Waiting Times in Accidents & Emergency (A&E)

Marco Scutari

Academic Year 2022–2023

Patients that present themselves to hospital A&E departments are prioritised based on the severity of their symptoms: this process is called *triage* and it is what you will try to model.

Some patients arrive in critical condition and need immediate attention; some can wait for a short time before treatment is administered; others need little or no medical treatment at all. Two important factors that may determine which category patients fall in are the type of *incident* (I) they were involved in and their *age* (A), since older people are physically more fragile and recover more slowly. These two variables can be taken to largely determine the *trauma score* (S) on a scale from 0 to 12. The trauma score is indicative of how urgently the patient requires treatment, so it is strongly indicative of how long the patient is likely to *wait in triage* (W). However, both the hospital's bed *occupancy rate* (O) and the *time of the day* (T) may prolong waiting times. Higher occupancy rates make it more difficult to find a free bed in the hospital to administer medical care. And some times of the day are busier than others, with more people presenting themselves at the A&E during the day than during the night.

Part 1: Construct the Network Structure

1. What arcs should be included in the DAG?
2. Consider the admission process. Firstly, the emergency staff determines the trauma score. Then (possibly different) staff goes through the patient queue in inverse-trauma-score order and either sends the patient home or try to find a bed in the appropriate hospital department. Does the network structure you propose separate these two phases of triaging?
3. Consider the waiting time W. What is its Markov blanket? Could it be smaller?

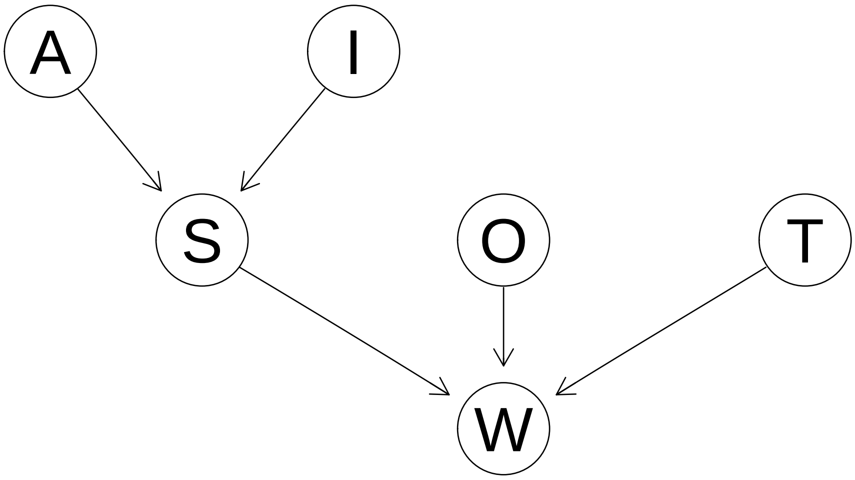
Commented Code for Part 1: Construct the Network Structure

1. From the description above I went for can the following relationships:

   $$\texttt{I} \rightarrow \texttt{S}, \qquad \texttt{A} \rightarrow \texttt{S}, \qquad \texttt{S} \rightarrow \texttt{W}, \qquad \texttt{O} \rightarrow \texttt{W}, \qquad \texttt{T} \rightarrow \texttt{W}.$$

   In pictures:

   ```
   aewaits.dag = model2network("[I][A][S|I:A][O][T][W|S:O:T]")
   graph.par(list(nodes = list(fontsize = 10)))
   graphviz.plot(aewaits.dag)
   ```



Could you consider adding a few more arcs? Definitely, but that would increase the complexity of the model without adding much to its expressiveness. We could argue that people of different ages are likely to have different types of accidents and add $\mathtt{A} \to \mathtt{I}$. We could also add $\mathtt{T} \to \mathtt{O}$ because patients are typically all discharged at a pre-determined time of the day.

2. An important characteristic of this DAG is that it divides our BN in two separate submodels: one for assigning the trauma score ($\{\mathtt{I}, \mathtt{A}\} \to \mathtt{S}$) and one for the waiting time itself ($\{\mathtt{S}, \mathtt{O}, \mathtt{T}\} \to \mathtt{W}$). Once the trauma score is determined, the waiting time is independent from both the incident type and the patient's age: it is easy to verify this is the case testing the d-separation of some nodes in different submodels with the `dsep()` function.

```
dsep(aewaits.dag, x = "A", y = "W", z = "S")
| [1] TRUE

dsep(aewaits.dag, x = "I", y = "O", z = "S")
| [1] TRUE
```

3. The Markov blanket of `W` is the following.

```
mb(aewaits.dag, "W")
| [1] "O" "S" "T"
```

It only contains the nodes in the second submodel.

Part 2: Pick the Most Appropriate Local Distributions

After constructing the DAG, you should now decide

1. which probability distributions to use for the local distribution of each node; and
2. its parameters.

Some hints:

- The type of accidents `A` is naturally a categorical variables. Using just a few, general categories is fine: *domestic incident*, *road traffic incident*, *work incident* and *other incident*.
- In most NHS documents, the age of patients is described in 5-years brackets for ages 0–100. Can you get away with less than 20 parameters?

- Trauma scores S are grouped in four brackets, denoted by color codes: *black* (0–2, beyond help), *red* (3–10, need immediate attention), *yellow* (10–11, can wait for a short time) and *green* (12, treatment can be delayed). It can be modelled either as a continuous or as a discrete variable, whichever is more parsimonious in terms of parameters.
- Different times of the day can be less or more busy: fewer patients arrive in the middle of the night. The T variable can be used to represent that with a cyclical distribution that peaks at noon and bottoms out at midnight. That is a bit of a simplification, admittedly. The density should looks as that below.

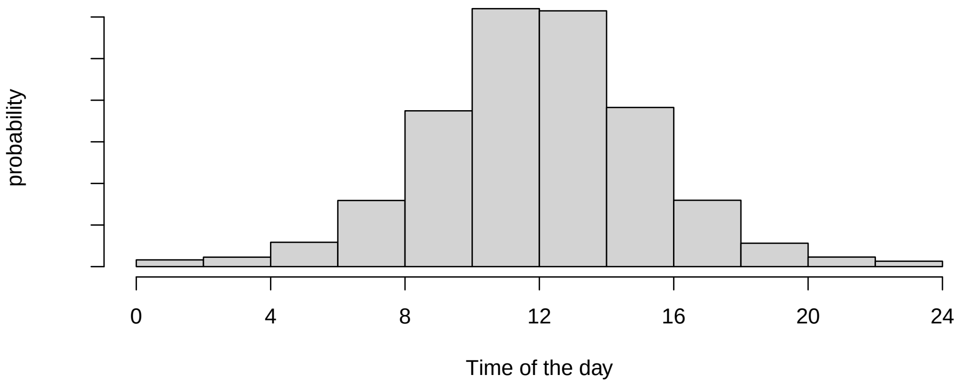


Exhibit #1:

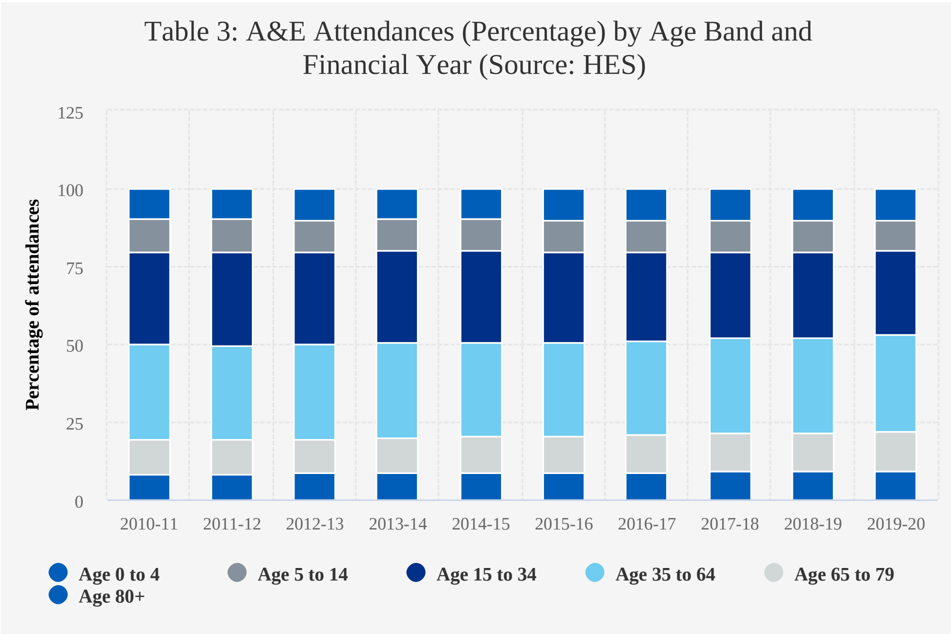


Exhibit #2:

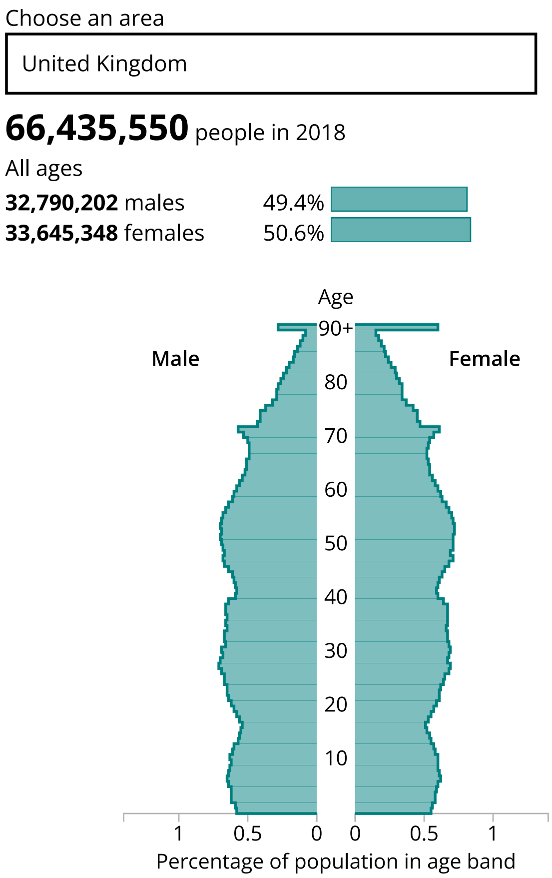
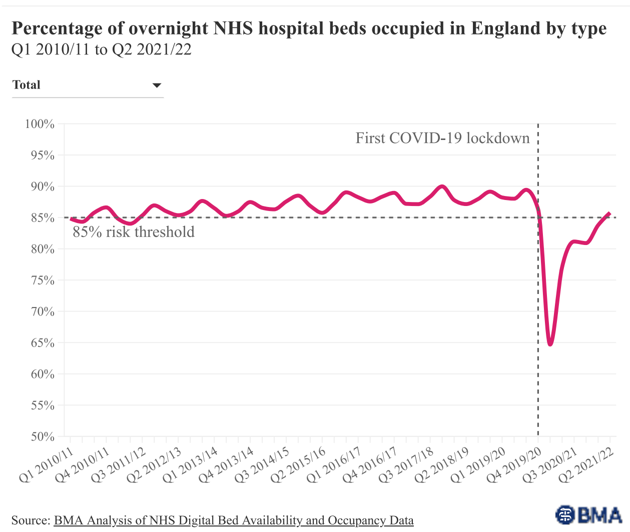


Exhibit #3:



Rising occupancy

While overall bed numbers have declined, occupancy rates have been rising.

Since 2010, average bed occupancy has consistently surpassed 85%, the level generally considered to be the point beyond which safety and efficiency are at risk.

Coming into the pandemic, England had an average occupancy of 90.2% in 2019/20. However, local variation in supply and demand have seen many trusts regularly exceeding 95% capacity in the winter months.

Exhibit #4:

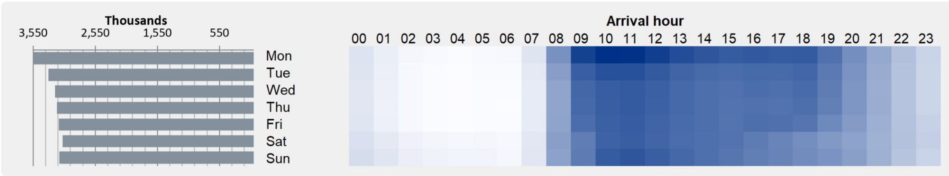


Table 11: A&E Attendances by Time of Arrival and Day of Week (Source: HES)

Exhibit #5:

Total Time in A&E from Hour of Arrival to Transfer, Admission or Discharge

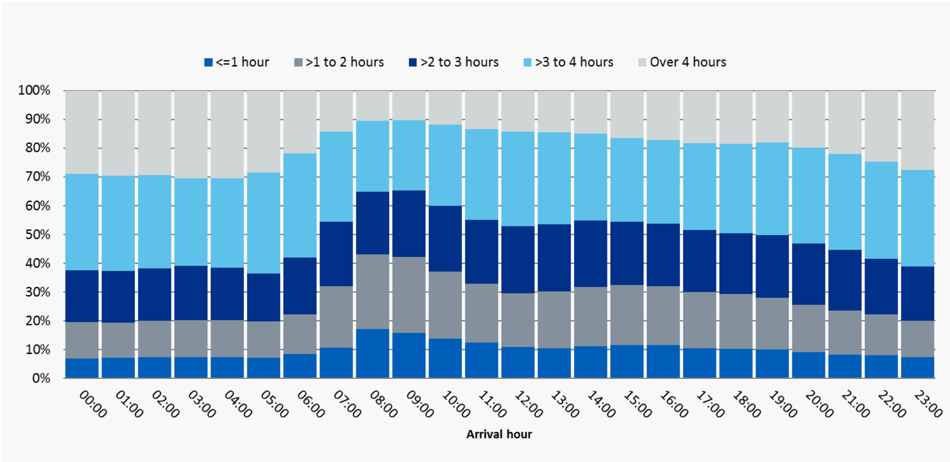
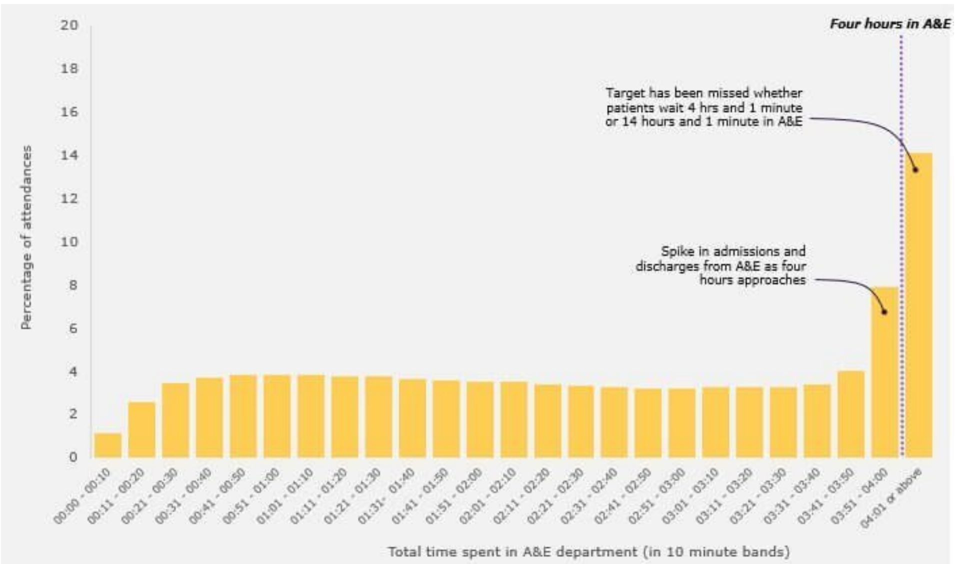


Exhibit #6:



Commented Code for Part 2: Pick the Most Appropriate Local Distributions

For the type of incident, consider four categories and model them with a multinomial random variable:

$$\mathtt{I} = \begin{cases} \text{domestic incident (}\texttt{domestic}\text{)} & \text{with probability} \quad 0.50 \\ \text{road traffic incident (}\texttt{road}\text{)} & \text{with probability} \quad 0.075 \\ \text{work incident (}\texttt{work}\text{)} & \text{with probability} \quad 0.25 \\ \text{other incident (}\texttt{other}\text{)} & \text{with probability} \quad 0.175 \end{cases}.$$

For the age, you can choose a Beta distribution multiplied by 100 and round it to unit values:

$$\mathtt{A} \sim \lceil Beta(1.1, 1.5) \cdot 100 \rceil.$$

This approach preserves the main features of the age distribution of people who attend A&E: the probability of attending A&E it is relatively flat between the ages of 30 and 70; it has a small peak for children aged 0 to 5, and it is larger for people in their 20s than it is for older adults; it gives increasing large probabilities as age increases beyond 70; 25% of patients are younger than 20 and 20% are older than 65. It is easy to check that the 0.25 and 0.80 quantiles roughly match the information above.

```
round(100 * qbeta(0.25, 1.1, 1.5))
```

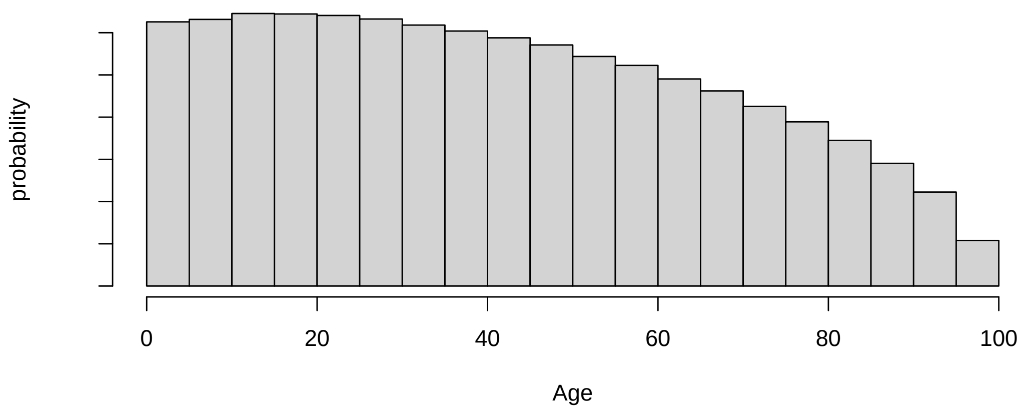
```
[1] 20
```

```
round(100 * qbeta(0.80, 1.1, 1.5))
```

```
[1] 68
```

Visually:

```
hist(round(100 * rbeta(10^6, shape1 = 1.1, shape2 = 1.5)),
  xlab = "Age", ylab = "probability", main = "", freq = FALSE,
  xlim = c(0, 100), axes = FALSE)
axis(2, label = FALSE)
axis(1)
```



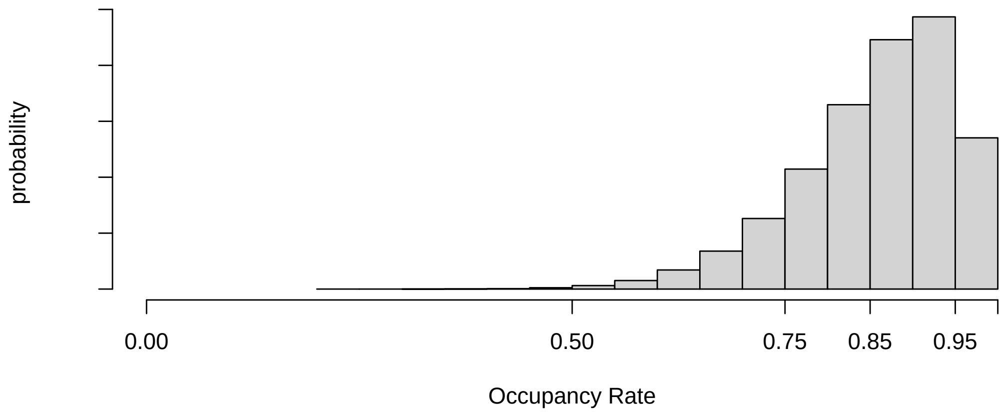
The trauma score is an integer number between 0 and 12, but scores between 0 and 2 are almost never used in practice. Hence $\mathtt{S}$ can be naturally modelled as $\mathtt{S} \sim 2 + Bi(p, 10)$ where the Binomial probability of each point is determined by a logistic regression:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathtt{A} + \beta_2 \mathbb{1}(\mathtt{I} = \mathtt{"road"}) + \beta_3(\mathtt{I} = \mathtt{"work"}) + \beta_4(\mathtt{I} = \mathtt{"other"}). \tag{1}$$

In other words, you can model the log-odds-ratio as a baseline (β_0, determined using domestic incidents since they are the most common) that increases with age ($\beta_1 \mathtt{A}$) and that is adjusted for the average severity of road incidents ($\beta_2 \mathbb{1}(\mathtt{I} = \mathtt{"road"})$), work incidents ($\beta_3 \mathbb{1}(\mathtt{I} = \mathtt{"work"})$) and other kinds of incidents ($\beta_4 \mathbb{1}(\mathtt{I} = \mathtt{"other"})$) relative to domestic incidents. Hence you can set the baseline $\beta_0 = 7$; decrease it by 1 point for each 20 years of age with $\beta_1 = -0.05$; and further decrease it by $\beta_3 = -4$, $\beta_4 = -3$, $\beta_5 = -1$ for the various types of incidents.

The occupancy rate is a proportion, so you can model it with a Beta distribution. We know from NHS statistics that the average is ≈ 0.90, and that the density should be concentrated between 0.80 and 0.99. This leads to an $\mathtt{O} \sim \mathit{Beta}\,(12, 2)$: it has expected value $\mathrm{E}(\mathtt{O}) \approx 0.89$ and $\mathrm{P}(\mathtt{O} \in [0.80, 0.99]) \approx 0.87$.

```
hist(rbeta(10^6, shape1 = 12, shape2 = 2),
  xlab = "Occupancy Rate", ylab = "probability", main = "", freq = FALSE,
  xlim = c(0, 1), axes = FALSE)
axis(2, label = FALSE)
axis(1, at = c(0, 0.5, 0.75, 0.85, 0.95, 1))
```



As for the time of the day T, you need a periodic function that can express the cyclic frequency of patient arrivals. One way to do this is to use a density function defined over a circle, which ensures continuity in the transition from one day to the next. One such distribution is the von Mises distribution, which is defined over $[-\pi, \pi]$ and has two parameters μ (where the peak is in each period) and κ (how sharp the peak is). Hence you can define T as

$$\mathtt{B} \sim \frac{\textit{vonMises}(0.001, 2) - \pi}{2\pi} \cdot 24,$$

which scales it to have period $[0, 24]$ to match a 24-hour clock. The parameter values $\mu = 0.001$ and $\kappa = 2$ make it so that new patients are most likely to present themselves between 10am and 12 noon (when NHS reports the highest daily attendance) and least likely to do so during the night.
Finally, you can choose a log-normal random variable for the waiting time in triage:

$$\log(\mathtt{W}) \sim N(\mu, \sigma^2) \text{ with } \mu = \gamma_0 + \gamma_1 \mathtt{O} + \gamma_2(12 - \mathtt{S}) + \gamma_3 \max\{0, 6 - |\mathtt{T} - 12|\}.$$

The baseline wait (with a completely free A&E, equivalent to $\mathtt{O} = \mathtt{T} = 0$ and $\mathtt{S} = 12$) may be given by $e^{\gamma_0} = 20$ minutes, so you can set $\gamma_0 = \log(20)$. Then assume that W doubles for every 0.50 occupancy, which means that $e^{\gamma_1 0.50} = 2$ and that $\gamma_1 = 2 \log 2$. For each trauma score point below the maximum, you can assume that W halves, which means $e^{\gamma_2 2} = 0.5$ and $\gamma_2 = -0.5 \log 2$. As for T, you can transform it as $6 - |\mathtt{T} - 12|$ to produce a triangular distribution peaking at 12 to link increases in waiting times to high daily attendance hours; and you should make sure the result is always non-negative. The resulting function is $\max\{0, 6 - |\mathtt{T} - 12|\}$. You can then set $\gamma_3 = 0.25 \log 2$, following the same reasoning as for γ_1 and γ_2; and choose $\sigma_2 = 1$ to control the spread of W.

Part 3: Construct and Validate the rstan Network

Construct the Stan model specification and compile it with the `stan_model()` function in **rstan**. Using the `sampling()` function, generate a large random sample from the model and use it to validate the model by answering the following questions.

- Is the frequency of the different brackets of trauma scores realistic? Only a small fraction of the patients should be classified with black codes, most should be split between less-serious red codes, yellow codes and green codes.
- Are the waiting times realistic? About 89% of the patients are seen within 4 hours (that is, 240 minutes), in line with NHS figures, and about half of the patients are seen within 1 hour. What about the (long) tail of the distribution?
- Intuitively, you would expect critical patients to wait much less than non-critical patients due to the trauma score ranking. Are they?

Commented Code for Part 3: Construct and Validate the rstan Network

The specification for the model I described in the solution to **Part II** is as follows.

```
data {
  vector[2] Ap; // shape parameters for the beta distribution.
  vector[4] Ip; // probabilities for incident types.
  vector[6] Sp; // regression coefficients, logistic regression.
  vector[2] Op; // parameters for the beta distribution.
  vector[2] Tp; // parameters for the von Mises distribution.
  vector[5] Wp; // regression coefficients, log-linear regression.
}
generated quantities {
  real A;
  int I;
  real S;
  real O;
  real W;
  real T;
  A = ceil(beta_rng(Ap[1], Ap[2]) * 100);
  I = categorical_rng(Ip);
  S = 2 + binomial_rng(10, inv_logit(Sp[1] + A * Sp[2] + Sp[2 + I]));
  O = beta_rng(Op[1], Op[2]);
  T = (von_mises_rng(Tp[1], Tp[2]) + pi()) / (2 * pi()) * 24;
  W = lognormal_rng(Wp[1] + O * Wp[2] + (12 - S) * Wp[3] +
                    fmax(6 - fabs(T - 12), 0) * Wp[4], Wp[5]);
}
```

After saving it as a string in a variable named `stancode`, you can compile the model and generate a random sample from it.

```r
library(rstan)
data.model = stan_model(model_code = stancode)
params = list(
  Ap = c(1.1, 1.5),
  Ip = c(0.075, 0.50, 0.25, 0.175),
  Sp = c(7, -0.05, 0, -4, -3, -1),
  Op = c(12, 2),
  Tp = c(0.001, 2),
  Wp = c(log(20), 2 * log(2), -0.5 * log(2), 0.25 * log(2), 1)
)
fit = sampling(data.model, algorithm = "Fixed_param",
         data = params, thin = 25, iter = 50000, seed = 42)
nodes = c("A", "I", "S", "O", "T", "W")
aewait = as.data.frame(extract(fit)[nodes])
```

1. As expected, only a small fraction of patients are classified with a black code (less than 3% have S $\leqslant 3$), and few are in the lower half of the red code bracket (about 16% have S $\in [3, 7]$). About 25% of patients are less-serious red codes and another 20% are yellow codes, leaving the remaining 35% as green codes. This is roughly in line with the 30% reported by the NHS some years ago outside of the peak of the flu season.

```r
S.cdf = ecdf(aewait$S)
S.cdf(c(3, 7, 10, 11, 12))
```
```
## [1] 0.0283 0.2050 0.4703 0.6550 1.0000
```

2. The empirical CDF suggests that the proportions of patients seen within 1 hour and within 4 hours are about right.

```r
W.cdf = ecdf(aewait$W)
W.cdf(c(10, 30, 60, 120, 180, 240))
```
```
## [1] 0.119 0.310 0.488 0.691 0.801 0.864
```

The long tail reveals that almost everybody is triaged within 24 hours (24 × 60 minutes = 1440), which also sounds about right outside of flu season.

```
1 - W.cdf(c(240, 360, 480, 960, 1440))
```

```
[1] 0.1365 0.0783 0.0517 0.0135 0.0050
```

3. Patients in critical conditions ($\texttt{S} \leqslant 3$) are seen as soon as they arrive: 73% within 10 minutes, almost all within 30 minutes.

```
nS = length(which(aewait$S <= 3))
length(which((aewait$S <= 3) & (aewait$W < 10))) / nS
```

```
[1] 0.726
```

```
length(which((aewait$S <= 3) & (aewait$W < 30))) / nS
```

```
[1] 0.92
```

On the other hand, patients with yellow and green codes ($\texttt{S} \geqslant 10$) make up the vast majority of those waiting for more than 4 hours.

```
nW = length(which(aewait$W > 240))
length(which((aewait$S >= 10) & (aewait$W > 240))) / nW
```

```
[1] 0.96
```