データを累積すると見えてくるもの

逆瀬川浩孝

1. はじめに

データはモデル分析の鍵を握っています．いきなりいくつかの特性量に集約する前に，まずはおおざっぱに全体としての様子を調べるために，累積してみる，ということがいろいろな情報を得る上で，有効な手段であることを，いくつかの適用例を通して調べます．

2. 経験分布関数

2.1 コルモゴロフスミルノフ検定

標本調査のように，ある母集団の性質を調べるためにデータが集められる，という状況を考えます．データの特徴をつかむために，平均分散などの特性量を計算してみるのも良いのですが，短兵急に情報を縮約しようとせずに，全体を眺めてみることも必要です．データ $x_1, x_2, \cdots, x_n$ を大きさの順に並べたものを $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ とすると，$x_{(1)}, x_{(2)}, \cdots, x_{(n)}$ で n 分の1ずつ増えていく階段関数を経験分布関数といいますが，それは n が十分に大きければ母集団の累積分布関数，に近づくことが保証されています．

このことを利用すると，あるデータセットから作られた経験分布関数とある分布の累積分布関数との「差」がある値以下ならば，その累積分布関数はデータセットの母集団分布といって良い，ということになりますから，「差」をうまく計量化すれば，それを検定統計量として使うことができます．この検定法がコルモゴロフスミルノフ検定です．

2.2 確率プロット

二つの曲線（折れ線）が似ているかどうかを判断するのは難しいので，それを視覚的に確認する方法を考えます．n 個のデータが，ある累積分布関数 $F(x)$ をもつ母集団からの標本かどうかを確かめる，という状況を想定しましょう．昇順に並んだデータの i 番目の値 $x_{(i)}$ は全体で下から $100i/n\%$ の位置にあるので，もしデータが $F(x)$ からのサンプルであったとすれば，$z_i=F^{-1}((i-0.5)/n)$ と同じようなばらつきをしているはずです．したがって，$\{(x_{(i)}, z_i), i=1, 2, \cdots, n\}$ をプロットすると，散布図は直線上に分布することが期待できます．直線かどうかならば簡単にチェックできます．この散布図を確率プロット，あるいはQ-Qプロットといいます．Qは分位点 Quantile の略です．

例えば，一様分布に従うサンプル20個を使ってそれらが正規母集団からの標本と見なしてよいかどうかを調べるために確率プロットを描いたのが図1です．$\{z_i\}$ の方が裾の方で間隔が広がってしまうため，散布図から直線を見いだすことには無理があります．

正規母集団に対するQ-Qプロットで縦軸の目盛り a の代わりに $F^{-1}(a)$ としたもの，ただし，$F^{-1}(a)$ の切りの良い数を目盛りにしたもの，は正規確率紙として知られ，視覚的に正規性を検証でき，それに加えてだいたいの平均標準偏差を見積もることができる優れものとして知られています．

同じような考え方でP-Pプロットという方法もあります．これは経験分布関数の値と，仮説母集団の累積分布関数の値の組み合わせが直線傾向をもっているかどうかを調べる方法です．$p_i=F(x_{(i)})$ として，$\left\{\left(\frac{i-0.5}{n}, p_i\right), i=1, 2, \cdots, n\right\}$ をプロットしたものがP-Pプロットです．

累積分布関数が決まりさえすれば，コンピュータを使って簡単にグラフを描くことができますから，いろいろな累積分布関数を使って，このような累積グラフを描いて，視覚的な判断を取り入れることは，選択するモデルの幅を広げることに役に立つでしょう．

3. 不平等曲線

3.1 ローレンツ曲線

データが正値の場合，データの値そのものを累積したグラフを描くと，別の特徴を見ることができます．データを大きさの順に並べるまでは同じですが，そこ

さかせがわ　ひろたか
早稲田大学 理工学術院
〒169-8555 新宿区大久保3-4-1

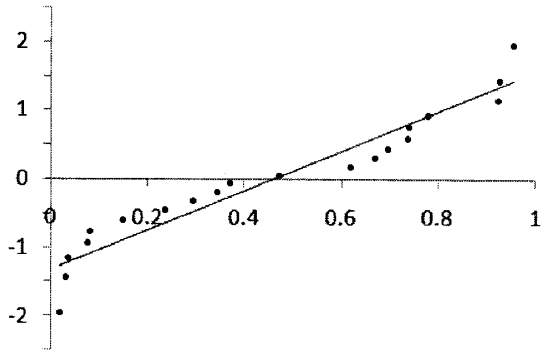

図1　Q-Qプロット実施例

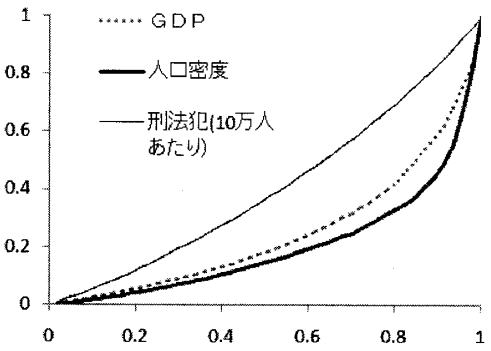

図2　都道府県別統計量の「集中」と「分散」

から，部分和 $A_k=x_{(1)}+x_{(2)}+\cdots+x_{(k)}$ を計算して，$(k/n, A_k/A_n)(k=0,1,2,\cdots,n)$ を折れ線で結んだものは，ローレンツ曲線として知られたグラフになります．

この曲線は，原点と (1,1) を通り，単調増加で下に凸になります．データのばらつきが大きくなると，曲線の立ち上がりが遅くなるので，彎曲度が大きくなる，という特徴があります．

一般に，密度関数を $f(x)$，累積分布関数を $F(x)$ としたとき，

$$L(x)=\frac{\int_0^{F^{-1}(x)} xf(x)dx}{\int_0^{\infty} xf(x)dx} \quad (0<x<1)$$

をローレンツ曲線といいます．データを使った折れ線グラフは，累積分布関数として経験分布関数を当てはめたものです．もともと，20世紀初頭の経済学者ローレンツが所得の不平等を分析したときに見いだした規則性を分布で表現したものです．横軸を人数の累積，縦軸を所得の累積とすると，完全に平等であれば45度の直線になり，一人が富を独占している場合は点 (1,0) を通る折れ線になり，一般の所得配分の場合はその間を通ることになります．また，所得配分がアンバランスなほど，彎曲度が大きくなる，ということを表しています．それをGini係数です．

もとは所得配分という特別な問題の分析から生まれていますが，所得＝富＝資産＝…と考えれば，どんなものでも「配分」「分布」のアンバランスを視覚的に表現するために有効な分析法であるといえるでしょう．例えば，図2は都道府県別の人口密度と県民総生産，10万人当たりの刑法犯発生率を一つのグラフに重ねて描いたものです．これにより，人口の大都市集中が富の集中よりも激しく，犯罪に関してはむしろ地方分散が進んでいることを伺わせます．

3.2　パレート曲線，ABC分析

資産＝在庫量と考えて，在庫管理にこの考え方を適用したのが，アメリカのデュランによって提唱されたABC分析と呼ばれる管理手法です．資産の多いものに注目する必要がある，ということから，ローレンツ曲線を180度回転して，$L(x)$ ではなく，$1-L(1-x)$ を使って分析を行います．

データでいえば，$(k/n, 1-A_{n-k}/A_n)(k=0,1,2,\cdots,n)$ を折れ線で結べば良いのです．このように，資産量の多いものから累積したグラフのことをパレート曲線といいます．パレート曲線は原点と点 (1,1) を通り，単調増加で上に凸の関数になります．ローレンツ曲線と同様に，データのばらつきが多いほど45度の直線から離れていきます．

在庫管理に適用するには，各品目ごとに在庫資産を計算し，それを上のデータと思ってパレート曲線を描き，累積資産額が70%以下の品目群をA群，累積資産額が90%以下の品目群からA群品目を除いたものをB群，残りをC群と分類します．A群の品目を重点的に管理する，C群はあまり手を掛けない管理法を行う，というように，資産額に応じて管理法を変えるのが良い，という考え方がABC分析です．

これは，パレートの法則，あるいは80-20ルール，という名前でも知られている，「少数の資産が大半の利益をもたらす」という経験的事実に基づいた指導原理です．パレートの法則はいろいろ言い換えることができます．所得の8割は2割の富裕層に属する，売り上げの8割は2割の顧客によってもたらされる，開発コストの8割は2割の工程から発生する，故障の8割は2割の原因から生じる，文章の8割は2割の単語から選ばれる，などなど．

多数派の8割の部分を重点的に管理すれば全体がうまく進行する，という考え方が在庫管理で成功してい

るならば，販売管理，開発管理，品質管理，などでも同じ考え方を適用して効果を上げることが期待できます．実際，例えば，品質管理ではエラーの頻度分布に対してこのパレート曲線の考え方を適用して，成果を挙げています．エラーの頻度を（負の）資産と考えて，資産の大きいものを重点管理する，という点で在庫管理と同じ発想です．

4. 待ち行列モデル

待ち行列モデルは普通，不特定多数のランダムな利用によって生じる施設の混雑具合を分析するために，確率過程モデルが用いられますが，サンプルパスを使った確率を伴わない分析法が時に有効な解析手段になります．

4.1 累積グラフと待ち行列過程

あるサービス施設の待ち行列の長さの時間推移は，横軸を時間として，客が到着すると上にジャンプし，サービスを終えて退去すると下にジャンプするという，おなじみのサンプルパスで表すことができますが，こうする代わりに，客が到着するたびに1ずつ増えるという到着の累積過程と，退去するたびに1ずつ増えるという退去の累積過程を重ねて一枚のグラフに描くことを考えます．システムに誰もいないという状態を原点として，そこから累積を始めれば，累積到着過程 $A(t)$ と累積退去過程 $D(t)$ のサンプルパスの差 $A(t)-D(t)$ は時刻 t の待ち行列の長さを表すことになります．

サンプルパスを描くためには，「到着」と「退去」が何か決まり，サービス施設が何か確定していれば良く，施設の内部がどうなっていようと，客がその中でどのようなサービスされようと，お構いなしにサンプルパスを描くことができます．サンプルパスですから，時間的に安定していなくても気にする必要はありませんし，確率も関係ありません．

この累積グラフを使って，待ち行列理論でもっとも基本的かつ重要な公式を導くことができます．十分長い時間 $[0, T]$ を観測して累積グラフを描いた場合，もし，待ち行列の長さが発散しないようであれば，二つのサンプルパスに挟まれた面積を定義することができます．それは $A(t)-D(t)$ を $[0, T]$ の区間で積分すれば求まりますが，もう一つ別の考え方が可能です．

サービス施設の中身はどうでも構わない，と書きましたが，ここで，仮に到着の順番に退去したと考えてみましょう．そうすると，時刻 t に到着した n 番目の客は $D(t')=n$ となる時刻 t' で退去したことになるので，n 番目の客の滞在時間を W_n とすると，$W_n=t'-t$ と表されます．したがって，二つのサンプルパスに挟まれた面積は客の滞在時間 $W_1, W_2, \cdots$ を合計したものという見方ができます．$[0, T]$ の間に N 人の客を処理したとすれば，N 人分の滞在時間の合計と，待ち行列長の時間積分が等しいという関係が導かれました．平均待ち行列長を L，平均滞在時間を W とすると，上の関係は

$$NW \approx TL$$

となります．

滞在時間の合計という量を導くために先着順にサービスされるという仮定を置きましたが，その仮定を置かない任意の退去順であっても，滞在時間の合計は先着順の場合のそれを同じになるということに注意してください．実際，n 番目に退去する客は n 番目に到着した客になりすまして出ていけば，n 番目に退去する客の滞在時間は先着順と同じになるからです．

十分大きな T を取れば N/T は単位時間当たりの到着客数とみなすことができるので，それを λ と置くと，上の式から

$$L=\lambda W$$

という等式が成り立つことが分かります．これはリトルの公式と呼ばれる，待ち行列モデルでもっとも有用な公式の一つで，その成り立ちから，実用的な意味で，ほとんどのシステムに対して成り立つ等式になっています．

4.2 テーマパークの人出

累積グラフの応用として，テーマパークの時間帯ごとの混雑具合を測る問題を考えてみましょう．スナップショットを取って数える，というのがもっともらしい方法ですが，混雑している現場を知らなくても，入りと出の数を押さえ，累積グラフを描くことによって「計算で」求めることができます．ある時刻の滞在人数は，それまでの累積入場者数から累積退場者数を引けば計算できるからです．

これはなんとか，理解できますが，では，平均滞在時間はどのように測ればよいでしょうか．入場券にICタグでも付けてデータを収集する，というように，個人個人の入退場の記録がないと，計算は難しいように思えますが，「平均」滞在時間であれば，手をかけずに計算する方法があります．リトルの公式を導く際に用いたように，累積グラフを横に切ってみれば，時間帯ごとに，入場した客の平均的な滞在時間を知るこ

とができるからです．

4.3 累積グラフとラッシュアワー問題

累積グラフを描くことによってマクロ的な分析の有効性を示したのがニューウェルの交通量分析です[1]．

一定の容量（単位時間あたり交通量の上限）をもつ道路にその容量を超える交通量が発生したとき，その渋滞はどのように広がり，いつ終息するか，ということを考える問題です．ある地点における累積到着車台数と累積通過車台数を重ねて描くことにします．ただし，累積到着車台数は，もし渋滞していなければその地点に到着したであろう車台数，と解釈することにします．実際には容量以下の交通量であっても，いろいろな要因で渋滞は発生しますが，マクロ的に水流のようなものだと考えれば，流入量が容量以下ならば溜まらずに流れていくと考えることができます．

流入量が容量を超えるとそこで滞留が発生し，再び流入量が容量以下になるまで渋滞は増え続けるでしょう．この渋滞が解消するのはいつか，ということは累積グラフを用いることによって簡単に求めることができます（図3）．実際，累積流入量の曲線を描いたとき，流入量が容量を上回るということは，累積流入量の傾きが容量を超えるところです．その時点を t_0 としましょう．累積流出量は時点 t_0 までは累積流入量をなぞっていますが，その点から容量の傾きでしか増えることができませんから，直線で増加することになり，累積流入量との乖離が始まります．

流入のピークは累積流入量の傾きがもっとも大きくなる時点 t_1 で，その後流入量は減少に転じ，そのうち，流入量が容量を下回るようになります．累積流入量の傾きが容量と等しくなる時点を t_2 としましょう．感覚的には渋滞のピークは流入のピーク時点 t_1 付近と思われがちですが，グラフから分かるように，渋滞の長さのピークは t_2 です．流入のピークでは容量を上回っていますから，単位時間当たりの積み残し量は最大になっていますが，ピークを過ぎると溜まる量が減るのであって，渋滞が解消に向かうのではありません．したがって，渋滞は増え続けます．解消に向かうのは流入量が容量を下回る点 t_2 ということが分かるでしょう．累積グラフを描いてみれば，このような説明はするまでもなく，一目瞭然です．そして，渋滞が完全になくなるのは累積流入量と累積流出量のグラフが再び交わる時点 t_3 となります．

この図から，制御の問題に見通しを立てることができます．道路交通の場合でいえば，流入が増えた場合

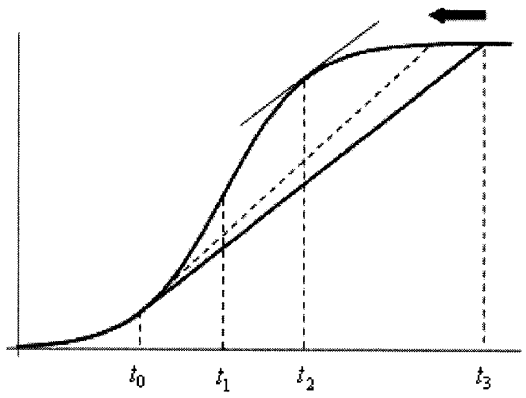

図3 渋滞の発生と解消

に信号を制御することで容量を増やす，というような工夫が考えられますが，容量を増やすということは，最大流出量の傾きを大きくするということにあたります．容量を増加させることで渋滞解消時点 t_3 を左に動かすことができますが，その増加のタイミングは早ければ早いほど効果が大きいということが分かります．

可能な限り早めに手当てすることによって，渋滞解消を早めることはもちろん，渋滞のピーク，すなわち，累積流入量と累積流出量の差をも小さくすることができることが分かります．逆に，その手当てが遅れると，渋滞発生時点も前倒しになり，より激しい渋滞を引き起こすことになります．

4.4 在庫グラフ

待ち行列をモノの在庫と読み替えると，累積グラフを使って在庫管理の問題を分かりやすく説明することができます．将来のある程度の需要予測ができている場合は，累積退去のグラフが与えられていると考え，在庫量，滞在時間が与えられた条件を満たすように累積流入量のグラフを描く問題になります．逆に，累積流入量が予測されている場合は，累積退去量を制御する問題になります．季節物商品の生産計画は前者，高速道路のサービスエリアでの浄化槽とか，ゴミ処理場の問題は後者の問題になるでしょう．

在庫量の見えない，次のような問題もあります．コンビニの納品時に行くと，プラスチックのケースに入った商品を棚に詰め込む作業に出くわします．あのプラスチックのケースは全体で何個くらいあるのでしょうか．品揃えについてどのような管理をしているのでしょうか．同じような問題を，居酒屋の裏に置いてある瓶ビールのケースについても見ることができます．両方とも，常にいろいろなところに出回っていて原材料を把握することは困難です．手持ちの在庫がなくな

ってきたら補充する，というくらいしか，方法はなさそうですが，これも累積グラフを描いて動きをおおざっぱにつかむことができます．

回収したケースについて個数を累積したものを$D(t)$，出荷したケースの累積個数を$A(t)+M$とします．Mは初期時点で市場に存在しているケースの数を表します．物流センターに溜まっているケースを在庫と考えずに，市場に出回っている未回収のケースを「在庫」と考えると，上の二つの累積カーブはMが適切に与えられていると，二つの曲線に挟まれた部分の縦軸方向が市場に出回っているケースの数，横軸方向が回収までの時間を表すことになります．現在の在庫と将来の必要量，回収率の将来見通しなどを見ながら，新たに補充すべき量とタイミングを決定する問題に対する解をこの累積グラフ上で模索するというアプローチは，数理モデルで考えるよりも，説得力があります．

5. おわりに

いくつかの例を通してデータを累積することによって分かることを調べてきました．データが与えられたら，まずは順番に並べて，グラフ化してみる，累積グラフを描いてみる，ということで，統計データ解析だけでは得られないモデル分析の可能性が広がるでしょう．

参考文献

[1] Newell, G. F.: 1982, *Applications of Queueing Theory*, Chapman and Hall.

[2] Erera, A. L., Lawson, T. W. and Daganzo, C. F.: 1998, A simple, generalized method for analysis of a traffic queue upstream of a bottleneck, *Transport. Res. Record* 1646, 132-140.