Statistical Curse of the Recurrent Second Half Rank

Jean Desbois[1,2], Stéphane Ouvry[1,2,3], Alexios Polychronakos[4]
[1] Université Paris-Sud, Laboratoire de Physique Théorique et Modèles Statistiques UMR 8626, F-91405 Orsay
[2] CNRS, LPTMS, F-91405 Orsay
[3] J-Venture Numerical Research Group, La Trinité sur Mer, Brittany
[4] Physics Department, City College of New York, NY

October 19, 2010

Abstract

In competitions involving many participants running many races the final rank is determined by the score of each participant, obtained by adding its ranks in each individual race. The "Statistical Curse of the Recurrent 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 formulae 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.

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. The typical example is boat racing, which can involve a large number of boats ($\sim$100) running a somehow 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 Recurrent 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 resuts for a small number of races are presented. A variant of the problem, in which the worst score for each participant is dropped, is also considered and solved for the case of two races.

2 Basic setup

Consider n_b boats racing n_r races. A boat i in the race k has an individual rank $n_{i,k} \in [1, n_b]$ (lower ranks represent better performance). The score of the boat i is the sum $n_i = \sum_{k=1}^{n_r} n_{i,k} \in [n_r, n_r n_b]$ 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 exaequo (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 $\{n_{ik}; i = 1, 2, ..., n_b\}$ is a random permutation of $\{1, 2, ..., n_b\}$ so that the $n_{i,k}$'s are correlated random variables (in particular $\sum_{i=1}^{n_b} n_{i,k} = n_b(n_b + 1)/2$), while n_{ik} and $n_{jk'}$ are uncorrelated for $k \neq k'$. We are interested in the probability distribution for boat i to have a final rank $m \in [1, n_b]$ 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 $\{1, 2, 3\}$ both for 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 $n_{1,2} = 2$, and
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 $m = 2$. 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 to 2. The 1/2 discrepancy is due to the fact that boats with equal total scores are all assigned the same higher final rank. E.g., 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.

More interestingly, cases (i), (ii) and (iii) give the same final rank probability distribution. This is particular to two races and would not be true any more for three or more races. The final rank probability distribution for a boat given its rank in each race would depend in this case on the full set of its ranks, and not just its total score. The final rank probability distribution given only its score would then be the average of the above distributions for all ranks consistent with its total score.

To simplify slightly the analysis, we consider from now on the set of n_b boats racing n_r races, plus an additional virtual boat with a given score $n_t \in [n_r, n_r(n_b+1)]$. We are interested in finding the probability for the virtual boat to have a final rank $m \in [1, n_b+1]$ given its score n_t compared to the set of scores $\{n_i; i = 1, 2, ..., n_b\}$ of the n_b boats.

3 The limit of many races

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

We start with a reminder of the Central Limit Theorem in the case of correlated random variables. Assume $\{x_{i,k}; i = 1, \ldots, n_b; k = 1, 2, \ldots, n_r\}$ to be correlated random variables such that

- they are independent for different k,
- the set $\{x_{1,k}, x_{2,k}, ..., x_{n_b,k}\}$ is distributed according to a joint density probablility distribution which is k-independent 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 CLT states that in the limit $n_r \gg 1$ the summed variables $x_i = \sum_{k=1}^{n_r} x_{i,k}$ are correlated gaussian random variables with $\langle x_i\rangle = n_r\rho_i$ and $\langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle = n_r\rho_{ij}$, that is, they are distributed in this limit according to the probability density

$$f(x_1, x_2, ..., x_{n_b}) = N \exp[-\frac{1}{2n_r}\sum_{i,j}\lambda_{ij}(x_i - n_r\rho_i)(x_j - n_r\rho_j)] \tag{1}$$

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

$$\rho_i = \frac{n_b+1}{2} \tag{2}$$

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

(off diagonal correlations are negative) so that

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

It follows that in the large number of races limit $\langle n_i\rangle = n_r\frac{n_b+1}{2}$ and $\langle n_i n_j\rangle - \langle n_i\rangle\langle n_j\rangle = n_r\rho_{ij}$.

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

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

with

$$\lambda = \frac{12}{n_r n_b(n_b+1)} \tag{6}$$

such that indeed $\langle n_i\rangle = n_r\rho_i$ and $\langle n_i n_j\rangle - \langle n_i\rangle\langle n_j\rangle = n_r\rho_{ij}$.

One can exponentiate the constraint $\delta(\sum_{i=1}^{n_b}(n_i - n_r(n_b+1)/2)$ so that

$$f(n_1,...,n_{n_b}) = \sqrt{\frac{n_b\lambda^{n_b-1}}{(2\pi)^{n_b+1}}}\int_{-\infty}^{\infty}\exp\left[-ik\sum_{i=1}^{n_b}(n_i - n_r\frac{n_b+1}{2}) - \frac{\lambda}{2}\sum_{i=1}^{n_b}(n_i - n_r\frac{n_b+1}{2})^2\right]dk \tag{7}$$

For a virtual boat with score n_t the probability to have a final rank m is the probability for $m-1$ boats among the n_b's to have a score $n_i < n_t$ and for the other $n_b - m + 1$'s to have a score $n_i \geq n_t$

$$P_{n_t}(m) = \binom{n_b}{m-1}\int_{-\infty}^{n_t} dn_1\dots dn_{m-1}\int_{n_t}^{\infty} dn_m\dots dn_{n_b} f(n_1,\dots,n_{n_b}) \tag{8}$$

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

$$P_{n_t}(m) = \binom{n_b}{m-1}\int_{-\infty}^{\infty} w_{n_t}(k)^{m-1}(1-w_{n_t}(k))^{n_b-m+1}\sqrt{\frac{n_b}{2\pi\lambda}}\exp\left[-\frac{n_b k^2}{2\lambda}\right]dk \tag{9}$$

where

$$w_{n_t}(k) = \sqrt{\frac{\lambda}{2\pi}}\int_{-\infty}^{n_t}\exp\left[-\frac{\lambda}{2}(n - n_r\frac{n_b+1}{2} + \frac{ik}{\lambda})^2\right]dn \tag{10}$$

If we further define

$$\bar{n}_t = \sqrt{\lambda}\left(n_t - \frac{n_r(n_b+1)}{2}\right) \tag{11}$$

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

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

with

$$w_{\bar n_t}(k) = \sqrt{\frac{1}{2\pi}} \int_{-\infty}^{\bar n_t} \exp\left[-\frac{(n+ik)^2}{2}\right] dn \tag{13}$$

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 $\sqrt{n_b/(2\pi)}\exp[-n_b k^2/2]$. We find in particular

$$\langle m\rangle = 1 + n_b \int_{-\infty}^{\infty} w_{\bar n_t}(k)\sqrt{\frac{n_b}{2\pi}} \exp\left[-\frac{n_b k^2}{2}\right] dk = 1 + n_b \mathcal{N}\left(\bar n_t \sqrt{\frac{n_b}{n_b-1}}\right) \tag{14}$$

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

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

We can go further by considering (12) in the large boat number limit $n_b \gg 1$. In this limit, n_t scales like n_b and thus $\bar n_t$ is n_b-independent: the n_b dependence of $P_{n_t}(m)$ is solely contained in the binomial coefficient and the exponents, not in $w_{\bar n_t}(k)$. Setting $r = m/n_b$ (the percentage rank) and using $n! \simeq \sqrt{2\pi n}(n/e)^n$ we obtain

$$P_{n_t}(r) = \frac{1}{\sqrt{2\pi r(1-r)}} \int_{-\infty}^{\infty} \exp\left[-n_b\left(r\ln\frac{r}{w_{\bar n_t}(k)} + (1-r)\ln\frac{1-r}{1-w_{\bar n_t}(k)} + k^2/2\right)\right] dk \tag{16}$$

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 $\bar n_t$

$$\bar r = \mathcal{N}(\bar n_t) \tag{17}$$

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

The fluctuations of r around $\bar r$ are obtained by expanding the exponent in (16) around $r = \bar r$ (one sets $r \simeq \bar r + \epsilon$) and around $k = 0$ so that

$$r\ln\frac{r}{w_{\bar n_t}(k)} + (1-r)\ln\frac{1-r}{1-w_{\bar n_t}(k)} \simeq \frac{(\epsilon - k w'_{\bar n_t}(0))^2}{2\bar r(1-\bar r)} \tag{18}$$

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

$$P_{n_t}(r) = \frac{1}{\sqrt{2\pi n_b(\bar r(1-\bar r) + w'_{\bar n_t}(0)^2)}} \exp\left[-\frac{n_b\epsilon^2}{2(\bar r(1-\bar r) + w'_{\bar n_t}(0)^2)}\right] \tag{19}$$

which is gaussian distributed around $\epsilon = 0$, i.e. $r = \bar r$, with variance $\frac{\bar r(1-\bar r)+w'_{\bar n_t}(0)^2}{n_b}$. Since

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

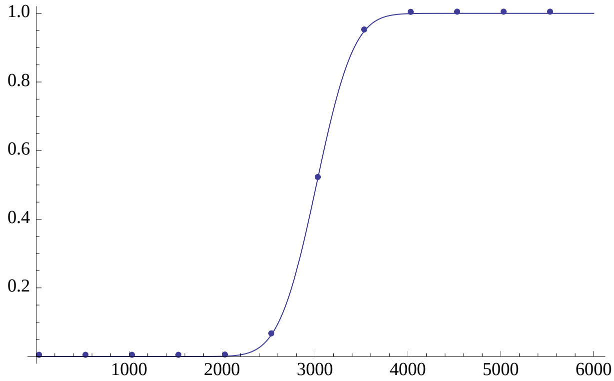

Figure 1: The final rank of the virtual boat among 200 boats racing 30 races : the continuous line is (17) and the points are numerical simuations for a score n_t ranging from 30 to 6000 by steps of 500. Both data in the curve and the simulation points have been divided by 200.

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

$$(\Delta m)^2 = n_b\, \phi(\bar{n}_t) \tag{21}$$

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 is essentially zero when $|x| > 3.5$ (see Fig. 2).

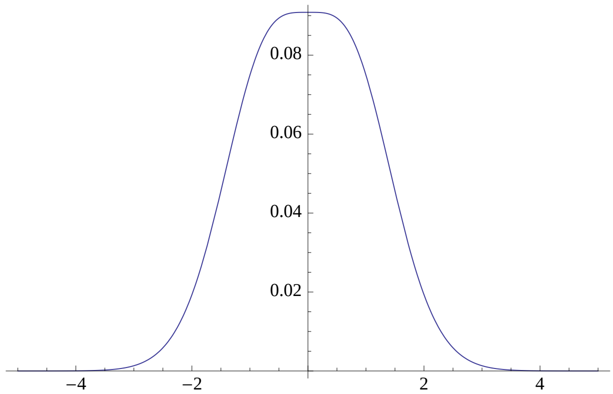

Figure 2: The Kollines function

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

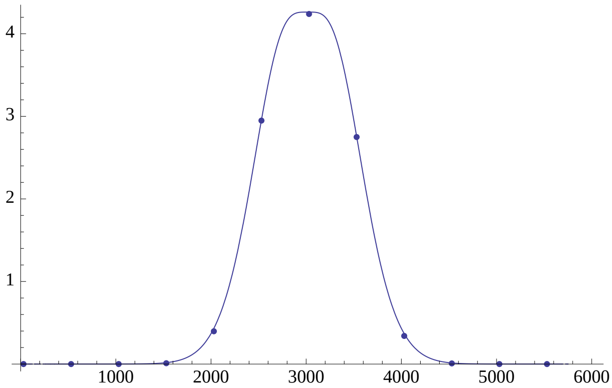

Figure 3: The standard deviation of the final rank of the virtual boat among 200 boats racing 30 races : the continuous line is the square root of the Kollines function and the points are numerical simuations for a score n_t ranging from 30 to 6000 by steps of 500.

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. Fig. 4 displays the mean final ranks and variances of the virtual boat among $n_b = 3, 4, .., 9$ boats racing 2 races. For a given n_b the score of the virtual boat spans the interval $[2, 2n_b + 1]$.

$$\begin{pmatrix} 1 & \frac{4}{3} & 2 & 3 & \frac{11}{3} & 4 \\ 0 & \frac{2}{9} & \frac{1}{3} & \frac{1}{3} & \frac{2}{9} & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{5}{4} & \frac{7}{4} & \frac{5}{2} & \frac{7}{2} & \frac{17}{4} & \frac{19}{4} & 5 \\ 0 & \frac{3}{16} & \frac{17}{48} & \frac{5}{12} & \frac{5}{12} & \frac{17}{48} & \frac{3}{16} & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{6}{5} & \frac{8}{5} & \frac{11}{5} & 3 & 4 & \frac{24}{5} & \frac{27}{5} & \frac{29}{5} & 6 \\ 0 & \frac{4}{25} & \frac{17}{50} & \frac{23}{50} & \frac{1}{2} & \frac{1}{2} & \frac{23}{50} & \frac{17}{50} & \frac{4}{25} & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{7}{6} & \frac{3}{2} & 2 & \frac{8}{3} & \frac{7}{2} & \frac{9}{2} & \frac{16}{3} & 6 & \frac{13}{2} & \frac{41}{6} & 7 \\ 0 & \frac{5}{36} & \frac{19}{60} & \frac{7}{15} & \frac{5}{9} & \frac{7}{12} & \frac{7}{12} & \frac{5}{9} & \frac{7}{15} & \frac{19}{60} & \frac{5}{36} & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{8}{7} & \frac{10}{7} & \frac{13}{7} & \frac{17}{7} & \frac{22}{7} & 4 & 5 & \frac{41}{7} & \frac{46}{7} & \frac{50}{7} & \frac{53}{7} & \frac{55}{7} & 8 \\ 0 & \frac{6}{49} & \frac{43}{147} & \frac{67}{147} & \frac{85}{147} & \frac{95}{147} & \frac{2}{3} & \frac{2}{3} & \frac{95}{147} & \frac{85}{147} & \frac{67}{147} & \frac{43}{147} & \frac{6}{49} & 0 \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{1} & \frac{\mathbf{9}}{\mathbf{8}} & \frac{\mathbf{11}}{\mathbf{8}} & \frac{\mathbf{7}}{\mathbf{4}} & \frac{\mathbf{9}}{\mathbf{4}} & \frac{\mathbf{23}}{\mathbf{8}} & \frac{\mathbf{29}}{\mathbf{8}} & \frac{\mathbf{9}}{\mathbf{2}} & \frac{\mathbf{11}}{\mathbf{2}} & \frac{\mathbf{51}}{\mathbf{8}} & \frac{\mathbf{57}}{\mathbf{8}} & \frac{\mathbf{31}}{\mathbf{4}} & \frac{\mathbf{33}}{\mathbf{4}} & \frac{\mathbf{69}}{\mathbf{8}} & \frac{\mathbf{71}}{\mathbf{8}} & \mathbf{9} \\ \mathbf{0} & \frac{\mathbf{7}}{\mathbf{64}} & \frac{\mathbf{121}}{\mathbf{448}} & \frac{\mathbf{7}}{\mathbf{16}} & \frac{\mathbf{65}}{\mathbf{112}} & \frac{\mathbf{305}}{\mathbf{448}} & \frac{\mathbf{47}}{\mathbf{64}} & \frac{\mathbf{3}}{\mathbf{4}} & \frac{\mathbf{3}}{\mathbf{4}} & \frac{\mathbf{47}}{\mathbf{64}} & \frac{\mathbf{305}}{\mathbf{448}} & \frac{\mathbf{65}}{\mathbf{112}} & \frac{\mathbf{7}}{\mathbf{16}} & \frac{\mathbf{121}}{\mathbf{448}} & \frac{\mathbf{7}}{\mathbf{64}} & \mathbf{0} \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{10}{9} & \frac{4}{3} & \frac{5}{3} & \frac{19}{9} & \frac{8}{3} & \frac{10}{3} & \frac{37}{9} & 5 & 6 & \frac{62}{9} & \frac{23}{3} & \frac{25}{3} & \frac{80}{9} & \frac{28}{3} & \frac{29}{3} & \frac{89}{9} & 10 \\ 0 & \frac{8}{81} & \frac{1}{4} & \frac{5}{12} & \frac{185}{324} & \frac{25}{36} & \frac{7}{9} & \frac{133}{162} & \frac{5}{6} & \frac{5}{6} & \frac{133}{162} & \frac{7}{9} & \frac{25}{36} & \frac{185}{324} & \frac{5}{12} & \frac{1}{4} & \frac{8}{81} & 0 \end{pmatrix}$$

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

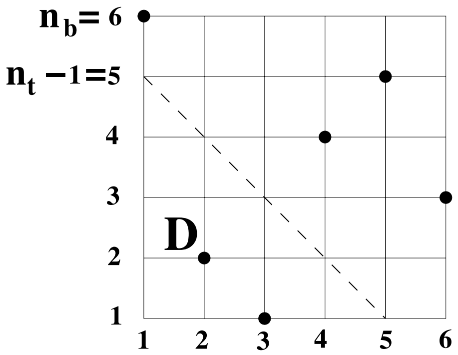

Figure 5: The sketch of an event for $n_r = 2$ and $n_b = 6$. A boat is represented by a point whose coordinates are its ranks in the two races. Here, we fix the score $n_t = 6$ of the virtual boat (dashed diagonal). There are 2 sites occupied in **D**. Thus, the rank of the virtual boat is $m = 3$ for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

4.1 Sketch and basic properties

For two races, the situation can be sketched by using a $n_b \times n_b$ square lattice as in Fig. 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.

The score n_t of the virtual boat is fixed and represented by the dashed diagonal. Let us call **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 $P_{n_t}(m) = \delta_{m,n_b+1}$ when $n_t \geq 2n_b + 1$. Moreover, from symmetry considerations,

$$P_{n_b+1-k}(m) = P_{n_b+2+k}(n_b + 2 - m) \ , \qquad k = 0, 1, ..., n_b - 1 \tag{23}$$

So, in the following, we will restrict n_t to the range $2 \leq n_t \leq n_b + 1$. 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 $n_t = n_b + 1$ or $n_b + 2$

$$P_{n_b+1}(m) = P_{n_b+1}(n_b + 1 - m) = P_{n_b+2}(m + 1) \ , \qquad m = 1, 2, ..., n_b \tag{24}$$

We will come back to this point later.

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

For $m = n_t - 1$, we observe (see Fig. 6) that there is only one possibility to occupy the $(n_t - 2)$ sites in **D**.
The $(n_b - n_t + 2)$ remaining occupied sites are distributed randomly on the sites of the $(n_b - n_t + 2)$ remaining lines and columns that are still available. Denoting $(u \equiv n_b - n_t + 2)$

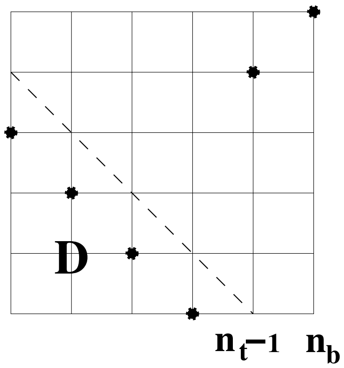

Figure 6: A configuration contributing to $P_{n_t}(m = n_t - 1)$. We have only one possibility for the $(n_t - 2)$ occupied sites under the dashed diagonal.

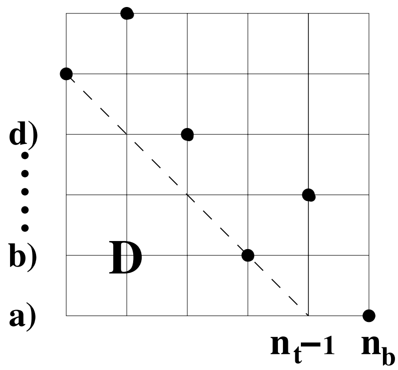

Figure 7: A configuration contributing to $P_{n_t}(1)$. No occupied site belongs to **D**. For each line a), b), ..., d), we have $n_b - n_t + 2$ possibilities for the occupied sites. The remaining occupied sites will generate the factor $P_{n_t}(n_t - 1)$. For further explanations, see the text.

one obtains

$$P_{n_t}(m = n_t - 1) = \frac{u!}{n_b!} \tag{25}$$

Now, for $m = 1$, there are no occupied sites in D. Let us fill (Fig. 7) the lines, starting from the bottom. On line (a), we have $n_b - n_t + 2$ ($\equiv u$) 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

$$P_{n_t}(1) = P_{n_t}(n_t - 1) \cdot \Phi_1(u) \qquad \text{with} \qquad \Phi_1(u) = u^{n_t - 2} \tag{26}$$

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

$$P_{n_t}(m) = P_{n_t}(n_t - 1)\Phi_m(u) \tag{27}$$

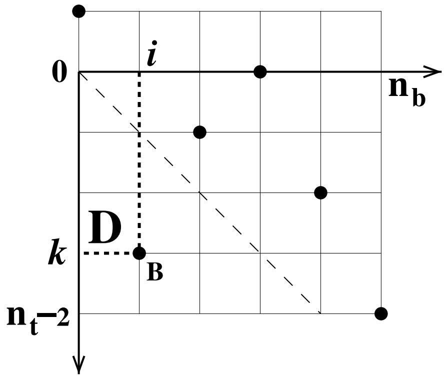

Figure 8: A configuration contributing to $P_{n_t}(2)$. The occupied site, B, in **D**, has coordinates i and k. For further explanations, see the text.

where $\Phi_m(u)$ is a polynomial in u with integer values[1].

For $m = 2$, there is one occupied site, B, in **D**.
With the coordinates (i, k) defined in Fig. 8, **D** is the domain $(0 \leq i \leq k-1;\ 1 \leq k \leq n_t - 2)$ so that

$$P_{n_t}(2) = P_{n_t}(n_t - 1). \sum_{\mathbf{D}} u^{n_t-2-k}(u+1)^{k-i-1}u^i = P_{n_t}(n_t - 1)\Phi_2(u) \tag{28}$$

with

$$\Phi_2(u) = (u+1)^{n_t-2}(u+1) - u^{n_t-2}(u + n_t - 1) \tag{29}$$

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

$$\Phi_3(u) = \frac{1}{2}\Big[(u+2)^{n_t-2}(u+1)(u+2) - 2(u+1)^{n_t-2}(u+1)(u+n_t-1) + \\ + u^{n_t-2}(u+n_t-1)(u+n_t-2)\Big] \tag{30}$$

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 by direct computation and will not be pursued along these lines.

4.3 Recursion relation and solution of the case $n_r = 2$

Looking at (26, 29, 30), we observe that, for $m \leq 2$, $\Phi_m(u)$ satisfies the recursion relation

$$\Phi_{m+1}(u) = \frac{1}{m}\Big((u+1)\Phi_m(u+1) - (u + n_t - m)\Phi_m(u)\Big) \tag{31}$$

[1]This is not true for $n_r \geq 3$.

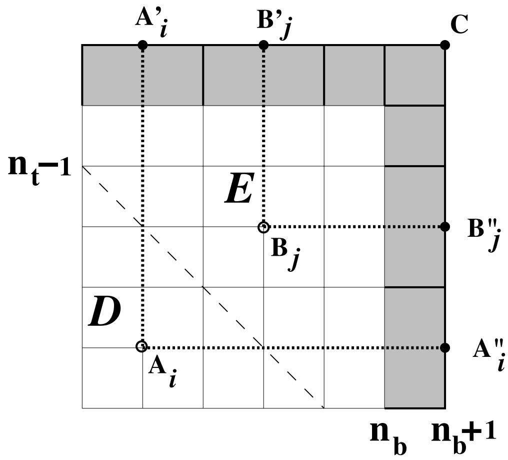

Figure 9: The 3 ways for producing a configuration contributing to $N'(m)$ (see the text for definition): i) start from a configuration contributing to $N(m+1)$, erase A_i and add A'_i and A''_i; ii) start from a configuration contributing to $N(m)$, erase B_j and add B'_j and B''_j; iii) start from a configuration contributing to $N(m)$ and add C.

We will now show that (31) holds in general.

Let us write $P_{n_t}(m) = \frac{u!}{n_b!}\Phi_m(u) = \frac{N(m)}{n_b!}$ where $N(m)$ is the number of configurations of the $n_b \times n_b$ square with $(m-1)$ occupied sites in **D**. Changing n_b into n_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 $P'_{n_t}(m) = \frac{(u+1)!}{(n_b+1)!}\Phi_m(u+1) = \frac{N'(m)}{(n_b+1)!}$ where $N'(m)$ is defined like $N(m)$ but for the $(n_b+1)\times(n_b+1)$ square lattice (Fig. 9).

$N'(m)$ receives three kinds of contributions:

i) Let us consider a configuration contributing to $N(m+1)$ (m occupied sites A_i in **D** – see Fig. 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 **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_b+1-m occupied sites B_j in **E** – see Fig. 9). By the same reasoning as in (i), we get $(n_b+1-m)N(m)$ configurations for $N'(m)$.

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

$$N'(m) = mN(m+1) + (n_b+2-m)N(m) \tag{32}$$

Reverting back to Φ_m's, it is straightforward to get (31). Equations (26) and (31) prove that $\Phi_m(u)$ has degree n_t-m-1.

Finally, solving the recursion equation, we get the exact solution for $n_r = 2$

$$P_{n_t}(m) = (n_b+1)\sum_{k=0}^{m-1}(-1)^k(n_b-n_t+m-k+1)^{n_t-2}\frac{(n_b-n_t+m-k+1)!}{k!(n_b-k+1)!(m-k-1)!} \quad (33)$$

with $2 \leq m+1 \leq n_t \leq n_b+1$ understood. We have checked (33) by a complete enumeration of the permutations up to n_b and $n_t = 10$.

Let us discuss the case $n_t = n_b + 1$. Equation (33) narrows down to

$$P_{n_t}(m) = n_t \sum_{k=0}^{m-1}(-1)^k \frac{(m-k)^{n_t-1}}{k!(n_t-k)!} \quad (34)$$

The moments are

$$\begin{aligned}\langle m^n\rangle &= n_t\left(\frac{\partial}{\partial\lambda'}\right)^n\Big|_{\lambda'=0}\left(\frac{\partial}{\partial\lambda}\right)^{n_t-1}\Big|_{\lambda=0}\sum_{m=1}^{n_t-1}\sum_{k=0}^{m-1}\frac{(-1)^k}{k!(n_t-k)!}e^{\lambda(m-k)}e^{\lambda' m}\\ &= \frac{1}{(n_t-1)!}\left(\frac{\partial}{\partial\lambda'}\right)^n\Big|_{\lambda'=0}\left(\frac{\partial}{\partial\lambda}\right)^{n_t-1}\Big|_{\lambda=0}\left[\frac{(1-e^{\lambda'})^{n_t}-(e^{\lambda+\lambda'}-e^{\lambda'})^{n_t}}{1-e^{\lambda+\lambda'}}\right] \quad (35)\end{aligned}$$

and in particular

$$\langle m\rangle = \frac{n_t}{2} \quad (36)$$

$$\langle(m-\langle m\rangle)^2\rangle = \frac{n_t}{12} \quad (37)$$

$$\langle(m-\langle m\rangle)^3\rangle = 0 \quad (38)$$

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 $n_t \leq n_b$

Starting from the equation (32), we get

$$mP_{n_t}(m+1) = (n_b+1)P'_{n_t}(m) - (n_b+2-m)P_{n_t}(m) \quad (39)$$

(recall that $P'_{n_t}(m)$ is the same as $P_{n_t}(m)$ but for n_b changed into n_b+1). Multipying both sides of (39) by m^k and summing over m, the recursion equation for the moments follows

$$(n_b+1-k) < m^k > + \sum_{p=0}^{k-1}\frac{(-1)^{k+1-p}(k+1)!}{p!(k+1-p)!} < m^p >= (n_b+1) < m^k >' \quad (40)$$

($< ... >$ refers to n_b and $< ... >'$ to $n_b + 1$).

For $k = 1$, setting $Z_{n_b} = n_b < m >$, we get $Z_{n_b} - Z_{n_b-1} = 1$ and, finally, $Z_{n_b} = Z_{n_t-1} + n_b - n_t + 1$. Computing Z_{n_t-1} with (36), we obtain the first moment

$$< m >= 1 + \frac{(n_t - 1)(n_t - 2)}{2n_b} \tag{41}$$

The other moments are obtained in a similar way. Equations (37), (38) and (40) lead to:

$$< (m- < m >)^2 > = \frac{(n_t - 1)(n_t - 2)}{12n_b^2(n_b - 1)} \left[3n_t^2 - n_t(9 + 8n_b) + 6(n_b + 1)^2\right], \quad n_b \geq 2 \tag{42}$$

$$< (m- < m >)^3 > = \frac{(n_t - 1)(n_t - 2)(n_t - n_b - 1)^2(n_t - n_b - 2)^2}{2n_b^3(n_b - 1)(n_b - 2)}, \qquad n_b \geq 3 \tag{43}$$

As expected, $< (m- < m >)^3 >$ vanishes for $n_t = n_b + 1$ or $n_b + 2$ (the distribution is symmetric); $< (m- < m >)^2 >$ and $< (m- < m >)^3 >$ vanish for $n_t = 1$ or 2 ($P_{1,2}(m) = \delta_{m,1}$).

5 The case $n_r \geq 3$

For the case of three or more races the problem is more complex. We can, however, establish some partial exact results. Fig. 10 demonstrates the stituation for three races, displaying the mean final ranks and variances of the virtual boat among $n_b = 3, 4, 5, 6$ boats. The score of the virtual boat spans the interval $[3, 3n_b + 1]$.

For $n_r = 3$ and $n_t \leq n_b + 2$, we established and checked numerically the recursion relation

$$N'(m) = (m + 1)mN(m + 2) + m(2n_b - 2m + 3)N(m + 1) + (n_b - m + 2)^2N(m) \tag{44}$$

More generally, for $n_r \geq 3$, we obtained the expression

$$< m >= 1 + \frac{(n_t - 1)!}{n_b^{n_r-1}(n_t - 1 - n_r)!n_r!} \qquad \text{for} \qquad n_r \leq n_t \leq n_b + n_r - 1 \tag{45}$$

$$\begin{pmatrix} 1 & \frac{10}{9} & \frac{13}{9} & \frac{19}{9} & \frac{26}{9} & \frac{32}{9} & \frac{35}{9} & 4 \\ 0 & \frac{8}{81} & \frac{20}{81} & \frac{17}{81} & \frac{17}{81} & \frac{20}{81} & \frac{8}{81} & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{17}{16} & \frac{5}{4} & \frac{13}{8} & \frac{9}{4} & 3 & \frac{15}{4} & \frac{35}{8} & \frac{19}{4} & \frac{79}{16} & 5 \\ 0 & \frac{15}{256} & \frac{3}{16} & \frac{61}{192} & \frac{47}{144} & \frac{11}{36} & \frac{47}{144} & \frac{61}{192} & \frac{3}{16} & \frac{15}{256} & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{26}{25} & \frac{29}{25} & \frac{7}{5} & \frac{9}{5} & \frac{12}{5} & \frac{78}{25} & \frac{97}{25} & \frac{23}{5} & \frac{26}{5} & \frac{28}{5} & \frac{146}{25} & \frac{149}{25} & 6 \\ 0 & \frac{24}{625} & \frac{84}{625} & \frac{27}{100} & \frac{39}{100} & \frac{21}{50} & \frac{507}{1250} & \frac{507}{1250} & \frac{21}{50} & \frac{39}{100} & \frac{27}{100} & \frac{84}{625} & \frac{24}{625} & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \frac{37}{36} & \frac{10}{9} & \frac{23}{18} & \frac{14}{9} & \frac{71}{36} & \frac{23}{9} & \frac{13}{4} & 4 & \frac{19}{4} & \frac{49}{9} & \frac{217}{36} & \frac{58}{9} & \frac{121}{18} & \frac{62}{9} & \frac{251}{36} & 7 \\ 0 & \frac{35}{1296} & \frac{8}{81} & \frac{1733}{8100} & \frac{707}{2025} & \frac{14\,987}{32\,400} & \frac{1022}{2025} & \frac{601}{1200} & \frac{37}{75} & \frac{601}{1200} & \frac{1022}{2025} & \frac{14\,987}{32\,400} & \frac{707}{2025} & \frac{1733}{8100} & \frac{8}{81} & \frac{35}{1296} & 0 \end{pmatrix}$$

Figure 10: By complete enumeration of all permutations: the mean final rank and variance for 3, 4, 5 and 6 boats and 3 races.

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 $[1, n_b + 1]$ 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 natural order: $\{1, 2, ..., n_b - 1, n_b\}$, ie $n_{i,1} = i$. (We will keep this order all along 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 $A = \{1, 2, ..., n_t - 2, n_t - 1\}$ and $B = \{n_t, n_t + 1, ..., n_b - 1, n_b\}$, we see that, taking, for the ordered[2] set of ranks $r_{i,2}$ in the second race, any permutation of A (for instance $\{n_t - 2, 2, 1, ..., n_t - 1\}$) followed by any permutation of B (for instance $\{n_b - 1, n_t, n_t + 1, ..., n_b\}$), we construct all the configurations leading to $m = n_t$. The number of such configurations is $(n_t - 1)! \times (n_b - n_t + 1)!$. Dividing by the total number of configurations $n_b!$, we get:

$$P_{n_t}(n_t) = \frac{1}{\begin{pmatrix} n_b \\ n_t - 1 \end{pmatrix}} \tag{46}$$

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 $m - n_t \leq n_b - n_t + 1$). 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 $(m - n_t)$

$$P_{n_t}(m) = \frac{\begin{pmatrix} n_t - 1 \\ m - n_t \end{pmatrix} \begin{pmatrix} n_b - n_t + 1 \\ m - n_t \end{pmatrix}}{\begin{pmatrix} n_b \\ n_t - 1 \end{pmatrix}} \tag{47}$$

with $n_t \leq m \leq \min\{2n_t - 1,\ n_b + 1\}$

Of course, this probability density is quite different from the one obtained in (33). In particular, it is interesting to note that the distribution (47) is unchanged when we replace, simultaneously, n_t by $n'_t = n_b + 2 - n_t$ and m by $m' = m + n_b + 2 - 2n_t$

$$P_{n'_t}(m') = P_{n_t}(m) \tag{48}$$

(Note that $n'_t - 1 = n_b + 1 - n_t$, $n_b - n'_t + 1 = n_t - 1$ and $m' - n'_t = m - n_t$. So, from (47), $P_{n_t}(m)$ is unchanged.)

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

When n_b is even, the distribution is symmetric for $n_t = \frac{n_b}{2} + 1$. Indeed

$$P_{\frac{n_b}{2}+1}(m) = \frac{\binom{\frac{n_b}{2}}{m - \frac{n_b}{2} - 1}^2}{\binom{n_b}{\frac{n_b}{2}}} = P_{\frac{n_b}{2}+1}\left(\frac{3n_b}{2} + 2 - m\right) \ , \qquad \frac{n_b}{2} + 1 \leq m \leq n_b + 1 \quad (49)$$

The moments of (47) are

$$< m > = n_t + \frac{(n_t - 1)(n_b - n_t + 1)}{n_b} \quad (50)$$

$$< (m- < m >)^2 > = \frac{(n_t - 1)^2 (n_b - n_t + 1)^2}{n_b^2 (n_b - 1)} \ , \qquad n_b \geq 2 \quad (51)$$

$$< (m- < m >)^3 > = -\frac{(n_t - 1)^2 (n_b - n_t + 1)^2 (n_b - 2n_t + 2)^2}{n_b^3 (n_b - 1)(n_b - 2)} \ , \qquad n_b \geq 3 \quad (52)$$

consistent with (48). Moreover, as expected, $< (m- < m >)^2 >$ and $< (m- < m >)^3 >$ vanish for $n_t = 1$ and $n_t = n_b + 1$. Finally, $< (m- < m >)^3 >$ vanishes for $n_t = \frac{n_b}{2} + 1$ when n_b is even (the distribution is symmetric, see (49)).

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 (2 or 3) races. We also demonstrated that the "Statistical Curse of the Recurrent 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, e.g., student ranks based on their results in many exams or quizes, rank 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 obtained results 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. Finally, the relevance 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.

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

References

[1] Feller, W. An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd ed. New York: Wiley, p. 229, 1968; Vol. 2, 3rd ed. New York: Wiley, 1971.

[2] Kallenberg, O. Foundations of Modern Probability. New York: Springer-Verlag, 1997.

[3] Arrow, K.J. : Social choice and individual values. New York: Wiley, 1951

[4] Geannakoplos J., Three Brief Proofs of Arrow's Impossibility Theorem, Economic Theory 26, 211, 2005