CENTRAL LIMIT THEOREMS FOR SOME SET PARTITION STATISTICS

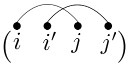
BOBBIE CHERN, PERSI DIACONIS, DANIEL M. KANE, AND ROBERT C. RHOADES

Abstract. We prove the conjectured limiting normality for the number of crossings of a uniformly chosen set partition of $[n] = \{1, 2, \dots, n\}$. The arguments use a novel stochastic representation and are also used to prove central limit theorems for the dimension index and the number of levels.

1. Introduction

Let λ be a partition of the set $[n] = \{1, 2, \dots, n\}$, so $1|2|3, 12|3, 13|2, 1|23, 123$ are the five partitions of [3]. The enumerative theory of "supercharacters" leads to the statistics

$$d(\lambda) = \sum_i (M_i - m_i + 1) \quad \text{and} \quad cr(\lambda) = \# \text{ of crossings of } \lambda. \tag{1.1}$$

In $d(\lambda)$, the sum is over the blocks of λ and M_i (m_i) is the largest (smallest) element of the block i. The statistic $cr(\lambda)$ counts $i < i' < j < j'$ with i, j adjacent elements of the same block and i', j' adjacent elements of the same block (). In a companion paper [3] the moments of $d(\lambda)$ and $cr(\lambda)$ are determined as explicit linear combinations of Bell numbers B_n. Numerical computations (see Figures 1 and 2) suggests that normalized by their mean and variance, these statistics have approximate normal distributions. Figures 1 – 3 are based on exact counts from our new algorithms [3]. Figures 1 and 2 suggest good agreement with the normal approximation for dimension index and crossings. Figure 3 shows slower convergence for levels and suggests a search for finite sample correction terms. We found the limiting normality challenging to prove using available techniques (eg. moments, Fristedt's method of conditioned limit theorems [9], or Stein's method [4]). Indeed, the limiting normality of $cr(\lambda)$ is conjectured in [13].

A key ingredient of the present paper is a stochastic algorithm for generating a random set partition due to Stam [24]. Supplementing this with some novel probabilistic ideas allows standard "delta method" techniques to finish the job.

Brief reviews of the extensive enumerative, algebraic and probabilistic aspects of set partitions are in [16] and [25]. The book of Mansour [18] contains applications to computer science and much else. An important paper combining many of the statistics we work with is [2]. The companion paper [3] has an extensive review. It also summarizes the literature on supercharacters. Briefly, these are natural characters χ_λ on the uni-upper triangular matrix group $U_n(\mathbb{F}_q)$ which are indexed by set partitions. The representation corresponding to χ_λ has dimension $q^{d(\lambda)}$. The (usual) inner product between χ_λ and χ_μ is $<\chi_\lambda, \chi_\mu>= q^{cr(\lambda)}\delta_{\lambda,\mu}$. This suggests understanding how $d(\lambda)$ and $cr(\lambda)$ vary for typical set partitions.

Date: December 15, 2014.

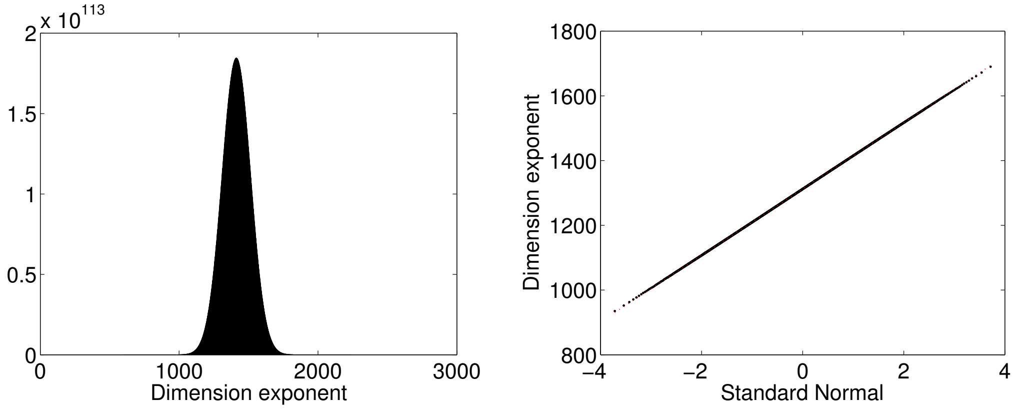


FIGURE 1. Histogram of the dimension exponent counts for $n = 100$ and the associated Q-Q plot.

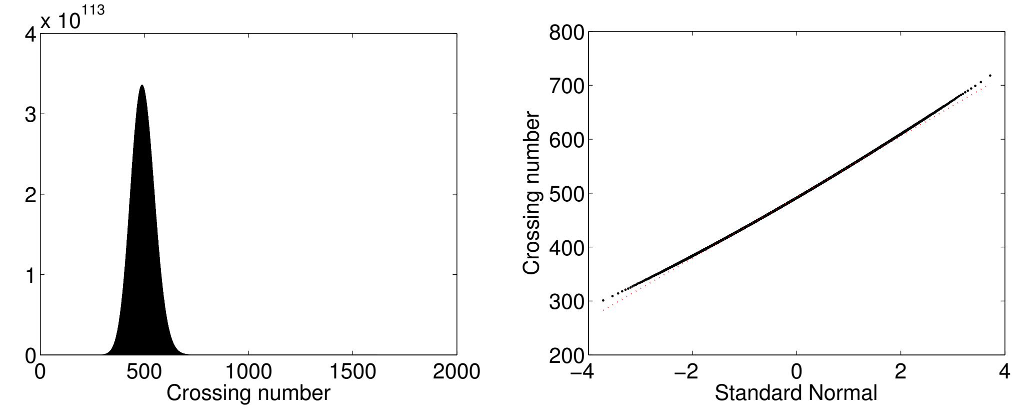


FIGURE 2. Histogram of the crossing number counts for $n = 100$ and the associated Q-Q plot.

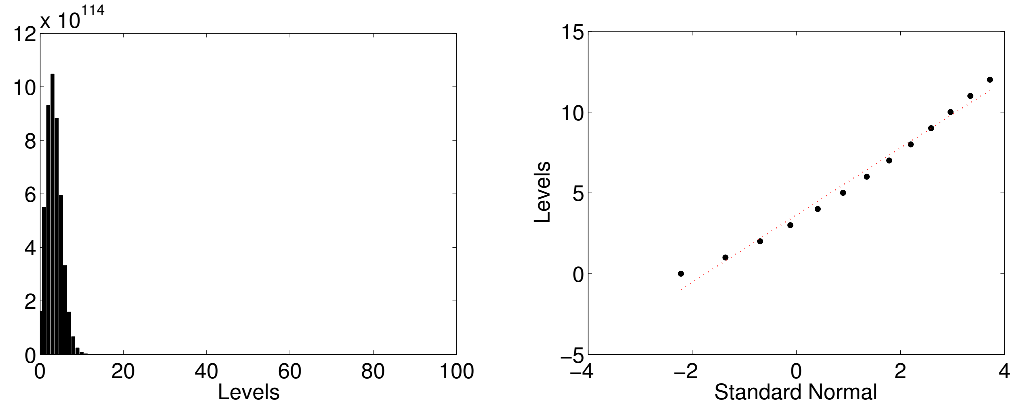


FIGURE 3. Histogram of the level counts for $n = 100$ and the associated Q-Q plot.

There are many codings of a set partition. One needed below codes λ as a sequence $x_1, x_2, \cdots, x_n$ with $x_i = j$ if and only if i is in block j of λ. Thus $135|24|6|7$ corresponds to $1,2,1,2,1,3,4$. If $a_i = x_i - 1$, $a_1, a_2, \cdots, a_n$ is a restricted growth sequence: $a_1 = 0$ and $a_{i+1} \leq 1 + \max(a_1, \cdots, a_i)$ for $1 \leq i \leq n-1$. This standard coding is discussed in [16, page 416]. For this coding, let

$$L(\lambda) = |\{i : x_{i+1} = x_i\}| \tag{1.2}$$

the number of levels of λ. This is used as an example of the present techniques. See [18, Chapter 4] for further references.

The main theorems proved use α_n, the positive real solution of $ue^u = n+1$ (so $\alpha_n = \log(n) - \log\log(n) + o(1)$ [8]). Let $\Pi(n)$ be the set of partitions of $[n]$. Throughout, λ is uniformly chosen in $\Pi(n)$.

Theorem 1.1. *The number of levels $L(\lambda)$ has $\mu_n^L = \mathbb{E}(L(\lambda)) = (n-1)\frac{B_{n-1}}{B_n} \sim \log(n)$ and $\left(\sigma_n^L\right)^2 = \mathrm{VAR}(L(\lambda)) = (n-1)\frac{B_{n-1}}{B_n} + n(n-1)\frac{B_{n-2}}{B_n} - (n-1)^2\frac{B_{n-1}^2}{B_n^2} \sim \log(n)$. Normalized by its mean and standard deviation, $L(\lambda)$ has an approximate standard normal distribution*

$$\mathbf{P}\left(\frac{L(\lambda)-\mu_n^L}{\sigma_n^L} \leq x\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\tau^2/2}d\tau$$

for all fixed x as $n \to \infty$.

Theorem 1.2. *The dimension index $d(\lambda)$ has $\mu_n^d = \mathbb{E}(d(\lambda)) = \frac{\alpha_n-2}{\alpha_n}n^2 + O\left(\frac{n}{\alpha_n}\right)$ and $\left(\sigma_n^d\right)^2 = \mathrm{VAR}(d(\lambda)) = \left(\frac{\alpha_n^2-7\alpha_n+17}{\alpha_n^3(\alpha_n+1)}\right)n^3 + O\left(\frac{n^2}{\alpha_n}\right)$. Normalized by its mean and standard deviation, $d(\lambda)$ has an approximate standard normal distribution*

$$\mathbf{P}\left(\frac{d(\lambda)-\mu_n^d}{\sigma_n^d} \leq x\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\tau^2/2}d\tau$$

for all fixed x as $n \to \infty$.

Theorem 1.3. *The number of crossings $cr(\lambda)$ has $\mu_n^{cr} = \mathbb{E}(cr(\lambda)) = \frac{2\alpha_n-5}{4\alpha_n^2}n^2 + O\left(\frac{n}{\alpha_n}\right)$ and $(\sigma_n^{cr})^2 = \mathrm{VAR}(cr(\lambda)) = \frac{3\alpha_n^2-22\alpha_n+56}{9\alpha_n^3(\alpha_n+1)}n^3 + O\left(\frac{n^2}{\alpha_n}\right)$. Normalized by its mean and standard deviation, $cr(\lambda)$ has an approximate standard normal distribution*

$$\mathbf{P}\left(\frac{cr(\lambda)-\mu_n^{cr}}{\sigma_n^{cr}} \leq x\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\tau^2/2}d\tau$$

for all fixed x as $n \to \infty$.

Section 2 of this paper explains Stam's algorithm and shows how it gives a useful heuristic picture of what a random set partition "looks like". The limit theorem for levels is proved in Section 3 as a simple illustration of our proof technique. The dimension index and number of crossings require further ideas. They are given separate proofs in Sections 4 and 5.

Notation

Throughout, we use the stochastic order symbols O_p and o_p. If X_n for $1 \leq n < \infty$ is a sequence of real valued random variables and a_n is a sequence of real numbers, write $X_n = O_p(a_n)$ if for every $\epsilon > 0$ and some $\eta > 0$, which may depend on ϵ, there is N so that $\mathbf{P}\{|X_n| \leq \eta\,|a_n|\} > 1-\epsilon$ for all $n > N$. Write $X_n = o_p(a_n)$ if for every $\epsilon > 0$ and $\eta > 0$ there is N so that $\mathbf{P}\{|X_n| < \eta\,|a_n|\} > 1-\epsilon$ for all $n > N$. For background, examples and many variations see Pratt [21], Lehman [17], or Serfling [22]. We say two sequences of random variables are weak star close if their distributions are close in Lévy metric.

2. Stam's algorithm and set partition heuristics

Write $\Pi(n)$ for the set partitions of $[n] = \{1, 2, \cdots, n\}$ and $B_n = |\Pi(n)|$ for the nth Bell number (sequence A000110 of Sloane's [23]). To help evaluate asymptotics it is helpful to have

$$\begin{aligned}
\frac{B_{n+1}}{B_n} &= \frac{n}{\alpha_n} + \frac{1}{2}\frac{\alpha_n}{(1+\alpha_n)^2} + O\left(\frac{\alpha_n}{n}\right)\\
\frac{B_{n+k}}{B_n} &= \frac{(n+k)!}{n!\alpha_n^k}\left(1 + O\left(\frac{k}{n\alpha_n}\right)\right)\\
\frac{\alpha_{n+k}}{\alpha_n} &= 1 + O\left(\frac{k}{n\log(n)}\right)
\end{aligned}$$

which are valid for fixed k as $n \to \infty$. See, for instance, [8]. Dobinski's identity [7, 20]

$$B_n = \frac{1}{e}\sum_{m=1}^{\infty}\frac{m^n}{m!} \tag{2.1}$$

shows that for fixed $n \in \{0, 1, 2, \cdots\}$

$$\mu_n(m) = \frac{1}{eB_n}\frac{m^n}{m!} \tag{2.2}$$

is a probability measure on $\{1, 2, 3, \cdots\}$. Stam [24] uses this measure to give an elegant algorithm for choosing a uniform random element of $\Pi(n)$.

Stam's Algorithm

(1) Choose M from μ_n.
(2) Drop n labelled balls uniformly into M boxes.
(3) Form a set partition λ of $[n]$ with i and j in the same block if and only if balls i and j are in the same box.

Of course, after choosing M and dropping balls, some of the boxes may be empty. Stam [24] shows that the number of empty boxes has (exactly) a Poisson distribution and is independent of the generated set partition. This implies that the number of boxes M drawn from μ_n at (2.2) has the same limiting distribution as the number of blocks in a random $\lambda \in \Pi(n)$. This is a well studied random variable. It will emerge that the fluctuations of M are the main source of randomness in Theorems 1.1 – 1.3. Results of Hwang [11] prove the following normal limit theorem (Hwang also has an error estimate).

Theorem 2.1. *For M chosen from μ_n of (2.2), as $n \to \infty$*

$$\mu_n^M := \mathbb{E}(M) = \frac{B_{n+1}}{B_n} = \frac{n}{\alpha_n} + O\left(\frac{1}{\alpha_n}\right)$$

and

$$\left(\sigma_n^M\right)^2 := \mathrm{VAR}(M) = \frac{B_{n+2}}{B_n} - \frac{B_{n+1}^2}{B_n^2} = \frac{n}{\alpha_n^2} + O\left(\frac{n}{\alpha_n^3}\right).$$

Normalized by its mean and standard deviation, M has an approximate standard normal distribution.

Heuristic I. Stam's algorithm gives a useful intuitive way to think about a random element of $\Pi(n)$. It behaves practically the same as a uniform multinomial allocation of n labelled balls into $m = n/\log(n)$ boxes. The arguments in the following sections make this precise. It appears to us that many of the features previously treated in the beautiful paper of Fristedt [9] can be treated by the present approach. Note that Fristedt treated features that only depend on block sizes (largest, smallest, number of boxes of size i). None of our statistics have this form.

Heuristic II. Fristedt's arguments randomize n. This makes the block variables, $N_i(\lambda) =$ # blocks of size i, independent allowing standard probability theorems to be used. At the end, a Tauberian argument (dePoissonization) is used to show that the theorems hold for fixed n. The present argument fixes n and randomizes the number of blocks. This results in a "balls in boxes" problem with many tools available. At the end, an Abelian argument shows that the appropriate limit theorem holds when m fluctuates. See [10] for background on this use of Abelian and Tauberian theorems. There are many variants of Poissonization in active use. We do not see how to abstract Stam's algorithm to other combinatorial structures.

We conclude this section with a simple illustration of Stam's algorithm. From (2.2), $\mu_n(m) = \frac{1}{eB_n}\frac{m^n}{m!}$ is a probability measure on $\{1, 2, 3, \cdots\}$. Thus for $-n < d < \infty$

$$\mathbb{E}_n(M^d) = \frac{1}{eB_n}\sum_{m=1}^{\infty}\frac{m^{n+d}}{m!} = \frac{B_{n+d}}{B_n}. \tag{2.3}$$

Let us apply this to compute the moments for $L(\lambda)$, the number of levels of $\lambda \in \Pi(n)$. From the definition (1.2), given M, $L(\lambda) = X_1 + \cdots + X_{n-1}$ where X_i is the indicator random variable of the event that balls i and $i+1$ are dropped into the same box. By inspection, the X_i are independent with $\mathbf{P}(X_i = 1) = \frac{1}{M}$. Thus

$$\mathbb{E}_n(L(\lambda)) = \mathbb{E}_n\mathbb{E}\left(L(\lambda|M)\right) = \mathbb{E}_n\left(\frac{n-1}{M}\right) = (n-1)\frac{B_{n-1}}{B_n}. \tag{2.4}$$

The standard identity

$$\mathrm{VAR}(Z) = \mathbb{E}(\mathrm{VAR}(Z|W)) + \mathrm{VAR}(\mathbb{E}(Z|W))$$

for any random variables Z and W such that the moments exist, shows that

$$\mathrm{VAR}_n(L(\lambda)) = (n-1)\frac{B_{n-1}}{B_n} + n(n-1)\frac{B_{n-2}}{B_n} - (n-1)^2\frac{B_{n-1}^2}{B_n^2}. \tag{2.5}$$

More generally, this provides an alternative approach to [3] for showing that the moments of statistics $T(\lambda)$ are shifted Bell polynomials. It requires $\mathbb{E}_n(T(\lambda)|m)$ to be a Laurent polynomial in m. As an example, Stam worked with $W_i(\lambda)$, the size of the block in λ containing i, $1 \leq i \leq n$. Then, any polynomial in the $\{W_i\}_{i=1}^n$ has expectation a shifted Bell polynomial; for example, W_i^k and W_iW_j. Stam proves that W_i is approximately normal.

3. Proof of Theorem 1.1

Theorem 1.1 is proved here as a simple illustration of our technique. Conditioning on M in Stam's algorithm, classical "balls in bins" central limit theorems are used to prove the limiting normality uniformly in M and standard δ-method arguments are used to complete the proof.

Proof of Theorem 1.1. The moments of the level statistic $L(\lambda)$ are computed in (2.4) and (2.5). Conditional on M, $L(\lambda) = X_1 + \cdots + X_{n-1}$ with X_i independent identically distributed binary variables with $\mathbf{P}(X_i = 1) = 1/M$. Thus conditioned on M,

$$\mathbb{E}(L(\lambda) \mid M) = \frac{n-1}{M}, \quad \text{and} \quad \mathrm{VAR}(L(\lambda) \mid M) = \frac{n-1}{M}\left(1 - \frac{1}{M}\right).$$

and, normalized by its conditional mean and variance, $L(\lambda)$ has a standard normal limiting distribution provided $n/M \to \infty$. In the present case, $M = M_n$ is a random variable. From Theorem 2.1, as n tends to infinity

$$\frac{M_n - \mu_n^M}{\sigma_n^M} \to N(0,1) \text{ with } \mu_n^M \sim \frac{n}{\alpha_n}, (\sigma_n^M)^2 \sim \frac{n}{\alpha_n^2}. \tag{3.1}$$

This implies

$$\frac{n}{M_n} = \alpha_n + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{3.2}$$

To be precise, write $M_n = \mu_n^M + Z_n\sigma_n^M$ with $Z_n = \frac{M_n - \mu_n^M}{\sigma_n^M}$. Then

$$\frac{n}{M_n} = \frac{n}{\mu_n^M + Z_n\sigma_n^M} = \frac{n}{\mu_n^M\left(1 + \frac{Z_n}{\sigma_n^M}\mu_n^M\right)} = \frac{n}{\mu_n}\left(1 + \frac{Z_n}{\mu_n^M}\sigma_n^M + O\left(\left(\frac{Z_n\sigma_n^M}{\mu_n^M}\right)^2\right)\right). \tag{3.3}$$

From Theorem 2.1, $n/\mu_n^M = \alpha_n + O(\alpha_n/n)$, $\sigma_n^M/\mu_n^M = O(1/\sqrt{n})$. Since $Z_n = O_p(1)$, (3.2) follows.

Thus, with probability close to 1 with respect to M we have that $L(\lambda)$ conditioned on M is weak star close to a Gaussian with mean

$$\mu^M = \frac{n-1}{M} = \alpha_n + O_p(n^{-1/2})$$

and standard deviation

$$\sigma^M = \sqrt{\frac{n-1}{M}\left(1 - \frac{1}{M}\right)} = \sqrt{\alpha_n} + O_p(n^{-1/2}).$$

Thus, with high probability over M, the conditional distribution on $L(\lambda)$ is weak star close to $N(\alpha_n, \sqrt{\alpha_n})$. Therefore, the overall distribution of $L(\lambda)$ is also close to this normal distribution.

□

4. Proof of Theorem 1.2

In outline, the proof proceeds by choosing a random $\lambda \in \Pi(n)$ using Stam's algorithm. Conditioning on the chosen m reduces the problem to a slightly non-standard balls in boxes problem. Given m, it is shown that $d(\lambda) = nm - 2m^2 + O_p(m^{3/2})$ so that the fluctuations in $d(\lambda)$ are driven by the fluctuations in m. These are asymptotically normally distributed with mean and variance $\left(\frac{n}{\alpha_n}, \frac{n}{\alpha_n^2}\right)$. From Theorem 2.1 above, a simple averaging argument completes the proof. The first proposition treats the balls in boxes argument. It proves more than is needed. The argument is useful for statistics such as $T(\lambda) = \sum_i M_i$ where the sum runs over the blocks of λ indexed by i and M_i is the maximum element in the ith block.

The first step in the proof is to prove the appropriate approximation conditional on m. While it would be of interest to explore this for general n, m, we content ourselves with proving what is needed for Theorem 1.2. From Theorem 2.1 the relevant values of m are $\frac{n}{\alpha_n} + \frac{c\sqrt{n}}{\log(n)}$ for large fixed values of c. This explains the choice in the next lemma.

Lemma 4.1. *Fix a large number C. Let n balls labeled $1, 2, \cdots, n$ be dropped uniformly at random into m boxes with $m = \frac{n}{\alpha_n} + \frac{c\sqrt{n}}{\log(n)}$. For $|c| \leq C$. Let*

$$D_n = \sum_{i=1}^{m} (M_i - m_i + 1)$$

with M_i the maximum label in box i and m_i the minimum label of box i. $M_i - m_i$ is omitted if box i is empty. Then $D_n = nm - 2m^2 + O_{p,C}(m^{3/2})$ uniformly in $|c| \leq C$.

Proof. Consider an infinite supply of balls labelled $1, 2, 3, \ldots$ dropped uniformly at random into m boxes. Let W_i $1 \leq i \leq m$ be the waiting time until i boxes have been filled. Thus $W_i = 1$, $W_2 - W_1$ is GEOMETRIC$(1/m)$, $W_3 - W_2$ is GEOMETRIC$(2/m)$, $\ldots$, $W_n - W_{n-1}$ is GEOMETRIC$((m-1)/m)$ and all these differences are independent. Here, if X is GEOMETRIC(θ), $\mathbf{P}(X = j) = \theta^{j-1}(1-\theta)$, $\mathbb{E}(X) = 1/\theta$, and VAR$(X) = \frac{1}{\theta}\left(\frac{1}{\theta} - 1\right)$. Let E_t be the number of empty boxes at time t and L_t be the largest ℓ so that $W_\ell \leq t$. If $L_t \leq m$ all boxes are non-empty at time t and $E_t = 0$. More generally, $L_t = m - E_t$.

The sum $\sum_{i=1}^m m_i$ is $W_1 + \cdots + W_{L_n}$. This sum may be controlled by showing that E_n is bounded with high probability and then bounding the sum by Chebychev bounds. The same argument works for $\sum_{i=1}^m M_i$. Toward this end, represent

$$E_n = \sum_{i=1}^{m} X_i \quad \text{where} \quad X_i = \begin{cases} 1 & \text{box } i \text{ is empty after } n \text{ balls} \\ 0 & \text{box } i \text{ is not empty after } n \text{ balls} \end{cases}.$$

$$\mathbb{E}(E_n) = m\left(1 - \frac{1}{m}\right)^n, \quad \text{VAR}(E_n) = m\left(1 - \frac{1}{m}\right)^n + m(m-1)\left(1 - \frac{1}{m}\right)^n - m^2\left(1 - \frac{1}{m}\right)^{2n}.$$

By elementary estimates

$$\mathbb{E}(E_n) = 1 + O\left(\frac{C}{\sqrt{n}}\right), \quad \text{VAR}(E_n) = 1 + O\left(\frac{C}{\sqrt{n}}\right). \tag{4.1}$$

Indeed, $m\left(1-\frac{1}{m}\right)^n = e^{\log(m)-\frac{n}{m}+O\left(\frac{n}{m^2}\right)}$. Using the assumption $m = \frac{n}{\alpha_n} + \frac{c\sqrt{n}}{\log(n)}$, $\log(m) = \alpha_n + O\left(\frac{c}{\sqrt{n}}\right)$, $\frac{n}{m} = \alpha_n + O\left(\frac{1}{\sqrt{n}\log(n)}\right)$. This gives the first result in (4.1), the second follows similarly. By classical results [1], E_n is approximately POISSON(1) distributed with an explicit total variation error but this is not needed.

Consider next

$$S_n = W_1 + \cdots + W_m = mW_1 + (m-1)(W_2 - W_1) + \cdots + 2(W_{m-1} - W_{m-2}) + (W_{m-1} - W_m).$$

$$\mathbb{E}(S_n) = \frac{m}{1} + \frac{m-1}{\frac{m-1}{m}} + \cdots + \frac{1}{\frac{1}{m}} = m^2 \tag{4.2}$$

$$\mathrm{VAR}(S_n) = \sum_{i=1}^{m-1}(m-i)^2\frac{m}{m-i}\left(\frac{m}{m-i}-1\right) = \sum_{i=1}^{m} mi = m\frac{m(m-1)}{2} \sim \frac{m^3}{2}. \tag{4.3}$$

Consider next the sum of the box maxima. Drop balls labelled $n, n-1, \cdots, 1$ sequentially into m boxes. If the new arrivals are at times $\widetilde{W_1}, \widetilde{W_2}, \cdots, \widetilde{W_m}$, the box maxima are $n - (\widetilde{W_1} - 1), n - (\widetilde{W_2} - 1), \ldots, n - (\widetilde{W_m} - 1)$. The sum

$$\widetilde{S_n} = \sum_{i=1}^{m} m_i = nm - \left(\widetilde{W_1} + \cdots + \widetilde{W_m}\right) + n.$$

Thus

$$\mathbb{E}(\widetilde{S_n}) = n(m+1) - m^2 \tag{4.4}$$

$$\mathrm{VAR}(\widetilde{S_n}) = m\frac{m(m-1)}{2}. \tag{4.5}$$

The random variable of interest is

$$D_n = \sum_{i=1}^{m}(M_i - m_i + 1) = \widetilde{S_n} - S_n - \sum_{i=L_n+1}^{m}\left(\widetilde{W_i} - W_i\right) + m.$$

The sum $\sum_{i=L_n+1}^{m}\widetilde{W_i} \leq E_n\widetilde{W_m}$. From the coupon collectors problem $\widetilde{W_m}$ is of stochastic order $m\log(m) \sim n$ and E_n is stochastically bounded. A similar argument holds with $\widetilde{W_i}$ replaced by W_i. It follows that the sum $\sum_{i=L_n+1}^{m}\left(\widetilde{W_i} - W_i\right) = O_p(n)$. Combining terms

$$D_n = \widetilde{S_n} - \widetilde{S_n} + m + O_p(n) = nm + \left(\widetilde{S_n} - \mathbb{E}(\widetilde{S_n})\right) - (S_n - \mathbb{E}(S_n)) + O_p(n).$$

By Chebychev's inequality $|S_n - \mathbb{E}(S_n)|$ and $\left|\widetilde{S_n} - \mathbb{E}(\widetilde{S_n})\right|$ are both $O_p(m^{3/2})$. It follows that $D_n = nm - 2m^2 + O_p(m^{3/2})$. □

Proof of Theorem 1.2. To finish the proof of Theorem 1.2 note that conditional on M

$$d(\lambda) = nM - 2M^2 + O_p(M^{3/2}).$$

This is weak star close to

$$N\left(\frac{n^2}{\alpha_n} - \frac{2n^2}{\alpha_n^2}, \frac{n^{3/2}}{\alpha_n}\right) + O_p\left(\frac{n}{\alpha_n}\right)^{3/2}.$$

Since $(n/\alpha_n)^{3/2}$ is much smaller than the standard deviation of the normal, this is in turn close to

$$N\left(\frac{n^2}{\alpha_n} - \frac{2n^2}{\alpha^2}, \frac{n^{3/2}}{\alpha_n}\right).$$

This completes the proof. □

5. Proof of Theorem 1.3

This section contains the proof of Theorem 1.3. Our approach is to compare the crossing statistic to the dimension statistic, which by Theorem 1.2 is known to be normally distributed.

Proof. To analyze the distribution of the crossing number, we compare it to the dimension index. We do this by producing a uniform random set partition λ in the following unusual way:

- Pick $M = m$, say,
- Pick a uniform random set partition μ for that m according to Stam's algorithm
- Let λ be a uniform random set partition conditional on the event that the set of minimum elements of blocks of λ is the set of minimum elements of blocks on μ and that the set of maximum elements of blocks of λ equals the set of maximum elements of blocks of μ.

This second step can be accomplished in the following way, assigning the elements of $[n]$ to blocks in order. We begin with no blocks and add elements to blocks one at a time, sometimes creating new blocks. If an element k, where k is the maximum element of some block of μ is added to a block in λ, we declare that block *closed*. After having assigned the first k elements to blocks in λ, we assign $k+1$ to a uniform random un-closed block, unless $k+1$ is the minimum element of some block of μ, in which case we assign $k+1$ to a new block of λ. This procedure clearly produces a uniform λ subject to the restriction on the minimum and maximum elements of blocks.

On the other hand, this method of choosing λ gives us a reasonable way to analyze $cr(\lambda)$. In particular, the crossing number of λ equals the number of pairs of a $j \in [n]$ and a block B in λ with

- $j \notin B$
- j not the first element of its block
- $\max(B) > j$
- The element of B immediately preceding j is larger than the element of j's block immediately preceding j

We note that this is easy to analyze given the procedure above for choosing λ. Suppose that when k is being added to λ that there are a_k blocks of λ currently open. If k is the first element of its block, then we have no crossings with $j = k$. Otherwise, we claim that the number of crossings with $j = k$ (which we call X_k) has distribution given by the discrete uniform random variable on $[0, a_k - 1]$. In particular, if the open blocks are $B_1, \ldots, B_{a_k}$ whose element immediately preceding k is $m_1 < m_2 < \ldots < m_{a_k}$, then $X_k = k - i$ if k is assigned to block B_i. Note furthermore, that the a_k are determined by μ and that the X_k

are independent. Since $cr(\lambda) = \sum_k X_k$ is a sum of independent random variables, it is easy to see that conditioned on μ that with high probability $cr(\lambda)$ is weak star close to

$$N\left(\sum_{k \text{ not a minimum}} \frac{a_k - 1}{2}, \sqrt{\sum_{k \text{ not a minimum}} \frac{a_k^2 - 1}{12}}\right).$$

We note that a given block contributes to a_k if and only if k is between is minimum and maximum values. Therefore,

$$\sum_{k=1}^{n}(a_k - 1) = \left(\sum_{i=1}^{m} M_i - m_i\right) - n = nm - 2m^2 + O_p(m^{3/2}).$$

On the other hand, the sum over a_k at the start of blocks is the number pairs of blocks that overlap. Note that for $m = n/\alpha_n + o_p(n/\log^2(n))$, that any given block has $n/2$ between its minimum and maximum with probability $1 - O(m^{-1/2})$. Thus, for m in this range, the expected number of pairs of non-overlapping blocks is $O(m^{3/2})$. Thus,

$$\sum_{k \text{ not a minimum}} \frac{a_k - 1}{2} = nm/2 - 5m^2/4 + O_p(m^{3/2}).$$

It is also easy to see that

$$\sum_{k \text{ not a minimum}} (a_k^2 - 1) = n^2 m(1 + o_p(1)) = \frac{n^3}{\alpha_n}(1 + o_p(1)).$$

Therefore, with probability approaching 1 over the choice of m, the distribution of λ conditioned on m is close to

$$N\left(nm/2 - 5m^2/4, \frac{n^{3/2}}{\sqrt{12\alpha_n}}\right).$$

This can be rewritten (up to small error) as the sum of $(n/2 + 5n/(2\alpha_n))(m - n/\alpha_n)$ and a variable with distribution

$$N\left(\frac{n^2}{2\alpha_n} - \frac{5n^2}{4\alpha_n^2}, \frac{n^{3/2}}{\sqrt{12}\alpha_n}\right).$$

On the other hand, by Theorem 2.1, $(n/2 + 5n/(2\alpha_n))(m - n/\alpha_n)$ is approximated by an independent normal weak star close to

$$N\left(0, \frac{n^{3/2}}{2\alpha_n}\right).$$

Thus, the distribution of $cr(\lambda)$ is close in cdf distance to this sum of independent normals, which is given by

$$N\left(\frac{n^2}{2\alpha_n} - \frac{5n^2}{4\alpha_n^2}, \frac{n^{3/2}}{\sqrt{3}\alpha_n}\right).$$

This completes the proof. □

References

[1] S. Chatterjee, P. Diaconis, and E. Meckes, *Exhchangable pairs and Poisson approximation.* Prob. Surveys (2005).

[2] W. Y. C. Chen, E. Y. P. Deng, R. R. X. Du, R. P. Stanley, *Crossings and nestings of matchings and partitions.* Trans. Amer. Math. Soc. **359** (4) (2007), 1555–1575.

[3] B. Chern, P. Diaconis, D. M. Kane, and R. C. Rhoades, *Closed expressions for set partition statistics.* Research in the Mathematical Sciences **1** 2, (2014).

[4] L. Chen, L. Goldstein, and Q-M. Shao, *Normal approximation by Stein's method*, Springer Verlag (2010).

[5] L. Chen, Q-M. Shao, *Normal Approximation under local dependence*, Ann. Probab. Volume 32, Number 3 (2004), 1727-2303

[6] P. Diaconis and I. M. Isaacs, *Supercharacters and superclasses for algebra groups.* Trans. Amer. Math. Soc. **360** (2008), 2359–2392.

[7] G. Dobinski, *Summirung der reine* $\sum n^m/m!$ *for* $m = 1, 2, 3, 4, 5, \ldots$. Grunet. Archiv. **61** (1877), 333–336.

[8] N. G. de Bruijn, *Asymptotic Methods in Analsysis.* Dover, N.Y.

[9] B. Fristedt, *The structure of random partitions of large sets.* Technical Report Dept. of Mathematics, University of Minnesota (1987), 86–154.

[10] G. H. Hardy, *Divergent series.* Oxford Univ. Press (1991).

[11] H.K. Hwang, *On Convergence Rates in the Central Limit Theorems for Combinatorial Structures.* European J. Combin. 19 (1998), no. 3, 329–343.

[12] A. Kasraoui, *Average values of some Z parameters in random set partitions.* Electronic J. Combinatorics, Volume 18, Issue 1 (2011) #P228.

[13] A. Kasraoui, *On the limiting distribution of some numbers of crossings in set partitions.* arxiv:1301,6546 (2013).

[14] A. Kasraoui and J. Zeng, *Distribution of crossings, nestings and alignments of two edges in matchings and partitions.* Electron. J. Combin. **13** (2006), no. 1, Research Paper 33, 12 pp.

[15] A. Knopfmacher, T. Mansour, and S. Wagner, *Records in set partitions.* Electronic Journal of Combinatorics **17** (2010) R109 (14pp.).

[16] D. Knuth, *The art of computer programming.* Vol 4A. Addison-Wesley.

[17] E. Lehman, *Elemnts of large sample theory.* Springer (1999).

[18] T. Mansour, *Combinatorics of set partitions.* Discrete Mathematics and its Applications (Boca Raton). CRC Press, Boca Raton, FL, 2013.

[19] A. Nijenhuis and H. S. Wilf, *Combinatorial algorithms. For computers and calculators.* Second edition. Computer Science and Applied Mathematics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1978.

[20] J. Pitman, *Some probabilistic aspects of set partitions.* Amer. Math. Monthly, **104** (1997), 201–209.

[21] J. Pratt, *On a general concept of "in probability".* Ann. Math. Statistics. **30** (1958), 549–558.

[22] R. Serfling, *Approximation theorems of mathematical statistics.* Wiley (2001).

[23] N. Sloane, Online Encyclopedia of Integer Sequences. `http://oeis.org/`

[24] A. J. Stam, *Generation of random partitions of a set by an urn model.* J. Combin. Theory A, **35** (1983), 231–240.

[25] R. P. Stanley, *Enumerative combinatorics. Volume 1.* Second edition. Cambridge Studies in Advanced Mathematics, 49. Cambridge University Press, Cambridge, (2012).

Stanford University, Department of Electrical Engineering, Stanford, CA 94305
E-mail address: bgchern@stanford.edu

Stanford University, Department of Mathematics and Statistics, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA
E-mail address: diaconis@math.stanford.edu

Stanford University, Department of Mathematics, Bldg 380, Stanford, CA 94305
E-mail address: dankane@math.stanford.edu

Center for Communications Research, Princeton, NJ 08540
E-mail address: rob.rhoades@gmail.com