AN EMERGENCY MEDICAL SYSTEM SIMULATION MODEL

James A. Fitzsimmons
Graduate School of Business
University of Texas at Austin
Austin, Texas 78712

Abstract

This paper presents a SIMSCRIPT simulation model designed for general use by health systems planners in evaluating existing or proposed emergency medical systems. An extensive validation of the model was performed using actual data from the San Fernando Valley area of Los Angeles. Appropriate operating procedures for use of the model are discussed. In conclusion, an application of the model in determining the desired number of ambulances for a system illustrates its usefulness in planning emergency medical systems.

Introduction

An emergency medical system will be defined by that sequence of events beginning with incident occurrence and ending with the transfer of the patient to an emergency medical facility where definitive medical care is available. A good case can be made for exempting the emergency facility since a large proportion of the emergency facility load is exogenous to the emergency ambulance service (i.e. people presenting themselves by walking or driving in). However, the number and location of emergency facilities is related to the performance of the emergency ambulance service system. Figure 1 describes the activities of the patient in such a system. Note the definition of response time, a system performance criterion peculiar to service systems with mobile servers. In emergency medical systems, response time is the elapsed time from notification of an emergency until an ambulance arrives at the scene. This important criterion of system performance has considerable political and psychological implications, in addition to being a component of the time to treatment for the patient.

As Figure 1 suggests, an emergency medical system can be conceptualized as a single queue, multi-server queuing system with mobile servers under central control. The distribution of incident arrival rates does fit the commonly assumed Poisson distribution. However, the mean rate of arrivals is time dependent with the peak of activity between 5 and 6 p.m. The distribution of service times is clearly not exponential. In fact, it is a function of the particular ambulances busy at the time a call is received. In addition, the service time distribution is not the same for every server except for the special case when all the ambulances are located at the same station.

These characteristics of a mobile server substantially complicate any analytical effort to model such a queuing system because the location of the servers and spatial distribution of incident locations influence the travel time which is a significant component of the service time. Because of the complex nature of the emergency medical system and desire to provide a general methodology for studying alternative dispatch, retrieval and deployment policies, computer simulation was selected as the most appropriate modeling technique.

Computer Simulation Model

The simulation model is written in SIMSCRIPT for at least the following reasons: (1) dynamic allocation of core storage; (2) list processing features; (3) generality and flexibility of the language; and (4) view of the world.

The model consists of two programs; an incident generator which creates input data and the main simulator which simulates the behavior of the system and creates a summary report of system performance at the completion of the simulation run.

A. Incident Generator

The incident generator program creates an exogenous events tape of all incident occurrences in chronological sequence together with the necessary descriptive information such as location and type of injury for each incident. The incident load is intended to reflect the actual cyclical pattern by the use of Poisson distributed hourly arrival rates which are a function of the time of day. The exact time of each incident occurrence is a random event

This research conducted at the University of California, Los Angeles, was funded by the U. S. Department of Transportation (Contract FH-11-6849).

determined by selecting the time within the hour from a uniform distribution.

B. Main Simulator

The simulation program will be described by explaining how the typical sequence of events from incident notification to the return of the ambulance to its station is simulated.

Notification

The exogenous input to the model is a simulated series of incident occurrences as they are reported to a public agency.

Vehicle Availability

After an incident has been reported and before the emergency medical system can respond, it must check the current availability of its rescue vehicles. The model has provided each vehicle with an attribute called STATS. The vehicle at any point in time must be in one of the following states:

1. idle at its station;
2. enroute to an incident;
3. at the scene of an incident or transporting patients to a hospital;
4. returning empty from the scene of an incident or the hospital;
5. cruising (helicopter);
6. out of service.

With this information, a set of immediately available or soon to be available vehicles can be identified for selection of that vehicle to respond to the incident.

If the vehicle happens to be enroute to a hospital or returning to its station, then a special subroutine called FIND is used to locate its current X and Y coordinates. A vehicle in transit is assumed to be on a line between its point of departure and destination at a distance from point of departure proportional to the time since departure.

Vehicle Selection and Dispatch

At this point, a decision is made to select from the set of available ambulances that vehicle to dispatch to the scene of the incident based on some "dispatch rule." The common dispatch rule in use is to send that ambulance, whether idle or transporting a patient, which will arrive at the scene first.

Arrival at Scene

The arrival time of an ambulance at the scene of an incident is scheduled in simulation time by calculating the travel time from the current location of the ambulance to the incident location and adding this travel delay to the time of dispatch.

Travel time is calculated using subroutine TRAVL which assumes the distance traveled between two points as being the sum of the rectangular displacements. This model is based on the typical orthogonal arrangement of streets, particularly in urban areas.

On Scene Care

Once the ambulance arrives at the scene, a further delay is incurred to account for first aid assistance and possible extrication. While at the scene of an incident a decision may be made to call for assistance which would initiate additional vehicle dispatches. The amount of time spent at the scene of an incident is a random variable. The Log Normal distribution was found to fit the distribution of empirical on-scene care time.

Very often injured persons may not need hospital care and thus are released at the scene after receiving first aid. The probability of an injured person being transported has been found to be a function of incident type.

Retrieval to the Hospital

For those cases that require transportation of the injured person to a medical facility, the question of which hospital to select arises. The common retrieval rule in use is to select the closest hospital. Again, the time of ambulance arrival at the hospital is scheduled to occur in simulated time after a calculated travel time, determined by subroutine TRAVL.

Ambulance Return to Station

Before the ambulance is free to return to its station, a short delay is incurred at the hospital to transfer the patient, make a report to the doctor, and process records. This delay is a random variable with a uniform distribution. If an incident is waiting in the dispatch queue at this time, the ambulance will be dispatched immediately upon leaving the hospital to the scene of the new incident. This is also true if the ambulance were assigned an incident while enroute to the hospital. However, if there are no outstanding incidents, then a deployment question arises. The deployment rule commonly used is to return the ambulance to its home station.

One final comment should be made concerning a special exogenous event routine called CHANGE. This routine is a special restart feature not provided with SIMSCRIPT which permits a series of simulation runs to be performed in succession without resubmitting the job. Very often a series of simulation experiments are desired in which a particular parameter is successively changed (e.g.

the number of ambulances in the system). Under the normal SIMSCRIPT operating procedure a separate run is required for each experiment with the associated cost and inconvenience. When a series of simulation runs are desired, exogenous event CHANGE is scheduled to occur at the end of each run after the last ambulance has returned to its station. The CHANGE routine is programmed to read in the new parameter changes, the random generator root is reset and TIME is set back to zero. Control is then returned to the event timing routine to begin the next simulation run.

Model Validation

To validate any kind of model, it is necessary to prove that the model represents a true abstraction of reality. It is not enough that a model gives results observable in the real world. The model should actually behave like the real world, particularly in its dynamic operation in time; otherwise one cannot be certain that results obtained under slightly different conditions will be useful.

An approach called "Multi-Stage Verification," suggested in an article by Naylor and Finger [3] is used to validate the simulation model. This method of verification consists of three stages: (1) formulation of a set of postulates describing the behavior of the system under study; (2) an attempt to verify these postulates using statistical tests; and finally (3) testing the model's ability to predict the behavior of the system.

A. First Stage: Postulate Formulation.

During the model development phase, efforts are made to derive the necessary and sufficient conditions for a complete description of system behavior. Determination of significant variables and estimation of associated parameters are conducted using statistical techniques where possible or professional guidance when required. Data gathered to estimate time on scene exemplifies the former; discussion with experts identifying the commonly used dispatch and retrieval policies typify the latter. The functional relationships are best discovered through actual field observations such as riding ambulances on emergency calls.

B. Second Stage: Postulate Verification

The internal consistency of the simulation model is verified for both the one ambulance system as well as a special multiple ambulance system for which analytical results are available.

Single Ambulance System

One ambulance with a speed of 36 m.p.h. is stationed at the hospital located at the center of a square service area with six mile sides. The mean incident occurrence rate is 18 incidents per day (.75 per hour) with a Poisson distribution. The location of the incidents in the service area is uniformly distributed. All incidents are transported and a FCFS queue discipline is used. On-scene care time is a constant 7 minutes and hospital transfer time is a constant 3 minutes. This special case is in fact an M/G/1 queuing system* for which steady state formulas are available. Using the following notation, formulas for this specific ambulance system were derived in [2] and are used here to calculate results for comparison with simulation estimates in Table 1.

a = mean incident occurrence rate per hour

C = Mean time at scene in minutes

r = mean distance in miles to the scene if incident location pattern is uniform and ambulance is at center of service area

T = mean transfer delay at hospital in minutes

v = average ambulance speed in m.p.h.

Mean Service Time:

$$1/s = 2r/v + C + T \qquad (1)$$

$$= 2\ (3/36)\ 60 + 7 + 3 = 20 \text{ minutes}$$

Traffic Intensity:

$$t = a/s \qquad (2)$$

$$= 20\ (.75)\ /60 = .25$$

Service Time Variance:

$$s^2 = 2\ (r/v)^2\ /3 \qquad (3)$$

$$= 2[(3/36)\ 60]^2\ /3$$

$$= 50/3 \text{ hours}^2$$

Mean Number of Incidents Waiting:

$$Lq = (a^2s^2 + t^2)\ /[2(1-t)] \qquad (4)$$

$$= [(.75)^2\ (50/3) + (1/4)^2]/[2(3/4)]$$

$$= 25/576$$

*A standardized shorthand is used in much of the queuing literature for identifying simple queuing systems. In the symbol A/B/C, C is the number of servers, while A and B indicate the arrival and service distributions, respectively. The common symbols used to identify distributions are: G, general independent; M, Poisson arrival rate or exponential service times; Ek (Erlangen), interarrival or service times distributed as a gamma distribution of order k; and D (deterministic), a schedule of arrivals or constant service times.

Mean Response Time:

$$R = Lq/a + r/v \qquad (5)$$

$$= (25/576) \ / \ (1/80) + (3/36)\ 60$$

$$= 8.5 \text{ min.}$$

Mean Time to Hospital:

$$H = Lq/a + 2r/v + C + T \qquad (6)$$

$$= 3.5 + 10 + 7 + 3$$

$$= 23.5 \text{ min.}$$

Multiple Ambulance System

Nine identical ambulances with a speed of 36 m.p.h. are stationed at the hospital located at the center of a square service area with six mile sides. The mean incident occurrence rate is 108 incidents per day (4.5 per hour) with a Poisson distribution. The location of the incidents in the service area is uniformly distributed. All incidents are transported and a FCFS queue discipline is used. On-scene care time is a constant 7 minutes and hospital transfer time is a constant 3 minutes.

Although only 9 ambulances were used, no waiting occurred in the system, and thus it was essentially an M/G/∞ system. Parzen [5] derives the following result for the probability of number of busy servers in such a system, a Poisson distribution with mean equal to the traffic intensity:

$$P(i) = \frac{t^i e^{-t}}{i!}$$

Where t = traffic intensity (7)

i = number of busy servers

Using Equation (7), with "t" equal to 1.5 in this case, the analytical frequency of number of busy servers in the system is compared to the simulation findings in Table 2. The mean system statistics are compared in Table 3.

A "Chi Square Goodness of Fit Test" is performed to statistically test the similarity of the simulation and analytical distributions. From the data in Table 2, Chi Square is calculated to be 4.082. However, Chi Square at the .05 significance level is equal to 11.1 with 5 degrees of freedom. Thus the hypothesis that the simulated number of busy ambulances were obtained from sampling from a Poisson population with mean of 1.5 can not be rejected. Therefore, the internal consistency of the simulation model is verified for both the one ambulance system as well as the multiple ambulance system when there is no waiting in the system.

C. Final Stage: Test of Predictive Ability

The model is developed using Fiscal Year 1967 data from the Fire Department Emergency Rescue Ambulance Companies located in the San Fernando Valley area of Los Angeles. A historical verification is performed by comparing the actual records of system performance for FY 1967 and the results generated by the model for the same time period. To statistically verify the predictive ability of the model a "Chi Square Goodness of Fit Test" is used to test the degree of similarity of the actual and generated response time distributions. The distribution of response times is an important system parameter for verification because it is affected by all the factors in the model.

The distribution of response times for the San Fernando Valley Fire Department Ambulance System is compared to the simulation results in Figure 2. It is clear for practical purposes that the simulated distribution of response times agrees with the actual data. Furthermore, using the Chi-Square Goodness of Fit Test, the hypothesis that the two samples (actual and simulated) came from the same population cannot be rejected.

Operating Procedures

A. Mean Value Estimation

The data generated in the simulation model should exhibit time correlation because the process is time-dependent. When positive time correlation is present in sample data, the use of classical statistical estimating techniques (which assume independent observations) will lead to under estimation of sampling variances and erroneous confidence intervals. In an effort to provide a valid estimate of the sampling variance, the total number of incidents in an experimental run were gathered consecutively into 30 subsample groups. The choice of 30 subsamples is predicated on the fact that a sample of 30 independent observations can be considered normally distributed. The variance of the grand mean $\bar{R}$ of this sample as discussed by Conway [1] and Nelson [4] is given by:

$$\text{Var}(\bar{R}) = g^2/n^2 \left[n + 2 \sum_{s=1}^{n-1} (n-s)\ A(s)\right] \qquad (8)$$

Where:

R_i is the mean response for subsample i

m is the number of observations in a subsample

n is the number of subsamples

g^2 is the variance of the subsample means

A(s) is the coefficient of autocorrelation between subsample means separated by (s-1) subsamples in the sequence of observations.

$$g^2 = \sum_{i=1}^{n} (R_i - \bar{R})^2 / n \qquad (9)$$

$$A(1) = \left[\sum_{n=1}^{n-1} (R_i \cdot R_{i+1}) \right] /(n-1)-\bar{R}^2] \; /g^2 \qquad (10)$$

$A(s) = A(1)^s$ assuming the geometric approximation model;

Otherwise:

$$A(s) = [E(R_i \cdot R_{i+s})-\bar{R}^2] \; /g^2 \qquad (11)$$

$$E(R_i \cdot R_{i+s}) = \left[\sum_{i=1}^{n-s} R_i \cdot R_{i+s} \right]/(n-s) \qquad (12)$$

The problem of determining the length of the simulation run then becomes one of selecting a subsample size of sufficient magnitude to provide the desired degree of accuracy in the mean value estimate (e.g., to provide a ± 5% confidence interval about the estimate of the mean with a 1% chance of error). Some experimentation is required to arrive at this value for "m" because the coefficient of autocorrelation is a peculiarity of both the system and its utilization. For the single ambulance system a subsample of 360 observations is necessary to meet the desired ± 5% confidence interval about the estimate of mean response. However, for the multiple ambulance system with a relatively smaller incident load, a subsample size of 108 observations provided an approximately ± 2% confidence interval about the estimate of the mean response.

From this experience the more conservative run length of 10,800 incidents is selected for all future experimentation. It is unlikely that the incident load on a single ambulance system will be greater than that used in the verification run because such cases would result in an unrealistic mean waiting time.

B. Sampling Procedure

The first problem associated with sampling is when to begin. Stochastic systems have an inherent warmup period or transient period during which time the behavior of the system is significantly influenced by the starting condition of the system. Traditionally, data collection is postponed until after some initialization period, thus allowing the system to reach a randomly chosen state which is not significantly influenced by an arbitrary starting condition. Because the incident load exhibits a daily cycle reaching a period of essentially no activity between the hours of 4 a.m. to 6 a.m., a one day initialization period was considered sufficient.

Finally a close replication sampling procedure is followed because of the comparative nature of the analysis. The principal purpose of system studies is the determination of the relative effect on system performance of varying parameter values and operating policies. To achieve this, the same incident load is replicated for all experiments, thus eliminating sampling variation from the input to the simulator.

C. Collection of Simulation Statistics

Incident Statistics

For each incident, the response time, waiting time, and time to hospital are collected. Then the following system statistics are calculated for each of these times:

- Mean
- Standard Deviation
- Distribution for selected fractile values
- Minimum and maximum
- Means by districts (census tract)
- Number in system probability distribution

Ambulance Statistics

The following statistics are collected for each ambulance:

- Total number of calls
- Number of calls taken by STATS at time of call
- Number of calls taken by district (census tract)
- Mean wait time, response time, time to hospital
- Mean utilization
- Mileage driven
- Mean time to scene and mean retrieval time

Hospital Statistics

For each hospital, the following statistic is collected:

- Total number of deliveries

Determining the Size of an Ambulance Fleet

An abstract ambulance system will be used to illustrate how the simulation model can aid a systems planner in determining the appropriate number of ambulances for adequate service.

A. Description of Abstract Model Used in Study

Although 24 hour and lessor cycles appear in the San Fernando Valley data, the arrival process for the abstract model is a time independent Poisson process with mean arrival rate of 30 calls per day. The distribution of call locations is considered uniform over the entire ambulance service area. The service area is square in shape with sides of 60,000 feet in length. A single emergency hospital is located at the center of the service area. The dispatch delay, of negligible value in reality, is considered to be zero. The

on scene care time, which has been found to have a log normal distribution in the San Fernando Valley, is treated as a constant with a mean value of 7.0 minutes. The transfer delay at the hospital, a uniformly distributed variate, is considered a constant of 3.0 minutes, also the mean value for the San Fernando Valley. The ambulances are surface vehicles with a mean speed of 3,333 feet per minute. The ambulance closest to the incident is dispatched and all incidents are transported to the hospital. Two ambulance deployment strategies are considered. In one case, all the ambulances are stationed at the hospital; in the other case, ambulances are dispersed throughout the service area. The two stochastic features retained in the model are the time between successive emergency calls and the corresponding location of the incident in the service area.

B. Experimental Results and Implications

Figure 3 shows the mean response time and waiting time for incidents in this abstract system as a function of the number of ambulances in service. As can be seen, the question of determining the desired number of ambulances for a system cannot be separated from considering their deployment. For the single station deployment strategy, the system in terms of the mean response criterion becomes saturated with ambulances with fleets of 4 or more. However, with dispersed deployment continued improvements in mean response can be achieved with increased fleet size. Perhaps the common practice of stationing ambulances at hospitals should be reconsidered in this light.

Concluding Remarks

The simulation model was developed as a general planning aid for evaluation of existing or proposed emergency medical systems. For this reason, the model is modular in design with the dispatching, retrieval and deployment sections contained in subroutines for ease in tailoring the model to fit any desired configuration. Perhaps the most valuable studies will be those that evaluate new configurations of medical care delivery systems using the latest technology such as helicopters, mobile intensive care units and paramedics.

References

[1] Conway, R. W., "Some Tactical Problems in Digital Simluation," Management Science, Vol. 10, No. 1, pp. 47-61 (October 1963).

[2] Fitzsimmons, James A., "Emergency Medical System: A Simulation Study and Computerized Method for Deployment of Ambulances," Ph.D. Dissertation, Graduate School of Business Administration, University of California, Los Angeles, (1970).

[3] Naylor, T.H. and J.M. Finger, "Verification of Computer Models," Management Science, Vol. 14, No. 2, pp. B-92 through B-101 (October 1967).

[4] Nelson, Rosser T., "Queuing Network Experiments with Varying Arrival and Service Processes," Naval Research Logistics Quarterly, Vol. 13, No. 3 (September 1966)

[5] Parzen, Emanuel, Stochastic Processes, Holden-Day, pp. 144-148 (1962).

[6] Savas, E. S., "Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service," Management Science, Vol. 15, No. 12, pp. B-608 through B-627 (August 1969).

Biography

James A. Fitzsimmons is Assistant Professor of Management at the University of Texas at Austin. He received a B.S.E. in industrial engineering from the University of Michigan, an M.B.A. from Western Michigan University and a Ph.D. in operations management from the University of California, Los Angeles. His present research interests include the application of management science to public systems analysis and techniques in computer simulation. He is a registered professional engineer in the State of Michigan, a member of the American Institute of Industrial Engineers and the Institute of Management Science.

TABLE I

SINGLE AMBULANCE VERIFICATION

Statistic	Analytical Result	Simulation Result Estimate	Simulation Result 99% Interval
Traffic Intensity (t)	.25	.252	-
Mean Wait (W)	3.50	3.546	3.089-4.003
Mean Response (R)	8.50	8.513	7.920-9.105
Mean Time to Hospital (H)	23.50	23.477	21.022-25.933

TABLE II

CHI SQUARE TEST OF NUMBER OF BUSY SERVERS

Number of Busy Servers	Poisson Dist t = 1.5	Expected Frequency	Simulated Frequency	Difference
0	.223	722	720	-2
1	.335	1085	1046	-39
2	.251	813	848	35
3	.125	405	402	-3
4	.047	152	159	7
5	.015	37	36	1
6 or more	.004	13	16	3

TABLE III

MULTIPLE AMBULANCE VERIFICATION

Statistic	Analytical Result	Simulation Result Estimate	Simulation Result 99% Interval
Mean Wait (W)	.00	.00	-
Mean Response (R)	5.00	4.99	4.90-5.08
Mean Time to Hospital (H)	20.00	19.98	19.80-20.16

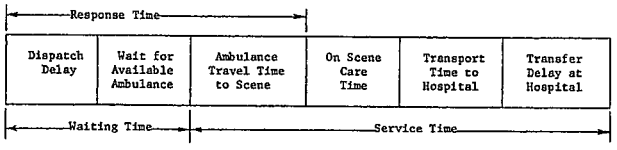

FIGURE 1 SEQUENCE OF EVENTS IN PATIENT SERVICE

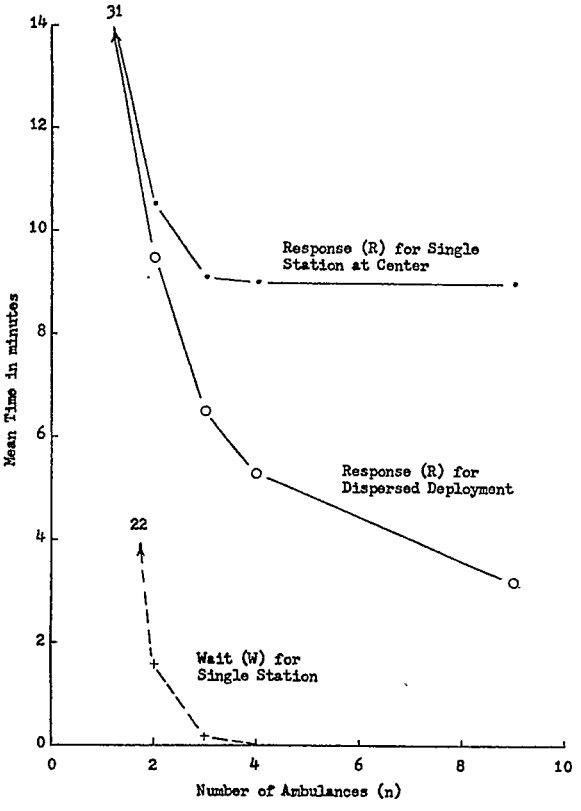

FIGURE 3 COMPARSION OF SINGLE AMBULANCE STATION AND DISPERSED LOCATIONS FOR UNIFORM DISTRIBUTION

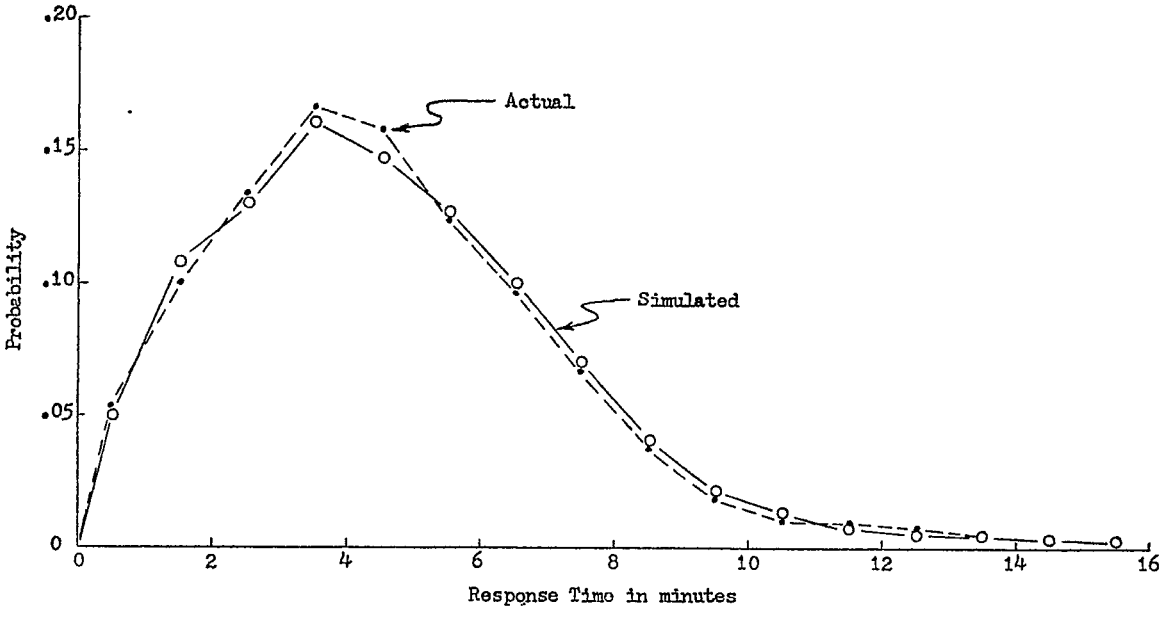

FIGURE 2 DISTRIBUTION OF RESPONSE TIMES FOR SFV AMBULANCE SYSTEM