Some Random Collections of Finite Subsets

J.F.C. Kingman

Let $X_1, X_2, \ldots$ be independent random variables having the same continuous distribution function F. For any n, define a family $\mathcal{A}_n$ of subsets of

$$I_n = \{1, 2, \ldots, n\} \tag{1}$$

by the following recipe: a subset A of I_n belongs to $\mathcal{A}_n$ if and only if, whenever $i < j$ and $i, j \in A$,

$$X_i < X_j. \tag{2}$$

The randomness of the X_i means that $\mathcal{A}_n$ is a random family of subsets of I_n, and it is clear that the distribution of $\mathcal{A}_n$ does not depend on F.

Hammersley (1972) studied the problem, proposed by Ulam, of finding at least the asymptotic distribution of the size of the largest set in $\mathcal{A}_n$, the random variable

$$L_n = \max\{|A|; A \in \mathcal{A}_n\}. \tag{3}$$

He showed that there is a constant c such that, with probability one,

$$L_n \sim c\sqrt{n} \tag{4}$$

as $n \to \infty$.

Hammersley conjectured that $c = 2$, but he was only able to prove that

$$\tfrac{1}{2}\pi \le c \le e. \tag{5}$$

These bounds are improved in Kingman (1973) to

$$(8/\pi)^{1/2} \le c \le \epsilon, \tag{6}$$

where $\epsilon = 2.49\ldots$ can be expressed as

$$\epsilon = \xi^{-1/2}(1-\xi)^{-1/2}, \tag{7}$$

where ξ is the positive root of

$$1 - \xi = e^{-2\xi}. \tag{8}$$

Much later Hammersley's conjecture was proved by Veršik and Kerov (1977), but by complex arguments very specific to the Ulam problem. By contrast, the arguments of Hammersley (1972) and Kingman (1973) are relatively crude, and for this reason apply to other problems. For instance, the upper bound in (5) follows from the obvious fact that, if $A \subset I_n$ has $|A| = r$, then

$$P(A \in \mathcal{A}_n) = \frac{1}{r!}. \tag{9}$$

This implies that the expected number of sets of size r in $\mathcal{A}_n$ is

$$\binom{n}{r}\frac{1}{r!},$$

and this is an upper bound for the probability that there is at least one such set. Thus

$$P(L_n \geq r) \leq \binom{n}{r}\frac{1}{r!}, \tag{10}$$

and Stirling's formula easily shows that this tends to zero as $n, r \to \infty$ in such a way that

$$\liminf rn^{-1/2} > e.$$

The sharpening in (6) is only a little more difficult, and makes use (in a way which will be described below) of the fact that $\mathcal{A}_n$ is *hereditary*: if $A \in \mathcal{A}_n$ and $A' \subset A$, then $A' \in \mathcal{A}_n$.

My interest in these arguments was revived when I encountered, in the context of a genetical problem, another random family with similar properties. Let Y_{ij} $(i, j = 1, 2, \dots)$ be random variables with a common continuous distribution function F. The Y_{ij} for $i \leq j$ are mutually independent, but the symmetry condition

$$Y_{ji} = Y_{ij} \tag{11}$$

is imposed. The family $\mathcal{A}_n$ is now defined by the requirement that $A \subset I_n$ belongs to $\mathcal{A}_n$ if and only if, for all $i, j \in A$,

$$Y_{ij} \geq \tfrac{1}{2}(Y_{ii} + Y_{jj}). \tag{12}$$

Clearly $\mathcal{A}_n$ is hereditary.

It is shown in Kingman (1988) that, if F corresponds to a uniform distribution, then

$$P(A \in \mathcal{A}_n) \leq \frac{1}{r!} \tag{13}$$

for any A of size r. This allows the arguments of the earlier papers to be carried through, to show that the size of the largest set in $\mathcal{A}_n$ is at most

$\epsilon\sqrt{n}$ for large n. Although crude, this result is of considerable significance in the genetical context. It is however specific to the particular uniform distribution, and the probability

$$P_r(F) = P(A \in \mathcal{A}_n) \tag{14}$$

depends on the choice of F (but not of course on the value of n). For some distributions the bound (13) can be improved; if F corresponds to a negative exponential distribution it is easy to compute that

$$P_r(F) = \left(\frac{2}{r+1}\right)^r \sim \frac{(2\pi r)^{1/2} e^{-1}}{r!}. \tag{15}$$

This ought to make it possible to improve on the coefficient ϵ in the upper bound. On the other hand, there are distributions for which $P_r(F)$ is of larger order than for the uniform distribution, and one may ask whether some cruder upper bound may then be established. Both of these questions are answered by the following theorem.

THEOREM 1. *For each n, let $\mathcal{A}_n$ be a random subset of $I_n = \{1, 2, \dots, n\}$ having the hereditary property*

$$A \in \mathcal{A}_n, A' \subset A \Rightarrow A \in \mathcal{A}_n. \tag{16}$$

Suppose that, for some constant α, and for sufficiently large n, r,

$$P(A \in \mathcal{A}_n) \leq \alpha^r / r! \tag{17}$$

for every $A \subset I_n$ of size $|A| = r$. Then the size of the largest set in $\mathcal{A}_n$,

$$L_n = \max\{|A|; A \in \mathcal{A}_n\}, \tag{18}$$

satisfies

$$\limsup_{n\to\infty} L_n n^{-1/2} \leq \alpha^{1/2}\epsilon \tag{19}$$

with probability one.

This formulation assumes that the $\mathcal{A}_n$ are all defined on the same probability space. If not, the same argument shows that (19) holds in probability.

PROOF: If $s \leq r \leq n$, the inequality $L_n \geq r$ implies that there is at least one set of size r in $\mathcal{A}_n$. The hereditary property shows that all subsets of this set are in $\mathcal{A}_n$, so that $\mathcal{A}_n$ contains at least $\binom{r}{s}$ sets of size s. But, by (17), the number of sets of size s in $\mathcal{A}_n$ has expectation at most

$$\binom{n}{s}\frac{\alpha^s}{s!},$$

so that

$$\binom{r}{s} P(L_n \geq r) \leq \binom{n}{s} \frac{\alpha^s}{s!}.$$

Hence

$$\log P(L_n \geq r) \leq \log n! - \log s! - \log(n-s)! + s\log\alpha - \log r! + \log(r-s)!.$$

In this inequality let $r, s, n \to \infty$ in such a way that $r \sim \rho n^{1/2}, s \sim \sigma n^{1/2}$ for constants $0 < \sigma < \rho$. Then Stirling's formula yields, after simplification, the inequality

$$\log P(L_n \geq r) \leq -\{\rho \log\rho + \sigma\log\sigma - (\rho-\sigma)\log(\rho-\sigma) \\ - \sigma\log\alpha - 2\sigma + o(1)\}n^{1/2}.$$

Hence, by the Borel-Cantelli lemma, $L_n \geq \rho n^{\frac{1}{2}}$ for only finitely many n, so long as

$$\rho\log\rho + \sigma\log\sigma - (\rho-\sigma)\log(\rho-\sigma) - \sigma\log\alpha - 2\sigma > 0. \tag{20}$$

It follows that, with probability one,

$$\limsup L_n n^{-1/2} \leq \rho, \tag{21}$$

so long as $\sigma < \rho$ can be chosen to satisfy (20). Putting $\sigma = \lambda\rho$ for $0 < \lambda < 1$, (20) becomes

$$\log\lambda - (\lambda^{-1} - 1)\log(1-\lambda) > 2 + \log\alpha - 2\log\rho. \tag{22}$$

The best choice of λ is that which maximises the left hand side; differentiation gives the equation

$$2\lambda + \log(1-\lambda) = 0,$$

and comparison with (8) shows that $\lambda = \xi$. With this value of λ, (22) becomes

$$2 + \log\xi + \log(1-\xi) > 2 + \log\alpha - 2\log\rho,$$

or

$$\rho > \{\alpha/\xi(1-\xi)\}^{1/2} = \alpha^{1/2}\epsilon.$$

Thus (21) holds for all $\rho > \alpha^{1/2}\epsilon$, and (19) is proved. ■

For example, if $\mathcal{A}_n$ is defined by (12), and if (for some constant α depending on F)

$$P_r(F) \leq \frac{\alpha^r}{r!} \tag{23}$$

for large r, then the size of the largest set in $\mathcal{A}_n$ is at most

$$(\alpha n)^{1/2}\epsilon \tag{24}$$

for large n. For the uniform distribution $\alpha = 1$, but (15) shows that, for the exponential distribution, (17) holds for any $\alpha > 2e^{-1}$, so that ϵ can be replaced by the smaller constant

$$(2e^{-1})^{1/2}\epsilon = 2.14\ldots\ .$$

It is an attractive conjecture that, for every continuous distribution F, there is a constant $\beta = \beta(F)$ such that

$$\lim_{r\to\infty}\{P_r(F)r!\}^{1/r} = \beta(F). \tag{25}$$

For any F for which this is true, Theorem 1 shows that, with probability one,

$$\limsup_{n\to\infty} L_n n^{-1/2} \le \beta(F)^{1/2}\epsilon. \tag{26}$$

Some insight into the way $P_r(F)$ (and thus $\beta(F)$ if it exists) depends on F can be gained by noting that, if the Y_i have distribution function F, the random variables $\phi(Y_i)$, if ϕ is a strictly increasing function, have distribution function

$$G(y) = F\{\phi^{-1}(y)\}. \tag{27}$$

If ϕ is convex, then (12) is implied by

$$\phi(Y_{ij}) > \tfrac{1}{2}\{\phi(Y_{ii}) + \phi(Y_{jj})\},$$

so that

$$P_r(F) > P_r(G).$$

The opposite inequality obtains if ϕ is concave.

This shows in particular that, if F has decreasing density on an interval (as does the exponential distribution), then $P_r(F)$ is less than for the uniform distribution, and so (13) remains valid.

It is natural to ask whether there are non-trivial bounds for $P_r(F)$ which hold for all F. The answer is given by the following theorem.

THEOREM 2. *For any continuous distribution function F, and any $r \ge 2$,*

$$\frac{2^r r!}{(2r)!} < P_r(F) < \frac{2^r}{(r+1)!}, \tag{28}$$

and these bounds are best possible.

The right hand inequality shows that (17) holds, for any F, with $\alpha = 2$. Hence (19) holds universally, if the right hand side is set at

$$2^{1/2}\epsilon = 3.52\ldots\ .$$

The inequalities (28) also show that, if $\beta(F)$ exists, it satisfies

$$\tfrac{1}{2} \le \beta(F) \le 2. \tag{29}$$

PROOF: Because F is continuous and non-decreasing,

$$\begin{aligned} P_r(F) &= P\Big\{Y_{ij} > \tfrac{1}{2}(Y_{ii} + Y_{jj}) \quad (i, j = 1, 2, \ldots, r)\Big\} \\ &\le P\Big\{Y_{ij} > \min(Y_{ii}, Y_{jj}) \quad (i, j = 1, 2, \ldots, r)\Big\} \\ &= E\Big\{\prod_{i<j}[1 - F(\min(Y_{ii}, Y_{jj}))]\Big\} \\ &= E\Big\{\prod_{i<j}\max(U_i, U_j)\Big\} \end{aligned}$$

where the random variables $U_i = 1 - F(Y_{ii}) \quad (i = 1, 2, \ldots, r)$ are independent and uniformly distributed on $(0, 1)$. It is easy to check by direct integration that this last expectation is $2^r/(r+1)!$, so that

$$P_r(F) \le 2^r/(r+1)!. \tag{30}$$

There is equality in (30) only if F is such that

$$Y_{ij} > \min(Y_{ii}, Y_{jj})$$

for all $i < j \le r$ implies

$$Y_{ij} > \tfrac{1}{2}(Y_{ii} + Y_{jj})$$

a.s. for all $i < j \le r$. This can only happen if, whenever $Y_{(1)} < Y_{(2)} < Y_{(3)}$ are the order statistics of a sample of size 3 from F, then

$$P\big\{Y_{(2)} > \tfrac{1}{2}(Y_{(1)} + Y_{(3)})\big\} = 1, \tag{31}$$

and this contradicts the continuity of F.

On the other hand, that (30) is best possible may be seen by considering

$$F(y) = 1 - (1-y)^{1/m} \quad (0 \le y \le 1), \tag{32}$$

where m is a large integer. For this choice of F,

$$\begin{aligned} P_r(F) &= E\Big\{\prod_{i<j}\big[1 - F\big(\tfrac{1}{2}(Y_{ii} + Y_{jj})\big)\big]\Big\} \\ &= E\Big\{\prod_{i<j}\big[1 - \tfrac{1}{2}(Y_{ii} + Y_{jj})\big]^{1/m}\Big\} \\ &= E\Big\{\prod_{i<j}\big[\tfrac{1}{2}(U_i^m + U_j^m)\big]^{1/m}\Big\} \\ &\to E\Big\{\prod_{i<j}\max(U_i, U_j)\Big\} = \frac{2^r}{(r+1)!} \end{aligned}$$

as $m \to \infty$.

The argument for the lower bound in (28) is exactly similar, starting from

$$P_r(F) \geq P\{Y_{ij} > \max(Y_{ii}, Y_{jj}) \quad (i, j = 1, 2, \ldots, r)\}.$$

The sharpness is established by taking

$$F(y) = y^{1/m} \quad (0 \leq y \leq 1), \tag{33}$$

and again letting $m \to \infty$. ■

References

Hammersley, J.M. (1972). A few seedlings of research. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* 1, 345–394.

Kingman, J.F.C. (1973). Subadditive ergodic theory. *Annals of Probability* 1, 883–909.

——— (1988). Typical polymorphisms maintained by selection at a single locus. *Journal of Applied Probability* 25A, 113–125.

Veršik, A.M. and Kerov, S.V. (1977). Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tables. *Soviet Mathematics Doklady* 18, 527–531.

Senate House
University of Bristol
Tyndall Avenue
Bristol BS8 1TH.