NOTA TÈCNICA

FIABILITAT ESTADÍSTICA DE LES ENQUESTES TERRITORIALS DE MOBILITAT

AUTORS

Xavier Roselló, Adjunt al Director Tècnic, ATM

ATM Àrea de Barcelona
Autoritat del Transport Metropolità

Àrea tècnica at

FEBRER DE 2011

CONTINGUTS

1 – JUSTIFICACIÓ

OBJECTE DE L'ESTUDI

Analitzar les condicions que han de complir les zones en què se subdivideix el territori respecte del nombre d'enquestes que s'hi ha realitzat perquè les estimacions siguin significatives. És a dir, quin nivell d'agregació del territori és compatible amb la construcció de matrius origen-destinació fiables.

Les enquestes de mobilitat quotidiana (EMQ) constitueixen una font d'informació encetada l'any 1996, i tenen per objectiu de conèixer tots els desplaçaments efectuats per la persona enquestada per qualsevol motiu al llarg d'un dia sencer.

La metodologia de realització ha evolucionat des dels inicis. Els anys 1996 i 2001 consistien en una enquesta domiciliària continguda en un quadern que el cap de família havia d'emplenar fent-hi constar els desplaçaments fets per tots els membres o només alguns de seleccionats al llarg de tota una setmana. Del 2006 ençà, se n'ha substituït la metodologia per l'enquesta telefònica amb suport CATI.

Aquest procediment escull aleatòriament una família i, tot seguit, un membre d'aquesta, a qui es demana d'enumerar tots els desplaçaments fets els dia anterior i, si es tracta d'un dilluns, també els d'un dels dos dies del cap de setmana precedent escollit a l'atzar. De cada desplaçament se sol·liciten els punts d'origen i de destinació, identificats per l'adreça postal, així com el motiu i el mode, la qual cosa aporta una informació exhaustiva de la mobilitat de les persones enquestades.

Les Enquestes de Mobilitat Quotidiana (EMQ) es duen a terme quinquennalment a una mostra gran de la població. Així, el 2006 la mostra va consistir en 106.000 persones més grans de 4 anys residents a Catalunya, les quals van informar sobre els 406.000 desplaçaments que havien fet en conjunt. L'univers considerat és el Padró d'habitants, que fa que es prenguin en consideració els residents a Catalunya però se n'excloguin els transeünts, entre els quals els turistes, generadors significatius de mobilitat.

L'EMQ06 suposava que cada desplaçament podia estar format per un cert nombre d'etapes, fins a un màxim de quatre. De cada etapa se'n coneixen els punts inicial i final, amb precisió d'adreça postal. Per a cada etapa es preveia fins a 21 modes de transport diferents. Pel que fa als motius, se'n considera un total de 23, 3 d'ocupacionals, 19 de personals i, el més freqüent de tots, la tornada a casa.

La Regió Metropolitana de Barcelona és el nucli central de Catalunya. Encara que només comprèn 1/10 de la superfície, la població assoleix 5 milions d'habitants dels 7,5 de tota Catalunya. En aquest territori l'EMQ06 va realitzar 40.000 enquestes, és a dir, un 40% del total. La resta de Catalunya té una taxa de mostratge superior a causa precisament de la seva dispersió demogràfica.

Les EMQ són enquestes de mobilitat que recullen tots els elements que cal saber d'un desplaçament. Per tant, són unes eines excel·lents per conèixer les tendències, les proporcions i les ràtios de tot tipus relatives a la mobilitat. No obstant, es revelen insuficients quan es tracta d'estimar matrius de mobilitat amb el nivell de desagregació que la planificació requereix. Tan sols amb una zonificació molt àmplia, on les zones siguin de la mida de comarques o grans municipis, poden inferir-se valors fiables dels fluxos entre zones. A mesura que hom redueix la grandària de les zones, també ho fa la mostra assignada a cada una de les caselles de la matriu corresponent a aquella zonificació i s'atenua de manera irremissible la capacitat predictiva de la mostra.

El dilema, doncs, és què cal mantenir agregat i què pot desagregar-se: els motius de desplaçament, els modes de transport o les zones del territori. I l'Estadística Inferencial ensenya que sovint no n'hi ha prou amb una sola concessió a l'agregació, sinó que cal fer-ne dues o totes tres.

FIGURA 1.
DADES GLOBALS RELATIVES A L'EMQ06

	CATALUNYA	RMB	Municipi de BARCELONA
Vegueries	7	1	-
Comarques	41	7	-
Municipis	947	164	1
Zones intramunicipals	1.142	330	63

L'objectiu del document present és analitzar les condicions que han de complir les zones en què se subdivideix el territori respecte del nombre d'enquestes que s'hi ha realitzat perquè les estimacions siguin significatives. O, dit en unes altres paraules, quin nivell d'agregació del territori és compatible amb la construcció de matrius origen - destinació fiables.

El document present és una versió refosa de textos diversos que s'han anat emetent a mesura que s'anava avançant en l'anàlisi, i que calia donar resposta als interrogants que apareixien al llarg de la recerca. S'ha optat per prescindir de tots ells i començar-ne la redacció de bell nou, tasca que ha donat com a fruit el document actual.

2 – PLANTEJAMENT TEÒRIC

Se suposa que una en zona determinada, dita ***i***, en un període de temps determinat s'originen un nombre de desplaçaments ***N***, que constitueixen la població.

Mitjançant les EMQ es coneix la destinació de ***n*** desplaçaments, que constitueixen la mostra. Els conceptes de població i mostra han de ser correlatius, és a dir, s'han de referir a un mateix concepte de desplaçaments. Així, si es volen tenir en compte els desplaçaments en vehicle privat per motiu estudi, la mostra a tenir en compte és només la referida a aquest mode i aquest motiu.

Els desplaçaments des de la zona origen determinada fins a una zona destinació genèrica, dita ***j***, segueixen una llei binomial. Sigui ***p*** la proporció mostral observada, és a dir, la proporció, respecte de la zona origen, de desplaçaments amb origen a la zona ***i*** i destinació a la zona **j**.

L'interval de confiança on es troba la proporció de la població, π, amb una confiança o probabilitat $1-\alpha$, ve donat per l'expressió següent que pot trobar-se a qualsevol manual d'Estadística:

És un interval de confiança simètric, centrat en l'estimador puntual, ***p***.

$$\hat{\pi} = p \pm z_{\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

S'anomena marge d'error, ***e***, al semi-interval de confiança, és a dir, a l'amplada d'aquest

Imagini's, com a exemple, el cas d'un flux des de ***i*** fins a ***j*** que representa el 8% de tots els

$$e = z_{\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

desplaçaments que tenen origen en aquella zona, i que aquesta estimació s'ha efectuat a partir d'una **mostra** realitzada a la zona origen que conté 155 desplaçaments. Si vol determinar-se l'interval amb una confiança del 90%, i per tant un risc del 10%, (z=1,65) aquest interval valdrà:

$$\pi = p \pm z_{\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0{,}08 \pm 1{,}65 \cdot \sqrt{\frac{0{,}08 \cdot (1-0{,}08)}{155}} = 0{,}08 \pm 0{,}036$$

és a dir, hi ha una probabilitat d'un 90% que la proporció del flux es trobi entre 4,4% (= 8% - 3,6%) i 11,6% (= 8% + 3,6%). És un interval que, dit de passada, sembla d'una amplada inacceptable als efectes de planejament.

3 – FORMULACIÓ DE LA CONDICIÓ

El nombre de fluxos entre dues zones és molt variable entre un lloc i un altre i, el que és més important, no pot parlar-se de valors grans i de valors petits de manera absoluta, ja que segons quin sigui l'objectiu i l'àmbit de l'estudi, un mateix valor serà gran en una situació mentre que en una altra seria petit. Cal, doncs, evitar criteris de significació basats en valors absoluts. Sembla preferible basar el criteri d'acceptació en aspectes relatius.

El criteri que es proposa consisteix a comparar l'estimació de la proporció ***p*** amb l'error associat ***e*** abans definit. Valors massa grans de ***e*** respecte de ***p*** faran una estimació inacceptable. És fàcil de veure que, a mesura que disminueix la proporció, l'error també ho fa, per bé que a un ritme inferior. Ho il·lustra la Figura 1, on s'ha suposat una mostra, *n*, de 2000 desplaçaments i un risc, α, del 10%

FIGURA 1.
RELACIÓ ENTRE ERROR I PROPORCIÓ DE LA MOSTRA

p PROPORCIÓ	*e* ERROR	*r=e/p* RELACIÓ ENTRE ERROR I PROPORCIÓ
50%	1,84%	3,7%
20%	1,48%	7,4%
10%	1,11%	11,1%
5%	0,80%	16,1%
2%	0,52%	25,8%
1%	0,37%	36,7%

S'ha suposat una mostra de 2000 desplaçaments i un risc del 10%.

Així, per un flux del 50%, és a dir, la meitat dels desplaçaments, l'error és del 1,8%, que només representa un 3,7% del valor estimat. En canvi, per a un flux de l'1%, valor força més habitual que l'anterior, l'error associat és del 0,37%, és a dir, un 36,7% de l'estimació.

El criteri que es proposa, doncs, és que per a qualsevol flux de desplaçaments, el quocient ***r*** *(=**e/p**)* sigui sempre inferior a un valor prefixat, γ. Aquesta condició es formalitza com segueix:

$$r = \frac{e}{p} = \frac{1}{p} \cdot z_{\alpha/2} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = z_{\alpha/2} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{\frac{1-p}{p}} \leq \gamma$$

Observi's que aquesta variable, *r*, està lligada estretament amb el Coeficient de Variació, *C*, definit com el quocient entre la desviació típica i la mitjana. És a dir:

$$C = \frac{s}{p} = \frac{1}{p} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = \frac{r}{z_{\alpha/2}}$$

La relació entre les dues variables és, doncs, constant, tot i que en el document es continua utilitzant la *r* per facilitat interpretativa.

Per tractar-la de manera correcta, però, cal anar un pas més enrere, i prendre en consideració les variables obtingudes directament de la mostra. Són les següents:

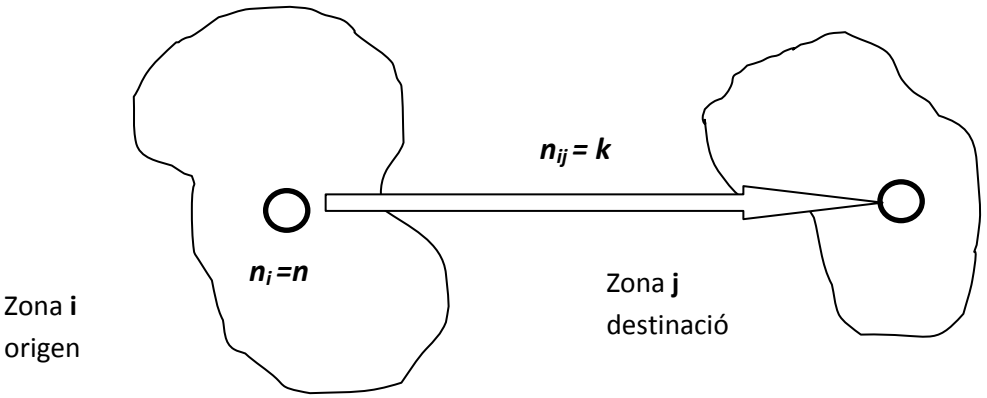
- ***n***: grandària de la mostra a la zona origen. Total de desplaçaments amb origen a la zona *i*.
- ***k***: elements de la mostra que tenen la destinació desitjada. Desplaçaments amb origen a la zona *i* i destinació a la zona *j*, també escrit n_{ij} per definició.
- ***p***: proporció mostral observada, és a dir, relació entre els desplaçaments amb la destinació desitjada i el total de desplaçaments originats en aquella zona. Coincideix amb el concepte utilitzat fins ara.

Evidentment es compleix:

$$p = \frac{k}{n} = \frac{n_{ij}}{n_i}$$

Les variables amb subíndex corresponen a la notació matricial habitual. En el document present se n'han definit unes d'equivalents per simplificar-ne la notació.

Esquemàticament:



Replantejant l'expressió de *r*, però substituint-hi la *p* per l'expressió definida més amunt s'obté:

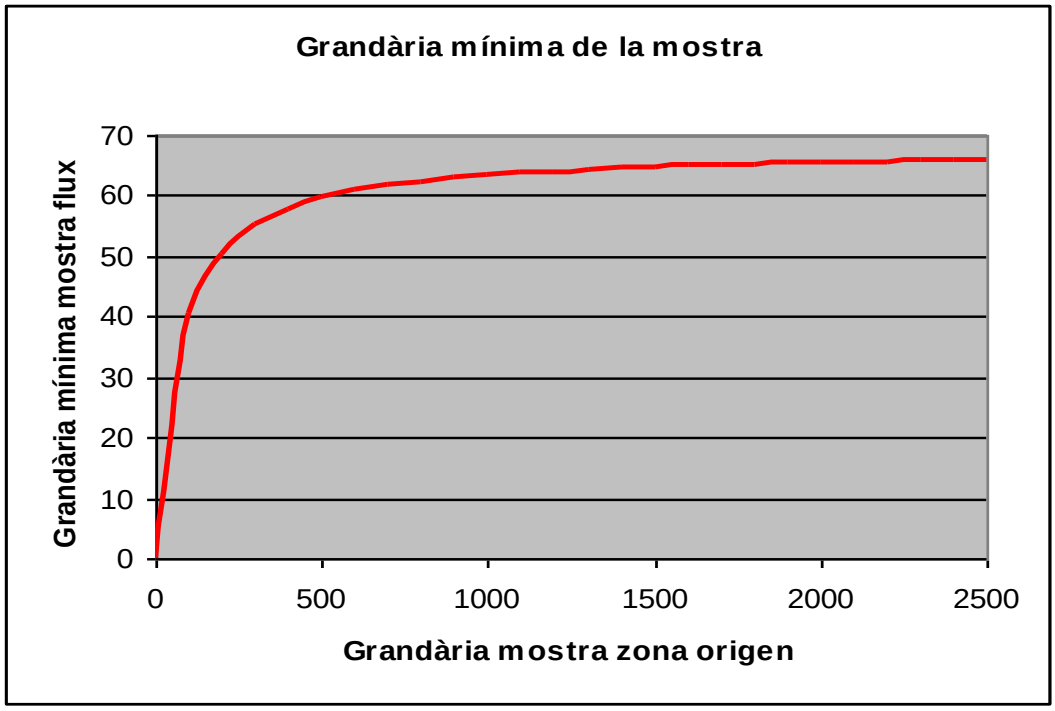
$$r = z_{\alpha/2} \cdot \sqrt{\frac{1-p}{pn}} = z_{\alpha/2} \cdot \sqrt{\frac{1-\frac{k}{n}}{\frac{k}{n}n}} = z_{\alpha/2} \cdot \sqrt{\frac{n-k}{kn}} \leq \gamma$$

D'on es dedueix el valor mínim de *k* que satisfà la desigualtat:

$$k \geq \frac{1}{\left(\frac{\gamma}{z_{\alpha/2}}\right)^2 + \frac{1}{n}}$$

que determina la talla mínima **en valor absolut** del flux de la mostra entre les zones *i* i *j* en funció de la mostra total obtinguda a la zona d'origen *i*. Es tracta d'una hipèrbola que té l'aspecte següent.

FIGURA 2.
RELACIÓ ENTRE ERROR I PROPORCIÓ DE LA MOSTRA



La mostra necessària perquè el flux d'una casella determinada sigui significatiu creix a mesura que ho fa la mostra total de la zona origen i tendeix asimptòticament al valor:

$$k = n_{ij} \geq \left(\frac{z_{\alpha/2}}{\gamma} \right)^2$$

Si s'adopten uns valors realistes com els següents:

- un **risc** del **10%** (α = 10% i z = 1,65)
- una **relació** entre error i proporció del **20%** (γ = 0,2)

la fórmula anterior esdevé:

$$k = n_{ij} \geq 68$$

Pot comprovar-se que a aquest valor només s'hi arriba en mostres força grans.

4 – DETERMINACIÓ DEL TERME D'ERROR

Fins al moment s'ha calculat l'error ***e*** a partir de l'aproximació normal a la llei binomial. Es tracta d'una aproximació que funciona bé per a valors grans d'***n*** i de ***k***, però que és inexacta quan aquests són petits, en què l'expressió anterior esdevé inaplicable. Malauradament, el cas de mostres petites és força freqüent a la pràctica. Per això, en aquest capítol s'han comparat els intervals d'error resultants aplicant la llei binomial, que són exactes, i l'aproximació normal, que només són aproximats. D'aquí es dedueix amb quines limitacions pot fer-se servir l'aproximació o quan cal anar a la utilització directa de la llei binomial.

4.1 - PLANTEJAMENT TEÒRIC

La llei binomial respon al model de les boles blanques i negres en una urna. Així, conegudes:

- ***π***: proporció de boles blanques a l'urna (paràmetre de l'univers)
- ***n***: nombre de boles extretes (grandària de la mostra),
- ***k***: nombre de boles blanques extretes a la mostra,

determina la probabilitat *P{k / π, n}* que s'esdevingui aquest fet, és a dir, que hi hagi **exactament *k*** boles blanques sabent que la proporció de la població és ***π*** i la mostra extreta és d'***n*** boles.

Però el problema a resoldre és el contrari: allò que es coneix és la grandària de la mostra, ***n*** i el nombre de boles blanques que conté aquesta, ***k***, i allò que es tracta de determinar és la proporció o interval de proporcions de boles blanques de la població, ***π,*** que és susceptible d'haver generat una mostra com aquesta amb una certa versemblança, és a dir, descartant les proporcions més improbables. De fet, es tracta de determinar l'interval de confiança més probable de *π*, que pot enunciar-se com segueix:

$$P\{\pi_L \le \pi \le \pi_U \mid n,k\} = 1-\alpha$$

Aplicant el teorema de Bayes, aquesta probabilitat és pot determinar de la manera següent:

$$P\{\pi \le \pi_0 \mid n,k\} = \frac{\int_0^{\pi_0} P\{k \mid n,\pi\} d\pi}{\int_0^1 P\{k \mid n,\pi\} d\pi}$$

on l'expressió sota el signe integral és la fórmula habitual de la llei binomial, és a dir:

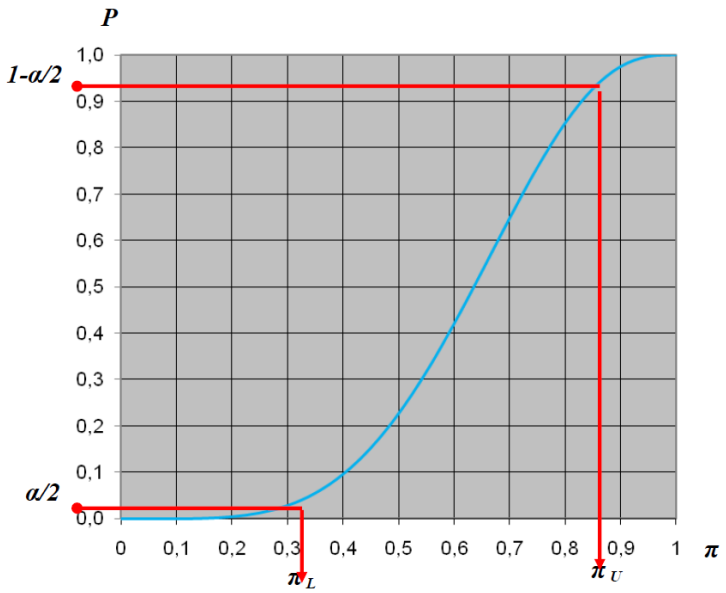
$$P\{k \mid n,\pi\} = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

A l'**Annex** s'explica de manera detallada l'expressió matemàtica de la probabilitat desitjada, però ara es fa èmfasi en la utilització de l'expressió.

4.2 - INFERÈNCIA ESTADÍSTICA

La funció $P\{\pi \leq \pi_0 \mid n,k\}$ serà una funció de distribució de probabilitat, que valdrà 0 per a π_0=0 i 1 per a π_0=1. Caldrà veure per a quins valors la probabilitat acumulada val precisament *α* i *1-α*, que correspondran respectivament als límits de l'interval de confiança, π_L i π_U definit més amunt. Observi's que per a cada parella de valors *n* i *k* es tindrà una funció diferent.

FIGURA 3.
EXEMPLE DE CORBA DE DISTRIBUCIÓ DE PROBABILITAT (ASSOCIADA A UNA MOSTRA DE *n*=6 I *k*=4)



Pot veure's que els límits de l'interval de confiança corresponen aproximadament a 0,33 i 0,87, que són els valors per sota dels quals hi ha respectivament un valor *α/2* (=5%) o *1-α/2* (=95%) de la probabilitat. És a dir, deixen a cada banda un 5% de probabilitat. D'aquesta manera es poden construir uns intervals lligats a cada ***n*** i a cada ***k*** havent fixat prèviament el risc α.

4.3 - DETERMINACIÓ DEL QUOCIENT R

Els intervals obtinguts per aquest procediment no són necessàriament simètrics respecte de l'estimador puntual central, *p* (= *k/n*). Tanmateix permeten determinar el valor de ***r*** definit al document inicial si se suposa que l'amplitud de l'interval de confiança és igual a ***2e***, per analogia amb la definició d'***e*** donada anteriorment. Per tant, si els límits de l'interval són π_L i π_U, i sabent que l'estimació puntual és $p = k/n = 0{,}67$, el quocient *r* es definirà com:

$$r = \frac{e}{p} = \frac{p_U - p_L}{2p}$$

A l'expressió anterior i a les següents s'han substituït els paràmetres π per les seves estimacions, *p*.

Continuant amb l'exemple, el valor de ***r*** en aquell cas seria:

$$r = \frac{p_U - p_L}{2p} = \frac{0{,}87 - 0{,}33}{2 \cdot 0{,}67} = 0{,}405$$

4.4 - TABULACIÓ

Per als diversos valors de n i de k ($\leq n$) s'han determinat p, e i r, tal com es mostra a la taula següent. També s'hi mostren els valors que s'obtindrien amb l'aproximació normal que, com pot veure's, són superiors i més poc ajustats. Igual que al llarg de tot el document s'ha suposat $\alpha = 10\%$ $\gamma = 0{,}2$.

FIGURA 4.
DETERMINACIÓ DELS INTERVALS DE CONFIANÇA EXACTE I APROXIMAT PER A DIVERSOS VALORS DE n I k

GRANDÀRIA DE LA MOSTRA	BOLES BLANQUES	VALORS EXACTES (BINOMIAL)					APROXIMACIÓ NORMAL	
n	k	p	p_L	p_U	$2e$	r	$2e$	r
4	3	75,0%	34,2%	92,4%	29,1%	38,8%	35,6%	47,5%
4	4	100,0%	55,0%	99,0%	22,0%	22,0%	0,0%	0,0%
5	4	80,0%	41,8%	93,7%	26,0%	32,4%	29,4%	36,8%
5	5	100,0%	60,6%	99.2%	19,3%	**19,3%**	0,0%	0,0%
6	5	83,3%	48,0%	94,6%	23,3%	28,0%	25,0%	30,0%
6	6	100,0%	65,2%	99.3%	17,1%	**17,1%**	0,0%	0,0%
7	6	85,7%	53,0%	95,3%	21,2%	24,7%	21,8%	25,4%
7	7	100,0%	68,7%	99.4%	15,4%	**15,4%**	0,0%	0,0%
8	7	87,5%	57,1%	95,9%	19,4%	22,2%	19,2%	22,0%
8	8	100,0%	71,7%	99.4%	13.9%	**13.9%**	0,0%	0,0%
9	8	88,9%	60,6%	96,3%	17,9%	20,1%	17,2%	19,4%
9	9	100,0%	74,1%	99.5%	12.7%	**12,7%**	0,0%	0,0%
10	8	80,0%	53,0%	92,1%	19,6%	24,4%	20,8%	26,0%
10	9	90,0%	63,6%	96,7%	16,6%	**18,4%**	15,6%	17,3%
10	10	100,0%	76,1%	99.5%	11,7%	**11,7%**	0,0%	0,0%
11	9	81,8%	56,2%	92,8%	18,3%	22,4%	19,1%	23,4%
11	10	90,9%	66,2%	97,0%	15,4%	**16,9%**	14,3%	15,7%
11	11	100,0%	77,9%	99,6%	10,9%	**10,9%**	0,0%	0,0%
12	10	83,3%	59,0%	93,4%	17,2%	20,6%	17,7%	21,2%
12	11	91,7%	68,4%	97,2%	14,4%	**15,7%**	13,1%	14,3%
12	12	100,0%	79,4%	99,6%	10,1%	**10,1%**	0,0%	0,0%
13	10	76,9%	53,4%	89,6%	18,1%	23,5%	19,2%	25,0%
13	11	84,6%	61,4%	93,9%	16,3%	**19,2%**	16,5%	19,5%
13	12	92,3%	70,3%	97,4%	13,6%	**14,7%**	12,2%	13,2%
13	13	100,0%	80,7%	99,6%	9,5%	**9,5%**	0,0%	0,0%

L'observació de la taula permet inferir-ne algunes conclusions interessants:

- Per obtenir una estimació acceptable, és a dir, amb $\gamma < 0{,}2$, cal una mostra mínima de $n = 5$. Per sota d'aquesta xifra no és possible cap estimació de cap casella amb origen a la zona esmentada.
- Per a mostres compreses entre 5 i 9 elements tan sols pot estimar-se una casella o destinació si la tenen com a tal tots els elements de la mostra, és a dir, si $k = n$. Altrament, tampoc no pot estimar-se cap casella de la fila.
- Per a $n \geq 10$, 11 i 12 tan sols pot estimar-se una casella de la fila sempre que la tinguin com a destinació tots els elements de la mostra o tots menys un, és a dir, $k \geq n\text{-}1$.
- A partir de $n \geq 13$, ja pot seguir-se la fórmula derivada de l'aproximació normal perquè, tot i que els intervals no coincideixen amb els exactes, el criteri de $\gamma \leq 0{,}2$ es compleix en els mateixos casos.

- L'estimació que se sol fer és puntual i el càlcul de l'interval de confiança s'utilitza només per decidir si l'estimació del flux d'una casella és acceptable o no. Aquesta és la raó de la regla donada més amunt.
- Tanmateix, si es volgués fer una estimació per interval caldria continuar utilitzant la fórmula exacta fins a valors força superiors d'*n* i de *k*.
- Els intervals exactes solen ser inferiors als aproximats si $k \leq n-2$. Tanmateix, si $k = n-1$ l'interval exacte és superior.
- Per a $k = n$, la fórmula aproximada és incapaç de calcular cap interval. En efecte, equival a demanar-se quina és la mínima proporció de boles blanques que acceptem que hi ha a l'urna sabent que totes les de la mostra han estat blanques. Observi's que l'interval de confiança no inclou el valor $p = 100\%$ perquè es considera improbable aquest valor en la població, és a dir, l'assumpció que tots els desplaçaments tenen la mateixa destinació.

El discurs precedent permet deduir quina és la mostra mínima necessària en una casella de la matriu de mobilitat per poder estimar-ne el nombre de desplaçaments amb un error relatiu inferior a ***γ***, que la pràctica ha fixat en 0,2.

n	**≤4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**≥13**
k_{min}	**-**	**5**	**6**	**7**	**8**	**9**	**9**	**10**	**11**	**Fórmula general**

Així doncs, el valor mínim de ***k*** (desplaçaments amb un mateix origen i una mateixa destinació) per a cada ***n*** (desplaçaments amb un mateix origen) ve donat a la taula següent:

D'11 endavant es fa servir l'aproximació normal, que s'ha estudiat en el capítol 4. Per a xifres molt grans pot prendre's el valor asimptòtic de l'expressió anterior que, amb els paràmetres utilitzats val 68.

5 - OBTENCIÓ DE LA MATRIU DE MOBILITAT

A continuació s'exposa un exemple del procediment, basat en dades fictícies, tot i que versemblants. Se suposa que és un territori dividit en 5 zones. Es parteix de la matriu mostral ***N***:

Matriu N

	1	2	3	4	5	n_i
1	2500	152	35	10	3	2.700
2	300	360	20	0	5	685
3	202	120	180	10	8	520
4	15	6	1	108	0	130
5	55	2	14	29	300	400
						4.435

Els coeficients d'elevació de cada zona, ***S***, o relació entre la mostra i la població, o invers de la taxa de mostratge, són:

1	185
2	150
3	120
4	80
5	55

L'aplicació de la fórmula del capítol 3 determina quin ha de ser el valor mínim de cada casella en funció de la mostra total de la zona, **K min**. Aquests valors són:

1	66
2	62
3	60
4	44
5	58

A partir d'aquí, la matriu aixecada o poblacional només podrà donar-se per aquelles caselles el valor de les quals superi el valor llindar definit més amunt. Altrament no se'n podrà fer l'estimació. Segons això, la matriu estimada resultant serà ***U***:

Matriu aixecada o estimada

	1	2	3	4	5	
1	462.500	28.120	-	-	-	499.500
2	45.000	54.000	-	-	-	102.750
3	24.240	14.400	21.600	-	-	62.400
4	-	-	-	8.640	-	10.400
5	-	-	-	-	16.500	22.000

Observi's que la suma de les caselles per files no coincideix amb els valors totals. Això és així perquè els valors totals són significatius malgrat no ser-ho els d'algunes caselles de la mateixa fila. Les caselles en blanc indiquen que l'estimació que se'n podria fer no és acceptable

6 – ANNEX

Deducció de la funció de distribució de la proporció poblacional a partir de les dades d'una mostra

En parlar de la llei binomial s'ha exposat que després d'extreure una mostra de ***n*** boles, ***k*** de les quals són blanques, la distribució de proporció de boles blanques de la població segueix l'expressió següent, deduïda aplicant del teorema de Bayes.

$$P\{p \le p_0 \mid n,k\} = \frac{\int_0^{p_0} P\{k \mid n,\, p\}dp}{\int_0^1 P\{k \mid n,\, p\}dp}$$

Les expressions sota el signe integral són la funció de probabilitat de la llei binomial, és a dir:

$$P\{p \le p_0 \mid n,k\} = \frac{\int_0^{p_0}\binom{n}{k}p^k(1-p)^{n-k}\,dp}{\int_0^1\binom{n}{k}p^k(1-p)^{n-k}\,dp} = \frac{\binom{n}{k}\int_0^{p_0}p^k(1-p)^{n-k}\,dp}{\binom{n}{k}\int_0^1 p^k(1-p)^{n-k}\,dp}$$

El **denominador** és una expressió que pot determinar-se si hom s'adona que és un cas particular de la funció Beta d'Euler el valor de la qual, al seu torn, es calcula a partir de la Gamma del mateix autor. En efecte:

$$\binom{n}{k}\int_0^1 p^k(1-p)^{n-k}\,dp = \binom{n}{k}\cdot B(k+1,n-k+1) = \binom{n}{k}\cdot\frac{\Gamma(k+1)\cdot\Gamma(n-k+1)}{\Gamma(n+2)} =$$

$$= \frac{n!}{k!\cdot(n-k)!}\cdot\frac{k!\cdot(n-k)!}{(n+1)!} = \frac{1}{n+1}$$

Observi's que el valor d'aquesta integral definida és només funció d'***n*** però independent de ***k.*** Per al càlcul de la resta de l'expressió, que de fet és el **numerador** amb un coeficient, que és l'invers del denominador, es determinen de manera progressiva o telescòpica, les diverses integrals indefinides, obtenint cada funció de *k* a partir de la funció corresponent al seu anterior.

Així, començant per ***k = n***:

$$P\{p \le p_0 \mid n,n\} = \frac{1}{1/n+1}\cdot\binom{n}{n}\cdot\int p^n\cdot(1-p)^0\,dp = (n+1)\cdot\frac{p^{n+1}}{n+1} = p^{n+1}$$

Per a ***k = n-1*** es resol una integració per parts aprofitant el resultat precedent:

$$P\{p \le p_0 \mid n,n-2\} = p^{n+1} + (n+1)\cdot p^n\cdot(1-p) + \frac{n(n+1)}{2}\,p^{n-1}\cdot(1-p)^2 =$$
$$= P\{p \le p_0 \mid n,n-1\} + \frac{(n+1)\cdot n}{2}\,p^{n-1}\cdot(1-p)^2$$

Cada funció pot expressar-se com l'anterior més un nou terme a mida que disminueix el valor de *k* en una unitat. Així, repetint el procés per a ***k = n-2:***

$$P\{p \le p_0 \mid n,n-1\} = (n+1)\cdot\binom{n}{n-1}\cdot\int p^{n-1}\cdot(1-p)^1\,dp =$$
$$= (n+1)\cdot n\left[\frac{1}{n}\,p^n\cdot(1-p) + \frac{1}{n}\int p^n\,dp\right] = (n+1)\cdot n\left[\frac{1}{n}\,p^n\cdot(1-p) + \frac{1}{n\cdot(n+1)}\,p^{n+1}\right] =$$
$$= p^{n+1} + (n+1)\cdot p^n\cdot(1-p) = P\{p \le p_0 \mid n,n\} + (n+1)\cdot p^n\cdot(1-p)$$

I per a ***k = n-3:***

$$P\{p \le p_0 \mid n,n-3\} = P\{p \le p_0 \mid n,n-2\} + \frac{(n+1)\cdot n\cdot(n-1)}{3!}\,p^{n-2}\cdot(1-p)^3$$

I per a ***k = n-4:***

$$P\{p \le p_0 \mid n,n-4\} = P\{p \le p_0 \mid n,n-3\} + \frac{(n+1)\cdot n\cdot(n-1)\cdot(n-2)}{4!}\,p^{n-3}\cdot(1-p)^4$$

I així successivament.

Un cop coneguda l'expressió analítica de la funció són fàcils de determinar els valors que resolen respectivament les equacions:

$$P\{p \le p_L \mid n,k\} = \alpha/2$$
$$P\{p \le p_U \mid n,k\} = 1-\alpha/2$$

per mètodes iteratius. Així és com s'ha calculat la taula del punt 4.5.