Final for EE 345 – Modeling and Simulation Stevens Institute of Technology Spring 2003	Name:
Pledge:	

April 28, 2003

The final is open book/open notes. Total value is 30 points. All questions are equally weighted. **Do any 10 questions. Do more than 10 for extra credit**. Some questions could be answered in more than one way. Only one answer is required, but extra credit will be given for identifying and explaining alternative answers. Some questions ask for N responses, extra credit will be given for more than N responses.

1. The inner product (or dot product) operator, •, is defined on linear vector spaces (e.g., a 3-dimensional coordinate systems) as:

   $$V_1 = (x_1, y_1, z_1)$$
   $$V_2 = (x_2, y_2, z_2)$$
   $$V_1 \bullet V_2 = x_1 \cdot x_2 + y_1 \cdot y_2 + z_1 \cdot z_2$$

   The inner product is defined for an arbitrary number of dimensions and provides a distance measure between two vectors in an N-dimensional space as:

   $$V_1 = (v_{1_1}, v_{1_2}, ..., v_{1_N})$$
   $$V_2 = (v_{2_1}, v_{2_2}, ..., v_{2_N})$$
   $$V_1 \bullet V_2 = \sum_{i=1}^{N=\dim(V_1)=\dim(V_2)} v_{1_i} \cdot v_{2i}$$

   (Note: while it is easiest to visualize a vector space built on unit length direction vectors, the definition works just as well with functions as the basis of the vector space, extending the summations to integrals. E.g., the Fourier series can be viewed as the inner product of a nearly arbitrary function and harmonically related sinusoids as basis vectors.)

   Apply this distance measure concept to random variables, as discussed in this class, to associate a physical meaning to some of the mathematical definitions we have covered.

   [1A1, 1A2, 1A3]
   For a discrete random variable, the covariance and correlation functions relate directly to the inner product. In the case of a zero-mean random variable,

   $$\mathrm{cov}(X,Y) = \tfrac{1}{N-1}\sum_{i}^{N} X_i Y_i = \tfrac{1}{N-1}(X \bullet Y)$$

   The correlation is equivalently the dot product, normalized by the variance of the random variables.
   For a continuous random variable, the functional vector space definition of dot product again directly corresponds to correlation and covariance definitions.
   In effect, when one calculates the covariance or correlation of two random processes, the *distance* between the two processes is being calculated.

2. You observe the interarrival times listed in the table below at a popular snack bar on the Snevets University campus. You have the option of modeling the distribution as exponential, normal or uniform. Describe the steps you would use to determine which would be the appropriate distribution to use. Extra credit: find the proper distribution

2.037	3.728	3.161	0.531	0.262	8.685	3.813	9.678	2.303	6.073
4.142	1.092	6.209	8.097	1.881	5.360	6.352	6.382	2.040	1.095
1.907	8.259	3.511	4.359	1.577	9.314	1.926	4.326	4.566	2.163
8.748	5.197	7.812	1.505	6.630	4.359	9.458	5.938	1.014	6.806
8.583	2.810	7.982	5.800	1.174	7.399	8.360	9.782	2.357	7.558

[3B2]

The distribution was actually generated from a uniform distribution with a mean of 5 minutes and a range of 0-10.

(1) generate histogram
(2) determine the mean and variance
(3) Generate 3 quantile-quantile plots, one for each the three types of distributions, with corresponding means and variances,
(4) Compare the q-q plots to find the one the looks most linear.
(5) Compute χ^2 statistic to compare to candidate distributions

This is the Mathcad worksheet used to generate the data and to form the q-q plots.

Note: the Mathcad worksheet recalculated the random numbers when it was reopened. The data listed does not correspond to the final, but the Q-Q plots do.

Snevets Snack Bar

Generate a set of random variates with a uniform distribution

$N := 50 \quad \mu := 5 \quad v := 5$

$R := \mathrm{runif}(N, \mu - v, \mu + v)$

$i := 0..9$

$D1_i := R_i \quad D2_i := R_{i+10} \quad D3_i := R_{i+20} \quad D4_i := R_{i+30} \quad D5_i := R_{i+40}$

$D1^T =$

	0	1	2	3	4	5	6	7	8	9
0	0.013	1.933	5.85	3.503	8.228	1.741	7.105	3.04	0.914	1.473

$D2^T =$

	0	1	2	3	4	5	6	7	8	9
0	9.885	1.191	0.089	5.317	6.018	1.662	4.508	0.571	7.833	5.199

$D3^T =$

	0	1	2	3	4	5	6	7	8	9
0	8.76	9.559	5.393	4.621	8.622	7.797	9.968	6.115	2.662	8.401

$D4^T =$

	0	1	2	3	4	5	6	7	8	9
0	3.759	6.772	0.088	2.759	5.879	8.376	4.849	7.437	4.58	7.444

$D5^T =$

	0	1	2	3	4	5	6	7	8	9
0	5.99	7.35	5.724	1.516	4.252	5.171	7.515	1.69	4.919	6.998

$S := \mathrm{sort}(R)$

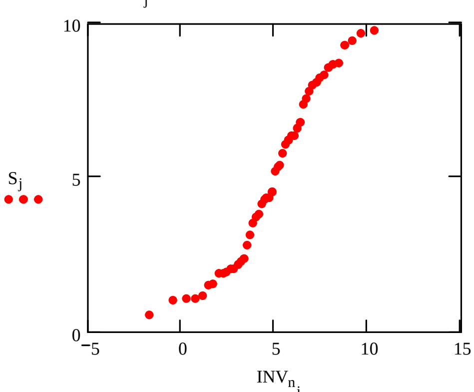
$\mathrm{mean}(R) = 4.871$

$\mathrm{var}(R) = 8.166$

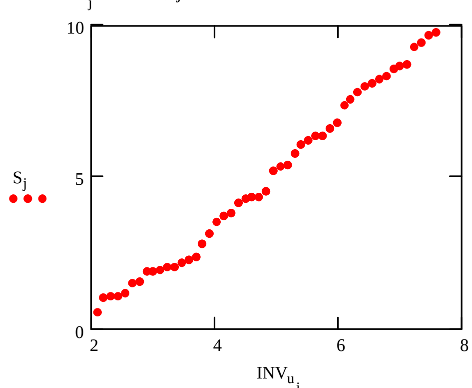
$j := 1..N$

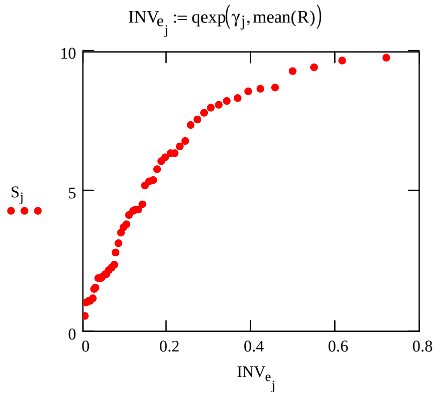
$$\gamma_j := \frac{j - \frac{1}{2}}{N}$$

$INV_{n_j} := \mathrm{qnorm}\left(\gamma_j, \mathrm{mean}(R), \sqrt{\mathrm{var}(R)}\right)$



$INV_{u_j} := \mathrm{qunif}\left(\gamma_j, \mathrm{mean}(R) - \sqrt{\mathrm{var}(R)}, \mathrm{mean}(R) + \sqrt{\mathrm{var}(R)}\right)$





Here, only the q-q plot with a uniform distribution looks nearly linear, so this would suggest that this is the best fit to the data.

3. You are designing an extremely complex system simulation that is stressing your computer's ability to generate random numbers fast enough. You have access to a hardware accelerator board for your machine that can implement any arbitrary digital logic function at very high speeds. You have decided to use a linear feedback shift register to generate numbers between 1 and 2^N-1, but you have been asked to verify if the numbers are actually usable as random numbers. Define at least three tests that you might perform to validate the usability of the numbers generated and how they would be used. More than 3 for extra credit.

[1A3, 3B2]

(1) if feasible, check the cycle length of the random number generator

(2) perform run tests on the output:
   a. runs of 1-0's,
   b. runs above and below mean

(3) measure the statistics of the random numbers and compare to uniform
   a. mean
   b. variance

(6) perform a q-q plot against a uniform distribution, verify linearity

(7) compute autocorrelation of data to find any unexpected patterns

(8) gap test – look for spacing between repeated digits

(9) frequency test

(10) play the signal into a wave device and listen for non-noiselike behavior, equivalently, you could create a two dimensional plot of the random data and look for visible patterns, e.g., with an x-axis corresponding to a subset of the data (first half, least significant bits, even data values, etc.) and the y-axis corresponding to the remainder of the data, plot a point corresponding to each pair of elements from the two subsets.

(11) use tools like Matlab to display the spectrum of the random numbers, looking for any unexpected nonuniformity in the spectrum

(12) perform a Turing test on the random data, asking experts to see if they can discern any difference between the pseudo-random sequence and truly random data.

4. Loam Depot, a local home center specializing in selling a designer dirt, has hired you to streamline their checkout process. They believe that their best, highest spending customers are experiencing inordinately long service times and are considering adding a separate checkout for these preferred customers. The table below lists the purchase amounts and service times for a set of 15 customers. Calculate the correlation between these data to verify or refute their belief.

Customer #	Purchase	Service time (minutes)	Customer #	Purchase	Service time (minutes)
1	$101	2.2	9	$137	4.3
2	$177	4.4	10	$159	7.2

3	$334	2.1	11	$495	7.6
4	$240	6.9	12	$148	8.1
5	$429	5.6	13	$104	5.3
6	$170	7.3	14	$313	3.5
7	$384	8.5	15	$341	7.6
8	$222	5.1			

[1A2, 1A3]

Here's the computation of the correlation between the series.

First, here is how the series was created:

Here is how the data was generated:

$N := 15$

Generate two sets of 15 random numbers between 0 and 1

$X_1 := \mathrm{runif}(N, 0, 1)$

$X_2 := \mathrm{runif}(N, 0, 1)$

Transform the first set of numbers between 100 and 500

$R_1 := X_1 \cdot 400 + 100$

Generate a second, correlated set of numbers

$\beta := 0.3$

$$R_2 := \frac{\beta \cdot (R_1 - 100) + X_2 \cdot 700}{100} + 1$$

$R_1 := \mathrm{floor}(R_1 + .5)$

$$R_2 := \frac{\mathrm{floor}(10R_2 + .5)}{10}$$

$R_1 =$

	0
0	101
1	177
2	334
3	240
4	429
5	170
6	384
7	222
8	137
9	159
10	495
11	148
12	104
13	313
14	341

$R_2 =$

	0
0	2.2
1	4.4
2	2.1
3	6.9
4	5.6
5	7.3
6	8.5
7	5.1
8	4.3
9	7.2
10	7.6
11	8.1
12	5.3
13	3.5
14	7.6

Here we calculate the correlation:

Calculate the correlation:

$$E(X) := \frac{\sum_{i=0}^{length(X)-1} X_i}{length(X)}$$

$\mu_1 := E(R_1)$ $\mu_1 = 250.267$

$\mu_2 := E(R_2)$ $\mu_2 = 5.713$

$$cov(X,Y) := \sum_{i=0}^{length(X)-1} \frac{(X_i - \mu_1)\cdot(Y_i - \mu_1)}{N}$$

$\sigma_1 := \sqrt{var(R_1)}$ $\sigma_2 := \sqrt{var(R_2)}$ $\sigma_1 = 120.253$ $\sigma_2 = 2.022$

$$correlation(X,Y) := \frac{cov(X,Y)}{\sqrt{var(X)}\cdot\sqrt{var(Y)}}$$

$correlation(R_1, R_2) = 0.242$

The result is a somewhat correlated set of values (.242). The claim is supported. It is not necessary to have correlation of 1 to say that two sets of numbers are correlation. It would be necessary to do an economic analysis to decide if this level of correlation is sufficient to justify a decision to have a high volume shopper checkout. Correlation levels near zero or negative would refute the conjecture.

5. It is hypothesized that the arrival times of students to the EE345 class are exponentially distributed. Use the χ^2-test with 6 class intervals and a 10% significance level to verify or refute this hypothesis, based on the arrival times listed in the table below.

1.372	25.313	14.722	24.705	6.895	18.663	18.106	17.213	5.488	12.937
9.174	49.332	16.838	3.352	20.504	24.789	59.959	5.250	23.780	12.782
12.525	10.511	14.046	6.059	6.606	16.262	20.879	33.903	16.500	9.859
10.062	3.391	15.738	14.395	5.959	0.859	51.341	1.871	35.150	3.750

[1A2, 3B2]

The data in the table above is generated with an exponential distribution as follows:

$n := 40$ $\lambda := .07$

$X := rexp(n, \lambda)$

$i := 0..9$ $D1_i := X_i$ $D2_i := X_{i+10}$ $D3_i := X_{i+20}$ $D4_i := X_{i+30}$

$D1^T =$

	0	1	2	3	4	5	6	7	8	9
0	1.372	25.313	14.722	24.705	6.895	18.663	18.106	17.213	5.488	12.937

$D2^T =$

	0	1	2	3	4	5	6	7	8	9
0	9.174	49.332	16.838	3.352	20.504	24.789	59.959	5.25	23.78	12.782

$D3^T =$

	0	1	2	3	4	5	6	7	8	9
0	12.525	10.511	14.046	6.059	6.606	16.262	20.879	33.903	16.5	9.859

$D4^T =$

	0	1	2	3	4	5	6	7	8	9
0	10.062	3.391	15.738	14.395	5.959	0.859	51.341	1.871	35.15	3.75

The χ^2 statistic is computed as follows:

$X_s := \text{sort}(X)$

$\lambda_{hat} := \frac{1}{\text{mean}(X)}$ $\quad \lambda_{hat} = 0.061$

$k := 6 \quad p := \frac{1}{k}$ $\quad p = 0.167$

$i := 0..k-1$ $\quad a_i := \frac{-1}{\lambda_{hat}} \cdot \ln(1 - i \cdot p)$

$a_k := \infty$ $\quad E_i := p \cdot n$

$O := \text{hist}(a, X_s)$ $\quad \text{term}_i := \frac{(O_i - E_i)^2}{E_i}$

$$a = \begin{pmatrix} 0 \\ 3.012 \\ 6.699 \\ 11.452 \\ 18.15 \\ 29.602 \\ 1 \times 10^{307} \end{pmatrix} \quad O = \begin{pmatrix} 3 \\ 8 \\ 5 \\ 12 \\ 7 \\ 5 \end{pmatrix} \quad E = \begin{pmatrix} 6.667 \\ 6.667 \\ 6.667 \\ 6.667 \\ 6.667 \\ 6.667 \end{pmatrix} \quad \text{term} = \begin{pmatrix} 2.017 \\ 0.267 \\ 0.417 \\ 4.267 \\ 0.017 \\ 0.417 \end{pmatrix}$$

$$\chi_{sq} := \sum_{i=0}^{k-1} \frac{(O_i - E_i)^2}{E_i} \qquad \chi_{sq} = 7.4$$

With 6 class intervals and the mean arrival rate estimated from the data, there are 4 degrees of freedom. The number of degrees of freedom depend on the number of classes, and not the number of samples for this test. At a 10% significance level, the χ^2 value from table A.6 is 7.78. The calculated value is less than this value, so the hypothesis that that data is exponentially distributed is not rejected.

Note: The hypotheses are:
H0: the data is exponentially distributed
H1: the data is not exponentially distributed.
Statistical tests can reject H0 or H1, but you can never prove the hypothesis or Accept it.

6. Why is the estimate of the mean value of an output process of a simulation more likely to be biased if it is averaged in time rather than averaged across individual runs of the simulation?

[1A3]
From one time interval to the next within any given run, it is likely that the state of the simulation is highly correlated. For instance, in a queuing system, the number of customers in the queue at the end of one time interval is likely to be very close to the number of customers in the queue at the start of the next time interval, since there will generally be a limited number of arrivals or departures.
On the other hand, the state of one instance (run) of a simulation is quite likely to be independent of the state of another instance (run) because a large number of run-to-run independent events led to the current state.
Averaging in time tends to capture the effects of the time correlation of the state, leading to a biased estimate, while averaging across runs tends to be averaging independent, and therefore inherently unbiased, views of the parameter.

7. In order to justify an increase in highway tolls for the 3-lane highway, the Garden State Parking authority has estimated a traffic volume of 75,000 cars per hour during the busy hour at the Podunk toll exit. Is this estimate reasonable on its face? State your assumptions and justify your conclusion.

[4A3]

Assuming that all the cars on the 3 lane highway exit the highway at the maximum legal speed of 65 mph, are, on average, 15 feet long and are driving bumper to bumper, the absolute theoretical maximum number of cars that could exit the highway would be:

(3 lanes)*(1 car/15 feet/lane)*(65 mph)*(5280 feet/mile) = 68640 cars/hour

obviously, cars per unit length, speed, and percentage exiting must reduce this number, so it is impossible for the estimate to be correct and unlikely for it to be very close.

Note that the question did not ask if the suggestion to raise tolls was justifiable or not, just whether the estimate made any sense.

8. A linear congruential random number generator with parameters X_0=23, a=9, c=0, m=63 is used to generate random numbers for a simulation. How many up and down runs does this generator create for a sequence of random numbers $X_1,\ldots,X_{20}$?

[1A2]

$X_0 := 23$

$c := 0$ $\qquad$ $S := 0$

$m := 63$

$a := 9$

$i := 1..20$

$X_i := \text{mod}(a \cdot X_{i-1} + c, m)$

$i := 0..9$

$D1_i := X_i$

$D2_i := X_{i+10}$

$D1^T =$

	0	1	2	3	4	5	6	7	8	9
0	23	18	36	9	18	36	9	18	36	9

$D2^T =$

	0	1	2	3	4	5	6	7	8	9
0	18	36	9	18	36	9	18	36	9	18

```
upruns(r) := | for  i ∈ 0.. length(r) − 2
             |    x_i ← if(r_{i+1} > r_i, 1, 0)
             | k ← 1
             | ups ← if(x_0 = 1, 1, 0)
             | while (k < length(x))
             |    | ups ← if[(x_{k−1} = 0) ∧ (x_k = 1), ups + 1, ups]
             |    | k ← k + 1
             | return  ups
```

$\text{downruns}(x) := \text{upruns}(-x)$

$k := 1..20$ $\qquad$ $XX_{k-1} := X_k$

$\text{upruns}(XX) = 7$

$\text{downruns}(XX) = 6$

9. Loam Depot has decided to *branch* out and start selling trees that customers may plant in the dirt that they buy, but they realize that they will need a number of flat-bed hand trucks to allow the customers to carry out their trees. Assuming an exponential customer arrival rate of 20 customers per hour, and assuming that the service time (the time that customers are driving the carts out to their cars or trucks) is uniformly distributed from 5-15 minutes, how many hand trucks will be needed to satisfy customers 95% of the time without waiting?

[1A2, 1A3]

This is an $M/G/\infty$ queue – the arrivals are exponential, the service time is uniform (a generalized distribution) and the number of servers is assumed infinite – self service customers where we are trying to determine the number of servers for a given QoS.

$\mu := \frac{1}{10}$ since the service time is uniformly distributed from 5-15, the mean is 10 minutes

$\lambda := \frac{1}{3}$ 20 customers per hour -> 1/3 customer per minute

$$P(n) := \frac{e^{\frac{-\lambda}{\mu}} \cdot \left(\frac{\lambda}{\mu}\right)^n}{n!}$$ the probability that there are n customers in the system

Find the average number of customers in the system. This is a lower bound for the number of carts/servers, c

$L := \frac{\lambda}{\mu}$ $L = 3.333$

The probability that the steady state number of customers in system is less than or equal to c is

$$P_{L.LE.c}(c) := \sum_{n=0}^{c} P(n)$$

Find the smallest c that causes the sum to exceed the QoS = .95

$P_{L.LE.c}(1) = 0.155$

$P_{L.LE.c}(2) = 0.353$

$P_{L.LE.c}(3) = 0.573$

$P_{L.LE.c}(4) = 0.756$

$P_{L.LE.c}(5) = 0.879$

$P_{L.LE.c}(6) = 0.947$

$P_{L.LE.c}(7) = 0.979$

I listed each value in order, rather than just the answer. If you were calculating these individually, you could sum over n until you crossed the threshold.

Here the number of carts needed is 7

10. The Third National Bank of Podunk has 6 tellers with an exponentially distributed service rate with a mean of 5 minutes. Customers arrive at the bank with exponentially distributed interarrival times with an average interarrival time of 1 minute. There is a single queue for customers to wait in, but the queue is restricted to 15 customers. What is the probability that a customer will arrive at the bank and find that the queue is full? In the interests of customer service, these overflow customers are redirected to the Able-Baker donut shop with a coupon for a free donut. What observations can you make about the arrival process at the donut shop?

[1A3]

The bank queue is an M/M/c/N queue. First, we must calculate the probability of zero queue length, P_0.

$c := 6 \quad \lambda := 1 \quad \mu := \frac{1}{5} \quad a := \frac{\lambda}{\mu} \quad N := 15$

$\rho := \frac{\lambda}{c \cdot \mu} \quad \rho = 0.833$

$$P_0 := \left(1 + \sum_{n=1}^{c} \frac{a^n}{n!} + \frac{a^c}{c!} \cdot \sum_{n=c+1}^{N} \rho^{n-c}\right)^{-1} \qquad P_0 = 4.985 \times 10^{-3}$$

Using P_0, we can find P_N for N=15.

$$P_N := \frac{a^N}{c! \cdot c^{N-c}} \cdot P_0 \qquad P_N = 0.021$$

About 2% of the customers arriving at the bank will find the queue length to be the maximum allowable. While the arrivals at the bank are exponentially distributed, since the arrivals at the donut shop are generated from the overflow condition, there is no mathematically tractable solution for the arrivals at the donut shop. If the bank queue happens to be less than 15 customers, none will be redirected to the donut shop, so there may be long intervals with no arrivals. On the other hand, if a customer arrives to find the queue full, the conditional probability that the next customer will also find the queue full is more than 2%. Thus, arrivals at the donut shop from the bank will generally be bursty and highly correlated, since they are only generated when the primary queue is at capacity. This is analogous to the traffic on overflow routes in a telephone switching network. It is very difficult to model this traffic. It is an interesting exercise to write a simulation of the bank queue overflows and examine the donut shop arrivals to see this bursty behavior.

11. You have written an event driven simulation for a single server system and are in the process of trying to debug the program which does not seem to be working correctly. You have printed out traces of the program execution and are examining a program generated simulation table. What are some of the things going wrong with this simulation?

Clock	Event	Interarrival time	Service time	FEL
1	Arrival(A)	2	4	(A,3) (D,5)
3	A	5	3	(D,5) (D,6)
4	A	12	12	(D,5) (D,6) (A,16) (D,16)
5	Departure(D)			(D,6) (A,16) (D,16)
6	D			(A,16) (D,16)
10	A	4	1	(D,11) (A,14) (A,16) (D,16)
11	A	7	5	(D,11) (A,14) (A,16) (D,16) (D,16) (A,18)
14	D			(A,14) (A,16) (D,16) (D,16) (A,18)
16	A	3	5	(A,16) (D,16) (D,16) (A,18)
18	A	1	2	(A,19) (D,20)
…	Simulation			

	continues			

[3B3]
This is, by design, a very buggy event table:
(1) Service times are getting scheduled at arrival time. How can you do this when you don't know the future status of the server? It makes more sense to schedule departure times at service completion (except for the first customer)
(2) The event at t=4 was not scheduled.
(3) At t=3, there was an arrival scheduled 5 units later, but no arrival at t=8
(4) The arrival at t=10 was not scheduled
(5) There was a departure scheduled for t=11. Where did it go?
(6) At t=11, there are multiple scheduled departures for t=16. How can this be with a single server system?
(7) At t=18+, there have been 7 arrivals and 3 departures. That means there should be 4 customers in the system, but at t=18, there were only one future departure event scheduled. This is because:
(8) At t=16, there was an arrival, but no corresponding departure scheduled
(9) The same happened at t=18

12. In a paper presented at a recent IEEE conference (L. Cimini, K. Leung, B. McNair, J. Winters,"Outdoor IEEE 802.11b Cellular Networks: MAC Protocol Design and Performance," *Proc. ICC 2002*, New York, NY, April 2002, if you're interested), we derived a model for the performance of the 802.11b CSMA wireless network in an outdoor environment using a distributed coordination function. We modeled the average duration (in μsec) from the first packet to the last in a busy period, $\overline{Y}$, as:

$$\overline{Y} = a - \frac{1-e^{-aG}}{aG}$$

where a is the average propagation delay between stations in μsec, and *G* is the total traffic average load in packets/μsec. Is this model valid on its face? Why or why not?

[1A2, 4A2, 4A3]
Despite the generous comments of some students, one should not always assume correctness of a model based on the person presenting it. Particularly, if that person is a devious professor who intentionally misstates a model for the purposes of creating a test question…

This model is not valid on its face.

(1) The units of $\overline{Y}$ must be time, but $\frac{1-e^{-aG}}{aG}$ is a dimensionless quantity, added to *a*, stated in units of time.
(2) As the load, *G*, goes to zero, this expression grows without bound since $1-e^{-aG}$ approaches 1. One would expect the expression to be strictly bounded and most likely to decrease with zero traffic load.
(3) As the load, *G*, increases without bound, this expression goes to zero. One would expect the delays to increase with increasing load.

ABET criteria
•1A1 - Recognize mathematical parameters as if they were physical variables and vice-versa
•1A2 - Be able to follow general mathematical concepts of derivation of engineering or scientific result and possess the mathematical skills to link those concepts
•1A3 - Be able to understand the relevance of the mathematical results to physical applications
~~**•1A4 - Be able to articulate algorithmic thinking through flow charts**~~
~~**•3B1 - Use software for preparing, transmitting, and displaying multimedia documents**~~
•3B2 - Have the ability to use computational tools for finding graphical, numerical, statistical, and analytic solutions to problems
•3B3 - Have the ability to use systems simulations appropriate to engineering practice

•4A2 - Be able to identify input, output, and operating variables as appropriate in various units
•4A3 - Be able to identify technical relationships between the input, output and variables and use the relationships to predict mutual changes.