TEACHING THE ESSENTIALS OF ELEMENTARY PROBABILITY THEORY OR HOW NOT TO CALCULATE THE MEAN OF THE NEGATIVE HYPERGEOMETRIC DISTRIBUTION

Peter Nüesch
Federal Institute of Technology, Mathematics Department
MA-Ecublens, CH-1015 Lausanne, Switzerland

1. Introduction

Brosch (1991) considers a binary urn model with r red balls out of a total of N. He lets his students simulate the length X of the game if one draws, without replacement, till k red balls have appeared. The students came up with the correct answer

$$E(X) = k \cdot \frac{N+1}{r+1}.$$

He obviously does not attempt to obtain this expectation in class but invites readers to do so. The two submitted solutions (Bach, 1991; Janous, 1991) use the probability law of X and calculate the mathematical expectation by direct summation of $\sum x\, P(X = x)$.

In the following we present a solution which reduces the problem to a simpler one, relates it to another problem that uses combinatorics sparingly and which has been assigned to the 15 to 18 year age group at an International Mathematics Olympiad. Not only does this make it accessible to students of a first elementary course in probability theory but at the same it gets across to the student a feeling of the randomness of the problem. This, of course, is a basic principle of any course in probability. A mere simulation as in (Brosch, 1991), or sophisticated summations as in (Bach, 1991) and (Janous, 1991), however beautiful mathematically, simply do not convey any sense of randomness. Along the way, many of the basic notions of elementary probability are used. Our solution then can be presented as a review example of these basic notions or as an accompanying example while introducing them.

In the following we will adapt the notation used in Bach (1991), Brosch (1991) and Janous (1991) to the accepted international standards as in Feller

(1968). In particular, X will denote the random variable and not the number of red balls in the urn.

2. Reducing the problem to a simpler one

It is clear that the waiting time X till the occurrence of the *kth* success is the sum

$$X = X_1 + X_2 + \dots + X_k, \tag{1}$$

where X_i is the waiting time after the $(i-1)st$ till the *ith* success. The proportionality of $E(X)$ with respect to k lets one hypothesize that the X_i are identically distributed. While a student has no difficulty in accepting this fact when sampling is done with replacement, he does have difficulty in exhaustive sampling. In an elementary course we would not try to prove the identical distribution property. The following easy example (Ross, 1988, p. 241, problem 2 and p. 323, problem 3) shows beautifully that it is so.

A bin of 5 transistors is known to contain two that are defective. The transistors are to be tested, one at a time, until the defective ones are identified. Denote by X_1 the number of tests made until the first defective is spotted and by X_2 the number of additional tests until the second defective is spotted; i) find the joint probability mass function of X_1 and X_2; ii) find the expected number of tests that are made.

While i) is helpful to establish dependence between the X_i, it is really ii) that interests us.

The sample space Ω consists of the $\binom{5}{2} = 10$ arrangements of the 5 transistors (D for defective, N for non-defective).

					x_1	x_2	y_1	y_2	y_3
D	*D*	*N*	*N*	*N*	1	1	3	1	1
D	*N*	*D*	*N*	*N*	1	2	2	2	1
D	*N*	*N*	*D*	*N*	1	3	2	1	2
D	*N*	*N*	*N*	*D*	1	4	2	1	1
N	*D*	*D*	*N*	*N*	2	1	1	3	1
N	*D*	*N*	*D*	*N*	2	2	1	2	2
N	*D*	*N*	*N*	*D*	2	3	1	2	1
N	*N*	*D*	*D*	*N*	3	1	1	1	3
N	*N*	*D*	*N*	*D*	3	2	1	1	2
N	*N*	*N*	*D*	*D*	4	1	1	1	1

Table 1

The probability laws for X_1 and X_2 are identical. (We observe the same fact for the random variables Y_j, the respective waiting times till the non-defective transistors show up). Here is an example of two (three) different random variables defined on the same sample space having the same distribution.

3. The probability law of X

First we restrict our attention to X_1. In order to simplify notation let us omit for the moment the subscript 1.

The event $\{X > x\}$ means that the waiting time up to success 1 is larger than x, or, equivalently, that the first x draws are failures. Thus the tail probability $P(X > x)$ is

$$P(X > x) = \frac{N-r}{N} \cdot \frac{N-r-1}{N-1} \cdot \frac{N-r-2}{N-2} \cdots \frac{N-r-x+1}{N-x+1}$$

$$= \frac{(N-r)_x}{(N)_x} . \tag{2}$$

Observe the analogy to the geometric case, i.e. when sampling is done with replacement

$$P(X > x) = \frac{(N-r)^x}{N^x} .$$

It follows that

$$\begin{aligned} P(X=x) &= P(X \le x) - P(X \le x-1) \\ &= 1 - P(X > x) - 1 + P(X > x-1) \\ &= P(X > x-1) - P(X > x) \\ &= \frac{(N-r)_{x-1}}{(N)_{x-1}} - \frac{(N-r)_x}{(N)_x} \\ &= \frac{(N-r)_{x-1}}{(N)_x} r \quad , \qquad x = 1, \ldots, N-r+1. \end{aligned} \tag{3}$$

Again, observe the analogy to the geometric distribution

$$P(X=x)=\frac{(N-r)^{x-1}}{N^x}r=q^{x-1}p \quad x=1,\ 2,\ldots$$
$$\text{where } p=\frac{r}{N} \text{ and } q=1-p.$$

(Most elementary texts treat the problem of the N identical keys of which only one opens a door. To the student's surprise the probability that the door opens on any one attempt is $1/N$ if keys already used are discarded. This is formula (3) for $r=1$).

The case $k>1$ leads to the negative binomial distribution when sampling is done with replacement

$$P(X=x)=\binom{x-1}{k-1}\frac{r^k(N-r)^{x-k}}{N^x}, \quad x=k,\ k+1,\ldots$$

and for sampling without replacement one obtains the analogous law

$$P(X=x)=\binom{x-1}{k-1}\frac{(r)_k(N-r)_{x-k}}{(N)_x}, \quad x=k,\ k+1,\ \ldots,\ N-r+k. \qquad (4)$$

It is the expectation of this distribution which we are asked to find. In the literature, e.g. Johnson and Kotz (1977), it has received the name "negative hypergeometric" distribution, in analogy to the fixed sample distributions with a variable number of successes, where "binomial" is used if sampling is done with replacement, and "hypergeometric" for sampling without replacement. (Strickly speaking, the name negative binomial was given to the number of failures before success k occurs. This stems from the fact that the binomial coefficient can be written as $\binom{-r}{k}$. Patil and Joshi (1968) make a further distinction between a negative and an inverse hypergeometric distribution).

We will not need the law for $k>1$. If we can find the mean of X_1 and thus of X_i, *the mean of* $X=\sum^k X_i$ *can be obtained without explicitly knowing the probability law of* X.

4. Relation to another problem

One finds in Ross (1988, p. 145, exercise 20) the following problem:

Balls numbered 1 through N are in an urn. Suppose that r, $r \leq N$, of them are randomly selected without replacement. Let Y denote the largest number selected. Find the probability mass function of Y.

To obtain the probability mass function of Y note that if y is the maximum number drawn all $r-1$ other balls in the sample must carry numbers less than r. Thus

$$P(Y=y)=\frac{\binom{y-1}{r-1}}{\binom{N}{r}}, \quad y=r, \ldots, N.$$

Since the probabilities add up to 1 we get the combinatorial identity

$$\sum_{y=r}^{N}\binom{y-1}{r-1}=\binom{N}{r} \tag{5}$$

which is the famous sum of entries on a 45° diagonal of the Pascal triangle.

We can equally easily compute the probability mass function of Z, the smallest number selected. By the same argument used for the law of Y, $r-1$ numbers are to be selected from the $N-z$ numbers larger than z. Thus

$$P(Z=z)=\frac{\binom{N-z}{r-1}}{\binom{N}{r}}, \quad z=1, \ldots, N-r+1.$$

The probability space Ω for $N=5$ and $r=2$ is

		y	z	$y-z$
1	2	2	1	1
1	3	3	1	2
1	4	4	1	3
1	5	5	1	4
2	3	3	2	1
2	4	4	2	2
2	5	5	2	3
3	4	4	3	1
3	5	5	3	2
4	5	5	4	1

Table 2

By comparing Table 2 to Table 1 one observes that

$$P(X_1 = x) = P(Z = z)$$

and

$$P(X_2 = x) = P(Y - Z = t).$$

The variable X_1, the position of the first detective item of r defective ones in a total of N, has the same probability law as the minimum Z of all samples of size r drawn without replacement from a total of N numbered items. Of course, this result can be obtained in a purely formal way. Instead of transforming (3)

$$P(X = x) = \frac{(N-r)_{x-1}}{(N)_x} \cdot r$$

into

$$P(X = x) = \frac{\binom{N-x}{r-1}}{\binom{N}{r}}, \quad x = 1,\ 2,\ \ldots,\ N-r+1 \tag{6}$$

it is easier to work with tail probabilities. We have (2)

$$P(X > x) = \frac{(N-r)_x}{(N)_x}$$

$$= \frac{\binom{N-r}{x}}{\binom{N}{x}} \qquad \text{using } (a)_b = \binom{a}{b} b! \tag{7}$$

$$= \frac{\binom{N-x}{N-r-x}}{\binom{N}{N-r}} \qquad \text{using } \binom{n}{j}\binom{j}{i} = \binom{n}{i}\binom{n-i}{j-i}$$

$$\text{for } n = N, i = x, j = N-r$$

$$= \frac{\binom{N-x}{r}}{\binom{N}{r}}, \quad x = 1, 2, \ldots,\ N-r \tag{8}$$

and

$$P(X=x)=P(X>x-1)-P(X>x)$$

$$=\frac{\binom{N-x+1}{r}-\binom{N-x}{r}}{\binom{N}{r}}$$

$$=\frac{\binom{N-x}{r-1}}{\binom{N}{r}}$$ using the basic relation of the Pascal triangle.

Remark: If one proceeds this way for the negative hypergeometric probability law (4), one obtains

$$P(X=x)=\frac{\binom{x-1}{k-1}\binom{N-x}{r-k}}{\binom{N}{r}}. \qquad (9)$$

This reduces for $k=1$ to the form (6) obtained above.

5. The mathematical expectation of X

We take all samples of size r and order the numbers

$$x_1 < x_2 < \ldots < x_h < \ldots < x_r.$$

One has

$$P(X_h=x)=\frac{\binom{x-1}{h-1}\binom{N-x}{r-h}}{\binom{N}{r}}, \quad x=1,\ldots,r \qquad (10)$$

since h − 1 numbers are smaller than x and $r-h$ numbers are larger than x. Since this is a probability law, and the probabilities are therefore summing up to 1, we found the combinatorial relation

$$\sum_{x}\binom{x-1}{b-1}\binom{N-x}{r-b}=\binom{N}{r}. \tag{11}$$

Observe that this relation is the sum of the negative hypergeometric probabilities (9) for $b = k$. In the literature, (11) is called the Vandermonde convolution, e.g. Riordan (1968, p. 8).

By using (11) and

$$\binom{a}{b}=\frac{a}{b}\binom{a-1}{b-1} \tag{12}$$

we get

$$E(X_1)=\frac{\sum x\binom{N-x}{r-1}}{\binom{N}{r}}$$
$$=\frac{N+1}{r+1}\ . \tag{13}$$

(13) holds for all i, i.e. $E(X_i) = (N+1)/(r+1)$. For the random variable X of (1) we get therefore

$$E(X)=\sum_{i=1}^{k}E(X_i)=k\cdot\frac{N+1}{r+1}\ .$$

Bibliography

Bach G. (1991), Lösung 2 zu "Eine Formel für den Erwartungswert", *Praxis der Mathematik*, 33, pp. 279-281.

Brosch H. (1991), Eine Formel für den Erwartungswert, *Praxis der Mathematik*, 33, pp. 132-133.

Feller W. (1968), *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd edition, Wiley, New York.

Janous W. (1991), Lösung 1 zu "Eine Formel für den Erwartungswert", *Praxis der Mathematik*, 33, pp. 278-279.

Johnson N.L. and Kotz S. (1977), *Urn Models and their Applications*, Wiley, New York.
Patil G.G. and Joshi S.W. (1968), *A Dictionary and Bibliography for Discrete Distributions*, Oliver & Boyd, Edimburgh.
Riordan J. (1968), *Combinatorial Identities*, Wiley, New York.
Ross S. (1988), A *First Course in Probability*, 3rd edition, Macmillan, New York.