Butlletí de la Societat Catalana de Matemàtiques
Vol. 22, núm. 1, 2007. Pàg. 5–22 DOI: 10.2436/20.2002.01.1

Rècords: Quina és la probabilitat d'obtenir-ne? Quan apareixen? Quins valors prenen?*

Xavier Bardina

Resum Donada una successió $X_1, X_2, \ldots, X_n, \ldots$ de variables aleatòries independents i idènticament distribuïdes amb distribució contínua, direm que la variable X_i és un *rècord* si pren un valor més gran que totes les anteriors. En aquest treball calcularem quin és el nombre esperat de rècords en una sèrie de mida n i quina és la probabilitat d'obtenir-hi exactament r rècords. Veurem alguns resultats sobre el temps d'espera fins a obtenir el rècord r-èsim i calcularem la distribució del valor que pren aquest rècord.

Paraules clau: rècords, valors extrems, distribucions contínues.

Classificació MSC2000: 60G70.

Les situacions extremes sempre han estat observades amb gran interès. En manuscrits antics ja es troben textos sobre estius extraordinàriament freds, sobre inundacions, sobre temperatures molt elevades, etc.

Naturalment hom s'interessa també en els rècords d'altres disciplines. Per exemple, *El llibre Guinness dels rècords* (vegeu [5]) és molt popular a tot el món i sovint se'n treuen noves versions.

Des del punt de vista probabilístic els rècords s'han utilitzat com a model d'esdeveniments extrems en matemàtica financera (vegeu [3]). El llibre d'Arnold, Balakrishnan i Nagaraja (vegeu [2]) és el llibre de referència de la teoria probabilística dels rècords, però a [1] trobem una introducció d'aquesta teoria adreçada a un públic menys especialitzat.

El punt de vista probabilístic d'aquesta teoria permet demostrar alguns resultats curiosos. Per exemple, en aquest treball veurem que si suposem que la variable *cota màxima de neu que cau a l'hivern* s'ajusta al nostre model, de mitjana, un nen d'onze anys haurà vist en la seva vida tres rècords mentre que un avi de vuitanta-tres anys n'haurà viscut només cinc. O veurem també que, si tenim una sèrie infinita, s'assoleixen tots els rècords amb probabilitat 1, però el temps d'espera fins a obtenir el segon rècord pot ser tan gran que no existeix l'esperança d'aquesta variable.

* Lliçó inaugural de la llicenciatura de matemàtiques de la UAB impartida el 18 d'octubre de 2006.

Aquesta lliçó s'estructura de la manera següent: a la introducció definirem què entenem per un rècord, fixarem les hipòtesis necessàries per tractar aquest tipus de problemes i veurem alguns exemples. A la secció 2 calcularem el nombre esperat de rècords en una sèrie de longitud n. Veurem que quan augmentem la mida de la sèrie cada cop és més difícil obtenir nous rècords. A la secció 3 trobarem una fórmula recurrent que ens permetrà calcular la probabilitat d'obtenir exactament r rècords en una sèrie de mida n. A la secció 4, el càlcul de la variància del nombre de rècords ens permetrà trobar fites sobre algunes probabilitats interessants. A la secció 5 veurem alguns resultats sobre els instants en què es produeixin els rècords. Finalment, a la secció 6 calcularem la distribució de la variable que ens dóna el valor que pren el rècord r-èsim.

1 Introducció

Considerem una sèrie de nombres reals $X_1, X_2, \ldots, X_n$. El primer valor X_1 el considerarem un rècord. A partir d'aquí, X_i serà considerat un rècord si és el valor més gran obtingut fins a l'instant i. És a dir, X_i serà un rècord si

$$X_i > \max(X_1, X_2, \ldots, X_{i-1}) \quad \text{per a } i \geq 2.$$

Si $X_1, X_2, \ldots, X_n$ és una sèrie aleatòria podem preguntar-nos diferents qüestions:

- Quina és la probabilitat d'obtenir exactament r rècords en la sèrie?
- Quin és el nombre esperat de rècords?
- Quant cal esperar per obtenir un cert nombre de rècords?
- Quin valor prenen els rècords?

Abans de continuar fixarem algunes hipòtesis per poder tractar aquests problemes:

i) Suposarem que $X_1, X_2, \ldots, X_n$ són variables aleatòries independents i idènticament distribuïdes (v.a.i.i.d.).

ii) Suposarem que la distribució d'aquestes variables és contínua.

La primera hipòtesi exclou els rècords de la majoria de disciplines esportives on els individus s'esforcen sistemàticament per obtenir cada cop millors resultats, és a dir, van a la recerca del rècord. Hi ha models per estudiar aquest tipus de situacions però no els tractarem en aquesta lliçó.

La segona hipòtesi és simplement tècnica i és per excloure la possibilitat que un rècord pugui ser igualat; com que les variables aleatòries $X_1, X_2, \ldots, X_n$ són contínues sabem que la probabilitat que siguin diferents és igual a 1. De fet, aquest és el motiu que fa que el cas discret sigui molt més complicat de tractar i alguns dels resultats que veurem del cas continu encara són desconeguts en el cas discret.

Ens interessarem per la variable aleatòria:

$$R_n = \text{nombre de rècords de la sèrie } X_1, X_2, \ldots, X_n.$$

I si tenim una sèrie infinita $X_1, X_2, \ldots, X_n, \ldots$ per les variables aleatòries:

N_n = instant en què s'obté el rècord n-èsim.
$\Delta N_n = N_n - N_{n-1}$
(temps entre el rècord $(n-1)$-èsim i el rècord n-èsim).
V_n = valor que pren el rècord n-èsim.
$\Delta V_n = V_n - V_{n-1}$
(increment del valor entre el rècord n-èsim i el $(n-1)$-èsim).

Una observació important és que les variables R_n, N_n i ΔN_n no depenen de la distribució contínua concreta que tinguem, només de com estan ordenades les variables en la sèrie.

1.1 Rècords mínims

En aquesta lliçó estudiarem els rècords màxims, però l'estudi també serveix per als rècords mínims atès que els rècords mínims de la sèrie $X_1, X_2, \ldots, X_n$ són rècords màxims si considerem la sèrie $-X_1, -X_2, \ldots, -X_n$.

Un exemple d'aquest cas és el següent: podem considerar que les velocitats (a les que circularien si no hi hagués cap obstacle) dels vehicles que van pel carril addicional d'una autopista segueixen les hipòtesis amb les quals treballem. És a dir, aquestes velocitats $X_1, X_2, \ldots, X_n$ són v.a.i.i.d. i la seva distribució és contínua.

Passat un cert nombre de quilòmetres, quan la velocitat d'un vehicle és un rècord mínim crea una caravana i, ja que no pot avançar, aquesta caravana estarà formada per tots els cotxes fins que aparegui un altre rècord mínim de velocitat que crearà novament una altra caravana, etc.

Segons l'estudi que veurem, si el flux de vehicles que entren al carril addicional és important, cada cop les caravanes de vehicles tindran una mida més gran.

1.2 Exemple: temperatures màximes i mínimes el 17 d'octubre a Barcelona

Les taules 1 i 2 contenen les temperatures màximes i mínimes que s'han enregistrat a Barcelona el 17 d'octubre des de l'any 1900 fins al 2006. Les dades s'han obtingut consultant l'hemeroteca del diari *La Vanguardia*. Hi ha alguns valors perduts corresponents a anys entre el 1935 i el 1947. El motiu és que el diari no va publicar la informació meteorològica.

A la taula 1 observem que els rècords màxims d'aquesta sèrie (temperatures màximes més elevades) s'han obtingut els anys 1900, 1916, 1926 i 1981. Les

variables que hem introduït prendrien els valors següents:

$$R_{97} = 4$$
$$N_1 = 1 \quad N_2 = 17 \quad N_3 = 27 \quad N_4 = 72$$
$$\Delta N_2 = 16 \quad \Delta N_3 = 10 \quad \Delta N_4 = 45$$
$$V_1 = 24,3 \quad V_2 = 24,9 \quad V_3 = 25,2 \quad V_4 = 25,4$$
$$\Delta V_2 = 0,6 \quad \Delta V_3 = 0,3 \quad \Delta V_4 = 0,2$$

	0	1	2	3	4	5	6	7	8	9
190_	24,3	17,2	21,8	23,7	21,3	20,7	19,6	17,0	21,7	23,4
191_	22,8	22,4	20,3	23,1	17,2	20,5	24,9	21,5	17,5	19,0
192_	23,2	24,4	21,5	23,8	20,7	23,1	25,2	18,7	20,2	22,8
193_	24,0	22,8	23,6	25,0	18,4	∗	21,0	∗	∗	∗
194_	∗	20,6	∗	19,4	∗	∗	∗	∗	24,8	22,4
195_	22,8	17,7	22,4	21,2	22,6	19,8	22,9	24,5	17,6	19,9
196_	16,6	22,5	19,0	22,0	18,6	19,6	22,0	25,2	20,2	20,2
197_	21,1	21,0	19,5	22,0	17,4	23,7	20,6	21,0	22,5	21,6
198_	16,0	25,4	23,4	18,6	19,0	20,0	20,6	20,2	25,4	21,8
199_	23,3	22,3	16,3	21,3	21,2	21,5	20,1	23,5	23,5	22,9
200_	21,2	23,5	21,5	18,8	20,2	22,1	21,9			

TAULA 1: Temperatures màximes a Barcelona el 17 d'octubre. Les temperatures són les publicades al diari el 18 d'octubre. Els anys en què el 18 d'octubre va caure en dilluns i no es va publicar el diari s'han agafat les temperatures del dia següent. (FONT: Hemeroteca de *La Vanguardia*.)

	0	1	2	3	4	5	6	7	8	9
190_	15,0	13,0	9,8	14,6	13,5	12,2	10,4	8,5	13,8	15,0
191_	12,8	13,0	11,2	14,5	12,1	13,6	11,2	8,5	8,3	6,0
192_	15,5	18,8	14,5	16,1	14,5	12,2	18,0	16,2	13,6	15,6
193_	18,4	15,6	14,8	11,3	10,1	∗	15,8	∗	∗	∗
194_	∗	14,9	∗	13,5	∗	∗	∗	∗	17,7	17,2
195_	17,3	14,4	15,2	14,8	15,4	13,2	16,8	15,8	14,0	15,0
196_	10,2	15,1	15,4	14,1	12,4	15,6	13,6	17,0	15,5	15,3
197_	15,5	16,3	15,7	18,2	13,6	19,9	16,9	16,3	15,6	14,0
198_	12,0	16,4	14,6	11,0	13,6	14,0	13,2	12,8	20,6	15,8
199_	16,4	15,0	11,4	15,3	16,3	16,2	12,6	18,4	17,4	16,8
200_	13,8	17,8	18,3	12,1	12,9	16,4	17,5			

TAULA 2: Temperatures mínimes a Barcelona el 17 d'octubre. Les temperatures són les publicades al diari el 18 d'octubre. Els anys en què el 18 d'octubre va caure en dilluns i no es va publicar el diari s'han agafat les temperatures del dia següent. (FONT: Hemeroteca de *La Vanguardia*.)

Podem també buscar els rècords mínims d'aquesta sèrie. És a dir, les temperatures màximes més petites obtingudes el 17 d'octubre a Barcelona. En total hi ha cinc rècords mínims que corresponen a 1900, 1901, 1907, 1960 i 1980.

Observant la taula 2, veiem que hi ha sis rècords mínims corresponents als anys 1900, 1901, 1902, 1907, 1918 i 1919. També observem que hi ha cinc rècords màxims que corresponen als anys 1900, 1920, 1921, 1975 i 1988.

2 Nombre esperat de rècords

Sigui $X_1, X_2, \ldots, X_n$ una sèrie de v.a.i.i.d. amb distribució contínua. Començarem calculant la probabilitat que el valor X_n sigui un rècord.

Les variables aleatòries $X_1, X_2, \ldots, X_n$ són diferents amb probabilitat 1 i, per tant, hi ha $n!$ maneres diferents d'ordenar-les. El fet que aquestes variables siguin independents i idènticament distribuïdes implica que la seva llei conjunta és simètrica. Aquesta simetria ens assegura que totes les ordenacions són equiprobables. Si X_n és un rècord, la resta de variables es poden ordenar de $(n-1)!$ maneres diferents equiprobables. Per tant, si denotem amb p_n la probabilitat que X_n sigui un rècord,

$$p_n = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

Introduïm ara les variables aleatòries següents de tipus Bernoulli,

$$Y_i := \begin{cases} 1 & \text{si } X_i \text{ és un rècord} \\ 0 & \text{altrament} \end{cases}$$

per $i = 1, \ldots, n$.

Observem que $E(Y_i) = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i = \frac{1}{i}$.

De manera que el nombre total de rècords de la sèrie $X_1, X_2, \ldots, X_n$, que denotarem amb R_n, serà igual a

$$R_n = Y_1 + Y_2 + \cdots + Y_n$$

i, per tant,

$$E(R_n) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} \frac{1}{i}.$$

La suma de la sèrie harmònica $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ tendeix a infinit. Però la convergència d'aquesta sèrie és molt lenta: si un ordinador sumés 10^6 termes de la sèrie per segon, després de $3,17 \times 10^{85}$ anys hauria sumat fins al terme $\frac{1}{10^{99}}$ i el resultat que hauria obtingut seria

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{10^{99}} = 228,5331\ldots$$

És a dir, encara que matemàticament parlant el nombre de rècords que s'obté tendeix a infinit, la convergència és molt i molt lenta.

A la taula següent veiem, per a determinats valors de N, el valor de n per al qual la suma de la sèrie és per primer cop més gran o igual que N. És a dir, per a cada N el valor de n per al qual el nombre esperat de rècords, $E(R_n)$, supera N per primera vegada:

N	2	3	4	5	6	7	8	9	10	100
n	4	11	31	83	227	616	1674	4550	12367	$1,5 \times 10^{43}$

TAULA 3: Per a cada N la taula mostra el valor de n per al qual el nombre esperat de rècords supera N per primera vegada.

Suposem que X_i representa, per exemple, la cota màxima de neu caiguda durant l'hivern de l'any i-èsim de vida d'una persona; de mitjana, un nen d'onze anys hauria vist en la seva vida tres rècords, mentre que als trenta-un anys hauria viscut quatre rècords i la mitjana de rècords que hauria vist un avi de vuitanta-tres anys seria només de cinc. La pregunta natural és si això no tindrà alguna cosa a veure amb el fet que algunes persones pensin que abans nevava més, feia més calor, etc.

En l'exemple de les temperatures, que hem vist en la secció 1.2, teníem $n = 97$ i el nombre esperat de rècords és de $5,1571$.

3 La probabilitat de r rècords

Tot seguit veurem com podem calcular la probabilitat que la sèrie $X_1, X_2, \ldots, X_n$ tingui exactament r rècords, $p_{r,n}$.

La probabilitat que s'obtingui un únic rècord és

$$p_{1,n} = P(R_n = 1) = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

En efecte, de les $n!$ possibles ordenacions equiprobables de n variables, en $(n-1)!$ el màxim de les variables apareix en la primera posició.

D'altra banda també és fàcil calcular la probabilitat d'obtenir exactament n rècords en la sèrie $X_1, X_2, \ldots, X_n$:

$$p_{n,n} = P(R_n = n) = \frac{1}{n!},$$

atès que hi ha una única ordenació de les $n!$ possibles on apareixen totes les variables en ordre creixent.

Per a la resta de probabilitats demostrarem el resultat següent que ens permetrà calcular-les de manera recursiva.

1 TEOREMA *Si n i r són dos enters positius arbitraris tals que $r \leq n$ aleshores*

$$p_{r,n} = \frac{n-1}{n} p_{r,n-1} + \frac{1}{n} p_{r-1,n-1},$$

on $p_{1,1} = 1$ i, perquè la fórmula tingui sempre sentit, definim $p_{r,0} = 0$ per a tot r.

PROVA $A_{i,j}$ representarà l'esdeveniment que hi ha exactament i rècords en la sèrie $X_1, X_2, \ldots, X_j$, és a dir, $A_{i,j} = \{R_j = i\}$. Sigui B_n l'esdeveniment que representa el fet que la variable X_n és un rècord. Aleshores,

$$\begin{aligned} p_{r,n} &= P(R_n = r) \\ &= P(A_{r,n-1} \cap B_n^c) + P(A_{r-1,n-1} \cap B_n). \end{aligned}$$

Usarem

$$P(A_{r,n-1} \cap B_n^c) = P(A_{r,n-1} | B_n^c) P(B_n^c).$$

La probabilitat de l'esdeveniment $A_{r,n-1}$, és a dir, la probabilitat de r rècords en la sèrie de les $n-1$ primeres variables, només depèn de com estan ordenades entre si i no del fet que la variable n-èsima sigui o no un rècord. Per tant, aquests dos esdeveniments són independents i

$$P(A_{r,n-1} | B_n^c) = P(A_{r,n-1}) = p_{r,n-1}.$$

D'altra banda, sabem que $P(B_n) = p_n = \frac{1}{n}$, per tant,

$$P(B_n^c) = 1 - P(B_n) = \frac{n-1}{n}.$$

Obtenim, doncs,

$$P(A_{r,n-1} \cap B_n^c) = \frac{n-1}{n} p_{r,n-1}.$$

Utilitzant arguments del mateix tipus,

$$\begin{aligned} P(A_{r-1,n-1} \cap B_n) &= P(A_{r-1,n-1} | B_n) P(B_n) \\ &= P(A_{r-1,n-1}) P(B_n) \\ &= \frac{1}{n} p_{r-1,n-1}. \end{aligned}$$

□

A partir d'aquesta fórmula podem calcular una taula dels valors de $p_{r,n}$ per a valors petits de n:

		r				
		1	2	3	4	5
n	1	1				
	2	$\frac{1}{2}$	$\frac{1}{2}$			
	3	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{6}$		
	4	$\frac{1}{4}$	$\frac{11}{24}$	$\frac{1}{4}$	$\frac{1}{24}$	
	5	$\frac{1}{5}$	$\frac{5}{12}$	$\frac{7}{24}$	$\frac{1}{12}$	$\frac{1}{120}$

TAULA 4: Probabilitats d'obtenir r rècords en una sèrie de mida n.

En l'exemple que hem vist de les temperatures $n = 97$ i utilitzant el teorema 1 s'obtenen les probabilitats següents:

$$\begin{aligned} P(R_{97} = 3) &= 0,1281 \\ P(R_{97} = 4) &= 0,1950 \\ P(R_{97} = 5) &= 0,2117 \\ P(R_{97} = 6) &= 0,1756 \\ P(R_{97} = 7) &= 0,1164 \end{aligned}$$

4 La variància del nombre de rècords

Podem calcular també la variància del nombre total de rècords R_n.

Recordem que havíem introduït les variables

$$Y_i := \begin{cases} 1 & \text{si } X_i \text{ és un rècord} \\ 0 & \text{altrament} \end{cases}$$

per a $i = 1, \ldots, n$, de manera que $R_n = Y_1 + Y_2 + \cdots + Y_n$, i havíem vist que $E(Y_i) = \frac{1}{i}$. De la mateixa manera podem calcular la variància d'aquestes variables,

$$\begin{aligned} \operatorname{Var}(Y_i) &= E(Y_i^2) - \big(E(Y_i)\big)^2 \\ &= \frac{1}{i} - \frac{1}{i^2}. \end{aligned}$$

Necessitarem el resultat següent:

2 TEOREMA *Si $i \neq j$ aleshores les variables Y_i i Y_j són independents.*

PROVA Siguin $1 \le i < j \le n$. Aleshores,

$$\begin{aligned}
&P(Y_i = 1, Y_j = 1)\\
&= P\{X_i = \max(X_1, \ldots, X_i), X_j = \max(X_1, \ldots, X_j)\}\\
&= P\{X_i = \max(X_1, \ldots, X_i) < X_j = \max(X_{i+1}, \ldots, X_j)\}\\
&= P\Big\{(X_i = \max(X_1, \ldots, X_i)) \cap (X_j = \max(X_{i+1}, \ldots, X_j))\\
&\quad \cap (\max(X_1, \ldots, X_i) < \max(X_{i+1}, \ldots, X_j))\Big\}\\
&= P\{(X_i = \max(X_1, \ldots, X_i)) \cap (X_j = \max(X_{i+1}, \ldots, X_j))\}\\
&\quad \times P\Big\{\max(X_1, \ldots, X_i) < \max(X_{i+1}, \ldots, X_j) | (X_i = \max(X_1, \ldots, X_i))\\
&\quad \cap (X_j = \max(X_{i+1}, \ldots, X_j))\}.
\end{aligned}$$

Observem que els dos esdeveniments que intervenen en la primera probabilitat són independents entre si. Pel que fa a la segona probabilitat, la posició que ocupa el màxim de les primeres i variables i la posició que ocupa el màxim de les següents $j - i$ variables és independent del fet que el primer d'aquests màxims sigui menor que el segon. Per tant, la segona probabilitat és igual a

$$P(\max(X_1, \ldots, X_i) < \max(X_{i+1}, \ldots, X_j)) = \frac{(j-i)\cdot(j-1)!}{j!} = \frac{j-i}{j},$$

ja que aquest esdeveniment és equivalent al fet que el màxim de les variables $X_1, \ldots, X_j$ correspongui a una de les variables $X_{i+1}, \ldots, X_j$. Per comptar quantes de les $j!$ possibles ordenacions satisfan aquesta condició fixem el màxim, que pot ocupar qualsevulla de les $j - i$ darreres posicions, i la resta de variables es poden ordenar de $(j-1)!$ maneres diferents.

Així,

$$P(Y_i = 1, Y_j = 1) = \frac{1}{i} \cdot \frac{1}{j-i} \cdot \frac{j-i}{j} = \frac{1}{i} \cdot \frac{1}{j} = P(Y_i = 1)P(Y_j = 1).$$

Com es tracta de variables de tipus Bernoulli, el que hem vist fins ara ja demostra la independència. □

3 TEOREMA *La variància del nombre de rècords fins a l'instant n, R_n, és*

$$\operatorname{Var}(R_n) = \sum_{i=1}^{n} \frac{1}{i} - \sum_{i=1}^{n} \frac{1}{i^2}.$$

PROVA Com les variables $Y_1, \ldots, Y_n$ són dos a dos independents

$$\operatorname{Var}(R_n) = \operatorname{Var}(Y_1 + \cdots + Y_n) = \sum_{i=1}^{n} \operatorname{Var}(Y_i) = \sum_{i=1}^{n} \left(\frac{1}{i} - \frac{1}{i^2}\right).$$

El fet de conèixer el valor de $E(R_n)$ i $\mathrm{Var}(R_n)$ es pot utilitzar per trobar fites d'algunes probabilitats sobre la variable R_n, nombre de rècords fins a l'instant n. Per la desigualtat de Txebixev sabem que per a qualsevulla constant positiva c,

$$P(|R_n - E(R_n)| < c) \geq 1 - \frac{\mathrm{Var}(R_n)}{c^2}.$$

Amb l'ajut d'un manipulador algebraic podem comprovar per exemple que quan $n = 97$, com en l'exemple de les temperatures,

$$E(R_{97}) = \sum_{i=1}^{97} \frac{1}{i} = 5,157072426,$$

$$\mathrm{Var}(R_{97}) = \sum_{i=1}^{97} \left(\frac{1}{i} - \frac{1}{i^2}\right) = 3,522394679.$$

Amb la constant $c = 3,84$ veiem, per exemple, que la probabilitat que R_{97} estigui entre 2 i 8 (ambdós inclosos) és més gran que 0,7611.

Com acostuma a passar amb la desigualtat de Txebixev, els valors exactes d'aquestes probabilitats són molt més grans que les fites que obtenim. Per exemple, per a $n = 97$ hem obtingut $P(2 \leq R_{97} \leq 8) \geq 0,7611$ mentre que el càlcul exacte d'aquesta probabilitat, utilitzant el teorema 1, és de 0,9434.

Com la variable aleatòria R_n és positiva, la desigualtat de Txebixev també ens dóna que per a qualsevulla constant positiva c,

$$P(R_n \geq c) \leq \frac{E(R_n^2)}{c^2},$$

i així, per exemple, $P(R_{97} \geq 10) \leq 0,3012$.

Podem millorar força aquesta fita si introduïm l'extensió de la desigualtat de Txebixev següent (vegeu [4]):

4 Teorema *Sigui X una variable aleatòria amb esperança $E(X)$ i variància $Var(X)$. Aleshores, per a tota constant c tal que $c > E(X)$,*

$$P(X \geq c) \leq \frac{Var(X)}{Var(X) + (c - E(X))^2}.$$

Aplicant aquest resultat trobem

$$P(R_{97} \geq 10) \leq 0,1306.$$

Tot i que hem rebaixat força la fita superior d'aquesta probabilitat encara estem lluny del valor exacte, que és 0,0171.

Prova (Teorema 4) Comencem demostrant que si tenim una funció g tal que

- $g(x) \geq 0$ per a tot x,
- $g(x) \geq a > 0$ per a tot $x \in I$,

on I és un interval, aleshores per a tota variable aleatòria Y,

$$P(Y \in I) \leq \frac{E(g(Y))}{a}.$$

En efecte, la demostració per al cas d'una variable aleatòria com la nostra, és a dir, una variable discreta que pren valors sobre els naturals, seria:

$$\begin{aligned} E(g(X)) &= \sum_{j=1}^{+\infty} g(j) \cdot P(Y = j) \\ &\geq \sum_{j \in \mathbb{N} \cap I} g(j) \cdot P(Y = j) \\ &\geq a \cdot P(Y \in I), \end{aligned}$$

i, per tant,

$$P(Y \in I) \leq \frac{E(g(Y))}{a}.$$

Si prenem com a funció g,

$$g(x) = (x + k)^2,$$

on k és una constant positiva, tenim $g(x) \geq 0$ per a tota x i $g(x) \geq (c + k)^2$ per a tota $x \geq c > 0$, per tant,

$$P(Y \geq c) \leq \frac{E((Y + k)^2)}{(c + k)^2}.$$

Suposem ara que la variable aleatòria Y satisfà

$$\begin{aligned} E(Y) &= 0, \\ E(Y^2) &= \sigma^2. \end{aligned}$$

Aleshores,

$$P(Y \geq c) \leq \frac{\sigma^2 + k^2}{(c + k)^2}.$$

Si busquem el mínim de la funció

$$h(k) = \frac{\sigma^2 + k^2}{(c + k)^2}$$

veurem que aquest mínim s'assoleix en el punt $k = \frac{\sigma^2}{c}$. Així,

$$P(Y \geq c) \leq \frac{\sigma^2 + \left(\frac{\sigma^2}{c}\right)^2}{\left(c + \frac{\sigma^2}{c}\right)^2} = \frac{\sigma^2}{c^2 + \sigma^2}.$$

Hem vist que per a tota constant $c > 0$, si Y és una variable aleatòria tal que $E(Y) = 0$ i $E(Y^2) = \sigma^2$ aleshores

$$P(Y \geq c) \leq \frac{\sigma^2}{\sigma^2 + c^2}.$$

Si X és una variable qualsevulla, tal que existeix la seva esperança i la seva variància, aleshores per a tota constant $c > E(X)$,

$$\begin{aligned} P(X \geq c) &= P(X - E(X) \geq c - E(X)) \\ &\leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + (c - E(X))^2}. \end{aligned}$$

□

5 Quan s'obtenen els rècords?

Ara introduïm la variable N_r que ens indicarà en quin instant s'ha produït el rècord r-èsim.

Per definició, $N_1 = 1$, ja que la primera observació és un rècord.

En aquesta secció haurem de suposar que la sèrie $X_1, X_2, \ldots$ és infinita.

Comencem considerant el problema de quan s'obté el segon rècord. Si s'obté en un temps m ($m \geq 2$), això significa que el valor de X_m és el més gran dels valors de $X_1, X_2, \ldots, X_m$ i que el valor de X_1 és el segon més gran després del de X_m.

Si tenim m variables poden ordenar-se de $m!$ maneres diferents, i en $(m-2)!$ d'aquestes ordenacions la variable més gran ocupa el darrer lloc i la segona més gran la primera posició, per tant,

$$P(N_2 = m) = \frac{(m-2)!}{m!} = \frac{1}{m(m-1)}.$$

Observem que

$$\sum_{m=2}^{+\infty} P(N_2 = m) = \sum_{m=2}^{+\infty} \left(\frac{1}{m(m-1)}\right) = \sum_{m=2}^{+\infty} \left(\frac{1}{m-1} - \frac{1}{m}\right) = 1,$$

ja que es tracta d'una sèrie telescòpica.

Això significa que el segon rècord apareix en un temps finit amb probabilitat 1, però malgrat això,

$$E(N_2) = \sum_{m=2}^{+\infty} m \cdot P(N_2 = m) = \sum_{m=2}^{+\infty} \frac{1}{m-1} = +\infty.$$

És a dir, encara que el segon rècord apareix en un temps finit, el temps d'espera fins que s'obté pot ser extremament llarg i de fet no existeix l'esperança de la variable temps d'espera fins a obtenir el segon rècord.

Pel que fa al temps d'espera fins a obtenir el rècord r-èsim, tenim el resultat següent:

5 TEOREMA *Si $r \geq 2$ i $m \geq r$ aleshores*

$$P(N_r = m) = \frac{p_{r-1,m-1}}{m}.$$

PROVA En efecte,

$$\begin{aligned} & P(N_r = m) \\ = {} & P(\{\text{les variables } X_1, \ldots, X_{m-1} \text{ contenen } r-1 \text{ rècords}\} \\ & \cap \{\text{la variable } X_m \text{ és un rècord}\}). \end{aligned}$$

Observem que aquests dos esdeveniments són independents ja que el fet que la variable X_m sigui més gran que les $m-1$ variables anteriors no ens diu res respecte al nombre de rècords que hi ha entre aquestes, ja que això només depèn de com estan ordenades aquestes $m-1$ variables. Per tant, la probabilitat que calculem és igual a

$$\begin{aligned} & P(\{\text{les variables } X_1, \ldots, X_{m-1} \text{ contenen } r-1 \text{ rècords}\}) \\ & \quad \times P(\{\text{la variable } X_m \text{ és un rècord}\}) \\ = {} & p_{r-1,m-1} \frac{1}{m}. \end{aligned}$$

□

Com $N_r \geq N_2$ per a $r \geq 2$, es té

$$E(N_r) \geq E(N_2) = +\infty,$$

en canvi, utilitzant el teorema 1 veiem que per a tot $r \geq 2$,

$$\begin{aligned} \sum_{m=r}^{+\infty} P(N_r = m) &= \sum_{m=r}^{+\infty} \frac{p_{r-1,m-1}}{m} \\ &= \sum_{m=r}^{+\infty} \left(p_{r,m} - \frac{m-1}{m} p_{r,m-1} \right) \\ &= \sum_{m=r}^{+\infty} \frac{p_{r,m}}{m+1} = \sum_{m'=r+1}^{+\infty} \frac{p_{r,m'-1}}{m'} \\ &= \sum_{m'=r+1}^{+\infty} P(N_{r+1} = m'), \end{aligned}$$

on hem utilitzat $p_{r,r-1} = 0$ (per definició) i que teníem una sèrie telescòpica. Com coneixem el valor del sumatori per al cas $r = 2$ es té que per a tot $r \geq 2$,

$$\sum_{m=r}^{+\infty} P(N_r = m) = \sum_{m=2}^{+\infty} P(N_2 = m) = 1.$$

Per estudiar el comportament de la variable N_r no podem utilitzar les esperances (com hem fet amb la variable R_n) perquè hem vist que no existeixen. Caldrà, doncs, estudiar-ho d'una altra manera.

Concretament, buscarem el nombre enter m_r tal que la suma

$$P(N_r = r) + P(N_r = r+1) + \cdots + P(N_r = m_r)$$

sigui per primera vegada igual o més gran que $\frac{1}{2}$.

Aquest nombre m_r correspon aproximadament a la idea de la mediana de la distribució de l'instant en què s'obté el rècord r-èsim.

Els resultats que s'obtenen són:

r	2	3	4	5	6	7	8	9
m_r	2	7	20	57	152	424	1166	3200

TAULA 5: Per a cada r la taula mostra el valor m_r que correspon al primer enter per al qual la probabilitat que s'hagin obtingut almenys r rècords en una sèrie de mida m_r supera $\frac{1}{2}$.

De nou veiem que els rècords apareixen en les sèries cada cop de manera menys i menys freqüent.

Per al cas de l'instant en què s'obté el segon rècord, podem demostrar un altre resultat curiós,

$$\begin{aligned} P(N_2 > n) &= \sum_{i=n+1}^{+\infty} P(N_2 = i) \\ &= \sum_{i=n+1}^{+\infty} \frac{1}{(i-1)i} \\ &= \sum_{i=n+1}^{+\infty} \left(\frac{1}{i-1} - \frac{1}{i}\right) \\ &= \frac{1}{n}. \end{aligned}$$

Recordem, finalment, que havíem definit també el temps entre dos rècords $\Delta N_n = N_n - N_{n-1}$.

Un altre resultat interessant (vegeu [2]) és la distribució asimptòtica del quocient $\frac{\Delta N_n}{N_n}$. Es pot veure que quan $n \to \infty$,

$$\frac{\Delta N_n}{N_n} \xrightarrow{d} \text{Unif}(0,1).$$

Això vol dir que per a n gran, la distribució de

$$\frac{\Delta N_n}{N_n} \approx \text{Unif}(0,1)$$

i, per tant, si n és gran

$$E\left(\frac{\Delta N_n}{N_n}\right) \approx \frac{1}{2}.$$

És a dir, de mitjana aproximadament la meitat del temps d'espera fins a obtenir el rècord n-èsim l'hem passat entre el rècord $(n-1)$-èsim i el rècord n-èsim.

6 Quin valor prenen els rècords?

En aquesta secció ens interessem pel valor que prenen els rècords. És a dir, volem conèixer quina és la distribució de la variable:

$$V_n = X_{N_n} = \text{ valor que pren el rècord } n\text{-èsim.}$$

Per fer-ho suposarem que la nostra sèrie $X_1, X_2, \ldots, X_n, \ldots$ està formada per v.a.i.i.d. amb funció de distribució F.

Amb el que hem vist en el darrer apartat podem calcular directament la funció de distribució de la variable V_2:

$$\begin{aligned} F_{V_2}(t) &= P(V_2 \le t) \\ &= P(X_{N_2} \le t) \\ &= \sum_{k=2}^{+\infty} P(\{X_{N_2} \le t\} \cap \{N_2 = k\}) \\ &= \sum_{k=2}^{+\infty} P(\{\max\{X_1, \ldots, X_k\} \le t\} \cap \{N_2 = k\}) \\ &= \sum_{k=2}^{+\infty} P(\max\{X_1, \ldots, X_k\} \le t) P(N_2 = k). \end{aligned}$$

Observem que

$$\begin{aligned} P(\max\{X_1, \ldots, X_k\} \le t) &= P(X_1 \le t, X_2 \le t, \ldots, X_k \le t) \\ &= [P(X_1 \le t)]^k \\ &= (F(t))^k. \end{aligned}$$

Per tant,

$$F_{V_2}(t) = \sum_{k=2}^{+\infty} \frac{1}{k(k-1)} (F(t))^k.$$

Si $F(t) = 1$ la suma d'aquesta sèrie hem vist que és 1. Podem, doncs, suposar que $0 \le F(t) < 1$. Utilitzant que per $0 < x \le 2$ el desenvolupament de la sèrie de Taylor del $\ln(x)$ és convergent i

$$\ln(x) = -\sum_{k=1}^{+\infty} \frac{(1-x)^k}{k}$$

i que $\frac{1}{k(k-1)} = \frac{1}{k-1} - \frac{1}{k}$ tenim que per $-1 \le x < 1$

$$\sum_{k=2}^{+\infty} \frac{x^k}{k(k-1)} = x + (1-x)\ln(1-x)$$

i, per tant, com que la funció de distribució pren valors entre 0 i 1,

$$F_{V_2}(t) = F(t) + (1 - F(t)) \ln(1 - F(t)).$$

Aquest mètode però, no es pot generalitzar per trobar la distribució de V_n per a qualsevol n, atès que el teorema 5 ens permetia calcular $P(N_n = m)$ però en funció d'unes probabilitats que s'havien de calcular de manera recurrent.

Veurem, però, una manera indirecta de calcular la distribució de V_n per a qualsevol n.

Considerem el cas particular en què les v.a. $X_1^*, X_2^*, \ldots, X_n^*, \ldots$ són exponencials de paràmetre 1. Totes les variables que apareguin en aquest cas particular les escriurem amb un asterisc per distingir-les del cas general que estudiem.

Utilitzant la propietat de falta de memòria de la distribució exponencial es pot demostrar que els increments entre rècords també són v.a.i.i.d. amb distribució exponencial de valor mitjà 1.

Es té, doncs,

$$V_1^* = X_1^*, \Delta V_2^* = V_2^* - V_1^*, \Delta V_3^* = V_3^* - V_2^*, \ldots$$

són v.a.i.i.d. amb distribució exponencial de valor mitjà 1.

Aleshores,

$$V_n^* = V_n^* - V_{n-1}^* + V_{n-1}^* - V_{n-2}^* + \cdots + V_2^* - V_1^* + V_1^*$$

és la suma de n v.a.i.i.d. amb distribució exponencial de valor mitjà 1 i, per tant,

$$V_n^* \sim \text{Gamma}(n, 1), \quad \text{per a } n = 1, 2, \ldots$$

Utilitzarem aquest resultat per trobar la distribució del valor del rècord n-èsim, V_n, per a qualsevol successió de v.a. $X_1, X_2, \ldots, X_n, \ldots$ amb funció de distribució contínua F.

Si X és una v.a. amb funció de distribució contínua F, aleshores la v.a.

$$Y := -\ln(1 - F(X))$$

té distribució exponencial de valor mitjà 1.

Farem la demostració del cas en què existeix la inversa de la funció de distribució F. Observem que

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(-\ln(1 - F(X)) \le y) \\ &= P(F(X) \le 1 - e^{-y}) \\ &= P\left(X \le F^{-1}(1 - e^{-y})\right) \\ &= F(F^{-1}(1 - e^{-y})) \\ &= 1 - e^{-y}, \end{aligned}$$

que és la funció de distribució exponencial de valor mitjà 1.

Quan no existeix la inversa, la pseudoinversa de F definida com

$$F^{-1}(u) := \inf\{x : F(x) \geq u\},$$

fa el paper de la inversa.

Així, doncs, tenim la igualtat en llei següent,

$$X \stackrel{d}{=} F^{-1}(1 - e^{-X^*}),$$

on $X^* \sim \text{Exp}(1)$.

Com que X és una funció monòtona creixent de X^*, els rècords de la sèrie $X_1, X_2, \ldots, X_n, \ldots$ es poden expressar en funció dels rècords de la sèrie $X_1^*, X_2^*, \ldots, X_n^*, \ldots$

Concretament tindrem

$$V_n \stackrel{d}{=} F^{-1}(1 - e^{-V_n^*}), \quad \text{per a } n = 1, 2, \ldots$$

Tornem al cas exponencial. Integrant per parts s'obté que la funció de supervivència de V_n^* (que recordem que segueix una distribució Gamma$(n, 1)$) és:

$$P(V_n^* > t) = e^{-t} \sum_{k=0}^{n-1} \frac{t^k}{k!}, \quad t > 0.$$

D'on podrem deduir la funció de supervivència per al cas general que estudiem:

$$\begin{aligned} P(V_n > t) &= P\left(F^{-1}(1 - e^{-V_n^*}) > t\right) \\ &= P(1 - e^{-V_n^*} > F(t)) \\ &= P\left(V_n^* > -\ln(1 - F(t))\right) \\ &= (1 - F(t)) \sum_{k=0}^{n-1} \frac{(-\ln(1 - F(t)))^k}{k!}. \end{aligned}$$

Observem que en el segon pas, per poder treure logaritmes, necessitem que $F(t) < 1$. Si $F(t) = 1$, obtenim $P(V_n > t) = 0$.

Referències

[1] ANDEL, J. *Mathematics of Chance*. John Wiley & Sons, Inc., 2001. Wiley Series in Probability and Statistics.

[2] ARNOLD, B. C.; BALAKRISHNAN, N.; NAGARAJA, H. N. *Records*. John Wiley & Sons, Inc., 1998. Wiley Series in Probability and Statistics.

[3] EMBRECHTS, P.; KLÜPPELBERG, C.; MIKOSCH, T. *Modelling extremal events for insurance and finance*. Springer, 1997.

[4] Feller, W. *Introducción a la teoría de las probabilidades y sus aplicaciones*. Limusa-Wiley, 1973.

[5] *Llibre Guinness dels rècords*. Versió electrònica: http://www.guinnessworldrecords.com

Departament de Matemàtiques
Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona
Xavier.Bardina@uab.cat