BUTLLETÍ DE LA SOCIETAT CATALANA DE MATEMÀTIQUES
Vol. 15, núm. 1, 2000. Pàg. 51–69

Punts en el pla: ordre o atzar?*

FRANCISCO MONTES, JORGE MATEU

1 Patrons de punts al pla

La figura 1 ens mostra la imatge d'un tarongerar en el qual resulta òbvia la distribució regular dels arbres. Els tarongers es planten seguint diferents estructures regulars; la de la imatge és coneguda en la Plana Baixa com el *quadre reial*, però si el llaurador vol aprofitar una mica més el terreny s'afegeix un arbre al centre del quadrat i tenim el què es coneix com *el cinc d'oros* o *dau*.[1]

FIGURA 1: Tarongerar a La Safor.

* Conferència pronunciada a la Tercera Trobada Matemàtica de la Societat Catalana de Matemàtiques, que va tenir lloc a la Universitat de València el març de 2000. Els autors volen fer palès el seu agraïment a la Societat Catalana de Matemàtiques per convidar-los a participar en aquesta III Trobada, i al company Guillermo Ayala, pels suggeriments i comentaris.

1 Això és el que ens ha dit el nostre amic Joan Monterde, que deu saber-ne perquè viu a Vila-Real.

Es tracta en aquest cas d'un patró puntual que ha estat produït artificialment per la intervenció humana; però també la natura produeix espontàniament distribucions puntuals semblants, com podem veure a la figura 2, en la qual es mostren els centres d'un conjunt de cèl·lules de l'endoteli humà. Aquesta estructura cel·lular té forma de mosaic i les seues característiques, la distribució observada n'és una, estan lligades a la forma i grandària de les cèl·lules.

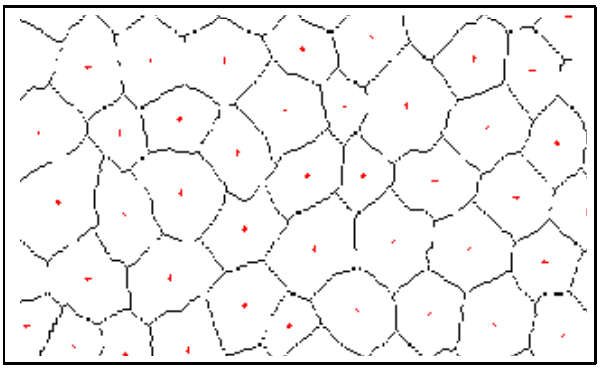

FIGURA 2: Distribució dels centres de les cèl·lules de l'endoteli.

En tots dos casos, la regularitat de l'estructura observada podria explicar-se per l'existència d'una mínima distància entre els objectes que la formen, conseqüència sens dubte de la impenetrabilitat de la matèria que es manifesta mitjançant una acció de rebuig entre els objectes, tot i que en el cas dels arbres impenetrabilitat i rebuig han estat reinterpretats per l'home.

Un patró puntual contrari sembla donar-se en la figura 3. Es tracta ara de les localitzacions de 62 plançons de sequoia distribuïts en un quadrat de 23 m^2, que ofereixen una clara imatge d'agregació que s'explica fàcilment si hi afegim informació complementària: el fet és que cada grup de plançons ha crescut al voltant de sequoies *mare* que sabem que estan presents en la regió estudiada, però que no apareixen representades en la imatge [3].

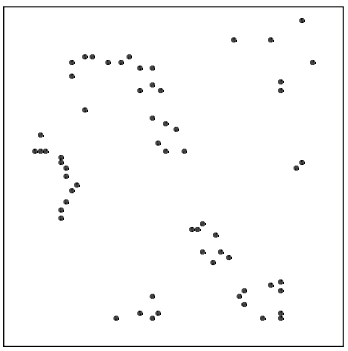

FIGURA 3: Localitzacions de 62 plançons de sequoia en un quadrat de 23 m^2.

Ni arbres ni cèl·lules no ofereixen cap dubte respecte del patró que la seua distribució espacial segueix. Una altra cosa serà trobar un model teòric que s'ajusti adequadament als patrons observats, però d'això no ens n'ocuparem en aquestes notes. El nostre objectiu és menys pretensiós, o potser caldria dir que és més primari. En efecte, tant si els objectes es rebutgen com si mostren atracció, en tots dos casos els patrons puntuals presenten una *interacció* entre els punts. El que es persegueix és descriure un patró amb absència d'interacció en el qual els punts s'hagen distribuït de manera *completament aleatòria*.[2] L'interès per conèixer en què consisteix una distribució d'aquestes característiques en el pla es justifica, si més no, per la conveniència de comptar amb un origen o model de referència, tot i que, com assenyala Diggle [3], l'aleatorietat espacial completa[3] representa una situació idealitzada pràcticament inassolible. Tanmateix, es tracta d'una hipòtesi acceptable com a primera aproximació, raó per la qual es fa servir com a *hipòtesi nul·la* en els tests estadístics per contrastar la presència d'interacció, però a més a més té un paper de divisòria entre els dos tipus de patrons anteriors, els *regulars* i els *agregats*.

La imatge de les localitzacions de 63 brots de pins negres japonesos ([8]) que ens mostra la figura 4 hauria de reforçar el nostre argument, en la mesura que no sembla possible, a primer cop d'ull, endevinar l'existència d'interacció de cap tipus entre aquestes localitzacions. Fiar-ho tot a la informació visual és massa perillós per massa subjectiu. És, doncs, convenient disposar d'un element de comparació.

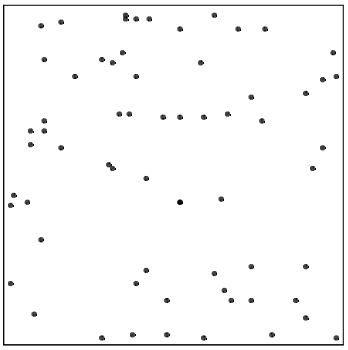

FIGURA 4: Localitzacions de 63 brots de pins negres japonesos en un quadrat de 5,7 m^2.

2 Aleatorietat espacial completa en el pla

Admès l'interès de disposar d'un patró de punts en el pla que representa l'absència d'interaccions, hauríem de trobar la manera de conèixer-ne característiques. En parlar de característiques aquí, ens estem referint a trobar un model probabilístic que descriga adequadament el seu comportament.

2 Potser caiguem en una contradicció terminològica en parlar d'un *patró completament aleatori*, perquè el que caracteritza la distribució completament aleatòria dels punts és, precisament, l'absència de cap patró.

3 La qualificació de l'aleatorietat com a *completa* és l'element clau en aquesta denominació, perquè qualsevol distribució de punts en el pla que no siga determinista, haurem d'adjectivar-la com d'aleatòria.

Aquesta modelització requereix en primer lloc esbrinar dos aspectes del problema:

1. com que parlar de model probabilístic és parlar de variables aleatòries, caldrà saber quina o quines són les variables d'interès lligades a la distribució de punts al pla, i
2. caldrà també trobar aquelles hipòtesis que millor tradueixen en termes probabilístics la idea de distribució completament aleatòria. Aquestes hipòtesis ens són imprescindibles per a poder obtenir la distribució de probabilitat de les variables aleatòries abans esmentades.

2.1 Variables lligades als patrons puntuals aleatoris

La distribució aleatòria de punts en el pla que puguem observar és una de les possibles manifestacions (*realitzacions* s'anomenen en inferència estadística) del mecanisme aleatori que l'ha generat. En la figura 5 podem veure com un mateix mecanisme aleatori es manifesta de manera diferent en dues realitzacions distintes.

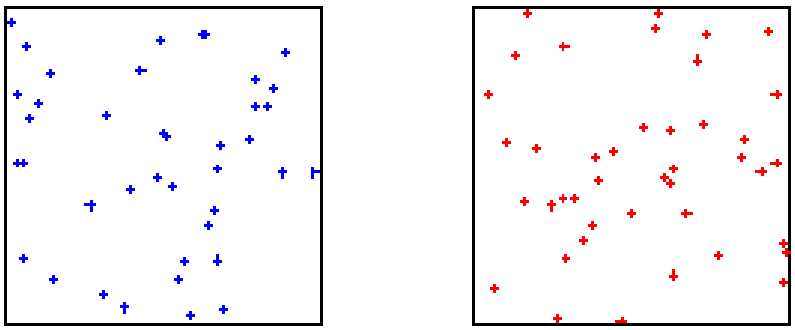

FIGURA 5: Dues realitzacions d'un mateix patró puntual aleatori.

Lligades al patró puntual aleatori podem considerar dues variables que, d'acord amb el que acabem de dir, són aleatòries. La *variable que compta els punts* que cauen a l'interior d'una determinada regió acotada del pla és una variable discreta, mentre que les *coordenades dels punts* són variables contínues. L'aproximació al problema que presentem en aquestes notes està basada en la primera, per bé queles unes i les altres estan relacionades.

2.2 Hipòtesis d'aleatorietat completa

Per tal d'evitar confusions caldrà introduir-hi una mica de nomenclatura. Ens interessen els *punts* que formen part del fenomen estudiat, la distribució dels quals ens ocupa; però aquests punts estan localitzats en *punts* del pla. Són dues categories de punts amb significats força diferents, que haurem de distingir. D'ara endavant ens referirem als punts que pertanyen a la distribució com a *esdeveniments*, mentre que els punts del pla seran senzillament *punts*.

L'estudi dels patrons aleatoris d'esdeveniments té interès per a nosaltres en la mesura que representen idealitzacions de fenòmens naturals: punts d'una superfície en els quals han incidit, per exemple, gotes de pluja, partícules atòmiques,

partícules de pols...; posicions d'arbres en un bosc; centres de cèl·lules distribuïdes en una preparació microscòpica; per parlar tan sols del cas del pla. Així les coses, sembla raonable pensar que no hi ha multiplicitat en el fenomen estudiat i que, en conseqüència, en cada punt del pla no podrem trobar més d'un esdeveniment. Tampoc és destrellatat pensar que el nombre d'esdeveniments (arbres, cèl·lules, partícules de pols...) que incideixen en una determinada regió acotada és finit, la qual cosa expressarem exigint que el nombre esperat d'esdeveniments siga finit.

Hipòtesi 1: *No hi ha esdeveniments múltiples, i el nombre esperat d'esdeveniments en una regió acotada del pla és finit.*

Pensem ara en les variables aleatòries que compten els esdeveniments que han caigut en dos subconjunts acotats qualssevol. En la figura 6 hem dibuixat dos d'aquests conjunts, A i B. Com que es tracta de dos conjunts amb intersecció buida, és raonable pensar que el nombre d'esdeveniments que cauen en l'un i en l'altre no s'influeixen mútuament. És a dir, si N_A i N_B designen les variables de comptatge respectives, n'estem postulant la independència.

Hipòtesi 2: *Si A i B són dos conjunts disjunts i N_A i N_B les variables aleatòries que compten els esdeveniments que hi trobem, N_A i N_B són independents.*

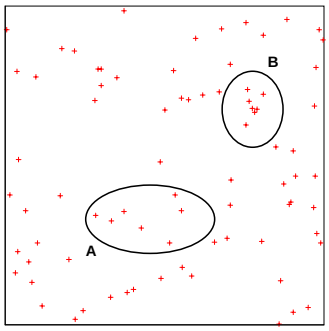

FIGURA 6: Incidència de punts distribuïts completament a l'atzar en el quadrat unitat en els subconjunts A i B.

Si la distribució dels esdeveniments és completament a l'atzar, un esdeveniment qualsevol no ha de mostrar cap preferència per un o altre subconjunt del pla a l'hora d'incidir-hi. Això sí, com més gran (major àrea) siga el subconjunt, més probable serà que hi vaja a parar. És a dir: a igual àrea, igual probabilitat que l'esdeveniment caiga en un conjunt o altre, independentment de la seua posició.

Hipòtesi 3: *Si A i B són dos conjunts disjunts amb la mateixa àrea, les probabilitats p_A i p_B, que un esdeveniment qualsevol hi incidesca, són iguals.*

Reparem que tota distribució aleatòria de punts en el pla dona lloc, per a cada subconjunt acotat del pla A, a la variable aleatòria N_A, el significat de la qual ja coneixem. En realitat estem parlant d'una família de variables aleatòries $\mathcal{N} = \{N_A,\ A \textit{ de Borel i acotat}\}$, que constitueixen allò que en probabilitat s'anomena un *procés*

estocàstic.[4] Conèixer el comportament probabilístic de la distribució aleatòria dels esdeveniments, és conèixer el comportament probabilístic del procés estocàstic que defineix, la qual cosa es pot fer mitjançant les *distribucions finitodimensionals*, que no són res més que les distribucions de probabilitat conjuntes de qualsevol subfamília finita, $N_{A_1}, N_{A_2}, \ldots, N_{A_m}$, de $\mathcal{N}$.

Tot seguit obtindrem les distribucions de probabilitat lligades a les variables N_A per al cas d'aleatorietat espacial completa. Presentarem dues aproximacions al problema:

- La primera és un desenvolupament que Pitman esbossa en el seu interessant text introductori a la probabilitat [9] tot recorrent a arguments probabilístics senzills.
- La segona utilitza un raonament analític basat en una senzilla equació diferencial, i pot trobar-se, amb lleugeres variants, en els textos de Feller [4], Gnedenko [6], Kingman [7] i Rényi [10].

En tots dos casos, és clar, les hipòtesis d'abans tenen un paper fonamental.

2.3 Una quadrícula i poc més

Suposem una distribució d'esdeveniments en el quadrat unitat Q i siga N el seu nombre. Establim successives particions de Q, Q_n, mitjançant n quadrats iguals, $n = 4, 16, 64, \ldots$, d'àrea $1/n$, i en cadascuna d'aquestes particions definim N_n com el nombre de quadrats de la partició que contenen algun esdeveniment. En la figura 7 es mostra Q i les tres primeres particions, Q_4, Q_{16} i Q_{64}, en les quals apareixen en gris els quadrats que contenen algun esdeveniment.

La seqüència N_n gaudeix de les dues propietats següents:

P1: $N_4 \leq N_{16} \leq N_{64} \leq \cdots$ perquè la construcció de les successives particions ens assegura que cada quadrat amb esdeveniments al seu interior dóna lloc, en la partició posterior, almenys a un altre quadrat que també conté algun esdeveniment (vegeu-ho en la figura 7), i

P2: la seqüència està acotada per N i, gràcies a la primera de les hipòtesis, que ens assegura que no hi ha multiplicitat en els esdeveniments, $N_n = N,\ \forall n \geq n_0$, on el valor de n_0 depèn de la distribució dels esdeveniments i serà distint per a cada realització del procés.

Com que N i les N_n són variables aleatòries, aquestes propietats ens permeten obtenir la distribució de probabilitat de N a partir de la d'aquelles[5] gràcies al límit,

$$P(N = k) = \lim_{n \to \infty} P(N_n = k). \tag{1}$$

La distribució de probabilitat de N_n i N. Com que tots els quadrats de la partició tenen la mateixa àrea, $1/n$, podem aplicar la hipòtesi 3, designar per p_n la probabilitat, comuna a tots els quadrats de la partició, de contenir un esdeveniment qualsevol. Però, tractant-se d'una partició, els quadrats són disjunts, i

4 El lector es deu haver adonat que A i B no poden ser qualssevol; han de ser conjunts de Borel del pla a més d'acotats, encara que és cert que per trobar conjunts del pla que no siguen de Borel cal recórrer a una d'aquelles sofisticades contruccions tan estimades en la nostra professió.

5 Vegeu els detalls tècnics en l'apèndix.

la hipòtesi 2 ens permet afirmar que la incidència o no d'un esdeveniment en un quadrat és una prova de Bernoulli amb probabilitat d'èxit p_n i independent de la dels altres quadrats. En resum, N_n es distribuirà com una binomial de paràmetres n i p_n, $N_n \sim B(n, p_n)$.

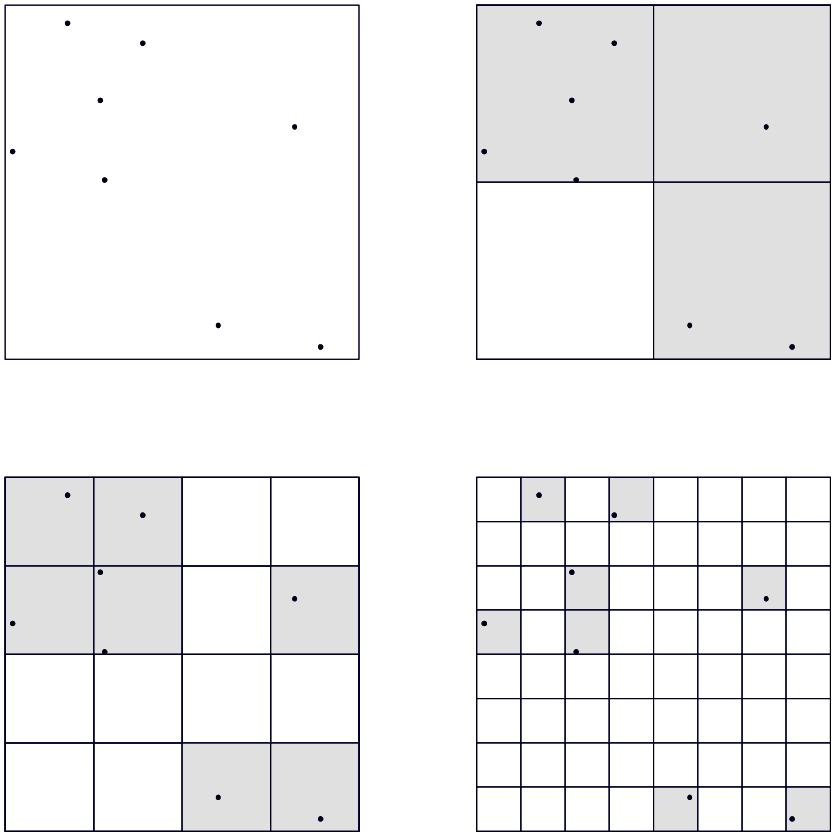

FIGURA 7: Digitalització d'un patró completament aleatori en el quadrat unitat.

Un conegut resultat de probabilitat afirma que

$$B(n, p_n) \xrightarrow{n\to\infty} P(\lambda),$$

on $P(\lambda)$ designa la distribució de Poisson de paràmetre λ, el valor del qual és

$$\lambda = \lim_{n\to\infty} np_n.$$

Però, és convergent la seqüència dels np_n? Per a cada N_n, $E(N_n) = np_n$, nombre esperat de quadrats de la partició que contenen algun esdeveniment. Pel que hem dit abans estarà acotat per $E(N)$ i hi convergirà, i per la primera de les hipòtesis tenim garantit que $E(N) < \infty$. Així doncs, la variable aleatòria

N es ditribueix com una Poisson de paràmetre λ, que representa el nombre esperat d'esdeveniments en Q, quadrat unitat.

La distribució de probabilitat de N_B, B qualsevol. Suposem en primer lloc que B és un subconjunt *simple* de Q, és a dir, està format per la unió de n_B quadrats d'àrea $1/n$, tal com es mostra a la figura 8.

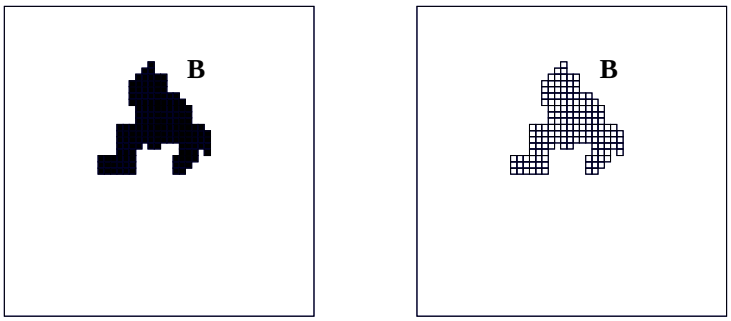

FIGURA 8: Un subconjunt de Q format per la unió de n_B subquadrats d'àrea $1/n$.

Podem emprar el raonament d'abans, tot substituint N per N_B i N_n per $(N_B)_n$. Arribarem a la conclusió que les $(N_B)_n \sim B(n_B, p_n)$ i, en conseqüència, $N_B \sim P(\lambda_B)$. Pel que fa al valor del paràmetre λ_B, adonem-nos que $n_B = n\|B\|$, on $\|B\| = \text{àrea}(B)$, perquè som dintre del quadrat unitat; així doncs,

$$\lambda_B = \lim_{n_B \to \infty} n_B p_n = \lim_{n \to \infty} n p_n \|B\| = \lambda \|B\|.$$

Per a un B qualsevol, podrem aproximar-lo interiorment mitjançant una seqüència de conjunts simples B_n, on

$$B_n \in \mathcal{P}(Q_n),\ B_n \subset B \ \text{ tal que } \ d(B, B_n) = \min_{B_i \in \mathcal{P}(Q_n)} d(B, B_i)\,.$$

Les variables aleatòries de comptatge associades a la família són Poisson de paràmetre λ_{B_n} i mitjançant un apropiat pas al límit[6] conclourem que

$$N_B \sim P(\lambda \|B\|). \tag{2}$$

Indepèndencia de N_A i N_B per a $A \cap B = \varnothing$. La indepèndencia de les distribucions de N_A i N_B és una conseqüència inmediata de la hipòtesi 2 i del fet que, com que són tots dos disjunts, les seqüències de conjunts simples que els aproximen també ho són.

La generalització d'aquestes conclusions, per a qualsevol subconjunt acotat del pla que puguem fer servir per a observar la distribució completament aleatòria d'esdeveniments, és immediata, i és coneguda com el Teorema de Poisson.

6 Els detalls tècnics d'aquest límit i el tipus de convergència involucrada es mostren en l'apèndix.

1 TEOREMA (TEOREMA DE POISSON) *Si una distribució d'esdeveniments en el pla verifica les hipòtesis 1, 2 i 3, existeix una constant λ tal que:*

1. *per a qualsevol conjunt acotat B, el nombre d'esdeveniments, N_B, que cauen dintre de B és una variable aleatòria Poisson amb paràmetre $\lambda\|B\|$,*
2. *per a conjunts acotats disjunts $B_1, \ldots, B_k$, les variables $N_{B_1}, \ldots N_{B_k}$ són mútuament independents.*

Diem aleshores que la distribució d'esdeveniments és un *procés de Poisson amb intensitat* λ, el nombre esperat d'esdeveniments per unitat d'àrea.

2.4 Una senzilla equació diferencial

La segona aproximació para esment en la distribució d'esdeveniments a una escala menuda. Pensem, per exemple, en un cercle A_t de radi t i considerem

$$\begin{aligned} p_n(t) &= P(N_{A_t} = n), \\ q_n(t) &= P(N_{A_t} \leq n). \end{aligned}$$

Com que N_{A_t} creix amb t, les funcions $p_n(t)$ i $q_n(t)$ són monòtones creixent i decreixent, respectivament, i com que totes dues estan acotades per la unitat, tindran un nombre de discontinuïtats a tot estirar numerable, i seran diferenciables quasi per totes parts.

A mesura que el radi del cercle creix, la variable N_{A_t} canviarà del valor n al valor $n+1$ quan la circumferència de A_t creue un dels esdeveniments. La probabilitat que això passe entre t i $t+h$ és la probabilitat que hi haja un esdeveniment en l'anell $A_h = A_{t+h} - A_t$. És a dir,

$$P(N_{A_{t+h}} = n+1 | N_{A_t} = n) = P(N_{A_h} = 1 | N_{A_t} = n), \tag{3}$$

probabilitat que serà menuda si h és menut.

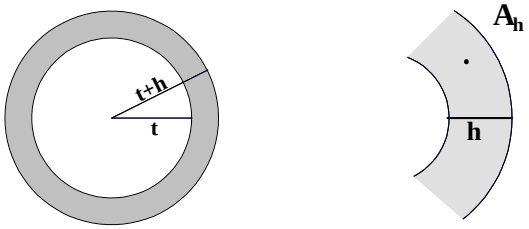

FIGURA 9: Els cercles A_t i $At + h$ i el seu anell circular A_h.

La hipòtesi 2, com que lliga la probabilitat d'incidència d'un esdeveniment a l'àrea del conjunt, ens permet escriure $P(N_{A_h} \geq k) \approx 0, \ \forall k \geq 2$, sempre que h siga

menut. En aquesta situació, si $\lambda(t) = E(N_{A_t})$, tindrem

$$\begin{aligned}\lambda(t+h) - \lambda(t) &= E(N_{A_{t+h}} - N_{A_t}) \\ &= \sum_{k\geq 1} kP(N_{A_{t+h}} - N_{A_t} = k) \\ &= \sum_{k\geq 1} kP(N_{A_h} = k) \approx P(N_{A_h} = 1). \end{aligned} \tag{4}$$

D'altra banda,

$$\begin{aligned} q_n(t) - q_n(t+h) &= (1 - q_n(t+h)) - (1 - q_n(t)) \\ &= P(N_{A_{t+h}} > n) - P(N_{A_t} > n) \\ &= P(\{N_{A_{t+h}} > n\} - \{N_{A_t} > n\}) \\ &= P(\{N_{A_{t+h}} > n\} \cap \{N_{A_t} \leq n\}) \\ &= \sum_{k=0}^{n} P(\{N_{A_{t+h}} > n\} \cap \{N_{A_t} = k\}) \\ &= \sum_{k=0}^{n} P(\{N_{A_h} > n-k\} \cap \{N_{A_t} = k\}) \\ &\approx P(\{N_{A_h} = 1\} \cap \{N_{A_t} = n\}) \\ &= P(N_{A_h} = 1 | N_{A_t} = n) P(N_{A_t} = n). \end{aligned} \tag{5}$$

La hipòtesi 3 ens assegura la independència de les variables de comptage quan els conjunts són disjunts; combinant-la amb (3), (4) i (5), podem escriure, sempre amb h menuda,

$$q_n(t) - q_n(t+h) = p_n(t)\{\lambda(t+h) - \lambda(t)\}.$$

Fent $h \to 0$,

$$-\frac{dq_n}{dt} = p_n \frac{d\lambda}{dt}.$$

Com que $p_n = q_n - q_{n-1},\ n \geq 1$ i $p_0 = q_0$, tindrem

$$\frac{dp_0}{dt} = -p_0 \frac{d\lambda}{dt}, \quad \frac{dp_n}{dt} = (p_{n-1} - p_n)\frac{d\lambda}{dt},\ n \geq 1. \tag{6}$$

De la primera equació de (6) deduïm

$$\frac{d}{dt}(\log p_0 + \lambda) = 0,$$

i com que $p_0(0) = 1$ i $\lambda(0) = 0$,

$$\log p_0 + \lambda = 0 \to p_0(t) = e^{-\lambda(t)},\ \forall t. \tag{7}$$

Com que

$$\frac{d}{dt}\{p_n e^{\lambda}\} = e^{\lambda}\left[\frac{dp_n}{dt} + p_n \frac{d\lambda}{dt}\right],$$

la segona equació de (6) pot també escriure's de la forma

$$\frac{d}{dt}\{p_n e^{\lambda}\} = p_{n-1} e^{\lambda} \frac{d\lambda}{dt},$$

i com que $p_n(0) = 0,\ \forall n \geq 1$,

$$p_n(t) = e^{-\lambda(t)} \int_0^t p_{n-1}(x) e^{\lambda(x)} \frac{d\lambda}{dx} dx.$$

Només cal integrar repetidament i comptar amb (7) per obtenir

$$p_n(t) = e^{-\lambda(t)} \frac{\lambda(t)^n}{n!},$$

i concloure que N_{A_t} té una distribució Poisson amb paràmetre $\lambda(t)$.

dia	**gen.**	**febr.**	**març**	**abr.**	**maig**	**juny**	**jul.**	**agost**	**set.**	**oct.**	**nov.**	**des.**
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

TAULA 1.

3 Un exemple: el sorteig del servei militar del 1970 als EUA

El sorteig que l'any 1940 es va fer als EUA per tal d'establir l'ordre d'incorporació al servei militar va rebre gran quantitat de crítiques per la falta de rigor amb què es va dur a terme. Analitzat amb un mínim de rigor científic, va quedar palès que les condicions d'equiprobabilitat i independència que calia exigir no hi eren satisfetes. Quan l'any 1969 es va tornar a plantejar la conveniència d'un sorteig, van sorgir tot un seguit de prejudicis, per vèncer els quals el president d'aleshores, Richard Nixon, va haver de publicar un decret on establia que les lleves per al servei militar de l'any

1970 haurien de fer-se mitjançant una selecció aleatòria de les dates de naixement dels xicots implicats.

Detalls sobre les precaucions que es varen prendre, sobre com es va dur a terme el sorteig, així com l'anàlisi rigorosa del resultat, poden consultar-se en un interessant i entretingut article de Fienberg [5]. En aquest article pot trobar-se l'ordre d'extracció de les 366 dates de naixement (la data del 29 de febrer també va ser considerada), que nosaltres reproduïm en la taula 1.

Provar que les condicions d'aleatorietat eren complertes era escaient, no tant perquè un decret ho exigia, com pel fet que les diferents lleves es farien en funció de les necessitats, i era previsible, com el Ministeri de Defensa i la mateixa Casa Blanca varen reconèixer, que no fos necessari incorporar els nascuts en les dates incloses en l'últim terç d'extraccions. Aquest darrer argument ens sembla concloent per justificar l'interès del treball de Fienberg.

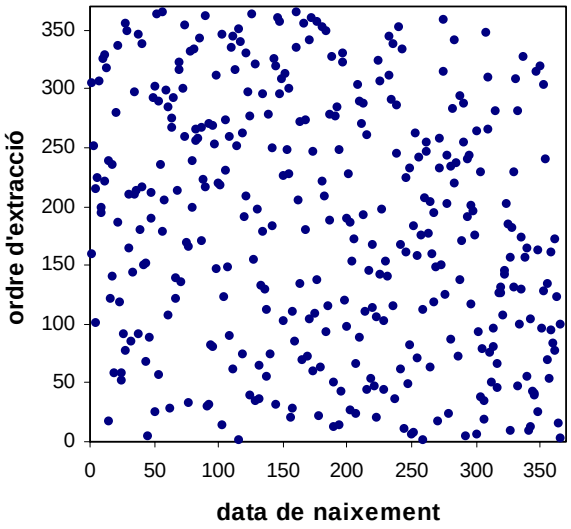

FIGURA 10: Gràfic de dispersió de l'ordre d'extracció i la data de naixement.

La compatibilitat del resultat amb l'exigida aleatorietat pot contrastar-se amb diferents mètodes (vegeu l'article de Fienberg o el de Corberán i Montes [2]), però fixarem la nostra atenció en aquell que estudia la distribució bidimensional d'esdeveniments associada al resultat. De què parlem? La figura 10 ens respon la qüestió. S'hi mostra la distribució d'esdeveniments que resulta de representar els 366 punts de coordenades (x, y), on x és la data de naixement expressada com una xifra entre 1 i 366 i y és l'ordre d'extracció d'aquesta data.

Si el sorteig ha estat com calia, el patró ha de ser completament aleatori[7] i no ha de mostrar cap interacció. Una primera anàlisi visual del gràfic mostra una certa mancança de punts als extrems de la bisectriu del primer quadrant i, potser també,

7 Algun lector pot pensar que aquesta situació no encaixa del tot amb el que hem contat abans. Doncs ben pensat, perquè els esdeveniments estan ara sobre un reticle i no sobre tot $\mathbb{R}^2$; però no es tracta de dur aquí el rigor fins a les darreres conseqüències, sinó de veure com aquesta eina estadística pot ajudar-nos en una situació aparentment molt allunyada del context habitual d'aquella.

una tendència a l'agregació, que tindria com a conseqüència la presència d'àrees en blanc relativament grans. Però es tracta, en qualsevol cas, d'apreciacions molt subjectives.

Hauríem de trobar un mètode per contrastar la hipòtesi d'aleatorietat completa. L'obtenció i exposició rigorosa d'un test d'hipòtesi apropiat va molt més enllà de l'objectiu d'aquestes ratlles. A més a més, cal dir que la cosa no és gens trivial per les dificultats que es troben per a obtenir la distribució de probabilitat dels possibles estadístics involucrats. Però el que sí que podem fer és un desenvolupament intuïtiu i pràctic d'un dels mètodes, el que utilitza les distàncies entre esdeveniments.

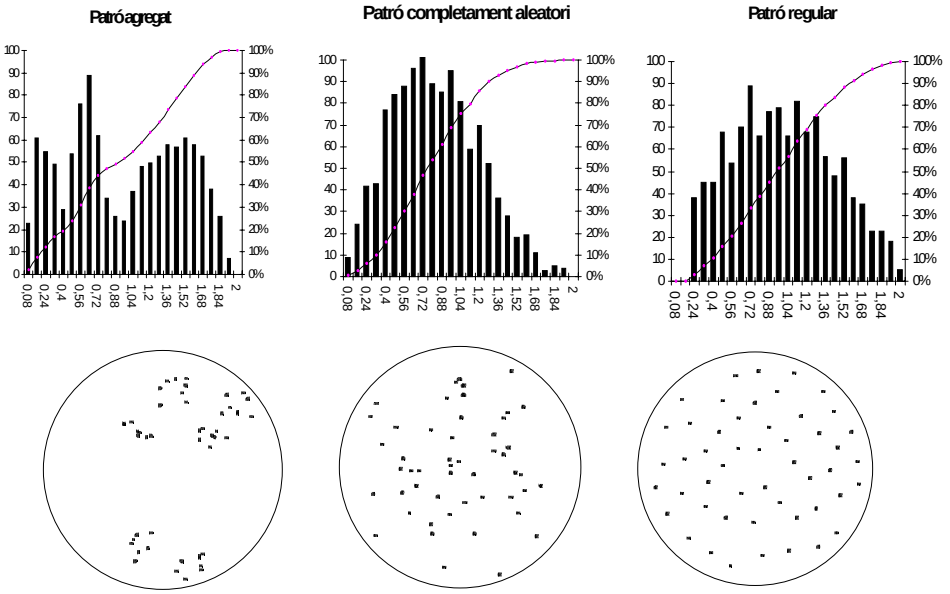

FIGURA 11: Histogrames i *fded* per a tres patrons simulats.

3.1 Distàncies entre esdeveniments

Una característica numèrica per a fer una descripció resumida del patró dels n esdeveniments observats en un recinte acotat és la *funció de distribució empírica (fded)* de les $\frac{1}{2}n(n-1)$ *distàncies* entre ells. Per veure la capacitat discriminadora de la *fded* observem la figura 11. S'hi mostren els histogrames i la *fded* de les distàncies entre els esdeveniments de les realitzacions que apareixen a sota de cada gràfic. En tots tres casos es tracta de recintes circulars de radi 1, i el nombre d'esdeveniments és molt semblant, 48 per al patró agregat i 50 per als altres dos. Pot observar-se que

- en el patró *agregat*, hi una major proporció de distàncies menudes i grans amb molta menor freqüència de distàncies intermèdies, la funció de distribució té un pendent més pronunciat en el seu inici,
- en el patró *regular*, les distàncies més menudes són inexistents o, en tot cas, molt menys freqüents que abans, mentre que les intermèdies tenen una presència major i amb un cert equilibri entre les diferents classes,

- en el patró *completament aleatori*, la distribució, encara que una mica segada, sembla la d'una normal, amb cues (valors extrems) de pes semblant i poc important, i una gradació de valors intermedis amb major freqüència dels valors centrals.

Admès el paper discriminador de la *fded*, toca ara obtenir-la. La seua definició és ben senzilla: en cada punt x, és la proporció de distàncies que no superen x,

$$\hat{F}(x) = \frac{\#\{d_{ij} \geq x\}}{\frac{1}{2}n(n-1)},$$

on # representa *el nombre de*. Si coneguéssem $F(x)$, la funció de distribució teòrica de les distàncies, una primera aproximació al contrast d'aleatorietat podria consistir a representar l'ordenada $\hat{F}(x)$ enfront de l'abscissa $F(x)$, que si el patró és compatible amb la hipòtesi d'aleatorietat completa hauria de donar lloc a alguna cosa molt semblant a la bisectriu del primer quadrant. Cal deixar clar que aquest és un mètode empíric i molt subjectiu, tret que la gràfica diferesca molt de l'esmentada bisectriu, i sovint es fa molt difícil prendre cap decisió.

Els mètodes convencionals de la inferència estadística exigeixen el coneixement de la distribució de $\hat{F}(x)$ sota la hipòtesi d'aleatorietat completa. Dissortadament això no és possible o, com a mínim, és molt complicat. Què hi podem fer? Les alternatives passen per emprar mètodes basats en simulacions de distribucions completament aleatòries d'esdeveniments de característiques semblants al nostre problema. Vegem-ne un parell i apliquem-los a les dades del sorteig.

3.1.1 Mètode de les envoltures superior i inferior de la fded Dir que el sorteig compleix les condicions d'aleatorietat equival a dir que l'ordre d'extracció ha estat una permutació aleatòria dels 366 primers naturals. El que farem és generar $n-1$ permutacions aleatòries i obtenir per a cadascuna la seua *fded*, $\hat{F}_i(x)$, $i = 2, \ldots, n$ i comparar-les totes amb la $\hat{F}_1(x)$ derivada del sorteig. Una manera senzilla de comparar-les es obtenir les *envoltures superior* i *inferior* mitjançant

$$S(x) = \max_{i=2,\ldots,n} \hat{F}_i(x), \qquad I(x) = \min_{i=2,\ldots,n} \hat{F}_i(x).$$

Si les dades del sorteig són completament aleatòries, $\hat{F}_1(x)$ no hauria de distingir-se de les $\hat{F}_i(x)$, $i = 2, \ldots, n$ i aleshores tindríem

$$P(\hat{F}_1(x) > S(x)) = P(\hat{F}_1(x) < I(x)) = \frac{1}{n}, \quad \forall x. \tag{8}$$

Si sistemàticament la $\hat{F}_1(x)$ és per dalt de $S(X)$ o per sota de $I(x)$, això serà un clar indici en contra de la hipòtesi establerta.

Per analitzar les dades del sorteig hem generat 99 permutacions aleatòries de l'1 al 366; això vol dir que les probabilitats de (8) valen 0,01, i és, per tant, molt improbable que $\hat{F}_1(x)$ supere $S(x)$ o siga inferior a $I(x)$. La gràfica conjunta de les tres funcions no és massa aclaridora, perquè les diferències són molt menudes i no poden apreciar-se, però si representem $\hat{F}_1(x) - I(X)$ i $\hat{F}_1(x) - S(x)$ en una mateixa gràfica, com hem fet a la figura 12, veurem que mentre que la primera és sempre positiva, la segona no sempre és negativa, perquè pren valors positius entre $x = 50$ i $x = 128$, aproximadament. Caldria concloure que l'extracció no va ser aleatòria, malgrat la cura que es va posar perquè així fos.

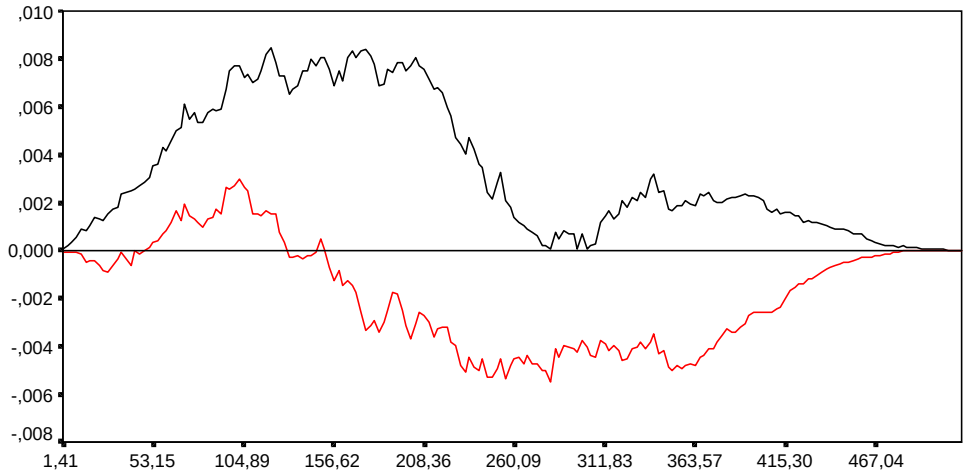

FIGURA 12: Comparació de $S(x)$, $\hat{F}_1(x)$ i $I(x)$.

3.1.2 Test de Montecarlo per a l'interval interquartílic L'altre mètode de contrast que proposem fa ús d'una característica numèrica associada a la *fded* de les distàncies, i forma part d'una família de mètodes coneguts genèricament com a *Tests de Montecarlo*, introduïts per Barnard en 1963 [1]. En essència, es tracta de comparar la característica numérica estudiada en el patró observat amb els valors d'aqueixa mateixa característica en cadascun dels $n-1$ patrons completament aleatoris que hem simulat, per a la qual cosa ordenarem de menor a major els n valors, l'observat i els simulats. Si la hipòtesi d'aleatorietat completa (hipòtesi nul·la) és satisfeta pel patró observat, el valor que la característica numérica hi pren podrà ocupar qualsevol lloc en l'ordenació amb la mateixa probabilitat, $1/n$. Tal com fem en el contrast d'hipòtesis clàssic, també ara valors extrems[8] en l'ordenació són evidències en contra de la hipòtesi nul·la. El métode permet també establir el nivell de significació en funció de l'ordre a partir del qual rebutgem la hipòtesi nul·la. Així, si en un test bilateral rebutgem la hipòtesi quan el valor observat es troba entre els k més grans o els k més menuts, el nivell de significació del nostre test serà $2k/n$.

Tornem a les dades del sorteig. Quan tenim un conjunt de dades numèriques, *l'interval interquatílic*, IIQ, és una mesura de dispersió definida mitjançant

$$IIQ = p_{75} - p_{25},$$

on p_{75} i p_{25} són, respectivament, el *tercer* i el *primer quartils*, que divideixen el conjunt de les observacions, pel que fa al seu nombre, en 3/4 i 1/4 el p_{75} i en 1/4 i 3/4 el p_{25}.

L'interval interquartílic de les $66795 = \frac{1}{2} \times 366 \times 365$ distàncies obtingudes a partir de les dades del sorteig val $IIQ = 139,5$. Per a les 99 simulacions els valors dels IIQ varien entre 138,4 i 136,4. Vol dir això que, en ordenar tots junts els valors, aquell és el major i per tant ocuparà la darrera posició. La conclusió és òbvia:

8 Segons que el contrast siga *unilateral* o *bilateral* tindrem compte d'un extrem o de tots dos.

rebutjarem la hipòtesi d'aleatorietat completa perquè el nivell de significació associat al valor de IIQ que hem trobat és del 2 %, donat que el contrast és bilateral, perquè la hipòtesi alternativa és la manca d'aleatorietat en l'extracció de les dates de naixement.

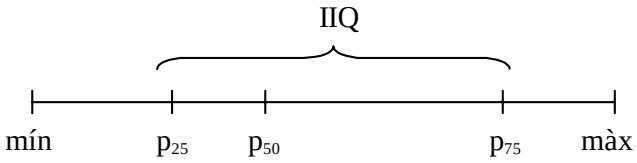

FIGURA 13: L'interval interquartílic.

Tot i que el contrast siga bilateral, el fet d'haver trobat un valor de l'IIQ tan gran ens dóna informació sobre el patró que les extraccions segueixen. Pensem que els patrons agregats tenen major dispersió que els regulars (potser pagarà la pena donar una nova ullada a la figura 11 per confirmar-ho), i això és el que sembla assenyalar-nos aquest valor. En els articles de Fienberg [5] i Corberán i Montes [2] es descriuen proves complementàries que confirmen aquesta darrera hipòtesi.

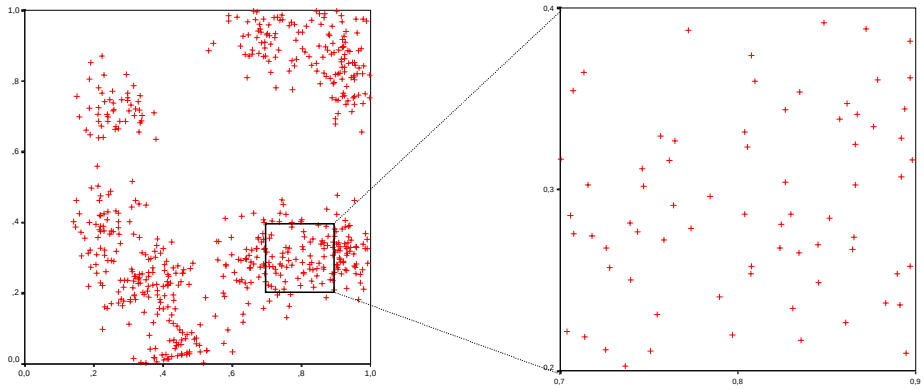

FIGURA 14: Un mateix patró observat a dues escales diferents.

4 I si res no fos el que sembla?

Per concloure, una mena d'advertència. Podria ser que les coses no fossen com ens sembla, perquè no hàgem sabut triar el punt d'observació apropiat. L'elecció de l'escala d'observació és fonamental, com queda ben palès en la figura 14. A l'esquerra es mostra la simulació d'un patró agregat en què els esdeveniments, com en el cas dels plançons de sequoia, s'han generat al voltant d'altres esdeveniments *pares* mitjançant un patró completament aleatori. Els *pares* no hi són representats i segueixen un patrò regular o de rebuig, és a dir, entre ells hi ha una distància mínima, la *distància de rebuig*.

Si observem la realització del procés de molt més a prop, a una escala més menuda, i ens fixem en el quadrat $[0.7,0.9]\times[0.2,0.4]$, el patró agregat desapareix i la

distribució dels esdeveniments que ens mostra el gràfic de la dreta sembla la d'un patró completament aleatori. Aquesta situació, artificial en l'exemple, és freqüent en molts processos biològics, raó per la qual és convenient, quan les dades ho permeten, analitzar el fenomen a dues escales clarament diferenciades.

Apèndix

A) La convergència de N_n a N

El límit (1) és cert si les N_n convergeixen en llei a N. Una manera indirecta de provar-ho és demostrar que hi ha convergència puntual de les N_n a N, perquè aquest tipus de convergència implica la convergència en llei. Vegem-ho.

Les variables N i N_n, $\forall n$ estan definides sobre l'espai de probabilitat $(\Omega, \mathcal{A}, P)$. A cadascun dels punts de l'espai correspon una realització distinta del procés estocàstic, $\Pi(\omega)$, que, observat a través del quadrat unitat, Q, donarà lloc a la seqüència numèrica $N_n(\omega)$ i al valor $N(\omega)$. Si

$$d_{min} = \min_{e_i, e_j \in \Pi(\omega)} d(e_i, e_j),$$

com que els quadrats que formen la partició Q_n tenen àrea $1/n$, en triar $n \geq 4^{n_0}$, amb $n_0 > 2/d_{min}^2$, tan sols podrà haver-hi un esdeveniment en cadascun dels quadrats de la partició i $N_n(\omega) = N(\omega)$. Així doncs,

$$\lim_{n\to\infty} N_n(\omega) = N(\omega), \quad \forall \omega \in \Omega,$$

i (1) està justificada perquè

$$N_n \xrightarrow{llei} N.$$

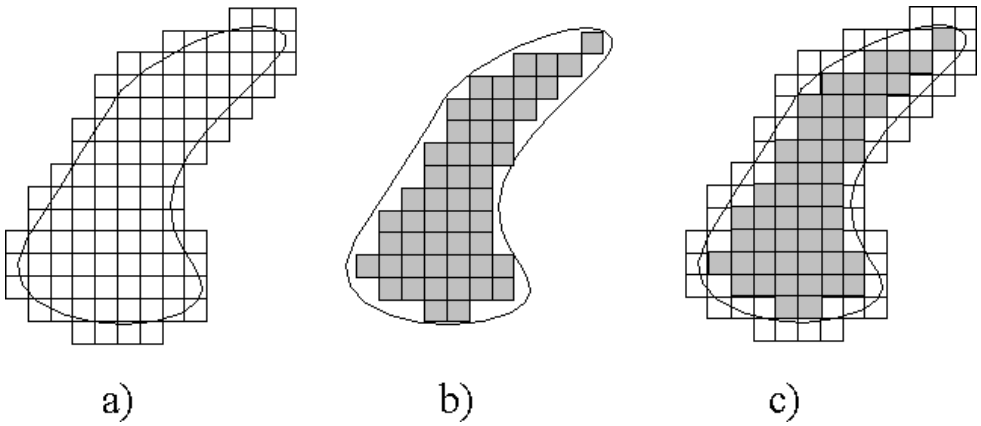

FIGURA 15: La figura mostra per a un conjunt qualsevol B, a) l'aproximació superior B_n^*, b) l'aproximació inferior B_n i c) la diferència entre totes dues, per a una determinada partició.

B) La convergència de N_{B_n} a N_B

També ara haurem de provar que

$$N_{B_n} \xrightarrow{llei} N_B,$$

perquè (2) siga certa.

Considerem, a més de la seqüència dels B_n, la dels B_n^*, definits mitjançant

$$B_n^* \in \mathcal{P}(Q_n),\ B \subset B_n^* \text{ tal que } d(B, B_n^*) = \min_{B_i^* \in \mathcal{P}(Q_n)} d(B, B_i^*).$$

Tenim que $B_n \subset B \subset B_n^*$.

Pel que fa a les variables aleatòries associades a totes dues famílies, considerem els conjunts

$$A_n^\epsilon = \{\omega;\ |N_{B_n^*}(\omega) - N_{B_n}(\omega)| \geq \epsilon\} = \{\omega;\ N_{B_n^* - B_n}(\omega) \geq \epsilon\}.$$

El conjunt $B_n^* - B_n$ és un conjunt simple i la variable $N_{B_n^* - B_n}$ es distribuirà com una Poisson de paràmetre $\lambda \|B_n^* - B_n\| = \lambda s_n$; aleshores

$$P(A_n^\epsilon) = P(N_{B_n^* - B_n} \geq \epsilon) = \sum_{k \geq \epsilon} e^{-\lambda s_n} \frac{(\lambda s_n)^k}{k!}.$$

Però $s_n = \|B_n^* - B_n\| \downarrow 0$ amb $n,\ \forall \epsilon$, i

$$P(A_n^\epsilon) \overset{n \to \infty}{\longrightarrow} 0, \quad \forall \epsilon.$$

El que hem provat és que la diferència $N_{B_n^*} - N_{B_n}$ convergeix en probabilitat a 0, i en conseqüència

$$N_{B_n} \xrightarrow{P} N_B.$$

Tan sols ens queda recordar la implicació

$$N_{B_n} \xrightarrow{P} N_B \Longrightarrow N_{B_n} \xrightarrow{llei} N_B$$

per concloure.

Referències

[1] Barnard, G. A. «Contribution to the discussion of Professor Bartlett's paper», *J. R. Statist. Soc. B* , 25:294, 1963.

[2] Corberán, A., Montes, F. «Perversiones y trampas de la probabilidad». *La Gaceta de la RSME*, pendent de publicació, 2000.

[3] Diggle, P. J. *Statistics Analysis for Spatial Point Patterns*. Academic Press, London, 1983.

[4] Feller, W. *An Introduction to Probability Theory ans Its Applications. Vol I.* 3a. edició. John Wiley, New York, 1968.

[5] Fienberg, S. E. «Randomization and Social Affairs: The 1970 Draft Lottery», *Science*, 171: 1970, 255–261.

[6] Gnedenko, B. *The Theory of Probability*. Mir, Moscow, 1978.

[7] Kingman, J. F. C. *Poisson Processes*. Oxford University Press, London, 1993.

[8] Numata, M. «Forest vegetation, particularly pine stands in the vicinity of Choshi-flora and vegetation at Choshi, Chiba Prefecture, VI» (en japonès). *Bull. Choshi Marine Lab.*, 6: 1964, 27–37.

[9] Pitman, J. *Probability*. Springer Verlag, New York, 1993.
[10] Rényi, A. *Probability Theory*. North-Holland, Amsterdam, 1970.

Francisco Montes Suay
Departament d'Estadística i Investigació Operativa
Universitat de València
E-46100 Burjassot, València
montes@uv.es

Jorge Mateu Mahiques
Departament de Matemàtiques
Universitat Jaume I
Campus Riu Sec
E-12071 Castelló
mateu@mat.uji.es