s'ha obtingut una **mostra aleatòria simple** quan el procés per a obtenir-la garanteix aquestes dues propietats:

1) Tots els elements de la població tenen la mateixa probabilitat de formar part de la mostra.

2) Els elements se seleccionen d'un a un i amb reposició, de manera que les seleccions es fan sempre sobre el total de la població.

Els efectes de la reposició

El fet que hi hagi reposició, és a dir, que un individu escollit pugui tornar a ser escollit té molta importància des del punt de vista teòric, encara que habitualment es considera que si la mostra és més petita que el 10% de la mesura de la població, tant és que actuem amb reposició o sense.

La primera condició assegura que no hi ha individus "privilegiats", que tinguin més tendència a estar representats que d'altres i, per tant, millora la representativitat en la mostra. La segona condició garanteix la independència de les seleccions, ja que el fet d'haver escollit un individu no modifica les possibilitats que els altres individus de la població siguin escollits.

Com podem obtenir una mostra aleatòria simple a partir d'una població donada? El primer que hem de fer és aconseguir una llista de tots els individus de la població. Habitualment, s'assigna un nombre a cada individu per a facilitar la feina; a continuació podem, per exemple:

1) Escriure cada nombre (1, 2, ..., *N*) en una papereta, introduir les paperetes en una urna, barrejar-les perfectament i anar traient *k* paperetes, d'una a una i vigilant de reintegrar la papereta a la urna després de cada extracció.

2) També es pot fer un sorteig amb boles numerades d'1 a *N*, a condició que les boles siguin reintegrades al bombo un cop n'han sortit. En aquest cas, el bombo ajuda (a força de donar-hi voltes) al fet que les boles estiguin repartides a l'atzar. Això és cert sempre que totes les boles siguin idèntiques (excepte en el nombre), pesin igual i també que el bombo sigui perfecte.

En tots dos casos resulta evident que se satisfan les dues condicions de mostra aleatòria simple.

En cas que no disposem de paperetes ni de boles ni bombos, podem optar per utilitzar les anomenades *taules de dígits aleatoris*, que és el procediment més habitual (si no utilitzen un ordinador, és clar) per a obtenir una mostra aleatòria simple dins una població finita.

3.1. Tria d'una mostra aleatòria: ús de taules de dígits aleatoris

Imaginem ara que disposem d'un cens de la nostra població (és a dir, d'una llista dels *N* individus de la població) i que volem extreure una mostra aleatòria d'aquesta població de *k* individus. Ens cal cercar un procediment que ens ajudi a extreure aquesta mostra de la manera més senzilla possible. La manera habitual consisteix a utilitzar una taula de dígits aleatoris.

Obtenció automàtica de mostres

La majoria de programes estadístics incorporen la possibilitat d'extreure mostres de la mida desitjada a partir del conjunt de les observacions d'una variable.

A continuació, mostrarem un exemple de com s'utilitzen aquest tipus de taules per a obtenir mostres aleatòries simples.

Imaginem que disposem de la llista dels 1.400 alumnes matriculats en una universitat i que volem extreure'n una mostra de dotze estudiants. El primer que cal fer és identificar cada estudiant amb un nombre.

A continuació mostrarem com obtenir dotze nombres de quatre dígits que determinaran els dotze individus seleccionats en la mostra. Ho farem de la manera següent:

Vegeu la taula de dígits aleatoris en l'annex d'aquesta sessió.

- Escollirem un punt per on començar a llegir la taula. Ho podem fer llançant un dau i començant la taula pel dígit corresponent al resultat del dau. Imaginem que traiem un tres; això vol dir que el dígit que està en la posició

tres (indicat en negreta) és el primer que considerem per calcular els altres individus de la mostra:

19223 95034 05756 28713 96409 12531 42544 82853

- Com que necessitem nombres de quatre xifres per a identificar els nostres alumnes, anirem formant grups de quatre xifres a partir del dígit que hem obtingut com a punt de partida; el primer grup de quatre xifres que obtenim a partir del punt de partida és el 2239. Com que aquest no es correspon amb cap individu de la població (només tenim 1.400 estudiants) ens el saltem i mirem la pròxima agrupació de quatre dígits, que és el 5034, que també saltem; en canvi, l'agrupació de quatre dígits següent és 0575, que sí es correspon amb un cert estudiant. Així, doncs, l'estudiant de nombre 575 serà el primer individu de la mostra. Continuarem el procés saltant el 6287 i incloent en la mostra com a segon estudiant el que té el nombre 1.396.

- Després, continuarem de la mateixa manera, fins a obtenir els dotze individus que necessitem en la mostra.

L'elaboració de taules

Com a curiositat, podem mencionar que la taula de dígits aleatoris que apareix en el llibre de D. Peña i J. Romo. *Introducción a la Estadística para las Ciencias Sociales* (Ed. McGraw-Hill) ha estat construïda:

"poniendo los números premiados en los sorteos de la lotería uno detrás de otro" (pàg. 268).

Observeu que la clau del procés consisteix en l'elaboració de les taules que, en definitiva, són les que han de garantir l'aleatorietat de tot el procés. Aquestes taules s'acostumen a generar per ordinador i han de superar diverses proves d'aleatorietat i d'independència entre els dígits que apareixen en les taules.

4. Mostreig sistemàtic

El mostreig sistemàtic és un procediment més simple d'obtenir una mostra. Suposem que volem seleccionar una mostra de mida k d'una població de n individus. Aquest procediment es basa en els punts següents:

1) Es numeren, com en el cas anterior, els individus de la població, d'1 a N.

Recordeu que...

... la part entera de x, denotada per $[x]$, s'obté truncant el nombre x a 0 decimals.

2) Es calcula $m = [N / k]$, on $[x]$ designa la part entera del nombre x.

3) Se selecciona a l'atzar un nombre entre 1 i m, que indicarà el primer individu que formarà part de la mostra.

4) Anem sumant m tantes vegades com calgui al nombre que indica el primer individu de la mostra i incloem en la mostra els individus que es corresponen amb els resultats d'aquestes sumes.

5) Tant el nombre que determina el primer individu que forma part de la mostra, com la quantitat m (que determina els intervals fixos que serveixen per a seleccionar els altres individus de la mostra), garanteixen que s'obtindrà

el nombre d'individus necessari en la mostra, a la vegada que es recorre tota la llista.

Exemples de mostreig sistemàtic

1) Mostra aleatòria de dotze individus amb mostreig sistemàtic

Suposem que volem obtenir una llista de dotze estudiants entre els 1.400 d'una certa universitat usant mostreig sistemàtic. Seguirem el procediment següent:

– Numerem els estudiants de l'1 al 1.400.

– Seleccionem a l'atzar un nombre menor o igual que $m = [1.400 / 12] = 116$; aquest nombre correspon al primer individu que seleccionarem per a la mostra. Suposem que utilitzant algun sistema que garanteixi l'aleatorietat, obtenim que el primer individu que apareix en la mostra és el que ocupa el lloc 10.

– A continuació, sumem tantes vegades com calgui $m = 116$ a partir del primer individu. Així la mostra estarà formada pels estudiants que ocupen les posicions

10, 10 + 116 = 126, 10 + 2 * 116 = 242, 10 + 3 * 116 = 358…

I així fins a obtenir els dotze individus de la mostra.

2) Mostra aleatòria de 400 individus usant mostreig sistemàtic sobre la guia de telèfons.

"Suposem que hi ha 834.781 abonats en les guies de telèfons de Barcelona i que volem una mostra aleatòria de 400 abonats [...] Com que 834.781 dividit per 400 és aproximadament 2.086, podem agafar cada 2.086è nombre de la guia, i això ens donarà una mostra de 400 abonats estesos al llarg de totes les entrades de la guia. Per a començar la selecció triem un nombre a l'altzar entre 1 i 2.086 [...]; suposem que aquest nombre és el 731. Busquem el nombre 731è en la guia i després l'entrada nombre 731 + 2.086 = 2.817, després 2.817 + 2.086 = = 4.903, i així successivament. [...] S'hi poden afegir algunes dreceres de sentit comú per a fer una tasca una mica més senzilla.

Comptar 2.086 entrades en la guia cada vegada és pesat, i un petit canvi en el disseny del mostreig anterior no hi resta validesa, sempre que el canvi s'estableixi a l'inici, abans que el mostreig comenci. Per exemple, suposem que, en comptar quantes entrades hi ha en unes quantes pàgines de la guia, trobem que la mitjana és de 205 entrades per pàgina, és a dir, 2.086 entrades fan unes 10 pàgines, amb un restant de 36. Després, des del punt inicial del mostreig, simplement compteu deu pàgines en la mateixa posició de la pàgina i després compteu 36 entrades per arribar a la unitat següent de la mostra".

M. Greenacre i F. Udina. *Estadística I*. Universitat Oberta de Catalunya

5. Mostreig estratificat

En els tipus de mostreig estudiats fins ara no s'ha tingut en compte el coneixement previ que poguéssim tenir de les característiques de la població. De fet, tant en el mostreig aleatori simple com en el sistemàtic prevalen l'aleatorietat de la mostra escollida sobre altres consideracions que en poden millorar la representativitat.

Ara estudiarem el concepte d'estratificació, concepte que integra, en la selecció de la mostra, el possible coneixement de la similitud dels valors de la variable en certs col·lectius (estrats).

Exemple de mostreig estratificat: connexió gratuïta a Internet

Suposem un cas extrem per il·lustrar millor el concepte de mostreig estratificat. Imaginem una població de 1.000 habitants en què només 500 individus disposen de connexió a In-

ternet a casa. Imaginem que els 500 que tenen connexió estan a favor de la gratuïtat de les trucades per a connectar-s'hi i que tots els qui no tenen connexió hi estan en contra. És a dir, en el global de la població una meitat està a favor de la gratuïtat i l'altra meitat hi està en contra.

Suposem que extraiem una mostra aleatòria simple de deu individus de la població; no seria gens estrany obtenir-ne sis a favor de la gratuïtat i quatre en contra (de la mateixa manera que, quan tiro una moneda deu cops, no resulta sorprenent obtenir sis cares i quatre creus). En aquest cas, en la mostra estan a favor de la gratuïtat un 60%, mentre que en la població la proporció dels que estan a favor de la gratuïtat és només del 50%. És evident, doncs, que el fet de disposar de connexió a casa condiciona la resposta donada a la pregunta, per la qual cosa seria convenient explorar la població segons estrats. En aquest cas en tenim dos: el dels qui tenen connexió i el dels qui no en tenen. Així, doncs, per a evitar el biaix, cada estrat hauria d'aportar el 50% de la mostra, tal com succeeix en la població.

Una manera fàcil d'evitar el biaix que apareix en els casos d'estratificació és fer que la mostra contingui les mateixes proporcions d'individus amb característiques comunes diferents; d'aquesta manera, les proporcions d'individus en les mostres seran les mateixes que en la població global.

Exemple de la necessitat dels estrats

Suposem que volem estudiar el temps que els estudiants de la UOC dediquen a navegar per Internet. Si acceptem, com afirmen alguns estudis, que els homes hi dediquen més temps que les dones, podríem desitjar que en la nostra mostra la proporció de gèneres fos la mateixa que en el total de la UOC, per a poder obtenir, així, una estimació més fiable. Així, doncs, hauríem d'estratificar per gènere.

En general, si sospitem que la resposta a una certa pregunta depèn d'una característica dels individus i disposem d'una llista de la població on s'indica aquesta característica, el mostreig estratificat intenta reproduir en la mostra l'estructura de la població en relació a aquesta característica.

Reproducció de l'estructura probable de la població en la mostra estratificada

Imaginem que sospitem que estar a favor o en contra del fet que l'AVE passi per l'aeroport de Barcelona depèn del lloc de residència (els qui viuen a Barcelona tenen tendència a estar-hi a favor i els qui viuen fora de Barcelona tenen tendència a estar-hi en contra). En aquest cas, la població està dividida en dos estrats (els qui viuen a Barcelona i els qui viuen fora). Si volem obtenir una mostra estratificada, escollirem dins de cada estrat, a partir d'un cens del seus membres, una mostra aleatòria simple per poder completar així una mostra de la població de la mesura desitjada. Escollirem les mesures de les mostres aleatòries dins de cada estrat de manera que en la mostra global estiguin representats els estrats en la mateixa proporció que en la població.

En el **mostreig estratificat** els individus de la població es divideixen en grups disjunts (anomenats **estrats**). La mostra s'obté seleccionant una mostra aleatòria simple dins de cada estrat.

En aquest tipus de mostreig es planteja el problema de quina ha de ser la mesura de la mostra dins de cada estrat. Habitualment s'acostuma a demanar que, en la mostra, el nombre d'individus de cada estrat estigui en proporció al pes de l'estrat dins la població. Aquest tipus de mostreig serà més precís que el mostreig aleatori simple, i més si els individus de cada estrat són molt similars entre ells i molt diferents dels individus d'altres estrats.

Exemple de mostreig estratificat: estratificació per curs

Si estudiem el temps de connexió a Internet dels alumnes de secundària dels instituts públics d'un cert barri, podem sospitar que el temps de connexió depèn del curs en què estan matriculats (bé perquè en algun curs es fa una assignatura d'informàtica que inclou l'ús d'Internet, bé perquè l'edat n'afavoreix l'ús). Si volem una mostra de 100 estudiants estratificada per

"curs en què estan matriculats", hauríem d'aconseguir el nombre d'alumnes matriculats a cada curs i distribuir els individus de la mostra segons la proporció d'alumnes a cada curs. Suposem que els alumnes es distribueixen d'aquesta manera:

		Percentatge sobre el total
1 ESO	200	16,19%
2 ESO	250	20,24%
3 ESO	260	21,05%
4 ESO	250	20,24%
1 batxillerat	150	12,15%
2 batxillerat	125	10,12%
	1.235	

Per a obtenir ara quants individus de la mostra han de ser de cada curs (estrat), repartirem el total d'individus en la mostra (cent) segons el percentatge que pertoca a cada estrat:

	Percentatge sobre el total	Individus en la mostra
1 ESO	16,19%	16
2 ESO	20,24%	20
3 ESO	21,05%	21
4 ESO	20,24%	20
1 batxillerat	12,15%	12
2 batxillerat	10,12%	10
		99

Observem que a causa de l'arrodoniment "hem perdut" un individu. Per a obtenir els cent que necessitem, afegirem un alumne a un estrat escollit a l'atzar. Després hauríem d'obtenir, a partir de la llista dels individus matriculats a cada curs, una mostra aleatòria simple de la mesura corresponent a cada curs.

Podem resumir el procediment per obtenir una mostra estratificada de la població de la forma següent:

1) Els individus de la població s'agrupen en estrats disjunts.

2) La mostra s'obté assignant un nombre d'individus a cada estrat i després se seleccionen els individus necessaris per mostreig aleatori simple dins de cada estrat.

3) El nombre d'individus que cada estrat aporta a la mostra final es pot decidir segons diferents criteris, normalment es fa proporcionalment a la mesura relativa de l'estrat en la població.

Entre els altres criteris per a l'obtenció de la mostra...

... que aporta cada estrat podem destacar aquell segons el qual tots els estrats aporten el mateix nombre d'individus a la mostra final.

6. Mostreig per conglomerats

Imaginem que volem estudiar el temps que els estudiants de secundària de TOTS els instituts de Catalunya dediquen a navegar per Internet, seleccionant una mostra de tres-cents estudiants. Ens podem trobar fàcilment amb dos problemes:

- És difícil disposar de la llista de tots els estudiants de secundària de tot Catalunya.

- Encara que tinguéssim la llista i seleccionéssim una mostra aleatòria simple de la població, la recollida d'informació seria molt complexa, amb el cost econòmic que això representa. En aquesta situació es podria donar el cas d'haver de visitar tres-cents instituts diferents per a entrevistar un únic alumne a cadascun d'aquests.

En casos com aquests, podem optar per utilitzar mostreig per conglomerats.

Conglomerats són unitats (normalment físiques o geogràfiques) en què es distribueixen els individus de la població que cal investigar. En el **mostreig per conglomerats** se selecciona una mostra aleatòria de conglomerats i, dins de cada conglomerat, se selecciona a l'atzar una mostra dels seus individus.

Normalment el mostreig per conglomerats es fa en diferents etapes.

Exemple de mostreig per conglomerats

En el cas dels instituts, i amb l'objectiu d'obtenir una mostra d'estudiants de tot Catalunya, podríem procedir de la manera següent:

1) Seleccionem a l'atzar unes quantes comarques de Catalunya (conglomerat = comarca).

2) Dins de cada comarca seleccionem a l'atzar alguns instituts (conglomerats = instituts).

3) Dins de cada institut seleccionem a l'atzar unes classes (conglomerats = classes).

4) En les classes seleccionades, i a partir de la llista de classe, seleccionem una mostra aleatòria simple dels seus estudiants. Observeu que la llista dels alumnes d'una classe és fàcil d'aconseguir, mentre que la llista de tots els estudiants de primer d'ESO en tots els instituts catalans serà més complicada d'obtenir.

5) Unint les mostres obtingudes a cadascuna de les classes obtenim la mostra global.

Aquest mètode simplifica molt la recollida d'informació mostral, però té també diversos inconvenients:

a) Si els conglomerats són molt diferents els uns amb els altres, i tenint en compte que no tots els conglomerats estan representats en la mostra final, la mostra pot perdre representativitat.

b) Cada conglomerat ha de contenir tanta diversitat com la mateixa població; dit d'una altra manera, a cada conglomerat han d'estar representats (com si es

tractés d'una mostra aleatòria simple) totes les característiques de la població. En cas que en algun conglomerat només hi hagués individus d'alguna característica particular, aquesta característica es podria veure sobrerepresentada o subrepresentada en la mostra (depenent de si el conglomerat aporta individus o no a la mostra final).

Podem resumir el procediment per a obtenir una mostra per conglomerats de la manera següent:

1) S'estudia la població distribuïda en conglomerats, que són agrupacions naturals dels individus.

2) Se seleccionen a l'atzar alguns dels conglomerats.

3) Dins els conglomerats seleccionats es pren una mostra aleatòria simple dels seus individus o bé s'escullen tots els individus del conglomerat.

4) La mostra final és la reunió de les mostres obtingudes a cada conglomerat.

Finalment, cal insistir en la diferència entre estrat i conglomerat:

Observeu...

... la diferència entre l'ús d'estrats i de conglomerats.

a) En els estrats es tenen en compte grups dins la població que cal investigar, els individus dels quals tenen característiques comunes i diferenciades dels individus d'altres grups.

b) Els conglomerats representen agrupacions de la població (per zones geogràfiques o per proximitat generalment) en què podem pensar que es reprodueixen totes les característiques de la població; així, doncs, els conglomerats han de ser similars els uns amb els altres i cadascun d'aquests ha de contenir individus tan heterogenis com si es tractés de la població.

D'altra banda, la construcció de la mostra garanteix que tots els estrats hi són representats, mentre que no tots els conglomerats aporten individus a la mostra final.

7. Mostreig polietàpic

A continuació, comentarem un tipus de mostreig que combina els mètodes d'estratificació i de conglomerats, cosa que fa que es millori la representativitat de la mostra, alhora que es manté la simplicitat en la recollida de les dades.

El problema es planteja quan els conglomerats resulten ser molt homogenis respecte la característica que cal tractar: per exemple, agrupacions físiques molt utilitzades, com barris o comarques, agrupen en alguns casos individus amb característiques socioeconòmiques similars. Això pot provocar un biaix en la mostra si per atzar només seleccionem barris o comarques d'una certa categoria socioeconòmica.

Exemple de biaix en el mostreig polietàpic

Si només seleccionem barris de classe alta (com els que tots coneixem a la nostra ciutat), els resultats seran molt diferents que si seleccionem barris no tan afavorits.

Exemple de mostreig polietàpic

En primer lloc, l'accés a Internet des de casa depèn de tenir ordinador a casa, i això depèn, en definitiva, del nivell socioeconòmic i cultural de les famílies. Resulta evident que no tots els barris seran similars respecte de les facilitats de connexió a casa.

Si volem mostrejar el temps que dediquen a navegar per Internet des de casa seva els estudiants d'una certa gran ciutat (per exemple, Barcelona) i prenem com a conglomerats els barris, podem trobar que els individus dins de cada barri tenen comportaments similars, mentre que el comportament dels individus de barris diferents també pot ser molt diferent. En els barris de classe alta és possible que les millors condicions econòmiques facin que els estudiants puguin accedir a Internet amb més facilitat que els d'altres barris i, per tant, dedicar-hi més temps.

Per a eliminar parcialment aquest problema, el que podem fer és agrupar els conglomerats per estrats.

Exemple d'agrupació de conglomerats per estrats

En el cas dels barris, els podem agrupar en estrats, en aquest cas per nivell socioeconòmic. Així, doncs, si volguéssim aplicar una estratificació polietàpica en l'exemple de l'accés a Internet des de casa, podríem procedir de la manera següent:

- Distribuïm els barris de Barcelona en *estrats* (per exemple: nivell econòmic alt, nivell econòmic mitjà, nivell econòmic baix). Dins de cada estrat escollim una mostra aleatòria de barris. En aquest cas, seria també convenient tenir en compte la població de cada barri, de manera que a cada estrat, barris més poblats tinguin més probabilitat de ser escollits que els menys poblats. També cal escollir almenys un conglomerat (barri en aquest cas) de cada estrat.

- Després, dins de cada barri seleccionem aleatòriament instituts (conglomerat = institut). En aquest pas, es podria procedir directament sense estratificar si considerem que els instituts de cada barri tendeixen a reproduir la població de cada barri; és a dir, en cas que tinguem motius per a pensar que els instituts de cada barri són similars els uns amb els altres i que contenen tanta diversitat entre els seus estudiants com la mateixa població del barri, podem optar per no estratificar. (Estratificar o no els instituts depèn de la informació de què disposem i, moltes vegades, de la experiència prèvia en aquest tipus d'estudis.)

- Després, continuaríem amb les classes i, finalment, seleccionaríem una mostra aleatòria dins de cada classe.

En resum, en el mostreig polietàpic es combinen la idea d'estratificació i la de conglomerats i s'estratifiquen els conglomerats considerats, per a després obtenir mostres aleatòries de conglomerats dins de cada estrat.

Acabarem la presentació d'aquest tipus de mostreig amb un altre exemple:

***Sesgo* és *biaix* en castellà.**

"Por ejemplo, se desea tomar una muestra de la población española para estudiar la proporción de personas que están de acuerdo con las relaciones prematrimoniales. Si suponemos que la edad y el sexo pueden influir en la opinión, deberíamos tomar una muestra donde todas las características sean las mismas que en la población base, lo que implica una muestra estratificada. Por otro lado, si suponemos que las provincias son homogéneas respecto a la opinión, podemos ahorrar muchos costes seleccionando al azar 4 provincias y dentro de cada una de ellas una muestra aleatória, o mejor estratificada. Este procedimiento tiene el inconveniente obvio de que si las provincias no son homogéneas respecto la opinión (por ejemplo las provincias más ricas tienen opinión distinta de las más pobres) tendremos sesgos (que evitaremos estratificando las provincias por riqueza).".

D. Peña. *Estadística. Modelos y métodos 1. Fundamentos*. Alianza Universidad Textos.

8. Mostreig per quotes

Quan la estratificació no és possible o resulta molt cara, i també en casos en què no disposem d'una llista de la població que cal investigar, es pot recórrer a l'anomenat *mostreig per quotes*.

Imaginem que ara volem estudiar el temps de connexió a Internet des de casa seva que dediquen tots els joves de Barcelona d'entre quinze i vint-i-tres anys. Fàcilment, podem sospitar que el gènere i l'edat (lligada a la consecució de feina) són característiques que influeixen en el temps de connexió. Si volguéssim estratificar, doncs, segons el gènere i l'edat, hauríem de començar per aconseguir una llista de la població d'entre quinze i vint-i-tres anys que inclogués, com a mínim, l'edat i el sexe. Si bé aquesta llista pot no estar disponible o resultar massa cara, les dades estadístiques referents a la distribució per edat i sexe d'una certa població acostumen a estar disponibles en els mateixos serveis municipals; d'aquest manera podríem saber la quantitat de joves a cadascuna de les edats considerades i la proporció de gèneres dins de cada grup d'edat. Amb aquestes dades seria molt fàcil deduir quants individus de cada combinació (edat-gènere) ha de contenir una mostra per a reproduir les proporcions reals de la població.

Exemple de mostreig per quotes: quotes per gènere i edat

Considerem el temps de connexió a Internet des de casa dels joves de Barcelona entre quinze i vint-i-tres anys. En el cas que ens ocupa obtenim dades com aquestes:

Edat d'any a any de la població de Barcelona per sexe

Edats	Total	Homes	Dones
15 anys	16.362	8.390	7.972
16 anys	17.340	8.837	8.503
17 anys	18.888	9.663	9.225
18 anys	20.338	10.209	10.129
19 anys	21.538	10.908	10.630
20 anys	22.813	11.614	11.199
21 anys	24.098	12.304	11.794
22 anys	23.862	12.087	11.775
23 anys	23.986	12.131	11.855
TOTAL	189.225	96.143	93.082

Font: Padró Municipal d'Habitants 1996. Departament d'Estadística. Ajuntament de Barcelona. Obtingut de la pàgina web www.bcn.es

A partir d'aquestes dades resulta clar que per a obtenir una mostra per quotes (de, per exemple, 300 individus), aquesta haurà de contenir (arrodonint convenientment) 300 * (8.390 / 189.225) = 13 homes de quinze anys, 300 * (7.972 / 189.225) = 13 dones de quinze anys, i així successivament.

El pas següent serà distribuir aquest nombre d'individus (o quotes) entre els entrevistadors, de manera que cada entrevistador haurà d'aconseguir tants individus de cada parella (edat-gènere) com li marquin les quotes assignades.

Distribució d'individus entre els entrevistadors

Si tenim tres entrevistadors, enviarem cadascun d'ells a diferents zones i encarregarem al primer que entrevisti sis homes de quinze anys, al segon, quatre homes de quinze anys i al tercer, tres, per exemple. Farem aquesta distribució per cada parella edat-gènere. Els entrevistadors aniran preguntant fins que omplin les corresponents quotes. Així, un cop el primer entrevistador ha aconseguit l'opinió de sis homes de quinze anys, ja no anotarà cap més resposta d'aquest col·lectiu.

En el **mostreig per quotes** es distribueixen els individus de la població en diferents categories i s'assigna un nombre d'individus a cada categoria, de manera que la proporció d'individus de cada categoria en la mostra sigui similar a la proporció dins la població. Un cop calculades aquestes proporcions, l'entrevistador rep instruccions sobre el nombre d'individus que ha d'entrevistar a cada categoria (quota). Quan ha esgotat els individus que té assignats d'una certa categoria, deixa de recollir dades d'aquesta i continua amb les altres categories.

Característiques del mostreig per quotes

En els casos estudiats fins ara, l'entrevistador (que es qui fa les preguntes) rebia una llista de les persones a les quals havia d'entrevistar. En el mostreig per quotes l'entrevistador intervé en la selecció de les persones de la mostra, respectant sempre les quotes que se li assignen.

El mostreig per quotes és molt utilitzat en estudis d'hàbits de consum, de marquèting, etc. per la seva simplicitat i per la facilitat en la recollida de dades (que també fa que sigui molt més barat). En general, la informació recollida no és prou fiable per a fer un estudi estadístic en profunditat, però sí que resulta útil per a fer una primera aproximació al tipus de resposta que podem obtenir i per a poder dissenyar una recollida de dades més fiable.

9. Resum

En aquesta sessió hem estudiat diferents maneres d'obtenir mostres a partir d'una població fixada. En la taula següent recollim els diferents mètodes introduïts, les seves principals característiques i un recordatori dels exemples que s'han tractat a cada mètode:

Tipus de mostreig	Breu descripció	Algun exemple
Aleatori simple	**1.** Tots els individus de la població tenen la mateixa probabilitat de ser escollits. **2.** Els individus se seleccionen de manera independent els uns dels altres.	**1.** Una mostra de dotze estudiants d'una universitat
Aleatori simple usant taules de dígits aleatoris	Se selecciona a l'atzar el primer individu de la mostra i els següents se seleccionen a partir de la taula.	**1.** Una mostra de dotze estudiants d'una universitat
Sistemàtic	Se selecciona a l'atzar el primer individu de la mostra i els següents se seleccionen a intervals fixos.	**1.** Una mostra de dotze estudiants de la UOC **2.** Mostra aleatòria de 400 individus sobre la guia de telèfons

Tipus de mostreig	Breu descripció	Algun exemple
Estratificat	Els individus de la població es divideixen en grups disjunts (estrats). La mostra s'obté seleccionant una mostra aleatòria simple dins de cada estrat.	Temps de connexió a Internet dels alumnes dels instituts públics d'un cert barri. Estratificació per cursos
Per conglomerats	Conglomerats són unitats (normalment físiques o geogràfiques) en què es distribueixen els individus de la població que cal investigar. En el mostreig per conglomerats se selecciona una mostra aleatòria de conglomerats i, dins de cada conglomerat, se selecciona a l'atzar una mostra dels seus individus.	Temps de connexió a Internet dels alumnes dels instituts públics de TOT Catalunya. Conglomerats: comarca, institut i classe.
Polietàpic	S'aplica estratificació als conglomerats.	Temps de connexió a Internet des de casa seva dels alumnes dels instituts públics de Barcelona. Estratifiquem els conglomerats (barris) i instituts, si cal.
Per quotes	Es distribueixen els individus de la població en diferents categories i s'assigna un nombre d'individus a cada categoria, de manera que la proporció d'individus de cada categoria en la mostra sigui similar a la proporció dins la població. L'entrevistador va seleccionant els individus de la mostra fins a omplir les quotes.	Temps de connexió a Internet des de casa dels joves de Barcelona entre quinze i vint-i-tres anys. Quotes per gènere i edat

Exercicis

1. Acabeu de completar la mostra aleatòria simple a partir de la taula de dígits aleatoris dels dotze estudiants matriculats a la Universitat.

2. Acabeu de completar la mostra sistemàtica dels dotze estudiants matriculats a la Universitat.

3. A partir de les dades de la taula "Edat d'any a any", expliqueu com obtindríem una mostra de 500 homes entre divuit i vint-i-un anys (tots dos inclosos).

a) Estratificada per edat

b) Per quotes segons la edat

4. Acabeu de completar la mostra per quotes de l'exemple del temps de connexió a Internet dels joves de Barcelona d'entre quinze i vint-i-tres anys.

5. En l'edició de Catalunya del dia 28-4-2001, el diari El Periódico publicava la fitxa tècnica següent d'un sondeig (d'àmbit estatal) efectuat amb motiu de les eleccions autonòmiques al País Basc del dia 13-5-2001.

> "Tipus de mostreig: estratificat per autonomies i dimensió del municipi. Selecció aleatòria de les vivendes. Selecció d'individus segons quotes de sexe i edat representatives de la població de cada comunitat".

Comenteu els tipus de mostreig que apareixen.

Solucionari

1. Si continuem explorant la taula de dígits aleatoris i anem marcant els grups de quatre dígits a partir de l'individu inicial (alternant subratllat i caixes), obtenim el següent:

19223	95034	05756	28713	96409 12531	45544	82853
73676	47150	99400	01927	27754 42648	82425	36290
45467	71709	77558	00095	32863 29485	82226	90056
52711	38889	93074	60227	40011 85848	48767	52573

D'aquestes agrupacions de quatre dígits, les següents són menors que 1.400 i, per tant, corresponen als quatre primers alumnes que són escollits per a formar la mostra:

575 1396 19 118

Continuant aquest procediment, obtenim que la mostra final estarà formada pels estudiants corresponents als nombres següents:

575 1396 19 118 895 1181 656 91 881 1238 463 206

2. Si anem sumant 126 al nombre corresponent al primer individu de la mostra (que era el nombre deu), obtenim que la mostra estarà formada pels individus corresponents als nombres següents:

10 126 242 358 474 590 706 822 938 1054 1170 1286

3. Les dades que interessen es troben en la taula següent, obtinguda a partir de l'original:

Edats	Homes	Nombre d'individus en la mostra de 500 corresponents a l'edat
18 anys	10.209	114
19 anys	10.908	121
20 anys	11.614	129
21 anys	12.304	137
Total	45.035	501

Arrodonint, passem de 500: podem eliminar un individu de la mostra o conservar els 501 individus. Amb aquests càlculs tenim el nombre d'homes de cada edat que hi ha d'haver en la mostra. Ara:

a) Si volem estratificar, hem d'aconseguir una llista amb els noms de tots els homes de entre divuit i vint-i-un anys de Barcelona, en què se n'indiqui l'edat, i seleccionar una mostra aleatòria simple de la mesura necessària a cada estrat (és a dir, a cada edat); podem extreure les mostres aleatòries amb ajuda de l'ordinador o de la taula de dígits aleatoris.

b) Enviarem els entrevistadors a diferents zones (intentant d'obtenir la màxima representativitat) amb instruccions precises de quants homes de cada edat ha d'entrevistar. La unió de les respostes obtingudes pels entrevistadors ha de contenir tants homes de cada edat com s'ha trobat en la taula anterior.

4. Completem la taula per obtenir el nombre d'individus de cada parella (gènere-edat) que ha de contenir la mostra. Aquestes quotes es distribueixen entre els entrevistadors.

Edats	TOTAL	Homes	Dones	En la mostra	
				Homes	Dones
15 anys	16.362	8.390	7.972	13	13
16 anys	17.340	8.837	8.503	14	13
17 anys	18.888	9.663	9.225	15	15
18 anys	20.338	10.209	10.129	16	16
19 anys	21.538	10.908	10.630	17	17
20 anys	22.813	11.614	11.199	18	18
21 anys	24.098	12.304	11.794	20	19
22 anys	23.862	12.087	11.775	19	19
23 anys	23.986	12.131	11.855	19	19
Total	189.225	96.143	93.082	151	149

5. S'ha volgut que en la mostra final la proporció d'individus de cada comunitat autònoma i de cada mesura de municipi fos igual a la proporció respecte de la població total. S'han utilitzat conglomerats (= "vivendes") seleccionades de manera aleatòria dins de cada estrat (comunitat i municipi). Finalment, els entrevistadors han anat a les "vivendes" seleccionades amb unes quotes de sexe i edat per garantir que a cada comunitat la mostra conté la mateixa proporció segons gènere i edat que el total de la població de la comunitat.

Annex

Dígits aleatoris							
19223	95034	05756	28713	96409	12531	42544	82853
73676	47150	99400	01927	27754	42648	82425	36290
45467	71709	77558	00095	32863	29485	82226	90056
52711	38889	93074	60227	40011	85848	48767	52573
95592	94007	69971	91481	60779	53791	17297	59335
68417	35013	15529	72765	85089	57067	50211	47487
82739	57890	20807	47511	81676	55300	94383	14893
60940	72024	17868	24943	61790	90656	87964	18883
36009	19365	15412	39638	85453	46816	83485	41979
38448	48789	18338	24697	39364	42006	76688	08708
81486	69487	60513	09297	00412	71238	27649	39950
59636	88804	04634	71197	19352	73089	84898	45785
62568	70206	40325	03699	71080	22553	11486	11776
45149	32992	75730	66280	03819	56202	02938	70915
61041	77684	94322	24709	73698	14526	31893	32592
14459	26056	31424	80371	65103	62253	50490	61181
38167	98532	62183	70632	23417	26185	41448	75532
73190	32533	04470	29669	84407	90785	65956	86382
95857	07118	87664	92099	58806	66979	98624	84826
35476	55972	39421	65850	04266	35435	43742	11937
71487	09984	29077	14863	61683	47052	62224	51025
13873	81598	95052	90908	73592	75186	87136	95761
54580	81507	27102	56027	55892	33063	41842	81868
71035	09001	43367	49497	72719	96758	27611	91596
96746	12149	37823	71868	18442	35119	62103	39244
96927	19931	36809	74192	77567	88741	48409	41903
53909	99477	25330	64359	40085	16925	85117	36071
15689	14227	06565	14374	13352	49367	81982	87209
36759	58984	68288	22913	18638	54303	00795	08727
69051	64817	87174	09517	84534	06489	87201	97245
05007	16632	81194	14873	04197	85576	45195	96565
68732	55259	84292	08796	43165	93739	31685	97150
45740	41807	65561	33302	07051	93623	18132	09547
27816	78416	18329	21337	35213	37741	04312	68508
66925	55658	39100	78458	11206	19876	87151	31260
08421	44753	77377	28744	75592	08563	79140	92454
53645	66812	61421	47836	12609	15373	98481	14592
66831	68908	40772	21558	47781	33586	79177	06928
55588	99404	70708	41098	43563	56934	48394	51719
12975	13258	13048	45144	72321	81940	00360	02428
96767	35964	23822	96012	94591	65194	50842	53372
72829	50232	97892	63408	77919	44575	24870	04178
88565	42628	17797	49376	61762	16953	88604	12724
62964	88145	83083	69453	46109	59505	69680	00900
19687	12633	57857	95806	09931	02150	43163	58636
37609	59057	66967	83401	60705	02384	90597	93600
54973	86278	88737	74351	47500	84552	19909	67181
00694	05977	19664	65441	20903	62371	22725	53340
71546	05233	53946	68743	72460	27601	45403	88692
07511	88915	41267	16853	84569	79367	32337	03316