Combinatorics, Probability and Computing (2021), **30**, pp. 105–123
doi:10.1017/S0963548320000309

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Breaking bivariate records

James Allen Fill[†]

Department of Applied Mathematics and Statistics, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
Email: jimfill@jhu.edu

(Received 24 January 2019; revised 1 June 2020; first published online 18 August 2020)

Abstract

We establish a fundamental property of bivariate Pareto records for independent observations uniformly distributed in the unit square. We prove that the asymptotic conditional distribution of the number of records broken by an observation given that the observation sets a record is Geometric with parameter 1/2.

2020 MSC Codes: Primary: 60D05; Secondary: 60F05, 60F15, 60G17

1. Introduction and main result

This paper proves an interesting phenomenon concerning the breaking of bivariate records first observed empirically by Daniel Q. Naiman, whom we thank for an introduction to the problem considered. We begin with some relevant definitions, taken (with trivial changes) from [4] and [3]. Although our attention in this paper will be focused on dimension $d = 2$ (see [3, Conjecture 2.3] for general d) and the approach we utilize seems to be limited to the bivariate case, we begin by giving definitions that apply for general dimension d.

Let $\mathbf{1}(E) = 1$ or 0 according as E is true or false. We write ln or L for natural logarithm, lg for binary logarithm, and log when the base does not matter. For d-dimensional vectors $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$, write $x \prec y$ to mean that $x_j < y_j$ for $j = 1, \ldots, d$. The notation $x \succ y$ means $y \prec x$.

Like Bai, Devroye, Hwang and Tsai [2], we find it more convenient (in particular, expressions encountered in their computations and ours are simpler) to consider (equivalently) record-*small*, rather than record-large, values. Let $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots$ be i.i.d. (independent and identically distributed) copies of a random vector $\mathbf{X}$ with independent coordinates, each uniformly distributed over the unit interval.

Definition 1.1.

(a) We say that $\mathbf{X}^{(n)}$ is a *Pareto record* (or simply *record*, or that $\mathbf{X}^{(n)}$ *sets* a record at time n) if $\mathbf{X}^{(n)} \not\succ \mathbf{X}^{(i)}$ for all $1 \leqslant i < n$.
(b) If $1 \leqslant j \leqslant n$, we say that $\mathbf{X}^{(j)}$ is a *current record* (or *remaining record*, or *minimum*) at time n if $\mathbf{X}^{(j)} \not\succ \mathbf{X}^{(i)}$ for all $i \in [n]$.

[†]Research supported by the Acheson J. Duncan Fund for the Advancement of Research in Statistics.

© The Author(s), 2020. Published by Cambridge University Press.

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

(c) If $0 \leqslant k \leqslant n$, we say that $\mathbf{X}^{(n)}$ *breaks* (or *kills*) k records if $X^{(n)}$ sets a record and there exist precisely k values j with $1 \leqslant j < n$ such that $\mathbf{X}^{(j)}$ is a current record at time $n-1$ but is not a current record at time n.

For $n \geqslant 1$ (or $n \geqslant 0$, with the obvious conventions), let R_n denote the number of records $\mathbf{X}^{(k)}$ with $1 \leqslant k \leqslant n$, and let r_n denote the number of remaining records at time n.

Here is the main result of this paper.

Theorem 1.1. *Suppose that independent bivariate observations, each uniformly distributed in* $(0,1)^2$*, arrive at times* $1, 2, \ldots$*. Let* $K_n = -1$ *if the nth observation is not a new record, and otherwise let* K_n *denote the number of remaining records killed by the nth observation. Then* K_n*, conditionally given* $K_n \geqslant 0$*, converges in distribution to* $G - 1$*, where* $G \sim \text{Geometric}(1/2)$*, as* $n \to \infty$*.*

Equivalently, the conclusion (with asymptotics throughout referring to $n \to \infty$) is that

$$\mathbb{P}\,(K_n = k \mid K_n \geqslant 0) \to 2^{-(k+1)} \quad \text{for each (fixed) integer } k \geqslant 0. \tag{1.1}$$

Here is an outline of the proof. In Section 2 we provide a simple and short proof of the well-known result that

$$\mathbb{P}\,(K_n \geqslant 0) = n^{-1} H_n, \quad n \geqslant 1,$$

where $H_n = \sum_{i=1}^{n} i^{-1}$ denotes the nth harmonic number. In Section 3 (see Theorem 3.1) we show that

$$|\,\mathbb{P}\,(K_n = k) - [2^{-(k+1)} n^{-1} H_n - (k-1)2^{-(k+2)} n^{-1}]| \leqslant \frac{1}{2} n^{-2} \tag{1.2}$$

for all $n \geqslant 1$ and all $k \geqslant 0$. The improvement

$$|\,\mathbb{P}\,(K_n = k \mid K_n \geqslant 0) - [2^{-(k+1)} + \alpha_{n,k}]| \leqslant \frac{1}{2} n^{-1} H_n^{-1} \tag{1.3}$$

to (1.1) then follows immediately, where $\alpha_{n,k}$ is a first-order correction term with

$$\alpha_{n,k} := -(k-1)2^{-(k+2)} H_n^{-1}$$

to the Geometric(1/2) probability mass function (PMF) $2^{-(k+1)}$. This improvement shows that approximation of the conditional PMF in Theorem 1.1 by the uncorrected Geometric(1/2) PMF has (for large n) vanishingly small relative error, not just for fixed k but for $k \equiv k_n = o(\log n)$. It also shows that the corrected approximation has small relative error for $k \leqslant \lg n + \lg\lg n - \omega(1)$. Of course we always have $K_n \leqslant r_{n-1}$, and by [4, Remark 4.3(b)] we have $r_n = O(\log n)$ almost surely; the corrected approximation thus gives small relative error for rather large values of k indeed.

As one might expect, the correction terms sum to 0. We observe that the correction is positive (and of largest magnitude in absolute-error terms) when $k = 0$, vanishes when $k = 1$, and is negative (and of non-increasing magnitude) when $k \geqslant 2$.

Formulation of Theorem 1.1 was motivated by [3, Table 1], reproduced here as Table 1. Table 1 tabulates, for the first 100 000 records generated in a single trial, the number of records that break k remaining records, for each value of k. The Geometric(1/2) pattern is striking. The precise relationship between Theorem 1.1 and the phenomenon observed in Table 1 is discussed in Section 4, where a main conjecture is stated and a possible plan for completing its proof is described.

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

Table 1. Results of a simulation experiment in which $M = 100\,000$ bivariate records are generated, and for each new record the number k of records it breaks is recorded. The number of records that break k current records is denoted by N_k, and $\tilde{p}_{M,k} = N_k/M$ is the proportion of the 100 000 records that break k records

k	N_k	$\tilde{p}_k$
0	50 334	0.50334
1	24 667	0.24667
2	12 507	0.12507
3	6 335	0.06335
4	3 040	0.03040
5	1 571	0.01571
6	782	0.00782
7	364	0.00364
8	202	0.00202
9	94	0.00094
10	48	0.00048
11	24	0.00024
12	18	0.00018
13	8	0.00008
14	4	0.00004
16	1	0.00001
17	0	0.00000
18	1	0.00001

Throughout, we denote the nth observation $\mathbf{X}^{(n)}$ simply by $\mathbf{X} = (X, Y)$ (note: *sub*scripted $\mathbf{X}$ will have a different later use) and, for any Borel subset S of $(0, 1)^2$, we denote the number of the first n observations falling in S by $N_n(S)$.

2. The probability that $K_n \geqslant 0$

In this section we compute the probability $\mathbb{P}(K_n \geqslant 0)$ (that the nth observation is a record) exactly. This result is already well known (see *e.g.* [2] or [4, (4.5), the sentence immediately preceding Definition 1.2, and the simple fact that $\mathbb{E}\, R_n = H_n$ in dimension 1]), but we give a proof for completeness.

Proposition 2.1. *For $n \geqslant 1$ we have*

$$\mathbb{P}(K_n \geqslant 0) = n^{-1} H_n.$$

Proof. Since $\{K_n \geqslant 0\} = \{\mathbf{X}^{(n)} \text{ is a record}\}$ is the event that none of the $n-1$ observations $\mathbf{X}^{(i)}$ with $1 \leqslant i < n$ lies to the southwest of $\mathbf{X}^{(n)}$, we have

$$\begin{aligned}\mathbb{P}(K_n \geqslant 0, \mathbf{X} \in \mathrm{d}\mathbf{x}) &= \mathbb{P}(N_{n-1}((0,x)\times(0,y)) = 0, \mathbf{X} \in \mathrm{d}\mathbf{x}) \\ &= \mathbb{P}(N_{n-1}((0,x)\times(0,y)) = 0)\ \mathbb{P}(\mathbf{X} \in \mathrm{d}\mathbf{x}) \\ &= (1-xy)^{n-1}\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

Integrating (and recalling the sum of a truncated geometric series), we therefore have

$$\begin{aligned}\mathbb{P}(K_n \geqslant 0) &= \int_{x=0}^{1}\int_{y=0}^{1}(1-xy)^{n-1}\,\mathrm{d}y\,\mathrm{d}x \\ &= n^{-1}\int_{x=0}^{1}x^{-1}[1-(1-x)^n]\,\mathrm{d}x \\ &= n^{-1}\sum_{j=0}^{n-1}\int_{x=0}^{1}(1-x)^j\,\mathrm{d}x \\ &= n^{-1}H_n,\end{aligned}$$

as claimed. □

3. The probability that $K_n = k$

In this section we compute $\mathbb{P}(K_n = k)$ for $k \geqslant 0$ exactly and produce the approximation (3.7) with its stated error bound.

3.1 The exact probability

Over the event $\{K_n = k\}$ (with $k \geqslant 0$), denote those remaining records at time $n-1$ broken by $\mathbf{X}$, in order from southeast to northwest (*i.e.* in decreasing order of first coordinate and increasing order of second coordinate) by $\mathbf{X}_1 = (X_1, Y_1), \ldots, \mathbf{X}_k = (X_k, Y_k)$. Note that if we read *all* the remaining records in order from southeast to northwest, then $\mathbf{X}_1, \ldots, \mathbf{X}_k$ appear consecutively.

If there are any remaining records at time $n-1$ with second coordinate smaller than Y, choose the largest such second coordinate Y_0 and denote the corresponding remaining record by $\mathbf{X}_0 = (X_0, Y_0)$ (and note that then $\mathbf{X}_0, \ldots, \mathbf{X}_k$ appear consecutively); otherwise, set $\mathbf{X}_0 = (X_0, Y_0) = \mathbf{e}_1 := (1, 0)$.

Similarly, if there are any remaining records at time $n-1$ with first coordinate smaller than X, choose the largest such first coordinate X_{k+1} and denote the corresponding remaining record by $\mathbf{X}_{k+1} = (X_{k+1}, Y_{k+1})$ (and note that then $\mathbf{X}_1, \ldots, \mathbf{X}_{k+1}$ appear consecutively); otherwise, set $\mathbf{X}_{k+1} = (X_{k+1}, Y_{k+1}) = \mathbf{e}_2 := (0, 1)$.

Observe that, (almost surely) over the event $\{K_n = k\}$, we have $X_k > X > X_{k+1}$ and $Y_1 > Y > Y_0$. In results that follow we will only need to treat three cases: (i) $\mathbf{X}_0 \neq \mathbf{e}_1$ and $\mathbf{X}_{k+1} \neq \mathbf{e}_2$, (ii) $\mathbf{X}_0 = \mathbf{e}_1$ and $\mathbf{X}_{k+1} \neq \mathbf{e}_2$, and (iii) $\mathbf{X}_0 = \mathbf{e}_1$ and $\mathbf{X}_{k+1} = \mathbf{e}_2$. The fourth case $\mathbf{X}_0 \neq \mathbf{e}_1$ and $\mathbf{X}_{k+1} = \mathbf{e}_2$ can be handled by symmetry with respect to the second case.

Our first result of this section specifies the exact joint distribution of $\mathbf{X}, \mathbf{X}_0, \ldots \mathbf{X}_{k+1}$. We write $n^{\underline{k}}$ for the falling factorial power

$$n(n-1)\cdots(n-k+1) = k!\binom{n}{k},$$

and we introduce the abbreviations

$$\sum_{j}^{k} := \sum_{i=j}^{k}(x_{i-1} - x_i)y_i, \quad \sum^{k} := \sum_{1}^{k}$$

for sums that will appear frequently below.

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

Proposition 3.1.

(i) *For* $n \geqslant k+3$ *and*

$$1 > x_0 > \cdots > x_k > x > x_{k+1} > 0 \quad \textit{and} \quad 0 < y_0 < y < y_1 < \cdots < y_{k+1} < 1,$$

we have

$$\mathbb{P}\,(K_n = k; \mathbf{X} \in \mathrm{d}\mathbf{x}; \mathbf{X}_i \in \mathrm{d}\mathbf{x}_i \ \textit{for}\ i = 0, \ldots, k+1)$$
$$= (n-1)^{\underline{k+2}} \left[1 - \left\{\sum^{k} + x_k y_{k+1}\right\}\right]^{n-(k+3)} \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{x}_0 \cdots \mathrm{d}\mathbf{x}_{k+1}.$$

(ii) *For* $n \geqslant k+2$ *and*

$$1 > x_1 > \cdots > x_k > x > x_{k+1} > 0 \quad \textit{and} \quad 0 < y < y_1 < \cdots < y_{k+1} < 1,$$

we have

$$\mathbb{P}\,(K_n = k; \mathbf{X} \in \mathrm{d}\mathbf{x}; \mathbf{X}_0 = \mathbf{e}_1; \mathbf{X}_i \in \mathrm{d}\mathbf{x}_i \ \textit{for}\ i = 1, \ldots, k+1)$$
$$= (n-1)^{\underline{k+1}} \left[1 - \left\{\sum^{k} + x_k y_{k+1}\right\}\right]^{n-(k+2)} \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_{k+1},$$

where $x_0 = 1$.

(iii) *For* $n \geqslant k+1$ *and*

$$1 > x_1 > \cdots > x_k > x > 0 \quad \textit{and} \quad 0 < y < y_1 < \cdots < y_k < 1,$$

we have

$$\mathbb{P}\,(K_n = k; \mathbf{X} \in \mathrm{d}\mathbf{x}; \mathbf{X}_0 = \mathbf{e}_1; \mathbf{X}_i \in \mathrm{d}\mathbf{x}_i \ \textit{for}\ i = 1, \ldots, k; \mathbf{X}_{k+1} = \mathbf{e}_2)$$
$$= (n-1)^{\underline{k}} \left[1 - \left\{\sum^{k} + x_k\right\}\right]^{n-(k+1)} \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_k,$$

where $x_0 = 1$.

Proof. We present only the proof of (i); the proofs of (ii) and (iii) are similar. We shall be slightly informal in regard to 'differentials' in our presentation. The key is that the event in question (almost surely) equals the event

$$\{N_{n-1}(\mathrm{d}\mathbf{x}_i) = 1 \text{ for } i = 0, \ldots, k+1; N_{n-1}(S) = 0; \mathbf{X} \in \mathrm{d}\mathbf{x}\}, \tag{3.1}$$

where S is the following disjoint union of rectangular regions:

$$S = \cup_{i=1}^{k} [(x_i, x_{i-1}) \times (0, y_i)] \cup [(0, x_k) \times (0, y_{k+1})].$$

See Figure 1. But the probability of the event (3.1) is

$$(n-1)^{\underline{k+2}} \left[\prod_{i=0}^{k+1} \mathrm{d}\mathbf{x}_i\right] \times [1 - \lambda(S)]^{n-(k+3)} \times \mathrm{d}\mathbf{x},$$

where $\lambda(S)$ is the Lebesgue measure (*i.e.* area) of S, and the displayed probability reduces easily to the claimed result. □

Remark 3.1. When $k = 0$, Proposition 3.1 is naturally and correctly interpreted as follows.

(i) For $n \geqslant 3$ and $1 > x_0 > x > x_1 > 0$ and $0 < y_0 < y < y_1 < 1$, we have

$$\mathbb{P}\,(K_n = 0; \mathbf{X} \in \mathrm{d}\mathbf{x}; \mathbf{X}_0 \in \mathrm{d}\mathbf{x}_0; \mathbf{X}_1 \in \mathrm{d}\mathbf{x}_1) = (n-1)^{\underline{2}} (1 - x_0 y_1)^{n-3}\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{x}_0\, \mathrm{d}\mathbf{x}_1.$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

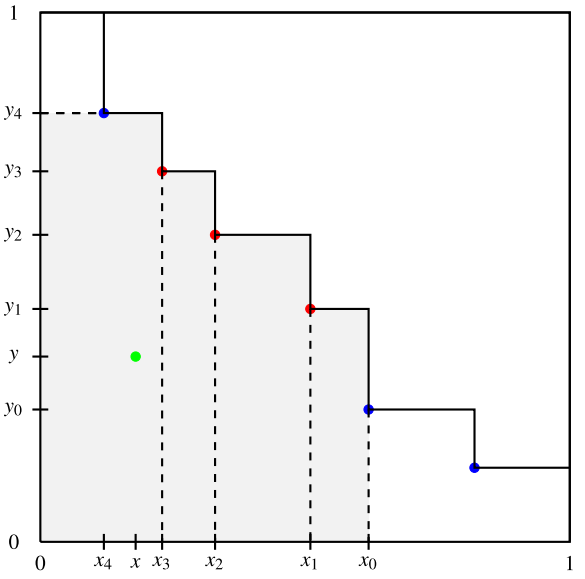

Figure 1. In this example, after $n-1$ observations, none of which fall in the shaded region S, there are $r_{n-1}=6$ remaining records. The nth observation, shown in green, breaks the $K_{n-1}=k=3$ remaining records shown in red but not the $r_{n-1}-K_{n-1}=3$ remaining records shown in blue.

(ii) For $n \geqslant 2$ and $1 > x > x_1 > 0$ and $0 < y < y_1 < 1$, we have

$$\mathbb{P}\,(K_n = 0; \mathbf{X} \in \mathrm{d}\mathbf{x}; \mathbf{X}_0 = \mathbf{e}_1; \mathbf{X}_1 \in \mathrm{d}\mathbf{x}_1) = (n-1)(1-y_1)^{n-2}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}_1.$$

(iii) For $n \geqslant 1$ and $1 > x > 0$ and $0 < y < 1$, we have

$$\mathbb{P}\,(K_n = 0; \mathbf{X} \in \mathrm{d}\mathbf{x}; \mathbf{X}_0 = \mathbf{e}_1; \mathbf{X}_1 = \mathbf{e}_2) = \mathbf{1}(n=1)\,\mathrm{d}\mathbf{x}.$$

To obtain an exact expression for $\mathbb{P}\,(K_n = k)$, one need only integrate out the variables $\mathbf{x}, \mathbf{x}_i$ in Proposition 3.1 to get

$$\mathbb{P}\,(K_n = k) = A_k + 2B_k + C_k, \tag{3.2}$$

where A_k, B_k and C_k (all of which also depend on n) correspond to parts (i), (ii) and (iii) of the proposition, respectively. For small values of k this can be done explicitly, but for general k we take an inductive approach. To get started on the induction, we first treat the case $k=0$.

3.2 The case $k=0$

Using Remark 3.1, we obtain the following result.

Proposition 3.2. *We have*

$$A_0 = \mathbf{1}(n \geqslant 3)\left[\frac{1}{2}n^{-1}H_n - \frac{3}{4}n^{-1}\right], \quad B_0 = \mathbf{1}(n \geqslant 2)\frac{1}{2}n^{-1}, \quad C_0 = \mathbf{1}(n=1),$$

and therefore

$$\mathbb{P}\,(K_n = 0) = \begin{cases} \dfrac{1}{2}n^{-1}H_n + \dfrac{1}{4}n^{-1} & \text{if } n \geqslant 2, \\ 1 & \text{if } n = 1. \end{cases}$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

Proof. Using Remark 3.1, we perform the computations in increasing order of difficulty. First, it is clear that $C_0 = 0$ for $n \geqslant 2$. Next, for $n \geqslant 2$ we have

$$\begin{aligned} B_0 &= \int_{\substack{1>x>x_1>0,\\ 0<y<y_1<1}} (n-1)(1-y_1)^{n-2}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}_1 \\ &= \frac{1}{2}(n-1)\int_{y_1=0}^{1} y_1(1-y_1)^{n-2}\,\mathrm{d}y_1 \\ &= \frac{1}{2}n^{-1}. \end{aligned}$$

Finally, for $n \geqslant 3$ we have

$$\begin{aligned} A_0 &= \int_{\substack{1>x_0>x>x_1>0,\\ 0<y_0<y<y_1<1}} (n-1)^{\underline{2}}(1-x_0\,y_1)^{n-3}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}_0\,\mathrm{d}\mathbf{x}_1 \\ &= \frac{1}{4}(n-1)^{\underline{2}}\int_{x_0=0}^{1}\int_{y_1=0}^{1} x_0^2\,y_1^2(1-x_0\,y_1)^{n-3}\,\mathrm{d}y_1\,\mathrm{d}x_0 \\ &= \frac{1}{4}(n-1)^{\underline{2}}\int_{x=0}^{1} x^{-1}\int_{z=0}^{x} z^2(1-z)^{n-3}\,\mathrm{d}z\,\mathrm{d}x \\ &= \frac{1}{2}n^{-1}\int_{x=0}^{1} x^{-1}[1-(1-x)^n]\,\mathrm{d}x - \frac{1}{2}\int_{x=0}^{1}(1-x)^{n-1}\,\mathrm{d}x - \frac{1}{4}(n-1)\int_{x=0}^{1} x(1-x)^{n-2}\,\mathrm{d}x, \end{aligned}$$

the final equality after two integrations by parts. Using the computation in the proof of Proposition 2.1 and the above computation of B_0, for $n \geqslant 3$ we therefore find

$$\begin{aligned} A_0 &= \frac{1}{2}\,\mathbb{P}\,(K_n \geqslant 0) - \frac{1}{2}n^{-1} - \frac{1}{2}B_0 \\ &= \frac{1}{2}n^{-1}H_n - \frac{1}{2}n^{-1} - \frac{1}{4}n^{-1} \\ &= \frac{1}{2}n^{-1}H_n - \frac{3}{4}n^{-1}. \end{aligned}$$

Now we just use (3.2) to establish the asserted expression for $\mathbb{P}\,(K_n = 0)$. □

3.3 Simplifications

The expressions obtained from Proposition 3.1 for A_k, B_k and C_k for $k \geqslant 1$ are easily simplified by integrating out the four variables x, x_{k+1}, y_0, y that do not appear in the integrand (when they do appear as variables). Here is the result.

Lemma 3.3. *Assume $k \geqslant 0$. Let A_k, B_k, C_k be defined as explained at (3.2).*

(i) *For $n \geqslant k+3$ we have*

$$\begin{aligned} A_k = \frac{1}{4}(n-1)^{\underline{k+2}} \\ \times \int_{\substack{1>x_0>\cdots>x_k>0,\\ 0<y_1<\cdots<y_{k+1}<1}} x_k^2 y_1^2\left[1-\left\{\sum^{k} +x_k y_{k+1}\right\}\right]^{n-(k+3)} \mathrm{d}x_0\,\mathrm{d}\mathbf{x}_1\cdots\mathrm{d}\mathbf{x}_k\,\mathrm{d}y_{k+1}. \end{aligned}$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

(ii) *For $n \geqslant k+2$ we have*

$$B_k = \frac{1}{2}(n-1)^{\underline{k+1}} \times \int_{\substack{1>x_1>\cdots>x_k>0,\\ 0<y_1<\cdots<y_{k+1}<1}} x_k^2 y_1 \left[1 - \left\{\sum^{k} + x_k y_{k+1}\right\}\right]^{n-(k+2)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_k \, \mathrm{d}y_{k+1},$$

where $x_0 = 1$ and if $k = 0$ then the integral is taken over $0 < y_1 < 1$.

(iii) *For $n \geqslant k+1$ we have*

$$C_k = (n-1)^{\underline{k}} \int_{\substack{1>x_1>\cdots>x_k>0,\\ 0<y_1<\cdots<y_k<1}} x_k y_1 \left[1 - \left\{\sum^{k} + x_k\right\}\right]^{n-(k+1)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_k,$$

where $x_0 = 1$ and if $k = 0$ then the interpretation is $C_0 = \mathbf{1}(n = 1)$.

Remark 3.2. Alternative expressions involving only finite sums are available for A_k, B_k, C_k by recasting the expressions in square brackets in Lemma 3.3 as finite sums of non-negative terms, expanding the integrand multinomially, and integrating the resulting polynomials explicitly. When this is done, one finds that A_k, B_k, C_k are all rational, as therefore are $\mathbb{P}(K_n = k)$ and $\mathbb{P}(K_n = k \mid K_n \geqslant 0)$.

Take C_k as an example. We have

$$1 - \left\{\sum^{k} + x_k\right\} = \sum_{i=1}^{k} (x_{i-1} - x_i)(1 - y_i),$$

and carrying out this procedure yields

$$C_k = n^{-2} \sum \prod_{i=1}^{k} \left(i + \sum_{\ell=k+1-i}^{k} j_\ell\right)^{-1},$$

where the indicated sum is taken over k-tuples $(j_1, \ldots, j_k)$ of non-negative integers summing to $n - (k+1)$ and the natural interpretation for $k = 0$ is $C_0 = \mathbf{1}(n = 1)$. Examples include

$$\begin{aligned} C_1 &= n^{-2}(n-1)^{-1}, \quad n \geqslant 2, \\ C_2 &= n^{-2}(n-1)^{-1} H_{n-2}, \quad n \geqslant 3, \\ C_{n-1} &= n^{-2} \prod_{i=1}^{n-2} i^{-1} = (n!\, n)^{-1}, \quad n \geqslant 1. \end{aligned} \tag{3.3}$$

Since our aim is to compute $\mathbb{P}(K_n = 0)$ up to additive error $O(n^{-2})$ for large n, the following lemma will suffice to treat the contributions C_k.

Lemma 3.4. *For $n \geqslant 1$, the probabilities $C_k \geqslant 0$ satisfy*

$$\sum_{k=0}^{\infty} C_k = \sum_{k=0}^{n-1} C_k = n^{-2}.$$

Proof. Recalling that r_n denotes the number of remaining records at time n, it is clear from the description of case (iii) leading up to Proposition 3.1 that

$$C_k = \mathbb{P}(r_{n-1} = k,\ K_n = k) = \mathbb{P}(r_{n-1} = k, K_n = r_{n-1}).$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

Therefore

$$\sum_{k=0}^{\infty} C_k = \mathbb{P}\,(K_n = r_{n-1}) = \mathbb{P}\,(\mathbf{X} \prec \mathbf{X}^{(i)} \text{ for all } 1 \leqslant i \leqslant n-1) = n^{-2}. \qquad \square$$

3.4 Recurrence relations

In this subsection we establish recurrence relations for A_k and B_k in the variable k, holding n fixed and treating the probabilities C_k as known.

Lemma 3.5. *For $k \geqslant 1$ we have*

(i) $A_k = \frac{1}{2}(A_{k-1} - B_k)$ *if* $n \geqslant k+3$,
(ii) $B_k = \frac{1}{2}(B_{k-1} - C_k)$ *if* $n \geqslant k+2$.

Proof. (i) Begin with the expression for A_k in Lemma 3.3 and integrate out the variable x_0. This gives

$$\begin{aligned} A_k = {} & \frac{1}{4}(n-1)^{\underline{k+1}} \\ & \times \left(\int_{\substack{1>x_1>\cdots>x_k>0,\\ 0<y_1<\cdots<y_{k+1}<1}} x_k^2 y_1 \left[1 - \left\{ \sum_{2}^{k} + x_k y_{k+1} \right\} \right]^{n-(k+2)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_k \, \mathrm{d}y_{k+1} \right. \\ & \left. - \int_{\substack{1>x_1>\cdots>x_k>0,\\ 0<y_1<\cdots<y_{k+1}<1}} x_k^2 y_1 \left[1 - \left\{ \sum_{1}^{k} + x_k y_{k+1} \right\} \right]^{n-(k+2)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_k \, \mathrm{d}y_{k+1} \right) \\ = {} & A'_k - A''_k \quad \text{(say)}, \end{aligned}$$

with $x_0 = 1$ in the subtracted integral. For A'_k, observe that the variable y_1 does not appear within the square brackets in the integrand. Thus, integrating out y_1 and then shifting variable names, we find

$$\begin{aligned} A'_k = {} & \frac{1}{8}(n-1)^{\underline{k+1}} \\ & \times \int_{\substack{1>x_1>\cdots>x_k>0,\\ 0<y_2<\cdots<y_{k+1}<1}} x_k^2 y_2^2 \left[1 - \left\{ \sum_{2}^{k} + x_k y_{k+1} \right\} \right]^{n-(k+2)} \mathrm{d}x_1 \, \mathrm{d}\mathbf{x}_2 \cdots \mathrm{d}\mathbf{x}_k \, \mathrm{d}y_{k+1} \\ = {} & \frac{1}{8}(n-1)^{\underline{k+1}} \int_{\substack{1>x_0>\cdots>x_{k-1}>0,\\ 0<y_1<\cdots<y_k<1}} x_{k-1}^2 y_1^2 \left[1 - \left\{ \sum^{k-1} + x_{k-1} y_k \right\} \right]^{n-(k+2)} \mathrm{d}x_0 \, \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_{k-1} \, \mathrm{d}y_k \\ = {} & \frac{1}{2} A_{k-1}, \end{aligned}$$

where the last equality follows from Lemma 3.3. We see also from Lemma 3.3 that $A''_k = \frac{1}{2}B_k$. This completes the proof of part (i).
(ii) The proof of part (ii) is similar. Begin with the expression for B_k in Lemma 3.3 and integrate out the variable y_{k+1}. This gives (with $x_0 = 1$)

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

$$B_k = \frac{1}{2}(n-1)^{\underline{k}} \left(\int_{\substack{1>x_1>\cdots>x_k>0,\\ 0<y_1<\cdots<y_k<1}} x_k y_1 \left[1 - \left\{ \sum^{k} + x_k y_k \right\} \right]^{n-(k+1)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_k \right.$$
$$\left. - \int_{\substack{1>x_1>\cdots>x_k>0,\\ 0<y_1<\cdots<y_k<1}} x_k y_1 \left[1 - \left\{ \sum^{k} + x_k \right\} \right]^{n-(k+1)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_k \right)$$
$$= B'_k - B''_k \text{ (say).}$$

For B'_k, observe that the expression within $\{\cdot\}$ equals $\sum^{k-1} + x_{k-1}y_k$, which does not depend on x_k. Thus, integrating out x_k, we find

$$B'_k = \frac{1}{4}(n-1)^{\underline{k}} \int_{\substack{1>x_1>\cdots>x_{k-1}>0,\\ 0<y_1<\cdots<y_k<1}} x_{k-1}^2 y_1 \left[1 - \left\{ \sum^{k-1} + x_{k-1}y_k \right\} \right]^{n-(k+1)} \mathrm{d}\mathbf{x}_1 \cdots \mathrm{d}\mathbf{x}_{k-1}\, \mathrm{d}y_k$$
$$= \frac{1}{2} B_{k-1},$$

where the last equality follows from Lemma 3.3. We see also from Lemma 3.3 that $B''_k = \frac{1}{2}C_k$. This completes the proof of part (ii). □

Using the 'initial conditions' delivered by Proposition 3.2, it is routine to solve the first-order linear recurrence relations of Lemma 3.5 in terms of the probabilities C_k.

Lemma 3.6. *For $n \geqslant 1$ and $k \geqslant 0$ we have*

$$A_k = \mathbf{1}(n \geqslant k+3) \left[2^{-k} A_0 - k 2^{-(k+1)} B_0 + \sum_{j=1}^{k} (k+1-j) 2^{-(k+2-j)} C_j \right], \tag{3.4}$$

$$B_k = \mathbf{1}(n \geqslant k+2) \left[2^{-k} B_0 - \sum_{j=1}^{k} 2^{-(k+1-j)} C_j \right]. \tag{3.5}$$

Proof. Clearly we have (3.5) and likewise

$$A_k = 2^{-k} A_0 - \sum_{j=1}^{k} 2^{-(k+1-j)} B_j. \tag{3.6}$$

Then, plugging (3.5) into (3.6) and rearranging yields (3.4). □

3.5 Approximation to the probability $\mathbb{P}(K_n = k)$, with error bound

Theorem 3.1. *For $n \geqslant 1$ and every $k \geqslant 0$ we have*

$$|\, \mathbb{P}(K_n = k) - [2^{-(k+1)} n^{-1} H_n - (k-1) 2^{-(k+2)} n^{-1}]| \leqslant \frac{1}{2} n^{-2}. \tag{3.7}$$

Proof. Recall from (3.2) that $\mathbb{P}(K_n = k) = A_k + 2B_k + C_k$. Substitute for A_k and B_k using Lemma 3.6, then substitute for A_0 and B_0 using Proposition 3.2, and finally rearrange.

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

For $0 \leqslant k \leqslant n-3$ this gives

$$\mathbb{P}(K_n = k) = 2^{-k}A_0 - (k-4)2^{-(k+1)}B_0 + \sum_{j=1}^{k-1}(k-3-j)2^{-(k+2-j)}C_j + \frac{1}{4}C_k$$
$$= 2^{-(k+1)}n^{-1}H_n - (k-1)2^{-(k+2)}n^{-1} + \sum_{j=1}^{k-1}(k-3-j)2^{-(k+2-j)}C_j + \frac{1}{4}C_k.$$

Denote the coefficient of C_j (with $1 \leqslant j \leqslant k$) by $c_{k,j}$. Note that $c_{k,j} \equiv c_{k-j}$ depends only on $k-j \geqslant 0$, and that $|c_i| \leqslant 1/4$ (with equality for $c_0 = 1/4$ and $c_1 = -1/4$). So Lemma 3.4 gives the bound on the remainder term (with half as big a constant).

For $k = n-2$ this gives

$$\mathbb{P}(K_n = k) = 2^{-k}n^{-1} - \sum_{j=1}^{k-1}2^{-(k-j)}C_j.$$

A simple argument omitted here shows that this differs from the approximation in the statement of the theorem by at most $\frac{1}{2}n^{-2}$ for all $n \geqslant 1$.

For $k = n-1$, this together with (3.3) gives

$$\mathbb{P}(K_n = k) = C_{n-1} = (n!\, n)^{-1}.$$

Now another simple and omitted argument shows that this differs from the approximation in the statement of the theorem by at most $\frac{1}{4}n^{-2}$ for all $n \geqslant 1$.

For $k \geqslant n$ we have $\mathbb{P}(K_n = k) = 0$, and another simple argument shows that this differs from the asserted approximation by at most $\frac{1}{2}n^{-2}$ provided $n \geqslant 6$, the worst case being $k = 7$ for $n = 6$ and $k = n$ for $n \geqslant 7$. Further, the bound can be checked directly for $n = 1, 2, 3, 4, 5$, the worst k in each of those cases again being $k = n$. □

Example 3.7. The matrix $C = C_{n,k}$ with $1 \leqslant n \leqslant 5$ and $0 \leqslant k \leqslant 4$ is

$$\begin{bmatrix} 1 & & & & \\ 0 & \frac{1}{4} & & & \\ 0 & \frac{1}{18} & \frac{1}{18} & & \\ 0 & \frac{1}{48} & \frac{1}{32} & \frac{1}{96} & \\ 0 & \frac{1}{100} & \frac{11}{600} & \frac{1}{100} & \frac{1}{600} \end{bmatrix}.$$

Observe that the nth row sums to n^{-2}, as noted at Lemma 3.4. The matrix with entries $\mathbb{P}(K_n = k)$ for the same values of n and k is

$$\begin{bmatrix} 1 & & & & \\ \frac{1}{2} & \frac{1}{4} & & & \\ \frac{7}{18} & \frac{1}{6} & \frac{1}{18} & & \\ \frac{31}{96} & \frac{13}{96} & \frac{5}{96} & \frac{1}{96} & \\ \frac{167}{600} & \frac{7}{60} & \frac{7}{150} & \frac{1}{75} & \frac{1}{600} \end{bmatrix}. \tag{3.8}$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

Observe that the nth row sums to $n^{-1}H_n$, as guaranteed by Proposition 2.1. The matrix with entries $\mathbb{P}(K_n = k \mid K_n \geqslant 0)$ is therefore

$$\begin{bmatrix} 1 & & & & \\ \frac{2}{3} & \frac{1}{3} & & & \\ \frac{7}{11} & \frac{3}{11} & \frac{1}{11} & & \\ \frac{31}{50} & \frac{13}{50} & \frac{5}{50} & \frac{1}{50} & \\ \frac{167}{274} & \frac{35}{137} & \frac{14}{137} & \frac{4}{137} & \frac{1}{274} \end{bmatrix},$$

with every row summing to unity.

Remark 3.3.

(a) Not that the optimal numerical constant appearing on the right in (3.7) is important to know, but it would appear from (3.8) and other computations that the optimal constant is 1/4, achieved in four cases: $n = 1, 2$ with $k = n - 1, n$.

(b) More importantly, we do not know whether the order n^{-2} of the error bound in Theorem 3.1 is asymptotically optimal. While the approximation is *perfect* for $k = 0$ if $n \geqslant 2$, for $k = 1$ it underestimates $\mathbb{P}(K_n = k)$ by $\frac{1}{4}C_1 = \frac{1}{4}n^{-2}(n-1)^{-1}$ if $n \geqslant 2$, and for $k = 2$ it underestimates by $\frac{1}{4}(C_2 - C_1) = \frac{1}{4}n^{-2}(n-1)^{-1}(H_{n-2} - 1)$ if $n \geqslant 3$. Thus the rate of convergence is $O(n^{-2})$ but $\Omega(n^{-3}\log n)$.

For fixed $k \geqslant 1$, we conjecture that the correct rate of convergence is $\Theta(n^{-3}(\log n)^{k-1})$, and more strongly that the error satisfies

$$[2^{-(k+1)}n^{-1}H_n - (k-1)2^{-(k+2)}n^{-1}] - \mathbb{P}(K_n = k) \sim -\frac{1}{4}C_k \sim n^{-3}\frac{(\mathrm{L}\, n)^{k-1}}{(k-1)!}$$

as $n \to \infty$. Since

$$\sup_{k \geqslant 1} \frac{(\mathrm{L}\, n)^{k-1}}{(k-1)!} = \Theta\left(\frac{n}{\sqrt{\log n}}\right),$$

this suggests that perhaps the optimal rate (uniformly in k) for Theorem 3.1 is the small improvement $\Theta(n^{-2}(\log n)^{-1/2})$.

4. Conjectures

The upshot of this section is that a variance bound would imply a Glivenko–Cantelli type theorem: Conjecture 4.5 would imply Conjecture 4.1.

4.1 The natural conjecture

While our main Theorem 1.1 does begin to explain how the Geometric(1/2) distribution arises in connection with the breaking of bivariate records, it is not the conjecture to which one is led by performing many independent trials of generating a large number M of records and, for each trial, watching the table such as Table 1 evolve as records are generated one at a time. A natural

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

conjecture concerns the fractions of records that break k remaining records, for various values of k. Accordingly, let

$$\tilde{p}_{M,k} := M^{-1} \sum_{m=1}^{M} \tilde{I}_{m,k},$$

where

$$\tilde{I}_{m,k} := \mathbf{1}(m\text{th record generated breaks precisely } k \text{ remaining records}).$$

A strong conjecture one might form is the following, of Glivenko–Cantelli type.

Conjecture 4.1. *The fractions $\tilde{p}_{M,k}$ of the first M records that break precisely k remaining records satisfy*

$$\sup_{k \geqslant 0} |\tilde{p}_{M,k} - 2^{-(k+1)}| \xrightarrow{\text{a.s.}} 0 \quad \text{as } M \to \infty.$$

In the remaining subsections we show how proving this conjecture can be reduced to an asymptotic variance calculation, and we leave that calculation for future research.

4.2 Uniformity in k

Of course, Conjecture 4.1 would have the following corollary, of strong law of large numbers type.

Conjecture 4.2. *For each fixed $k \geqslant 0$, the fraction $\tilde{p}_{M,k}$ of the first M records that breaks precisely k remaining records satisfies*

$$\tilde{p}_{M,k} \xrightarrow{\text{a.s.}} 2^{-(k+1)} \quad \text{as } M \to \infty.$$

But it is standard to check that Conjecture 4.2 also implies Conjecture 4.1. For completeness, here is a proof, with all claims holding almost surely. Let $\varepsilon_{M,k} \geqslant 0$ denote the random variable $|\tilde{p}_{M,k} - 2^{-(k+1)}|$. Then, for any $K \geqslant 0$, we have

$$\varepsilon_M := \sup_{k \geqslant 0} \varepsilon_{M,k} = \max\Big\{ \max_{k \leqslant K} \varepsilon_{M,k},\ \sup_{k > K} \varepsilon_{M,k} \Big\},$$

and so

$$\limsup_{M \to \infty} \varepsilon_M = \limsup_{M \to \infty} \sup_{k > K} \varepsilon_{M,k}$$

by Conjecture 4.2. But

$$\sup_{k > K} \varepsilon_{M,k} \leqslant \sum_{k > K} \tilde{p}_{M,k} + 2^{-(K+1)} = 1 - \sum_{k \leqslant K} \tilde{p}_{M,k} + 2^{-(K+1)}.$$

Therefore

$$\limsup_{M \to \infty} \varepsilon_M \leqslant 1 - \sum_{k \leqslant K} \lim_{M \to \infty} \tilde{p}_{M,k} + 2^{-(K+1)} = 1 - \sum_{k \leqslant K} 2^{-(k+1)} + 2^{-(K+1)} = 2^{-K}.$$

Letting $K \to \infty$ completes the proof. □

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

4.3 Time change

We show next that Conjecture 4.2 would follow from the following 'observations-time' conjecture. Let

$$R_{n,k} := \sum_{i=1}^{n} I_{i,k}, \tag{4.1}$$

where

$$I_{i,k} := \mathbf{1}(K_i = k).$$

Note that

$$R_n = \sum_{k \geqslant 0} R_{n,k},$$

and define

$$p_{n,k} := \frac{R_{n,k}}{R_n}.$$

Conjecture 4.3. *For each fixed $k \geqslant 0$, we have*

$$p_{n,k} \xrightarrow{\text{a.s.}} 2^{-(k+1)} \quad \text{as } n \to \infty.$$

Here is a proof that Conjecture 4.3 implies Conjecture 4.2. Working in observations-time, for $m \geqslant 1$, let T_m denote the time at which the mth record is set, so that $R_{T_m} = m$ for all m. In similar fashion, $R_{T_M,k} = \sum_{m=1}^{M} \widetilde{I}_{m,k}$. Thus Conjecture 4.2 follows from Conjecture 4.3 simply by looking at the sequence (T_m) of n-values. □

4.4 Expectations

Conjecture 4.3 is certainly plausible, because, as we prove in this subsection, with

$$\rho_{n,k} := \mathbb{E}\, R_{n,k}, \quad \rho_n := \mathbb{E}\, R_n, \quad \phi_{n,k} := \frac{\rho_{n,k}}{\rho_n}$$

we have

$$\phi_{n,k} \to 2^{-(k+1)} \quad \text{as } n \to \infty. \tag{4.2}$$

In the statement of the following lemma, we refer (indirectly) to the second-order harmonic numbers

$$H_n^{(2)} = \frac{\pi^2}{6} - (1 + o(1))n^{-1} \quad \text{as } n \to \infty, \quad \text{where } H_n^{(r)} := \sum_{i=1}^{n} i^{-r}$$

(aside: we shall encounter the fourth-order harmonic numbers in Section 4.6), and (directly) to the second-order *Roman harmonic numbers* (see [5] and references [16, 22, 23] therein)

$$\begin{aligned} c_n^{(2)} &:= \sum_{i=1}^{n} i^{-1} H_i \\ &= \frac{1}{2}(H_n^2 + H_n^{(2)}) \\ &= \frac{1}{2}(\mathrm{L}\, n)^2 + \gamma\, \mathrm{L}\, n + \frac{1}{2}\left(\frac{\pi^2}{6} + \gamma^2\right) + O(n^{-1} \log n). \end{aligned}$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

The lemma shows that

$$\hat{\rho}_{n,k} := 2^{-(k+1)} c_n^{(2)} - (k-1)2^{-(k+2)} H_n$$

gives a good approximation to $\rho_{n,k}$.

Lemma 4.1. *For $n \geqslant 1$ we have*

$$\rho_n = c_n^{(2)} \tag{4.3}$$

and, for every $k \geqslant 0$, also

$$|\hat{\rho}_{n,k} - \rho_{n,k}| \leqslant \frac{1}{2} H_n^{(2)} < 1. \tag{4.4}$$

Proof. For (4.3), just sum the result of Proposition 2.1 (with n replaced by i) over i from 1 to n. For (4.4), apply the same operation to (3.7) in Theorem 3.1, observing $\pi^2/12 < 1$. □

Remark 4.1. From Lemma 4.1 it is an immediate corollary that

$$\sup_{k \geqslant 0} \left| \phi_{n,k} - \left[2^{-(k+1)} - (k-1)2^{-(k+2)} \frac{H_n}{c_n^{(2)}} \right] \right| < \frac{1}{c_n^{(2)}} \sim (\mathrm{L}\, n)^{-2}.$$

In particular, (4.2) holds, uniformly in k.

4.5 Reduction to a variance calculation

In light of Lemma 4.1, to establish $p_{n,k} \xrightarrow{\mathrm{P}} 2^{-(k+1)}$ as $n \to \infty$ it would be sufficient to establish concentration of measure for the distributions of the denominator R_n and the numerator $R_{n,k}$ of $p_{n,k}$; for example, by means of variance bounds combined with Chebyshev's inequality. As we will explain in this subsection, we already know about the variance of R_n, and if we were to bound the variance of $R_{n,k}$ in suitably similar fashion we could prove not only convergence in probability but also the almost sure convergence of Conjecture 4.2.

The following results concerning R_n are implied by [4, Theorems 4.1(b), 4.2(a)] (with the mean, variance and central limit theorem results there taken from [2] and [1]) after specializing to our present case of dimension $d = 2$.

Lemma 4.2. *Let Φ denote the standard normal distribution function. The number R_n of records set through time n satisfies*

$$\rho_n = \mathbb{E}\, R_n = \frac{1}{2}(\mathrm{L}\, n)^2 + \gamma\, \mathrm{L}\, n + \left(\frac{\pi^2}{12} + \frac{1}{2}\gamma^2 \right) + o(1),$$

$$\sigma_n^2 := \mathrm{Var}\, R_n \sim \left(\frac{\pi^2}{6} + \gamma^2 \right) (\mathrm{L}\, n)^2,$$

$$\sup_x \left| \mathbb{P}\left(\frac{R_n - \rho_n}{\sigma_n} < x \right) - \Phi(x) \right| = O((\log n)^{-1/2} (\log\log n)^3),$$

$$\mathbb{P}\left(|R_n - \rho_n| \geqslant (\mathrm{L}\, n)^{(3/2)+\varepsilon} \text{ infinitely often} \right) = 0 \quad \text{if } \varepsilon > 0, \tag{4.5}$$

and consequently

$$\frac{R_n}{\rho_n} \xrightarrow{\text{a.s.}} 1. \tag{4.6}$$

□

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

A careful review of the proof of (4.5) (a first Borel–Cantelli argument applied along a geometrically increasing sequence of times), which immediately implies (4.6), shows that to establish (4.5) it is sufficient to know that the samples paths of the process R are non-decreasing, that

$$\rho_n = a(\mathrm{L}\, n)^2 + b(\mathrm{L}\, n) + O(1)$$

for some constants $a > 0$ and b, that $\sigma_n^2 = O((\log n)^2)$, and that

$$\rho_n - \rho_{n-1} = \Theta(n^{-1} \log n).$$

Now observe, for each fixed $k \geqslant 0$, that the sample paths of the process $R_{\cdot,k}$ are non-decreasing, that

$$\rho_{n,k} = a_k(\mathrm{L}\, n)^2 + b_k(\mathrm{L}\, n) + O(1)$$

with $a_k = 2^{-(k+2)} > 0$ and $b_k = -2^{-(k+2)}(k - 2\gamma - 1)$, and that

$$\rho_{n,k} - \rho_{n-1,k} = \mathbb{P}\,(K_n = k) = \Theta(n^{-1} \log n),$$

with the last equality holding by Theorem 3.1. Thus the analogues of (4.5)–(4.6) for $R_{\cdot,k}$ hold if we can establish that

$$\sigma_{n,k}^2 := \operatorname{Var} R_{n,k} \tag{4.7}$$

satisfies $\sigma_{n,k}^2 = O((\log n)^2)$, which (in light of the known corresponding result for R) seems eminently reasonable to conjecture.

Conjecture 4.4. *For each fixed $k \geqslant 0$, the variance $\sigma_{n,k}^2$ defined at (4.7) satisfies*

$$\sigma_{n,k}^2 = O((\log n)^2).$$

A summary of this subsection is that Conjecture 4.4 would imply Conjecture 4.3 and therefore also Conjecture 4.1.

Remark 4.2.

(a) Use of the refinement (4.5) to (4.6) shows that Conjecture 4.4 would imply the refinement

$$p_{n,k} = 2^{-(k+1)}[1 + O((\log n)^{-(1/2)+\varepsilon})] \quad \text{a.s.}$$

of Conjecture 4.3 for each fixed $k \geqslant 0$ and any $\varepsilon > 0$.

(b) More than Conjecture 4.4, we conjecture that for each fixed $k \geqslant 0$ we have

$$\sigma_{n,k}^2 \sim s_k^2(\mathrm{L}\, n)^2$$

for some constants $s_k^2 > 0$ satisfying $s_k^2 \to 0$ as $k \to \infty$ (likely with $s_k \equiv 2^{-(k+1)}s$, letting $s^2 := (\pi^2/6) + \gamma^2$), and that there is asymptotic normality for $R_{n,k}$. It seems reasonable to conjecture that, moreover, the random vector $(R_{n,1}, \ldots, R_{n,k})$ enjoys full-dimensional asymptotic k-variate normality.

(c) It may be that the random variables $R_{n,k}$ are positively correlated for fixed n as k varies, the idea being that larger values of R_n (more records) should lead to larger values of $R_{n,k}$ (more records that break k remaining records) for every k. If this positive correlation were to be known, then Conjecture 4.4 would follow immediately, without the need for additional calculations. Indeed, for large n and fixed k we would then have

$$\sigma_{n,k}^2 \leqslant \sum_{j=1}^{n} \sigma_{n,j}^2 \leqslant \sigma_n^2 \sim s^2(\mathrm{L}\, n)^2.$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

4.6 Reduction of the variance calculation

Corresponding to the breakdown into cases utilized in Section 3, observe that $I_{n,k} = \mathbf{1}(K_n = k)$ satisfies

$$I_{n,k} = I_{n,k}^{(0)} + I_{n,k}^{(1)} + I_{n,k}^{(2)} + I_{n,k}^{(1,2)},$$

where the four terms here are the respective indicators of the events

$$\{K_n = k,\ \mathbf{X}^{(n)} \text{ does not set a record in either coordinate}\},$$
$$\{K_n = k,\ \mathbf{X}^{(n)} \text{ sets a record in the first coordinate but not the second}\},$$
$$\{K_n = k,\ \mathbf{X}^{(n)} \text{ sets a record in the second coordinate but not the first}\},$$
$$\{K_n = k,\ \mathbf{X}^{(n)} \text{ sets a record in both coordinates}\}.$$

By analogy with (4.1), define respective record counts $R_{n,k}^{(0)}, R_{n,k}^{(1)}, R_{n,k}^{(2)}, R_{n,k}^{(1,2)}$, so that

$$R_{n,k} = R_{n,k}^{(0)} + R_{n,k}^{(1)} + R_{n,k}^{(2)} + R_{n,k}^{(1,2)}. \tag{4.8}$$

It thus seems daunting to calculate $\sigma_{n,k}^2$ to prove Conjecture 4.4. But in this subsection we argue by means of suitable control of all but the first term in (4.8) that

$$\sigma_{n,k}^2 = \operatorname{Var} R_{n,k}^{(0)} + O((\log n)^2),$$

for fixed k, thus reducing proof of Conjecture 4.4 to proof of the following simpler conjecture.

Conjecture 4.5. *For each fixed $k \geqslant 0$, we have*

$$\operatorname{Var} R_{n,k}^{(0)} = O((\log n)^2).$$

Here is a proof that Conjecture 4.5 would imply Conjecture 4.4. By the triangle inequality for L^2-norm $\|\cdot\|_2$, in obvious notation we have

$$\sigma_{n,k} - \sigma_{n,k}^{(0)} \leqslant \sigma_{n,k}^{(1)} + \sigma_{n,k}^{(2)} + \sigma_{n,k}^{(1,2)} = 2\sigma_{n,k}^{(1)} + \sigma_{n,k}^{(1,2)}. \tag{4.9}$$

But with $R_n^{(1)}$ counting the number of records through time n in the first coordinate, we have

$$\begin{aligned}
\operatorname{Var} R_{n,k}^{(1)} &\leqslant \|R_{n,k}^{(1)}\|_2^2 \\
&\leqslant \|R_n^{(1)}\|_2^2 \\
&= [\,\mathbb{E}\, R_n^{(1)}]^2 + \operatorname{Var} R_n^{(1)} \\
&= H_n^2 + [H_n - H_n^{(2)}] \\
&= O((\log n)^2),
\end{aligned} \tag{4.10}$$

and with $R_n^{(1,2)}$ counting the number of observations through time n that set a record in both coordinates, we have

$$\begin{aligned}
\operatorname{Var} R_{n,k}^{(1,2)} &\leqslant \|R_{n,k}^{(1,2)}\|_2^2 \\
&\leqslant \|R_n^{(1,2)}\|_2^2
\end{aligned}$$

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

$$\begin{aligned} &= [\,\mathbb{E}\, R_n^{(1,2)}]^2 + \operatorname{Var} R_n^{(1,2)} \\ &= (H_n^{(2)})^2 + [H_n^{(2)} - H_n^{(4)}] \\ &= O(1) \\ &= o((\log n)^2). \end{aligned} \tag{4.11}$$

Thus, returning to (4.9) and applying the inequality $(a+b)^2 \leqslant 2(a^2+b^2)$, we find

$$\sigma_{n,k}^2 \leqslant [\sigma_{n,k}^{(0)} + O(\log n)]^2 \leqslant 2 \operatorname{Var} R_{n,k}^{(0)} + O((\log n)^2),$$

and so Conjecture 4.5 would imply Conjecture 4.4. □

Remark 4.3.

(a) Observe that $R_{n,0}^{(1,2)} = 1$ for every $n \geqslant 1$, and so $\operatorname{Var} R_{n,0}^{(1,2)} = 0$. For $k \geqslant 1$, we claim that (4.11) can be strengthened to $\operatorname{Var} R_{n,k}^{(1,2)} = \Theta(1)$. To establish the lower bound $\operatorname{Var} R_{n,k}^{(1,2)} = \Omega(1)$ matching the upper bound (4.11), we perform two computations. The first, valid for $n \geqslant 2k+1$, is that

$$\mathbb{P}\,(R_{n,k}^{(1,2)} \geqslant 2) \geqslant \mathbb{P}\,(R_{2k+1,k}^{(1,2)} = 2) = \mathbb{P}\,(R_{k+1,k}^{(1,2)} = 1, R_{2k+1,k}^{(1,2)} = 2) > 0,$$

and the other, valid for $n \geqslant k+1$, is that

$$\begin{aligned} \mathbb{P}\,(R_{n,k}^{(1,2)} = 1) &\geqslant \mathbb{P}\,(R_{k+1,k}^{(1,2)} = 1,\ R_{n,k}^{(1,2)} = 1) \\ &\geqslant \mathbb{P}\,(R_{k+1,k}^{(1,2)} = 1)\,\mathbb{P}\,(R_{n-k}^{(1,2)} = 1) \\ &= \mathbb{P}\,(R_{k+1,k}^{(1,2)} = 1) \prod_{i=2}^{n-k} (1 - i^{-2}) \\ &= \frac{1}{2}\,\mathbb{P}\,(R_{k+1,k}^{(1,2)} = 1)[1 + (n-k)]^{-1} \\ &\geqslant \frac{1}{2}\,\mathbb{P}\,(R_{k+1,k}^{(1,2)} = 1) \\ &> 0. \end{aligned}$$

(b) We conjecture that (4.10) can be strengthened to $\operatorname{Var} R_{n,k}^{(1)} = \Theta(\log n)$. If we knew even the upper bound $\operatorname{Var} R_{n,k}^{(1)} = O(\log n)$, then it would follow from (4.9) and the matching upper bound on $\sigma_{n,k}^{(0)} - \sigma_{n,k}$ that

$$\sigma_{n,k} = \sigma_{n,k}^{(0)} + O((\log n)^{1/2}).$$

In that way, if one could prove the conjecture that $\sigma_{n,k}^{(0)} \sim s_k \,\mathrm{L}\, n$ for some constant $s_k > 0$, then the same leading-order asymptotics would apply to $\sigma_{n,k}$.

Acknowledgements

We thank Vince Lyzinski, Daniel Q. Naiman, Fred Torcaso and an anonymous referee for helpful comments, and Daniel Q. Naiman for producing Figure 1.

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press

References

[1] Bai, Z.-D., Chao, C.-C., Hwang, H.-K. and Liang, W.-Q. (1998) On the variance of the number of maxima in random vectors and its applications. *Ann. Appl. Probab.* **8** 886–895.
[2] Bai, Z.-D., Devroye, L., Hwang, H.-K. and Tsai, T.-H. (2005) Maxima in hypercubes. *Random Struct. Algorithms* **27** 290–309.
[3] Fill, J. A. and Naiman, D. Q. (2020) The Pareto record frontier. *Electron. J. Probab.* **25** 1–24.
[4] Fill, J. A. and Naiman, D. Q. (2019) The Pareto record frontier. arXiv:1901.05620.
[5] Sesma, J. (2017) The Roman harmonic numbers revisited. *J. Number Theory* **180** 544–565.

Cite this article: Fill JA (2021). Breaking bivariate records. *Combinatorics, Probability and Computing* **30**, 105–123. https://doi.org/10.1017/S0963548320000309

https://doi.org/10.1017/S0963548320000309 Published online by Cambridge University Press