入門講座

やさしい待ち行列(3)——ランダムネスと待ち時間

高橋　幸雄

前回はバスや電車が等間隔運転をすることが、客の待ちをずいぶん減らしていることをみました。逆にいえば、ランダムネスが待ち時間を大きくしているのです。今回は、スタンダードな待ち行列モデルを使って、ランダムネスと待ち時間の関係やいろいろなモデル間の関係などについて考えてみましょう。

1. 待ち行列モデル

待ち行列モデルというのは右の図1~図3のようなものです。行列ができるところは多くの場合このような形でモデル化することができます。

たとえば、スーパーのレジやJRの緑の窓口などは図3のような待ち行列ができますし、銀行の自動預金払戻機コーナーは図2のように行列を1本にしているところが増えました。駅のタクシー乗り場などは図1のようになっているところが多いでしょう。また、一列に並んではいなくても、病院の待合い室、銀行の窓口などのように、実質的に図1や図2のような待ち行列を作っているところもあります。

待ち行列を作るのは人間ばかりではありません。工場では加工を待つ部品やできあがった製品があちこちで行列していますし、高速道路の料金所では車が行列を作ってゲートを通過するのを待っています。

2. 単一窓口モデル M/G/1

まずはいちばん簡単な図1のモデルから考えることにしましょう。待ち行列モデルは、サービスを受ける“客”(図の○印)と客をサービスする“窓口”、それに窓口に入りきらない客が待つための“待ち行列”で構成されています。図1のように窓口がひとつのモデルを単一窓口モデルと呼んでいます。

たかはし ゆきお　東京工業大学 大学院 情報理工学研究科 数理・計算科学専攻
〒152 東京都目黒区大岡山 2-12-1

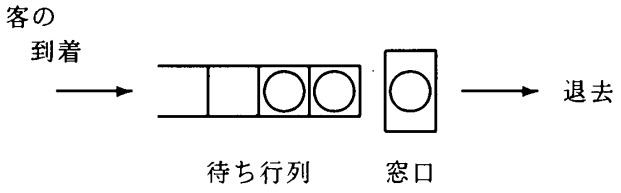

図1: 単一窓口待ち行列モデル

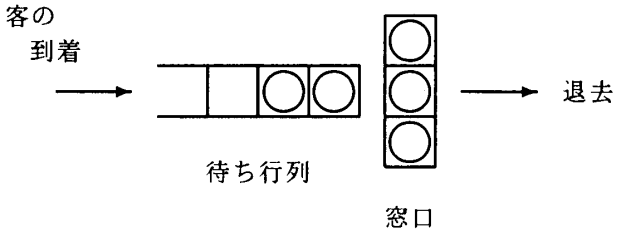

図2: 複数窓口待ち行列モデル

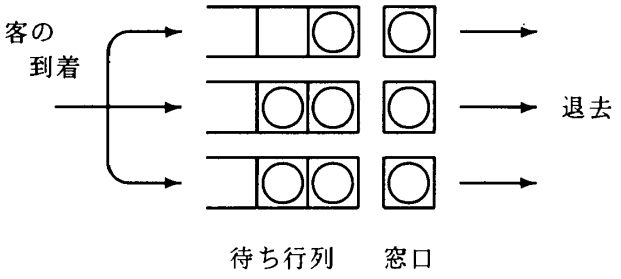

図3: 並列待ち行列モデル

モデルの挙動をきちんと記述するために、客の到着やサービスに関していくつかの仮定が必要です。

客はパラメータλのポアソン過程にしたがって1人ずつ到着するものとしましょう。これは前回お話ししたように、客がランダムに到着する、ということをモデル化したもので、客の到着間隔は平均が$1/\lambda$の指数分布にしたがいます。客のサービス時間はひとりひとり異なることが多いので、互いに独立で同一の分布にしたがう確率変数(列)であるものとします。これらのサービス時間は到着過程とも独立です。

他の客がだれもいないときに到着した客はそのまま窓口でサービスをうけはじめます。他の客のサービ

ス中に到着した客は、待ち行列の最後尾に並んで自分の順番を待たなければなりません。窓口でひとりの客のサービスが終了するとその客はシステムを去り、待ち行列の先頭の客が (もしいれば) 窓口に進んでサービスをうけはじめます。

このように客がサービスをうける順番などを定めたルールを “サービス規律” と呼びます。このモデルのサービス規律は “先着順” です。では最初のクイズ、

> **クイズ 1**
> このような待ち行列モデルでは、サービス時間が一定の場合と指数分布にしたがう場合とでは、どちらが客の待ち時間が長くなるでしょう。

“待ち時間” というのは、客が到着してからサービスが開始されるまで、待ち行列で待っている時間のことです。前回のバスや電車の待ち時間では、ランダムな程度の大きい指数分布の方が平均待ち時間が長くなりました。ここはどうでしょう。

指数分布の場合、たまに非常に長いサービス時間がかかる客が到着して、その後からきた客を長いこと待たせます。そのためこの待ち行列モデルでも指数分布のときの方が一定の場合より平均待ち時間が大きいことが予想されます。しかしどの程度長くなるかはモデルをきちんと解析してみなければわかりません。

解析はすこしやっかいですので 6 節に回して、結果だけ先に見てしまいましょう。S を客のサービス時間を表す確率変数として、$\rho = \lambda \mathrm{E}(S)$ とおきます。$\rho < 1$ ならば、平均待ち時間 W_q はつぎのポラチェック–ヒンチンの公式で与えられます。

$$W_q = \frac{\rho}{1-\rho}\frac{\mathrm{E}(S^2)}{2\,\mathrm{E}(S)} \tag{1}$$

これは $\rho/(1-\rho)$ という係数が掛かっていることを除けば、前回のバスの待ち時間のときと全く同じです。したがってサービス時間が指数分布にしたがうときは一定のときの 2 倍待たされることになります。やはりランダムネスは待ち時間を大幅に増やすのです。

利用率　ρ

式 (1) の係数 $\rho/(1-\rho)$ は、この待ち行列モデルのもうひとつの重要な特徴を示しています。

$\rho = \lambda \mathrm{E}(S)$ と定義しましたが、λ は単位時間当たりに到着する客の数ですので、これは単位時間あたりサービスが行われる時間、つまりサービスされている時間の割合、を表しています。そこでこの ρ を “利用率”

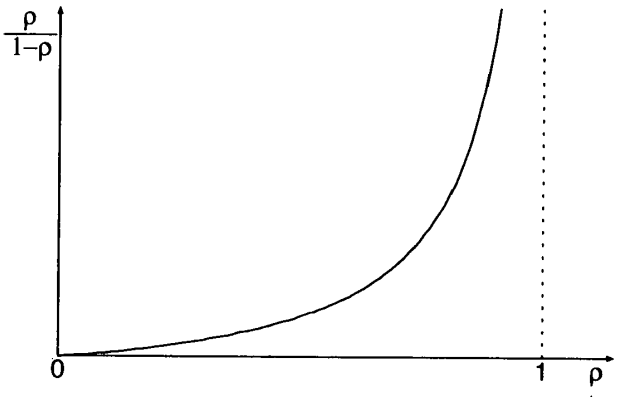

図 4: 利用率と平均待ち時間

と呼びます。T を十分長い時間として、$\rho = \dfrac{T\lambda}{T/\mathrm{E}(S)}$ と書いてみると、分子は T の間に到着する客数、分母は T の間にサービス可能な客数です。したがって

$$\rho = \frac{\text{サービス要求量}}{\text{処理能力}} \tag{2}$$

とも解釈できます。このことから $\rho < 1$ という条件は、必ず必要であることがわかります。もし $\rho > 1$ ならば、サービス要求量に処理能力が追いつかず、待ち行列は際限なく長くなってしまうでしょう。

式 (1) に戻ると、$\rho/(1-\rho)$ という係数から、平均待ち時間は ρ が小さいときはほぼ ρ に比例して増加し、ρ が 1 に近づくと急速に大きくなることがわかります (図 4)。したがって待ち時間を短くするには

i) 到着率 λ を小さくして客の数を減らす
ii) 平均サービス時間 $\mathrm{E}(S)$ を小さくする
iii) サービス時間を一定に近づける

のが有効です。とくにはじめの 2 つは効果的です。

このモデルではポアソン到着を仮定していますが、“再生到着” といって、客の到着間隔が互いに独立で同一の分布にしたがうモデルもあります (4 節)。そのようなモデルを解析してみると、到着間隔のランダムネスも少ない方がよく、

iv) 到着間隔を一定に近づける

とするのも待ち時間を減らすのに効果的であることがわかります。歯医者さんが予約制にしているのは、この手を使っているわけです。

3. 複数窓口モデル M/M/c

つぎに、図 2 の場合を考えてみましょう。今度は窓口が複数あます。このような複数窓口モデルは数学的に複雑で、ポアソン到着、指数サービスというような

ごく特別な場合しか簡単には解析できません。まずはそのごく簡単な場合をみてみましょう。

客はパラメータが λ のポアソン過程にしたがって到着します。窓口の数は c 個で、客のサービス時間はパラメータが μ の指数分布にしたがい、互いに、また到着過程とも独立です。サービス規律は先着順です。

このモデルを使って、たとえば次のような問題を考えることができます。

クイズ 2

処理速度の速い計算機 A と処理速度がその半分の計算機 B があります。もし計算機 B の価格が A の半分だったとしたら、計算機 A を 1 台購入するのと計算機 B を 2 台購入するのとでは、どちらがよいでしょうか。

これをみるには、上のモデルを $c = 1$ の場合と $c = 2$ の場合について解析して、ターンアラウンドタイムに相当する系内滞在時間の平均を求めてみるとよいでしょう。ここで系内滞在時間というのは、客が到着してからシステムを去るまでの時間で、待ち時間とサービス時間の和を指します。

7 節で示すように、マルコフ連鎖の手法を用いると、系内に n 人の客がいる確率 p_n は、系内にだれもいない確率 p_0 を用いて

$$p_n = \begin{cases} \dfrac{c^n \rho^n}{n!} p_0 & 0 \le n \le c \text{ のとき} \\ \dfrac{c^c \rho^n}{c!} p_0 & n \ge c \text{ のとき} \end{cases} \tag{3}$$

と表せます。ただしこの場合の利用率は (2) から $\rho = \lambda / c\mu$ とします。これらの p_n の和が 1 であることを使うと p_0 が決定できて、$\rho < 1$ ならば

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{c^n \rho^n}{n!} + \frac{c^{c-1} \rho^c}{(c-1)!(1-\rho)} \right]^{-1} \tag{4}$$

となります。(3) と (4) から、平均系内人数 L や平均行列長 L_q の式を、さらにはリトルの公式を用いて平均系内滞在時間 W や平均待ち時間 W_q の式を導くことができます。たとえば平均待ち時間は

$$W_q = \frac{c^{c-1} \rho^c}{(c-1)!(1-\rho)^2 \mu} p_0 \tag{5}$$

で与えられ、平均系内滞在時間 W はこの W_q に平均サービス時間 μ^{-1} を加えることによって求められます。

上のクイズのケースを実際に計算してみましょう。計算機 A では $c = 1$, $\mu = 1$、計算機 B では $c = 2$, $\mu = 0.5$ として、平均待ち時間 W_q および平均系内滞在時間 W を計算したのが表 1 です。この表をみると、W_q は計算機 B の方が小さく、ターンアラウンドタイムに相当する W は計算機 A の方が小さいことがわかります。これは次のように解釈できるでしょう。

表 1: 計算機の平均系内滞在時間の比較

	平均待ち時間 W_q		平均系内滞在時間 W	
ρ	計算機 A	計算機 B	計算機 A	計算機 B
0.1	0.111	0.020	1.111	2.020
0.2	0.250	0.083	1.250	2.083
0.3	0.429	0.198	1.429	2.198
0.4	0.667	0.381	1.667	2.381
0.5	1.000	0.667	2.000	2.667
0.6	1.500	1.125	2.500	3.125
0.7	2.333	1.922	3.333	3.922
0.8	4.000	3.556	5.000	5.556
0.9	9.000	8.526	10.000	10.526

計算機 A は窓口が 1 つなので、たまたまサービス時間の長い客がくると、その後ろの客は長いこと待たされます。計算機 B は窓口が 2 つなので、ときにサービス時間の長い客がひとつの窓口を塞いでも、もう一方の窓口が客をさばいて、客の待ち時間はそう大きくはなりません。これが複数窓口の効果です。

ただ、サービス速度が計算機 A の方が 2 倍速いので、サービス時間の長い客が窓口を塞ぐ時間は半分ですし、自分自身のサービス時間も半分になります。結局、この効果が待ち時間の長さをカバーして、平均滞在時間を計算機 B よりも短くしてしまうのです。

このようにターンアラウンドタイムを考える限り、速いマシンを買った方が得、とくに ρ が小さいときはずっと得、ということがわかります。実際には遅いマシンの方が価格も安いでしょうし、信頼性などの点からも計算機 B の方がよいと判断されることも多いかと思います。

4. ケンドールの記号

待ち行列理論では、図 1 や図 2 のような標準的なモデルを統一的な記号で表す習慣があります。これは“ケンドールの記号”と呼ばれるもので、A/B/c という形で書かれます。ここで A は到着間隔の分布を、B はサービス時間分布を、c は窓口の数を表します。

A や B としてはつぎのような記号を用います。

M : 指数分布 (Markov)

D : 一定分布 (単位分布) (Deterministic)

PH : 相型分布 (PHase-type)

G : 一般分布 (General)

たとえば2節で扱ったのはM/G/1モデルでしたし、3節で扱ったのはM/M/cモデルでした。ケンドールの記号で表されるモデルは、到着間隔が互いに独立で同一の分布にしたがっていること (再生到着) を暗黙のうちに仮定しています。最近では到着過程としてもう少し一般的な点過程を考えることも行われています。

じつは基本的なモデルであるM/G/cモデルはまだ数学的に解析ができていません。これを近似的にでも扱おうと、近似式と数値計算法が研究されました。

M/G/cモデルの平均待ち時間に対しては、リー–ロントンの近似式と呼ばれる有名な近似式があります。

$$W_q(M/G/c) = \frac{\mathrm{E}(S^2)}{2\,\mathrm{E}(S)} W_q(M/M/c) \qquad (6)$$

ここで$W_q(M/M/c)$は、同じ利用率ρをもつM/M/1モデルの平均待ち時間です。この式は (1) のある種の一般化になっていますが、この式をベースにいくつもの近似式が提案されています。

また数値解析法の研究もはじまりました。待ち行列モデルが数学的に解析できなくても、計算機を用いてその特性量が計算できるアルゴリズムがあれば、モデルが解けたと考えてもいいのではないか、という考えに基づくものです。そして相型分布 (PH) が提案され、さらにはPH/PH/cモデルを効率よく計算するアルゴリズムなども開発されています。

相型分布というのは、大雑把にいえば指数分布を組み合わせて作られる分布で、どんな正の確率分布も理論的には望みの精度で近似することができます。今では、実用上必要とされるほとんどの標準型待ち行列モデルは、適当なPH/PH/cモデルで近似することによって数値的に解析できるようになってきています。

5. 並列待ち行列モデル

今度は図3の並列待ち行列モデルを考えてみましょう。簡単のため、ここでも客はパラメータがλのポアソン過程にしたがって到着し、平均が$1/\mu$の指数分布にしたがってサービスされるものとします。

客が到着したとき、その客はもっとも短い待ち行列に並ぶことにします。そして、待っている客の移動の違いによって、つぎの2つのモデルを考えます。

モデルI : 一度並んだら、待ち行列は変えない

モデルII : サービス中の客プラス待ち行列の長さに2以上の違いができたら、長い方の待ち行列の最後尾の客は短い方へ移る

モデルIは高速道路の料金所などでよくみられるもので、一度並んでしまうともう隣へ移るわけにはいきません。これに対してモデルIIはスーパーのレジなどでみられるもので、客は自由に待ち行列を行ったり来たりできます。

ところで、モデルIIは図2の標準型M/M/cとかなり似ています。ここでは図2のモデルをモデルIIIと呼んで、これら3つのモデルを比較してみましょう。

クイズ 3

モデルI、II、IIIでは、平均待ち時間W_qはどれが短く、どれが長いでしょう。

モデルIIとIIIでは、待っている客がひとりでもいれば窓口は必ずサービスを行います。ところがモデルIでは待ち行列を移ることができないので、隣の窓口が空いていても自分の待ち行列で待たなければなりません。そのため、平均待ち時間はモデルIで少し長くなります。解析してみると、モデルIIとIIIでは平均待ち時間は同じになります。では

クイズ 4

モデルIIとIIIでは、客の待ち時間はどう違うのでしょう。

モデルIIIでは客の到着順にサービスが行われますが、モデルIIでは運の悪い客はあちこちの待ち行列を渡り歩くことになります。ときには後からきた客の後ろになることさえあります。そのため運の悪い客とよい客の差が大きくなります。これは待ち時間の分散に反映されるはずです。

具体的に数値でこのことを見てみましょう。表2は、窓口の数が$c=3$、平均サービス時間が$1/\mu=1$、利用率が$\rho=\lambda/3\mu=0.8$のケースの結果です。W_qが平均待ち時間でV_qが待ち時間の分散です。実際、W_qはモデルIで少し大きく、モデルIIとIIIでは同じです。またV_qはモデルIIの方がモデルIIIより少し大きくなっています。

この表からみると、3つのモデルの中ではモデルIIIがもっとも優れていることがわかります。したがって、これが標準型モデルとして選ばれているのです。ただしこれは待ち行列から窓口へ進む時間が無視できる

表 2: 並列待ち行列モデル : 待ち時間の平均と分散

	モデル I	モデル II	モデル III
W_q	1.286	1.079	1.079
V_q	3.677	2.974	2.432

ときの話です。この時間が多少ともかかるような場合には、モデル II の方がよくなります。

銀行の ATM でモデル III を採用しているのは、ATM 機へ進むのに多少時間がかかっても、客の間の公平性を保ち、さらには ATM を操作している客の暗証番号が後ろの人から見えないようにするのに効果的だからです。スーパーのレジでは、サービスの効率性がもっとも重んじられるので、モデル II を使っているのでしょう。高速道路の料金所がモデル I になっているのは、自動車が横方向には動けないことによるものです。このようにしてみると、それぞれ理由があって待ち行列を作っていることがよくわかります。

6. ポラチェック–ヒンチンの公式

最後に、待ち行列の解析がどのように行われているのか、その簡単な例を 2 つ紹介しましょう。数学が嫌いな方は読み飛ばしていただいても結構です。

第 1 回の累積到着·退去曲線とリトルの公式、前回のポアソン過程とバスの平均待ち時間の公式を使って、M/G/1 の平均待ち時間 W_q を与えるポラチェック–ヒンチンの公式 (1) を証明することができます。

ポアソン過程というのは、区間 $(0, n/\lambda)$ の中に n 個の点（これが客の到着時刻に相当します）を互いに独立に一様分布にしたがってばらまいた確率過程の $n \to \infty$ とした極限でした。そこで、われわれのモデルでも客はこのようにして到着するものと考えて、とくに最後の n 個目の点に相当する客に注目します。

図 5 は、この n 番目の客を除いた $n-1$ 人の客がシステムに到着してサービスを受け、出ていったときの累積到着·退去曲線を表しています。横線のハッチが入っているところが待ち行列で待っている部分です。

n 番目の客が到着したときに系内にいた客の数 (サービス中の客と待ち行列で待っていた客の合計) を図 5 のように確率変数 N で表すことにしましょう。さらに 2 節と同様に客のサービス時間を表す確率変数を S、利用率を $\rho = \lambda \mathrm{E}(S) < 1$ とします。

このとき図 5 からつぎのことがわかります。n 番目

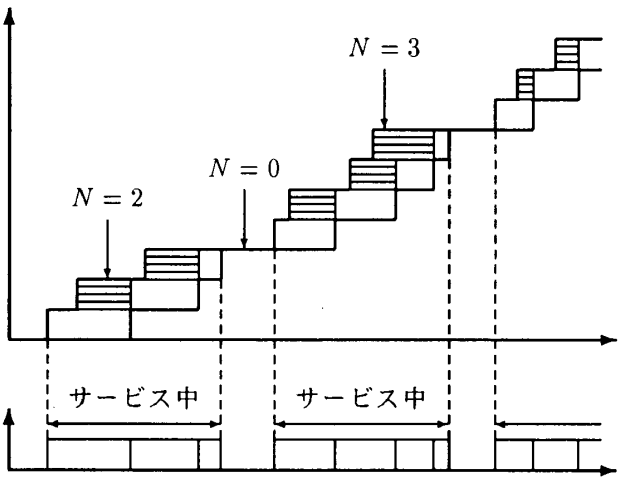

図 5: M/G/1 モデルにおける累積到着 · 退去曲線

の客の到着時点において

a) $N = 0$ となる確率は $1 - \rho$

b) $N > 0$ ならば、サービス中の客の残りサービス時間の平均は $\mathrm{E}(S^2)/2\,\mathrm{E}(S)$

a) は、長さ n/λ の区間の中で $n-1$ 人の客をサービスするので、$\mathrm{P}(N > 0)$ は $n \to \infty$ のとき ρ となることからわかります。また b) は、ある客のサービス中に到着したという条件から、前回にお話ししたバスの平均待ち時間と同じ構造になることを利用して導けます。

この n 番目の客の (Nによって条件づけられた) 平均系内滞在時間は、$N = 0$ のときは $\mathrm{E}(S)$ (自分のサービス時間の平均)、$N = l > 0$ のときは $\mathrm{E}(S^2)/2\,\mathrm{E}(S) + l\,\mathrm{E}(S)$ (サービス中の客の残り平均サービス時間と、待っている客および自分自身の l 人分の平均サービス時間、の合計) となります。これらに確率 $\mathrm{P}(N = l)$ を掛けて加えると、n 番目の客の平均滞在時間は

$$\begin{aligned} W &= \mathrm{E}(S)\,\mathrm{P}(N=0) \\ &\quad + \sum_{l=1}^{\infty} \left[\frac{\mathrm{E}(S^2)}{2\,\mathrm{E}(S)} + l\,\mathrm{E}(S) \right] \mathrm{P}(N=l) \\ &= \rho \frac{\mathrm{E}(S^2)}{2\,\mathrm{E}(S)} + (1 - \rho + \mathrm{E}(N))\,\mathrm{E}(S) \qquad (7) \end{aligned}$$

となります。ここで $\mathrm{P}(N=0) = 1-\rho$、$\sum_{l=0}^{\infty} \mathrm{P}(N = l) = 1$、$\sum_{l=0}^{\infty} l\,\mathrm{P}(N=l) = \mathrm{E}(N)$ を使っています。

$\mathrm{E}(N)$ は、$n \to \infty$ の極限では平均系内客数 L と一致しますし、W は一般の客の平均系内滞在時間と同じはずです。そこでリトルの公式 $L = \lambda W$ を使って $\mathrm{E}(N)$ を λW で置き換えると、(7) は W に関する方程式になります。これを解くと

$$W = \frac{\rho}{1-\rho} \frac{\mathrm{E}(S^2)}{2\,\mathrm{E}(S)} + \mathrm{E}(S) \qquad (8)$$

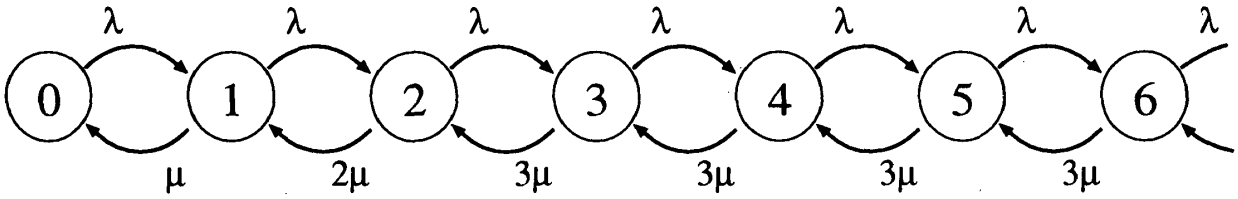

図 6: M/M/3 モデル：状態推移図

となり、これに平均待ち時間と平均系内滞在時間の関係 $W = W_q + \mathrm{E}(S)$ を使うと 2 節の (1) が導けます。

7. マルコフ連鎖と M/M/c の解析

指数分布の性質を巧みに使った数学的モデルに“時間連続的なマルコフ連鎖”があります。これは、たかだか可算個の“状態”の間をシステムを表す“粒子”が動き回るモデルで、おのおのの状態での滞在時間は与えられたパラメータの指数分布にしたがい、滞在時間が終了するとそれぞれ与えられた確率で他の状態へ推移する、というものです。

指数分布は“無記憶性”という特殊な性質を持っていることを前回お話ししましたが、もうひとつ、互いに独立な確率変数の minimum の分布がやはり指数分布になる、という特徴があります。

いま X と Y を互いに独立でそれぞれパラメータが α と β の指数分布にしたがう確率変数として、$Z = \min\{X, Y\}$ とすると、Z の分布はパラメータが $\alpha+\beta$ の指数分布になります[1]。さらに、$Z = z$ という条件のもとで X の方が min をとるという条件付き確率は $\alpha/(\alpha+\beta)$ です[2]。これらの性質は 3 個以上の確率変数の場合にも容易に拡張できます。

このことを念頭に置いて、図 2 の系内人数 5 人が、つぎに 4 人になったり 6 人になったりする時間や確率について考えてみましょう。

系内人数が変化する要因としては

a) 客がひとり到着する

b) 窓口 1 でサービスが終了する

c) 窓口 2 でサービスが終了する

d) 窓口 3 でサービスが終了する

の 4 つが考えられます。指数分布は無記憶性を持っていますから、それぞれが起こるまでの時間はみな指数分布にしたがっていて、しかも独立です。すると上の最小値の分布の議論から、どれかが起こるまでの時間はパラメータが $\lambda+3\mu$ の指数分布にしたがい、それが客の到着である確率は $\lambda/(\lambda+3\mu)$、いずれかの窓口におけるサービス終了である確率は $3\mu/(\lambda+3\mu)$ であることがわかります。これは上で述べたマルコフ連鎖の枠組みにぴったり当てはまります。

マルコフ連鎖の確率的構造は図 6 のような推移図で表現されます。円の中の数字が系内人数を表す状態のラベルで、矢印につけられた λ や 3μ などは、そのような推移が起こるまでの時間分布のパラメータです。状態 2 から 状態 1 への矢印には 3μ でなく 2μ がついているのは、系内に 2 人しか客がいないときには、2 つの窓口でしかサービスが行われていないことを反映しています。

このようなマルコフ連鎖では、定常状態確率 (十分時間がたったときの各状態にいる確率) p_n が存在して、それらは確率をあたかも水の流れのように考えて、そのフローバランスから求められることが知られています。とくに図 6 のような形をしているときは、隣あった状態間のフローバランスが成り立っていて、たとえば状態 5 から状態 6 へ短い時間 dt の間に流れる確率の量は $\lambda p_5\, dt$、状態 6 から状態 5 へ流れる量は $3\mu p_6\, dt$ となります。これらを等しいとおくと

$$\lambda p_5 = 3\mu p_6 \quad \text{より} \qquad p_6 = \frac{\lambda}{3\mu} p_5 = \rho p_5$$

となります。このような式を各状態の組にたいして求めると、(3) の式が得られるのです。

待ち行列理論の書物はたくさんありますが、ここでは代表的なものとして、下の [1] を紹介しておきます。本稿でも計算や図の一部は大学院生の大原久樹君にお願いしました。次回は待ち行列のコントロールと待ち行列ネットワークについて解説する予定です。

参考文献

[1] 森村英典, 大前義次,「応用待ち行列理論」, 日科技連, 1975.

---

[1] $\mathrm{P}(Z > z) = \mathrm{P}(X > z \text{ and } Y > z)$
$= \mathrm{P}(X > z)\,\mathrm{P}(Y > z) = e^{-\alpha z} e^{-\beta z} = e^{-(\alpha+\beta)z}$

[2] $\mathrm{P}(X < Y \mid z < Z < z + dz)$
$= \mathrm{P}(z < X < z + dz, Y > z + dz) / \mathrm{P}(z < Z < z + dz)$
$= \dfrac{\alpha e^{-\alpha z}\, dz\, e^{-\beta z}}{(\alpha+\beta) e^{-(\alpha+\beta)z}\, dz} = \dfrac{\alpha}{\alpha+\beta}$