COMPUTERS & CALCULATORS

Co-Edited by

Thomas M. Green
Contra Costa College
San Pablo, California 94806

and

Glenn D. Allinger
Montana State University
Bozeman, Montana 59714

In this section readers are encouraged to share their experiences with computers and calculators as they apply to the two-year college mathematics curriculum. There is special interest in innovative uses of these tools to solve problems, to present concepts, and to define new directions for curriculum development. All material for this section should be sent to Thomas Green. (See inside cover for details on submitting manuscripts.) Be sure to include with your paper a copy of the computer or calculator program and a successful run and output.

Binomial Baseball

Eugene M. Levin

Eugene Levin received his Ph.D. in Physics from New York University in 1967 and is Associate Professor of Physics at York College of the City University of New York. He became interested in the statistics of hits and wins in baseball during six seasons as a Little League manager.

Programmable calculators can be used to play simulated baseball games. Student access to programmable calculators and computer terminals, coupled with a familiarity with baseball, provides an opportunity to enhance their understanding of the binomial distribution in statistics. The use of random numbers to determine hits and outs introduces the student to the Monte Carlo method [1] in a familiar context. For various hit probabilities, the average number of hits per game agrees well with the expected values computed from the negative binomial distribution.

Our programmed baseball game proceeds by determining each batter's performance by comparing a random number with a given set of "hit probabilities" for that batter. All batters on team X are assigned a common "probability of getting a hit," p_x, and likewise with team Y.

A nine-inning game is played in the following manner: The first player of team X comes to bat: a random number r is generated, such that $0 < r \leqslant 1.0$. This number is compared with four numbers $H < T < D < P_x$ which represent thresholds for home run, triple, double, and single base hit, respectively.

Home Run:	$0 < r \leqslant H$
Triple:	$H < r \leqslant T$
Double:	$T < r \leqslant D$
Single:	$D < r \leqslant P_x$
Out:	$P_x < r \leqslant 1.0$

If, for example, the next random number generated indicates the batter gets a double (two-base hit), any runner at first base is advanced to third base, and the batter to second. Runners score upon reaching home plate, and three "outs" retire the sides, clearing all bases but retaining the accumulated score. The opposing team, with its "hit probability," p_y, comes to bat. The same values of H, T, and D are used for both teams. At the end of the nine innings, the game is over and the two scores are compared. Extra innings are played in the event of a tied score after nine innings. Other features of baseball—stolen bases, errors, walks, double plays, etc., could be introduced in a more sophisticated program, but this would complicate the game analysis. The program used on the H.P. 9820 calculator is shown in the appendix.

Distribution of "Hits". The standard binomial distribution describes the probability of obtaining X successes in N trials, given a "success" probability p per trial. The sum of these probabilities from $X = 0$ to $X = N$ is 1.00 [2]. Calling a "hit" a success, we are interested in the distribution of "number of hits per game," regardless of their being single-, double-, triple-base hits or home runs. Our case differs from the standard distribution in two ways:

1. It is the total number of failures ("outs") in a nine-inning game, rather than trials ("at bats"), which is fixed.
2. The last batter on each team, by definition, can only strike out (or fly out or ground out). Thus, in a nine-inning game there are only 26 of the 27 outs which are freely distributable among the hits.

This situation is described by the "Negative Binomial Probability Distribution"[3] which gives the probability of getting h successes (hits) prior to the 27th failure (out) in a nine-inning game:

$$P(p,h) = \frac{(26+h)!}{26!h!} \cdot p^h \cdot (1-p)^{27}. \tag{1}$$

The sum of terms $P(p,h)$, from h = zero to infinity, equals 1.00. In principle there is no upper limit to the number of hits achievable by either team in a baseball game. However, the probability of achieving many hits (e.g., $h > 25$), for $p \leqslant 0.300$, is quite small. It is easy to construct a program which evaluates this sum over the range $h = 0$ to $h = 25$ for a fixed p. It is instructive for the student to include this summation in a program designed to evaluate the expected value of h:

$$\langle h \rangle = \sum_{h=0}^{h_{\max}} h \cdot P(p,h). \tag{2}$$

Neglecting terms involving $h > h_{\max}$ may be justified heuristically.

Results. Six 50-game series were run, with p_x fixed at 0.200, and p_y varied from 0.200 in the first series to 0.300 in the final series. The multiple-base-hit thresholds were fixed at $H = 0.04$, $T = 0.10$, $D = 0.20$. For each series, the frequency, N, with which team Y obtained hits was plotted against h; these histograms are shown in Figure 1, along with the corresponding negative binomial distributions.

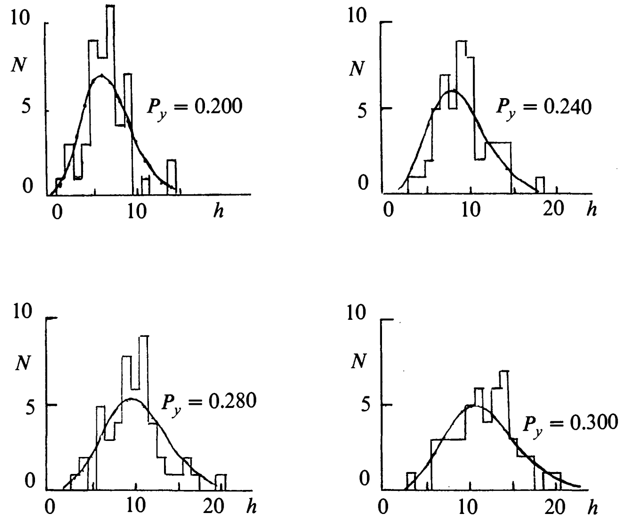

Figure 1. Histograms: Number of games N in which team Y gets h hits, for four 50-game series, with various p_y. Curves represent the corresponding negative binomial distributions, vertically scaled × 50.

As expected, increasing the hit probability from 0.200 to 0.300 shifts the peak to the right. The mean values of h were found, using:

$$\bar{h} = \sum_{h=0}^{h_{\max}} h \cdot N_h / \Sigma N_h \tag{3}$$

and are presented (Table 1) with the corresponding values of $\langle h \rangle$ (from (2)); the agreement is quite good. The final column depicts the data for a larger sample (350 games), with $p_y = 0.200$.

Table 1. Expected values of the number of hits per game $\langle h \rangle$ from negative binomial distribution, compared with mean values obtained in six 50-game series and one 350-game series.

hit probability p	.200	.220	.240	.260	.280	.300	.200
number of games	50	50	50	50	50	50	350
$\langle h \rangle_{\text{binom.}}$	6.8	7.8	8.5	9.5	10.5	11.5	6.75
$\bar{h}_{\text{exper.}}$	6.5	7.6	9.1	9.8	9.9	11.7	6.76

Extra-inning games have been included in the histograms (of hits per nine-inning game) by dividing the total number of hits (for team Y) by the total "at bats" to give the batting average for that game. The integral number of hits that would yield a batting average, in a nine-inning game, closest to that value, was taken as the appropriate number of hits.

Some game statistics are shown in Table 2, for a set of six 50-game series with p_x constant at 0.200, and p_y ranging from 0.200 to 0.300. The range of values of p_y is larger than the spread of team batting averages within the American and National Leagues (.248 to .279 as of this writing). The probabilities of extra-base hits are also inflated. This was designed to yield higher-scoring games and to compensate for absent elements (stolen bases, errors, walks, hit batters, etc.) which can contribute to higher scores. The same procedure was carried out for the last column with realistic major league values [4] for the extra-base-hit thresholds (see Table 2).

Table 2. Game statistics for six 50-game series, for various p_y (columns 1–6). Extra-base-hit thresholds are $H = 0.04$, T = 0.10, $D = 0.20$, producing multiple-base-hit probabilities: (home run, 0.04), (triple, 0.06), (double, 0.10). Probability of single-base hit, $P_{\text{single}} = (p\text{-}D)$. A 50-game series between Oakland ($p_x = .248$) and Detroit ($p_y = .279$) is shown in column 7, with realistic threshold values: $H = 0.023$, $T = 0.028$, $D = 0.066$ for both teams.

Team X							
P_x	.200	.200	.200	.200	.200	.200	.248
number of games won	29	28	20	16	15	14	17
total runs scored	144	147	131	131	140	135	73
batting average	.188	.204	.192	.196	.198	.187	.242
Team Y							
P_y	.200	.220	.240	.260	.280	.300	.279
number of games won	21	22	30	34	35	36	33
total runs scored	136	150	171	177	217	229	110
batting average	.196	.206	.243	.242	.274	.298	.282
number of extra inning games	13	12	7	9	9	9	9

Batting Averages and Runs Scored. Team batting averages (number of hits/number of times at bat) are found to be rather close to the hit probabilities (Table 2), as one would expect for a fairly large sample. Baseball games are won, of course, by the team with the most runs, not hits. However, fortuitous grouping of hits can lead to runs. About one-seventh of all computer games were won by the team with fewer hits.

An attempt to determine the expected number of runs in a nine-inning game, directly from probability theory, would be quite difficult. One complicating factor is that at the end of a half-inning (at the third "out") runners may be (and frequently are) left on bases. While it has been assumed here that the division of the game into innings does not affect "hit" statistics, it seems certain to affect scoring statistics (runs per team). Our Monte Carlo technique is well suited to explore the effect of hit probability and multiple-base-hit thresholds on scoring.

The relationship between single-base-hit probability (p_y-D) and runs scored per game is explored as follows. The probabilities of a home run, triple, and double (H, T-H, and D-T, respectively) remain unchanged in each case, as p_y (and therefore p_y-D) is varied. For the data of Table 2 the mean number of runs per game is plotted (Figure 2) against the probability of a single-base hit (p_y-D) (curve B). From data gathered from another set of six 50-game series, using higher threshold values ($H = 0.04$, $T = 0.10$, $D = 0.20$), another graph is similarly plotted

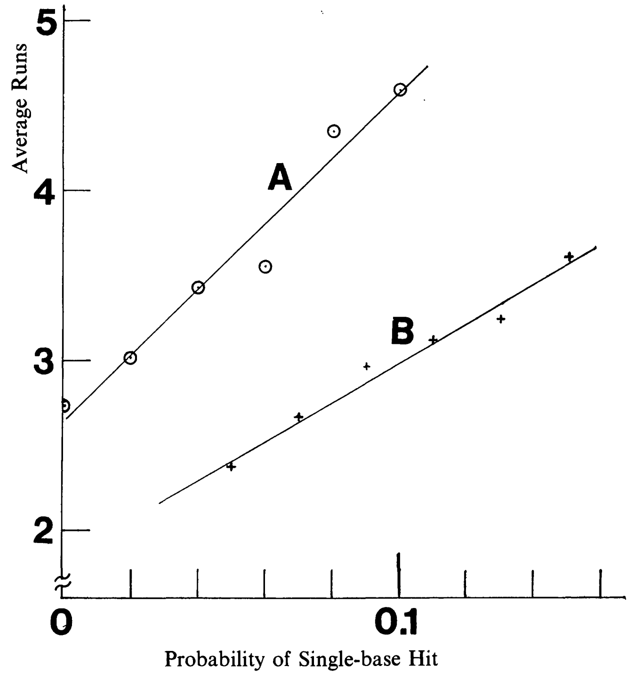

Figure 2. Mean number of runs scored per game vs. single-base-hit probability (p_y-D), with linear least square fits. Two sets of data are plotted, for which the respective home run, triple, and two-base-hit probabilities are: Curve A: 0.04, 0.06, 0.10; Curve B: 0.04, 0.02, 0.09.

(curve A). For simplicity, the data are fitted to straight lines (least square fits). A least square program was used which calculates the y-intercept, slope, and the uncertainty in the slope [5]; the values are shown in Table 3.

Table 3. Variation in single-base-hit probability (= inverse slope) corresponding to one run per game. Data of Figure 2.

	Threshold Values				
Curve	H	T	D	Slope	Inverse Slope
A	.04	.1	.2	19.2 ± 1.9	.052
B	.04	.06	.15	11.6 ± 0.9	.086

The significance of this calculation may be understood in terms of the team manager's dilemma: there are several key "long-ball" hitters on his team who commonly hit doubles and triples but may be only fair fielders. The complete team must include some excellent key fielders whose hitting probabilities may vary inversely to their fielding abilities. He can improve his team defensively by sacrificing single-base-hit probability. The inverse of the slope (Table 3, last column) represents the amount of single-base-hit probability which, if added to the team, will yield one additional run per game (on the average). Because of his superior double and triple hitting, the manager of team A (curve A, Figure 2) need only increase the team single-base-hit probability by 0.052 to yield one more (expected) run per game. The team B manager (curve B, Figure 2) can add one additional expected run per game by increasing the team single-base-hit probability by 0.086, but only by sacrificing defensive ability.

It is known that baseball is amenable to statistical studies, and indeed may be the most statistically analyzed of all sports. I have tried to show, in this article, that baseball can serve as an interesting vehicle for an introduction to probability and statistics. There is probably no upper limit, except that imposed by computer time and cost, to the degree of sophistication that can be introduced into the programming of game play and analysis.

Acknowledgment. The author is greatly indebted to Dr. Joseph Malkevitch for his critical comments and suggestions.

REFERENCES

1. D. E. Raeside, An introduction to Monte Carlo methods, Am. J. Phys. 42 (1974) 21.
2. Philip R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969, p. 31.
3. A. S. Ukena, Statistics Today, Harper & Row, New York, 1978, p. 134.
4. Courtesy of Public Relations Dept., New York Yankees Baseball Club, River Ave., Bronx, N. Y.
5. J. Higbie, Uncertainty in a least-square fit, Am. J. Phys., 46 (1978) 945.

Appendix

Program for running a sequence of six series of 'baseball games', using the Hewlett-Packard desk calculator (Model 9820 A). Each fifty-game series is 'played' with fixed values of p_x and p_y. The value of p_y is increased by 0.020 at the end of each series (step 28 in program), p_x remaining fixed throughout the sequence.

Step	Instruction
0	prt "A PROB HIT =", R17
1	prt "B PROB HIT =", R18
2	1 + R22 → R22; 0 → R11 → R12 → R13 → R23 → R24 → R25 → R26
3	0 → R1 → R2 → R3 → R4 → R5 → R6 → R7 → R8 → R9 → R10; 3 → R0
4	(R19 + π)**4 → R19; R19-int R19 → R19
5	if flg 0; 1 + R24 → R24; R18 → R8; go to +2
6	1 + R23 → R23; R17 → R8
7	if R8 > R19; go to 30
8	RO − 1 → RO
9	if RO ≠ RO; go to 4
10	if flg 0; R4 + R5 + R6 + R7 + R12 → R12; R13 + 1 → R13; clr flg 0: go to +2
11	R4 + R5 + R6 + R7 + R11 → R11; set flg 0; go to 3
12	if R13 ⩽ 8; go to 3
13	if R13 = R12; set flg 1; go to 3
14	prt "GAME NO", R22
15	prt "TEAM X RUNS =", R11
16	prt "TEAM Y RUNS =", R12
17	prt "TEAM X BATAV =", R25/R23; R25/R23 + R27 → R27
18	prt "TEAM Y BATAV =", R26/R24; R26/R24 + R28 → R28
19	if flg 1: R29 + 1 → R29; clr flg 1
20	if R11 > R12; 1 + R20 → R20; go to +2
21	R21 + 1 → R21
22	if R22 ⩽ 49; go to 2
23	prt "TEAM X WINS =", R20; 0 → R20
24	prt "TEAM Y WINS =", R21; 0 → R21
25	prt "XTRA INNING GAMES =", R29; 0 → R29
26	prt "TEAM X AVG OF AVGS =", R27/R22; 0 → R27
27	prt "TEAM Y AVG OF AVGS =", R28/R22; 0 → R28
28	0 → R22; R18 + .020 → R18
29	go to 0
30	if flg 0; R26 + 1 → R26; go to +2
31	R25 + 1 → R25
32	if R14 > R19; 4 → R10; go to +4
33	if R15 > R19; 3 → R10; go to +3
34	if R16 > R19; 2 → R10; go to +2
35	1 → R10
36	if R3 > 0; 1 + R4 → R4; R3 − 1 → R3
37	if R2 > 0; 1 → R(R10 + 2); R2 − 1 → R2
38	if R1 > 0; 1 → R(R10 + 1): R1 − 1 → R1
39	1 → RR10
40	0 → R10; go to 4