The Use of Consulting Projects in Undergraduate Statistics Education

Laurence D. Robinson and Ryan R. Rahrig
Department of Mathematics and Statistics, Ohio Northern University, Ada, OH 45810

Abstract

In the summer of 2013, Operations Analyst Paul Erford of the Marathon Petroleum Company, contacted Dr. Ryan Rahrig, Assistant Professor of Statistics at Ohio Northern University, asking him to look at a probability problem that had arisen in his work. Dr. Rahrig forwarded this request to Dr. Laurence Robinson, Associate Professor of Statistics at Ohio Northern University. Together Dr. Rahrig and Dr. Robinson determined the nature of solutions to this type of probability problem. In this paper we (Dr. Rahrig and Dr. Robinson) discuss the solution to the probability problem, and discuss its use as a valuable teaching tool for our mathematical statistics majors at Ohio Northern University.

Key Words: Probability problem, Consulting project, Teaching tool

1. Emails Sent by Mr. Erford

Hey Ryan,

Let's say you have a population set of 200 people. Out of those 200 people, the names of 125 people are drawn as winners in each drawing. There are 9 consecutive drawings in all. I'm trying to determine what is the probability that exactly 1 person wins in all nine drawings, the probability that exactly 2 people win in all nine drawings, the probability that exactly 3 people win in all nine drawings, etc., all the way to the probability that 125 people win in all nine drawings. Do you know a formula that I can use to determine these probabilities?

Thanks,
Paul Erford

Dr. Robinson asked Dr. Rahrig why an engineer at Marathon Oil would be interested in this particular problem, and he in turn sent an email to Mr. Erford asking "what's the real context of the problem?" Mr. Erford responded:

Hey Ryan,

The context of the problem is that there are certain pipeline systems that administer a 'lottery' system similar to what is described below. If a new shipper wins 9 months in a row, they will become a shipper with a higher status that states they are guaranteed service on the pipeline. This guaranteed service will have to take away service from somewhere, and that somewhere is typically a shipper that has lots more available service on the pipeline. By knowing the probability that x people could graduate with 200

new shippers, we can use those probabilities to manage the risk of people taking away some of our service.

Thanks,
Paul Erford

2. An Interesting Problem for Undergraduates

We first discuss some factors that make this a particularly attractive and suitable problem for use with undergraduate students. It is a concisely stated problem that students can easily comprehend. It is a real-life problem posed by a major corporation (ONU has the added benefit of it being a local corporation). This problem is accessible to students at a variety of levels, but still presents a challenge. As we will show, the solution can be obtained by using only concepts covered in an introductory probability course, but it can be more elegantly described using more advanced theory such as Markov chains.

Most of the time, the expectation of an undergraduate research project is not that the student will have a significant breakthrough. Instead, it is that the student will be challenged and rewarded with the opportunity to synthesize information from various areas of study while exploring his or her own ideas. This problem provides such an opportunity – it is difficult enough to present several obstacles to initial classical approaches, yet is still able to be solved by examination from alternative perspectives.

3. A Common Initial Obstacle

It is common for students to first approach this problem simply by counting equally likely outcomes. That is, the goal is to first count the total number of (equally likely) possible outcomes and then determine the number of outcomes that make up the event of interest. Then a simple ratio computes the desired probability. This is certainly a reasonable first attempt as it is a common technique that students use to solve many problems in their introductory probability course. However, for this example, a student will quickly realize that there are $\binom{200}{125}^9 \approx \left(1.6885 * 10^{56}\right)^9$ total possible outcomes, and determining how many of these consist of exactly 1, 2, 3, … winners on all 9 drawings is a non-trivial task.

It is at this point that the advisor may suggest looking at it from an alternative viewpoint. Consider the random variable X_k which represents the number of people who have been selected on all of the first k drawings. Thus, X_3 denotes the number of people who are selected on all of the first $r = 3$ drawings and X_9 denotes the number of people who are selected on all of the first $r = 9$ drawings.

Another piece of advice to the student at this point may be to work with a "smaller problem" rather than the original problem. The solution to the smaller problem may suggest a solution to the larger, original problem.

4. Obtaining the Solution to a "Smaller Problem"

Rather than work with the original problem posed, for which the parameters used were N = population size = 200, n = sample size = 125, and r = number of samplings = 9, we choose to work with a "more manageable" set of parameters, specifically $N = 5$, $n = 3$, and $r = 3$.

The solution to this smaller problem is:

i	0	1	2	3
$\Pr(X_3 = i)$	0.18	0.57	0.24	0.01

We now explain one way to obtain this solution. If a student is unable to reach this solution on his/her own, an additional hint would be to consider the distribution of X_1 and then the distribution of X_{k+1} given X_k.

The probability distribution of the random variable X_1 is easily found. X_1 cannot equal 0, 1, or 2, but rather must equal 3. This is easily seen to be true, since when only a single sampling of $n = 3$ people is made, all 3 selected people must have been selected in all samplings – of which there has only been 1.

i	0	1	2	3
$\Pr(X_1 = i)$	0	0	0	1

Next consider the probability distribution of the random variable X_2. The probability that $X_2 = 3$ equals the probability that the same 3 people chosen in the first drawing will be chosen again. There are $\binom{5}{3} = 10$ possible samples, thus $\Pr(X_2 = 3) = 1/10 = 0.1$. The probability that $X_2 = 2$ equals the probability that exactly 2 out of the 3 people chosen in the first drawing will be chosen again and 1 of the 2 not picked in the first drawing will be chosen. There are $3 \times 2 = 6$ ways this can happen so $\Pr(X_2 = 2) = 6/10 = 0.6$. The probability that $X_2 = 1$ equals the probability that exactly 1 out of the 3 people chosen in the first drawing will be chosen again and the 2 people not drawn in the first drawing are chosen. So $\Pr(X_2 = 1) = 3/10 = 0.3$. $\Pr(X_2 = 0) = 0$, since at least 1 of the 3 chosen in the first drawing will be chosen in the second. Thus, we have:

i	0	1	2	3
$\Pr(X_2 = i)$	0	0.3	0.6	0.1

Now we can investigate the probability distribution of the random variable X_3. First consider $\Pr(X_3 = 0)$. We find this probability for each of the 4 cases: $X_2 = 0$, $X_2 = 1$, $X_2 = 2$, and $X_2 = 3$. $\Pr(X_3 = 0 \mid X_2 = 0) = 1$, because if there were no repeat winners for the first 2 drawings, there will not be any repeat winners of the first 3 drawings. $\Pr(X_3 = 0 \mid X_2 = 1)$ is the probability that the 1 repeat winner from the first 2 drawings will not be drawn in the third. There are $\binom{4}{3} = 4$ ways this can happen, so $\Pr(X_3 = 0 \mid X_2 = 1) = 4/10$. $\Pr(X_3 = 0 \mid X_2 = 2)$ is the probability that the 2 repeat winners from the first 2 drawings will not be drawn in the third. There is only 1 way this can happen (choose the other 3 people), so $\Pr(X_3 = 0 \mid X_2 = 2) = 1/10$. $\Pr(X_3 = 0 \mid X_2 = 3)$ is the probability that the 3 repeat winners from the first 2 drawings will not be drawn in the third. This is not possible since at least 1 of the 3 must be chosen, so $\Pr(X_3 = 0 \mid X_2 = 3) = 0$.

The probability that $X_3 = 0$ can now be found using the "law of total probability":

$$\begin{aligned}\Pr(X_3 = 0) = \quad & \Pr(X_3 = 0 \mid X_2 = 0)\ \Pr(X_3 = 0) \\ & + \ \Pr(X_3 = 0 \mid X_2 = 1)\ \Pr(X_3 = 1) \\ & + \ \Pr(X_3 = 0 \mid X_2 = 2)\ \Pr(X_3 = 2) \\ & + \ \Pr(X_3 = 0 \mid X_2 = 3)\ \Pr(X_3 = 3)\end{aligned}$$

$$= 1\times 0 \ + \ .4\times .3 \ + \ .1\times .6 \ + \ 0\times .1 \ = \ .18$$

The remaining probabilities concerning X_3 can be found similarly.

5. Organizing the "Smaller Solution"

The next step for the student to develop is the organization of the smaller solution so that the solution to the larger problem may be determined. The above solution can be obtained from the following matrix equations:

$$\begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 1.0 \end{bmatrix} = \begin{bmatrix} \Pr(X_1 = 0) \\ \Pr(X_1 = 1) \\ \Pr(X_1 = 2) \\ \Pr(X_1 = 3) \end{bmatrix}$$

$$\begin{bmatrix} 1.0 & 0.4 & 0.1 & 0.0 \\ 0.0 & 0.6 & 0.6 & 0.3 \\ 0.0 & 0.0 & 0.3 & 0.6 \\ 0.0 & 0.0 & 0.0 & 0.1 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 1.0 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.3 \\ 0.6 \\ 0.1 \end{bmatrix} = \begin{bmatrix} \Pr(X_2 = 0) \\ \Pr(X_2 = 1) \\ \Pr(X_2 = 2) \\ \Pr(X_2 = 3) \end{bmatrix}$$

$$\begin{bmatrix} 1.0 & 0.4 & 0.1 & 0.0 \\ 0.0 & 0.6 & 0.6 & 0.3 \\ 0.0 & 0.0 & 0.3 & 0.6 \\ 0.0 & 0.0 & 0.0 & 0.1 \end{bmatrix} \begin{bmatrix} 0.0 \\ 0.3 \\ 0.6 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 0.18 \\ 0.57 \\ 0.24 \\ 0.01 \end{bmatrix} = \begin{bmatrix} \Pr(X_3 = 0) \\ \Pr(X_3 = 1) \\ \Pr(X_3 = 2) \\ \Pr(X_3 = 3) \end{bmatrix}.$$

The first vector given above represents the probability distribution of the random variable X_1. The second vector given above represents the probability distribution vector of the random variable X_2, and is obtained by pre-multiplying the X_1 probability distribution vector by the given matrix. Similarly, the third vector represents the probability distribution of the random variable X_3 (the desired solution), and is obtained by pre-multiplying the X_2 probability distribution vector by the same matrix.

In general, the probability distribution vector of X_{k+1} can be obtained by pre-multiplying the probability distribution vector of X_k by the same matrix **M**, with

$$\mathbf{M} = \begin{bmatrix} 1.0 & 0.4 & 0.1 & 0.0 \\ 0.0 & 0.6 & 0.6 & 0.3 \\ 0.0 & 0.0 & 0.3 & 0.6 \\ 0.0 & 0.0 & 0.0 & 0.1 \end{bmatrix}.$$

That is, we have the following matrix equation which applies for all integer values of *k*:

$$\begin{bmatrix} 1.0 & 0.4 & 0.1 & 0.0 \\ 0.0 & 0.6 & 0.6 & 0.3 \\ 0.0 & 0.0 & 0.3 & 0.6 \\ 0.0 & 0.0 & 0.0 & 0.1 \end{bmatrix} \begin{bmatrix} \Pr(X_k = 0) \\ \Pr(X_k = 1) \\ \Pr(X_k = 2) \\ \Pr(X_k = 3) \end{bmatrix} = \begin{bmatrix} \Pr(X_{k+1} = 0) \\ \Pr(X_{k+1} = 1) \\ \Pr(X_{k+1} = 2) \\ \Pr(X_{k+1} = 3) \end{bmatrix}$$

The matrix **M** can be expressed as follows:

$$\begin{bmatrix} \Pr(X_{k+1} = 0 \mid X_k = 0) & \Pr(X_{k+1} = 0 \mid X_k = 1) & \Pr(X_{k+1} = 0 \mid X_k = 2) & \Pr(X_{k+1} = 0 \mid X_k = 3) \\ \Pr(X_{k+1} = 1 \mid X_k = 0) & \Pr(X_{k+1} = 1 \mid X_k = 1) & \Pr(X_{k+1} = 1 \mid X_k = 2) & \Pr(X_{k+1} = 1 \mid X_k = 3) \\ \Pr(X_{k+1} = 2 \mid X_k = 0) & \Pr(X_{k+1} = 2 \mid X_k = 1) & \Pr(X_{k+1} = 2 \mid X_k = 2) & \Pr(X_{k+1} = 2 \mid X_k = 3) \\ \Pr(X_{k+1} = 3 \mid X_k = 0) & \Pr(X_{k+1} = 3 \mid X_k = 1) & \Pr(X_{k+1} = 3 \mid X_k = 2) & \Pr(X_{k+1} = 3 \mid X_k = 3) \end{bmatrix}$$

With regard to the first column of **M**, we have:

$$\begin{bmatrix} \Pr(X_{k+1} = 0 \mid X_k = 0) \\ \Pr(X_{k+1} = 1 \mid X_k = 0) \\ \Pr(X_{k+1} = 2 \mid X_k = 0) \\ \Pr(X_{k+1} = 3 \mid X_k = 0) \end{bmatrix} = \begin{bmatrix} 1.0 \\ 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$$

Here we see that if after *k* samplings there are 0 people who have been selected every time, then after sampling *k*+1 there must also be 0 people who have been selected every time.

With regard to the second column of **M**, we have:

$$\begin{bmatrix} \Pr(X_{k+1} = 0 \mid X_k = 1) \\ \Pr(X_{k+1} = 1 \mid X_k = 1) \\ \Pr(X_{k+1} = 2 \mid X_k = 1) \\ \Pr(X_{k+1} = 3 \mid X_k = 1) \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.6 \\ 0.0 \\ 0.0 \end{bmatrix}$$

Here we see that if after *k* samplings there is 1 person who has been selected every time, then after sampling *k*+1 there must either be 1 or 0 people who have been selected every time. The probability $\Pr(X_{k+1} = 0 \mid X_k = 1)$ is computed as:

$$\binom{1}{0}\binom{4}{3} \div \binom{5}{3} = (1)(4) / 10 = 0.4.$$

The computation of this probability can be understood as follows: From a dichotomous population of 5 individuals, 1 of "type 1" (having been selected on all prior samplings) and 4 of "type 2" (having <u>not</u> been selected on all prior samplings), we have obtained a simple random sample of size 3 without replacement. The probability computed is the

hypergeometric probability that the single "type 1" individual is not selected and that 3 of the 4 "type 2" individuals are selected.

Similarly, the probability $\Pr(X_{k+1} = 1 \mid X_k = 1)$ is the hypergeometric probability that the single "type 1" individual is selected and that only 2 of the 4 "type 2" individuals are selected, and is computed as:

$$\binom{1}{1}\binom{4}{2} \div \binom{5}{3} = (1)(6)/10 = 0.6.$$

We can interpret the third and fourth columns of **M** in a similar fashion. With regard to the third column of **M**, we have:

$$\begin{bmatrix} \Pr(X_{k+1} = 0 \mid X_k = 2) \\ \Pr(X_{k+1} = 1 \mid X_k = 2) \\ \Pr(X_{k+1} = 2 \mid X_k = 2) \\ \Pr(X_{k+1} = 3 \mid X_k = 2) \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \\ 0.0 \end{bmatrix}$$

The first 3 probabilities constitute the hypergeometric distribution associated with a simple random sample of size 3 (without replacement) from a dichotomous population of size 5, consisting of 2 "type 1" and 3 "type 2" individuals. Of course it is impossible to select 3 "type 1" individuals from a population containing only 2 "type 1" individuals, hence the result that $\Pr(X_{k+1} = 3 \mid X_k = 2) = 0$.

With regard to the fourth column of **M**, we have:

$$\begin{bmatrix} \Pr(X_{k+1} = 0 \mid X_k = 3) \\ \Pr(X_{k+1} = 1 \mid X_k = 3) \\ \Pr(X_{k+1} = 2 \mid X_k = 3) \\ \Pr(X_{k+1} = 3 \mid X_k = 3) \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.3 \\ 0.6 \\ 0.1 \end{bmatrix}$$

All 4 probabilities constitute the hypergeometric distribution associated with a simple random sample of size 3 (without replacement) from a dichotomous population of size 5, consisting of 3 "type 1" and 2 "type 2" individuals. It should be noted that for this particular case the sampling must select at least 1 "type 1" individual, hence the result that $\Pr(X_{k+1} = 0 \mid X_k = 3) = 0$.

Denoting the probabilities of matrix **M** as $p_{ij} = \Pr(X_{k+1} = i \mid X_k = j)$, the matrix **M** can be expressed as:

$$\mathbf{M} = \begin{bmatrix} p_{00} & p_{01} & p_{02} & p_{03} \\ p_{10} & p_{11} & p_{12} & p_{13} \\ p_{20} & p_{21} & p_{22} & p_{23} \\ p_{30} & p_{31} & p_{32} & p_{33} \end{bmatrix}$$

where
$$p_{ij} = \begin{cases} \dfrac{\binom{j}{i}\binom{N-j}{n-i}}{\binom{N}{n}} & \text{if } i \le j \text{ and } n-i \le N-j \\ 0 & \text{otherwise} \end{cases}$$

Also, let us denote the probability distribution associated with the random variable X_r as

$$\boldsymbol{\pi_r} = \begin{bmatrix} \Pr(X_r = 0) \\ \Pr(X_r = 1) \\ \Pr(X_r = 2) \\ \Pr(X_r = 3) \end{bmatrix}$$

Thus, the probability distribution associated with the random variable X_{k+1} can be obtained from the probability distribution associated with the random variable X_k as $\mathbf{M}\boldsymbol{\pi_k} = \boldsymbol{\pi_{k+1}}$. However, given $\boldsymbol{\pi_1}$ one can utilize the recursive nature of the formula for $\boldsymbol{\pi_{k+1}}$ to determine that $\boldsymbol{\pi_{k+1}} = \mathbf{M}^k \boldsymbol{\pi_1}$.

6. Results for Mr. Erford's Original Problem

The original problem had the following parameters: $N = 200$, $n = 125$, and $r = 9$. In this case,

$$\boldsymbol{\pi_1} = \begin{bmatrix} \Pr(X_1 = 0) \\ \Pr(X_1 = 1) \\ \vdots \\ \Pr(X_1 = 125) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

and

$$\mathbf{M} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0,125} \\ p_{10} & p_{11} & \cdots & p_{1,125} \\ \vdots & \vdots & \ddots & \vdots \\ p_{125,0} & p_{125,1} & \cdots & p_{125,125} \end{bmatrix}$$

where
$$p_{ij} = \begin{cases} \dfrac{\binom{j}{i}\binom{200-j}{125-i}}{\binom{200}{125}} & \text{if } i \le j \text{ and } 125-i \le 200-j \\ 0 & \text{otherwise} \end{cases}$$

The solution for the particular set of parameters specified is as follows:

i	$\Pr(X_9 = i)$		i	$\Pr(X_9 = i)$		i	$\Pr(X_9 = i)$
0	.0470		4	.1704		8	.0050
1	.1513		5	.0953		9	.0013
2	.2351		6	.0429		10	.0003
3	.2352		7	.0159		11	.0001

It should be noted that this table actually pertains to values of i extending to a maximum possible value of $n = 125$. However, for all values of i greater than 11, the associated probabilities are 0 (to 4 decimal places), and hence have be omitted from the table.

7. A Markov Chain Problem

The knowledgeable student who has taken more advanced courses may recognize that the random process discussed here is simply a discrete-time homogeneous Markov chain, with the matrix **M** being the transition matrix. It is a Markov chain because the probability distribution of the random variable X_{k+1} depends only on the immediately preceding random variable X_k. It is time-homogeneous because the transition matrix is constant from one sampling to the next.

For students unfamiliar with Markov Chain theory, this problem can serve as an introduction once the student's interest has been piqued after solving the problem using only first principles of probability. Having gone through the entire process of deriving a solution, the student will more greatly appreciate the compactness and elegance of the Markov Chain approach.

8. Conclusion

This consulting project can be used as a teaching tool in many contexts in the future. It will make an excellent senior capstone project for a mathematical statistics major in our department, and can also be included in a variety of courses, such as applied probability, stochastic processes, statistical computing, and statistical consulting. While this problem certainly can be posed and its solution be presented to the students within one class period, it is our opinion that the best approach is to allow the students to have time to explore and investigate it as we have outlined in this paper. The process of experiencing both the joys and challenges of solving a problem independently is critical to the development of mathematical thinking.

References

1. Chin Long Chiang. *An Introduction to Stochastic Processes and their Applications.* Robert E. Krieger Publishing Co., New York, 1980.