COMPUTER SIMULATIONS TO MOTIVATE AND/OR CONFIRM THEORETICAL CONCEPTS

Elliot A. Tanis
Hope College
Holland, Michigan 49423

Abstract

Undergraduate students do not always have the theoretical background to prove and/or appreciate certain concepts. Computer simulations can provide greater understanding and will motivate some students to delve more deeply into the theory. In addition, questions and extensions of the original problem are often raised by the simulations. Some possible examples include the following: Let $Z_1, Z_2, Z_3$ be a random sample of size three from a standard normal distribution. What can be said about the distribution of $Z_2/\sqrt{(Z_1^2 + Z_3^2)/2}$? (If $Z_3$ would be in the numerator, it would have a t-distribution.) On the average, how many "random numbers" must be added together so that their sum exceeds one? When sampling from the standard normal distribution, what is an unbiased estimator for the standard deviation $\sigma$? What is the average length of a confidence interval for $\mu$ when sampling from a normal distribution with unknown variance? Is it really possible to simulate observations of the order statistics in order? For the "birthday problem," how many students have to give their birthdays, on the average, to obtain a match?

1 Introduction

Undergraduate students who are taking their first course in mathematical statistics and probability often lack the theoretical tools, the appreciation, and/or the inspiration to delve into certain theoretical concepts. In order to give them a better feeling for probabilistic and/or statistical concepts, it is often helpful if they write a computer program to simulate the problem. In the writing of the computer program for the simulation, they increase their own understanding of the problem and the necessary hypotheses, and sometimes further questions arise that enhance their learning.

In this paper some illustrative problems are given. Problems such as these can be solved by students in a laboratory setting, as regular assignment problems, or as extra credit problems. It is helpful to have subroutines that provide graphical support for some of the problems. For others, a good random number generator is all that is needed.
Some of these examples were motivated by journal articles, some are based on textbook exercises, some grew out of talks that were given at professional meetings, others were motivated by problems that were raised in conversations with participants at meetings. They have all been used with Hope College students as exercises, test questions, and/or demonstrations.

2 Examples

A. Let $Z_1$, $Z_2$, $Z_3$ be a random sample of size three from a standard normal distribution. It is well known that $T = Z_3/\sqrt{(Z_1^2 + Z_2^2)/2}$ has a t distribution with 2 degrees of freedom. Suppose that we let $W = Z_3/\sqrt{(Z_1^2 + Z_2^2)/2}$. What can be said about the distribution of $W$? The following program simulates values of $W$. (All of the programs in this paper are written in BASIC for an IBM-PC.)

100 SCREEN 2: CLS: RANDOMIZE TIMER: KEY OFF
110 PRINT "This program simulates N observations of
    Z2/SQR((Z1^2 + Z2^2)/2)"
120 PRINT :PRINT "The empirical data is then compared graphically with
    the theoretical distribution."
130 PRINT :PRINT "Input the number of repetitions (up to 200);\n    N = ";: INPUT N
140 IF N > 0 AND N < 201 THEN 160
150 GOTO 130
160 DIM X(200)
170 PRINT
180 FOR K = 1 TO N
190 GOSUB 25500 : REM SNRML - THE STANDARD NORMAL DISTRIBUTION -
    RETURNS XX
200 Z1 = XX
210 GOSUB 25500 : REM SNRML
220 Z2 = XX
230 X(K) = Z2/SQR((Z1^2 + Z2^2)/2)
240 PRINT USING "###";K;: PRINT " ";: PRINT USING
    "####";X(K);: PRINT " ";
250 NEXT K
260 GOSUB 5000 : REM GRAPHC - A GRAPHING ROUTINE FOR CONTINUOUS DATA
270 END
2000 REM THE PDF IS DEFINED HERE, V AS A FUNCTION OF U
2010 IF U <= -SQR(2) + .00001 OR U >= SQR(2) - .00001
    THEN V = 0: RETURN
2020 PI = 3.14159
2030 V = 1/PI/SQR(2 - U^2): RETURN

2
2200 REM THE DISTRIBUTION FUNCTION IS DEFINED HERE, V AS A
FUNCTION OF U
2210 PI = 3.14159
2220 IF U < -SQR(2) THEN V = 0: RETURN
2230 IF U < 0 THEN V = (1/PI)*ATN(SQR(2/U^2 - 1)): RETURN
2240 IF U = 0 THEN V = .5: RETURN
2250 IF U < SQR(2) THEN V = 1 - (1/PI)*ATN(SQR(2/U^2 - 1)): RETURN
2260 V = 1: RETURN

Sample output based on 200 observations of W is illustrated in the following figures.

To find the distribution of W theoretically, we first note that if \( V = Z_1/Z_2 \), then
V has a Cauchy distribution. It is also easy to see that \( -\sqrt{2} < W < \sqrt{2} \). The
distribution function of W, for \( 0 < w < \sqrt{2} \), is

\[
F(w) = P(W \leq w) = 1 - P(W > w) \\
= 1 - P(0 < V < \sqrt{2/w^2 - 1}) \\
= 1 - (1/\pi)\arctan(\sqrt{2/w^2 - 1}), \ 0 < w < \sqrt{2}
\]

Because of the symmetry of this distribution,

\[
F(w) = (1/\pi)\arctan(\sqrt{2/w^2 - 1}), \ -\sqrt{2} < w < 0
\]

It follows that the p.d.f. of W is given by
\[ f(w) = \frac{1}{\pi \sqrt{2 - w^2}}, \quad -\sqrt{2} < w < \sqrt{2} \]

B. Let that random variable \( U \) have a uniform distribution on the interval \((0,1)\). Given a sequence of independent observations of \( U: u_1, u_2, u_3, \ldots \), let the random variable \( X = \min\{k: u_1 + u_2 + \ldots + u_k > 1\} \). That is, \( X \) is the smallest number of "random numbers" that must be added together so that their sum exceeds one. We shall find the expected value of \( X, E(X) \). That is, on the average, how many random numbers must be added together so that their sum exceeds one?

This problem is very easy to simulate.

100 CLS: RANDOMIZE TIMER: KEY OFF: SCREEN 2
110 DIM X(200), F(20)
120 PRINT "This program simulates observations of X, the minimum number of random numbers"
130 PRINT "that must be added together to obtain a sum that is greater than one."
140 PRINT :PRINT " k \sum x k \sum x k \sum x k \sum x"
150 PRINT: N = 200
160 FOR K = 1 TO 200
170 SUM = 0
180 SUM = SUM + RND
190 X(K) = X(K) + 1
200 IF SUM < 1 THEN 180
210 PRINT USING "#####";K; :PRINT USING "#####";SUM;
220 IF K = INT(K/4)*4 THEN PRINT
230 S = S + X(K)
240 SQ = SQ + X(K)^2
250 NEXT K
260 XBAR = S/200
270 SD = SQR(((200*SQ - S*S)/200)/199)
280 PRINT :PRINT "The sample mean and sample standard deviation are:"
290 PRINT :PRINT "xbar = ";XBAR,"s = ";SD
300 PRINT :PRINT "The distribution mean and standard deviation are:"
310 PRINT
320 DEN = 1
330 FOR K = 2 TO 20
340 DEN = DEN*K
350 F(K) = (K-1)/DEN
360 NEXT K
370 EX = 0: EX2 = 0
380 FOR K = 2 TO 20
390   EX = EX + K*F(K)
400   EX2 = EX2 + K*K*F(K)
410   NEXT K
420   MU = EX
430   SIGMA = SQR(EX2 - MU^2)
440   PRINT "mu = ";MU,";"sigma = ";SIGMA
450 PRINT :INPUT "Hit ENTER to see graphs of the data",A$
460   Gosub 5000: REM GRAPHD - A GRAPHING ROUTINE FOR DISCRETE DATA
470   END

2000 REM The p.d.f. of X defined by V as a function of U
2010 IF U < 2 OR U > 20 THEN V = 0: RETURN
2020 V = F(U): RETURN
2200 REM The distribution function of X defined by V as a function of U
2210 IF U < 2 THEN V = 0: RETURN
2220 IF U > 20 THEN V = 1: RETURN
2230 V = F(U)
2240 IF U = 2 THEN RETURN
2250 FOR K = 3 TO U
2260   V = V + F(K)
2270   NEXT K
2280 RETURN

Typical output along with two graphical displays follows.

<table>
<thead>
<tr>
<th>k</th>
<th>x</th>
<th>k</th>
<th>x</th>
<th>k</th>
<th>x</th>
<th>k</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.3223</td>
<td>2</td>
<td>1.1791</td>
<td>3</td>
<td>1.0596</td>
<td>4</td>
<td>1.0596</td>
</tr>
<tr>
<td>5</td>
<td>1.3223</td>
<td>6</td>
<td>1.2804</td>
<td>7</td>
<td>1.1222</td>
<td>8</td>
<td>1.2804</td>
</tr>
<tr>
<td>13</td>
<td>1.3223</td>
<td>14</td>
<td>1.4575</td>
<td>15</td>
<td>1.4575</td>
<td>16</td>
<td>1.4575</td>
</tr>
<tr>
<td>21</td>
<td>1.3223</td>
<td>22</td>
<td>1.2804</td>
<td>23</td>
<td>1.2804</td>
<td>24</td>
<td>1.2804</td>
</tr>
<tr>
<td>25</td>
<td>1.3223</td>
<td>26</td>
<td>1.2804</td>
<td>27</td>
<td>1.2804</td>
<td>28</td>
<td>1.2804</td>
</tr>
<tr>
<td>29</td>
<td>1.3223</td>
<td>30</td>
<td>1.2804</td>
<td>31</td>
<td>1.2804</td>
<td>32</td>
<td>1.2804</td>
</tr>
<tr>
<td>34</td>
<td>1.3223</td>
<td>35</td>
<td>1.2804</td>
<td>36</td>
<td>1.2804</td>
<td>37</td>
<td>1.2804</td>
</tr>
<tr>
<td>37</td>
<td>1.3223</td>
<td>38</td>
<td>1.2804</td>
<td>39</td>
<td>1.2804</td>
<td>40</td>
<td>1.2804</td>
</tr>
<tr>
<td>41</td>
<td>1.3223</td>
<td>42</td>
<td>1.2804</td>
<td>43</td>
<td>1.2804</td>
<td>44</td>
<td>1.2804</td>
</tr>
<tr>
<td>45</td>
<td>1.3223</td>
<td>46</td>
<td>1.2804</td>
<td>47</td>
<td>1.2804</td>
<td>48</td>
<td>1.2804</td>
</tr>
<tr>
<td>49</td>
<td>1.3223</td>
<td>50</td>
<td>1.2804</td>
<td>51</td>
<td>1.2804</td>
<td>52</td>
<td>1.2804</td>
</tr>
<tr>
<td>53</td>
<td>1.3223</td>
<td>54</td>
<td>1.2804</td>
<td>55</td>
<td>1.2804</td>
<td>56</td>
<td>1.2804</td>
</tr>
<tr>
<td>57</td>
<td>1.3223</td>
<td>58</td>
<td>1.2804</td>
<td>59</td>
<td>1.2804</td>
<td>60</td>
<td>1.2804</td>
</tr>
<tr>
<td>61</td>
<td>1.3223</td>
<td>62</td>
<td>1.2804</td>
<td>63</td>
<td>1.2804</td>
<td>64</td>
<td>1.2804</td>
</tr>
<tr>
<td>65</td>
<td>1.3223</td>
<td>66</td>
<td>1.2804</td>
<td>67</td>
<td>1.2804</td>
<td>68</td>
<td>1.2804</td>
</tr>
<tr>
<td>69</td>
<td>1.3223</td>
<td>70</td>
<td>1.2804</td>
<td>71</td>
<td>1.2804</td>
<td>72</td>
<td>1.2804</td>
</tr>
<tr>
<td>73</td>
<td>1.3223</td>
<td>74</td>
<td>1.2804</td>
<td>75</td>
<td>1.2804</td>
<td>76</td>
<td>1.2804</td>
</tr>
<tr>
<td>77</td>
<td>1.3223</td>
<td>78</td>
<td>1.2804</td>
<td>79</td>
<td>1.2804</td>
<td>80</td>
<td>1.2804</td>
</tr>
</tbody>
</table>

The sample mean and sample standard deviation are:

xbar = 2.81  s = .909671

The distribution mean and standard deviation are:

mu = 2.718282  sigma = .857026

Hit ENTER to see graphs of the data.
Hopefully, after students see the results of several simulations, they will guess that $E(X) = e$.

An outline of the theoretical solution is as follows:

1. Show that $P(U_1 + U_2 + ... + U_k < 1) = \frac{1}{k!}$

2. Show that the distribution function of $X$, defined at a positive integer $k$, is

   $$F(k) = P(X \leq k) = 1 - \frac{1}{k!}$$

3. Show that the p.d.f. of $X$ is

   $$f(x) = \frac{x-1}{x!}, \quad x = 2, 3, 4, ...$$

4. Show that $E(X) = e$.

C. Let $X_1, X_2, ..., X_n$ be a random sample of size $n$ from a normal distribution $N(\mu, \sigma^2)$. Then

   $$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

is an unbiased estimator of $\sigma^2$. This is easy to show using the fact that $(n-1)S^2/\sigma^2$ has a chi-square distribution with $n - 1$ degrees of freedom. A more challenging problem is to find a constant $c$ so that $cS$ is an unbiased estimator of $\sigma$. It can be shown that (e.g., see Exercise 5.2-7 in [1])

   $$c = \frac{\Gamma\left(\frac{n-1}{2}\right) \sqrt{n-1}}{\Gamma\left(\frac{n}{2}\right) \sqrt{2}}$$
It is then interesting to simulate these results. For example, let the distribution of $X$ be $N(75,400)$. The following program simulates 100 samples of size 5 from this distribution, and for each, calculates the values of $s^2$ and $cs$. The averages of these 100 $s^2$s and $cs$s is then found. They should be close to 400 and 20, respectively. It is advantageous to have several students perform this simulation and then combine the results. It is also instructive to have the sample variances printed on the screen as they are calculated so that students learn to appreciate the large variance of the sample variance, which in this case is $\text{Var}(S^2) = 80,000$.

100 CLS: RANDOMIZE TIMER: KEY OFF: SCREEN 2
110 PRINT "This program simulates 100 samples of size 5 from a normal distribution"
120 PRINT "N(75,400). For each sample the value of $s^2$ and the value of 
130 PRINT "$(8/3)/SQR(2*PI)$s is calculated, the unbiased estimates 
of the variance and"
140 PRINT "standard deviation, respectively. The averages of these 
estimates are found. They should be close to 400 and 20, respectively."
150 C = 8/3/SQR(2*3.141592653#)
155 PRINT :PRINT "k  s^2  cs  k  s^2  cs"
160 FOR K = 1 TO 100
170 SUMX = 0: SUMX2 = 0
180 FOR J = 1 TO 5
190 Z = 0
200 FOR L = 1 TO 12
210 Z = Z + RND
220 NEXT L
230 X = 20*(Z - 6) + 75
240 SUMX = SUMX + X
250 SUMX2 = SUMX2 + X^2
260 NEXT J
270 SVAR = (5*SUMX2 - SUMX^2)/20
280 CS = C*SQR(SVAR)
290 PRINT USING "###";K;: PRINT USING "###.###";SVAR;:
295 IF K = INT(K/2)*2 THEN PRINT
300 SSVAR = SSVAR + SVAR
310 SCS = SCS + CS
320 NEXT K
330 PRINT :PRINT "The average of the sample variances is ";SSVAR/100
340 PRINT :PRINT "The average of the unbiased standard deviations is ";SCS/100
D. Related to the last problem is that of finding the length of a confidence interval for the mean of a normal distribution. Let \( X_1, X_2, ..., X_n \) be a random sample from a normal distribution, \( N(\mu, \sigma^2) \). If \( \sigma \) is known, then the length of a \( 100(1 - \alpha)% \) confidence interval for \( \mu \) is

\[
\text{length} = 2z_{\alpha/2} \sigma / \sqrt{n}
\]

If \( \sigma \) is unknown and the sample size is small, we cannot give the length of the confidence interval. However, we can find the expected value of the length. That is,

\[
E(\text{length}) = \frac{2 \Gamma \left( \frac{n}{2} \right) \sqrt{2}}{\Gamma \left( \frac{n-1}{2} \right) \sqrt{n-1} \sqrt{n}} \frac{\sigma}{t_{\alpha/2}(n-1)}
\]

If, for example, \( n = 5 \), \( (1 - \alpha) = 0.90 \), \( \mu = 50 \), and \( \sigma^2 = 16 \), the length of the confidence interval using the known variance is

\[
\text{length} = 2(1.645)(4)/\sqrt{5} = 5.8853
\]

while the average length, when the variance is unknown and the \( t \) confidence interval is used, is (e.g., see Exercise 5.3-8 in [1])

\[
E(\text{length}) = \frac{\frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi} \sqrt{2}}{\sqrt{4}} \frac{4}{\sqrt{5}}(2.132) = 7.1699
\]

It is instructive to compare the confidence intervals of the two types. Both types give 90% confidence intervals and on the average the \( z \) intervals are shorter. However the lengths of the \( t \) intervals are sometimes very short and sometimes quite long. The
following graphical display illustrates this, with the t-intervals on the left and the z-intervals on the right.

E. When studying order statistics, it is possible to simulate values of the order statistics by first simulating $n$ observations, $x_1, x_2, ..., x_n$, and then ordering these observations. A more challenging and interesting problem is to simulate the values of the order statistics in order — that is, first simulate the value of the smallest order statistic, then the second smallest, etc. This would be appropriate in a life testing experiment where you observe the first $m$ "deaths" and $m \leq n$. The following program was motivated by an article by Lurie and Hartley [2], will do this simulation.

100 CLS: RANDOMIZE TIMER: KEY OFF: SCREEN 2
110 INPUT "N = "; N
120 INPUT "M = "; M
130 DIM Y(N)
140 REM Define the distribution function of X
150 DEF FNA(X) = (GIVE THE DEFINITION AS A FUNCTION OF X)
160 REM Define the inverse of the distribution function of X
170 DEF FNB(W) = (GIVE THE DEFINITION AS A FUNCTION OF W)
180 W = 1 - RND^(1/N)
190 Y(1) = FNB(W)
200 PRINT "1",Y(1)
210 REM Generate the next order statistics recursively using conditional dist.
220 FOR I = 1 TO M - 1
230 X = Y(I)
240 W = 1 - (1 - FNA(X)*RND^(1/(N - I))
250 Y(I+1) = FNB(W)
260 PRINT I,Y(I)
270 NEXT I
280 END
F. In many probability and statistics classes the "birthday problem" is discussed. That is, what is the probability that at least two students share a common birthday. If the class is small, say around 20, this is easily checked. It is also possible to check for a common birthday in a large class. However, since there is a large probability of a common birthday, a different question could be considered, namely, how many students have to give their birthdays, on the average, before there is a match. That is, suppose that the students are numbered 1, 2, 3, ... and they give their birthdays. Let the random variable $X$ equal the number of the first student whose birthday matches one of the earlier students. What is the value of $\mu = E(X)$? What is the p.d.f. of $X$?

We shall discuss a modification of this problem and then give a simulation of it. Consider successive rolls of a fair "$M$-sided die." Let the random variable $X$ equal the minimum number of rolls required so that one of the faces is observed twice. That is, $X$ is the roll on which the first match occurs. Then the p.d.f. of $X$ is

$$f(x) = \frac{M}{M} \frac{M - 1}{M} \frac{M - 2}{M} \frac{M - 3}{M} \ldots \frac{M - (x - 2)}{M} \frac{x - 1}{M}, \quad x = 2, 3, 4, ..., M + 1$$

Thus

$$E(X) = \sum_{x=2}^{M+1} x f(x)$$

When $M = 365$ (the birthday problem), $E(X) = 24.6166$. What else can be said about the distribution of $X$? It is interesting to simulate observations of $X$ for different values of $M$. The following program will do this. Sample output is given when $M = 20$. For a more complete discussion, see [3].

100 CLS: RANDOMIZE TIMER: SCREEN 2: KEY OFF
110 LOCATE 1,30:PRINT "Birthday Problem"
120 PRINT :PRINT "This program simulates the roll of an $M$-sided die until a match occurs. The"!
130 PRINT "random variable $X$ is equal to the number of rolls required to obtain this first"
140 PRINT "match. A value of $M = 365$ simulates the birthday problem. The user may select the number of faces on the die (up to 365) and the number of repetitions (up to 400). The empirical and theoretical results can be compared graphically."
150 DIM X(400), F(366), FL(365)
160 PRINT :INPUT "The number of faces on the die is $M =$ "; M
170 IF M > 0 AND M < 366 THEN 200
180 PRINT "Select a value for $M$ that is greater than 0 and less than 366."
190 GOTO 160
200 PRINT :INPUT "The number of repetitions is $N =$ "; N
210 IF N > 0 AND N < 401 THEN 240
220 PRINT "Select a value for $N$ that is greater than 0 and less than 401."
230 GOTO 200

10
CLS
PRINT "Number of rolls to obtain a match with a ";M;"-sided die."
PRINT
FOR K = 1 TO N
FOR J = 1 TO M: FL(J) = 0: NEXT J
D = INT(M*RND) + 1
X(K) = X(K) + 1
IF FL(D) = 1 THEN 340
FL(D) = 1
GOTO 290
SUM = SUM + X(K)
SUMSQ = SUMSQ + X(K)*X(K)
PRINT K;"-";X(K),
NEXT K
XBAR = SUM/N
SVAR = (N*SUMSQ - SUM*SUM)/N/(N - 1)
PRINT :PRINT "The sample mean is ";XBAR
PRINT "The sample variance is ";SVAR
PRINT "The sample standard deviation is ";SQR(SVAR)
PP = 1
FOR K = 2 TO M + 1
PP = PP*(M + 2 - K)/M
F(K) = PP*(K - 1)/M
MU = MU + K*F(K)
E2 = E2 + K*K*F(K)
NEXT K
PRINT :PRINT "The theoretical mean is MU =";MU
PRINT "The theoretical variance is VAR(X) =";E2 - MU*MU
PRINT "The standard deviation is SIGMA = ";SQR(E2 - MU*MU)
PRINT :PRINT "Hit ENTER to continue.";:INPUT A$
GOSUB 5000 : REM GRAPHD
END
IF U < 2 OR U > M + 1 THEN V = 0: RETURN
V = F(U): RETURN
IF U < 2 THEN V = 0: RETURN
IF U > M + 1 THEN V = 1: RETURN
V = F(2)
IF U = 2 THEN RETURN
FOR K = 3 TO U
V = V + F(K)
NEXT K
RETURN
3 Conclusion

The examples are just that, examples. They can be expanded, modified, and improved upon. And many other examples could be given.

These examples are available on a disk for an IBM PC. If you are interested in receiving a copy of this disk, sent $10 to the author to cover the cost. The programs are written in BASIC and they can be modified easily.

4 References

