A generalization of Stirling numbers and distribution of phylogenetic trees

Éva Czabarka — University of South Carolina, USA

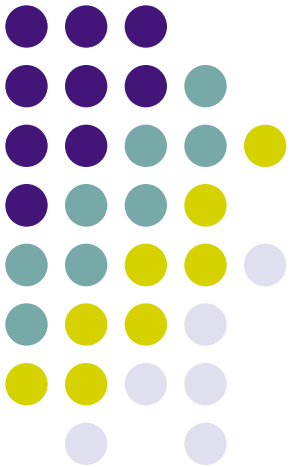
Results joint with:

Péter L. Erdős — Rényi Institute of Mathematics, Hungary

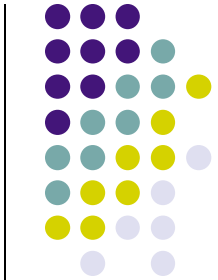
Virginia Johnson — University of South Carolina, USA

Anne Kupczok — Center for Integrative Bioinformatics, Austria

László A. Székely — University of South Carolina, USA



Counting (unordered) trees



- A species tree (phylogenetic tree) is a leaf-labeled tree where the labels occur with multiplicity 1, root is not of degree 1, non-root, non-leaf vertices have degree at least 3.
- Two trees are the same if there is a (label/root preserving) isomorphism between them.
- 1889 Cayley: formula for unrooted trees with n vertices
- 1923 Wedderburn-Etherington: functional equation for OGF of rooted unlabeled binary trees
- 1948 Otter: for unlabeled trees T $\quad p_T - q_T + s_T = 1$

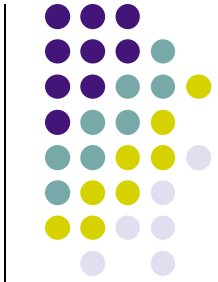
  a way to relate rooted trees to unrooted trees

  asymptotics for rooted & unrooted unlabeled trees on n leaves
- 1967 Cavalli-Sforza – Edwards, 1974 Dobson, 1976 Phipps: formula for rooted and unrooted leaf-labeled binary trees with n leaves
- Flajolet: asymptotics for the total number of phylogenetic trees on n leaves (Schröder's 4th problem 1870), developed combstruct

Rooted leaf-labeled trees and partitions

- $F(n,k)$: number of rooted leaf-labeled trees with k labeled leaves and n non-root vertices
  - Root may or may not have degree 1, and is not a leaf
  - Degree 2 vertices are not supressed
- $S(n,j)$: Strirling number of the second kind, number of partitions of an n element set into j (nonempty) partition classes
- 1983 Erdős-Székely: $F(n,k)=S(n,\ n-k+1)$ with a bijection, under which outdegrees of non-leaf vertices correspond to class sizes.
- Results on $S(n,j)$ translate to results on $F(n,k)$
- 1968 Dobson: $S(n,j)$ unimodal
- Klarner, 1968 Lieb: $S(n,j)$ strictly log concave (implies unimodal)

Normality results



- $A(n,j) \geq 0$, $j=1,\ldots,d_n$, $E_n=A(n,1)+\ldots+A(n,d_n)$; Z_n r.v. with $P(Z_n=j)=A(n,j)/E_n$
- $A(n,j)$ is asymptotically normal:

$$\frac{1}{E_n}\sum_{j=1}^{\lfloor \mathbb{E}(Z_n)+x\sigma(Z_n)\rfloor} A(n,j) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt \text{ uniformly in } x$$

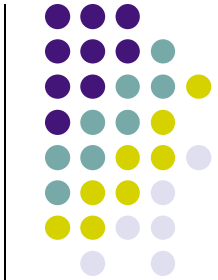
- 1967 Harper: (CLT) asymptotic normality of $S(n,j)$,

$$\mathbb{E}(Z_n)=\frac{B_{n+1}}{B_n}-1 \qquad \sigma^2(Z_n)=\frac{B_{n+2}}{B_n}-\left(\frac{B_{n+1}}{B_n}\right)^2-1$$

- 1977 Canfield: (LLT) asymptotic normality, strict log-concavity and $\sigma(Z_n) \to \infty$ implies

$$\lim_{n\to\infty}\frac{\sigma(Z_n)}{E_n}A\left(n,\lfloor \mathbb{E}(Z_n)+x\sigma(Z_n)\rfloor\right)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2} \text{ uniformly in } x$$

Harper-Canfield method to show asymptotic normality

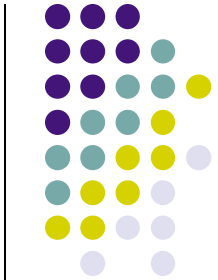


- Consider the generating function $A_n(z) = \sum_k A(n,k) z^k$

$$E(Z_n) = \frac{A_n'(1)}{A_n(1)} \text{ and } \sigma^2(Z_n) = \frac{A_n'(1)}{A_n(1)} + \left(\frac{A_n'(z)}{A_n(z)} \right)' \Bigg|_{z=1}$$

- If $A_n(z)$ only has nonpositive real roots, Z_n is the sum of independent Bernoulli r.v.s
- If in addition $\sigma^2(Z_n) \to \infty$, then $A(n,k)$ is asymptotically normal by the Lindeberg-Feller theorem
- By Canfield's remark srict log-concavity of $A(n,k)$ implies LLT

Asymptotics for Bell numbers



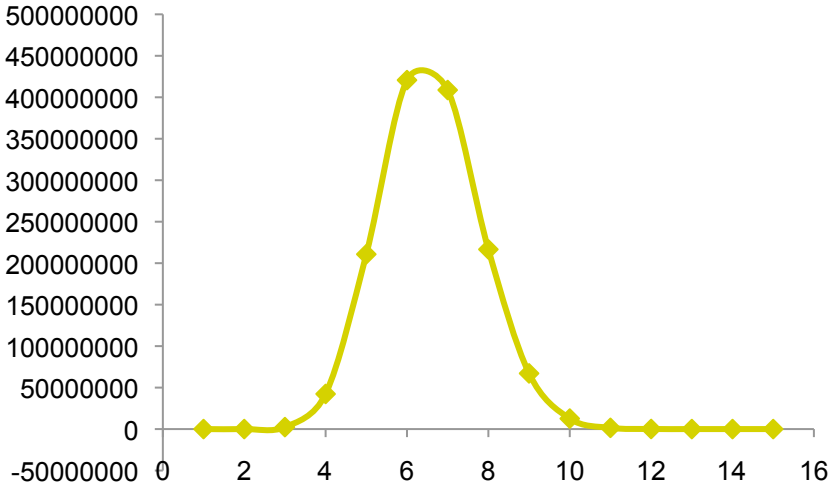
- B_n is the number of partitions of an n-element set
- 1955 Moser-Wyman: asymptotics for B_n
- 1994 Canfield: As $n -> \infty$,

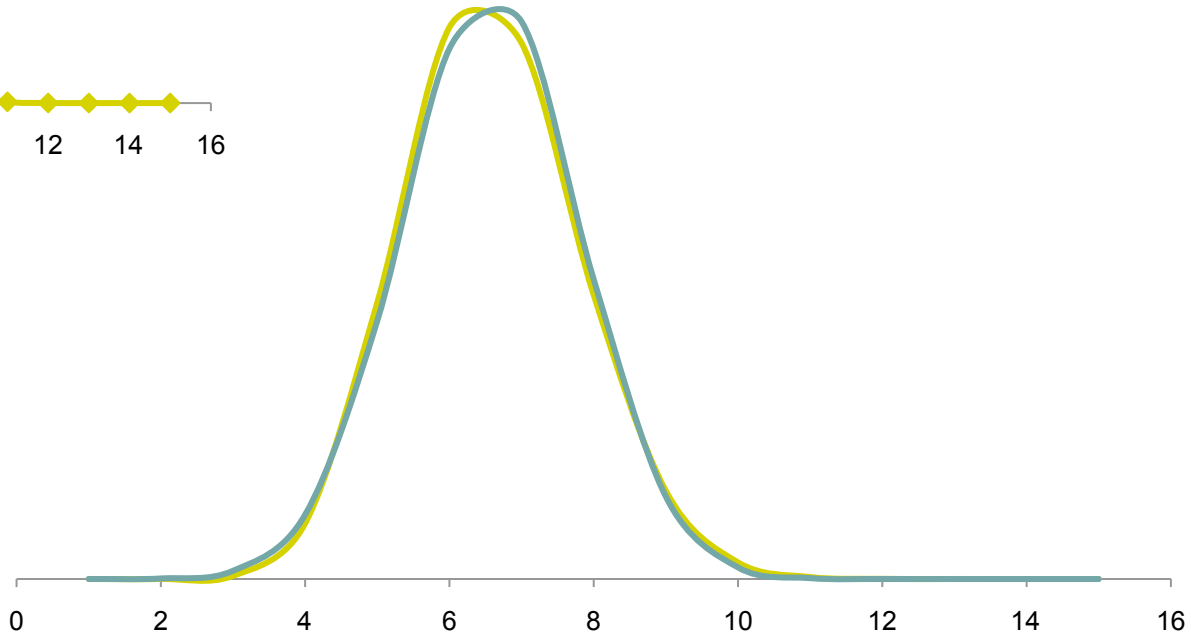
$$B_{n+h} = \frac{(n+h)!}{r^{n+h}} \frac{e^{e^r-1}}{\sqrt{2\pi B}} \left(1 + \frac{\sum_{i=0}^{2} h^i P_i}{e^r} + \frac{\sum_{j=0}^{4} h^j Q_j}{e^{2r}} + O\left(e^{-3r}\right) \right)$$

  where $r=r(n)$ is the solution for $re^r=n$; $h = O(\log(n))$, $P_i, Q_i = \Theta(r^{-i})$ are explicit rational functions of r

- $r = \log(n) - \log\log(n) + O(1)$

$S(15,k)$ "mathematical tree"
$16-k$ leaves, k internal vertices





Rooted species trees and partitions

- 1983 Erdős-Székely: $F(n,k)=S(n, n-k+1)$ with a bijection, under which outdegrees of non-leaf vertices correspond to class sizes.
- The relevant partitions have class size at least 2
- $S^*(n,k)=$ the number of partitions of an n-element set into k partitions, each partition class has size at least 2; $1 \leq k \leq n/2$
- $B^*_n=$ the number of partitions of an n-element set, each partition class has size at least 2
- $F^*(n,k)$: number of rooted leaf-labeled trees with k labeled leaves and n non-root vertices, outdegrees of nonleaf vertices are at least 2
- Still $F^*(n,k)=S^*(n, n-k+1)$

Distribution of $S^*(n,m)$

$$B_n = B^*_{n+1} + B^*_n$$

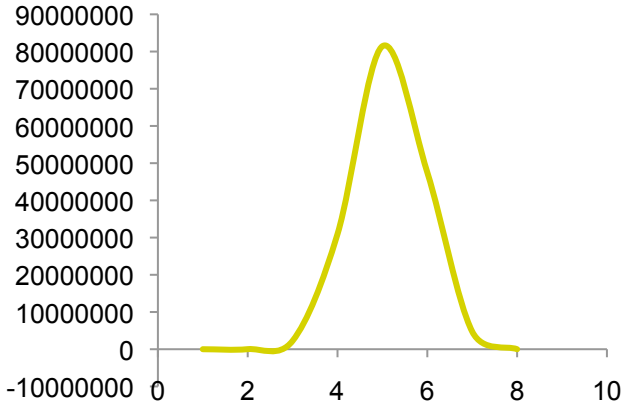
$$B^*_{n+1} = (-1)^n + \sum_{i=1}^{n} (-1)^{n-i} B_i$$

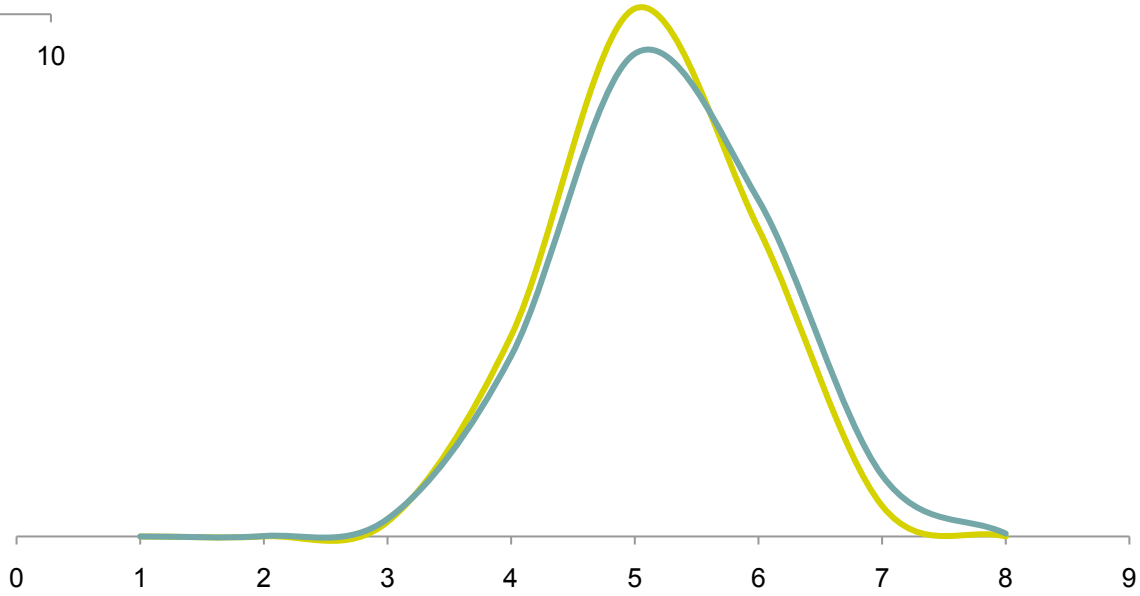
$$E(Z^*_n) = \frac{n}{r} - r - \frac{1}{2r} + \frac{1}{2r(r+1)^2} + O\left(\frac{1}{n}\right)$$

$$\sigma^2(Z^*_n) = \frac{n}{r(r-1)} - r + 1 + \frac{2}{r+1} - \frac{1}{2(r+1)^1} - \frac{1}{2(r+1)^3} + \frac{1}{(r+1)^4} + O\left(\frac{1}{n}\right)$$

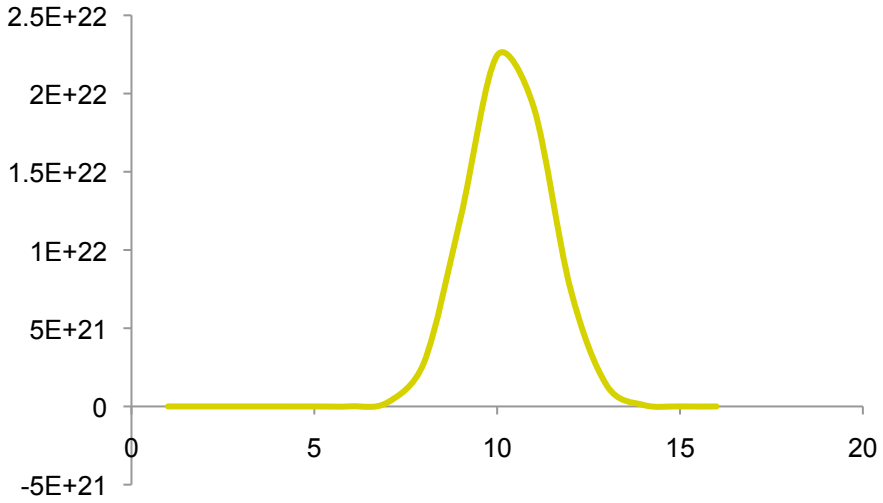
- $S^*(n,m)$ is asymptotically normal and has LLT – it is SLC, the generating polynomial has different nonpositive real roots and $\sigma(Z_n) \to \infty$
- This means that if we fix the number of non-root vertices and count the number of trees with k internal vertices the resulting distribution tends to normal

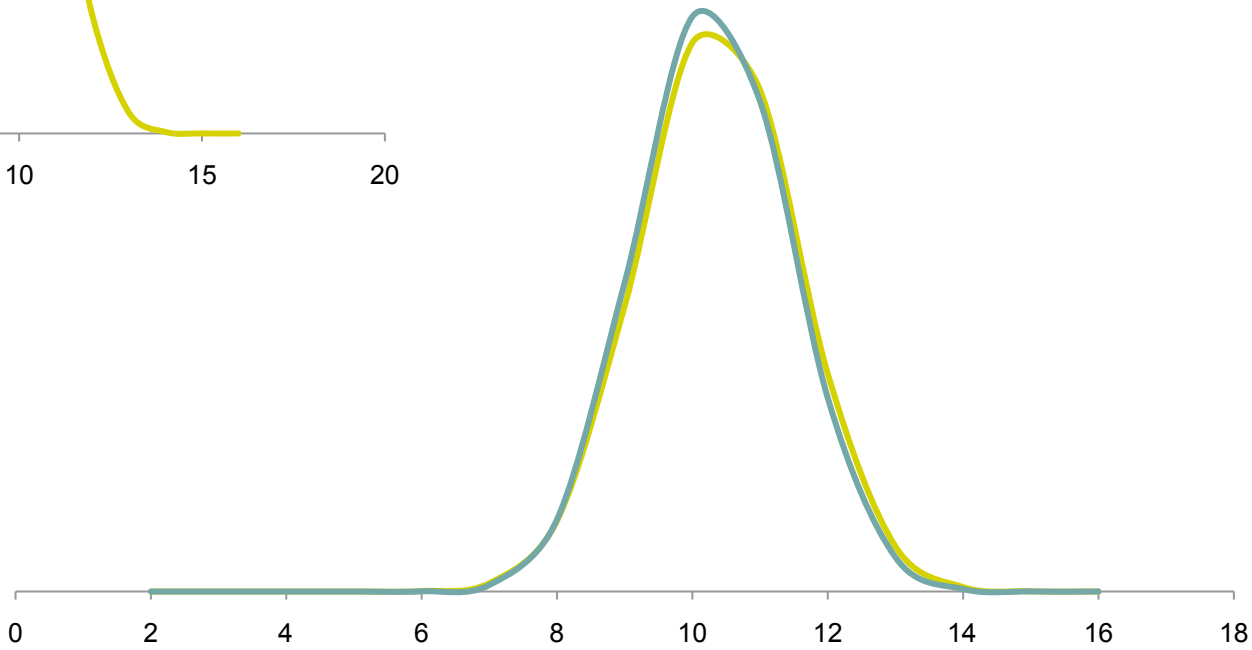
$S^*(15,k)$
$16-k$ leaves, k internal vertices





$S^*(30,k)$
$31-k$ leaves, k internal vertices





Distribution of $S^*(n+m,m)$

- Relevant distribution: number of leaves are fixed, and the number of non-leaf points vary;

  $T_{n+1,m} = F^*(n+m,n+1) = S^*(n+m, m)$, $m=1,2,\ldots,n$

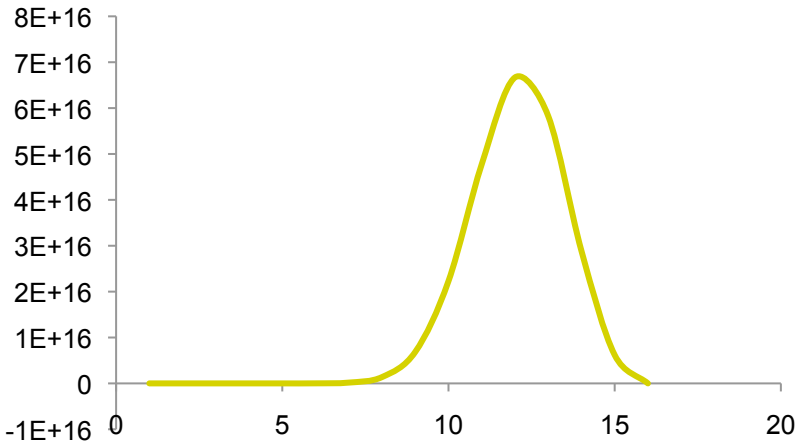
- $F(n+m,n+1) = S(n+m, m)$, $m=1,2,\ldots$

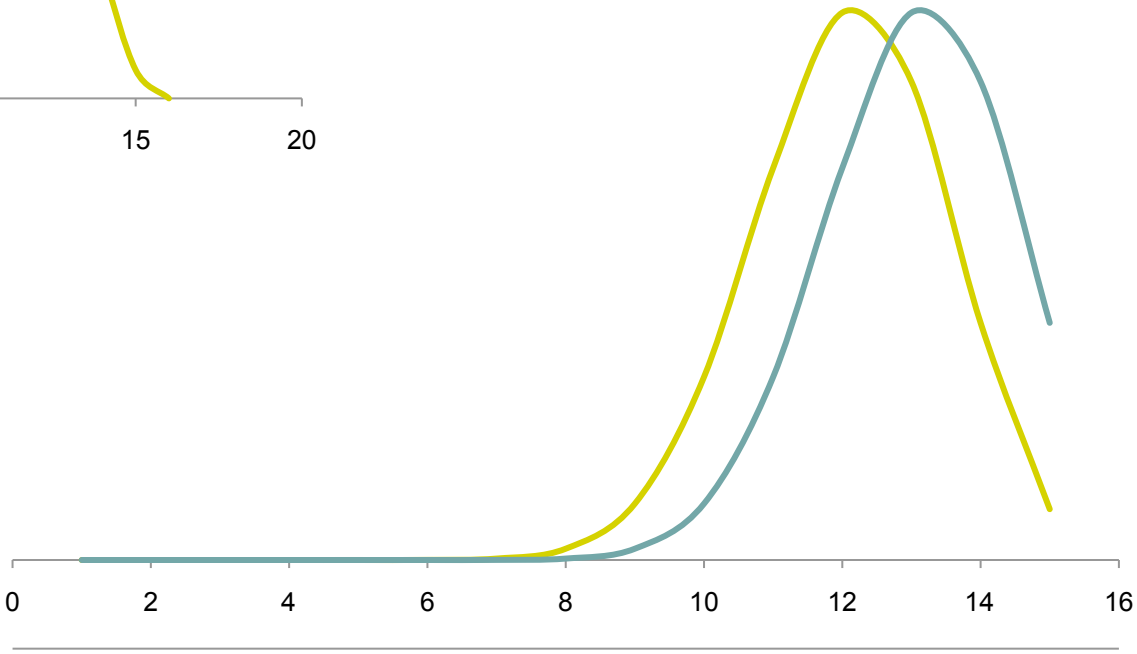
$$E(Z_n) = \frac{1-\rho}{2\rho} n + \frac{.75 - \ln 2}{\rho} + O\left(\frac{1}{n}\right) \quad \text{Flajolet}$$

$$\sigma^2(Z_n) = \frac{n}{4}\left(\frac{1}{\rho^2} - \frac{2}{\rho} - 1\right) + \frac{1 - 2\rho - 2\rho^2}{8\rho^2} + O(1), \text{ where } \rho = -1 + 2\ln(2)$$

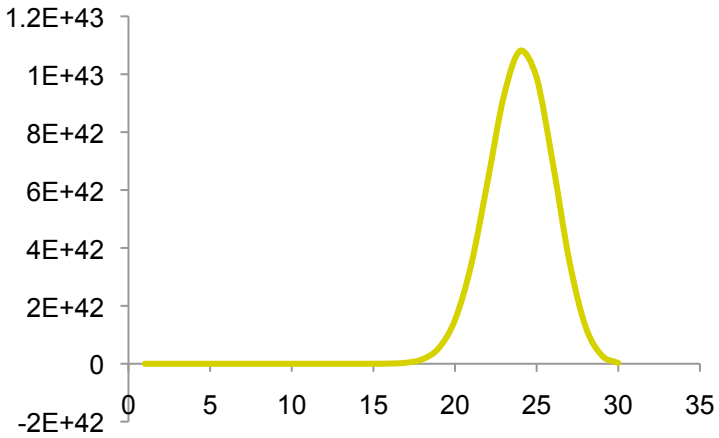
- $T_{n,m}$ is asymptotically normal & has LLT – it is SLC, the generating polynomial has different nonpositive real roots and $\sigma(Z_n) \to \infty$

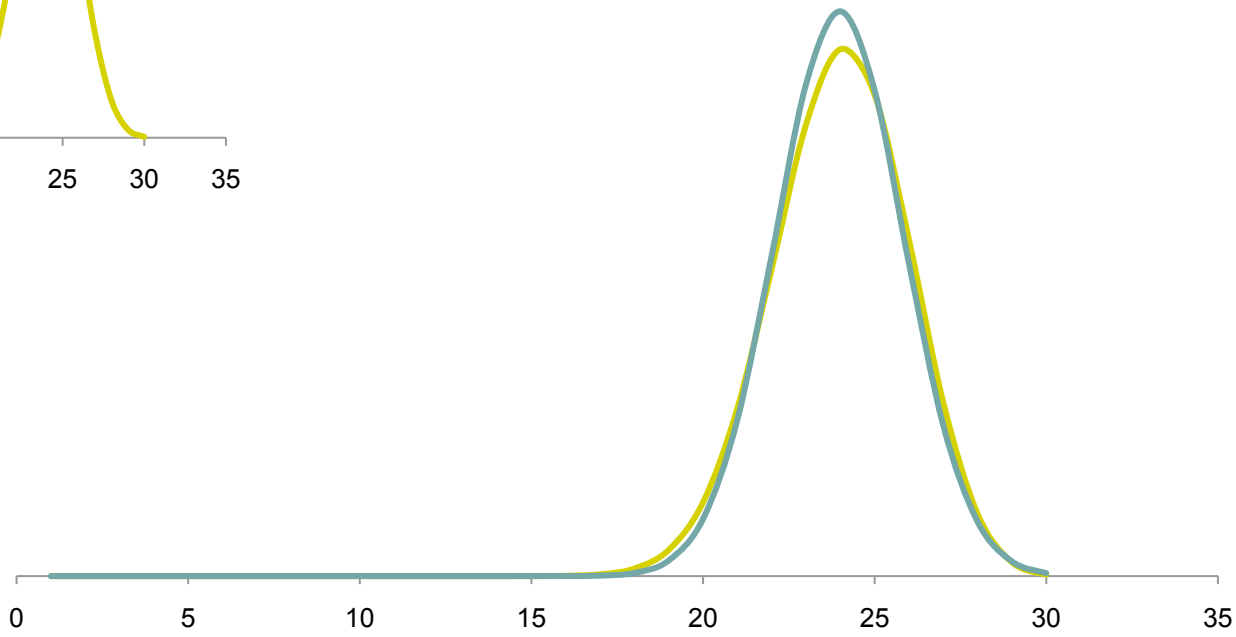
$$S^*(15+k,k)=T_{16,k}$$


8E+16
7E+16
6E+16
5E+16
4E+16
3E+16
2E+16
1E+16
0
-1E+16
0
5
10
15
20


0
2
4
6
8
10
12
14
16

$$S^*(30+k,k)=T_{31,k}$$


1.2E+43
1E+43
8E+42
6E+42
4E+42
2E+42
0
-2E+42
0
5
10
15
20
25
30
35


0
5
10
15
20
25
30
35