Health Economists Study Group, January 1999.

Ambulance Response Times

Alastair Fischer, Paul O'Halloran and Peter Littlejohns

Health Care Evaluation Unit, Public Health Sciences Department, St George's Hospital Medical School, London SW17 0RE

The government has set a standard that 75% of emergency calls to the ambulance service be responded to within 8 minutes. Using data from the Surrey Ambulance Service, we estimate the mean response time when there are n ambulances available and ready for use, where n can range from 0 to 35. From this, we estimate the reduction in response time for the addition of one ambulance. From there, we estimate the marginal cost of one second in both mean response time and 75^{th} percentile response time. This provides a powerful tool for resource allocation, because now an innovation in ambulance usage can be compared with the spending of additional resources.

Ambulance Response Times

Introduction

The demand for ambulance services in Britain has been increasing at around 7% per year during the 1990s. Such a rate of increase is not sustainable in the long term. This has prompted the authorities to seek innovative ways to improve the service, to avoid either the cost increases or the deterioration in service that would otherwise accompany the increase in demand. The purpose of this paper is to provide the necessary statistical background and an economic framework, within which the likely effect of innovations can be judged.

Improvements in automated data capture in British ambulance services over the past few years now make it possible to analyse their performance in meeting time targets.
A national target for meeting a response time for life-threatening calls of 8 minutes or less in 75% of cases has been set for the fiscal year 2000/1. ("Response time" refers to the time elapsed from when the ambulance control room answers an incoming "999" call to when the ambulance arrives at its destination.)

We have used data from the 1997-8 records of calls to the Surrey Ambulance Service (SAS) to analyse response times over that period. Over 10 months from April 1, 1997 to January 31, 1998, 75,239 calls were attended by a service whose full complement of ambulances was 37.

The parameters obtained from the analysis performed are not only useful as measures of ambulance performance, but may also be used as inputs into simulation models which attempt to site ambulances to minimise response times.

When ambulances in one area are being fully utilised, response times will increase, as ambulances from further afield must be diverted to serve that area. After incorporating the effect of traffic, we find that the average response time declines by about 9.7 seconds for each additional ambulance on call but not in use. This figure is shown to be of great

importance for the planning of ambulance services. We use it to predict the effect on response times of reducing the total all-round trip time, increasing the time in lay-bys, introducing prioritisation of ambulance calls and estimating the effect of demand increases. It is planned to extend the analysis to two or three levels to incorporate time and location effects.

Background

Surrey is an English county that borders the southern edge of London. Its area is 641 square miles and its population in 1998 was 1.1 million people, a density of 1,700 per square mile. The southern part of the county is more rural, although Gatwick airport is on the southern boundary.

The SAS, whose headquarters are in Banstead, an outer suburb of London, operates up to 37 ambulances at a time from 19 bases. During most days, up to 32 ambulances are used during the day, dropping to 19 in the early hours of the morning, and averaging 26. As the occasion demands, crews operate over the county border, and similarly, crews from other counties operate in Surrey. Of the unusual features of Surrey with respect to the ambulance service, up to 11 ambulances attend Gatwick Airport at times of full alert. Besides this, Heathrow Airport is just over the northern border of Surrey, and in emergency situations may also call ambulances from Surrey. The M25 London beltway motorway runs across the northern part of the county. It is both an opportunity for quick transportation and a potential hazard for multiple vehicle accidents.

When the ambulances at a particular base are all in use, ambulances not in use in adjacent areas will be sent to "cover" for that area. This may cause a domino effect, of ambulances further afield covering for the ambulance which is providing the primary cover. (When all the ambulances except one are in use, that ambulance is sent to the most central point (transportation-wise) in the county, the junction of the M25 motorway and the A3 arterial road.) Ambulances on cover may travel all the way to another base, whose ambulances are already in use, but may also stand in a lay-by, part of the way to the base.

Even where no ambulance from a base is in use, an ambulance from that base may be required to stand in a lay-by as part of a strategy of covering the county more evenly, rather than at a base. In this way, it can reach some of the population more quickly than otherwise.

Of the 75,000 calls, 59,000 of them were co-called "emergency" calls, and 16,000 "urgent" calls. The urgent calls comprise mainly inter-hospital transportation of patients and calls by GPs to transport patients to hospital appointments. Emergency calls receive priority over urgent calls, which are generally answered within two hours. At the time that the data refer to, emergency calls were not prioritised. However, a prioritisation scheme is in the process of being introduced, and is likely to be in place by the time this is read.

Data

The data consisted of the log of all calls made by the ambulance service over the period of ten months beginning April 1, 1997. Each call is represented by a row of alpha and numeric columns. The data collected for each patient include the times the call was received, was activated, of arrival at the patient, of leaving for hospital (or somewhere else), of arrival at the hospital, of leaving hospital, of arrival at the station and of becoming ready for the next call. As well as this, the ambulance call-sign, the names of the patient and the hospital (coded), what was wrong with the patient and some other details were also recorded. From this, activation time, response time and all-round trip time can all be calculated.

Analysis

1. Preliminary analysis of response times

In most of what follows, when more than one ambulance has been called, the response time has been calculated for only the first ambulance called. This is because sometimes the second ambulance is called only after the first has arrived, and therefore with a

considerable lag, the time being measured from the first contact with the ambulance service.

In what follows, we examine *average* response times rather than the response time of the 75th percentile, which is the industry standard. There are four reasons for this: first, the mean contains more information than the 75th percentile and second, it is arguably more easily interpreted. The other two reasons have to do with technicalities: response times at the time the data were collected were expressed in whole minutes, so that, for example, both 8 minutes 0 seconds and 8 minutes 59 seconds were recorded as 8 minutes. Thus 75th percentiles were recorded to the nearest minute over a number of observations, whereas mean times over a number of observations could be estimated to the nearest second, with an error of only several seconds. Fourth, in the regression equations that follow, the regression of response time (TM) against number of available ambulances (n) is carried out for each of the 55,319 observations with a valid response time as data points, and not on the 35 mean response times for the 1 to 35 available ambulances. It is not possible to do this calculation for the 75th percentile on an individual call basis, and the regression would have to be carried out using only 35 data points.

Nevertheless, the answers for the mean response times have been translated into 75th percentile response times using the regression equation

$$75M = 0.064 + 1.171TM \qquad R^2 = 0.974 \qquad \text{.....(1)}$$
$$(0.331)\ (0.035) \qquad \text{(SDs in parentheses)}$$

where $75M$ is the 75th percentile response time. This therefore allows quite an accurate translation of mean response times to 75th percentile response times. The mean response time for the whole data set was 8.87 minutes = 8'52" (standard deviation = 0.99 seconds), and the 75th percentile for the set was calculated to be 10.44 minutes = 10'26" from equation (1), with a standard deviation of 3.01 seconds. Only 51.7% of all calls were answered in 8 minutes. If 75% of calls were to be answered in 8 minutes, this implies that the mean response time would have to have fallen to 6.78 minutes, that is, a fall of 2.09 minutes.

1.1 Time of day and day of week

Response times tend to be highest late at night and in the very early hours of the morning, mainly reflecting the smaller number of ambulances on duty at those times. The effect is more marked on Friday night/Saturday morning and on Saturday night/Sunday morning. This results in response times for the whole of Saturday and Sunday being higher than those for the rest of the week. These things are shown in Figures 1 to 4. (Subsequently, the SAS has redeployed its workforce and now makes more crews available overnight on weekends.) Figure 5 shows the frequency of response times.

Figures 1 to 4 here.
Also histogram of response time frequencies (Fig 5)

1.2 Location

The following mean response times were recorded for the 19 ambulance stations in Surrey, together with the number of calls undertaken by each station. The times are lower for more densely populated areas, as would be expected. See Table 1.

Table 1 about here

To minimise the mean response time for the whole of Surrey, it is not optimal for all stations to have the same mean response time. The reason for this is that if a sparsely populated area has a higher-than-average response time, relocating additional resources to it would lower the response time for the few calls that area would receive, but raise it for the many calls a more-densely populated area would receive. This would in most cases raise the overall average response time.

1.3 Ambulances away from home base or in lay-bys.

To facilitate ambulance dispatch, the SAS has divided Surrey into 367 zones. Calls answered from zones normally served from a particular base will in general have a lower response time than calls answered from elsewhere. Since there is no "home" base associated with any zone, we created a home base for this analysis from the data themselves, as follows. For each zone, we found the base which answered more calls than any other base, and the most-frequent base became that zone's home base. The mean response time for home-base calls was 7.79 minutes, and for non-home base calls was 10.86 minutes.

Response times for ambulances in lay-bys are subject to two different effects. The first we shall call the "in-cab" effect. Ambulance crews in lay-bys are already in their seats, ready to go. They can thus respond a good 35 to 40 seconds faster than crews at base, who need to leave their common room, walk to their vehicle, open the door, sit down, shut the door, put on seatbelts and turn on the ignition. The magnitude of the "in-cab" effect can be seen from the difference in activation times for ambulances at base compared with those at lay-bys, as given in Table 2. It averages 0.59 minutes (35 seconds) for ambulances going to home zones and 0.67 minutes (40 seconds) for those going to non-home zones.

The second effect is the location effect. This may go in either direction. If ambulance X is put on cover because most other ambulances are in use (called the "cover effect"), the mean response time can be expected to be high. However, if it has gone to a lay-by because another ambulance or ambulances are already at ambulance X's base, it could be expected that the mean response time would be relatively low, as the area to be covered by ambulance X would be smaller than average. (We call this the "spread effect").

The data do not readily allow the distinction between a covering ambulance and a spreading ambulance at a lay-by. However, the four mean response times of Table 2 give some indication of these effects:

Table 2 here

From Table 2, it may be seen that for journeys to home zones, ambulances in lay-bys have a shorter activation time. However, this is offset by longer travel time than those at the station, apparently reflecting the return to home zones of ambulances on cover, rather than the spreading of ambulances within "home" territory. For journeys to non-home bases, however, layby ambulances have an advantage in both activation time and travel time.

2. The effect of an additional ambulance.

The data were already in chronological order of receipt of call. At activation time of an ambulance, a new variable was created with the value of (-1) to represent its departure from the available fleet. When the ambulance returned and was once more available for use, a different new variable was given the value of (+1) to represent its addition to the available fleet. The return times were then put into a separate file and sorted into chronological order. The (+) and (-) files were then merged chronologically, and a running tally kept of the number of ambulances being added to and subtracted from the fleet of available ambulances. Similarly, a third file containing the numbers of ambulances beginning and ending their shifts (+ for beginning and – for ending) was merged with the file that had already been merged, to obtain the numbers of available ambulances at all times. In this way, the number of available ambulances (not in use at the time of the call) was associated with the response time of the ambulance, for all calls throughout the year.

Figure 6 gives the mean response time (TM^*) for each n (number of available ambulances not in use). Negative values of n represent the size of the queue at that time. There were very few observations with negative n, so the standard error for the mean response time for each negative value of n is quite high. However, for $n > 2$, the number of observations for each n rises from over one hundred to several thousand when n is between 20 and 30.

Figure 6 here

In part because of the large number of observations comprising TM^* for each n, the graph shows a remarkably steady decline in TM^* as n increases, from $n = 2$ to $n = 34$.

2.1 Statistical analysis

To describe the data more succinctly, several different functional forms were tried, and are listed below:

Linear regression:

$TM^* = 10.72 - 0.094n$ $\quad R^2 = 0.018$
(0.06) (0.003) (SDs in parentheses)

Piecewise linear regression:

$TM^* = 13.06 - 0.276n, \quad -3 < n < 11 \quad R^2 = 0.013$
(0.40) (0.046)

$TM^* = 10.23 - 0.072n, \quad n > 11 \quad R^2 = 0.009$
(0.07) (0.003)

Negative exponential (decay function):

$TM^* = 9.69e^{-0.0096n} \quad TM^* > 0 \quad R^2 = 0.013$
(0.07)(0.0003)

The weighted average decrease in mean response times, for an increase of one extra ambulance, *derived directly from the data*, is 0.096 minutes = 5.8 seconds. This may be compared with the coefficients obtained from the regression equations (transformed where necessary). This comparison is made in Table 3.

Table 3 here

2.2. The effect of traffic on response times.

Increased traffic density will lengthen mean response times. We therefore extend the regression model of response times against number of ambulances available, to include a dummy variable for each hour of the day, to represent traffic congestion. Further dummies for each day of school term (one at the start and the other at the finish of each school day) and to denote the hours of darkness each day of the year could also have been used, but so far this work has not been done.

The linear equation was estimated to be: .

$$TM^* = 11.49 - 0.162n + 0.18\text{hour02} - 0.01\text{hour03} - 0.02\text{hour04} - 0.10\text{hour05} - 0.23\text{hour06} - 0.003\text{hour07} - 0.19\text{hour08} + 0.83\text{hour09} + 0.43\text{hour10} + 0.28\text{hour11} + 0.47\text{hour12} + 0.70\text{hour13} + 0.38\text{hour14} + 0.57\text{hour15} + 0.71\text{hour16} + 1.02\text{hour17} + 1.22\text{hour18} + 1.37\text{hour19} + 0.86\text{hour20} - 0.17\text{hour21} - 0.45\text{hour22} - 0.40\text{hour23} + 0.14\text{hour24}$$

$$R^2 = 0.032$$

(0.10) (0.0045) (0.12 to 0.15 for the hour variables) (SDs in parentheses)

(The exponential equation gave similar results, and the piecewise linear equation has not been calculated.)

What is apparent is that the inclusion of hourly variables has altered the coefficient on *n* quite significantly, in the linear case from -0.100 to –0.162 minutes, or from -6 to –9.7 seconds. As it has turned out, this is the main reason for the inclusion of hourly variables in the analysis.

2.3 Economic analysis

From the analysis already undertaken, we assume that an additional ambulance reduces mean response time by 9.7 seconds. Given the data, this is the most conservative assumption we can make. An extra ambulance, run fully staffed every hour of the day for a year, would cost about £300,000 per year (about $US 500,000 at current exchange rates). That implies that a reduction of one second in mean response time would cost £31,000 ($52,000) if it were accomplished by an increase in resources. This number is of course scale-dependent: a service twice as big would have to pay something of the order of double this amount for a one-second reduction in mean response time. Further, as successive additional ambulances are added to the fleet, the reduction in time savings will itself decline. For example, for the tenth additional ambulance, we estimate a reduction in mean response time of about 6.5 seconds, which in turn implies a time cost of £46,000 per second.

These findings have implications for resource allocation, as innovations in the ambulance service can now be costed by comparing their time savings with those of saving the same time by increasing the number of ambulances.

3. Comparing the time savings of extra resources with those of innovations

3.1. Increasing the time in laybys

We have seen that ambulances in laybys have an activation time which is about 36 seconds shorter than that of those starting from base. Currently, (29.4%) of ambulance journeys start from laybys. If this could be increased to 60% (the extra 30.6% are assumed to be sitting in the ambulance at the ambulance station itself, and not being given the advantage of being "spread"), then the average response time would decline by 30.6% of 36 = 11.0 seconds. This time reduction could alternatively be achieved by adding 11.0/9.7 = 1.13 additional ambulances, at a cost of £340,000 per year. The number of additional hours in laybys would be of the order of 46,000 per year, or 92,000 crew hours. Thus, it would appear that, as an alternative to putting on extra ambulances, the ambulance service would be no worse off if it were to pay its crew up to £3 per hour extra, for each additional hour spent in laybys, over and above the time already spent there. The second benefit of being in a layby, in addition to the "in cab" effect, is the benefit obtained from spreading ambulances more evenly across the county. Since we have been unable to distinguish "covering" from "spreading", however, we cannot estimate the spreading effect of a 30.6 percentage point increase in layby time with any accuracy. Rough estimates suggest a small figure, since most of the multiple ambulances will already be covering.

3.2 Saving time on the round trip

From the data, the average time between calls is 2.4 hours. Since a round trip for a call takes on average 45 minutes, there are 99 minutes of slack time between calls. If the all-round trip time could be reduced by one minute to 44 minutes, slack time would increase by 1 minute to 100 minutes, or by 1%. But the slack time is also the time that an ambulance has available to answer any new calls, so this would increase by 1%. It would therefore be equivalent to having 1% more ambulances, or 0.26 ambulances in 26. If 0.26 of an extra ambulance were

available, we would cut about 2.5 seconds from average response time. Each minute reduction in average all-round trip time would be worth £78,000 per year in resources that would not need to be spent by the SAS. This is relatively small.

3.3 Prioritisation of calls

In an exercise carried out before prioritisation was undertaken, ambulance crew recorded whether the patient's problem was life-threatening or not, for the whole of the period under consideration. This was done after having seen the patient, so cannot readily be compared with the telephone prioritisation to be carried out in future in the control room. Only 9.5% of patients were categorised as "A" (life threatened), whereas it is expected that about a third of all calls will be so categorised by telephone, erring on the side of caution.

We asked ourselves the question: " What if the existing fleet answered *only* these 9.5% of calls, and left the other 90.5% unanswered forever?" We were able to see how many ambulances would have been available and not in use if this had happened. Suppose n ambulances were available when all calls were answered, and that n_A would have been available if only the A calls had been answered. We then assumed that the response time for an A call would be reduced by the difference between the average response time when n ambulances were available and the average response time when n_A would have been available. For example, suppose that an A call had taken 9 minutes when there were 20 ambulances available and not in use. The average response time when there were 20 ambulances available and not in use was 8'43". If the only calls answered were A calls, then there would have been, say, 30 ambulances available and not in use. The average response time when 30 ambulances were available was 8'09". The reduction in average response time in increasing the number of available, not-in-use ambulances from 20 to 30 is therefore 34 seconds, so we subtracted this amount from the 9 minutes. We repeated this analysis for all A calls.

We found that when only the A calls were answered, the average response time was reduced from 8'57" to 8'05", a difference of only 54 seconds. We estimate that the response time for the 75th percentile would reduce by 63 seconds, to 9'23". This does not go anywhere near meeting the target of responding to 75% of calls within 8 minutes.

If ambulances were to answer B calls (eventually) as well, the time savings would be substantially smaller than this. The analysis of prioritisation will be undertaken more fully in a separate paper.

4. Other comparisons that can be made using this model

4.1. Effect of dummy variables (traffic, darkness, and school holidays)

In section 2.2, it was found that the evening rush hour traffic increased response times, *ceteris paribus*, by up to one minute compared with the daily average. This implies that up to 6 extra ambulances would have to be employed at those times to keep response times the same as they would have been, in the absence of this traffic (*ceteris paribus*). However, there is little that can be done about traffic. The importance of this effect, as already stated, is that it has changed the slope coefficient of n, the number of ambulances available and not in use. So far, we have not investigated the variable "darkness" to see whether ambulances take longer to find patients when it is dark. It would be worth investigating, because if better street numbering could reduce night response times by even a second or two, a project of this kind may be worth undertaking.

4.2. Increase in demand

The model can also be used to predict the effect of an increase in demand for ambulance services. However, this is somewhat more difficult to analyse than the

other changes to the system already examined, because not only does ambulance usage increase, but the variance of its usage also increases. We therefore intend to report on this issue in more detail separately. Suffice to say that, by our calculations, an increase of 7% in usage would require a lower limit of 2.2% increase in the number of available ambulances and an upper limit of 4.6% increase. If no additional funds were available, this would increase mean response time by between 5.5 and 11.5 seconds.

5. Discussion

The model of ambulance usage developed here may help authorities decide how to allocate resources. It is able to give moderately accurate estimates of the cost of reducing average response times by providing additional ambulances. In providing a framework for comparing innovations against spending additional money to lower response times, it shows clearly that response time targets are unlikely to be met without the spending of additional resources.

We summarise the effects of various innovations in Table 4. In order to meet the target of 75% of calls in 8 minutes, recall that a decrease in mean response time of 2'05" is required.

Table 4 here

Solutions to the problem will therefore lie in either revising response time targets, reducing demand (although this is shown – on its own - not to have an effect great enough to meet the present target), and/or significantly changing the nature of the service to include motor cycle rapid first response. We have not tried to model this last type of response in this paper, as we have been looking at changes in the *status quo* situation. It is an open question as to whether radical changes can meet response time targets without an increase in resources, though the orders of magnitude presented here suggest that they are likely to fall somewhat short. The question should, however, be amenable to the framework devised here.

Table 1
Number of calls and mean response times for each ambulance station

Ambulance Station	Number of calls	Mean response time (minutes & seconds)
Chertsey	4788	9.26
Ashford	3500	8.41
Farnborough	6462	8.50
Knaphill	3189	9.33
Walton	3832	8.45
Esher Fire Station	1888	9.07
Godalming	2471	9.05
Cranleigh	1290	10.07
Farnham	4592	9.08
Guildford	4540	8.50
Haslemere	1551	10.01
Redhill	3691	8.25
Dorking	1790	9.11
Epsom	5027	8.11
Gatwick	4238	6.47
Leatherhead	2777	9.34
Warlingham	2069	8.54
Godstone	1633	9.59

Table 2

Category	Activation time	Response time	Travel time
1	2.44	7.79	5.33
2	1.85	7.79	5.94
3	2.61	11.49	8.88
4	1.94	10.22	8.28

Categories:

1. At home base, going to a home zone.
2. At lay-by, going to a home zone.
3. At home base, going to a non-home zone.
4. At lay-by, going to a non-home zone.

Table 3 :

Comparison of the slopes of models 1-3 at 5 ambulances and 25 ambulances available

	Change in response time (in seconds)		
		Model	
Number of available ambulances	1	2	3
5	-5.6	-16.6	-5.3
25	-5.6	-4.3	-4.4

Models :

Model 1. Simple linear regression model
Model 2. Piecewise linear regression model (-3 < ambulances <= 10, ambulances > 10)
Model 3. Negative exponential (decay function)

Table 4

Effect of a change in conditions on mean response time

Condition	Time savings (seconds)
One additional ambulance	10
Sit crew in ambulance: from current 30% to 40% 50% (etc)	 3.7 7.4 (plus an additional 3.7 for each 10%)
Spread ambulances when more than one at base	? (not much, because of limited opportunities to do this)
Prioritise: categorise 60% as A calls 30% " " " 10% " " "	28 49 63
Reduce round trip time of 45 minutes by one min	2.5

Figure 1

Mean response time during each hour of the day

Monday to Thursday

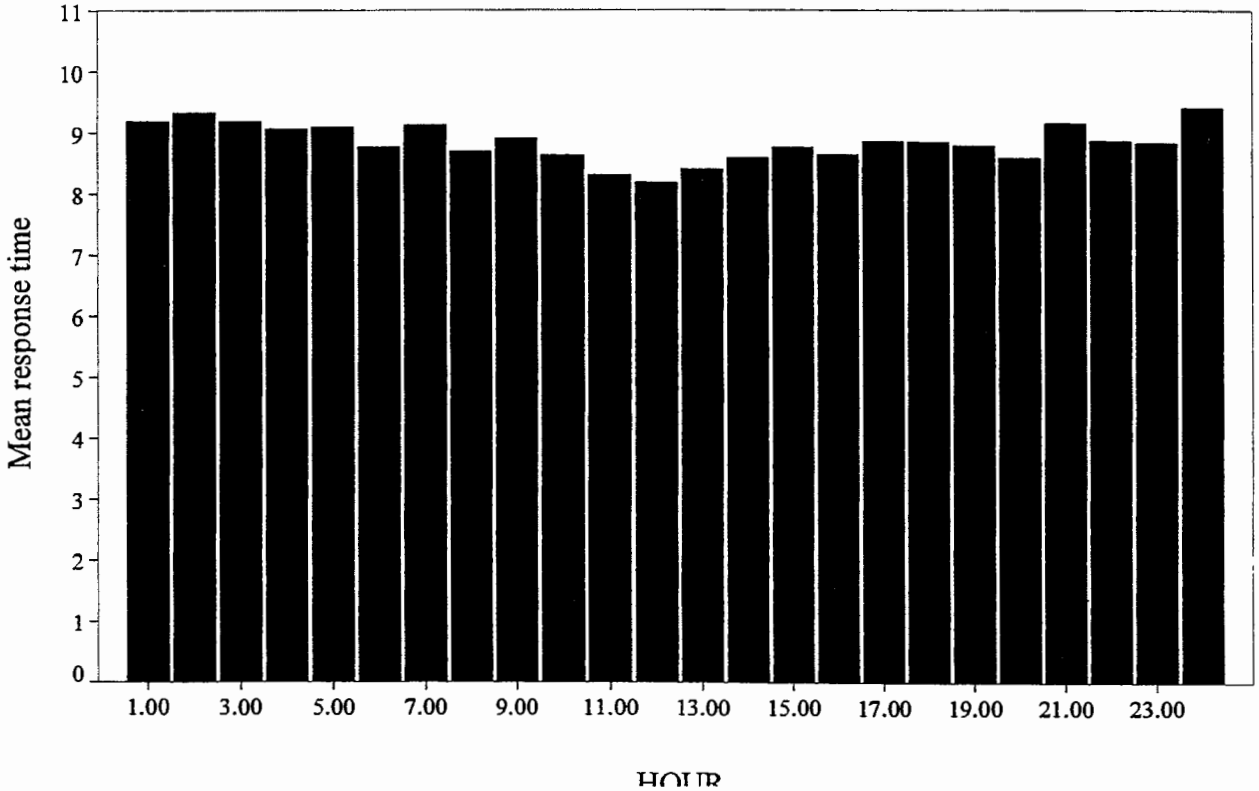

Figure 2

Mean response time during each hour of the day

Friday

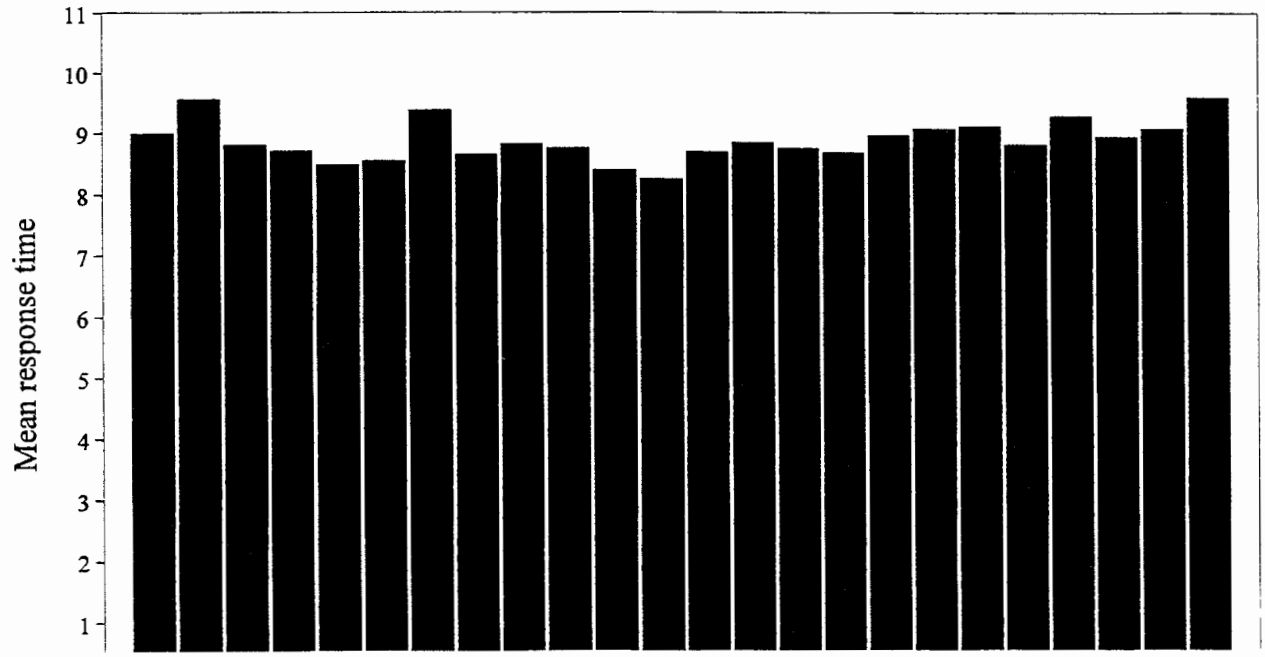

Figure 3

Mean response time for each hour of the day

Saturday

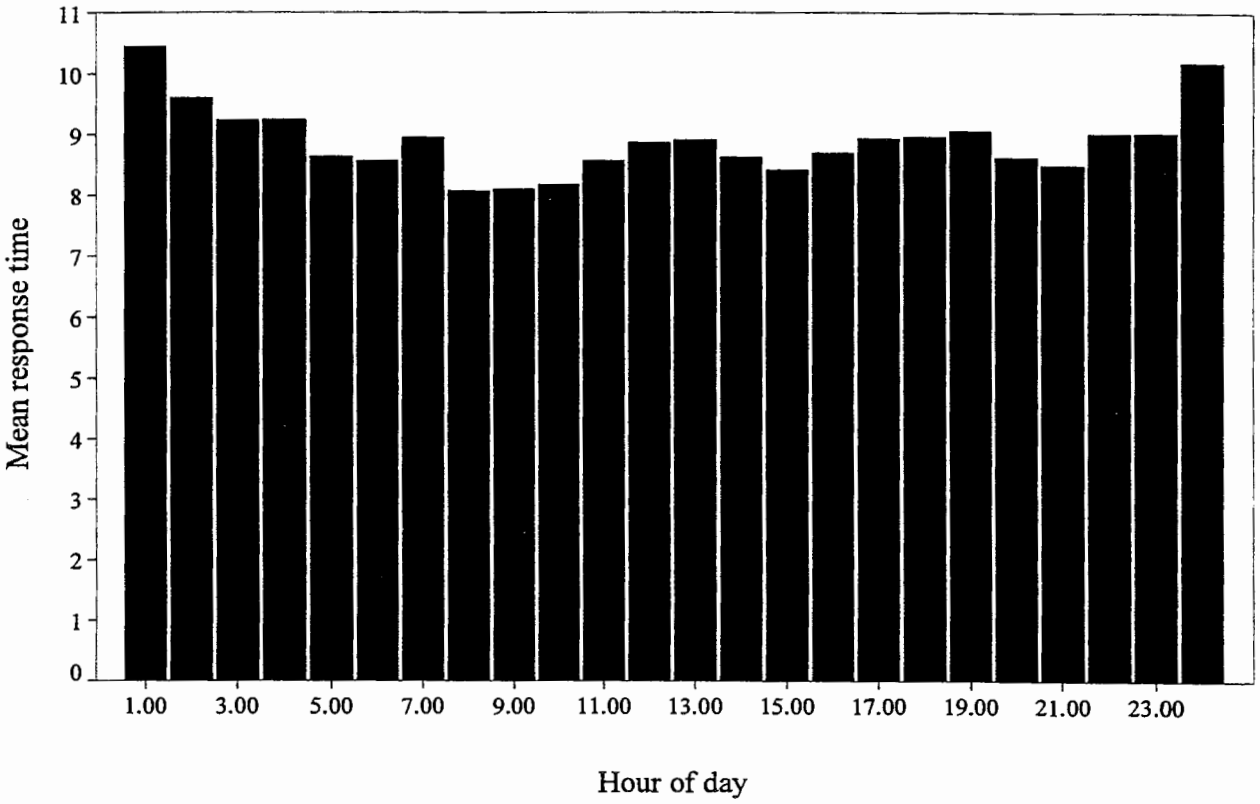

Figure 4

Mean response time during each hour of the day

Sunday

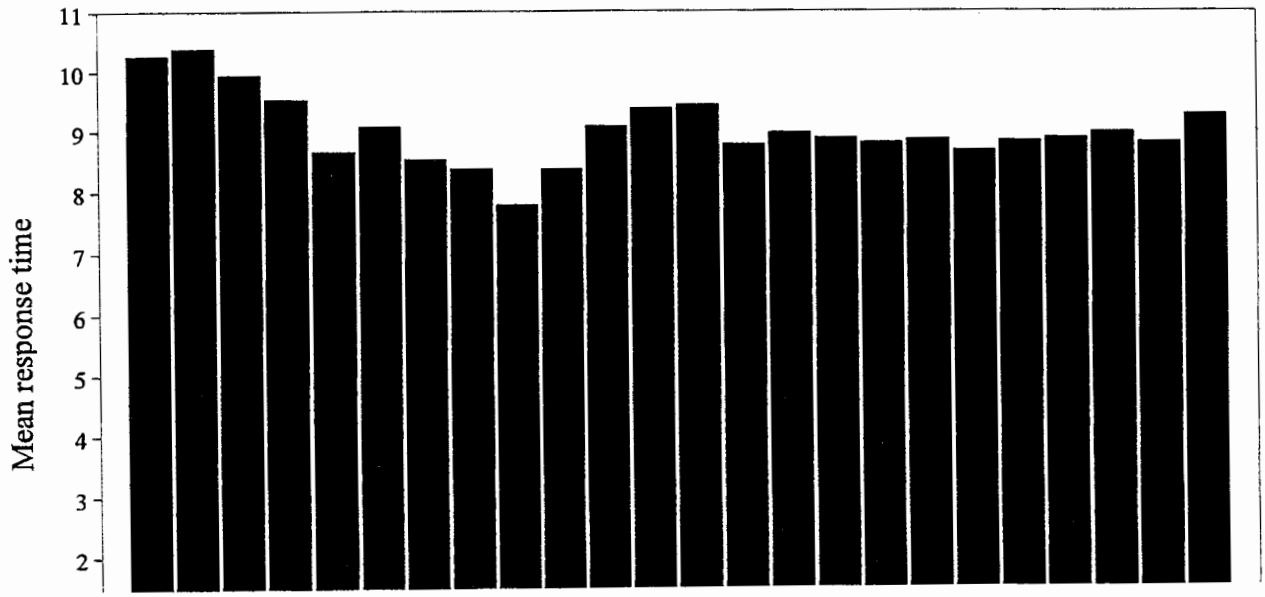

Figure 5

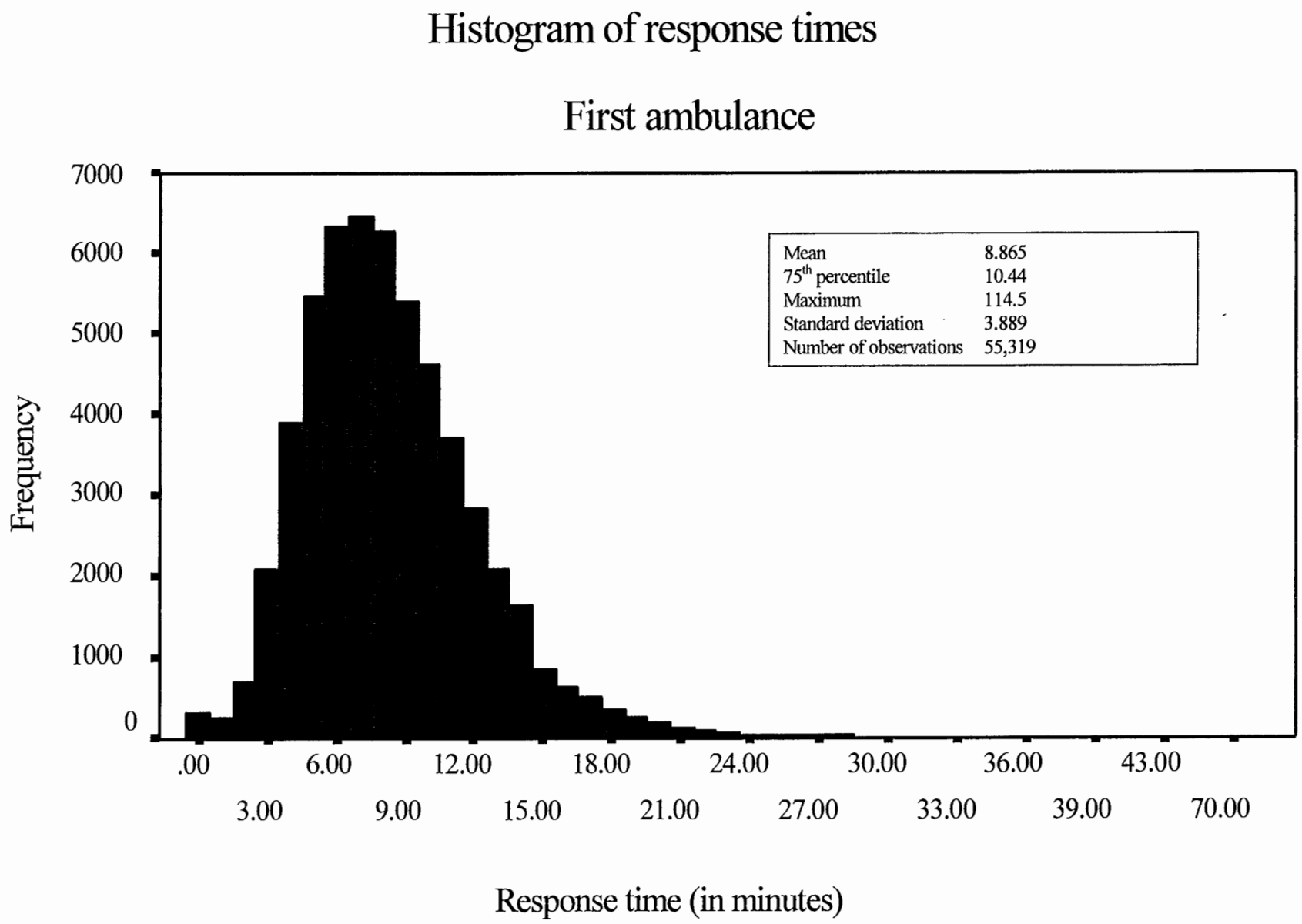
Histogram of response times
First ambulance
Frequency
7000
6000
5000
4000
3000
2000
1000
0
Mean 8.865
75th percentile 10.44
Maximum 114.5
Standard deviation 3.889
Number of observations 55,319
.00
3.00
6.00
9.00
12.00
15.00
18.00
21.00
24.00
27.00
30.00
33.00
36.00
39.00
43.00
70.00
Response time (in minutes)

Figure 6

Mean response time by number of available ambulances not in use

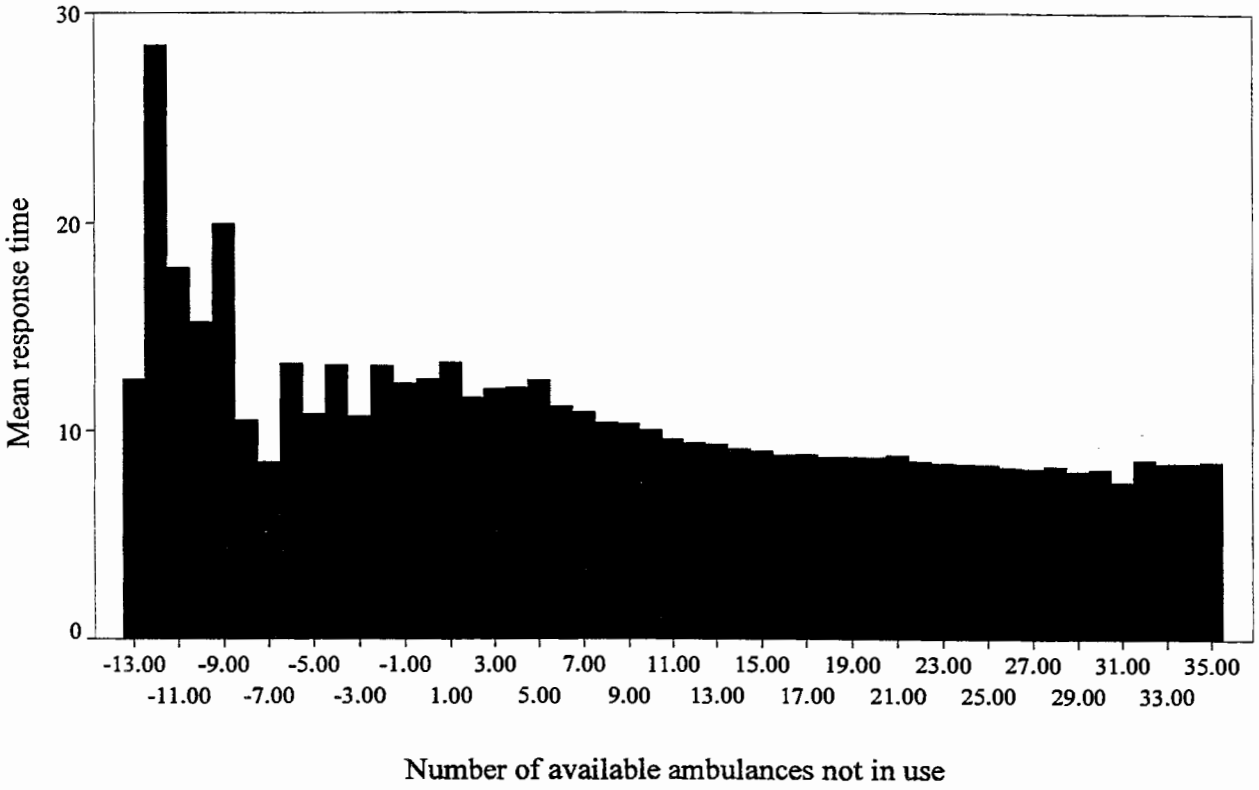

Figure 7

Effect of traffic at each hour of the day controlling for the number of available ambulances