Elem. Math. 56 (2001) 163 – 169
0013-6018/01/040163-7 $ 1.50+0.20/0

Elemente der Mathematik

Estimating the size of a union of random subsets of fixed cardinality

Michael Barot and José Antonio de la Peña

José Antonio de la Peña got his Ph.D. from UNAM, México in 1983. He made a postdoctoral stay at the University of Zurich, Switzerland from 1984 to 1986. Since then he has a research position at the Instituto de Matemáticas, UNAM. His main research area is the representation theory of algebras but he has also done some work in combinatorics. At this moment, he is Director of the Instituto de Matemáticas, UNAM.

Michael Barot, born in 1966 in Schaffhausen, Switzerland, obtained his degree from University of Zurich in 1994 and his Ph.D. from UNAM, México in 1997. Since 1998 he is an associated researcher of the Instituto de Matemáticas, UNAM. His main fields of interest are representation theory of algebras and quadratic forms.

1 Introduction and result

1.1 The Problem. Our problem can be simply explained as an urn problem. Suppose that we have an urn with N white balls and repeat the following procedure s times: take k balls out of the urn, color them black and put them back. How many black balls do we expect to find in the urn at the end?

Certainly, the problem may be reformulated in the following easy model. Let $\mathcal{N}$ be a fixed set with N elements and denote by $\mathcal{P}_k(\mathcal{N})$ the set of all subsets of $\mathcal{N}$ containing k elements. We ask then for the probability that the union of s elements of $\mathcal{P}_k(\mathcal{N})$ contains

Die Motivation für die vorliegende Arbeit hat ihren Ursprung in der Methode indirekter Umfragen, bei denen die befragten Personen nicht Auskunft über sich selbst, sondern über eine feste Anzahl von „Freunden“ geben. Dies führt zur Frage nach der Anzahl der Personen, über die insgesamt Informationen gesammelt worden sind. Dementsprechend wird in dieser Arbeit von der folgenden Situation ausgegangen. Es wird zufällig eine bestimmte Anzahl von Teilmengen derselben Kardinalität einer gegebenen Menge ausgewählt und die Vereinigung dieser Teilmengen gebildet. Die Kardinalität dieser Vereinigung wird als Zufallsvariable gewählt. Für diese Zufallsvariable werden dann die Wahrscheinlichkeitsverteilung, die Erwartung und die Varianz explizit berechnet. Dazu wird die Technik der erzeugenden Funktionen herangezogen.

exactly i elements if each element of $\mathcal{P}_k(\mathcal{N})$ has the same probability to be chosen. More precisely, let $\mathcal{S}_{s,k}(\mathcal{N})$ be the set of all s-tuples in $\mathcal{P}_k(\mathcal{N})$ and p the uniform probability measure in $\mathcal{S}_{s,k}(\mathcal{N})$. Denote by $\mathbf{X} : \mathcal{S}_{s,k}(\mathcal{N}) \to \mathbb{N}$ the discrete random variable given by $\mathbf{X}(A) = |\bigcup_{i=1}^{s} A_i|$. In this work, we give an explicit formula for the probability $\mathrm{P}(\mathbf{X} = i)$, the expectation $\mathrm{E}(\mathbf{X})$ and the variance $\mathrm{V}(\mathbf{X})$.

Our motivation for this problem comes from the technique of indirect polls, where each interviewed person is asked to give information about “friends" instead about her/himself. This technique was originally suggested by Killworth, Johnson, McCarty, Shelley and Bernard in situations where a direct question might well lead to misleading results because of the stigmatizing character of the question as for example “Are you infected with the AIDS-virus?", see [1] and [2] for details. However, the mathematical model underlying their approach is far more complicated since they do not fix the number of “friends" about which each person is asked.

1.2 Result. Since k, s and N may vary, we denote by $\mathbf{X}_{s,k,N}$ the corresponding random variable.

Theorem *With the above notation, we have*

$$\mathrm{P}(\mathbf{X}_{s,k,N} = i) = \frac{\binom{N}{i}}{\binom{N}{k}^s} \sum_{\ell=0}^{i-k} (-1)^{\ell} \binom{i}{\ell} \binom{i-\ell}{k}^s,$$

$$\mathrm{E}(\mathbf{X}_{s,k,N}) = N(1 - \omega_{s,k,N})$$

and

$$\mathrm{V}(\mathbf{X}_{s,k,N}) = N(N-1)\omega_{s,k,N}\omega_{s,k,N-1} - N^2\omega_{s,k,N}^2 + N\omega_{s,k,N},$$

where $\omega_{s,k,N} = \left(1 - \frac{k}{N}\right)^s$.

The article is organized as follows. In Section 2 we prove some technical lemmas about binomial coefficients and in Section 3 we prove our theorem. We thankfully acknowledge support from CONACyT.

2 Preparing lemmas

Lemma 2.1 *For any natural numbers* $k \leq j \leq i$ *we have*

$$\sum_{t=i-k}^{i} (-1)^{t-j} \binom{t}{j} \binom{k}{i-t} = (-1)^{i-j} \binom{i-k}{j-k}.$$

Proof. If $k = 0$ the result is obvious, and if $k = 1$ then we have $\binom{i-1}{j-1} = \binom{i}{j} - \binom{i-1}{j}$, again the result. Assume now that the formula holds for k. Then we have

$$\begin{aligned}
(-1)^{i-j}\binom{i-k-1}{j-k-1} &= (-1)^{i-j}\binom{i-k}{j-k} - (-1)^{i-j}\binom{i-k-1}{j-k} \\
&= \sum_{t=i-k}^{i}(-1)^{t-j}\binom{t}{j}\binom{k}{i-t} + \sum_{t=i-1-k}^{i-1}(-1)^{t-j}\binom{t}{j}\binom{k}{i-1-t} \\
&= (-1)^{i-j}\binom{i}{j} + \sum_{t=i-k}^{i-1}(-1)^{t-j}\binom{t}{j}\left[\binom{k}{i-t} + \binom{k}{i-1-t}\right] \\
&\quad + (-1)^{i-1-k-j}\binom{i-1-k}{j} \\
&= \sum_{t=i-(k+1)}^{i}(-1)^{t-j}\binom{t}{j}\binom{k+1}{i-t}.
\end{aligned}$$

Hence the result follows by induction. □

Lemma 2.2 *For any natural numbers $k \leq i$ we have*

$$\sum_{j=i-k}^{i}(-1)^{j-k}\binom{j-1}{k-1}\binom{k}{i-j} = \begin{cases} 1 & \textit{if } i = k, \\ 0 & \textit{else.} \end{cases}$$

Proof. If we substitute $\binom{j-1}{k-1}$ by $\binom{j}{k} - \binom{j-1}{k}$ we obtain for the left-hand side $\sum_{j=i-k}^{i}(-1)^{j-k}\binom{j}{k}\binom{k}{i-j} - \sum_{j=i-k}^{i}(-1)^{j-k}\binom{j-1}{k}\binom{k}{i-j}$. By Lemma 1, the first summand equals $(-1)^{i-k}\binom{i-k}{0}$, whereas the second summand is zero if $i = k$ and otherwise equals $-(-1)^{(i-1)-k}\binom{(i-1)-k}{0}$. Hence the result follows. □

Lemma 2.3 *For any natural number $j \leq N$, we have*

a) $$\sum_{i=j}^{N}(-1)^{i-j}i\binom{N-j}{i-j} = \begin{cases} 0 & \textit{for } j \leq N-2, \\ -1 & \textit{for } j = N-1, \\ N & \textit{for } j = N, \end{cases}$$

b) $$\sum_{i=j}^{N}(-1)^{i-j}i^2\binom{N-j}{i-j} = \begin{cases} 0 & \textit{for } j \leq N-3, \\ 2 & \textit{for } j = N-2, \\ 1-2N & \textit{for } j = N-1, \\ N^2 & \textit{for } j = N. \end{cases}$$

Proof. Set $f_{j,N}(x) = \sum_{i=j}^{N}(-1)^{i-j}\binom{N-j}{i-j}x^i$. Observe that $\sum_{i=j}^{N}(-1)^{i-j}i\binom{N-j}{i-j} = \frac{\partial}{\partial x}f_{j,N}(1)$ and that $f_{j,N}(x) = (-1)^{N-j}x^j(x-1)^{N-j}$. Thus, part (a) follows straightforward by differentiating $f_{j,N}(x)$ once and (b) follows also easily by differentiating $f_{j,N}(x)$ twice and combining the outcome with the first result. □

3 Proof

3.1 Probability distribution

Proof. We first express $\mathrm{P}(\mathbf{X}_{s,k,N} = i)$ as fraction of "good" events over the total number of "possible" events. The latter is simply $\binom{N}{k}^s$, so let $\mathrm{N}(\mathbf{X}_{s,k,N} = i) = \binom{N}{k}^s \mathrm{P}(\mathbf{X}_{s,k,N} = i)$, the number of "good" events. Since there are $\binom{N}{i}$ ways to fix a subset of cardinality i in P, we have

$$\mathrm{N}(\mathbf{X}_{s,k,N} = i) = \binom{N}{i} n_{s,k}(i)$$

where $n_{s,k}(i)$ is the number of ways, how s subsets of cardinality k, out of a set of cardinality i, can be chosen such, that their union is the whole set. For the forthcoming it will be convenient to define

$$n_{0,k}(i) := (-1)^{i-k}\binom{i-1}{k-1},$$

since then the following reduction formula holds for all $s \geq 1$:

$$n_{s,k}(i) = \sum_{j=i-k}^{i} \binom{i}{j} n_{s-1,k}(j) \binom{j}{k-i+j}. \tag{1}$$

In fact, if $s > 1$, the first $s-1$ subsets form a union U of cardinality $j \in \{i-k, \ldots, i\}$ (there are $n_{s-1,k}(j)$ ways to do so) and $\binom{i}{j}$ ways to fix a subset of cardinality j inside a set of cardinality i. The last subset must then contain all $i - j$ remaining elements which do not belong to U, and the other $k - i + j$ elements may be chosen freely in U. In the remaining case, where $s = 1$, we observe that $\binom{i}{j}\binom{j}{i-k} = \binom{i}{k}\binom{k}{i-j}$. Therefore, the left-hand side equals $\binom{i}{k}\sum_{j=i-k}^{i}(-1)^{j-k}\binom{j-1}{k-1}\binom{k}{i-j}$, so by Lemma 2.2, it equals 1 if $i = k$ and 0 otherwise, just like $n_{1,k}(i)$.

We now consider the generating function

$$h_{k,i}(x) = \sum_{s=0}^{\infty} \frac{1}{s!} n_{s,k}(i) x^s.$$

We calculate the formal derivative with respect to x using (1):

$$\begin{aligned}
\frac{\partial}{\partial x} h_{k,i}(x) &= \sum_{s=1}^{\infty} \frac{s}{s!} n_{s,k}(i) x^{s-1} \\
&= \sum_{s=0}^{\infty} \frac{1}{s!} n_{s+1,k}(i) x^s \\
&= \sum_{s=0}^{\infty} \frac{1}{s!} \sum_{j=i-k}^{i} \binom{i}{j} n_{s,k}(j) \binom{j}{k-i+j} x^s \\
&= \sum_{s=0}^{\infty} \binom{i}{k} \sum_{j=i-k}^{i} \binom{k}{i-j} \frac{1}{s!} n_{s,k}(j) x^s \\
&= \binom{i}{k} \sum_{j=i-k}^{i} \binom{k}{i-j} h_{k,j}(x).
\end{aligned}$$

In other words, the family $h_{k,i}$ satisfies the following system of equations

$$\frac{\partial}{\partial x} f_{k,i}(x) = \binom{i}{k} \sum_{j=i-k}^{i} \binom{k}{i-j} f_{k,j}(x). \tag{2}$$

We verify that the functions

$$g_{k,i}(x) = \sum_{j=k}^{i} (-1)^{i-j} \binom{i}{j} \mathrm{e}^{\binom{j}{k} x}$$

also satisfy (2). Indeed,

$$\begin{aligned}
\frac{\partial}{\partial x} g_{k,i}(x) &= \sum_{j=k}^{i} (-1)^{i-j} \binom{i}{j} \binom{j}{k} \mathrm{e}^{\binom{j}{k} x} \\
&= \binom{i}{k} \sum_{j=k}^{i} (-1)^{i-j} \binom{i-k}{j-k} \mathrm{e}^{\binom{j}{k} x} \\
&= \binom{i}{k} \sum_{j=k}^{i} \sum_{t=i-k}^{i} (-1)^{t-j} \binom{t}{j} \binom{k}{i-t} \mathrm{e}^{\binom{j}{k} x} \qquad \text{(by Lemma 2.1)} \\
&= \binom{i}{k} \sum_{t=i-k}^{i} \sum_{j=k}^{t} (-1)^{t-j} \binom{t}{j} \binom{k}{i-t} \mathrm{e}^{\binom{j}{k} x} \\
&= \binom{i}{k} \sum_{t=i-k}^{i} \binom{k}{i-t} g_{k,t}(x).
\end{aligned}$$

It is easy to check that $g_{0,0}(x) = h_{0,0}(x) = \mathrm{e}^x$ and $g_{k,0}(x) = h_{k,0}(x) = 0$ for $k > 0$ and that for all k and i, $g_{k,i}(0) = h_{k,i}(0) = n_{0,k}(i)$. Therefore, we get $g_{k,i} = h_{k,i}$ for all k and i.

Since

$$g_{k,i}(x) = \sum_{s=0}^{\infty} \frac{1}{s!} \sum_{j=k}^{i} (-1)^{i-j} \binom{i}{j} \binom{j}{k}^{s} x^{s},$$

we obtain

$$n_{s,k} = \sum_{j=k}^{i} (-1)^{i-j} \binom{i}{j} \binom{j}{k}^{s},$$

hence the result. □

3.2 Expectation

Proof. By definition, we have

$$\mathrm{E}(\mathbf{X}_{s,k,N}) = \sum_{i=k}^{N} i\mathrm{P}(\mathbf{X}_{s,k,N} = i).$$

Define

$$E(x) = \sum_{s=0}^{\infty} \frac{1}{s!}\mathrm{E}(\mathbf{X}_{s,k,N})x^s.$$

Then, if we set $x' = \frac{x}{\binom{N}{k}}$, we have

$$\begin{aligned}
E(x) &= \sum_{s=1}^{\infty} \frac{1}{s!} \sum_{i=k}^{N} i\mathrm{P}(\mathbf{X}_{s,k,N} = i)x^s \\
&= \sum_{i=k}^{N} i \sum_{s=1}^{\infty} \frac{1}{s!} \frac{\binom{N}{i}}{\binom{N}{k}^s} n_{s,k}(i)x^s \\
&= \sum_{i=k}^{N} i \binom{N}{i} h_{k,i}(x') \\
&= \sum_{i=1}^{N} \sum_{j=k}^{i} i \binom{N}{i} (-1)^{i-j} \binom{i}{j} \mathrm{e}^{\binom{j}{k}x'} \qquad \text{(since } h_{k,i} = g_{k,i}) \\
&= \sum_{j=k}^{N} \left[\sum_{i=1}^{N} (-1)^{i-j} i \binom{N}{i} \binom{i}{j} \right] \mathrm{e}^{\binom{j}{k}x'} \\
&= \sum_{j=k}^{N} \binom{N}{j} \left[\sum_{i=1}^{N} (-1)^{i-j} i \binom{N-j}{i-j} \right] \mathrm{e}^{\binom{j}{k}x'} \\
&= -N\mathrm{e}^{\binom{N-1}{k}x'} + N\mathrm{e}^{\binom{N}{k}x'} \qquad \text{(by Lemma 2.3(a))} \\
&= N\left[-\mathrm{e}^{(1-\frac{k}{N})x} + \mathrm{e}^x \right] \\
&= N\left[\sum_{s=1}^{\infty} \frac{1}{s!} \left(1 - (1 - \frac{k}{N})^s \right) x^s \right].
\end{aligned}$$

Therefore, we have $\mathrm{E}(\mathbf{X}_{s,k,N}) = N(1 - (1 - \frac{k}{N})^s)$, which completes the proof. □

3.3 Variance

Proof. By definition, we have

$$\begin{aligned}
\mathrm{V}(\mathbf{X}_{s,k,N} = i) &= \sum_{i=1}^{\infty} (i - \mathrm{E}(\mathbf{X}_{s,k,N}))^2\, \mathrm{P}(\mathbf{X}_{s,k,N} = i) \\
&= \sum_{i=1}^{\infty} i^2\mathrm{P}(\mathbf{X}_{s,k,N} = i) - \mathrm{E}(\mathbf{X}_{s,k,N})^2,
\end{aligned}$$

so we define

$$V(x) = \sum_{s=1}^{\infty} \frac{1}{s!} \sum_{i=1}^{N} i^2 \mathrm{P}(\mathbf{X}_{s,k,N} = i) x^s.$$

In the following, the first equation follows by the same arguments as in 3.2, whereas the second is due to Lemma 2.3(b). Again, we set $x' = \frac{x}{\binom{N}{k}}$.

$$\begin{aligned} V(x) &= \sum_{j=k}^{N} \binom{N}{j} \left[\sum_{i=j}^{N} (-1)^{i-j} i^2 \binom{N-j}{i-j} \right] \mathrm{e}^{\binom{j}{k} x'} \\ &= 2\binom{N}{N-2} \mathrm{e}^{\binom{N-2}{k} x'} + (1-2N) N \mathrm{e}^{\binom{N-1}{N} x'} + N^2 \mathrm{e}^{\binom{N}{k} x'} \\ &= N(N-1) \mathrm{e}^{(1-\frac{k}{N})(1-\frac{k}{N-1})x} + (1-2N) N \mathrm{e}^{(1-\frac{k}{N})x} + N^2 \mathrm{e}^x \\ &= N \left[\sum_{s=0}^{\infty} \frac{1}{s!} \left((N-1)(1-\frac{k}{N})^s (1-\frac{k}{N-1})^s + (1-2N)(1-\frac{k}{N})^s + N \right) x^s \right]. \end{aligned}$$

Thus, by comparing coefficients, we obtain the explicit formula for the variance of $\mathbf{X}_{s,k,N}$ as given in our theorem. □

References

[1] P. Killworth, E. Johnson, C. McCarty, G. A. Shelley, R. Bernard: *A social Network Approach to Estimating Seroprevalence in the United States*. Preprint.

[2] P. Killworth, E. Johnson, C. McCarty, G. A. Shelley, R. Bernard: *Estimation of seroprevalence, rape and homelessness in the U.S. using a social network approach*. Preprint.

Michael Barot
Instituto de Matemáticas
Universidad Nacional Autonoma de México
México, D.F., 04510, MEXICO
e-mail: barot@matem.unam.mx

José Antonio de la Peña
Instituto de Matemáticas
Universidad Nacional Autonoma de México
México, D.F., 04510, MEXICO
e-mail: jap@penelope.matem.unam.mx