Statistical Curse of the Second Half Rank

Jean Desbois, Alexios Polychronakos, S.O.

JSTAT 2011 P01025

and some recent developments

a problem from real life which can lead to a pretty much involved combinatorics : ranking expectations in sailing boats regattas

example : the Spi Ouest France at la Trinité sur Mer (Brittany, each Easter)

involve a "large" number of identical boats $n_b \sim 100$

running a "large" number of races $n_r \sim 10 = 2, 3$ races per day during 4 days


FRA
504
ESP 1291

in each race each boat gets a rank $1 \leq \text{rank} \leq 100$

no equal rank (no ex-aequo)

how to determine the final rank of a boat (and thus the winner) :

1) for each boat add its ranks in each race $\rightarrow$ its score n_t

here $n_b = 100$ and $n_r = 10 \Rightarrow 10 \leq n_t \leq 1000$

$n_t = 10 \rightarrow$ lowest score always 1^{rst}

$n_t = 1000 \rightarrow$ highest score always 100^{th}

$n_t = 10 \times 50 = 500 \rightarrow$ middle score

2) order the scores $\Rightarrow$ final rank :

the boat with lowest score $\Rightarrow$ winner 1^{rst}

the next boat after the winner $\Rightarrow$ second 2^{nd}

$\ldots$

what is the problem ?

for example consider the ranks of a given boat to be

$51, 67, 76, 66, 55, 39, 67, 59, 66, 54 \to$ its score $n_t = 600$

clearly this boat has a mean rank $\frac{600}{10} = 60$

$\to$ on average it has been 60^{th}

$\to$ one might naively expect its final rank to be around 60^{th}

no way : its final rank will rather be around 70^{th} $\to$ ”curse”

see Spi Ouest 2009 data :

60	496.00	3 J *X. Bourrut Lacouture*	53.00	60.00	35.00	58.00	66.00	(69.00)	62.00	63.00	36.00	63.00
61	499.00	Atout Nautisme *M. Bolou*	52.00	(91.00)	67.00	42.00	48.00	26.00	91.00	49.00	91.00	33.00
62	500.00	Bmw Sailing Cup N°8 *B. Le Rossignol*	28.00	76.00	76.00	(78.00)	57.00	15.00	61.00	74.00	46.00	67.00
63	501.00	Jeroboam Marine Lorient *X. Bonvarlet*	55.00	64.00	52.00	65.00	49.00	53.00	52.00	60.00	51.00	(73.00)
64	509.00	Icam - Olac *A. Dal*	58.00	68.00	45.00	70.00	64.00	48.00	60.00	61.00	(75.00)	35.00
65	509.00	Bmw Sailing Cup N°11 *R. Lebohec*	68.00	55.00	65.00	57.00	(91.00)	40.00	55.00	65.00	60.00	44.00
66	510.00	Ste Morbihannaise de Navigation *H. Dubois*	46.00	62.00	38.00	76.00	(91.00)	37.00	73.00	48.00	72.00	58.00
67	515.00	Bmw Sailing Cup N°12 *M. Dolle*	(91.00)	57.00	68.00	61.00	29.00	61.00	54.00	58.00	67.00	60.00
68	518.00	J' Marine - Marine Lorient *G. Lautredou*	22.00	65.00	56.00	67.00	54.00	(75.00)	66.00	57.00	63.00	68.00
69	521.00	Cholet *C. Bore*	63.00	59.00	51.00	(68.00)	61.00	63.00	64.00	53.00	48.00	59.00
70	524.00	J-Venture *M. Le Borgne*	51.00	67.00	(76.00)	66.00	55.00	39.00	67.00	59.00	66.00	54.00
71	543.00	Bmw Sailing Cup N°2 *O. Tarle*	50.00	74.00	50.00	64.00	39.00	72.00	68.00	70.00	56.00	(91.00)
72	553.00	Penac'h *B.Jaud*	66.00	61.00	(73.00)	72.00	56.00	65.00	56.00	56.00	57.00	64.00
73	560.00	Ymir Junior *H. Schilling*	45.00	66.00	64.00	56.00	53.00	44.00	68.00	73.00	(91.00)	91.00
74	564.00	Art & Stamps *G. Le Baud*	77.00	70.00	(88.00)	71.00	72.00	77.00	59.00	40.00	58.00	40.00
75	571.00	Marine Lorient *P. Coindreau*	47.00	(91.00)	49.00	77.00	52.00	67.00	75.00	77.00	62.00	65.00
76	608.00	Denis Pelfresne *N. Barre*	46.00	63.00	78.00	80.00	70.00	(85.00)	74.00	66.00	61.00	70.00
77	630.00	J'Mini *A. Ponsar*	(91.00)	73.00	81.00	83.00	32.00	82.00	63.00	67.00	74.00	75.00
78	645.00	Jade Hisse Cn Pornic 1 *R. Romano*	54.00	78.00	75.00	75.00	67.00	73.00	79.00	(91.00)	64.00	80.00
79	653.00	Mazda *G. Tarin*	70.00	77.00	72.00	62.00	69.00	76.00	(83.00)	72.00	76.00	79.00
80	655.00	Ldt	57.00	56.00	83.00	69.00	(91.00)	78.00	70.00	91.00	73.00	78.00

NB :

$51, 67, (76), 66, 55, 39, 67, 59, 66, 54$

implies that the highest rank (76) is not taken into account

$\Rightarrow$ 9 races : $51, 67, \quad\quad , 66, 55, 39, 67, 59, 66, 54 \rightarrow$ score $n_t = 524$

$\Rightarrow$ mean rank $\frac{524}{9} = 58$

on average $58^{\text{th}} \rightarrow 70^{\text{th}}$ even worse a qualitative explanation of this ”curse” is simple :

given the ranks of the boat : $51, 67, 76, 66, 55, 39, 67, 59, 66, 54$

assume that the ranks of the other boats are random variables with uniform distribution

random ranks : a good assumption if the crews are more or less equally worthy (which is in part the case)

since no ex aequo it means :

ranks of the other boats = a random permutation

in the first race : random permutation of $(1, 2, 3, ..., 50, 52, ..., 100)$

in the second race : random permutation of $(1, 2, 3, ..., 66, 68, ..., 100)$

$\ldots$

each race is obviously independent from the others

$\rightarrow$ a score is a sum of 10 independent random variables

10 is already a large number in probability calculus :

$\rightarrow$ Central Limit Theorem applies

$\rightarrow$ scores are random variables with gaussian probability density centered around the middle score $10 \times 50 = 500$

gaussian distribution $\Rightarrow$ a lot a boats with scores packed around 500

if the score of a boat is > 500
its final rank is pushed upward from its mean rank
$\Rightarrow$ **statistical ”curse”**

on the contrary if the score of a boat is < 500
its final rank is pushed downward from its mean rank
$\Rightarrow$ **statistical ”blessing”**

write things more precisely : namely given the score n_t of a boat what is the probability distribution $P_{n_t}(m)$ for its final rank to be m ?

a complication : $P_{n_t}(m)$ does not depend only on the score n_t of the boat but also on its ranks in each race

for example : $n_r = 3$, $n_b = 3$ with a boat with score $n_t = 6$

it is very easy to check by complete enumeration that
$P_{6=2+2+2}(m) \neq P_{6=1+2+3}(m)$ (distributions are similar but different)

$\rightarrow$ a simplification : consider n_b boats with random ranks

i.e ranks = random permutation of $(1, 2, 3, \ldots, n_b)$

$\oplus$ an additional/virtual boat specified only by its score n_t

$\rightarrow$ same question : given the score n_t of a virtual boat what is the probability distribution $P_{n_t}(m)$ for its final rank to be m ?

$\rightarrow$ almost the same but simpler call $n_{i,k}$ rank of the boat i in a given race k ($1 \leq i \leq n_b$ and $1 \leq k \leq n_r$)

$$\langle n_{i,k} \rangle = \frac{1+n_b}{2}$$

no ex-aequo in race k : $\Rightarrow$ the $n_{i,k}$'s are correlated random variables

sum rule $$\sum_{i=1}^{n_b} n_{i,k} = 1+2+3+\ldots+n_b = \frac{n_b(1+n_b)}{2}$$

$$\langle n_{i,k} n_{j,k} \rangle - \langle n_{i,k} \rangle \langle n_{j,k} \rangle = \frac{1+n_b}{12}(n_b \delta_{i,j} - 1)$$

$n_{i,k} \Rightarrow$ score of boat $i = \sum_{k=1}^{n_r} n_{i,k} \equiv n_i$ and middle score $= n_r \frac{1+n_b}{2}$

large n_r limit $\rightarrow$ Central Limit Theorem for correlated random variables

$\Rightarrow$ joint density probability distribution

$$f(n_1, \ldots, n_{n_b}) =$$

$$\sqrt{2\pi\lambda n_b}\left(\sqrt{\frac{1}{2\pi\lambda}}\right)^{n_b} \delta\left(\sum_{i=1}^{n_b}(n_i - n_r\frac{1+n_b}{2})\right) \exp\left[-\frac{1}{2\lambda}\sum_{i=1}^{n_b}(n_i - n_r\frac{1+n_b}{2})^2\right]$$

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

for a virtual boat with score n_t :

$P_{n_t}(m)$ 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 $n_i \geq n_t$

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

take also large number of boats limit $\rightarrow$ saddle point approximation to finally get $\langle m \rangle$ = cumulative probability distribution of a normal variable

$$\langle m \rangle = \frac{n_b}{\sqrt{2\pi\lambda}} \int_{-\infty}^{\bar{n}_t} \exp\left[-\frac{n^2}{2\lambda}\right] dn$$

$$\bar{n}_t = n_t - n_r \frac{(1+n_b)}{2}$$

$$n_r \leq n_t \leq n_r n_b \rightarrow -n_r \frac{n_b}{2} \leq \bar{n}_t \leq n_r \frac{n_b}{2}$$

$\langle r \rangle = \frac{\langle m \rangle}{n_b}$

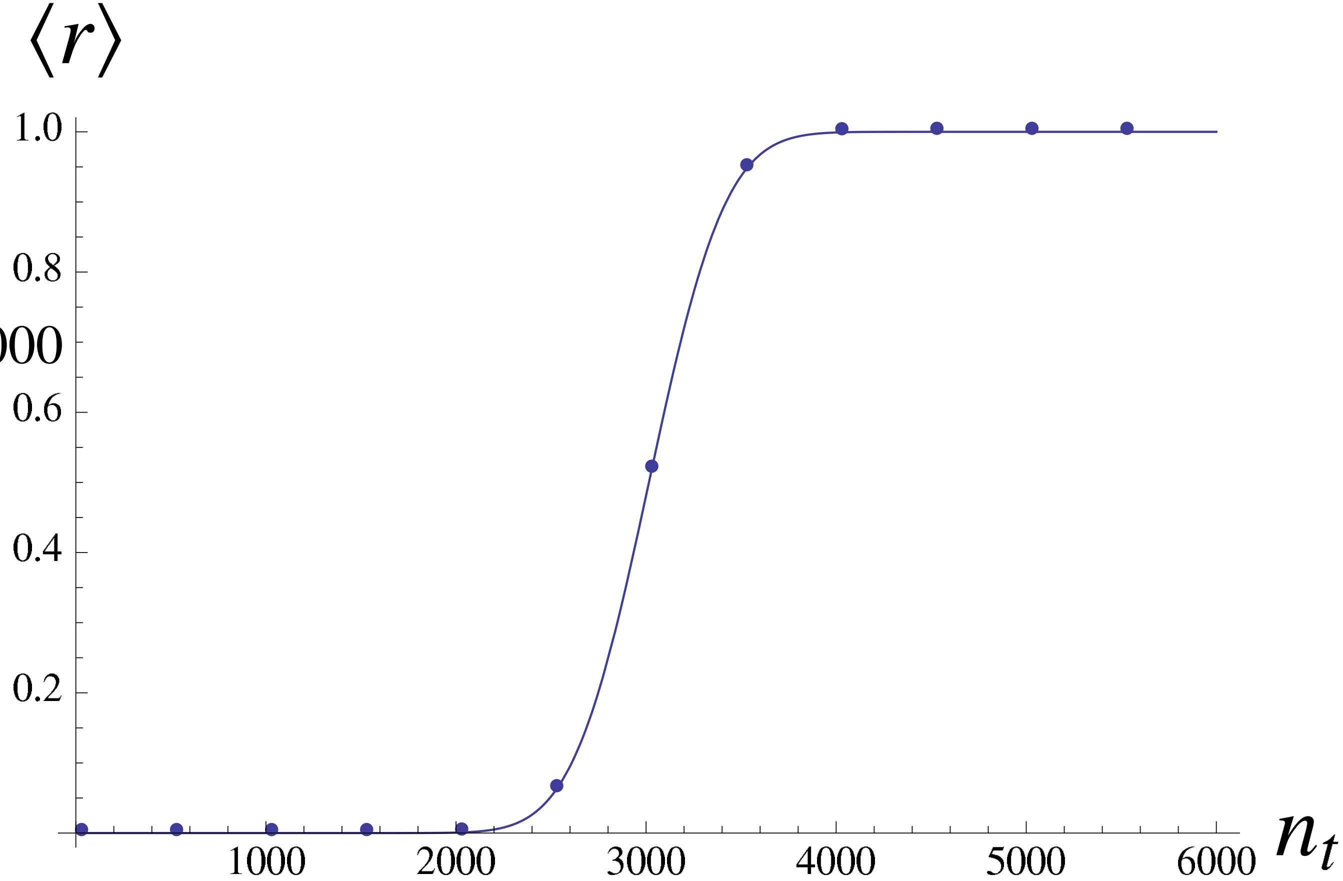
$n_b = 200$

$n_r = 30$

middle score $= 3000$

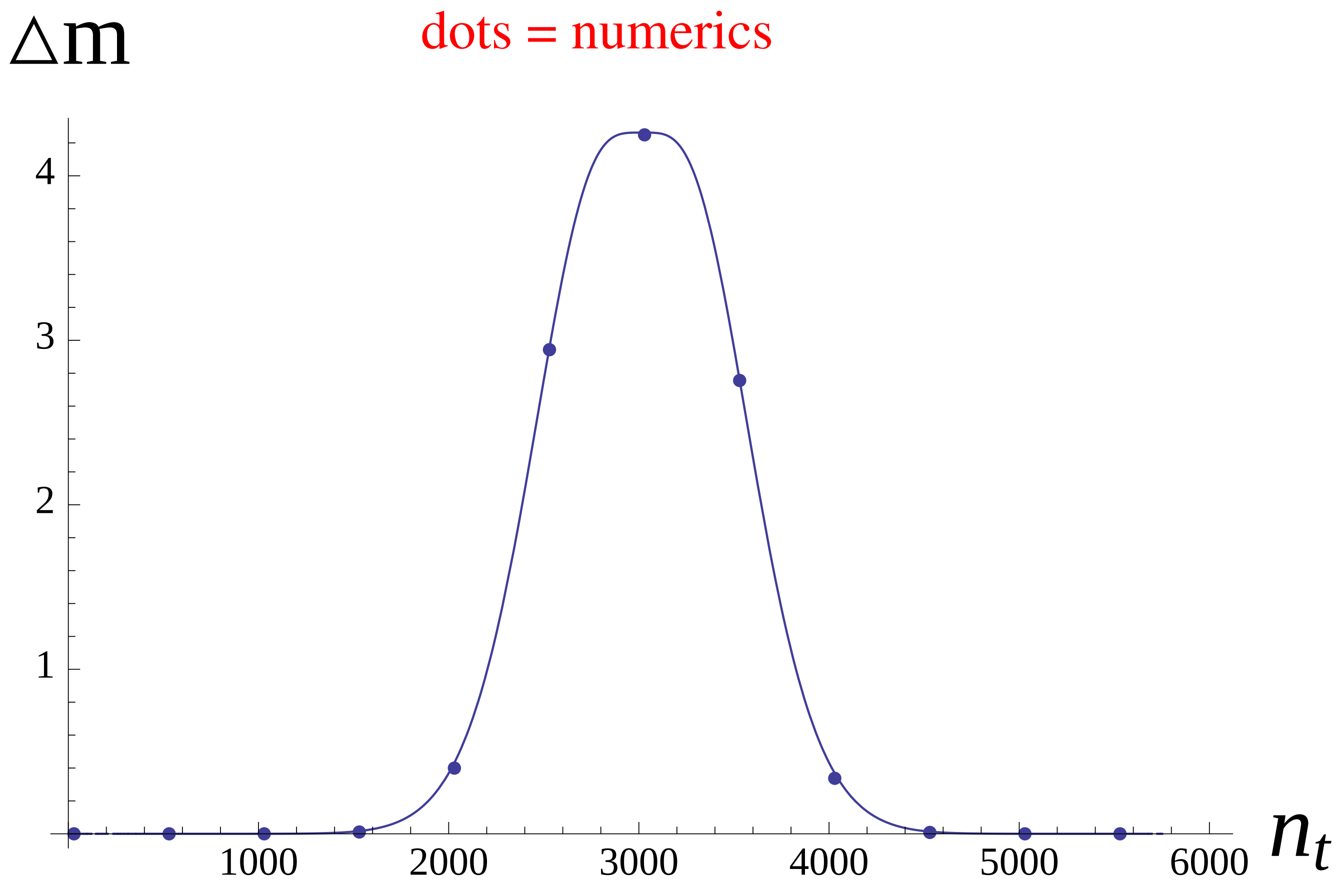
dots = numerics

blessing/curse

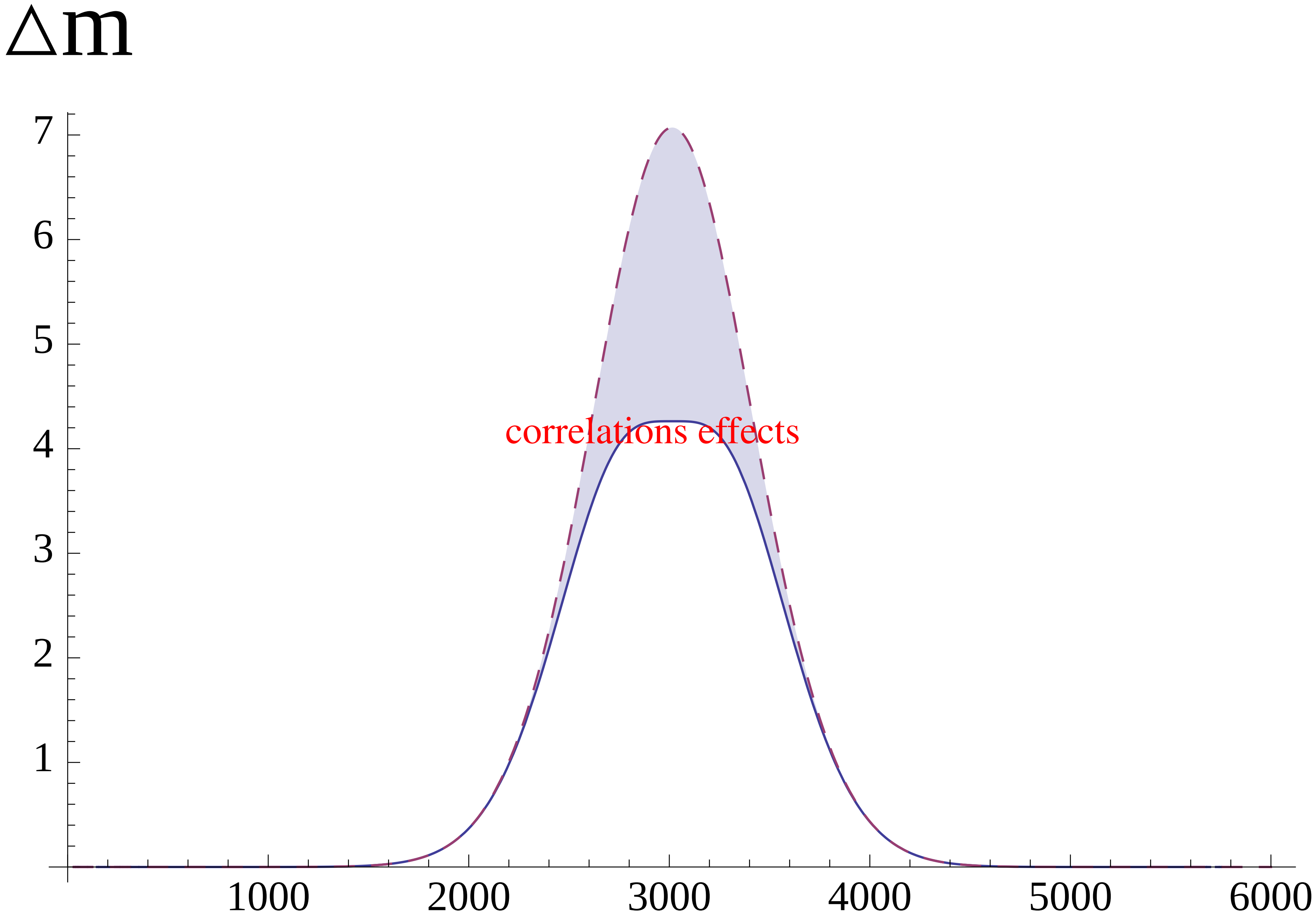


variance

$$\frac{(\Delta m)^2}{n_b} = \frac{1}{\sqrt{2\pi\lambda}} \int_{-\infty}^{\bar{n}_t} \exp\left[-\frac{n^2}{2\lambda}\right] dn \frac{1}{\sqrt{2\pi\lambda}} \int_{-\infty}^{-\bar{n}_t} \exp\left[-\frac{n^2}{2\lambda}\right] dn - \frac{1}{2\pi} \exp\left[-\frac{\bar{n}_t^2}{\lambda}\right]$$



Δm
dots = numerics
4
3
2
1
1000
2000
3000
4000
5000
6000
n_t


Δm
7
6
5
4
3
2
1
correlations effects
1000
2000
3000
4000
5000
6000

now consider small number of races $n_r = 2, 3, \ldots$ and boats $n_b = 1, 2, \ldots$

$\Rightarrow$ combinatorics problem

the simplest case $n_r = 2$: like a "2-body" problem

$\Rightarrow$ exact solution for $P_{n_t}(m)$

how to proceed :

i) represent possible configurations of ranks in the two races by points on a $n_b \times n_b$ lattice

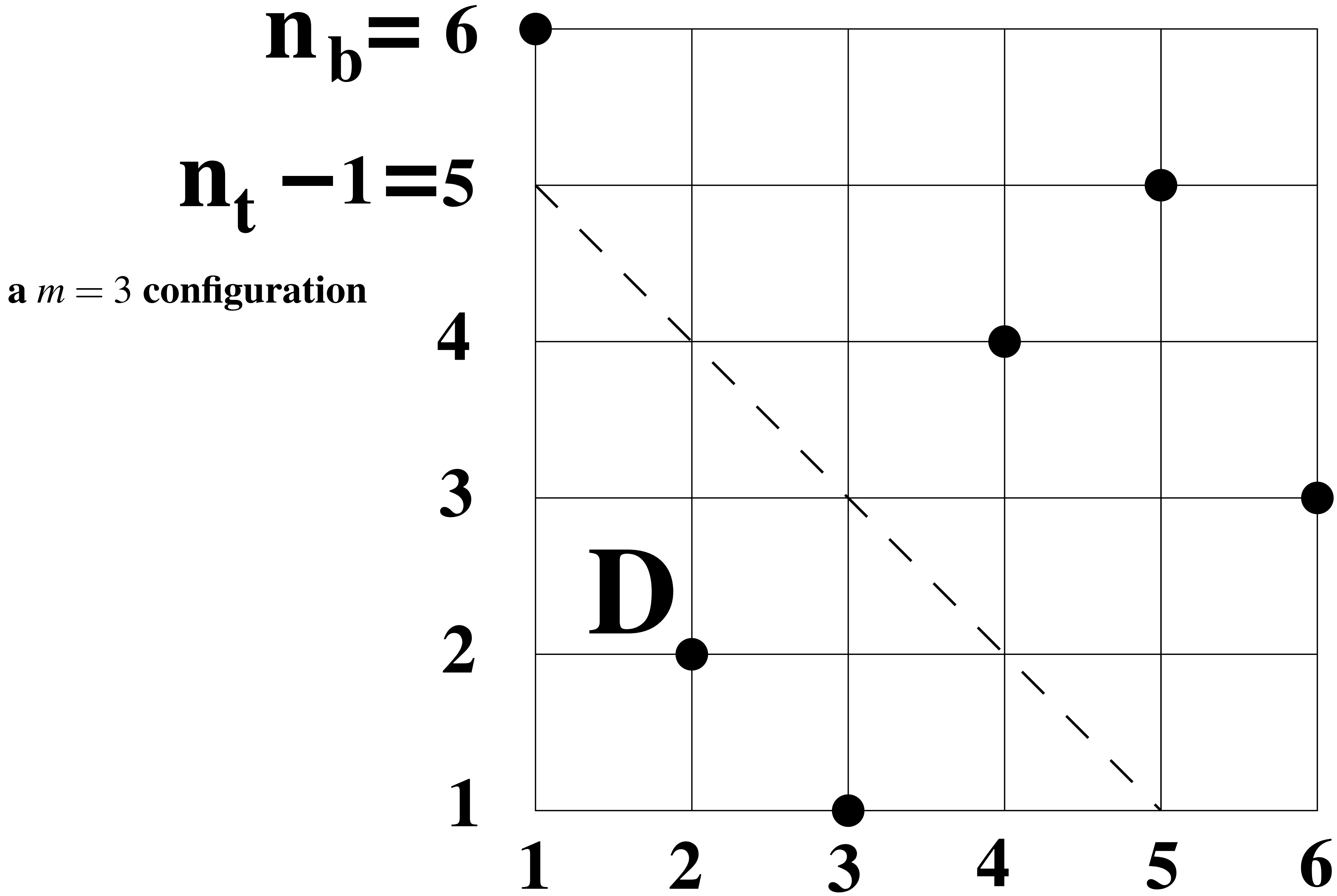
2 races $\leftrightarrow$ square lattice, 3 races $\leftrightarrow$ cubic lattice, $\ldots$

no ex aequo $\Rightarrow$ 1 point per line and per column

in general for n_b boats and n_r races $\rightarrow (n_b!)^{n_r-1}$ such configurations

ii) enumerate the configurations with $m-1$ points below the diagonal n_t

$\Rightarrow$ final rank m


$n_b = 6$
$n_t - 1 = 5$
a $m = 3$ configuration
4
3
D
2
1
1
2
3
4
5
6

combinatorics (not easy) :

for $2 \leq n_t \leq 1 + n_b$

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

for $2+n_b \leq n_t \leq 2n_b+1$ by symmetry $P_{n_b+1-k}(n_b+2-m) = P_{n_b+2+k}(m)$

for the middle score $n_t = 2\frac{1+n_b}{2} = 1+n_b$

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

```
Table[p[5, nt, m], {nt, 2, 6}, {m, 1, 6}]
```

$$\left\{\{1, 0, 0, 0, 0, 0\}, \left\{\frac{4}{5}, \frac{1}{5}, 0, 0, 0, 0\right\}, \left\{\frac{9}{20}, \frac{1}{2}, \frac{1}{20}, 0, 0, 0\right\}, \left\{\frac{2}{15}, \frac{11}{20}, \frac{3}{10}, \frac{1}{60}, 0, 0\right\}, \left\{\frac{1}{120}, \frac{13}{60}, \frac{11}{20}, \frac{13}{60}, \frac{1}{120}, 0\right\}\right\}$$

```
Table[p[nb, nt = 1 + nb, m] nb!, {nb, 1, 7}, {m, 1, nb + 1}]
```

{{1, 0}, {1, 1, 0}, {1, 4, 1, 0}, {1, 11, 11, 1, 0}, {1, 26, 66, 26, 1, 0},
{1, 57, 302, 302, 57, 1, 0}, {1, 120, 1191, 2416, 1191, 120, 1, 0}}

→ Eulerian numbers

$$\alpha = \frac{1}{1(p-1)}$$

$$\beta = \frac{p+1}{1.2\,(p-1)^2}$$

$$\gamma = \frac{pp+4p+1}{1.2.3\,(p-1)^3}$$

$$\delta = \frac{p^3+11p^2+11p+1}{1.2.3.4\,(p-1)^4}$$

$$\varepsilon = \frac{p^4+26p^3+66p^2+26p+1}{1.2.3.4.5\,(p-1)^5}$$

$$\zeta = \frac{p^5+57p^4+302p^3+302p^2+57p+1}{1.2.3.4.5.6\,(p-1)^6}$$

$$\eta = \frac{p^6+120p^5+1191p^4+2416p^3+1191p^2+120p+1}{1.2.3.4.5.6.7\,(p-1)^7}$$

&c.

L. Euler, 1755.

Eulerian Polynomials

$$\frac{A_n(p)/p}{n!(p-1)^n} \quad (1 \le n \le 7)$$

Eulerian number =

the number of permutations of the numbers 1 to n in which exactly m elements are greater than the previous element (permutations with m ”ascents”)

n	***m***	**Permutations**
1	0	(1)
2	0	(2, 1)
	1	(1, **2**)
3	0	(3, 2, 1)
	1	(1, **3**, 2) (2, 1, **3**) (2, **3**, 1) (3, 1, **2**)
	2	(1, **2**, **3**)

$n = 4 \quad (1,4,2,3) \rightarrow m = 2$

generating function

$$g[x, y] = \frac{e^x (-1 + y)}{-e^{x y} + e^x y}$$

Series[g[x, y], {x, 0, 6}]

$$1 + x + \frac{1}{2} (1 + y) x^2 + \frac{1}{6} \left(1 + 4 y + y^2\right) x^3 + \frac{1}{24} \left(1 + 11 y + 11 y^2 + y^3\right) x^4 + \frac{1}{120} \left(1 + 26 y + 66 y^2 + 26 y^3 + y^4\right) x^5 + \frac{1}{720} \left(1 + 57 y + 302 y^2 + 302 y^3 + 57 y^4 + y^5\right) x^6 + O[x]^7$$

why Eulerian numbers should play a role here seems a mystery

but : an other way to look at things by rewriting

$$P_{n_t}(m) = \frac{1}{n_b!}\sum_{i=m}(-1)^{i+m} n_{n_t}(i)(1+n_b-i)!\binom{i-1}{m-1}$$

$n_{n_t}(i) =$ Stirling partition numbers : count in how many ways can the numbers $(1,2,\ldots,n_t-1)$ be partitioned in i groups

example $n_t = 5 : \rightarrow 1$ way to split the numbers $(1,2,3,4)$ into 1 group

$\rightarrow 7$ ways to split the numbers $(1,2,3,4)$ into 2 groups

$(1),(2,3,4);(2),(1,3,4);(3),(1,2,4);(4),(1,2,3);(1,2),(3,4);(1,3),(2,4);(1,4),(2,3)$

$\rightarrow 6$ ways to split the numbers $(1,2,3,4)$ into 3 groups

$(1),(2),(3,4);(1),(3),(2,4);(1),(4),(2,3);(2),(3),(1,4);(2),(4),(1,3);(3),(4),(1,2)$

$\rightarrow 1$ way to split the numbers $(1,2,3,4)$ into 4 groups

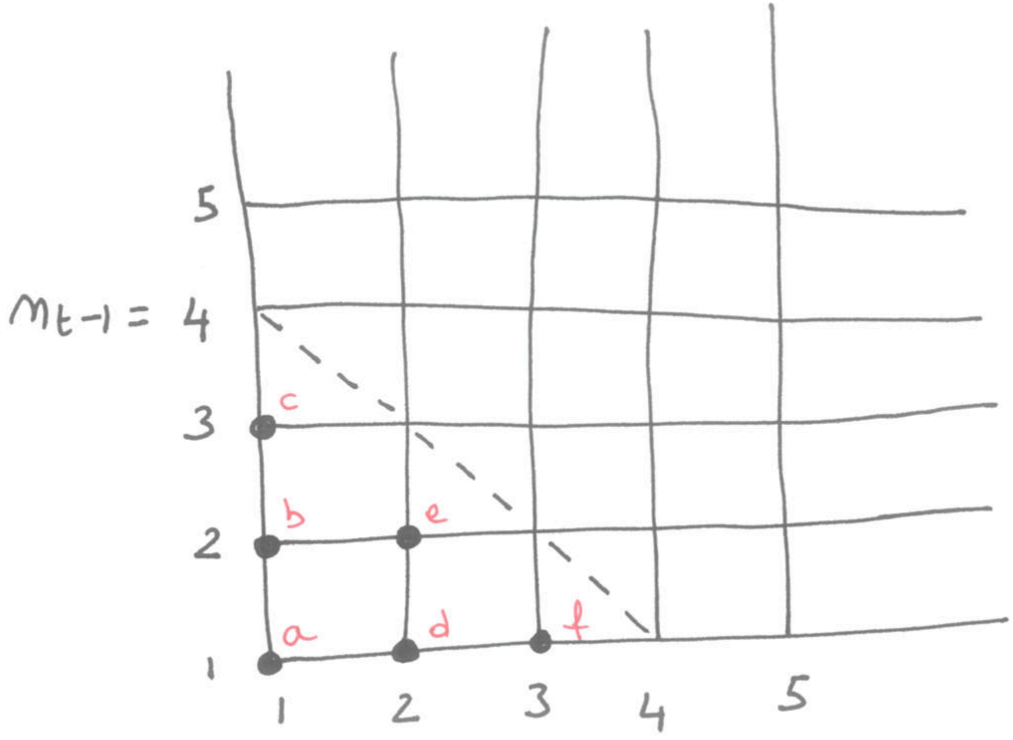
$$n_t = 5 \to 1, 6, 7, 1$$

why Stirling numbers should play a role here seems again a mystery

they appear from graph counting considerations on the configuration lattice :

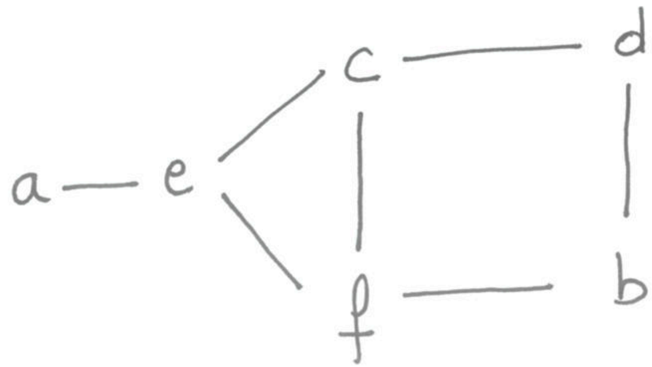
for example for $n_t = 5$

consider all the points below the diagonal



a $\begin{vmatrix} 1 \\ 1 \end{vmatrix}$	b $\begin{vmatrix} 1 \\ 2 \end{vmatrix}$	c $\begin{vmatrix} 1 \\ 3 \end{vmatrix}$	d $\begin{vmatrix} 2 \\ 1 \end{vmatrix}$	e $\begin{vmatrix} 2 \\ 2 \end{vmatrix}$	f $\begin{vmatrix} 3 \\ 1 \end{vmatrix}$

$\Rightarrow$



$\Rightarrow$ 6, 7, 1

so from graph counting

$n_{n_t+1}(i+1) =$ under the diagonal n_t number of subgraphs with i points fully connected

$\rightarrow$ recurrence relation :

either 0 point on the diagonal $n_t - 1 \rightarrow n_{n_t}(i+1)\binom{n_t-1}{0}$

either 1 point on the diagonal $n_t - 1 \rightarrow n_{n_t-1}(i)\binom{n_t-1}{1}$

either 2 points on the diagonal $n_t - 1 \rightarrow n_{n_t-2}(i-1)\binom{n_t-1}{2}$

$\dots$

$\Rightarrow n_{n_t+1}(i+1) = \sum_{k=0}^{i} n_{n_t-k}(i+1-k)\binom{n_t-1}{k}$

$\Leftrightarrow$ recurrence relation for Stirling partition numbers there is indeed a one to one correspondance between Stirling partition (in fact second class Stirling) numbers and Eulerian numbers

$$\mathrm{Eulerian}[\mathrm{n},\mathrm{k}] = \sum_{j=1}^{k+1} (-1)^{k-j+1} \binom{n-j}{n-k-1} j!\ \mathrm{Stirling}[n, j]$$

why all this ?

$$2 \text{ races}: \; P_{n_t}(m) = \frac{1}{n_b!} \sum_{i=m} (-1)^{i+m} n_{n_t}(i) (1+n_b-i)! \binom{i-1}{m-1}$$

$$\rightarrow n_r \text{ races}: \; P_{n_t}(m) = \frac{1}{(n_b!)^{n_r-1}} \sum_{i=m} (-1)^{i+m} n_{n_t}(i, n_r) (1+n_b-i)!^{n_r-1} \binom{i-1}{m-1}$$

how to calculate $n_{n_t}(i, n_r)$? $\rightarrow$ work in progress