ON A CLASSICAL PROBLEM OF PROBABILITY THEORY

by

P. ERDŐS and A. RÉNYI

We consider the following classical "urn-problem". Suppose that there are n urns given, and that balls are placed at random in these urns one after the other. Let us suppose that the urns are labelled with the numbers $1, 2, \ldots, n$ and let ξ_j be equal to k if the j-th ball is placed into the k-th urn. We suppose that the random variables $\xi_1, \xi_2, \ldots, \xi_N, \ldots$ are independent, and $\mathbf{P}(\xi_j = k) = \frac{1}{n}$ for $j = 1, 2, \ldots$ and $k = 1, 2, \ldots, n$. By other words each ball may be placed in any of the urns with the same probability and the choices of the urns for the different balls are independent. We continue this process so long till there are at least m balls in every urn ($m = 1, 2, \ldots$). What can be said about the number of balls which are needed to achieve this goal?

We denote the number in question (which is of course a random variable) by $\nu_m(n)$. The "dixie cup"-problem considered in [1] is clearly equivalent with the above problem. In [1] the mean value $\mathbf{M}(\nu_m(n))$ of $\nu_m(n)$ has been evaluated (here and in what follows $\mathbf{M}(\ )$ denotes the mean value of the random variable in the brackets) and it has been shown that

$$\text{(1)} \qquad \mathbf{M}(\nu_m(n)) = n \log n + (m-1)\, n \operatorname{loglog} n + n \cdot C_m + o(n)$$

where C_m is a constant, depending on m. (The value of C_m is not given in [1]).

In the present note we shall go a step further and determine asymptotically the probability distribution of $\nu_m(n)$; we shall prove that for every real x we have

$$\text{(2)} \qquad \lim_{n \to +\infty} \mathbf{P}\left(\frac{\nu_m(n)}{n} < \log n + (m-1) \operatorname{loglog} n + x\right) = \exp\left(-\frac{e^{-x}}{(m-1)!}\right).$$

(Here and in what follows $\mathbf{P}(\,.\,)$ denotes the probability of the event in the brackets.)

(1) can be deduced from (2); moreover we obtain from (2)

$$\text{(3)} \qquad C_m = c - \log(m-1)! \qquad (m = 1, 2, \ldots)$$

where c is Euler's constant, that is

$$\text{(4)} \qquad c = \int_0^1 \frac{1-e^{-t}}{t}\, dt - \int_1^{+\infty} \frac{e^{-t}}{t}\, dt = 0{,}5772 \ldots .$$

To prove (2) we shall consider the following related problem: Let $x_k(n, N)$, denote the number of balls in the k-th urn ($k = 1, 2, \ldots, n$) after distributing N balls among the urns that is, we put $x_k(n, N) = \sum_{\substack{\xi_j = k \\ 1 \leqq j \leqq N}} 1$. Let us put

$$\mu(n, N) = \min_{1 \leqq k \leqq n} x_k(n, N). \tag{5}$$

We have evidently

$$\mathbf{P}\big(\nu_m(n) > N\big) = \mathbf{P}\big(\mu(n, N) < m\big). \tag{6}$$

Thus to prove (2) it is sufficient to show that putting

$$N(n) = n \log n + (m-1)\, n \operatorname{loglog} n + xn + o(n) \tag{7}$$

(where $o(n)$ is an arbitrary function of n which is of smaller order of magnitude than n and is such that $N(n)$ is a positive integer for all n) we have

$$\lim_{n \to +\infty} \mathbf{P}\big(\mu(n, N(n)) < m\big) = 1 - \exp\left(-\frac{e^{-x}}{(m-1)!}\right). \tag{8}$$

Now clearly we have for $j \leqq m-2$

$$\mathbf{P}\big(\mu(n, N(n)) = j\big) \leqq n \binom{N(n)}{j} \frac{1}{n^j} \left(1 - \frac{1}{n}\right)^{N(n)-j} = O\left(\frac{1}{(\log n)^{m-1-j}}\right) \tag{9}$$

and thus

$$\lim_{n \to +\infty} \mathbf{P}\big(\mu(n, N(n)) < m-1\big) = 0. \tag{10}$$

Denoting by $A_m(n)$ the event that there is at least one k for which $x_k(n, N(n)) = m-1$, we have clearly

$$\big|\mathbf{P}\big(\mu(n, N(n)) < m\big) - \mathbf{P}\big(A_m(n)\big)\big| \leqq \mathbf{P}\big(\mu(n, N(n)) < m-1\big).$$

Thus to prove (8) it suffices to show that

$$\lim_{n \to +\infty} \mathbf{P}\big(A_m(n)\big) = 1 - \exp\left(-\frac{e^{-x}}{(m-1)!}\right). \tag{11}$$

But clearly

$$\mathbf{P}\big(A_m(n)\big) = \sum_{k=1}^{n} \binom{n}{k} (-1)^{k-1} W_k(n) \tag{12}$$

where $W_k(n)$ is the probability of the event that k prescribed urns contain exactly $m-1$ balls. Now evidently

$$W_k(n) = \frac{N(n)!}{[(m-1)!]^k \big(N(n) - (m-1)k\big)!} \cdot \frac{1}{n^{(m-1)k}} \left(1 - \frac{k}{n}\right)^{N(n)-(m-1)k} \tag{13}$$

and therefore

$$\lim_{n \to +\infty} \binom{n}{k} W_k(n) = \frac{\left(\frac{e^{-x}}{(m-1)!}\right)^k}{k!} \qquad (k = 1, 2, \ldots) \tag{14}$$

It is easy to see that if we stop after taking an odd resp. even number of terms on the right of (12), we get a number which is greater resp. smaller than the left hand side of (12). It follows therefore from (14) that (11) holds. As mentioned above, with respect to (10) this implies (8) and taking (6) into account (2) follows.

To deduce (3) we note first that putting

$$F_m(x) = \exp\left(-\frac{e^{-x}}{(m-1)!}\right). \tag{15}$$

we have, with respect to (4),

$$\int_{-\infty}^{+\infty} x d\, F_m(x) = c - \log (m-1)!\,. \tag{16}$$

Now it is easy to show that in the present case the limit of the mean value is equal to the mean value of the limiting distribution that is

$$\lim_{n\to+\infty} \mathbf{M}\left(\frac{\nu_m(n)}{n} - \log n - (m-1)\operatorname{loglog} n\right) = \int_{-\infty}^{+\infty} x d\, F_m(x)\,, \tag{17}$$

which proves (3).

Let us mention that in view of (3), (2) can be written also in the form

$$\lim_{n\to+\infty} \mathbf{P}\left(\frac{\nu_m(n) - \mathbf{M}(\nu_m(n))}{n} < x\right) = e^{-e^{-x-c}}\,, \tag{2'}$$

which shows that the limit distribution of $\frac{\nu_m(n) - \mathbf{M}(\nu_m(n))}{n}$ does not depend on m.

It should be mentioned that for the special case $m = 1$ (2) can be deduced in an other, more straightforward way, namely by the method by which the explicit formula

$$\mathbf{M}(\nu_1(n)) = n\left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}\right) \tag{18}$$

is proved in [2]. Let us denote by $\nu_1(n, k)$ $(k = 1, 2, \ldots, n)$ the number of balls which are necessary in order that exactly k urns should contain at least one ball. Clearly $\nu_1(n, 1) = 1$, $\nu_1(n, n) = \nu_1(n)$ and the random variables

$$\delta_1 = 1, \quad \delta_k = \nu_1(n, k) - \nu_1(n, k-1) \quad (k = 2, 3, \ldots, n) \tag{19}$$

are independent. We have further

$$\mathbf{P}(\delta_k = j) = p_k(1 - p_k)^{j-1} \qquad (j = 1, 2, \ldots) \tag{20}$$

where

$$p_k = \frac{n-k+1}{n} \qquad (k = 2, 3, \ldots, n).$$

Thus it follows that the characteristic function $\varphi_n(t)$ of

$$\eta_n = \frac{\nu_1(n) - n \sum\limits_{k=1}^{n} \frac{1}{k}}{n} = \frac{\delta_2 + \delta_3 + \ldots + \delta_n - n \sum\limits_{k=1}^{n=1} \frac{1}{k}}{n}$$

is given by

$$\varphi_n(t) = \mathbf{M}(e^{it\eta_n}) = \frac{1}{\prod\limits_{h=1}^{n} e^{\frac{it}{h}} \left(1 + \frac{n}{h}\left(e^{-\frac{it}{n}} - 1\right)\right)} \tag{21}$$

and thus

$$\lim_{n \to +\infty} \varphi_n(t) = \frac{1}{\prod\limits_{h=1}^{\infty} e^{\frac{it}{h}} \left(1 - \frac{it}{h}\right)}. \tag{22}$$

By the classical product representation of the gamma function it follows that

$$\lim_{n \to +\infty} \varphi_n(t) = \Gamma(1 - it)\, e^{-itc} \tag{23}$$

where c is again Euler's constant. As however by the integral representations of the gamma function we have

$$\int\limits_{-\infty}^{+\infty} e^{ixt}\, d\, F_1(x) = \Gamma(1 - it) \tag{24}$$

where $F_1(x)$ is defined by (15), it follows that

$$\lim_{n \to +\infty} \mathbf{P}\left(\frac{\nu_1(n) - n\left(1 + \frac{1}{2} + \ldots + \frac{1}{n} - c\right)}{n} < x\right) = F_1(x) = e^{-e^{-x}}, \tag{25}$$

and as it is well known that

$$1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} = \log n + c + o(1) \tag{26}$$

we obtain

$$\lim_{n \to +\infty} \mathbf{P}\left(\frac{\nu_1(n) - n \log n}{n} < x\right) = e^{-e^{-x}}. \tag{27}$$

Thus we obtained a second proof of (2) for $m = 1$.

Finally we consider the following problem: Let $\omega_k(n, N)$ denote the number of urns containing exactly k balls, if we place at random N balls into n urns. Let us investigate the asymptotic distribution of $\omega_{m-1}(n, N(n))$ where $N(n)$ is given by (7). (By other words we take so many balls that the probability that there should be less than $m-1$ balls in any of the urns should tend to 0). By using the well known formula of CH. JORDAN [3] we have

$$\mathbf{P}\left(\omega_{m-1}(n, N(n))=k\right)=\sum_{j=0}^{n}(-1)^{j} W_{k+j}(n)\binom{k+j}{k}\binom{n}{k+j}. \tag{28}$$

Thus it follows from (14) that

$$\lim_{n\to+\infty} \mathbf{P}\left(\omega_{m-1}(n, N(n))=k\right)=\frac{\lambda^{k} e^{-\lambda}}{k!} \tag{29}$$

where

$$\lambda=\frac{e^{-x}}{(m-1)!}. \tag{30}$$

Thus the number of urns containing exactly $(m-1)$ balls will be in the limit distributed according to Poisson's law with mean value $\frac{e^{-x}}{(m-1)!}$. In the special case $m=1$ this result states that if we distribute $n \log n + n + o(n)$ balls among n urns then the number $\omega_0(n, N(n))$ of empty urns will for $n \to \infty$ in the limit be distributed according to Poisson's law with mean value e^{-x}. This special case was mentioned already by S. BERNSTEIN [4] (see also [5] Ch. IV. Problem No. 8.).

It is an interesting problem to investigate the limiting distribution of $\nu_m(n)$ when m increases together with n, but we can not go into this question here.

Finally we mention that the problem treated above is analogous to a problem concerning random graphs which we considered recently (see [6]).

(Received February 7, 1961.)

REFERENCES

[1] NEWMAN, D. J.—SHEPP, L.: "The double dixie cup problem." *The American Mathematical Monthly* **67** (1960), 58—61.

[2] FELLER, W.: *Introduction to Probability Theory*, Vol. I. New York, 1950. p. 213.

[3] JORDAN, CH.: «Le théoreme de probabilité de Poincaré generalisé au cas de plusieurs variables indépendantes.» *Acta Sci. Math. (Szeged)* **7** (1934) 103—111.

[4] BERNSTEIN, S. N.: *Teoria Veroiatnostei.* (In Russian) Moscow, 1945. p. 75—76.

[5] RÉNYI, A.: *Valószínűségszámítás* (Textbook of probability theory, in Hungarian) Budapest, 1954. p. 134.

[6] ERDŐS, P.—RÉNYI, A.: "On the strength of connectedness of a random graph." Acta Math. Acad. Sci. Hung. **12** (1961) 261—267.

ОБ ОДНОМ КЛАССИЧЕСКОМ ПРОБЛЕМЕ ТЕОРИИ ВЕРОЯТНОСТЕЙ

P. ERDŐS и A. RÉNYI

Резюме

В n ящиков брошено наудачу N дробинок. Пусть в ящике номера $k(k = 1, 2, \ldots, n)$ попадает $x_k(n, N)$ дробинок. Положим

$$\mu(n, N) = \min_{1 \leq k \leq n} x_k(n, N) .$$

Доказывается что если

$$N(n) = n \log n + (m - 1)\, n \operatorname{loglog} n + xn + o(n) ,$$

где m целое положительное число, то имеет место

$$\lim_{n \to +\infty} \mathbf{P}\bigl(\mu(n, N(n)) < m\bigr) = 1 - \exp\left(- \frac{e^{-x}}{(m-1)!}\right).$$