The statistical curse of the second half-rank

J. Stat. Mech. (2011) P01025

(http://iopscience.iop.org/1742-5468/2011/01/P01025)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

The statistical curse of the second half-rank

Jean Desbois[1,2], **Stéphane Ouvry**[1,2] **and Alexios Polychronakos**[3]

[1] Laboratoire de Physique Théorique et Modèles Statistiques, Université Paris-Sud, UMR 8626, F-91405 Orsay, France
[2] CNRS, LPTMS, F-91405 Orsay, France
[3] Physics Department, City College of New York, NY, USA
E-mail: jean.desbois@u-psud.fr, stephane.ouvry@lptms.u-psud.fr and alexios@sci.ccny.cuny.edu

Received 4 November 2010
Accepted 23 December 2010
Published 20 January 2011

Online at stacks.iop.org/JSTAT/2011/P01025
doi:10.1088/1742-5468/2011/01/P01025

Abstract. In competitions involving many participants running many races the final rank is determined by the score of each participant, obtained by adding their ranks in each individual race. The 'statistical curse of the second half-rank' is the observation that if the score of a participant is even modestly worse than the middle score, then its final rank will be much worse (that is, much further away from the middle rank) than might have been expected. We give an explanation of this effect for the case of a large number of races using the central limit theorem. We present exact quantitative results in this limit and demonstrate that the score probability distribution will be Gaussian with scores packing near the center. We also derive the final rank probability distribution for the case of two races and we present some exact formulas verified by numerical simulations for the case of three races. The variant in which the worst result of each boat is dropped from its final score is also analyzed and solved for the case of two races.

Keywords: stochastic processes

 1742-5468/11/P01025+18$33.00

Contents

1. Introduction

In competitive individual sports involving many participants it is in some cases standard practice to have several races and determine the final rank for each participant by taking the sum of its ranks in each individual race, thereby defining its score. By comparing the scores of the participants a final rank can be decided among them. Typical examples are regattas, which can involve a large number of sailing boats (∼100), running a somewhat large number of consecutive races (≥10).

An empirical observation of long-time participants is that, if their scores are even slightly below the average, their final rank will be much worse than expected. This frustrating fact, which we may call the 'statistical curse of the second half-rank', is analyzed in this work and argued to be due to statistical fluctuations in the results of the races, on top of the inherent worth of the participants. Using some simplifying assumptions we demonstrate that it can be explained by a version of the central limit theorem [1, 2] for correlated random variables. A general result for a large number of participants and races is derived. Some exact results for a small number of races are presented. A variant of the problem, in which the worst rank for each participant is dropped, is also considered and solved for the case of two races.

2. The basic setup

Consider n_b boats racing n_r races. A boat i in the race k has an individual rank $n_{i,k} \in [1, n_\mathrm{b}]$ (lower ranks represent better performance). The score of the boat i is the sum $n_i = \sum_{k=1}^{n_\mathrm{r}} n_{i,k} \in [n_\mathrm{r}, n_\mathrm{r} n_\mathrm{b}]$ of its individual ranks in each race. The final rank of boat i is determined by the place occupied by its score n_i among the scores of the other boats n_j, with $j \neq i$.

For reasons of simplicity we assume that in a given race the ranks are uniformly distributed random variables with no *ex aequo* (that is, all boats are inherently equally worthy and there are no ties). We shall also take the ranks in different races to be independent random variables. It follows that for the race k the set $\{n_{ik}; i = 1, 2, \ldots, n_\mathrm{b}\}$ is a random permutation of $\{1, 2, \ldots, n_\mathrm{b}\}$, so the $n_{i,k}$s are correlated random variables (in particular $\sum_{i=1}^{n_\mathrm{b}} n_{i,k} = n_\mathrm{b}(n_\mathrm{b}+1)/2$), while n_{ik} and $n_{jk'}$ are uncorrelated for $k \neq k'$. We are interested in the probability distribution for boat i having a final rank $m \in [1, n_\mathrm{b}]$ given its score.

Let us illustrate this situation in the simple case of three boats racing two races. We have to take all random permutations of $\{1, 2, 3\}$ for both the first and the second race, and to add them to determine the possible scores of the three boats. It is easy to see that for, say, boat 1 to have a score $n_{1,1} + n_{1,2} = 4$ there are twelve possibilities:

(i) four instances where $n_{1,1} = 1$ and $n_{1,2} = 3$,
(ii) four instances where $n_{1,1} = 2$ and $n_{1,2} = 2$,
(iii) four instances where $n_{1,1} = 3$ and $n_{1,2} = 1$.

In each of these three cases (i), (ii) and (iii), one finds that boat 1 has an equal probability 1/2 for its final rank to be either $m = 1$ or 2. Its mean rank follows as $\langle m \rangle = (1/2)(1+2) = 3/2$. Clearly the score 4 is precisely the middle of the set $\{2, 3, 4, 5, 6\}$ and $\langle m \rangle = 3/2$ is indeed close[4] to the middle rank 2.

More interestingly, cases (i), (ii) and (iii) give the same final rank probability distribution. This means that the final rank probability distribution depends only on the score of boat 1, and not on its individual ranks in each of the two races consistent with its score. This fact is particular to two races and would no longer be true for three or more races. The final rank probability distribution for boat 1 given its score would depend in this case on the full set of its ranks in each race, and not just on its score. The final rank probability distribution for boat 1 should then be defined as the average of the above distributions for all sets of ranks consistent with its score.

To avoid this additional averaging and slightly simplify the analysis, we consider from now on n_b boats racing n_r races, plus an additional virtual boat which is only specified by its score $n_t \in [n_\mathrm{r}, n_\mathrm{r}(n_\mathrm{b}+1)]$. We are interested in finding the probability distribution for this virtual boat to have a final rank $m \in [1, n_\mathrm{b}+1]$ given its score n_t when it is compared to the set of scores $\{n_i; i = 1, 2, \ldots, n_\mathrm{b}\}$ of the n_b boats. By definition this probability distribution will then depend only on three variables: n_b, the number of boats; n_r, the number of races; and n_t, the score of the virtual boat that we are interested in.

3. The limit of many races

The problem simplifies when some of the parameters determining the size of the system become so large that we can use central limit-type results. In this section we consider the limit in which the number of races becomes large.

[4] The 1/2 discrepancy is due to the fact that boats with equal scores are all assigned the same final rank. For example, two boats tying in the first place are assigned a rank of 1, while the next boat would have a rank of 3. If, instead, the two top boats were assigned a rank of 1.5 (the average of 1 and 2) we would have obtained $\langle m \rangle = 2$. This effect, at any rate, will be important only for a small number of boats.

We start with a reminder of the central limit theorem in the case of correlated random variables. Assume $\{x_{i,k}; i = 1, \ldots, n_{\rm b}; k = 1, 2, \ldots, n_{\rm r}\}$ to be correlated random variables such that

- they are independent for different k,
- the set $\{x_{1,k}, x_{2,k}, \ldots, x_{n_{\rm b},k}\}$ is distributed according to a joint density probability distribution which is independent of k and whose first two moments (mean and covariance) are $\langle x_{i,k}\rangle = \rho_i$ and $\langle x_{i,k}x_{j,k}\rangle - \langle x_{i,k}\rangle\langle x_{j,k}\rangle = \rho_{ij}$.

The central limit theorem states that in the limit $n_{\rm r} \gg 1$ the summed variables $x_i = \sum_{k=1}^{n_{\rm r}} x_{i,k}$ are correlated Gaussian random variables with $\langle x_i\rangle = n_{\rm r}\rho_i$ and $\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle = n_{\rm r}\rho_{ij}$, that is, they are distributed in this limit according to the probability density

$$f(x_1, x_2, \ldots, x_{n_{\rm b}}) = N \exp\left[-\frac{1}{2n_{\rm r}}\sum_{i,j}\lambda_{ij}(x_i - n_{\rm r}\rho_i)(x_j - n_{\rm r}\rho_j)\right] \tag{1}$$

where N is a normalization constant. The matrix $[\lambda]$ is the inverse of the covariance matrix $[\rho]$, assuming that $[\rho]$ is non-singular.

In the race problem, $x_{i,k} = n_{i,k}$ and $x_i = n_i$: one has

$$\rho_i = \frac{n_{\rm b}+1}{2} \tag{2}$$

$$\rho_{ii} = \frac{n_{\rm b}^2 - 1}{12}, \qquad \rho_{ij} = -\frac{n_{\rm b}+1}{12} \quad (i \neq j) \tag{3}$$

(off diagonal correlations are negative), so

$$\rho_{ij} = \frac{n_{\rm b}+1}{12}(n_{\rm b}\delta_{i,j} - 1) \qquad i, j \in [1, \ldots, n_{\rm b}]. \tag{4}$$

It follows that in the large number of races limit $\langle n_i\rangle = n_{\rm r}(n_{\rm b}+1)/2$ and $\langle n_i n_j\rangle - \langle n_i\rangle\langle n_j\rangle = n_{\rm r}\rho_{ij}$.

The covariance matrix $[\rho]$ is singular with a single zero-eigenvalue eigenvector $(1, 1, \ldots, 1)$. Any vector perpendicular to $(1, 1, \ldots, 1)$, that is, such that the sum of its entries is 0, is an eigenvector with eigenvalue $n_{\rm r}(n_{\rm b}+1)/2$. The fact that $(1, 1, \ldots, 1)$ is a zero-eigenvalue eigenvector signals that the variable $\sum_{i=1}^{n_{\rm b}} n_i = n_{\rm r}n_{\rm b}(n_{\rm b}+1)/2$ is deterministic. It must be 'taken out' of the set of the scores before finding the large $n_{\rm r}$ limit. We arrive at the density probability distribution

$$f(n_1, \ldots, n_{n_{\rm b}}) = \sqrt{\frac{2\pi n_{\rm b}}{\lambda}}\left(\sqrt{\frac{\lambda}{2\pi}}\right)^{n_{\rm b}} \delta\left(\sum_{i=1}^{n_{\rm b}} n_i - \frac{6}{\lambda}\right) \exp\left[-\frac{\lambda}{2}\sum_{i=1}^{n_{\rm b}}\left(n_i - n_{\rm r}\frac{n_{\rm b}+1}{2}\right)^2\right] \tag{5}$$

with

$$\lambda = \frac{12}{n_{\rm r}n_{\rm b}(n_{\rm b}+1)} \tag{6}$$

such that we do indeed have $\langle n_i\rangle = n_{\rm r}\rho_i$ and $\langle n_i n_j\rangle - \langle n_i\rangle\langle n_j\rangle = n_{\rm r}\rho_{ij}$.

doi:10.1088/1742-5468/2011/01/P01025

J. Stat. Mech. (2011) P01025

One can exponentiate the constraint $\delta(\sum_{i=1}^{n_\mathrm{b}}(n_i - n_\mathrm{r}(n_\mathrm{b}+1)/2)$, so

$$f(n_1,\ldots,n_{n_\mathrm{b}}) = \sqrt{\frac{n_\mathrm{b}\lambda^{n_\mathrm{b}-1}}{(2\pi)^{n_\mathrm{b}+1}}} \times \int_{-\infty}^{\infty} \exp\left[-\mathrm{i}k\sum_{i=1}^{n_\mathrm{b}}\left(n_i - n_\mathrm{r}\frac{n_\mathrm{b}+1}{2}\right) - \frac{\lambda}{2}\sum_{i=1}^{n_\mathrm{b}}\left(n_i - n_\mathrm{r}\frac{n_\mathrm{b}+1}{2}\right)^2\right]\mathrm{d}k. \quad (7)$$

For a virtual boat with score n_t the probability of having a final rank m is the probability for $m-1$ boats among the total n_b of them to have a score $n_i < n_t$ and for the other $n_\mathrm{b}-m+1$ of them to have a score $n_i \geq n_t$:

$$P_{n_t}(m) = \binom{n_\mathrm{b}}{m-1}\int_{-\infty}^{n_t}\mathrm{d}n_1\cdots\mathrm{d}n_{m-1}\int_{n_t}^{\infty}\mathrm{d}n_m\cdots\mathrm{d}n_{n_\mathrm{b}} f(n_1,\ldots,n_{n_\mathrm{b}}) \quad (8)$$

which obviously satisfies $\sum_{m=1}^{n_\mathrm{b}+1} P_{n_t}(m) = 1$. It can be rewritten as

$$P_{n_t}(m) = \binom{n_\mathrm{b}}{m-1}\int_{-\infty}^{\infty} w_{n_t}(k)^{m-1}(1-w_{n_t}(k))^{n_\mathrm{b}-m+1}\sqrt{\frac{n_\mathrm{b}}{2\pi\lambda}}\exp\left[-\frac{n_\mathrm{b}k^2}{2\lambda}\right]\mathrm{d}k \quad (9)$$

where

$$w_{n_t}(k) = \sqrt{\frac{\lambda}{2\pi}}\int_{-\infty}^{n_t}\exp\left[-\frac{\lambda}{2}\left(n - n_\mathrm{r}\frac{n_\mathrm{b}+1}{2} + \frac{\mathrm{i}k}{\lambda}\right)^2\right]\mathrm{d}n. \quad (10)$$

If we further define

$$\bar{n}_t = \sqrt{\lambda}\left(n_t - \frac{n_\mathrm{r}(n_\mathrm{b}+1)}{2}\right) \quad (11)$$

and absorb $1/\sqrt{\lambda}$ in k, (9) becomes

$$P_{n_t}(m) = \binom{n_\mathrm{b}}{m-1}\int_{-\infty}^{\infty} w_{\bar{n}_t}(k)^{m-1}(1-w_{\bar{n}_t}(k))^{n_\mathrm{b}-m+1}\sqrt{\frac{n_\mathrm{b}}{2\pi}}\exp\left[-\frac{n_\mathrm{b}k^2}{2}\right]\mathrm{d}k \quad (12)$$

with

$$w_{\bar{n}_t}(k) = \sqrt{\frac{1}{2\pi}}\int_{-\infty}^{\bar{n}_t}\exp\left[-\frac{(n+\mathrm{i}k)^2}{2}\right]\mathrm{d}n. \quad (13)$$

The probability distribution (12) is of binomial form but with a k-dependent 'pseudo-probability' $w_{\bar{n}_t}(k)$, and k normally distributed according to $\sqrt{n_\mathrm{b}/(2\pi)}\exp[-n_\mathrm{b}k^2/2]$. We find in particular

$$\langle m\rangle = 1 + n_\mathrm{b}\int_{-\infty}^{\infty} w_{\bar{n}_t}(k)\sqrt{\frac{n_\mathrm{b}}{2\pi}}\exp\left[-\frac{n_\mathrm{b}k^2}{2}\right]\mathrm{d}k = 1 + n_\mathrm{b}\mathcal{N}\left(\bar{n}_t\sqrt{\frac{n_\mathrm{b}}{n_\mathrm{b}-1}}\right) \quad (14)$$

where $\mathcal{N}(x)$ is the cumulative probability distribution of a normal variable

$$\mathcal{N}(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left[-\frac{n^2}{2}\right]\mathrm{d}n. \quad (15)$$

We can go further by considering (12) in the large boat number limit $n_\mathrm{b} \gg 1$. In this limit, n_t scales like n_b and thus $\bar{n}_t$ is independent of n_b: the n_b dependence of $P_{n_t}(m)$

J. Stat. Mech. (2011) P01025

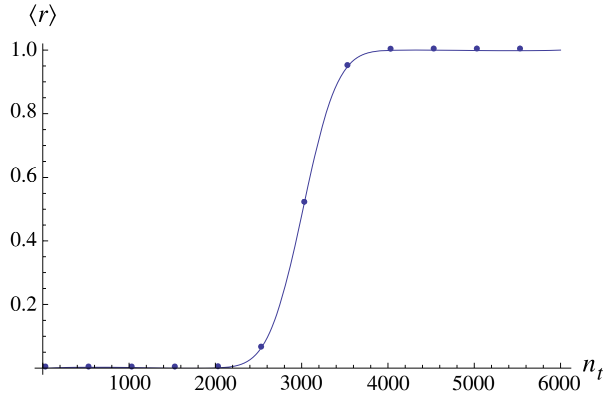

Figure 1. For 200 boats racing 30 races, the final percentage rank of the virtual boat: the continuous line is the final percentage rank $\langle r\rangle$ in (17) and the points are numerical simulations for a score n_t ranging from 30 to 6000 in steps of 500. The 'statistical curse (blessing) of the second (first) half-rank' effect is clearly visible.

is solely contained in the binomial coefficient and the exponents, not in $w_{\bar{n}_t}(k)$. Setting $r = m/n_\mathrm{b}$ (the final percentage rank) and using $n! \simeq \sqrt{2\pi n}(n/e)^n$ we obtain

$$P_{n_t}(r) = \frac{1}{\sqrt{2\pi r(1-r)}} \times \int_{-\infty}^{\infty} \exp\left[-n_\mathrm{b}\left(r\ln\frac{r}{w_{\bar{n}_t}(k)} + (1-r)\ln\frac{1-r}{1-w_{\bar{n}_t}(k)} + k^2/2\right)\right]\,\mathrm{d}k. \quad (16)$$

In (16) the exponent of the integrand is negative except when $k = 0$ and $r = w_{\bar{n}_t}(k)$: for large n_b a saddle point approximation yields that $P_{n_t}(r)$ vanishes except when r is taken to be $w_{\bar{n}_t}(0)$. It follows that the final rank of the virtual boat is essentially fixed by its score $\bar{n}_t$

$$\langle r\rangle = \mathcal{N}(\bar{n}_t) \quad (17)$$

as expected from (13), (14) in the large n_b limit and shown in figure 1 for 200 boats racing 30 races.

The fluctuations of r around $\langle r\rangle$ are obtained by expanding the exponent in (16) around $r = \langle r\rangle$ (one sets $r \simeq \langle r\rangle + \epsilon$) and around $k = 0$ as follows:

$$r\ln\frac{r}{w_{\bar{n}_t}(k)} + (1-r)\ln\frac{1-r}{1-w_{\bar{n}_t}(k)} \simeq \frac{(\epsilon - k w'_{\bar{n}_t}(0))^2}{2\langle r\rangle(1-\langle r\rangle)} \quad (18)$$

where $w'_{\bar{n}_t}(0)$ is the derivative of $w_{\bar{n}_t}(k)$ at $k = 0$. The integration over k in (16) finally yields

$$P_{n_t}(r) = \frac{1}{\sqrt{2\pi n_\mathrm{b}(\langle r\rangle(1-\langle r\rangle) + w'_{\bar{n}_t}(0)^2)}} \exp\left[-\frac{n_\mathrm{b}\epsilon^2}{2(\langle r\rangle(1-\langle r\rangle) + w'_{\bar{n}_t}(0)^2)}\right] \quad (19)$$

J. Stat. Mech. (2011) P01025

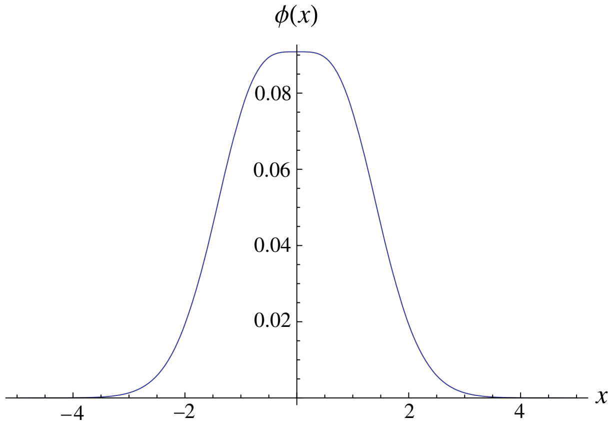

Figure 2. The Kollines function.

which is Gaussian distributed around $\epsilon = 0$, i.e. $r = \langle r \rangle$, with variance $(\langle r \rangle(1 - \langle r \rangle) + w'_{\bar{n}_t}(0)^2)/n_\mathrm{b}$. Since

$$w'_{\bar{n}_t}(k) = \mathrm{i}\sqrt{\frac{1}{2\pi}} \exp\left[-\frac{(\mathrm{i}k + \bar{n}_t)^2}{2}\right] \tag{20}$$

and $\langle r \rangle = w_{\bar{n}_t}(0)$, $1 - \langle r \rangle = 1 - w_{\bar{n}_t}(0) = w_{-\bar{n}_t}(0)$ we eventually get for the variance

$$(\Delta m)^2 = n_\mathrm{b}\, \phi(\bar{n}_t). \tag{21}$$

In the above we introduced the Kollines function

$$\phi(x) = \mathcal{N}(x)\mathcal{N}(-x) - \frac{1}{2\pi} \exp[-x^2]. \tag{22}$$

It is positive, very flat around $x = 0$ (the first three derivatives vanish at $x = 0$) and essentially zero when $|x| > 3.5$ (see figure 2).

It follows that when $|n_t - n_\mathrm{r}(n_\mathrm{b} + 1)/2| \gg 3.5/\sqrt{\lambda}$ ($\simeq 3.5 n_\mathrm{b}\sqrt{n_\mathrm{r}/12}$) the final rank has no fluctuation. It is only when $|n_t - n_\mathrm{r}(n_\mathrm{b} + 1)/2| < 3.5/\sqrt{\lambda}$ that $\Delta m \simeq \sqrt{n_\mathrm{b}}$, as illustrated in figure 3 for 200 boats racing 30 races.

As said above, (12) has a binomial form but with a k-dependent 'pseudo-probability', a direct consequence of the correlation of the random variables n_i as encoded in (5). A naive approach could have been to ignore the correlations and to use instead the standard central limit theorem for uncorrelated variables and, consequently, the standard binomial distribution. In this approach, focusing on any given boat, say boat i, among the n_b boats, the probability distribution for its score n_i should have been, instead of (5),

$$f(n_i) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(n_i - \langle n_i \rangle)^2\right] \tag{23}$$

where, from (2) and the diagonal part of (4), $\langle n_i \rangle = n_\mathrm{r}(n_\mathrm{b} + 1)/2$ and $\sigma^2 = n_\mathrm{r}(n_\mathrm{b}^2 - 1)/12$. The binomial law for the final rank of the virtual boat should have read, instead of (8),

$$P_{n_t}(m) = \binom{n_\mathrm{b}}{m-1} \left(\int_{-\infty}^{n_t} f(n)\, \mathrm{d}n\right)^{m-1} \left(1 - \int_{-\infty}^{n_t} f(n)\, \mathrm{d}n\right)^{n_\mathrm{b}-m+1} \tag{24}$$

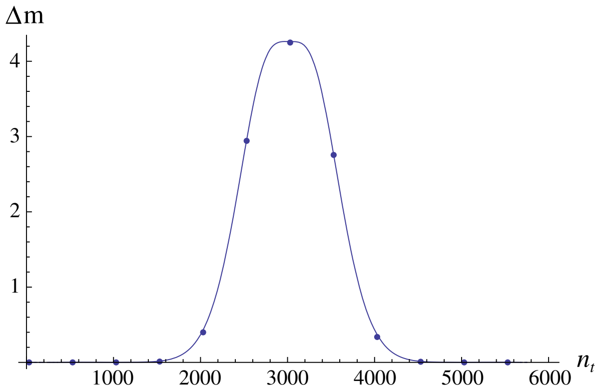

Figure 3. For 200 boats racing 30 races, the standard deviation of the final rank of the virtual boat: the continuous line is the square root of the Kollines function in (21) and the points are numerical simulations for a score n_t ranging from 30 to 6000 in steps of 500.

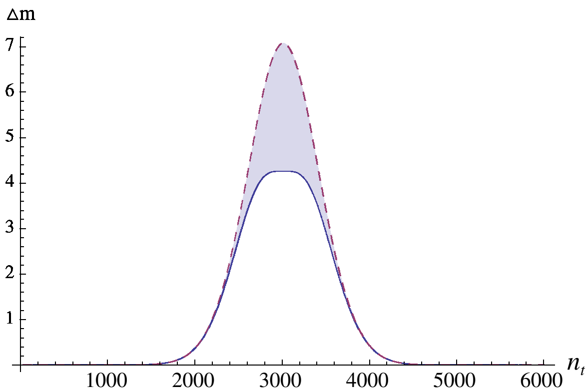

Figure 4. For 200 boats racing 30 races, the standard deviation of the final rank of the virtual boat with a score n_t ranging from 30 to 6000: the dashed line is the naive standard deviation as given in (26) and the continuous line represents the correct deviation as given by (21). The effect of correlations is to strongly decrease the fluctuations (shaded area).

and, therefore,

$$\langle m \rangle = 1 + n_{\rm b} \int_{-\infty}^{n_t} f(n)\, {\rm d}n = 1 + n_{\rm b} \mathcal{N}\left(\frac{n_t - \langle n_i \rangle}{\sigma} \right) \tag{25}$$

and

$$(\Delta m)^2 = n_{\rm b} \int_{-\infty}^{n_t} f(n)\, {\rm d}n \left(1 - \int_{-\infty}^{n_t} f(n)\, {\rm d}n \right) = n_{\rm b} \mathcal{N}\left(\frac{n_t - \langle n_i \rangle}{\sigma} \right) \mathcal{N}\left(- \frac{n_t - \langle n_i \rangle}{\sigma} \right). \tag{26}$$

When $n_{\rm b}$ is large, $(n_t - \langle n_i \rangle)/\sigma \simeq \bar{n}_t$, so, on the one hand, from (25), $\langle r \rangle = \langle m \rangle / n_{\rm b}$ coincides with (17), but, on the other hand, from (26), $(\Delta m)^2$ differs from (21): the effect of correlations essentially ends up decreasing the variance by an amount equal to $(1/2\pi) \exp[-(\bar{n}_t)^2]$ as shown in figure 4 for again 200 boats racing 30 races.

Table 1. By complete enumeration of all permutations: the mean final rank and variance for $3, 4, \ldots, 9$ boats and 2 races.

n_t	2	3	4	5	6	7												
$\langle m\rangle$	1	$\frac{4}{3}$	2	3	$\frac{11}{3}$	4												
$(\Delta m)^2$	0	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{9}$	0												
n_t	2	3	4	5	6	7	8	9										
$\langle m\rangle$	1	$\frac{5}{4}$	$\frac{7}{4}$	$\frac{5}{2}$	$\frac{7}{2}$	$\frac{17}{4}$	$\frac{19}{4}$	5										
$(\Delta m)^2$	0	$\frac{3}{16}$	$\frac{17}{48}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{17}{48}$	$\frac{3}{16}$	0										
n_t	2	3	4	5	6	7	8	9	10	11								
$\langle m\rangle$	1	$\frac{6}{5}$	$\frac{8}{5}$	$\frac{11}{5}$	3	4	$\frac{24}{5}$	$\frac{27}{5}$	$\frac{29}{5}$	6								
$(\Delta m)^2$	0	$\frac{4}{25}$	$\frac{17}{50}$	$\frac{23}{50}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{23}{50}$	$\frac{17}{50}$	$\frac{4}{25}$	0								
n_t	2	3	4	5	6	7	8	9	10	11	12	13						
$\langle m\rangle$	1	$\frac{7}{6}$	$\frac{3}{2}$	2	$\frac{8}{3}$	$\frac{7}{2}$	$\frac{9}{2}$	$\frac{16}{3}$	6	$\frac{13}{2}$	$\frac{41}{6}$	7						
$(\Delta m)^2$	0	$\frac{5}{36}$	$\frac{19}{60}$	$\frac{7}{15}$	$\frac{5}{9}$	$\frac{7}{12}$	$\frac{7}{12}$	$\frac{5}{9}$	$\frac{7}{15}$	$\frac{19}{60}$	$\frac{5}{36}$	0						
n_t	2	3	4	5	6	7	8	9	10	11	12	13	14	15				
$\langle m\rangle$	1	$\frac{8}{7}$	$\frac{10}{7}$	$\frac{13}{7}$	$\frac{17}{7}$	$\frac{22}{7}$	4	5	$\frac{41}{7}$	$\frac{46}{7}$	$\frac{50}{7}$	$\frac{53}{7}$	$\frac{55}{7}$	8				
$(\Delta m)^2$	0	$\frac{6}{49}$	$\frac{43}{147}$	$\frac{67}{147}$	$\frac{85}{147}$	$\frac{95}{147}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{95}{147}$	$\frac{85}{147}$	$\frac{67}{147}$	$\frac{43}{147}$	$\frac{6}{49}$	0				
n_t	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17		
$\langle m\rangle$	1	$\frac{9}{8}$	$\frac{11}{8}$	$\frac{7}{4}$	$\frac{9}{4}$	$\frac{23}{8}$	$\frac{29}{8}$	$\frac{9}{2}$	$\frac{11}{2}$	$\frac{51}{8}$	$\frac{57}{8}$	$\frac{31}{4}$	$\frac{33}{4}$	$\frac{69}{8}$	$\frac{71}{8}$	9		
$(\Delta m)^2$	0	$\frac{7}{64}$	$\frac{121}{448}$	$\frac{7}{16}$	$\frac{65}{112}$	$\frac{305}{448}$	$\frac{47}{64}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{47}{64}$	$\frac{305}{448}$	$\frac{65}{112}$	$\frac{7}{16}$	$\frac{121}{448}$	$\frac{7}{64}$	0		
n_t	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$\langle m\rangle$	1	$\frac{10}{9}$	$\frac{4}{3}$	$\frac{5}{3}$	$\frac{19}{9}$	$\frac{8}{3}$	$\frac{10}{3}$	$\frac{37}{9}$	5	6	$\frac{62}{9}$	$\frac{23}{3}$	$\frac{25}{3}$	$\frac{80}{9}$	$\frac{28}{3}$	$\frac{29}{3}$	$\frac{89}{9}$	10
$(\Delta m)^2$	0	$\frac{8}{81}$	$\frac{1}{4}$	$\frac{5}{12}$	$\frac{185}{324}$	$\frac{25}{36}$	$\frac{7}{9}$	$\frac{133}{162}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{133}{162}$	$\frac{7}{9}$	$\frac{25}{36}$	$\frac{185}{324}$	$\frac{5}{12}$	$\frac{1}{4}$	$\frac{8}{81}$	0

4. Small race number: the case $n_r = 2$

The problem without the benefit of the large-n_r limit becomes harder and, for generic n_r, is not amenable to an explicit solution. For the case of few races, however, we can obtain exact results.

In the present section we deal with the case $n_r = 2$, for which we can find the exact solution. Table 1 displays the mean final ranks and variances of the virtual boat for $n_b = 3$–9 boats racing two races. For a given n_b the score of the virtual boat spans the interval $[2, 2n_b + 1]$.

4.1. A sketch and basic properties

For two races, the situation can be sketched by using a $n_b \times n_b$ square lattice as in figure 5 for $n_b = 6$.

The two coordinates correspond to the ranks of a boat in each one of the two races. So, each boat will be represented by an occupied site. It follows that each line and each column will be occupied once and only once. This leads to $n_b!$ possible configurations.

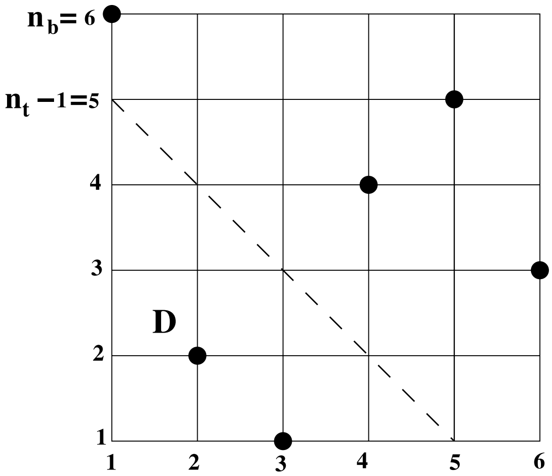

Figure 5. The sketch of an event for $n_{\rm r} = 2$ and $n_{\rm b} = 6$. A boat is represented by a point whose coordinates are its ranks in the two races. Here, we fix the score $n_t = 6$ of the virtual boat (dashed diagonal). There are two sites occupied in **D**. Thus, the rank of the virtual boat is $m = 3$ for this event.

The score n_t of the virtual boat is fixed and represented by the dashed diagonal. Let us name as **D** the domain under the diagonal. The rank of the virtual boat is equal to m when $(m-1)$ sites are occupied in **D**. We have obviously $P_{n_t}(m) = \delta_{m,1}$ when $n_t \leq 2$ and $P_{n_t}(m) = \delta_{m,n_{\rm b}+1}$ when $n_t \geq 2n_{\rm b} + 1$. Moreover, from symmetry considerations,

$$P_{n_{\rm b}+1-k}(m) = P_{n_{\rm b}+2+k}(n_{\rm b} + 2 - m), \qquad k = 0, 1, \ldots, n_{\rm b} - 1. \tag{27}$$

So, in the following, we will restrict n_t to the range $2 \leq n_t \leq n_{\rm b} + 1$. In that case, it is easy to realize that only $(n_t - 2)$ columns (or lines) are available in **D**. This implies for m the restriction $1 \leq m \leq n_t - 1$.

We also observe that the distribution is symmetric for $n_t = n_{\rm b} + 1$ or $n_{\rm b} + 2$

$$P_{n_{\rm b}+1}(m) = P_{n_{\rm b}+1}(n_{\rm b} + 1 - m) = P_{n_{\rm b}+2}(m + 1), \qquad m = 1, 2, \ldots, n_{\rm b}. \tag{28}$$

4.2. Direct computations of $P_{n_t}(m)$ for some m

For $m = n_t - 1$, there is only one possibility for occupying the $(n_t - 2)$ sites in **D** (see figure 6) The $(n_{\rm b} - n_t + 2)$ remaining occupied sites are distributed randomly on the sites of the $(n_{\rm b} - n_t + 2)$ remaining lines and columns that are still available. Defining $u \equiv n_{\rm b} - n_t + 2$, one obtains

$$P_{n_t}(m = n_t - 1) = \frac{u!}{n_{\rm b}!}. \tag{29}$$

Now, for $m = 1$, there are no occupied sites in D. Let us fill (figure 7) the lines, starting from the bottom. On line (a), we have $n_{\rm b} - n_t + 2$ ($\equiv u$) available sites; on line (b), we still have u available sites (because of the site occupied in line (a)); and so on, up to line (d). Moreover, from the u upper lines, we still get a factor $u!$.

Finally

$$P_{n_t}(1) = P_{n_t}(n_t - 1) \cdot \Phi_1(u) \qquad \text{with } \Phi_1(u) = u^{n_t - 2}. \tag{30}$$

J. Stat. Mech. (2011) P01025

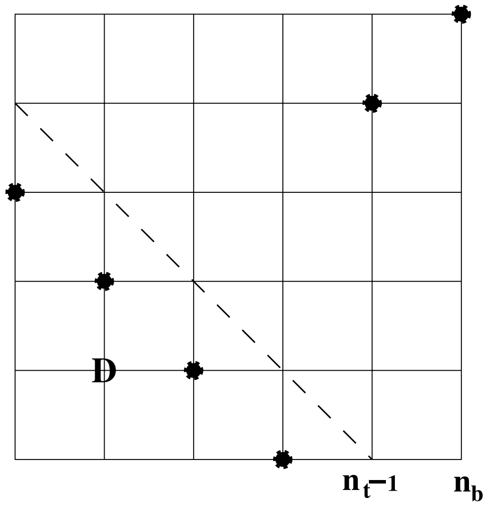

Figure 6. A configuration contributing to $P_{n_t}(m = n_t - 1)$. We have only one possibility for the $(n_t - 2)$ occupied sites under the dashed diagonal.

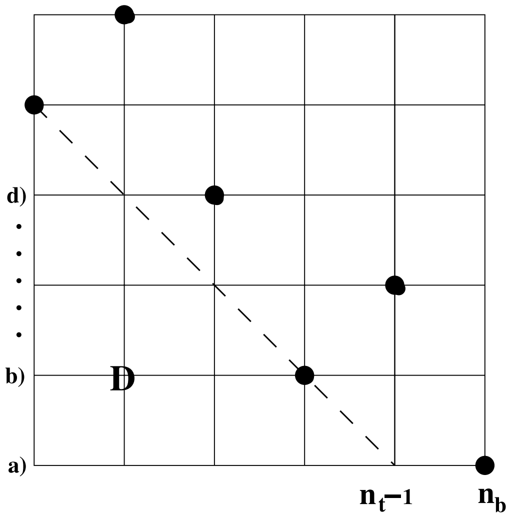

Figure 7. A configuration contributing to $P_{n_t}(1)$. No occupied site belongs to **D**. For each line (a), (b), ..., (d), we have $n_{\mathrm{b}} - n_t + 2$ possibilities for the occupied sites. The remaining occupied sites will generate the factor $P_{n_t}(n_t - 1)$. For further explanations, see the text.

It is easy to see, from the above considerations, that, for $1 \leq m \leq n_t - 1$,

$$P_{n_t}(m) = P_{n_t}(n_t - 1)\Phi_m(u) \tag{31}$$

where $\Phi_m(u)$ is a polynomial in u with integer values[5].

For $m = 2$, there is one occupied site, B, in **D**.

[5] This is not true for $n_{\mathrm{r}} \geq 3$.

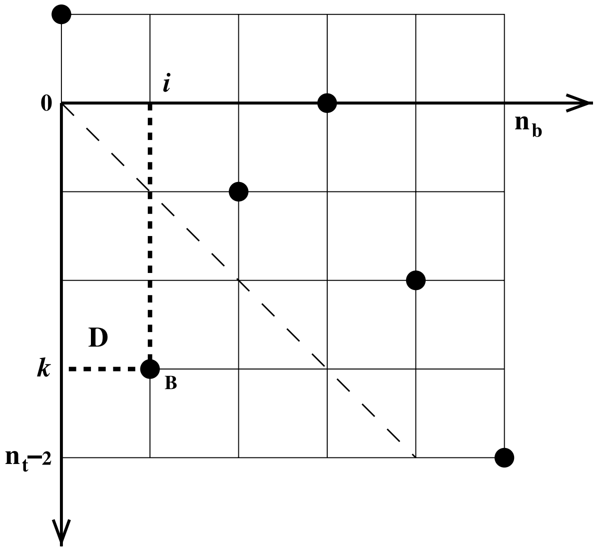

Figure 8. A configuration contributing to $P_{n_t}(2)$. The occupied site, B, in $\mathbf{D}$, has coordinates i and k. For further explanations, see the text.

With the coordinates (i,k) defined in figure 8, $\mathbf{D}$ is the domain $(0 \leq i \leq k-1; 1 \leq k \leq n_t-2)$, so

$$P_{n_t}(2) = P_{n_t}(n_t-1).\sum_{\mathbf{D}} u^{n_t-2-k}(u+1)^{k-i-1}u^i = P_{n_t}(n_t-1)\Phi_2(u) \tag{32}$$

with

$$\Phi_2(u) = (u+1)^{n_t-2}(u+1) - u^{n_t-2}(u+n_t-1). \tag{33}$$

The computation for $m=3$ is more involved because the relative position of the two occupied sites in $\mathbf{D}$ plays an important role in the expression for the terms to be summed. One gets

$$\Phi_3(u) = \tfrac{1}{2}[(u+2)^{n_t-2}(u+1)(u+2) - 2(u+1)^{n_t-2}(u+1)(u+n_t-1) \\ + u^{n_t-2}(u+n_t-1)(u+n_t-2)]. \tag{34}$$

It is worth noting that, despite the apparent complexity of $\Phi_3(u)$, the degree of $\Phi_m(u)$ decreases when m increases. We will clarify this point later.

The case $m=4$ seems out of reach for direct computation and will not be pursued along these lines.

4.3. The recursion relation and the solution for the case $n_r = 2$

Looking at (30), (33), (34), we observe that, for $m \leq 2$, $\Phi_m(u)$ satisfies the recursion relation

$$\Phi_{m+1}(u) = \frac{1}{m}((u+1)\Phi_m(u+1) - (u+n_t-m)\Phi_m(u)). \tag{35}$$

We will now show that (35) holds in general.

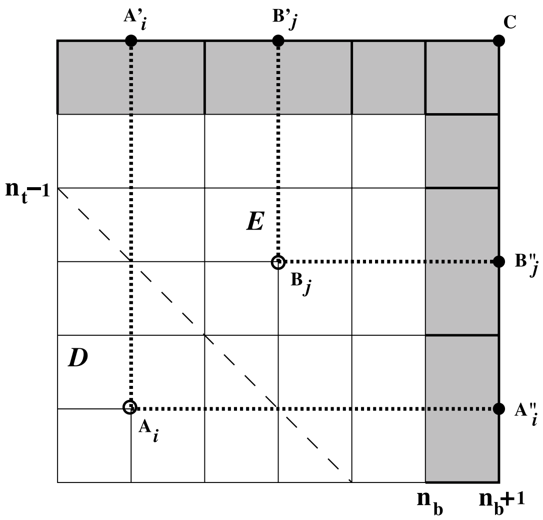

Figure 9. The three ways for producing a configuration contributing to $N'(m)$ (see the text for the definition): (i) start from a configuration contributing to $N(m+1)$, erase A_i and add A'_i and A''_i; (ii) start from a configuration contributing to $N(m)$, erase B_j and add B'_j and B''_j; (iii) start from a configuration contributing to $N(m)$ and add C.

Let us write $P_{n_t}(m) = (u!/n_{\rm b}!)\Phi_m(u) = N(m)/n_{\rm b}!$ where $N(m)$ is the number of configurations of the $n_{\rm b} \times n_{\rm b}$ square with $(m-1)$ occupied sites in $\mathbf{D}$. Changing $n_{\rm b}$ into $n_{\rm b}+1$ (which amounts to changing u into $u+1$ while keeping n_t unchanged), we call $P'_{n_t}(m)$ the new probability distribution $P'_{n_t}(m) = ((u+1)!/(n_{\rm b}+1)!)\Phi_m(u+1) = N'(m)/(n_{\rm b}+1)!$ where $N'(m)$ is defined like $N(m)$ but for the $(n_{\rm b}+1) \times (n_{\rm b}+1)$ square lattice (figure 9).

$N'(m)$ receives three contributions:

(i) Let us consider a configuration contributing to $N(m+1)$ (m occupied sites A_i in $\mathbf{D}$—see figure 9). The replacement of A_i by A'_i and A''_i produces a configuration contributing to $N'(m)$ (only $(m-1)$ occupied sites in $\mathbf{D}$; all the columns and lines of the biggest square are occupied once). Since we can choose any of the A_i s before applying this procedure, we get a contribution $mN(m+1)$ to $N'(m)$.

(ii) Let us next consider a configuration contributing to $N(m)$ ($n_{\rm b}+1-m$ occupied sites B_j in $\mathbf{E}$—see figure 9). By the same reasoning as in (i), we get $(n_{\rm b}+1-m)N(m)$ configurations for $N'(m)$.

(iii) To each configuration contributing to $N(m)$, we can add an occupied site in C (see figure 9). This produces the contribution $N(m)$ to $N'(m)$.

Summing the above contributions leads to

$$N'(m) = mN(m+1) + (n_{\rm b}+2-m)N(m). \tag{36}$$

Reverting back to the Φ_m s, it is straightforward to get (35). Equations (30) and (35) prove that $\Phi_m(u)$ has degree $n_t - m - 1$. Finally, solving the recursion equation, we get

J. Stat. Mech. (2011) P01025

the exact solution for $n_\mathrm{r} = 2$:

$$P_{n_t}(m) = (n_\mathrm{b}+1)\sum_{k=0}^{m-1}(-1)^k(n_\mathrm{b}-n_t+m-k+1)^{n_t-2}\frac{(n_\mathrm{b}-n_t+m-k+1)!}{k!(n_\mathrm{b}-k+1)!(m-k-1)!} \qquad (37)$$

with $2 \le m+1 \le n_t \le n_\mathrm{b}+1$ understood. We have checked (37) by a complete enumeration of the permutations up to n_b and $n_t = 10$.

Let us discuss the case $n_t = n_\mathrm{b}+1$. Equation (37) narrows down to

$$P_{n_t}(m) = n_t\sum_{k=0}^{m-1}(-1)^k\frac{(m-k)^{n_t-1}}{k!(n_t-k)!}. \qquad (38)$$

The moments are

$$\langle m^n\rangle = n_t\left(\frac{\partial}{\partial\lambda'}\right)^n\Big|_{\lambda'=0}\left(\frac{\partial}{\partial\lambda}\right)^{n_t-1}\Big|_{\lambda=0}\sum_{m=1}^{n_t-1}\sum_{k=0}^{m-1}\frac{(-1)^k}{k!(n_t-k)!}\mathrm{e}^{\lambda(m-k)}\mathrm{e}^{\lambda' m}$$
$$= \frac{1}{(n_t-1)!}\left(\frac{\partial}{\partial\lambda'}\right)^n\Big|_{\lambda'=0}\left(\frac{\partial}{\partial\lambda}\right)^{n_t-1}\Big|_{\lambda=0}\left[\frac{(1-\mathrm{e}^{\lambda'})^{n_t}-(\mathrm{e}^{\lambda+\lambda'}-\mathrm{e}^{\lambda'})^{n_t}}{1-\mathrm{e}^{\lambda+\lambda'}}\right] \qquad (39)$$

and in particular

$$\langle m\rangle = \frac{n_t}{2} \qquad (40)$$

$$\langle(m-\langle m\rangle)^2\rangle = \frac{n_t}{12} \qquad (41)$$

$$\langle(m-\langle m\rangle)^3\rangle = 0. \qquad (42)$$

We recover the fact that $P_{n_t}(m)$ is symmetric. These results will be especially useful in the next section.

4.4. Computations of the first three moments for $n_t \le n_\mathrm{b}$

Starting from equation (36), we get

$$mP_{n_t}(m+1) = (n_\mathrm{b}+1)P'_{n_t}(m) - (n_\mathrm{b}+2-m)P_{n_t}(m) \qquad (43)$$

(recall that $P'_{n_t}(m)$ is the same as $P_{n_t}(m)$ but for n_b changed into $n_\mathrm{b}+1$). Multiplying both sides of (43) by m^k and summing over m, the recursion equation for the moments follows:

$$(n_\mathrm{b}+1-k)\langle m^k\rangle + \sum_{p=0}^{k-1}\frac{(-1)^{k+1-p}(k+1)!}{p!(k+1-p)!}\langle m^p\rangle = (n_\mathrm{b}+1)\langle m^k\rangle' \qquad (44)$$

($\langle\cdots\rangle$ refers to n_b and $\langle\cdots\rangle'$ to $n_\mathrm{b}+1$).

For $k=1$, setting $Z_{n_\mathrm{b}} = n_\mathrm{b}\langle m\rangle$, we get $Z_{n_\mathrm{b}} - Z_{n_\mathrm{b}-1} = 1$ and, finally, $Z_{n_\mathrm{b}} = Z_{n_t-1} + n_\mathrm{b} - n_t + 1$. Computing Z_{n_t-1} with (40), we obtain the first moment

$$\langle m\rangle = 1 + \frac{(n_t-1)(n_t-2)}{2n_\mathrm{b}}. \qquad (45)$$

J. Stat. Mech. (2011) P01025

Table 2. By complete enumeration of all permutations: the mean final rank and variance for 3, 4, 5 and 6 boats and 3 races.

n_t	3	4	5	6	7	8	9	10									
$\langle m \rangle$	1	$\frac{10}{9}$	$\frac{13}{9}$	$\frac{19}{9}$	$\frac{26}{9}$	$\frac{32}{9}$	$\frac{35}{9}$	4									
$(\Delta m)^2$	0	$\frac{8}{81}$	$\frac{20}{81}$	$\frac{17}{81}$	$\frac{17}{81}$	$\frac{20}{81}$	$\frac{8}{81}$	0									
n_t	3	4	5	6	7	8	9	10	11	12	13						
$\langle m \rangle$	1	$\frac{17}{16}$	$\frac{5}{4}$	$\frac{13}{8}$	$\frac{9}{4}$	3	$\frac{15}{4}$	$\frac{35}{8}$	$\frac{19}{4}$	$\frac{79}{16}$	5						
$(\Delta m)^2$	0	$\frac{15}{256}$	$\frac{3}{16}$	$\frac{61}{192}$	$\frac{47}{144}$	$\frac{11}{36}$	$\frac{47}{144}$	$\frac{61}{192}$	$\frac{3}{16}$	$\frac{15}{256}$	0						
n_t	3	4	5	6	7	8	9	10	11	12	13	14	15	16			
$\langle m \rangle$	1	$\frac{26}{25}$	$\frac{29}{25}$	$\frac{7}{5}$	$\frac{9}{5}$	$\frac{12}{5}$	$\frac{78}{25}$	$\frac{97}{25}$	$\frac{23}{5}$	$\frac{26}{5}$	$\frac{28}{5}$	$\frac{146}{25}$	$\frac{149}{25}$	6			
$(\Delta m)^2$	0	$\frac{24}{625}$	$\frac{84}{625}$	$\frac{27}{100}$	$\frac{39}{100}$	$\frac{21}{50}$	$\frac{507}{1250}$	$\frac{507}{1250}$	$\frac{21}{50}$	$\frac{39}{100}$	$\frac{27}{100}$	$\frac{84}{625}$	$\frac{24}{625}$	0			
n_t	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$\langle m \rangle$	1	$\frac{37}{36}$	$\frac{10}{9}$	$\frac{23}{18}$	$\frac{14}{9}$	$\frac{71}{36}$	$\frac{23}{9}$	$\frac{13}{4}$	4	$\frac{19}{4}$	$\frac{49}{9}$	$\frac{217}{36}$	$\frac{58}{9}$	$\frac{121}{18}$	$\frac{62}{9}$	$\frac{251}{36}$	7
$(\Delta m)^2$	0	$\frac{35}{1296}$	$\frac{8}{81}$	$\frac{1733}{8100}$	$\frac{707}{2025}$	$\frac{14987}{32400}$	$\frac{1022}{2025}$	$\frac{601}{1200}$	$\frac{37}{75}$	$\frac{601}{1200}$	$\frac{1022}{2025}$	$\frac{14987}{32400}$	$\frac{707}{2025}$	$\frac{1733}{8100}$	$\frac{8}{81}$	$\frac{35}{1296}$	0

The other moments are obtained in a similar way. Equations (41), (42) and (44) lead to

$$\langle (m - \langle m \rangle)^2 \rangle = \frac{(n_t - 1)(n_t - 2)}{12 n_{\mathrm{b}}^2 (n_{\mathrm{b}} - 1)} \left[3n_t^2 - n_t(9 + 8n_{\mathrm{b}}) + 6(n_{\mathrm{b}} + 1)^2 \right], \qquad n_{\mathrm{b}} \geq 2 \tag{46}$$

$$\langle (m - \langle m \rangle)^3 \rangle = \frac{(n_t - 1)(n_t - 2)(n_t - n_{\mathrm{b}} - 1)^2 (n_t - n_{\mathrm{b}} - 2)^2}{2 n_{\mathrm{b}}^3 (n_{\mathrm{b}} - 1)(n_{\mathrm{b}} - 2)}, \qquad n_{\mathrm{b}} \geq 3. \tag{47}$$

As expected, $\langle (m - \langle m \rangle)^3 \rangle$ vanishes for $n_t = n_{\mathrm{b}} + 1$ or $n_{\mathrm{b}} + 2$ (the distribution is symmetric); $\langle (m - \langle m \rangle)^2 \rangle$ and $\langle (m - \langle m \rangle)^3 \rangle$ vanish for $n_t = 1$ or 2 ($P_{1,2}(m) = \delta_{m,1}$).

5. The case $n_{\mathrm{r}} \geq 3$

For the case of three or more races the problem is more involved. We can, however, establish some partial exact results. Table 2 displays in the case of three races for $n_{\mathrm{b}} = 3$–6 boats the mean final ranks and variances of the virtual boat (its score spans the interval $[3, 3n_{\mathrm{b}} + 1]$).

For $n_{\mathrm{r}} = 3$, one has to consider cubes instead of squares as was the case for $n_{\mathrm{r}} = 2$. The probability distribution should be rewritten now as $P_{n_t}(m) = N(m)/(n_{\mathrm{b}}!)^2$. Following the same line of reasoning as in section 4.3, one can find a recursion relation valid for $n_t \leq n_{\mathrm{b}} + 2$:

$$N'(m) = (m + 1)mN(m + 2) + m(2n_{\mathrm{b}} - 2m + 3)N(m + 1) + (n_{\mathrm{b}} - m + 2)^2 N(m). \tag{48}$$

The recursion (48) has been checked numerically by complete enumeration (see table 2).

More generally, for $n_{\mathrm{r}} \geq 3$, one obtains the expression

$$\langle m \rangle = 1 + \frac{(n_t - 1)!}{n_{\mathrm{b}}^{n_{\mathrm{r}} - 1} (n_t - 1 - n_{\mathrm{r}})! n_{\mathrm{r}}!} \qquad \text{for } n_{\mathrm{r}} \leq n_t \leq n_{\mathrm{b}} + n_{\mathrm{r}} - 1. \tag{49}$$

6. Two races with the worst individual rank dropped

We conclude our analysis with a variant of the original problem, also used in competitions, for the specific case of two races.

Specifically, suppose that, for each boat, we drop the greatest rank (worst result) obtained in the two races. For instance, if the boat i had ranks $n_{i,1} = 2$ and $n_{i,2} = 5$, we only retain the score $n_i = 2$. The virtual boat has a fixed score n_t in the range $[1, n_{\mathrm{b}} + 1]$ and, as before, its rank is m when $(m - 1)$ boats have scores n_i smaller than n_t.

It is obvious that $m \geq n_t$. Indeed, without loss of generality, we can consider that the ranks $n_{i,1}$ obtained in the first race are arranged in the natural order: $\{1, 2, \ldots, n_{\mathrm{b}} - 1, n_{\mathrm{b}}\}$, i.e. $n_{i,1} = i$. (We will keep this order throughout this section.) Now, from $n_i \leq n_{i,1}$, it is easy to realize that at least $(n_t - 1)$ boats will have scores n_i smaller than n_t; thus $m \geq n_t$.

Defining the ordered sets $A = \{1, 2, \ldots, n_t - 2, n_t - 1\}$ and $B = \{n_t, n_t + 1, \ldots, n_{\mathrm{b}} - 1, n_{\mathrm{b}}\}$, we see that, taking, for the ordered[6] set of ranks $r_{i,2}$ in the second race, any permutation of A (for instance $\{n_t - 2, 2, 1, \ldots, n_t - 1\}$) followed by any permutation of B (for instance $\{n_{\mathrm{b}} - 1, n_t, n_t + 1, \ldots, n_{\mathrm{b}}\}$), we construct all the configurations leading to $m = n_t$. The number of such configurations is $(n_t - 1)! \times (n_{\mathrm{b}} - n_t + 1)!$. Dividing by the total number of configurations $n_{\mathrm{b}}!$, we get

$$P_{n_t}(n_t) = \frac{1}{\dbinom{n_{\mathrm{b}}}{n_t - 1}}. \tag{50}$$

For $m > n_t$, we start from the naturally ordered sets A and B and exchange $(m - n_t)$ elements of A with $(m - n_t)$ elements of B (of course, $m - n_t \leq n_t - 1$ and $m - n_t \leq n_{\mathrm{b}} - n_t + 1$). So, we get the sets A' and B'. Taking, for the ordered set of ranks in the second race, any permutation of A' followed by any permutation of B', we get all the configurations leading to the rank m for the virtual boat. We eventually obtain a hypergeometric law for the random variable $(m - n_t)$:

$$P_{n_t}(m) = \frac{\dbinom{n_t - 1}{m - n_t}\dbinom{n_{\mathrm{b}} - n_t + 1}{m - n_t}}{\dbinom{n_{\mathrm{b}}}{n_t - 1}} \tag{51}$$

with $n_t \leq m \leq \min\{2n_t - 1, n_{\mathrm{b}} + 1\}$.

Of course, this probability density is quite different from the one obtained in (37). In particular, it is interesting to note that the distribution (51) is unchanged when we replace, simultaneously, n_t by $n'_t = n_{\mathrm{b}} + 2 - n_t$ and m by $m' = m + n_{\mathrm{b}} + 2 - 2n_t$:

$$P_{n'_t}(m') = P_{n_t}(m). \tag{52}$$

(Note that $n'_t - 1 = n_{\mathrm{b}} + 1 - n_t$, $n_{\mathrm{b}} - n'_t + 1 = n_t - 1$ and $m' - n'_t = m - n_t$. So, from (51), $P_{n_t}(m)$ is unchanged.)

[6] Here, 'ordered' does not mean 'in the natural order' but simply that we take into account the order when we enumerate the elements of the set (i.e., $\{a, b, \ldots\} \neq \{b, a, \ldots\}$).

J. Stat. Mech. (2011) P01025

When $n_{\rm b}$ is even, the distribution is symmetric for $n_t = (n_{\rm b}/2) + 1$. Indeed

$$P_{(n_{\rm b}/2)+1}(m) = \frac{\binom{\frac{n_{\rm b}}{2}}{m - \frac{n_{\rm b}}{2} - 1}^2}{\binom{n_{\rm b}}{\frac{n_{\rm b}}{2}}} = P_{(n_{\rm b}/2)+1}\left(\frac{3n_{\rm b}}{2} + 2 - m\right), \qquad (53)$$

$$\frac{n_{\rm b}}{2} + 1 \le m \le n_{\rm b} + 1.$$

The moments of (51) are

$$\langle m \rangle = n_t + \frac{(n_t - 1)(n_{\rm b} - n_t + 1)}{n_{\rm b}} \qquad (54)$$

$$\langle (m - \langle m \rangle)^2 \rangle = \frac{(n_t - 1)^2 (n_{\rm b} - n_t + 1)^2}{n_{\rm b}^2 (n_{\rm b} - 1)}, \qquad n_{\rm b} \ge 2 \qquad (55)$$

$$\langle (m - \langle m \rangle)^3 \rangle = -\frac{(n_t - 1)^2 (n_{\rm b} - n_t + 1)^2 (n_{\rm b} - 2n_t + 2)^2}{n_{\rm b}^3 (n_{\rm b} - 1)(n_{\rm b} - 2)}, \qquad n_{\rm b} \ge 3 \qquad (56)$$

consistent with (52). Moreover, as expected, $\langle (m - \langle m \rangle)^2 \rangle$ and $\langle (m - \langle m \rangle)^3 \rangle$ vanish for $n_t = 1$ and $n_t = n_{\rm b} + 1$. Finally, $\langle (m - \langle m \rangle)^3 \rangle$ vanishes for $n_t = (n_{\rm b}/2) + 1$ when $n_{\rm b}$ is even (the distribution is symmetric; see (53)).

7. Conclusions

We demonstrated that the problem of determining the final rank distribution for a boat in a set of races given its total score can be explicitly solved in two distinct situations: for a large number of races, and for a few (two or three) races. We also demonstrated that the 'statistical curse of the second half-rank' effect can be attributed to statistical averaging in the case of many races.

Although we obtained our results in the context and language of boat racing, they are clearly applicable in several similar situations, such as for student ranks based on their results in many exams or quizzes, ranks of candidates for positions or awards when they are reviewed and ranked by many independent evaluators, and voting results when voters submit a rank of the choices.

There are many open issues and unsolved problems for further investigation. The exact result for an arbitrary number of races (greater than 2) is not known. Further, the results obtained are based on the simplifying assumption that all boats are equally worthy (all ranks in each race are equally probable). One could examine the situation in which boats have *a priori* different inherent worths, handicapping the probabilities for the ranks, and see to what extent the 'statistical curse' effect also emerges here.

Finally, the relevance of our results to and relation of our results with well-known difficulties in rank situations, such as Arrow's impossibility theorem [3, 4], would be an interesting topic for further investigation. Indeed, we can consider each race as a 'voter' and the rank of the boats in each race as the preference list of the voter. The conflict in Arrow's theorem arises from trying to adopt a social preference ('constitution') that respects three reasonable but, in fact, excessive requirements; namely transitivity, unanimity and independence of 'irrelevant alternatives'. The assignment of a total ranking, in our case, as the sum of the rankings produces a unique and unambiguous constitution, which obviously respects unanimity and transitivity but violates the 'irrelevant alternatives' condition.

An interesting question would be that of whether the presence of a large number of voters with correlated preferences, as in the case of our races, essentially resolves the issue. Specifically, can we identify in that limit a nontrivial and non-biased constitution that has a vanishingly small probability of violating the third condition (that is, the measure of violations as a fraction of the total space of statistical alternatives goes to zero)? And, in particular, is the sum of rankings such a constitution in the large number of voters limit?

Acknowledgments

SO acknowledges useful conversations with H Hilhorst, M Leborgne, S Mashkevich and S Matveenko, and thanks the City College of New York for hospitality during the completion of this work. AP acknowledges the hospitality of CNRS and Université Paris-Sud during the initial stages of this work. The research of AP is supported by an NSF grant.

References

[1] Feller W, 1968 *An Introduction to Probability Theory and its Applications* 3rd edn, vol 1 (New York: Wiley) p 229
Feller W, 1971 *An Introduction to Probability Theory and its Applications* 3rd edn, vol 2 (New York: Wiley)
[2] Kallenberg O, 1997 *Foundations of Modern Probability* (New York: Springer)
[3] Arrow K J, 1951 *Social Choice and Individual Values* (New York: Wiley)
[4] Geannakoplos J, *Three brief proofs of Arrow's impossibility theorem*, 2005 *Econ. Theory* **26** 211