しかし，特に医学現象解析においては説明変量間に従属関係があっても，判別・予測問題に，臨床的立場から必要な変量が他の説明変量と独立でない場合でも線形モデル式の要因として選択しなければならないことが非常に多い．このような場合の解決策として，主成分分析の情報を利用して判別分析を行ない，問題を解決することがある．ただし，共通の分散共分散行列（または，相関行列）に対して主成分分析を施す際に，群間の変動と群内の変動の方向によっては，分散 σ_Y^2（寄与率）の必ずしも大きい主成分 Y を用いれば良いとは限らないが，重み ω_j, $j=1, 2,\ldots, m$（説明変量 X_j と主成分 Y との相関係数 r_{X_jY}）が現象の判別に関与している主成分軸であることを現象の専門的知識から選別することで，判別・予測問題に対して多変量データ解析が成功することが非常に多い．

このことは，質的データの多変量的解析の数量化の方法論でも同様な問題を生じる．数量化の場合，説明アイテム間の従属関係が強いアイテムを次式の線形モデル式へ選択すると，

$$Y=\sum_{j=1}^{m}\sum_{k=1}^{l_j} w_{(jk)}\,\delta_{(jk)}\,.$$

外的基準の数量を予測する第I類，外的基準の項目・区分を判別・予測する第II類では，強い従属関係の説明アイテム・カテゴリー（$C_{(j1)}C_{(j2)}\cdots C_{(jl_j)}$），（$C_{(u1)}C_{(u2)}\cdots C_{(ul_u)}$）に与える数量（$\omega_{(j1)}\omega_{(j2)}\cdots\omega_{(jl_j)}$），（$\omega_{(u1)}\omega_{(u2)}\cdots\omega_{(ul_u)}$）が要因的解釈において正反対の符号を持った数量となる．これを解決するには，主成分判別分析と同様に，数量化第III類の解析解を利用して分析する．最適な成分軸は試みの基準での重相関，相関比，判別的中率，分割表の AIC 等の評価値，および先験的情報で選択する．この実証的研究として，循環器系の健康診断の事例を示した．

「日本人の国民性調査」のサンプリング計画

中 村 隆

1988年秋に実施された「第8回 日本人の国民性調査」のサンプリング計画について，パーソナルコンピュータ上のデータベースシステムで作成したサンプリング計画用プログラム SAMPLAN/PC を操作しながら報告を行なった．

「日本人の国民性調査」は，昭和28年から5年ごとに統計数理研究所が実施している全国規模の継続調査である（母集団は20歳以上の有権者）．計画標本の大きさは6,000であり，これを折半してK調査票（継続質問中心）とM調査票（新しい質問中心）に割り当てて訪問面接式調査を行なっている．調査対象者のサンプリング方法は，層別多段抽出法であり，ほぼ同一の方法を現在まで続けている．

サンプリング計画は，全国3,300余りの市区町村の有権者数と層別に必要な情報の収集から始まる（この段階で，毎日新聞社世論調査部のお世話になった）．今回は，ラップトップコンピュータを持ち込み，SAMPLAN/PC によって市区町村データベースを構築した．このような方法をとることによってデータの入力や整合性のチェックが柔軟に行なえた．

次の段階は，市区町村の層別である．従来は県別を主要な基準として区部（6層），市部（29層），郡部（20層），沖縄（1層）の計56層に分けていたが，地点間での標本の大きさのばらつきが大きくなるので，今回は層を区部，人口20万人以上の市部，人口20万人未満の市部，郡部，沖縄の5層に簡略化した．ただし，従来とも各層の中ではさらに県別・人口規模別に市区町村を並べ換えている．層の組替えなどの検討も SAMPLAN/PC を使うことによって自由に行なえた．

層別が終了すると，実際に第1次抽出単位である地点（市区町村）を抽出する．層ごとに物理乱数を用いてスタート番号を決め，市区町村を等間隔で抽出していく．SAMPLAN/PC によって，抽出された市区町村名，有権者数，スタート番号，割当サンプル数などの一覧が得られる．前回は大型計算機によって同様の抽出を行なったが，今回は漢字で市区町村名が得られるという利点が大きかった．また，サンプリング計画の細部での変更に対しても迅速に対応ができた．

最後は，第2次抽出単位である投票区の抽出である．抽出された市区町村について投票区別の有権者数を入力し，スタート番号を含む投票区を抽出する．市区町村でのスタート番号は投票区でのスタート番号へ引き継がれる．投票区別のデータを予め入力しておくのは1回限りのサンプリングでは不経済なので，市区町村を抽出した段階で入力している．

最終的な調査対象者の抽出は，各地点を担当した調査員が面接の直前に各市区町村の選挙管理委員会を訪れて行なう．指定された投票区を見つけ，その中のスタート番号にあたる有権者から10人あるいは20人おきに調査対象者を転記する．このあたりの現状についてはまた別のトピックになる．

以上述べたように，サンプリング計画はほぼ決まりきった手順で能率的に行なえるようになってきているが，社会調査は，回収率の低下に現れているように調査環境の悪化の中にある．このような状況にどのように対処していけばよいか，統計学がどのような役割を果たすことができるのか，取り組むべき課題は大きい．

パターン分類と e_{ij} 型数量化

林　　文

数量化の方法のうち，外的基準のない場合を扱うものに，数量化III類（パターン分類の数量化）と数量化IV類（e_{ij} 型数量化）がある．e_{ij} 型数量化は，項目同志の関係 —— 何等かの親近性を表す数値 —— のデータをもとに項目間の近さを再現する空間配置を求めるもので，データに何の制約も無いため，種々の分析の初期値等として適用対象が広い．一方，パターン分類は，個人が選択項目の中からいくつかを選択したときの反応パターンの似たものを集めて，人と項目を同時に分類するものである．このように二つの方法は出発点の考え方は異なるが，パターン分類の数量化も，解く数式は項目間のクロス集計による項目間関係を用いることになる．そこで，個人の項目選択のデータに対し，項目間の同時選択の頻度を用いて親近性として，ある適当な定義を与えて e_{ij} 型数量化を実行すると，パターン分類と同じ解が得られる場合がある．その一例として次のことが言える．

e_{ij} 型数量化の親近性として個人の回答データから

$$e_{ij}=\frac{1}{2N}\frac{d_{ij}-\dfrac{d_i d_j}{N}}{\sqrt{\dfrac{d_i}{N}}\sqrt{\dfrac{d_j}{N}}}$$

をとることとする．ここで，各個人の選択した項目数が一定という条件（条件1）を仮定しておく．また，d_i：第 i 項目を選択した人数，d_{ij}：第 i 項目と第 j 項目を同時選択した人数とする．この e_{ij} は明らかに i, j 項目間の親近性の尺度となっている．このとき一致する例として，各項目を選択した人の数が一定（条件2）のとき，また，条件2を満たす完全scalableデータのときにも一致する．条件1だけのとき，解くべき固有方程式の対角線上の要素は一致しないが，それ以外は一致し，解はscaleを調整するとそれほど異ならない．条件1，2を共に満たさないデータに対しても，一般には解は異なるが，相互の位置関係が全く異なるということはない．

このほかにも解が同じになる定義やデータの条件は何か，また，解の異なりかたを調べることは，数量化の意味を考える上で必要であると考える．