STA 113 Fall 2003
I. H. Dinwoodie

October 7 Notes on Assignment 2: Monte Carlo Method

1. Assume for illustration that the lifetimes in problem 1 are exponential with mean 10 hours. Then the number of failures \( N \) in \([0, 60]\) is a random variable with the Poisson distribution. The number of parts required is \( 1 + \lambda \), assuming a part that fails at time 60 must be replaced. So the expected cost is \( E(10(1 + \lambda)) = 10(1 + E(\lambda)) \) and \( E(\lambda) = 60/10 \).

2. The Monte Carlo method simulates the failure process a number of times \( n \) independently, say \( n = 100 \), and computes a value of \( \lambda \) on each, say \( \lambda_1, \lambda_2, \ldots, \lambda_{100} \). Then the law of large numbers says

\[
E(\lambda) \approx \frac{1}{100}(\lambda_1 + \cdots + \lambda_{100}).
\]

In Matlab, the following procedure implements the Monte Carlo approach. It simulates 100 histories in 100 columns of 20 failures, and looks at the number of failures in \([0, 60]\) in each column. One has to be sure that 20 failures is enough to take us beyond 60 hours. It should be enough, because on average it should take us out to \( 10 \times 20 = 200 \) hours. Below, \( t \) is a matrix with the independent lifetimes in 100 columns, \( s \) is their cumulative sum in each column, corresponding to the clock time of each failure, \( f \) flags whether the values in \( s \) are no more than 60, and \( N \) counts the number of values in \( s \) in each column that are no more than 60. In other words, \( N \) is the number of failures in \([0, 60]\).

\[
\begin{align*}
t &= \text{exprnd}(10, 20, 100) \\
s &= \text{cumsum}(t) \\
f &= (s <= 60) \\
N &= \text{sum}(f) \\
\text{mean}(N)
\end{align*}
\]

The sample mean of the simulated counts is our Monte Carlo “point estimate” of \( E(\lambda) \). How close is it really to \( E(\lambda) \)? Do a histogram of the counts \( N_1, \ldots, N_{100} \) to see if the counts are approximately bell-shaped in distribution, then go ahead and use the confidence interval on p. 292. It says roughly that the true value is within \( 2 \times s / \sqrt{n} \) with 95% chance, where \( s \approx \sigma = \sqrt{\lambda} = 6 \).
This is not very accurate with a sample size of \( n = 100 \). To get the sample mean \( \hat{N} = \text{mean}(n) \) within .05 of the (unknown) true value \( E(N) \) (with very high certainty), we can use the ideas of sample size design on p. 293. If \( 2 \times \sigma/\sqrt{n} \leq .05 \) we better make \( n \geq 10,000 \). So we should have used

\[
\texttt{t=exprnd(10,20,10000);}
\]

Use 10000 columns in your experiments. It is important to use the semicolon “;” at the end!

To get the corresponding costs, it is easiest to

\[
\begin{align*}
\texttt{>> C}&=10*(1+N); \\
\texttt{>> mean(C)} \\
\texttt{>> [a,b,ci]=ttest(C)}
\end{align*}
\]

In your answers, report the confidence interval of p. 292, which is easily computed in Matlab as the value \( \text{ci} \) above from the command \texttt{ttest}.

At the end, one sees that although the two strategies look superficially similar, the exact costs are different.

The math computation is much harder with any distribution other than the exponential for survival times, because only the exponential gives a simple distribution for the number \( N \) of failures up to a certain time. But the Monte Carlo approach does not get more difficult, so it is perfect for part 2.

3. To simulate from the Gamma distribution, use \texttt{gamrnd}. The meaning of the parameters is not clear from Matlab help. The Matlab website says that the parameters \( A, B \) match up with \( \alpha, \beta \) in our book on p. 175 (you have to be careful because reciprocals sometimes are used). So for a mean 10 Gamma distribution with \( \alpha = 2 \) use

\[
\texttt{gamrnd(2,5,20,10000);}
\]

4. In your answers, please put the expected costs and the confidence intervals after a title page, and include the Matlab diary and the histograms of the counts to justify the confidence interval at the end as an appendix. The answers for 2 will vary slightly from person to person because the answer is random, but all answers should be within the range suggested by the confidence interval for a sample size of \( n = 10000 \).