A (Simplified) Classic Queuing Theory Result

Math 394

A famed anecdote has a bank manager stumble into a queuing theory talk in the 60s, and walking out determined to change how his bank handled customers in line - and the single feeder line was introduced in real life. In short, the result that impressed the manager can be loosely stated as follows: under some assumption, customers faced with n tellers (e.g., at the bank, or the DMV), will be served, on average, faster if they wait in a single line, with the front customer being served by the first teller that completes its previous service, rather than break up in n lines, one per teller. Here is a much simplified setting that shows how this can be the case.

1 A Simple Model

Assume for simplicity $n = 2$. Also, assume that the two tellers have serving times that are independent, identically distributed, with exponential distributions, with parameters λ_1, and λ_2. Also assume that, as you enter the premises, there are $2m$ customers in line. In one scenario, the customers are divided in two lines of m individuals each. In the other, they are waiting in a single line. As a customer arrives, she will become the $2m + 1$-st customer in line in the second scenario, and, lacking any additional information, will choose one of the two lines with probability $\frac{1}{2}$ in the first. The question is which strategy will result in the shortest expected waiting time until service.

1.1 Separate Lines

Each customer in line will be served for a random time, which we assume to be exponentially distributed with the parameter associated with the teller. Thus, line 1 will take $\sum_{i=1}^{m} T_i$ to empty, with the T_i's independent random variables, with parameter λ_1. The expected waiting time is then $\sum_{i=1}^{m} E[T_i] = \frac{m}{\lambda_1}$. Similarly, for line 2, so that the expected waiting time (the two lines are equally likely to be chosen) is

$$m\left(\frac{1}{2}\frac{1}{\lambda_1} + \frac{1}{2}\frac{1}{\lambda_2}\right) = \frac{m}{2}\frac{\Lambda_1 + \lambda_2}{\lambda_1\lambda_2} \tag{1}$$

.

Remark: We have used, implicitly, the fact that conditional probabilities are probabilities. Hence we can take expectations with respect to them, obtaining what are, reasonably, called *conditional expectations*. In our case, calling C_i the event of choosing line i, the total waiting time is computed as

$$E[T] = E[T|C_1]\,P[C_1] + E[T|C_2]\,P[C_2]$$

1.2 Single Feeder Line

If we have only one line, each successive customer is served by the first teller that finishes its previous service. Given that all service times are independent and exponential, any time a new service is started, the remaining service time at both tellers is again exponential with parameter λ_i! Hence, the line moves at times defined by $\min(S_1, S_2)$ where S_i is the service time of teller i. We know that the minimum of two independent exponential random variables is exponential, with parameter the sum of the parameters. Hence, since we have $2m$ customers in line, the total waiting time will be the sum of $2m$ such variables, and the expectation will be

$$\frac{2m}{\lambda_1 + \lambda_2} \tag{2}$$

Comparing (1) and (2), we have

$$\frac{m}{2} \cdot \frac{\lambda_1 + \lambda_2}{\lambda_1 \lambda_2} - 2m \frac{1}{\lambda_1 + \lambda_2} = \frac{m}{2} \frac{(\lambda_1 + \lambda_2)^2 - 4\lambda_1\lambda_2}{\lambda_1\lambda_2(\lambda_1 + \lambda_2)} = \frac{m}{2} \frac{(\lambda_1 - \lambda_2)^2}{\lambda_1\lambda_2(\lambda_1 + \lambda_2)} \geq 0$$

with equality only if $\lambda_1 = \lambda_2$. Note that the single feeder line will never perform worse than the alternative setup, and will be better as soon as one line has a faster teller (but we don't know which one it is).

2 Conclusions

The result holds under more general conditions[1], but the main feature is that the advantage of the single feeder line depends on two factors: the two lines move at different average speeds, and we don't know which is the fastest. This is the main reason why single feeder lines have not been adopted in grocery stores, for example: in principle, looking at the basket of the customers already in line you can guess which line will be the fastest, barring exceptional circumstances.

[1] For example, more general distributions can be chosen for the service times, and we can allow the two lines to have different lengths, for example having the new customer choose the shortest line. If you will take a course in queuing theory you will most certainly get a more thorough discussion of this result