Computer Simulations to Motivate Understanding

Elliot A. Tanis
Hope College
Holland, Michigan 49423

Introduction

Students who are taking their first course in probability and/or mathematical statistics often lack the theoretical tools, appreciation, background, or experience for attacking the solution of a problem. And even if they are able to solve a problem theoretically, they do not always have an intuitive understanding of the result.

There are various ways in which the computer can be used to provide this understanding and motivation for further study. These include:

- Use simulation to provide insight into a theoretical solution.
- Use simulation to confirm and/or gain a better understanding of the theoretical solution.
- Simulate a plausible solution when the necessary theoretical tools are not known and perhaps motivate further study.
- Let the computer calculate the mean, variance, and standard deviation when a closed form solution is not known.

Some examples from a variety of sources are given that illustrate various uses of the computer. These examples are motivated by journal articles, textbook exercises, talks at professional meetings, and conversations with statisticians. Computer programs that you write to solve these problems will usually be quite short. They require only a decent "random number generator" such as the built in RND in IBM BASIC or GWBASIC. Graphical output will enhance several of the solutions. (See the Conclusions Section.) In addition, physical experiments can sometimes be used to help understand the problems.
Examples

1. On the average, what is the minimum number of random numbers that must be added together so that their sum exceeds 1? That is, if \( u_1, u_2, u_3, \ldots \) is a sequence of independent random numbers that are selected from the interval \((0,1)\) and \(X = \min\{k : u_1 + u_2 + \cdots + u_k > 1\}\), what is the value of \(E(X)\)?

Using a table of “random numbers” or a “random number generator” on a calculator, it is easy to simulate observations of \(X\). This is an interesting problem for an entire class to simulate, and in the process, a large number of observations can be found rather quickly. Writing the results obtained by individual class members on the board will help students see not only the randomness of such an experiment, but will also provide additional data for guessing the answer. Furthermore, if you tally the numbers of times that \(x = 2, 3, 4, \ldots\) are observed, the students will perhaps be able to guess that the probability function of \(X\) is

\[
f(x) = P(X = x) = \frac{x - 1}{x!}, \quad x = 2, 3, 4, \ldots.
\]

It is instructive to count and tally and calculate by hand, but time constraints also suggest that a computer simulation could provide a larger number of repetitions of this experiment more easily and rather quickly. Questions about the appropriate sample size can be discussed so that students begin to think about the number of repetitions that should be used.

A program to simulate this experiment is short and the simulation does not take very long. To see how a computer program can be written to solve a problem like this, consider the following program in BASIC for an IBM (compatible) computer:
100 CLS: RANDOMIZE TIMER: KEY OFF
110 N = 500: REM N is the number of repetitions
120 FOR K = 1 TO N
130 X = 0: SUM = 0
140 SUM = SUM + RND
150 X = X + 1
160 IF SUM < 1 THEN 140
170 PRINT X,
180 SUMOFX = SUMOFX + X
190 NEXT K
200 XBAR = SUMOFX/N
210 PRINT
220 PRINT "The sample mean of the x's is "; XBAR
230 END

On each of five successive runs of this program, the average numbers of random numbers needed so that their sums exceeded 1 were 2.714, 2.690, 2.736, 2.728, and 2.736.

After the students have guessed that the answer to the question is $E(X) = e$, hopefully they will become intrigued by the problem and will want to prove this result theoretically. This problem appeared on The William Lowell Putnam examination in 1958. Some of the places in which the solution appears and extensions of this problem are discussed are given in [1,3,4,5,6,7,10].

2. An urn contains $n$ balls that are numbered from 1 to $n$. Take a random sample of size $n$ from the urn, one at a time. A match occurs if ball numbered $k$ is selected on draw $k$. Let $A$ be the event that at least one match is observed. Show that

$$P(A) = 1 - \left(1 - \frac{1}{n}\right)^n$$

when sampling with replacement, and

$$P(A) = 1 - \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!}\right)$$

when sampling without replacement. In addition, show that

$$\lim_{n \to \infty} P(A) = \lim_{n \to \infty} \left[1 - \left(1 - \frac{1}{n}\right)^n\right] = 1 - \frac{1}{e}$$
and
\[
\lim_{n \to \infty} P(A) = \lim_{n \to \infty} \left[ 1 - \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!}\right) \right] = 1 - \frac{1}{e}.
\]
Thus for "large" \( n \), \( P(A) \approx 1 - e^{-1} \) when sampling either with or without replacement.

There are several ways to simulate this problem physically. You could write consecutive integers from 1 to \( n \) on \( n \) balls or on \( n \) slips of paper, place them in a container, and select a sample of size \( n \), sampling either with or without replacement. Show that the proportion of trials on which at least one match occurs is approximately \( 1 - 1/e \) when \( n \) is sufficiently large.

Another way to simulate sampling with replacement is by rolling an \( n \) sided die \( n \) times, checking whether face \( k \) is observed on roll \( k \), and counting the number of trials on which at least one match occurs.

Sampling without replacement can be simulated by shuffling each of two identical decks of \( n \) cards and then comparing them to see whether the cards in position \( k \) match. Again calculate the proportion of shuffles that lead to at least one match.

Physically simulating this experiment in class provides an excellent opportunity to get the class involved. It also helps them to really understand the problem. However, to obtain a sufficient number of repetitions, a computer program could be written to simulate this problem. Care must be taken when simulating sampling without replacement.

The following table gives the results of several repetitions of this experiment so that you can also see the randomness in the estimates. Each of these repetitions was based on 200 trials with \( n = 6 \).

<table>
<thead>
<tr>
<th></th>
<th>Estimates Based on 200 Repetitions</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling With Replacement</td>
<td>0.705 0.685 0.620 0.650 0.660 0.670</td>
<td>0.6651</td>
</tr>
<tr>
<td>Sampling Without Replacement</td>
<td>0.645 0.580 0.625 0.655 0.675 0.640</td>
<td>0.6319</td>
</tr>
</tbody>
</table>
Using simulation, or calculating probabilities using the formulas, the students can decide how greatly the probability is affected by whether the sampling is done with or without replacement. The students can also determine the effect of \( n \) on the probability and decide when "\( n \) is sufficiently large" for the probability of at least one match to equal, approximately, \( 1 - 1/e = 1 - 1/2.71828 \ldots = 0.6321 \). For a more complete discussion, see [9].

3. A modification of the "birthday problem" can be stated as follows. 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

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

Find the value of \( \mu = E(X) \) for several values of \( M \).

Note that the mean of \( X \) is given by

\[
\mu = E(X) = \sum_{x=2}^{M+1} xf(x)
\]

so that it is not easy to solve for \( \mu \). However, by using a short computer program, the value of \( \mu \) can be found easily for different values of \( M \). For example, suppose that you have a large number of people in a room and you ask each person to give their birthday. On the average, how many people do you have to ask before a match occurs? It is interesting to show that \( \mu = 24.6166 \) when \( M = 365 \).

Because 6-sided dice are readily available, this problem could be simulated by letting \( M = 6 \) and rolling such a die. The students could compare probabilities and relative frequencies for \( x = 2, 3, 4, 5, 6, 7 \). In addition, the sample mean could be compared with the theoretical mean \( \mu = 3.775 \). See [8] for a more complete solution.
4. In New Zealand a coin has a Kiwi on one side and Queen Elizabeth II on the other side. Flip such a coin successively.

(a) Let \( X \) equal the number of flips that are required to observe the same face on consecutive flips. Show that the mean and variance of \( X \) are 3 and 2, respectively.

(b) Let \( Y \) equal the number of flips that are required to observe Kiwis on consecutive flips. Show that the mean and variance of \( Y \) are 6 and 22, respectively.

This problem is perhaps too advanced theoretically for some students. However, a physical simulation is very easy. For part (a), simply count the number of flips of a standard coin that are required to observe the same face, heads-heads or tails-tails, on consecutive flips. For part (b), count the number of flips required to observe heads-heads, for example, on consecutive flips. For each experiment, calculate the sample means and sample variances.

For part (a), the p.d.f. of the random variable \( X \) is

\[
f(x) = \left( \frac{1}{2} \right)^{x-1}, \quad x = 2, 3, 4, \ldots
\]

If you know that the mean and variance of a geometric random variable \( W \) with p.d.f.

\[
h(w) = \left( \frac{1}{2} \right)^{w-1} \left( \frac{1}{2} \right), \quad w = 1, 2, 3, \ldots
\]

are \( \mu_W = 1/(1/2) = 2 \) and \( \sigma^2_W = (1/2)/(1/2)^2 = 2 \), it follows that \( \mu_X = 3 \) and \( \sigma^2_X = 2 \).

For (b), let \( f_n \) equal the \( n^{th} \) Fibonacci number where \( f_1 = 1, f_2 = 1, \) and \( f_n = f_{n-1} + f_{n-2} \) for \( n = 3, 4, 5, \ldots \). Then the p.d.f. of \( Y \) is

\[
g(y) = \frac{f_{y-1}}{2^y}, \quad y = 2, 3, 4, \ldots
\]

It is possible to find the mean and variance of \( Y \) using sums of infinite series. However, a computer program can also be written to calculate the sums for you, and thus show that the mean and variance are 6 and 22, respectively. Extensions of this problem are given in [2].
5. Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from a normal distribution with mean $\mu$ and variance $\sigma^2$.

(a) Show that an unbiased estimator of $\sigma^2$ is

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

(b) Show that an unbiased estimator of $\sigma$ is $cS$ where

$$c = \frac{\Gamma \left( \frac{n-1}{2} \right) \sqrt{n-1}}{\Gamma \left( \frac{3}{2} \right) \sqrt{2}}.$$ 

To illustrate this result, it is necessary for you to be able to simulate observations from a normal distribution. There are several ways to do this. We will look at three of them.

- This first method is perhaps the easiest and most intriguing. If $U$ is a uniform random variable on the interval $(0,1)$, that is, the value of $U$ is a “random number,” then

$$Z = \frac{U^{0.135} - (1-U)^{0.135}}{0.1975}$$

is approximately normal with mean 0 and variance 1, i.e, $N(0,1)$.

- If $U$ has a uniform distribution on the interval $(0,1)$, then the mean and variance of $U$ are $1/2$ and $1/12$, respectively. Applying the Central Limit Theorem, the sum of 12 observations of $U$ has a distribution that is approximately normal with mean $12(1/2) = 6$ and variance $12(1/12) = 1$. Thus, the sum of 12 random numbers minus 6 gives the value of an approximate standard normal random variable. That is,

$$Z = \sum_{i=1}^{12} RND - 6,$$

where each of the 12 RND’s represents a different “random number” selected from the interval $(0,1)$, has an approximate $N(0,1)$ distribution.
• The Box-Muller method gives an exact theoretical method for obtaining observations of standard normal random variables. Let $U$ and $V$ be independent uniform random variables on the interval $(0,1)$. Let

$$X = \sqrt{(-2)(\ln U) \cos(2\pi V)},$$

and

$$Y = \sqrt{(-2)(\ln U) \sin(2\pi V)}.$$  

Then $X$ and $Y$ have independent $N(0, 1)$ distributions.

To go from a standard normal random variable $Z$ to a normal random variable $X$ with mean $\mu$ and standard deviation $\sigma$, let

$$X = \sigma Z + \mu.$$  

It is now possible to write a computer program that would, for example, simulate $n = 5$ observations from a normal distribution with mean $\mu = 75$ and standard deviation $\sigma = 20$. If 100 samples of size 5 were simulated, and for each of them the sample mean ($\bar{X}$), sample variance ($s^2$), and $cs$ were calculated, it should be true that the average of the $\bar{X}$’s is close to 75, the average of the $s^2$’s is close to 400, and the average of the $cs$’s is close to 20 where

$$c = \frac{\Gamma(\frac{5-1}{2})}{\sqrt{2}} \frac{\sqrt{5-1}}{\Gamma(\frac{5}{2})} = \frac{2}{(3/2)(1/2)\sqrt{\pi\sqrt{2}}} = \frac{8}{3\sqrt{2\pi}}.$$  

The simulation also illustrates that the variance of the sample variance is very large. In fact, since we know that the distribution of $(n - 1)S^2/\sigma^2$ is chi-square with $n - 1$ degrees of freedom,

$$\text{Var}(S^2) = \text{Var} \left( \frac{\sigma^2}{n-1} \left( \frac{(n-1)S^2}{\sigma^2} \right) \right) = \left( \frac{\sigma^2}{n-1} \right)^2 2(n-1).$$

So for our example,

$$\text{Var}(S^2) = \left( \frac{400}{4} \right)^2 (2)(4) = 80000.$$  

The following table shows the output for just 8 repetitions of this experiment.
|
|---|---|---|---|---|---|---|---|
| Simulation Results |  |  |  |  |  |  |
|   | 69.33 | 71.96 | 83.34 | 76.91 | 71.90 | 79.75 | 68.21 | 79.82 |
| $s^2$ | 177.69 | 477.72 | 1062.98 | 249.19 | 679.15 | 188.23 | 815.41 | 570.11 |
| $cs$ | 14.18 | 23.25 | 34.69 | 16.79 | 27.72 | 14.60 | 30.38 | 25.40 |
|  |  |  |  |  |  |  |  | 75.15 |
|  |  |  |  |  |  |  |  | 527.56 |
|  |  |  |  |  |  |  |  | 23.38 |

See [10] for a BASIC program and additional sample output.

For those of you who are not familiar with the gamma function, all that is needed for this example is that, when $n$ is an integer,

$$\Gamma(n) = (n - 1)!$$

and when $n$ is odd, say $n = 2k + 1$,

$$\Gamma\left(\frac{n}{2}\right) = \Gamma\left(\frac{2k + 1}{2}\right) = \frac{2k - 1}{2} \frac{2k - 3}{2} \ldots \frac{1}{2} \sqrt{\pi}.$$ 

6. Let $X_1, X_2, \ldots, X_5$ be a random sample of size $n = 5$ from a normal distribution with mean $\mu$ and variance $\sigma^2$. The endpoints for a 90% confidence interval for $\mu$ are

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{5}}$$

when $\sigma$ is known and

$$\bar{x} \pm 2.132 \frac{s}{\sqrt{5}}$$

when $\sigma$ is unknown. Note that the critical $t$ value replaces the critical $z$ value and $s$ replaces $\sigma$. Which of these intervals is shorter?

First, it should be made clear that both intervals are 90% confidence intervals. And it is not too difficult to empirically find the answer to the question. The simulation depends on your ability to sample from a normal distribution. A method for doing this is given in the solution for the last example.

When the standard deviation is known, the length of the $z$ confidence interval for $\mu$ is

$$\text{length} = 2(1.645)\sigma/\sqrt{5} = 1.471 \sigma.$$ 

When the standard deviation is unknown, the length of the $t$ confidence interval will vary from sample to sample, some being quite short and others very long. Simulation will illustrate
this. So in a sense, the answer to the question is that sometimes the $z$ interval is shorter and sometimes the $t$ interval is shorter. However, on the average, the expected length of the $t$ interval is longer. Using the result from the last example, we have that

$$E(\text{length}) = 2(2.132) \frac{\Gamma\left(\frac{5}{2}\right)\sqrt{2}}{\Gamma\left(\frac{3.5}{2}\right)\sqrt{5} - 1} \frac{\sigma}{\sqrt{5}} = 1.792\sigma.$$  

A graphical comparison of the two types of confidence intervals is extremely helpful. The following graphical display from [10] illustrates this with the $t$-intervals on the left and the corresponding $z$-intervals on the right.

![Graphical display showing confidence intervals](image)

**Implementation**

Examples like those that have been described can be used in a variety of ways. One of the most effective ways is to offer an optional computer based laboratory along with a probability and statistics class. In this laboratory, each of the students would be expected to write a computer program to solve each problem. It is helpful in this setting if programs for the graphical output are provided for the students. The computer programs and the results of the students' simulations should be shared with all of the students in the class. Students often come up with ingenious ways for solving a problem. They should also be encouraged to modify the questions that are asked and to raise new questions.
If it is not possible to add a laboratory, a few students could solve the problems as an independent study project. It would then be possible for the student or the professor to present the solution to the class, getting the students in the class involved as much as possible in the solution of the problem, perhaps through physical simulation.

And, of course, the professor could solve a problem alone and present its solution to a class.

Much of the learning comes about by writing the computer program, rather than just watching the output from the program. Thus, the effectiveness of this approach increases as the student involvement increases.

Conclusions

Hopefully these examples give you some new ideas about ways in which the computer can be used. Additional suggestions and examples are given in other sections of this volume.

A computer disk for an IBM (compatible) computer is available that contains computer solutions for the listed problems as well as several others. Some of these solutions contain graphical output. If you would like to receive a copy of this disk for $10 to cover the cost, specify whether you would like a 5\(1/4\)” or a 3\(1/2\)” disk, and send your request to the author.

References


