待ち行列理論の応用：コールセンターを例に

河西　憲一

待ち行列理論はその名が示すように日常生活で経験する待ち行列についてなにがしかの実益に即した知恵を授けてくれるような期待感を持たせてくれる．一方で待ち行列理論は，混雑現象一般を数理の立場から解明することを指向した応用数学あるいは応用確率論として捉えることもでき，数学的にも洗練されたその内容を理解するためにはやや障壁が高い感があるのも否めないのではないだろうか．誕生からおよそ100年が経過した現在でもその理論の奥の深さを伝えてくれるが，本稿では待ち行列理論が誕生した頃に回帰して，初期の成果が現代的なシステムにどのように生かされているかについて紹介する．また，その有効性についても検討する．

キーワード：待ち行列理論，コールセンター，ポアソン過程，指数分布，定常解析

1. はじめに

サービスを提供するところに待ち行列ができることは我々の日常生活でよく見られる光景である．スーパーのレジでの待ち行列しかり，昼時の（繁盛している？）飲食店前の行列しかりである．病院の待合室も整然とは並んでいないが，患者が診察の順番を待っていることには違いない．人が列をなすとか，大勢待機しているだけが混雑や待つことの現れではない．目にはしかと映らないけれども，明らかに混み合っていると感じられる場合もある．通信に関係するシステムがよい例である．その中でもコールセンターは混雑の度合いに応じてつながりにくくなったり，また待ち時間が長くなったりするため，混雑を感知しやすいシステムともいえる．

このような混雑に伴う現象を分析する道具として待ち行列理論が考えられてきた．本稿では待ち行列理論の標準的な教科書に掲載されている成果がどの程度現実のシステムに使えるのかについて，コールセンターを例に探っていく．もともと待ち行列理論が誕生した背景には電話交換機の適切な設備数について答えを得ること，すなわち**容量設計**が目標にあった．幸いにして，コールセンターは電話を主体としたシステムであるので待ち行列理論の成果が生かされ，また生かされやすい具体的な例と思われる．

なお，コールセンターとは企業が顧客からの問い合わせなどを受け付ける拠点として設置する専門部署である．近年は顧客満足度を向上させるための中心的戦略拠点としてコールセンターを捉える傾向があり，頻繁に混雑するコールセンターを放置することは企業にとって看過できない．快適なコールセンターの構築に，初歩の待ち行列理論が果たす役割は決して軽くない．

2. コールセンター：概観

本稿で想定するコールセンターとは，顧客がコールセンターの外部から電話を通じてサービス（例えば，航空券の予約など）の提供を求める，インバウンド型のコールセンターとする．このようなコールセンターを待ち行列システムとして捉えるときの構成要素について整理しておく．まず，コールセンターには外部からかかってくる電話をつなぐ（大抵の場合複数の）電話回線が用意されている．この電話回線のことを本稿では**外線**と呼ぶことにする．外線はPBX（Private Brach eXchange）と呼ばれる構内交換機に収容されている．コールセンターに電話をかける顧客はこの外線のうち空いている電話回線を通じてまずPBXにつながることになる．つながった顧客からの要求は初めにIVR（Interactive Voice Response）と呼ばれる音声自動応答装置に接続されて処理されることが多い．このIVRは，例えば「フライトスケジュールを知りたい場合はボタン1を押してください」などの定型ガイダンスを自動的に流し応答する装置と考えればよい．

かわにし　けんいち
群馬大学 大学院工学研究科
〒376-8515 桐生市天神町1-5-1

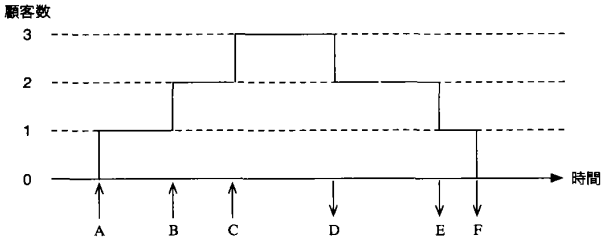

図1　顧客数の時間変化

IVRでは対応しきれない要求はコールセンターの応対者と実際に話しながら電話応対を通じて処理される．コールセンターで電話応対業務に従事する応対者のことを本稿では**エージェント**と呼ぶことにする．エージェントの人数はコールセンターの規模にもよるが，数十人から多い場合は百人を超える程度にも及ぶことも珍しくない．PBXにつながった顧客は内線を通じてエージェントと接続される．どのエージェントに接続するかを決める際に，ACD（Automatic Call Distributor）と呼ばれる特殊な装置を使う．ACDはいわば顧客の要求をエージェントに分配する役割を果たし，様々な基準に基づいてエージェントを選択することが可能である．一人の顧客に対してエージェント一人で要求が完結することもあればそうでない場合もある．完結しない場合，さらに別のエージェントに接続される．その際にもACD機能を使って分配することが可能である．最初に受け付けるエージェントの一群を一次受付，次のエージェント群を二次受付のように呼ぶこともある．

3.　コールセンターの待ち行列解析

前節ではコールセンターの構成要素を概観した．現代のコールセンターは機能も豊富で様々な装置から構築されている複雑なシステムであることが伺える．このような複雑なシステムに待ち行列理論の考え方を適用し，コールセンターに繋がる確率だとか，つながってからの待ち時間の期待値などを評価するのが本節の目的である．

待ち行列理論の考え方を説明する前に，まずコールセンター内に滞在する顧客数の時間変化を観察してみよう．例えば，図1のような場合が考えられるであろう．図中のA，B，Cは顧客がコールセンターに電話をかけてつながった時点，すなわちコールセンターに到着した時点を表す．逆にD，E，Fは顧客が電話を切ってコールセンターから退去した時点を表す．当然のことではあるが，図1から，到着時点では顧客数は増加し，退去時点では減少することが読み取れるであろう．また，到着時点が密集していれば短い期間内で急激に顧客数が増大し，逆にまばらに散っていれば顧客数が少ないまま推移することが多いと考えられる．退去時点についても，例えばEとFの時点が共によりDの時点に近ければ，より早く顧客数はゼロになるであろう．このような定性的な把握ではなく定量的な理解をするために，待ち行列理論では到着時点の間隔や退去時点の間隔について確率的な挙動を仮定することが多い．その中でもはじめの第一歩と目されるのが，ポアソン到着と指数サービス時間の仮定である．

「ポアソン到着」とは，顧客の到着時点が（定常）ポアソン過程と呼ばれる確率過程にしたがうことを意味する．この場合，図1のAとBの間隔やBとCの間隔など，相隣り合う到着時点の間隔を表す確率変数Xが

$$\Pr\{X \leq t\} = 1 - e^{-\lambda t} \tag{1}$$

にしたがうと仮定することになる．ここで，$\lambda > 0$は顧客の到着率と呼ばれ，単位時間当たりに到着する顧客数の平均を意味する．さらに「指数サービス時間」とは，顧客の一人ひとりがコールセンターから受けるサービスの期間がある確率変数Sで表すことができ，その確率分布が

$$\Pr\{S \leq t\} = 1 - e^{-\mu t} \tag{2}$$

で与えられることを意味する．指数分布の平均を思い出せば，顧客のサービス時間の期待値は$1/\mu$で与えられることになる．また，μは単位時間当たりにサービスが完了する処理率（サービス終了率）とも捉えることが可能である．コールセンターの場合，サービスの期間の具体的な例としてはエージェントとの会話時間を念頭に置けばよいであろう．

顧客の到着過程とサービス時間を確率的に決めることは待ち行列理論の手法を適用する上で必要ではあるが十分ではない．コールセンターの場合，エージェントの人数であるとか，外線数であるとか，さらには顧客をサービスする場合の順序であるとか，これらの条件によってシステムの振る舞いが異なるからである．本稿では，エージェントがc人存在し，外線がN本用意されているコールセンターを考える．顧客はコールセンターにつながった時点で一人でもエージェントが空いていれば直ちにサービスを受け，そのサービス終了後に他のエージェントから別にサービスを受けることなくコールセンターから退去すると仮定する．空きエージェントがない場合は到着の順序通りに（先着

順に）待つ．ただし，すべての外線が使われているときはコールセンターにつながらないまま直ちに退去するものとする．Nがcに等しいか大きいと仮定すると，待つことができる顧客の最大人数は$N-c$人である．

このようなモデルは待ち行列理論では$M/M/c/N$システムとして知られている．仮にコールセンターに電話をかけてエージェントにつながったとしよう．このとき，顧客は外線を保留すると同時にまたエージェントも「保留」していると言える．この同時保留する点が$M/M/c/N$システムでは上手く表現されている．

3.1 定常分布と率保存則

図1からも分かるように，コールセンター内に滞在している顧客数は時々刻々変化する．時間変化をつぶさに追跡し，ある時刻でコールセンターにつながらない確率を解析することも原理的には可能であるが，実際的な待ち行列解析では十分時間が経過し安定した定常状態に着目することが多い．定常状態での顧客数の確率分布を**定常分布**と呼ぶことにする．定常分布は必ず存在するわけではないが，ここでは仮に存在することを前提とし，p_nによってコールセンター内の顧客数がn人（ただし，$0\leq n\leq N$）である定常分布を表すとする．

ひとたび定常分布の存在を認めてしまえば，その値を計算することは容易である．定常分布については

$$\lambda p_n=\min(n+1, c)\mu p_{n+1}, 0\leq n<N, \tag{3}$$

が成立し，正規化条件$\sum_{n=0}^{N}p_n=1$とあわせて逐次的に解くことにより一意に定まってしまうからである．式(3)は定常状態において成立する**平衡方程式**をよりどころとし，左辺と右辺とが拮抗し釣り合うことを意味する（図2参照）．

左辺は定常状態において顧客数がnである状態から$n+1$へ流出する単位時間あたりの確率の流れ（あるいは確率流，確率フローとも）を意味する．実際，顧客数がnである定常分布がp_nであり，顧客が到着率λで到着すると人数が一つ増えて$n+1$になることに注意すれば了解できるであろう．同様に右辺は顧客数が$n+1$である状態からnへ流出する確率の流れを示している．この場合は顧客数が減るので到着率ではなく退去率を考えなければならない．顧客の退去はサービスが終了することにより生じる．退去率はサービスを受けている顧客数に応じて変わり，式(3)の右辺の$\min(n+1, c)\mu$はサービスを提供しているエージェントの人数倍の処理速度で顧客が退去することを反映している．このような直感的な説明はサービス時間が指数分布にしたがう場合に論拠立てることができる[1]．なお，式(3)はマルコフ過程における詳細釣り合いの式として古くから知られ，また今日では一般的な条件で成立する**率保存則**[2]からも理解できる．

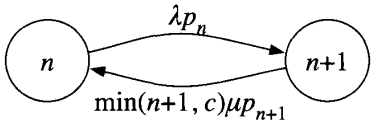

図2 状態遷移

3.2 性能評価指標

コールセンターを$M/M/c/N$システムとして捉え，その定常分布が分かるとシステムの性能評価指標を評価することができる．例えば，式(3)から定常状態において顧客の数がNに等しい確率，すなわちコールセンターが満杯である確率p_Nが次のように与えられる．

$$p_N=\frac{a^N}{c^{N-c}c!}p_0. \tag{4}$$

ただし，$a\equiv\lambda/\mu$であり，また

$$p_0=\left[\sum_{n=0}^{c-1}\frac{a^n}{n!}+\sum_{n=c}^{N}\frac{a^c}{c^{n-c}c!}\right]^{-1}, \tag{5}$$

である．確率p_Nは定常分布であり，したがって任意の時点で観測した場合に顧客数がNとなる確率に等しくなる．直感的には非常に長い時間コールセンターを観測し，その観測時間に対して顧客数がNとなる時間の比とも捉えることができる量である．このことから確率p_Nは時間平均という側面も併せ持つ．時間平均としての定常分布はコールセンターを運用する側の視点に立った評価指標である．例えば，p_Nはコールセンターの外線がすべて使われている割合（稼働率）と考えられ，この値が高ければ効率的といえる．

外線の稼働率は無視できない評価指標ではあるものの，コールセンターを利用する顧客の視点からはややかけ離れている．なぜならば，顧客は電話をかけてみて初めてコールセンターの様子（つながるのかつながらないのかなど）が分かるはずだからである．ということは電話をかけた時点，すなわち顧客の到着時点で観測したときの状態が分かればありがたい．幸いなことにポアソン到着の場合はPASTA (Poisson Arrivals See Time Averages)[3]と呼ばれる性質が成り立ち，到着時点の確率分布は定常分布に等しい．したがって，p_Nは顧客がコールセンターに電話をかけても待ち室に入れず，つながることもできない確率

表1　$\Pr\{W_q \leq 20\}$ と p_N（括弧内）の数値例

	$N=18$	$N=19$	$N=20$
$c=12$	80% (2.9%)	78% (2.3%)	76% (1.9%)
$c=13$	90% (1.9%)	88% (1.4%)	87% (1.1%)

（損失率）を与える．

たとえコールセンターにつながったとしても，即座にエージェントが応答するとは限らず，待たされる場合もある．$M/M/c/N$ システムは顧客の待ちについても情報を与えてくれる．実際，先着順でサービスを受ける顧客の**待ち時間** W_q の確率分布について，

$$\Pr\{W_q \leq t\} = 1 - \sum_{n=c}^{N-1} q_n \sum_{i=0}^{n-c} \frac{(c\mu t)^i}{i!} e^{-c\mu t}, \tag{6}$$

のように表すことができる．ただし，q_n はつながった顧客がコールセンター内に n 人の顧客を見いだす確率であり，$q_n = p_n/(1-p_N), 0 \leq n < N$ で与えられる．

ここで，$M/M/c/N$ システムに基づいてコールセンターの容量設計について検討してみよう．1時間に平均300人の顧客から問い合わせがあり，エージェントとの応対時間が平均2分であるようなコールセンターを考える．設計基準として応答時間80/20ルール，すなわち「20秒以内に応答する確率が80%以上」がよく使われる．表1から，エージェントを12人配置すると18回線であれば応答時間80/20ルールを満足することが分かる．しかしながら，損失率はおよそ3%に上ることも見て取れる．損失率の基準としては概ね1%を目標とするのが標準的であり，この場合回線数が不足気味である．一方で，20回線用意すれば損失率が1%台に下がることが分かる．さらに，エージェントを一人増やせば損失率がほぼ1%に等しくなり，かつ応答時間80/20ルールも満足し，妥当な線といえよう．

4. 現実との整合性

$M/M/c/N$ システムはコールセンターの特徴をよく捉えているものの，実際にその前提が成立するか吟味する必要がある．まず，顧客の到着がポアソン過程にしたがうことが要求される点に注意しよう．ポアソン過程の数学的な定義は専門書[1]などに譲ることにして，大雑把にその性質を表現すれば「でたらめに発生する」であろうか．潜在的な顧客数が非常に多く，いつどの顧客が電話をかけてくるか全く予想できない場合が想定できればポアソン過程はよいモデルといえる．

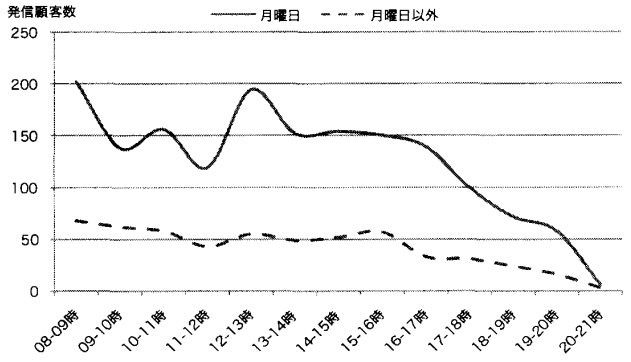

図3　仮想トラヒック

サービス時間が指数分布にしたがうかは要検討事項である．周知のように，平均が $1/\mu$ の指数分布の確率密度関数は $\mu e^{-\mu t}, t \geq 0$ であり，その最大値は $t=0$ で与えられ，t が大きくなるにつれて確率密度関数の値は急速に減少する．コールセンターでのサービス時間をエージェントとの応対時間とするならば，いわば短い時間で完了する顧客が一番多くて，長い問い合わせになればなるほどその割合が激減する場合になる．実際の応対時間のデータと見比べてみると必ずしも指数分布が適当であるとはいえず，むしろ対数正規分布と呼ばれる確率密度関数がよいとも報告されている[4]．

定常性が認められるかどうかも検討の対象となる．定常性の可否を左右する大きな要因の一つが顧客の到着率である．そもそもポアソン過程では到着率が時間に依存せず一定であることが暗に仮定されている．到着率は単位時間当たりに到着した顧客数であることを思い出せば，定常性が現実的に成立するか疑問を抱いてもおかしくない．例えば，時間ごとに電話をかける顧客数が変化することは容易に考えられるだろう．「週明けの月曜日はコールセンターの受付が始まる時間帯で問い合わせが多く，その後いったん収まるもののお昼休み直後に再びピークを迎える．月曜日以外は午前に問い合わせが多いものの緩やかに減少しながら比較的落ち着いて推移する（図3参照）」このようなシナリオは電話をかけるという行為が我々の日常活動と関連していることから想像に難くない．この場合，ピーク時の到着率を基準に設計することは実際的な処方箋でもある．

5. おわりに

コールセンターを例に待ち行列理論の応用を述べた．この小文が待ち行列理論に興味を抱くきっかけになれば幸甚である．

参考文献

[1] 高橋敬隆, 山本尚生, 吉野秀明, 戸田彰, わかりやすい待ち行列システム—理論と実践—, ㈳電子情報通信学会, 2003.

[2] M. Miyazawa, "Rate conservation laws: A survey," *Queueing Systems*, **15**, 1-58 (1994).

[3] R. W. Wolff, "Poisson arrivals see time averages," *Operations Research*, **30**, 223-231 (1982).

[4] N. Gans, G. Koole and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing and Service Operations Management*, **5**, 79-141 (2003).