FROM ALGORITHM TO THEOREM

SOME PHILOSOPHY SUPPOSE THAT YOU MEET A NEW COMBINATORIAL OBJECT: 'PARKING FUNCTIONS' OR 'THE LAH NUMBERS' OR 'BELL NUMBERS' HOW DO YOU BEGIN TO COME TO TERMS AND WORK WITH THEM? STANDARD MOVES ARE

- HOW MANY ARE THERE? (BOTH FOR SMALL n AND ASYMPTOTICALLY)
- WHAT ARE NATURAL FEATURES
- WHICH OBJECTS HAVE EXTREME FEATURES (MAX, MIN)
- IS THERE A NATURAL DISTANCE BETWEEN TWO OBJECTS (METRICS).

I WOULD LIKE TO ADD ANOTHER BASIC QUESTION

"HOW DO I PICK A RANDOM OBJECT IN MY SET (ON A COMPUTER)"

THIS TO BE FOLLOWED BY 'WHAT DOES A TYPICAL OBJECT LOOK LIKE?"

THE MAIN POINT OF THIS WEEKS LECTURES IS THAT AN EXPLICIT ALGORITHM CAN OFTEN BE HARNESSED TO MAKE PROOFS OF THEOREMS. THUS GOING FROM ALGORITHM TO THEOREM.

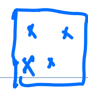
THERE IS A DIFFERENT LINE OF WORK (QUITE DIFFERENT) GOING FROM THEOREM TO ALGORITHM. THIS WEEK WE WILL STICK TO A TO T.

WE HAVE ALREADY SEEN SEVERAL EXAMPLES OF ALGORITHMS YIELDING THEOREMS IN OUR WORK ON RANDOM PERMUTATIONS

- THE SUBGROUP ALGORITHM (DIACONIS - SHAHSHAHANI) SPECIALIZED TO S_n BECOMES UNIFORMLY PICK $1 \le I_1 \le n$, TRANSPOSE $(1, I_1)$, THEN PICK I_2 $2 \le I_2 \le n$ UNIFORMLY AND TRANSPOSE $(2, I_2), \cdots$, PICK I_{n-1} UNIFORMLY IN $n-1 \le I_{n-1} \le n$ AND UNIFORMLY TRANSPOSE $(n-1, I_{n-1})$. IF σ_i IS THE iTH TRANSPOSITION $\sigma = \sigma_1 \sigma_2 \cdots \sigma_{n-1}$ IS UNIFORM ON S_n. THERE ARE MYRIAD USES INCLUDING: LET $X_i = \begin{cases} 1 & 1/i \\ 0 & - \end{cases}$. THEN $X_1X_2 + X_2X_3 + \cdots X_{n-1}X_n \cdots \sim \text{Poiss}(1)$.

- Building σ up sequentially in cycle notation, at stage i putting i to the right of a randomly picked previously placed element or using it to start a new cycle on its own (with probability $1/i$). This was useful in getting the distribution of the number of cycles and as part of the Chinese Restaurant Process $1,\ (12),\ (12)(3),\ (142)(3), \ldots$

- Placing $1, 2, 3, \ldots$ down in a row sequentially, inserting i in one of the i available places (before all, between or to the right of all. This was useful in getting the distribution of the number of inversions $1,\ 21,\ 231,\ 2431, \ldots$

- Putting $(x_1, y_1), \ldots, (x_n, y_n)$ $\quad 0 \le x_i, y_i \le 1$ down in the unit square uniformly at random. Then, ordered by x values define $\sigma(i)$ by the rank of y_i among all the y values. This was useful in deriving the distribution of the longest increasing subsequence.  $\leftrightarrow \begin{matrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{matrix}$

There are many further examples coming. Of course, one benefit of having an efficient algorithm for sampling -- you can draw a random sample and make a histogram of the feature(s) of interest (for the values of n of interest).

I'm going to begin by treating an example in some detail. Here, a non-standard sampling algorithm really saved the day.

(2) SET PARTITIONS. LET $B(n)$, THE n^{th} BELL NUMBERS, DENOTE THE NUMBER OF PARTITIONS OF $[n]$ INTO DISJOINT BLOCKS -- ORDER WITHIN OR BETWEEN BLOCKS DOESN'T MATTER. THUS, $B(3) = 5$ FROM

$$1/2/3, \quad 12/3, \quad 13/2, \quad 1/23, \quad 123$$

$B(4) = 15$, $B(5) = 52$ (!). THE FIRST FEW ARE (A000110)

n	0	1	2	3	4	5	6	7
$B(n)$	1	1	2	5	15	52	203	4140

THERE IS A HUGE CLASSICAL LITERATURE ON SET PARTITIONS; THE CHAPTER IN GRAHAM-KNUTH-PATASHNIK'S 'CONCRETE MATHEMATICS' OR RICHARD STANLEY'S 'ENUMERATIVE COMBINATORICS' OR THE 700 PAGE BOOK BY MANSOUR SHOULD GET YOU STARTED. THERE IS ALSO A FAIRLY RICH PROBABILISTIC LITERATURE AROUND 'PICK A SET PARTITION AT RANDOM, WHAT DOES IT LOOK LIKE?' THE WONDERFUL ARTICLE BY FRISTEDT [ ] USES CONDITIONED LIMIT THEORY MUCH AS WE HAVE ABOVE TO GIVE A UNIFIED TREATMENT OF MANY NATURAL FEATURES (NUMBER OF BLOCKS, NUMBER OF BLOCKS OF SIZE i, LARGEST BLOCK, ...). REFERENCES TO MORE ESOTERIC STATISTICS ARE IN [ ] [ ].

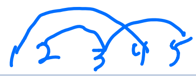
ALL OF THESE REFERENCES ONLY STUDY BLOCK SIZES. FOR QUESTIONS ARISING IN THE REPRESENTATIONS OF THE GROUP $U_n(q)$ -- UNI-UPPER TRIANGULAR MATRICES WITH ENTRIES IN $\mathbb{F}_q$, I NEEDED TO KNOW ABOUT FEATURES THAT DEPEND ON THE ENTRIES WITHIN A BLOCK; FOR EXAMPLE, FOR THE i^{th} BLOCK, LET M_i BE THE LARGEST NUMBER IN THE BLOCK, m_i THE SMALLEST NUMBER IN THE BLOCK. THE DIMENSION INDEX FOR THE SET PARTITION λ IS

$$d(\lambda) = \sum_i (M_i - m_i + 1) - n$$

A DIFFERENT STATISTIC, NOT CAPTURED BY BLOCK SIZES IS THE CROSSING NUMBER

$$C(\lambda) = \# \text{ CROSSINGS OF } \lambda$$

HERE, IF YOU GRAPH λ BY PUTTING AN ARC FROM i TO j IF j FOLLOWS i IN A BLOCK (ORDER THE NUMBERS IN A BLOCK FOR CONVENIENCE)

EXAMPLE $\lambda = \{1,4\},\{2,3,5\} \leftrightarrow$  $c(\lambda) = 2$.

ONE MORE NON STANDARD STATISTIC: CODE λ UP AS $x_1 x_2 \cdots x_n$ WITH $x_i = j$ IF i IS IN BLOCK j OF λ. SO $\{1,4\},\{2,3,5\} \leftrightarrow 12212$. LET

$$L(\lambda) = |\{i : x_{i+1} = x_i\}| \quad \text{THE NUMBER OF LEVELS OF } \lambda.$$

FOR EACH OF THESE, ONE MAY ASK ABOUT THE MEAN, VARIANCE AND LIMITING DISTRIBUTION. STANDARD TECHNIQUES (MOMENTS, STEINS METHOD, GENERATING FUNCTIONS, ---) BROKE DOWN; THE 'ALGORITHMS TO THEOREMS' APPROACH WORKED AND GAVE A CLEAR HEURISTIC PICTURE OF WHAT A TYPICAL SET PARTITION LOOKS LIKE.

(3) STAMS ALGORITHM WRITE $\Pi(n)$ FOR THE SET PARTITIONS OF n, SO $B(n) = |\Pi(n)|$. HOW DO YOU CHOOSE $\lambda \in \Pi(n)$ UNIFORMLY? ITS NOT EVEN CLEAR HOW LARGE $B(n)$ IS (ROUGHLY $\left(\frac{n}{\log n}\right)!$). THERE IS A (STRANGE) FORMULA OF DOBINSKI:

$$B_n = \frac{1}{e} \sum_{m=1}^{\infty} \frac{m^n}{m!} \qquad \text{(WHAT'S } e \text{ DOING HERE)}$$

INDEED, TAKE A LOOK AT THIS; IT SAYS $1 = B_1 \overset{?}{=} \frac{1}{e} \sum_{1}^{\infty} \frac{1}{(m-1)!}$; OK, $2 \overset{?}{=} \frac{1}{e} \sum_{m=1}^{\infty} \frac{m}{(m-1)!}$?

OK, WHATEVER ITS TRUE, DOBINSKI SAYS THAT

$$(1) \qquad \mu_n(m) = \frac{1}{e B_n} \frac{m^n}{m!}$$

IS A PROBABILITY DISTRIBUTION ON $\{1,2,3,\cdots\}$. STAM'S ALGORITHM USES $\mu_n(m)$ TO GIVE AN ELEGANT ALGORITHM FOR CHOOSING A RANDOM ELEMENT OF $\Pi(n)$:

STAM'S ALGORITHM

(1) CHOOSE m FROM μ_n

(2) DROP n LABLED BALLS UNIFORMLY INTO m BOXS

(3) FORM λ WITH i AND j IN THE SAME BLOCK IFF THEY ARE IN THE SAME BOX.

OF COURSE, AFTER CHOOSING m AND DROPPING BALLS, SOME OF THE BOXS MAY BE EMPTY. STAM SHOWS THAT THE NUMBER OF EMPTY BOXS HAS (EXACTLY) A POISSON DISTRIBUTION AND IS INDEPENDENT OF λ. STAM'S PAPER IS VERY CLEARLY WRITTEN BUT THE ALGORITHM IS STILL MAGICAL. THE NICE ARTICLE BY JIM PITMAN (1997) 'SOME PROBABALISTIC ASPECTS OF SET PARTITIONS' AMER. MATH. MONTHLY, MAKES MANY CONNECTIONS, BUT (a) I STILL FIND IT MAGICAL (b) IT'S BEEN HARD TO GENERALIZE (WHAT DISTRIBUTION ON $\Pi(n)$ IS INDUCED IF BOSE-EINSTEIN ALLOCATION IS USED IN STEP (2)?).

HERE IS AN ILLUSTRATION OF USING STAMS ALGORITHM TO COMPUTING MOMENTS

FOR M IN (1): FOR $-n < \alpha < \infty$

$$(2) \qquad E(m^{\alpha}) = \frac{1}{eB_n} \sum_{m=1}^{\infty} \frac{m^{n+\alpha}}{m!} = \frac{B_{n+\alpha}}{B_n}$$

NEXT CONSIDER $L(\lambda)$, THE NUMBER OF LEVELS OF λ, DEFINED ABOVE. GIVEN m, $L(\lambda) = X_1 + X_2 + \cdots + X_{n-1}$ WHERE X_i IS THE INDICATOR OF THE EVENT THAT BALLS i AND $i+1$ ARE DROPPED INTO THE SAME BOX. BY INSPECTION, THE X_i ARE INDEPENDENT WITH $P(X_i = 1) = \frac{1}{m}$. THUS

$$E_n(L(\lambda)) = E_n(E(L(m)) = E_n\left(\frac{n-1}{m}\right) = (n-1)\frac{B_{n-1}}{B_n} \sim \log n$$

THE STANDARD IDENTITY $\mathrm{VAR}(Z) = E(\mathrm{VAR}(Z|W)) + \mathrm{VAR}(E(Z|W))$ GIVES

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

PLEASE NOTE SEVERAL THINGS: UP TO DIVISION BY B_n, THE MOMENTS OF L

ARE SHIFTED BELL POLYNOMIALS; POLYNOMIALS IN $B_n, B_{n-1} \cdots B_{n-k}$ WITH COEFFICIENTS THAT ARE POLYNOMIALS IN n. THIS HOLDS ALSO FOR THE STATISTICS $d(\lambda)$, $C(\lambda)$ AND ANY REASONABLE FUNCTION OF THE COUNTS $N_i(\lambda)$ (# i-BLOCKS). INDEED IT HOLDS VERY GENERALLY; SEE MY PAPER ON MOMENTS OF FUNCTIONS OF SET PARTITIONS WITH CHERN, KANE AND RHODES.

MUCH MORE IMPORTANT STAM'S ALGORITHM GIVES AN ACCURATE, SIMPLE, HEURISTIC PICTURE OF 'WHAT A TYPICAL SET PARTITION LOOKS LIKE': TO GOOD APPROXIMATION IT LOOKS LIKE THIS

'DROP n BALLS UNIFORMLY AT RANDOM (MULTINOMIAL ALLOCATION) INTO $m = \frac{n}{\log n}$ BOXES. PUT i, j IN SAME BLOCK IF THEY ARE IN THE SAME BOX. THAT'S IT!. THIS MEANS THAT ALL OF OUR ACCUMULATED WISDOM ABOUT BALLS IN BOXES CAN BE HARNESSED TO SUGGEST THEOREMS.

(4) THE DISTRIBUTION OF L. I WANT TO USE STAM'S ALGORITHM TO PROVE A LIMIT THEOREM. AN EASY CASE IS THE LIMITING DISTRIBUTION OF THE NUMBER OF LEVELS L DEFINED ABOVE. RECALL THAT, GIVEN m, $L = X_1 + \cdots + X_{n-1}$ WITH X_i BINARY VARIATES, INDEPENDENT, WITH $P(X_i = 1) = 1/m$.

BECAUSE OF THE NEAT MOMENT FORMULA (2) ABOVE

$$\mu_n^m = E(m) = \frac{B_{n+1}}{B_n} \sim n/\log n, \quad (\sigma_n^m)^2 = \mathrm{VAR}_n(m) = \frac{B_{n+2}}{B_n} - \left(\frac{B_{n+1}}{B_n}\right)^2 \sim n/(\log n)^2,$$

AND, NORMALIZED BY ITS MEAN AND STANDARD DEVIATION $m \Rightarrow \mathcal{N}(0,1)$

HERE, I HAVE USED STANDARD ASYMPTOTICS OF BELL NUMBERS. TO STATE THEM MORE CAREFULLY, DEFINE α_n AS THE UNIQUE POSITIVE REAL SOLUTION OF $ue^u = n+1$ (SO $\alpha_n = \log n - \log\log n + o(1)$). THEN (SEE e.g. [de BRUIJN, ASYMPTOTIC ANALYSIS]),

$$\frac{B_{n+h}}{B_n} = \frac{(n+h)!}{n!\,\alpha_n^h}\left(1+O\left(\frac{h}{n\alpha_n}\right)\right)$$

$$\frac{\alpha_{n+h}}{\alpha_n} = 1 + O\left(\frac{h}{n\log n}\right)$$

WE WANT TO PROVE

<u>THEOREM</u> THE NUMBER OF LEVELS $L(\lambda)$ HAS $E_n(L) = \mu_n^L \sim \log n$, $\mathrm{Var}_n(L) = (\sigma_n^L)^2 \sim \log n$ AND, NORMALIZED BY ITS MEAN AND VARIANCE, $\frac{L-\mu_n}{\sigma_n} \Rightarrow \mathcal{N}(0,1)$.

<u>PROOF</u> THE MOMENTS OF L ARE COMPUTED ABOVE. CONDITIONAL ON m, $L = X_1+\dots+X_{n-1}$ WITH X_i $1\le i\le n-1$ INDEPENDENT 0/1 VARIATES HAVING $P(X_i=1) = \frac{1}{m}$. THUS

$$E(L\,|\,m) = \frac{n-1}{m}, \quad \mathrm{Var}(L\,|\,m) = \frac{n-1}{m}\left(1-\frac{1}{m}\right)$$

AND, NORMALIZED BY ITS MEAN AND STANDARD DEVIATION, L HAS A LIMITING NORMAL DISTRIBUTION, PROVIDED $n/m \to \infty$. THIS SHOWS, BY (a):

$$\text{(b)} \qquad \frac{n}{m_n} = \alpha_n + O_p\left(\frac{1}{\sqrt{n}}\right).$$

MORE PRECISELY, WRITE $m_n = \mu_n^m + Z_n\sigma_n^m$ (SO $Z_n = \frac{m_n-\mu_n^m}{\sigma_n^m}$).

THEN

$$\frac{n}{m_n} = \frac{n}{\mu_n^m + Z_n \sigma_n^m} = \frac{n}{\mu_n^m\left(1+\frac{Z_n}{\mu_n^m}\sigma_n^m\right)} = \frac{n}{\mu_n^m}\left(1+\frac{Z_n\sigma_n^m}{\mu_n^m}+O\left(\left(\frac{Z_n\sigma_n^m}{\mu_n^m}\right)^2\right)\right).$$

OUR MOMENT CALCULATIONS GIVE

$$\frac{n}{\mu_n^m} = \alpha_n + O\left(\frac{\alpha_n}{n}\right), \quad \frac{\sigma_n^m}{\mu_n^m} = O\left(\frac{1}{\sqrt{n}}\right), \quad Z_n = O_p(1) \text{ SO (b) FOLLOWS.}$$

WITH PROBABILITY CLOSE TO 1, GIVEN m, L IS CLOSE TO A GAUSSIAN WITH MEAN $\mu_n^m = \frac{n-1}{m} = \alpha_n + O_p(1/\sqrt{n})$, $\sigma_n^m = \sqrt{\alpha_n} + O_p(1/\sqrt{n})$. THUS, WITH HIGH PROBABILITY $L(\partial)$ IS (WEAK STAR) CLOSE TO $N(\alpha_n, \sqrt{\alpha_n})$ UNCONDITIONALLY. $\square$

THE ARGUMENTS FOR THE LIMITING DISTRIBUTION OF $d(\partial)$ AND $C(\partial)$ ARE SIMILAR BUT REQUIRE SOME NEW IDEAS. SEE MY PAPERS WITH CHERN-KANE-RHODES. THE POINT FOR TODAY IS; THE DISTRIBUTION OF $C(\partial)$ WAS AN OPEN PROBLEM. STANDARD METHODS DIDN'T WORK, AND HAVING AN ALGORITHMIC WAY OF CHOOSING ∂ SAVED THE DAY.

(6) <u>SOME MOTIVATION</u> ONE OF THE THINGS I HOPE YOU LEARN IN THIS COURSE IS THAT COMBINATORIAL PROBLEMS (EVEN INTERESTING NEW ONES) ARISE IN APPLICATIONS. THE STATISTICS $d(\partial)$, $C(\partial)$, $L(\partial)$ SEEM STRANGE; WHO CARES?

MY INTEREST STARTED WITH A RANDOM WALK PROBLEM ON $U_n(q)$ THE GROUP OF UNI-UPPER TRIANGULAR MATRICES WITH ENTRIES IN $\mathbb{F}_q$. THIS IS THE SYLOW-p SUBGROUP OF GL_n AND A 'UNIVERSAL p-GROUP' JUST AS EVERY FINITE GROUP OF ALL SUFFICIENTLY LARGE SYMMETRIC GROUPS S_n, EVERY p-GROUP IS A SUBGROUP OF $U_n(p)$.

I was studying the simplest random walk on $U_n(p)$: Start at id, each step, pick a row at random, and add or subtract it to the row above. How long does this walk take to get random? One approach to studying random walk on groups is to close up the generating set under conjugation, then use character theory to study the conjugation invariant walk, finally use comparison theory to go back to the original walk.

OK, what are the characters (and conjugacy classes) of $U_n(p)$? It turns out that nobody knows and in a real sense, you can prove that nobody will ever know (!). What to do? Well, it turns out that there is a cruder 'super character theory' due to Carlos André (and now many others). This lumps together certain conjugacy classes and sums certain characters. This new theory is 'nice' there is even a closed formula for a super-character on a superclass.

It turns out that the superclasses (and characters) are indexed by set partitions (just like the characters of S_n are indexed by partitions). There are fascinating combinatorial rules (analogs of tableaux combinatorics) and a full analog of the duality between the representation theory of the symmetric group and symmetric function theory. For superclass theory, symmetric functions are replaced by symmetric functions in non-commuting variables. The dimension of the super character indexed by $\lambda \in \Pi(n)$ is $q^{d(\lambda)}$. The super characters are orthogonal with $\langle \chi_\lambda | \chi_\mu \rangle = \delta_{\lambda\mu} q^{c(\lambda)}$.
This is the motivation for studying $d(\lambda)$ and $c(\lambda)$.