

A case study on scheduling MRI facilities

Computational Research Skills (EBS4043/EBS4044)

Academic Year 2025-2026

1 Introduction

In a hospital, MRI facilities are both scarce and expensive. Therefore, under-utilization is expensive, but planning patients too tight, will lead to high waiting times, which is also not desirable. Therefore, there is a need to schedule patients efficiently so that the facilities have a high utilization and all patients can be helped within a reasonable amount of time.

Two types of patients need to use the MRI facilities and each type has its own characteristics in terms of duration of the scan and the number of patients per week. The management of the hospital under consideration in this case study, has one MRI facility for each patient type. However, management is considering to use all facilities for all types of patients. Of course, the medical staff is used to the current procedures and are not willing to change them. However, the staff of two of the MRI facilities are willing to consider to change the procedures, as long as they are convinced that it will lead to a significant improvement.

Therefore, the management is asking you to analyze historical data and do a simulation to determine whether merging these two MRI facilities is indeed beneficial.

2 Part I: Statistical Analysis of the Data

The management of the hospital is providing you with historical data on the two patient types. The data shows the date and time at which the call was made as well as the duration of the scan. The data are available over a month for both patient groups.

Your goal is the design and implement a simulation study that can be used for the following two purposes: (1) finding an appropriate length of the timeslots for both types of patients and (2) generate new data that can be used to compare the old and new scheduling policy.

As a first step, you will focus on a statistical analysis of the historical data. The hospital would like to know how many patients to expect each day, and how long their scans will take. Having more experience with the first time of page, hospital staff knows more about these data than the second type.

Specifically, the following information is known:

- All series can be treated as independent and identically distributed.
- For the first type of patients (denoted as 'Type 1' in the dataset), it is known that the durations of the scans are normally distributed, though the mean and standard deviation are unknown.

- For the first type, it is known that the number of patients per day is distributed according to a Poisson distribution, so that the time between two consecutive arrivals is exponentially distributed. The mean of the distribution is unknown though.
- For the second group of patients, the hospital does not know the distributions of scan durations and arrival times.

Your goal is to estimate the unknown (elements of the) distributions. We want to pay particular attention to measuring the uncertainty of the estimates in an appropriate way. For this purpose, we will use the **bootstrap**. It is up to you to decide which type of bootstrap study you are going to design: a parametric or non-parametric study (or perhaps a mix?).

Note also that the hospital is not necessarily interested in estimates for the parameters of the distributions. Instead, they care about estimates of quantities that they can interpret. When designing your statistical analysis, make sure that you keep this point in mind when you report on your results. Trying to motivate what quantities would be **relevant** for the hospital, and provide estimates (along with uncertainty) for these.

While the first group of patients can be dealt with relatively straightforwardly, the second group is a bit more complicated. To give you some suggestions, consider the following. It is possible to estimate relevant quantities (Think of means, quantiles or probabilities of crossing some threshold) without using a particular distribution; for example, using so-called 'plug-in' estimators (see the video). Alternatively, you could try fitting a distribution that you think is (at least approximately) correct. In either case, such an approach begs the question, how accurate these estimators are if indeed we do not know the correct distribution. You could investigate this, using a small, Monte Carlo simulation study. The hospital is likely to be sceptical about your approach for the second group, so it would be good to back your analysis up with such a simulation study.

3 Part 2: Optimizing scheduling policies

To provide some evidence on whether or not using the two MRI facilities for the two patient types is beneficial, the hospital management asks you to design a *discrete event simulation*, using the results of the data analysis.

Patients can call for an appointment every (working)day from 8:00h until 17:00h. When a patient calls to make an appointment, they can not be scheduled on a timeslot at the same day. The MRI facilities operate starting every (working)day from 8:00h and finish as soon as all patients of that had their scans; using the facility after 17:00 is considered overtime.

Patients of the first type each are scheduled for a time slot of fixed length that **you must decide on** based on the analysis of the data and patients of the second type are scheduled for a potentially different, but fixed time slot. The goal is to set up a schedule so that all scans finish by 17:00h.

Management would like you to analyze both the performance of the old system as well as the performance of the system in which the two facilities can be used for both patient types. As they have no clear idea on how the performance is measured, they also ask you **to clearly describe which indicators you are using to measure the performance**.

In the old system, a patient that calls in to make an appointment is assigned the earliest feasible available time slot (on the next working day) on their dedicated machine, according to the above mentioned rules.

In the proposed new system, as soon as a patient calls in to make an appointment, the patient is assigned to the earliest available time slot on one of the two machines; again satisfying the conditions that a patient can be scheduled at the earliest the next day and the time slot should ideally finish not later than 17:00.

Some further things to think about:

- Ideally, the hospital would like you to give them some advice about how to choose the time slots as well. This is relevant in both the new and the old system.
- While it is okay if you for a large part disconnect the statistical part from the discrete event study (you obviously need some input parameters, but you could treat those as given), things get really interesting if you integrate the two parts. What role could the uncertainty about your estimates play in the discrete event simulation? Could there perhaps still be any role for the bootstrap in the second part?

4 The Dataset

The dataset contains the following variables:

- **Date:** The date at which the call for an appointment was made.
- **Time:** The time at which the call for an appointment was made. The time is recorded in fractions of hours; for example, a call made at 9:30 would be recorded as '9.5'.
- **Duration:** The duration of the scan in fractions of hours.
- **PatientType:** The type of patient needing the scan.

Note: when calculating arrival times, you may simply assume that no time elapsed outside of working hours. Hence, the time between a call at 16:45 on day 1 and 8:15 on day 2 is simply a half an hour.

5 Assignment

Write a report of at most 10 pages (not including appendices) on both the data analysis as well as the analysis of the two policies. Make sure that the report has a decent introduction, body and conclusions. Write the report in such a way that that it tells the full story of your analysis, specifically focusing on the interpretation of your analysis. (E.g., the hospital staff should be able to read your report and get the main message that is relevant for them.)

You are free to use whatever programming language you prefer. Python or R are perhaps the most logical options, but this is entirely up to you. Your code does not have to be included in the report, but we would like you to make your code available as open source via GitHub or GitLab.

Working with a group on joint code is a lot easier if you use Git rather than via Dropbox or sharing by email. Moreover, it is a way of working that is used a lot in various data science jobs in industry. We therefore strongly encourage you to use it as well and gain experience with this.