

# Random Walks as Motivational Material in Introductory Statistics and Probability Courses

Lynn A. Fisher\* and Donald St. P. Richards†

May 1, 2004

## Abstract

Recent articles have described the advantages of teaching elementary statistics and probability classes using approaches which encourage greater student engagement, including experimentation, with the subject matter. We describe our experiences in introducing the subject of random walks to small groups of high-school and first-year college students. As we show in this article, the topic of random walks provides a superb way for instructors to introduce a class to elementary simulation problems, calculation of expectations and measures of variability for geometric distributions, real-world interpretation and consequences for the divergence of infinite series, and the behavior of random walks on restricted sets in the plane. Most enchantingly, all facets of this journey are entirely accessible to an involved class of students equipped with minimal knowledge of calculus. Based on our experiences, we strongly recommend student involvement in the teaching of introductory concepts to small classes.

## 1 Introduction

In recent articles Hogg (1996), Magel (1996), Gelman and Glickman (2000), and other authors described some advantages accruing to students and faculty when introductory statistics and probability classes utilize approaches which encourage student involvement. The *Chance* project (Snell, 1994) is a well-known example of an involvement-based approach, and Magel (1996) has given several references to pedagogical studies of involvement-based approaches.

In this paper we describe our experiences in introducing basic concepts from probability theory, notably, elementary aspects of random walks, to high-school and first-year college students. We demonstrate that the topic of random walks is sufficiently broad-based that it can be used to explain to students concepts ranging from high-school to graduate levels. As we show in this article, an appealing feature of the topic of random walks is that it contains sufficiently many surprising twists to enchant students for extended periods of time.

---

\*Lynn Fisher teaches mathematics and computer science at Woodstock Union High School, Woodstock, VT 05091, and is a graduate student in statistics at the Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405.

†Donald Richards teaches statistics and probability in the Department of Statistics, Penn State University, University Park, PA 16802. This work was supported in part by the National Science Foundation, grant DMS-9703705.

*Key words and phrases.* *Chance* database project, collaborative learning, Jack-or-Jill controversy, involvement-based teaching, statistical education.

## 2 Jack or Jill?

While teaching introductory courses in statistics and probability, we introduced our students to the “Jack or Jill” controversy, a topic related to the issue of prospective parents who wish to determine the sex of their offspring. The “Jack or Jill” topic is well-known to adherents of the popular *Chance* method of teaching statistics. Under the *Chance* approach, class meetings place emphasis on group-wide discussions rather than the customary lecture method, and students maintain diaries containing their reflections on the subject material, questions, and homework assignments. For further discussion of this topic, we refer to Assignment 12 at the *Chance* web page given in the reference to Snell (1994).

In our classes, we introduced the “Jack or Jill” controversy to our students as follows:

*Suppose a society values sons more than daughters. In this society, couples continue to have children until they produce a son, at which point they stop having children.*

After the students had read this item and discussed it among themselves, they were asked to vote “Yes,” “No,” or “Undecided” on the following question:

*Over the long term, does this family-planning scheme lead to a society in which boys outnumber girls?*

All students were required to vote, so their level of involvement was at its maximum. The question, perhaps because of its very nature, also revealed a mild tension between female and male students. Nevertheless, we did not sense that voting was done largely along gender lines.

After collecting the vote, a prominent defender of each viewpoint was asked to make a brief presentation to the class to explain the vote of their faction. After a summary by the instructor, class members gave general agreement to the basic assumption of independence of gender from birth to birth; further, they agreed to assume that there was a 50% probability of “success” (i.e., the arrival of a boy) on each individual “trial.” Then, it was elementary to derive the geometric distribution of  $X$ , the number of daughters in a randomly chosen family. Indeed, this discussion marked the introduction of the geometric distribution to the class.

The concept of  $E(X)$ , the “mean” or long-term average value of  $X$ , is next introduced. In fact, after additional discussion, students often formulated by themselves the concept of  $E(X)$ . Then we proved, by the standard method of summing the usual infinite series, that  $E(X) = 1$ , demonstrating that the average family will have one girl and one boy.

At this point, it was not uncommon for women in our classes *literally* to stand up and cheer loudly, indicating, perhaps, that the “mild tension” referred to above was not so mild after all. Some disgruntled opponents then argued in favor of modifying the assumption of a 50% probability of success, arguing that 50% is too high and should be replaced by a lower number,  $p$ . However, they were quickly discouraged from doing so by more astute colleagues. These colleagues argued correctly that a lower value of  $p$  leads to larger families, for it takes a longer time for the first male to arrive; and since each family has only one boy then boys, on average, will be outnumbered by girls. These arguments were buttressed by further calculations with the geometric distribution, proving to the class that for general  $p$ , the average family would have  $1/p$  children, and only one of these would be a boy. Hence, the percentage of boys in an average family would be  $1 \div (1/p) = p$ .

These episodes indicated that in introductory courses at the high-school and college levels, involved students will devote their mental energies keenly to the formulation of concepts such as “probability distribution” and “long-term average value,” and their societal implications.

At the high-school level, a small number of students had some difficulty in summing the infinite series for  $E(X)$ . In accordance with the students’ backgrounds, we then provided various proofs for calculating the expectation.

The long-term outcome of the above child-bearing scheme was demonstrated to students by the following question, suggested by the related discussion on the *Chance* web page:

*Simulate the outcomes of fifty families. What is the overall ratio of sons to daughters? Construct a histogram for the number of sons and daughters, and calculate the sample mean and standard deviation of the number of sons and daughters.*

In the classroom, students performed these simulations by tossing coins, rolling dice, or using a calculator (e.g., the *TI-83 PLUS* calculator). Computation of the ratio of sons to daughters and other summary statistics were usually done with a calculator, and histogram sketches were done by hand or by calculator. In computer laboratory sessions outside the classroom, we taught the students to carry out the simulations, numerical calculations, and histogram construction using computer programs, e.g., MINITAB. Particularly for classes at the high-school level, we recommend the use of the *True BASIC* software package (Catlin, 1996, Venit and Schleiffers, 1999) owing to its relative simplicity for all levels of students, the ease with which it can be modified to generate additional data whenever necessary, the ease of producing graphic output (which was particularly critical to our exploration of random walks in the latter part of this unit), its very low cost, and the ease with which it can be translated into other languages. With each program have included documentation so that a reader who is unfamiliar with *BASIC* syntax can easily understand the algorithm sufficiently well in order to translate the code into a more familiar language.

We provided to our classes the following *True BASIC* program which simulates the outcomes of fifty families which opt to have children until the arrival of a boy. The program is written as simply as possible since, up to this stage, students were assumed to have no programming experience. Naturally, the program signals the arrival of girls and boys by coloring pixels in pink and blue, respectively. The program simulates the outcomes for `numFamilies`, a specified number of families, and displays the mean family size and the proportion of boys.

```
! Program #1: Jack_or_Jill?
! True BASIC program to simulate fifty families which have children until
! the arrival of a boy. The program displays the births in each family,
! the average family size, and the percentage of boys in the family.
```

```
RANDOMIZE                ! random seed (new pseudo-random numbers)
LET numFamilies = 50      ! declare number of experiments
LET count = 0             ! initialize total population of children
FOR i = 1 to numFamilies
  DO                      ! repeat the following:
    IF rnd < .5 then      ! if pseudo-random number is less than .5
      LET sex = 0         ! "0" corresponds to birth of girl
      SET COLOR 13        ! pink on most systems
      PRINT "G";         ! display result of this birth
```

```

ELSE                                ! if pseudo-random number of .5 or greater
  LET sex = 1                        ! "1" corresponds to birth of boy
  SET COLOR "blue"
  PRINT "B";                          ! display result of this birth
END IF
LET count = count + 1              ! increment total population size
LOOP until sex = 1                  ! continue until a boy is born
PRINT                               ! output on a new line for the new family
NEXT i
! Display statistics:
PRINT
PRINT "The mean number of children per family is: ";count/numFamilies
PRINT "The percent of children that are boys is: ";numFamilies/count*100;"%"
END

```

When these simulation outcomes were depicted on the students' computer screens by histograms, it was immediately clear to students that  $E(X)$  was unlikely to be as large as five, say. In our experience, it was useful for students to see the compatibility between the mathematical result that  $E(X) = 1$  and the simulated values of  $E(X)$ .

At this stage, we initiated a discussion of the spread or variability of the histograms. We then defined the standard deviation and variance of the random variable  $X$ , estimated them heuristically from the histograms, and concluded this part of the course by calculating the standard deviation and variance of  $X$  from its probability distribution.

### 3 More Jack than Jill?

The next part of the course was motivated by an intriguing question, often posed by students:

*Suppose that families develop an even stronger preference for sons, to the extent that they contemplate having children until the number of sons first exceeds the number of daughters. What are the consequences for society of such a preference?*

To answer this question, we introduced our students to the concept of a restricted random walk on the plane. Students are asked to depict a family with no children as located at the origin. As each child arrives, the family takes a unit step to the right if the child is a girl, or a unit step upwards if the child is a boy. In this way, the random walk proceeds in the portion of the positive quadrant below the line  $y = x$ , and the journey ends once the path crosses the line  $y = x$ . The outcome is that each family's outcome is identified with a random walk on the lattice points in the nonnegative quadrant.

For example, consider a family whose trials ended (some would say "began"! ) with the outcome GGBBGBB, so that the family then had three girls and four boys; the corresponding random walk is depicted by the diagram in Figure 1.

Please place Figure 1 here
----------------------------

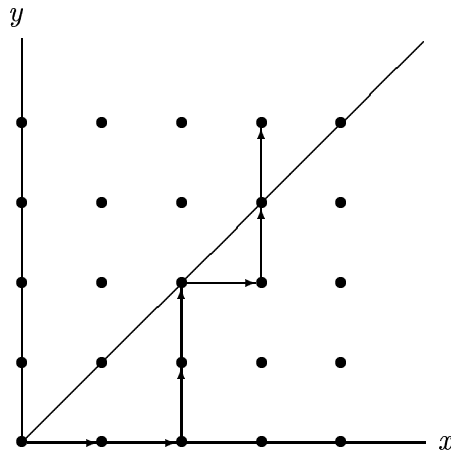


Figure 1: Random walk for a family with outcome GGBBGBB

Having identified the more-Jack-than-Jill problem with a restricted random walk in the nonnegative quadrant, we helped our students to write a *True BASIC* program to simulate a single family which continues to have children until the number of boys first exceeds the number of girls. An example is given in the following program, which generates births repeatedly, increments the counter `numGirls` or `numBoys` according to whether the latest birth is a girl or boy, respectively, and halts the process when `numBoys` first exceeds `numGirls`.

```
! Program #2: More_Jack_than_Jill_Random_Walk
! A random walk which models births until more boys than girls are present
! in a family. Displays the family size and the number of boys and girls.

RANDOMIZE                               ! random seed (new pseudo-random numbers)

! Standard window coordinates, with origin in lower left corner.
! Program uses increments in pixels on the current system's screen to
! ensure that vertical and horizontal increments of equal length will
! have equal length, regardless of screen on which the program is run.
SET MODE "graphics"
ASK PIXELS px, py                       ! px = # of pixels across horizontal
                                         ! py = # of pixels across vertical

SET WINDOW -10, px-1, -10, py-1
LET step = 10                            ! step size in pixels
SET COLOR "red"
PLOT 0,0; py-1,py-1                     ! display y = x line
LET numBoys = 0                          ! initialize counters of number of boys
LET numGirls = 0                         ! and number of girls

LET x = 0                                ! initialize random walker at (0,0)
LET y = 0
```

```

PLOT x,y;                ! display initial position of random walker
                        ! Semi-colon is True BASIC syntax for holding
                        ! the pen down after plotting the point
DO                        ! repeat following:
  IF rnd < .5 then       ! if pseudo-random number is less than .5
    SET COLOR 13         ! set color to pink
    LET numGirls = numGirls + 1 ! increment number of girls
    LET x = x + step     ! increment x-coordinate of random walker
  ELSE                   ! if pseudo-random number is .5 or greater
    SET COLOR "blue"    ! set color to blue
    LET numBoys = numBoys + 1 ! increment number of boys
    LET y = y + step     ! increment y-coordinate of random walker
  END IF
  PLOT x,y;            ! draw line segment to new position of
                        ! random walker, hold pen down.
LOOP until numBoys > numGirls ! stop when number of boys > number of girls.
PRINT "The family size is"; numBoys + numGirls;","
PRINT "with"; numBoys ;"boys, and"; numGirls; "girls."
END

```

At this stage, two remarkable events occurred. Some students reported simulated families with gigantic numbers of children. Other students reported that their computers crashed suddenly; on further investigation, they discovered that the computer malfunctions were caused by family sizes too large for their computers to handle.

To explain these phenomena, we denoted by  $X$  and  $Y$  the number of girls and boys, respectively, in a randomly chosen family. Then we wish to derive the probability distribution of  $N = X + Y$ , the total number of children. We again adopted the usual assumptions of independence of gender from birth to birth and the 50% probability of “success” on each individual trial. After viewing the combined results of the class’ computer simulations, students were asked to guess the value of  $E(N)$ , the expected number of children in a family. Despite many hints from us, few students are willing to hazard guesses as large as one million, say. It then comes with much surprise and amazement to the students when they are informed that

$$E(N) = \infty. \quad (3.1)$$

At this stage, we divide the exposition into two parts, one for students at the high-school level, and the other for college students.

### 3.1 Students at the high-school level

To explain (3.1) to our students, we noted that since families cease to have children when  $Y = X + 1$ , it follows that  $N = X + Y = 2X + 1$ . Therefore  $N$  attains odd integer values only, and so we need to calculate  $P(N = 2k + 1)$  for  $k = 0, 1, 2, \dots$ .

We usually began by posing to our classes the problem of calculating, e.g.,  $P(N = 7)$ . A short discussion results in the identification of any such family’s outcome as one of a number of equally likely paths. Then our students deduced that

$$P(N = 7) = \frac{\text{The number of lattice paths from } (0, 0) \text{ to } (3, 3) \text{ in the region } \{x \geq y \geq 0\}}{2^7}.$$

Here, it was straightforward for students to determine that there are five such paths; therefore  $P(N = 7) = 5/2^7$ . Students were also reminded by this example that the probability distribution of  $N$  was determined completely by the enumeration of lattice paths in the region  $\{x \geq y \geq 0\}$ .

In the general case, to calculate  $W(m, n)$ , the number of lattice paths from  $(0, 0)$  to  $(m, n)$  which stay in the region  $\{x \geq y \geq 0\}$ , we demonstrated to the class three properties of  $W$ . We noted that

$$W(1, 0) = W(1, 1) = 1, \quad (3.2)$$

for there is only one path from  $(0, 0)$  to  $(1, 0)$  or from  $(0, 0)$  to  $(1, 1)$ . Next,

$$W(m, m) = W(m, m - 1), \quad m \geq 1, \quad (3.3)$$

for there only one way to walk from  $(m, m)$  to  $(m, m - 1)$ . Last, we pointed out that

$$W(m, n) = W(m - 1, n) + W(m, n - 1), \quad m > n \geq 1, \quad (3.4)$$

since one may arrive at the point  $(m, n)$  in two ways only, either by walking from  $(m - 1, n)$  or from  $(m, n - 1)$ . We then taught our students how to apply these recurrence relations to calculate values of  $W(m, n)$  for numerous values of  $m$  and  $n$ , and provided them with the general formula

$$W(m, n) = \frac{(m+n)!(m-n+1)}{(m+1)!n!}. \quad (3.5)$$

Using this formula, we deduced that

$$P(N = 2k + 1) = \frac{W(k, k)}{2^{2k+1}} = \frac{(2k)!}{(k+1)!k!2^{2k+1}}, \quad (3.6)$$

$k = 0, 1, 2, \dots$  and, in turn, that

$$E(N) = \sum_{k=0}^{\infty} (2k+1) \frac{(2k)!}{(k+1)!k!2^{2k+1}}. \quad (3.7)$$

Here, we asked our students to accept that, for large  $k$ , the  $k$ th term in this series is approximately equal to  $1/(\pi k)^{1/2}$ ; as can be imagined, students also used their calculators to check the accuracy of this result. With this approximation in hand, we noted that for sufficiently large  $k$ , the  $k$ th term was greater than  $1/2k$ . By appeal to the divergence of the harmonic series, we were able to explain to our students the divergence of (3.1).

Our experience was that students never fail to be amazed by the result that  $E(N) = \infty$ . For many of them, it marked the first instance in which the concept of infinity is more than simply an abstract, academic notion in the classroom.

To impress our students further, we again raised the issue of whether sons will outnumber daughters in the society. Here, we calculated  $E(X/Y)$ , the expected ratio of girls to boys in a randomly chosen family. By writing  $X/Y = (N - 1)/(N + 1)$ , we then expressed the expectation as an infinite series. In the classroom, we simply used *True BASIC* to sum a large number of terms of the series, thereby obtaining an approximation to the exact value,

$$E(X/Y) = 2 \ln 2 - 1 = 0.3863, \quad (3.8)$$

a result which will be established in the Appendix. Students then saw that the society will achieve its purpose: There will be more sons than daughters, with about thirty-nine daughters for every one hundred sons in the average family. But it will be a pyrrhic victory, for the society almost certainly will be bankrupted by the costs of dealing with such gigantic families.

### 3.2 Students at the college level

For students at the college level, all the material in the previous subsection was covered in class and at a higher level of rigor. Thus, we first derived the recurrence relations (3.2) - (3.5), and applied them to calculate various values of  $W(m, n)$ . It was our experience that, when given this problem as a homework assignment, many students at this level could formulate a conjecture for the general formula for  $W(m, n)$  after examining numerous special cases. In the case of students having knowledge of mathematical induction, we solved the recurrence relations and deduced the formula (3.5).

Next, we stated Stirling's formula and thereby approximated each factorial term in (3.6). We deduced that, as  $k \rightarrow \infty$ , the terms in the series (3.7) are approximately equal to  $1/(\pi k)^{1/2}$ ; hence the series diverges. Also, as regards the calculation of  $E(X/Y)$ , the expected ratio of girls to boys in a random family, we observed that

$$E\left(\frac{X}{Y}\right) = E\left(\frac{N-1}{N+1}\right) = 1 - E\left(\frac{2}{N+1}\right) = 1 - \frac{1}{2} \sum_{k=0}^{\infty} \frac{1}{k+1} \frac{(2k)!}{(k+1)! k! 2^{2k}}. \quad (3.9)$$

The convergence of this series can again be verified using Stirling's formula. We also obtained the exact value (3.8) of this expectation by representing the infinite series in (3.9) as a double integral, the details being provided in the Appendix.

In our classes, we concluded the presentation of each topic with wide-ranging review sessions in which students discussed the material and raised questions about all aspects of the topic. The resulting discussions raised many issues appropriate for review in an involvement-based manner. We now describe some of the issues arising from our lectures on random walks, and we will point the reader toward related references.

With regard to (3.8), some students found it paradoxical that  $E(X+Y) = \infty$  even though  $E(X/Y) < \infty$ . Here, we reminded students that  $0 \leq X/Y = (N-1)/(N+1) < 1$ , and then it was clear that  $E(X/Y) < 1$ .

Some students asked whether the conclusions would change if  $p$ , the probability of success, were different from 50%. In subsequent discussions, we first determined that, for general  $p$ ,

$$P(N = 2k + 1) = W(k, k) p^{k+1} (1-p)^k, \quad (3.10)$$

$k = 0, 1, 2, \dots$ , for in a family with  $2k + 1$  children, there will be  $k + 1$  boys and  $k$  girls. By summing these probabilities, we can see that  $P(N < \infty) = 1$  for  $p \geq 1/2$ ; therefore, with probability one, any family's random walk will end. Moreover, the expected value of  $N$  is

$$E(N) = \sum_{k=0}^{\infty} (2k + 1) \frac{(2k)!}{(k+1)! k!} p^{k+1} (1-p)^k. \quad (3.11)$$

In class, we used *True BASIC* to sum a large number of terms of this series for various values of  $p$ . We provided to students the following table of values of  $E(N)$  near the critical value  $p = 1/2$ . This table illustrated the fact that  $E(N)$  is finite for all  $p > 1/2$ , and it demonstrated the rapid increase in  $E(N)$  as  $p$  approaches  $1/2$ . We also noted, as shown in the Appendix, that this series has the closed-form expression

$$E(N) = 1/(2p - 1), \quad (3.12)$$

For  $p < 1/2$ , we again sum the probabilities in (3.10) and find that  $P(N < \infty) < 1$ . Equivalently, the probability that a randomly chosen family will have infinitely many children

Table 1: Values of  $E(N)$  near  $p = 1/2$ 

$p$	0.50	0.505	0.51	0.53	0.55	0.60
$E(N)$	$\infty$	100	50	16.67	10	5

is non-zero, so it follows that  $E(N) = \infty$ . As we show in the Appendix,  $P(N < \infty) = p/(1-p)$ , hence  $P(N = \infty) = (1 - 2p)/(1 - p)$ .

Another aspect of our discussions centered on the numbers

$$W(k, k) = \frac{(2k)!}{(k+1)!k!} = \frac{1}{k+1} \binom{2k}{k}$$

which appeared in (3.6). We mentioned that these are the well-known *Catalan numbers*. We also used this as an opportunity to introduce our students to some of the history of the Catalan numbers, including their prominence in the work of Euler; and their ubiquity in statistics, combinatorics and probability, particularly in the study of random walks (Mohanty, 1979).

Owing to the imbroglio concerning the counting of votes in Florida during the 2000 U.S. presidential elections, students were particularly interested to learn of the connection between the Catalan numbers and the classical ballot problem. In this problem, it is desired to calculate the probability that the winning candidate in a two-candidate election holds the lead throughout the counting of ballots. Students easily made the connection between the ballot problem and the more-Jack-than-Jill problem. In the case of more advanced students, we referred to applications of the ballot problem to statistical rank-order tests (Feller, 1969, p. 69) and to variations of the classical ballot problem (e.g., Chao and Severo, 1991; Mohanty, 1979), and we briefly discussed André's reflection principle (Feller, 1969, p. 70; Zeilberger, 1983) and the general area of random walks on Weyl chambers (Gessel and Zeilberger, 1992; Richards and Gross, 1995) and other restricted sets.

Other topics which arose in some of our class discussions included the connections between random walks and statistical inference. We described applications to the problem of estimating the number of fish species in a lake (Lehmann, 1983, p. 92) and, more generally, to sequential analysis. Students learnt that the values of  $E(N)$ , and especially the fact that  $E(N) < \infty$  only for the case in which  $p \neq 1/2$ , had practical implications for sequential binomial sampling plans (Lehmann, 1983, p. 134, Problem 3.9).

## 4 Concluding Remarks

The teaching methods and material described in this article were utilized by us in *Chance*-type and other introductory classes at the high-school and college levels, and in several summer programs for highly gifted high-school students. We greatly enjoyed teaching these classes along the lines described earlier.

In the course evaluations, some students described the material presented here as “riveting.” Students felt broadly that they enjoyed the class, that it motivated self-learning and discovery due to the necessity for all students to participate in the process, and that small classes would generally benefit greatly from this style of teaching. Some students mentioned that they were so imbued with their experiences in this class that they found themselves broaching the idea of a career in the statistical sciences.

In the case of moderately large classes, we feel that this material could still be presented, particularly so if students were able to attend small-enrollment discussion sections run by a faculty member or by well-trained graduate teaching assistants. In such a setting, it is imperative to set aside sufficient time to ensure that students are comfortable enough with the chosen programming software in order to be able to write and modify programs.

In conclusion, it is clear that the exploratory teaching style described in this article is more time-consuming than the traditional approach of formal lectures combined with reading assignments and exercises. Nevertheless, we opine that the pay-off in terms of increased student comprehension and excitement about the subject is well worth the time expense.

**Acknowledgments.** We are grateful to the Editor, an Associate Editor, and Gábor Székely for comments on an earlier version of this manuscript, leading to a greatly improved revision.

## References

- Andrews, G. E., Askey, R., and Roy, R. (1999). *Special Functions*. Cambridge University Press, New York.
- Catlin, A. (1996). *Let's Program It ... In True BASIC*, 3rd. edition. True BASIC Press, West Lebanon, N.H.
- Chao, C.-C., and Severo, N. C. (1991). Distributions of ballot problem random variables. *Ann. Appl. Prob.*, **23**, 586–597.
- W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. I*, second edition, Wiley, New York, 1969.
- Gelman, A., and Glickman, M. E. (2000). Some class-participation demonstrations for introductory probability and statistics. *J. Educat. Behavioral Statist.*, **25**, 84–100.
- Gessel, I. M., and Zeilberger, D. (1992). Random walk in a Weyl chamber. *Proc. Amer. Math. Soc.*, **115**, 27–31.
- Hogg, R. V. (1991). Statistical education: Improvements are badly needed. *The American Statistician*, **45**, 342–343.
- Lackritz, J. R. (1997). Increasing student participation in large introductory statistics classes. *The American Statistician*, **51**, 210–210.
- Lehmann, E. (1983). *Theory of Point Estimation*. Wiley, New York.
- Magel, R. C. (1996). Increasing student participation in large introductory statistics classes. *The American Statistician*, **50**, 51–56.
- Mohanty, S. G. (1979). *Lattice Path Counting and Applications*. Academic Press, New York.
- Richards, D. St. P., and Gross, K. I. (1995). Total positivity, harmonic analysis and random walks on Weyl chambers. *Contemp. Math.*, **191**, 153–161.
- Snell, J. L. (1994). See *The Chance Database Project*; Assignment 12, “Jack or Jill,” at <http://www.dartmouth.edu/~chance/course/Syllabi/94D/94D.html> (based on an article from *Lancet*, March 20, 1993).
- Venit, S. M., and Schleifers, S. M. (1999). *Programming in True BASIC: Problem Solving with Structure and Style*. PWS Publishing, Pacific Grove, CA.
- Zeilberger, D. (1983). André’s reflection proof generalized to the many-candidate ballot problem. *Discrete Math.*, **44**, 325–326.

## 5 Appendix

We establish (3.8) by evaluating the series in (3.9). First, we note that this series may be represented as a double integral,

$$\sum_{k=0}^{\infty} \frac{1}{k+1} \frac{(2k)!}{(k+1)! k! 2^{2k}} = \int_0^1 \int_0^1 (1-tu)^{-1/2} du dt. \quad (5.1)$$

This is proved by expanding the integrand  $(1-tu)^{-1/2}$  in a power series,

$$(1-tu)^{-1/2} = \sum_{k=0}^{\infty} \frac{(2k)!}{k! 2^{2k}} t^k u^k$$

and then integrating term by term. On the other hand, the integral in (5.1) may be evaluated by direct integration. Since

$$\int_0^1 (1-tu)^{-1/2} du = \frac{2[1-(1-t)^{1/2}]}{t},$$

$0 < t < 1$ , then the integral (5.1) equals

$$\int_0^1 \frac{2[1-(1-t)^{1/2}]}{t} dt.$$

Substituting  $s = 1 - t^2$ , we find that this latter integral equals

$$4 \int_0^1 \frac{s}{1+s} ds = 4(1 - \ln 2).$$

By (3.9), we obtain  $E(X/Y) = 1 - 2(1 - \ln 2) = 2 \ln 2 - 1$ , as stated in (3.8).

The sum of the probabilities in (3.10) over all  $0 \leq k < \infty$  may be readily derived from the theory of hypergeometric series. For any real number  $a$  and nonnegative integer  $k$ , define the *rising factorial*  $(a)_0 = 1$  and  $(a)_k = a(a+1) \cdots (a+k-1)$ ,  $k \geq 1$ . It is straightforward to verify that  $(2k)! = 2^{2k}(1/2)_k(1)_k$  and  $(k+1)! = (2)_k$ , and then we obtain

$$P(N < \infty) = p \sum_{0 \leq k < \infty} \frac{(1/2)_k(1)_k}{k!(2)_k} (4p(1-p))^k = p \cdot {}_2F_1 \left( \begin{matrix} 1/2, 1 \\ 2 \end{matrix}; 4p(1-p) \right)$$

where  ${}_2F_1$  denotes the Gaussian hypergeometric function. Applying the first formula in Problem 39 of Andrews, et al. (1992), p. 185, we obtain

$$P(N < \infty) = p \cdot \frac{2}{1 + \sqrt{1 - 4p(1-p)}} = \frac{2p}{1 + |1 - 2p|}.$$

For  $p \geq 1/2$  this expression equals 1, and for  $p < 1/2$  it equals  $p/(1-p)$ . Therefore, for  $p < 1/2$ ,  $P(N = \infty) = 1 - P(N < \infty) = (1 - 2p)/(1 - p)$ .

We show similarly that the series (3.11), for  $p > 1/2$ , has the closed-form expression (3.12). Noting that  $(2k+1) \cdot (2k)! = 2^{2k}(1)_k(3/2)_k$  and  $(k+1)! = (2)_k$ , we find that (3.11) becomes

$$E(N) = p \sum_{k=0}^{\infty} \frac{(1)_k(3/2)_k}{k!(2)_k} (4p(1-p))^k = p \cdot {}_2F_1 \left( \begin{matrix} 1, 3/2 \\ 2 \end{matrix}; 4p(1-p) \right).$$

By the second formula in Problem 39 of Andrews, et al. (1992), p. 185, it follows that

$$E(N) = \frac{p}{\sqrt{1-4p(1-p)}} \cdot \frac{2}{1 + \sqrt{1-4p(1-p)}} = \frac{1}{2p-1},$$

which is (3.12).