# Exploring simple appointment based queuing system

MH4702 Group 9

Semester 1, AY 18/19

This project is done by:

- James Zephaniah Hadimaja (U1540059C)
- Joevhan (U1640048D)
- Lee Yik Siang (U1640262L)
- Phua Si Jia (U1540765E)
- Yohanes Alfredo Phoa (U1520179D)

## Contents

# 1 Introduction

A good number of the literature with regards to appointment based queuing systems are centered around the healthcare industry [2][3][4]. When considering a good solution to such queuing systems, there are some considerations that we should take into account, such as the scheduled interval is too close, resulting in long waiting time for the patient, or scheduled too far apart such that the servers are under-utilized. We consider the problem of obtaining a good appointment schedule of $n$ independent arrivals of customers with exponential service time. In this report, by simulating 4 appointment based models, we will explain the usefulness of our models in aiding the decision making the process to certain specialist clinics and dentists.

# 2 Methodology

In this report we will explore 4 appointment-based queuing systems. They are $D/M/1$, $D/M/1$ with no show ($D^*/M/1$), $D/M/2$ with no show ($D^*/M/2$) or 2 parallel server system with no show and 2 Sequential Servers with no show.

## 2.1 Problem Definition

In general, the queuing system in small-scale specialist/dental clinics can be approximated using $D/M/1$. This is because many of these small-scale establishments have only 1 doctor to provide the medical service. If the administrative process is very fast, we may assume time spent doing these things to be negligible and we can leave it out of our simulations. We can also assume the service time follows an exponential distribution as the service time provided can vary according to the type of service rendered and that we assume that each person's service time is independent of the others. For simplicity, we assume that the probability of *no show* to be constant and it follows a Bernoulli distribution. We follow a first come first served policy, meaning that if a person's slot is earlier, they will be served first.

Three main performance criteria that we will be focusing on are the patient's mean waiting time, doctors' mean overtime, and mean idle time. We choose to look at the mean waiting time of each customer because of its relation to customer satisfaction as well as customer loyalty which is important to many establishments [1]. Overtime and idle time both incur sizable costs on the businesses themselves [4] and thus we need to include it in our considerations. We can evaluate these 3 criteria simultaneously by taking the weighted sum of the meaning waiting time, mean idle time, and mean overtime [2][4]. The weights for these 3 criteria are 0.5, 1, and 1 respectively. We place more emphasis on minimizing wastage in server utilization and additional costs associated with overtime. We shall call this weighted sum "total costs" (not to be confused with the total costs in the traditional monetary sense). Note that when we say total costs increase we are implying system performance becomes worse (based on the weights assigned). Another possible criterion that we can look at is the number of people in the system (including the current customer), which we will go into the details later on in the report.

## 2.2 Simulations

We will only briefly touch on how we computed our performance criteria, performed variance reduction, and sensitivity analysis as the rest of the details can be found in the excel file *"Projectscript"*. For $D/M/1$, $D^*/M/1$ models, we compute the waiting time for each patient by taking the difference between their arrival time and service starting time. To calculate the idle time between consecutive patients, we take the difference between the previous patient's service ending time and the current patient's arrival time. Additionally, if the last patient leaves before the end of the day we consider the remaining time to closing as idle time. Daily overtime is calculated by taking the last patient's service ending time minus daily working hours, if this value is negative we take it that no overtime is incurred. For $D^*/M/2$ and 2 Sequential server model, since we assume both doctors leave together, we need to multiply this value by 2. It is important to make such a distinction between idle time and overtime as their associated costs can be different in practical settings. The idle time for each server in 2 sequential model is calculated similarly to the 1 server models. For $D^*/M/2$, we calculate our daily idle time by taking the maximum of the last patient's service ending time and working hours per day, multiplying it by 2 and subtracting the total service time for the day. Lastly, for both mean waiting time and idle time, we just take the average waiting and idle time over the people (who showed up) and for mean overtime, an average of overtime over the number of days under consideration.

We perform our variance reduction through the use of antithetic variables. The results are recorded in the spreadsheets *"Basic Appt (Var Red)"*, *"No Show (Var Red)"*, *"No Show w 2 Servers (Var Red)"*, and *"No show & seq servers (Var Red)"*. In general, the larger *"Random Number $U1$"* generated, the more likely it is to exceed the probability of showing up and hence the patient becomes a no-show which decreases waiting time, thus waiting time is a decreasing function with respect to $U1$. However for *"Random Number $U2$"* (and *"Random Number $U3$"* for 2 sequential servers model), the larger the values of $U2$ and $U3$, the longer the corresponding service time, potentially increasing waiting time for subsequent customers (since the longer service time might "eat" into their appointment). Thus, expected waiting time is an increasing function with respect to $U2$ and $U3$. With what we have above, we can perform variance reduction to get more accurate results for the expected mean waiting time.

In terms of sensitivity analysis, we choose to perform a sensitivity analysis on only the scheduled arrival interval because it is one of the parameters that can be easily controlled by the clinics themselves. Changes in service rates may not always be easy to control (for example, how efficient a doctor can be after becoming more proficient is outside of their control). We can also perform sensitivity analysis on the associated weights assigned to the 3 performance criteria, but we usually already have an idea of what we want to focus on (for example focus on patient's satisfaction or focus on reducing wastage) when performing such simulations so we do not usually need to see how varying the weights affect total costs.

# 3 Findings

In this section, we will present our findings for the four appointment models mentioned earlier. For the purpose of easy comparison, we will be using data from the excel file *"Static"*. We fixed each patient to arrive every 0.5 hour, the daily working hours to be 8 hours, and the service rate for all servers are set at 3 patients per hour. Furthermore, for models that account for no shows, we set our probability of showing up to be 0.8. All mean waiting time results used here are obtained from performing variance reduction while the rest are taken from the normal simulation spread sheet (without variance reduction). Our simulations in this section is performed using Microsoft Excel 2016 and our sensitivity analysis is performed using *"Data Table"* under *"What-If Analysis"*. Our findings are summarized in the table below:

Table 1: Simulation Results

| Queuing Model | Mean Waiting Time | Mean Idle Time | Mean Overtime | Total Costs |
|---|---|---|---|---|
| $D/M/1$ | 0.146 | 0.221 | 0.609 | 0.895 |
| $D^*/M/1$ | 0.101 | 0.340 | 0.504 | 0.894 |
| $D^*/M/2$ | 0.00161 | 0.970 | 0.928 | 1.898 |
| 2 Sequential Servers (No Show) | 0.299 | 0.750 | 2.256 | 2.753 |

## 3.1 Accounting for no shows in D/M/1 system

Based on our simulations, it was observed that by setting the probability of not showing up at 0.2, the mean waiting time for each patient decreased by around 30 percent. This is because if a patient does not show up, the doctor can use the extra time till the next patient's arrival to finish up their service with their current patient and/or have more idle time. Thus, when the next patient enters the clinics, their waiting time would be naturally shorter. At the same time, mean overtime also decreased by around 17 percent, the reason for this is also similar to what we have mentioned earlier, if fewer patients show up, then doctors will have more time to finish up their service with other patients and thus reducing the need to work overtime. However, mean idle time increased by almost 54 percent (as explained earlier). When taking into account the weights that we assigned to each of these performance criteria, our total costs are actually around the same for both of these systems. The benefits of reduced expected waiting time and mean overtime has balanced out the cost of increased idle time. So at the current set of parameters, the performance of the clinic would not be badly affected by the patient's no show.

## 3.2 Adding one more server in the no show system

Now suppose we add one more doctor, the mean waiting time decreases by 98 percent as compared to before as having more doctors means fewer customers need to wait for their service or wait a shorter time for their service to begin. However, mean overtime actually increased in terms of man-hours. As mentioned earlier, both doctors are assumed to leave at the same time so the overtime is multiplied by 2. So if we just look at the individual doctor's mean overtime, the amount of overtime per doctor decreased by about 8 percent by adding one more doctor. However, this decrease is not as much as what we initially thought because of how we arranged the patients' arrival. By arranging the last patient to come at the closing time, we are inevitably resulting in some overtime being incurred. In our computations, our mean idle time is computed together for the 2 patients in the 2 server system, thus to approximate individual doctor's idle time, we will just divide this value by 2. We observe that the addition of 1 more doctor increased mean idle time per doctor increased by around 43 percent, this is because having more people to share the same load would generally mean that each doctor has a lighter workload. Furthermore, we can work out that the $D^*/M/1$ model is unlikely to be at full capacity given that idle time is around 20.4 minutes per patient who showed up. If on average, 80 percent of customers showed up daily (roughly 12 customers), we have around 245 minutes (4.08 hours) of idle time which is easily half the working hours. So the additional doctor is likely to increase the idle time per doctor as well. Performance-wise, the total costs will increase from 0.894 to 1.898, indicating that system performance is not as good.

## 3.3 Two Sequential Servers and Two Parallel Servers

Now suppose that the doctor wants to decide between establishing a triage station, or employing another doctor to share his/her workload. In the queuing system for 2 sequential servers (nurse in the triage in service 1 and doctor in service 2), our mean waiting time $(0.299 hours)$ is much higher than that for $D^*/M/2$ (0.00161 hours). Additionally, the mean waiting for doctor's service (service 2) to start in the sequential server's system is around 0.197 hours which constitute the bulk of their mean waiting time. The reason for the large difference in mean waiting time is that in the sequential servers queuing system, each stage has only 1 server so we need to wait for that particular server in that stage to complete their service for previous patient before we can start our service for that stage. In contrast, for the $D^*/M/2$ model, both servers (doctors) are providing the same service and thus we just need to wait for either of them to be done to start our service which leads to a substantially lower mean waiting time. The mean overtime of the parallel servers queuing system (0.928 hours) is also substantially lower than that of the sequential servers queuing system (2.26 hours). As mentioned earlier, having more people to share the same workload will enable the doctors to more quickly serve all the patients, leading to less overtime required. In contrast, in the sequential server's system, as two stages of service are distinct from each other so each server still has to serve everyone that shows up, thereby resulting in longer overtime hours. However, mean idle time for the 2 doctors in the parallel server's system (0.970 hours) is more than the mean idle time for the 2 servers in the sequential servers queuing system (0.750 hours). The possible reason for this is similar to what we have already mentioned earlier. If we are to consider the overall system performances, we would find that the total costs of the sequential servers queuing system (2.753) are higher than that of the parallel servers queuing system (1.898). Meaning performance wise, employing another doctor is better than establishing a triage.

## 3.4 Further Analysis

In this section, we will explore how we can use the number of people in the system (including current customer) as a possible performance criterion. We plotted the histograms of the frequency distribution of the number of patients in the system (including the current patient), for all 4 models and we present the plots in the figures below. We observe that from Figure 1 and 2 that majority of the time, the current patient is the only person in the system when they enter. This provides us with an indication that the proportion of patient who gets served immediately upon entering the system is very high, implying that the probability of having to wait is rather low. In Figure 3, we can get draw the same conclusion from taking observing the frequency of patients in the first two bins. In fact, almost all the patient encounter the situation where there are 2 or less patients in the system (including themselves) when they enter, thus implying probability of having to wait for service is even lower. Interestingly, we see that the graph in Figure 4 is relatively more evenly spread out. A possible reason for this can be that the patients have to spend longer time within the system in the 2 sequential server model as there are 2 stages of services they need to go through (and there are 2 potential stages of waiting), whereas for the other 3 models we only have 1 stage. So with the same arrival interval, there would likely be more people in the system in when i enter, since people clear out of the system at a slower pace. While we are unable to safely determine what is threshold to conclude that we definitely need to wait, but we can use what we observed here in conjunction with Figure 1, to loosely conclude that we are likely to have a higher probability of needing to wait for service here. The lower the probability of waiting, the higher the level of patients' satisfaction in general, and is important if we are evaluating the system from a patient-centric approach.

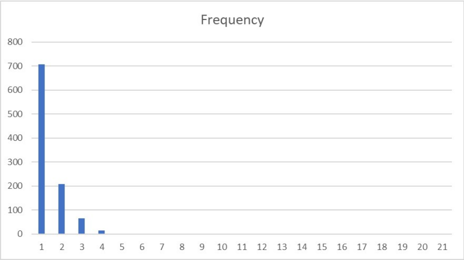

Figure 1: D/M/1

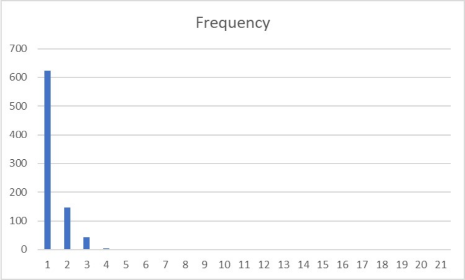

Figure 2: $D^*/M/1$

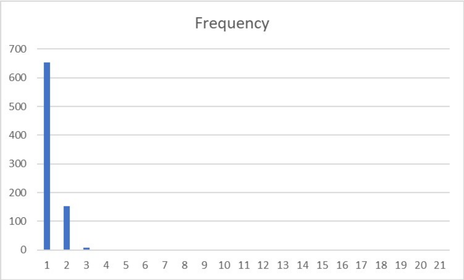

Figure 3: $D^*/M/2$

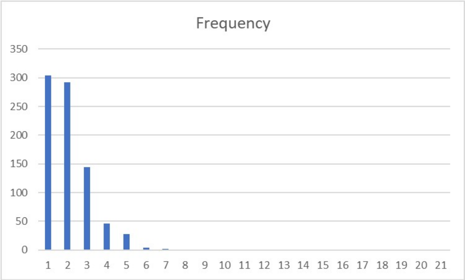

Figure 4: 2 Sequential Servers (No Show)

# 4 Applications of Findings and Sensitivity Analysis

Comparing between $D/M/1$ and $D^*/M/1$ can help these clinics predict the effects of appointment no shows on system performance (ie total costs), and utilize sensitivity analysis to help them find a better scheduling plan. Accounting for no shows, and based on our sensitivity analysis for $D^*/M/1$, by changing the interval of arrival for each patient from 0.5 hours to 0.6 hours, we would be able to reduce total costs by about 26 percent. Furthermore, if at some point they wish to evaluate whether it is worthwhile to employ another doctor for their establishment, we can use the results from $D^*/M/2$ to help them estimate the increase in costs and compare systems' performances (although it is unlikely they need to employ another doctor since we already worked out that that the system is currently under-utilized). Should they decide to employ another doctor, they can then use sensitivity analysis for $D^*/M/2$ to help them determine that by changing interval of arrival from 0.5 hours to 0.3 hours, they can reduce total costs by around 59 percent. As mentioned earlier with the 2 sequential servers with no show model they can also evaluate whether it is more worthwhile to employ another doctor or set up an additional station providing another service (for instance a triage).

# 5 Optimization

This section covers how the optimization on our simulation problem is done. Our optimization was done by using Bayesian Optimization with Python instead of using Microsoft Excel. This is because as the number of observations grow, the excel's solver performance speed decrease. Our simulations here is performed on Python Version 3.6.7 (Required packages are "Pandas", "Numpy" and "bayesian-optimization"). The source code file is "source.py".

## 5.1 Bayesian Optimization

Bayesian optimization works by constructing a posterior distribution of functions (gaussian process) that best describes the function that we want to optimize. As the number of observations grows, the posterior distribution improves, and the algorithm becomes more certain of which regions in parameter space are worth exploring and which are not. As we iterate over and over, the algorithm balances its needs of exploration and exploitation by taking into account what it knows about the target function. At each step a Gaussian Process is fitted to the known samples (points previously explored), and the posterior distribution, combined with a exploration strategy (such as UCB (Upper Confidence Bound), or EI (Expected Improvement)), are used to determine the next point that should be explored. This process is designed to minimize the number of steps required to find a combination of parameters that are close to the optimal combination. To do so, this method uses a proxy optimization problem (finding the maximum of the acquisition function) that, albeit still a hard problem, is cheaper (in the computational sense) and common tools can be employed. Therefore Bayesian Optimization is most adequate for situations where sampling the function to be optimized is a very expensive endeavor.

On a more realistic setting, business entities sought to increase its generating profit and sustain the business in the long-run. Consider a simplified scenario of a clinic for our optimization problem. The optimization problem is defined as maximizing the clinic profit based on the following assumptions :

Each day a customer will arrive at the clinic at a fixed interval. Customers are not be able to arrange appointments beyond the opening hours. The patient can choose not to show up with a certain probability. The patients are assumed to arrive on time and will wait for any length of time for his/her service and will not balk from the queue. There is no maximum length for the queue. Each patient who shows up for their appointment is liable to pay a fixed fee. If a customer chooses not to show up, he/she will not pay for the service fee and hence will not generate any income. The doctor's skill is measured by his or her service rate. A more skilled doctor will have a higher service rate and able to serve patients faster. The doctor is also entitled to overtime pay in the event he/she operates beyond his/her assigned work hours. Other costs are included as hourly operational cost. This cost will be incurred as long as the clinic is open.

## 5.2 Results

For our system optimization, we design the following scheme. The clinic will open for 8 hours, customer fee is set as 600. The hiring cost for the doctor is $1000 + 1800$ multiplied with service rate and the doctor will be given a fifth of his daily wage as hourly compensation for his or her overtime. The hourly operational cost is set as 300 per hour. Hence, the optimization objective function is to maximize. We decided to vary only the doctor's service rate and scheduling interval. Each simulation or function calls during the optimization will simulate the clinic for 30 days and calculate the necessary calculation of profit. We run our scenario with the base parameter of service rate of 3 and scheduling interval of 0.5 hours as our baseline. We assess the performance in terms of profit by calling the function 100 times and aggregate the results. For the case with 0 probability of not showing up, We found that on average the clinic will earn 14621.185 monthly with a standard deviation of 4011.349. After optimization, we found that the optimal parameter is with service rate of 3.243 and scheduling interval of 0.226. After the optimization, the clinic able to yield 200125.833 on average monthly with a standard deviation of 15943.814. For the with 0.2 probability of patient not showing up. The clinic yields $-39155.920$ on average monthly with a standard deviation of 5847.893. The best parameter for this scenario is with service rate of 2.993 and scheduling interval of 0.269. After optimization, the clinic yields 172976.861 on average monthly, with a standard deviation of 11870.862. The queuing system performance after optimization is compiled in this table.

Table 2: Simulation Results for Optimization

| Queuing Model | Mean Waiting Time | Mean Idle Time | Mean Overtime |
|---|---|---|---|
| $D/M/1$ | 0.181 | 0.156 | 2.904 |
| $D^*/M/1$ | 0.103 | 0.228 | 1.203 |

# 6 Conclusion

In summary, we have looked at 4 appointment systems and evaluated each system's performance based on the criteria that we specified (mean waiting time, mean idle, mean overtime and number of patient in the system, including the current patient). We also explained how such information can be useful to specialist/dentist clinics when making business decisions. Additionally, we also brought in alternative methods of simulating this problem (with optimization) in Section 5. Nonetheless, we recognize that many of the queuing systems in the health care industry are more complex than what we have discussed in this report and that further exploration is required to study those systems.

# References

[1] Frédéric Bielen and Nathalie Demoulin. “Waiting time influence on the satisfaction-loyalty relationship in services”. In: *Journal of Service Theory and Practice* 17.2 (2007), pp. 174–193.

[2] Tugba Cayirli and Emre Veral. “Outpatient Scheduling In Health Care: A Review Of Literature”. In: *Production And Operations Management* 12.4 (2003), pp. 519–549.

[3] Diwakar Gupta and Brian Denton. “Appointment scheduling in health care: Challenges and opportunities”. In: *IIE Transactions* 40.9 (2008), 800–819.

[4] Chongjun Yan Jiafu Tang and Richard Y.K. Fung. “Optimal appointment scheduling with no-shows and exponential service time considering overtime work”. In: *Journal of Management Analytics* 1.2 (2014), pp. 99–129.