2008 年度前期　応用統計学　第１２回

因子分析 (1) －因子分析とは

「体格」と「成績」には相関がある，というのは本当でしょうか？

そんな馬鹿なと思うかもしれませんが，小学生全体を対象に同じ問題の試験をすれば，高学年生のほうが，体格が大きく成績がよいことが多いでしょうから，体格と成績には相関があることになります．つまり，この例では，体格と成績の相関は実は「見かけ上の相関」であり，この相関は，「学年」というもうひとつの量を考えると，「『学年』と『体格』との相関」「『学年』と『成績』との相関」の２つの相関で説明できることになります．

因子分析は，この例の「学年」のように「複数の量の間の相関が，それらに共通な量との相関によって説明できる」というモデルを考え，この「共通な量」－「因子」と言います－を見つける方法です．この方法は，心理学の分野での，「表に現れる人の行動」には「背後にある心理的要因」があるのではないか，という考えから発達したものです．

因子を用いたモデル

例えば，数学，英語，理科，国語という４つの科目で行う試験を考えてみましょう．これらの科目の成績の間には，正の相関があるという仮説が考えられます．つまり，ある科目の成績がよい人は，他の科目の成績もよいと想像できます．

因子分析では，このように４科目間の相関があると考えられるとき，「実は，１つの『学力』という量が背後にあって，『学力』とそれぞれの科目の成績とに正の相関があるので，このような４科目間の相関が現れる」と考えます．極端な場合，もしも一人の受験生の点数が４科目とも同じであれば，各受験生の成績は４科目で表す必要はなく，１つの「学力」というデータになります．そんなに極端な場合でなくても，４科目間の相関が強ければ，各受験生の成績は「学力」というひとつの量で，かなりの程度表現できるはずです．

この場合の「学力」のように，測定はしていないけれども，相関のある複数の量を一度に説明できると考えられる量を，**因子（共通因子）**とよびます．

また，数学と理科の成績の間には強い正の相関があり，英語と国語の成績の間にも強い相関があるという仮説も考えられます．このように２科目間の相関が２組あるならば，「各科目の成績は，２つの因子で表現できる」と考えます．この場合，元の４科目のデータにおける各受験生の成績は，２つの因子（例えば「数理能力」と「言語能力」）でほぼ表現できると考えられます．このように，たくさんの項目（すなわち変量）からなるデータを少数の因子を使って表現し，測定されたデータの背後にある「そのデータが現れる仕組みのモデル」を見つけ出す方法を**因子分析**といいます．例えば，４科目の成績が，１つの「学力因子」であらわせるというモデルを考えたとしましょう．すると，ある受験生（例えば受験番号 i の「i 君」）の成績は

（i 君の数学の成績）＝（数学への影響の度合）×（i 君の「学力因子」の点数）＋（誤差）
（i 君の英語の成績）＝（英語への影響の度合）×（i 君の「学力因子」の点数）＋（誤差）
（i 君の理科の成績）＝（理科への影響の度合）×（i 君の「学力因子」の点数）＋（誤差）
（i 君の国語の成績）＝（国語への影響の度合）×（i 君の「学力因子」の点数）＋（誤差）

とあらわせることになります．

ここで，(「学力因子」の点数）はもちろん各受験生で異なりますが，これは「i君の学力」を表すものですから，i君という１人の人についてはどの科目についても同じになっています．一方，((科目）への影響の度合）は，(「学力因子」の点数）が実際の各科目の成績にどのくらい影響するかを表しているもので，科目間では異なりますが，ひとつの科目についてはどの受験生にとっても同じになっています．

ここでいう（(科目）への影響の度合）をその科目の**因子負荷量**，(i君の「学力因子」の点数）をi君の**因子得点**といいます．各科目の因子負荷量を求めることで,「学力因子」がどの程度実際の各科目の成績に影響しているかを調べることができます．また，各人の因子得点を調べることによって，各人の潜在的な「学力」の評価ができます．実際に因子分析を用いる場面では，個人の評価よりも，測定されたデータについて研究者が考えたモデルの妥当性を検証することが多いため，因子負荷量を調べることが重視される場合が多いです．

複数の因子がある場合

では，共通因子が「学力」ただ１つなのではなく,「数理能力」と「言語能力」の２つあるモデルを想定した場合を考えてみましょう．この場合，i君の成績と因子との関係は次のようになります．

(i君の数学の成績）＝（数理因子の数学への影響の度合）×（i君の「数理因子」の点数）
＋（言語因子の数学への影響の度合）×（i君の「言語因子」の点数）＋（誤差）

(i君の英語の成績）＝（数理因子の英語への影響の度合）×（i君の「数理因子」の点数）
＋（言語因子の英語への影響の度合）×（i君の「言語因子」の点数）＋（誤差）

(i君の理科の成績）＝（数理因子の理科への影響の度合）×（i君の「数理因子」の点数）
＋（言語因子の理科への影響の度合）×（i君の「言語因子」の点数）＋（誤差）

(i君の国語の成績）＝（数理因子の国語への影響の度合）×（i君の「数理因子」の点数）
＋（言語因子の国語への影響の度合）×（i君の「言語因子」の点数）＋（誤差）

だんだん複雑になるので，記号で書いてゆきましょう．i君についての，各科目の成績と，各科目における誤差（**独自因子**といいます）を，それぞれひとつにまとめてベクトル $\boldsymbol{z}_i, \boldsymbol{e}_i$ で表すことにしましょう．つまり，

$$\boldsymbol{z}_i = \begin{pmatrix} z_i(\text{数学}) \\ z_i(\text{英語}) \\ z_i(\text{理科}) \\ z_i(\text{国語}) \end{pmatrix},\ \boldsymbol{e}_i = \begin{pmatrix} e_i(\text{数学}) \\ e_i(\text{英語}) \\ e_i(\text{理科}) \\ e_i(\text{国語}) \end{pmatrix} \tag{1}$$

とするわけです．また，i君の因子得点も，(i君の「数理因子」の点数）と（i君の「言語因子」の点数）をひとまとめにして，次のようにベクトル $\boldsymbol{f}_i$ で表すことにしましょう．

$$\boldsymbol{f}_i = \begin{pmatrix} f_i[\text{数理}] \\ f_i[\text{言語}] \end{pmatrix} \tag{2}$$

さらに，因子負荷量を，例えば（言語因子の数学への影響の度合）を a[言語](数学) と書くようにすると，上の成績・因子負荷量・因子得点・誤差の関係は，次のような行列とベクトルの計算になります．

$$\begin{pmatrix} z_i(\text{数学}) \\ z_i(\text{英語}) \\ z_i(\text{理科}) \\ z_i(\text{国語}) \end{pmatrix} = \begin{pmatrix} a[\text{数理}](\text{数学}) & a[\text{言語}](\text{数学}) \\ a[\text{数理}](\text{英語}) & a[\text{言語}](\text{英語}) \\ a[\text{数理}](\text{理科}) & a[\text{言語}](\text{理科}) \\ a[\text{数理}](\text{国語}) & a[\text{言語}](\text{国語}) \end{pmatrix} \begin{pmatrix} f_i[\text{数理}] \\ f_i[\text{言語}] \end{pmatrix} + \begin{pmatrix} e_i(\text{数学}) \\ e_i(\text{英語}) \\ e_i(\text{理科}) \\ e_i(\text{国語}) \end{pmatrix} \tag{3}$$

(3) 式に現れる行列（**因子負荷行列**とよびます）を A で表すと，(3) 式の関係は，

$$\boldsymbol{z}_i = A\boldsymbol{f}_i + \boldsymbol{e}_i \tag{4}$$

という式で表せます．

データを調べることによってわかるのは，各受験生の成績 $\boldsymbol{z}_i$ だけです．さらに，科目の数ははじめからわかっており，共通因子の数はモデルを考えた時点で決めました．因子分析の計算をひとことでいうと，これだけのことしかわかっていないときに，(4) 式を満たす因子負荷量 A と因子得点 f_i を求めること，となります．

共通性

因子負荷量や因子得点を，たったこれだけのことしかわからない状況で求めるには，いくつかの仮定をする必要があります．

- 各受験生の成績 $\boldsymbol{z}_i$ は，どの科目についても，平均 0，分散 1 になるように標準化されているものとします．
- 因子得点の単位・尺度は自由に決められるので，因子得点 $\boldsymbol{f}_i$ も，どの因子についても，各々平均 0，分散 1 に標準化されているものとします．
- 独自因子得点 $\boldsymbol{e}_i$ は，どの科目についても，いずれも平均 0 とします．
- 各独自因子得点の分散を

$$d^2 = \begin{pmatrix} d^2(\text{数学}) \\ d^2(\text{英語}) \\ d^2(\text{理科}) \\ d^2(\text{国語}) \end{pmatrix} \tag{5}$$

  であらわすことにします．
- 異なる科目の独自因子どうしの間，および独自因子と共通因子との間は相関がないとします．
- さらに，ここでは，共通因子どうしも（つまり「数理能力」と「言語能力」の間で）相関がないとします（このような因子を**直交因子**といいます）．

さて，例えば数学の成績の，受験生全体での分散 $V(z(\text{数学}))$ を求めると，(3) 式から

$$z_i(\text{数学}) = a[\text{数理}](\text{数学}) \times f_i[\text{数理}] + a[\text{言語}](\text{数学}) \times f_i[\text{言語}] + e_i(\text{数学}) \tag{6}$$

ですから，

$$V(z(\text{数学})) = \{a[\text{数理}](\text{数学})\}^2 \times V(f[\text{数理}]) + \{a[\text{言語}](\text{数学})\}^2 \times V(f[\text{言語}]) + V(e(\text{数学})) \tag{7}$$

となります[1]．ここで，上の仮定から，どの科目の成績もその分散は 1 で，また因子得点の分散も，どの因子についても 1 です．また，独自因子得点の分散は (5) 式で表されています．これらを用いると，

$$1 = \{a[\text{数理}](\text{数学})\}^2 + \{a[\text{言語}](\text{数学})\}^2 + d^2(\text{数学}) \tag{8}$$

[1]受験生全体のデータから，ひとつの「分散」が求められますから，添字の i はなくなっています．

ですから，

$$1 - d^2(\text{数学}) = \{a[\text{数理}](\text{数学})\}^2 + \{a[\text{言語}](\text{数学})\}^2 \tag{9}$$

という関係があることがわかります．この両辺の値のことを，(「数学」という科目における）**共通性**といいます．数学の共通性は，数学の成績の分散のうち，ここで用いた２つの共通因子で何パーセントが表現できたかを表しています．共通性が大きいほど，その科目の成績は今考えている共通因子を用いたモデルでうまく表せていることを意味しています．

因子負荷行列の推定

さて，因子負荷行列 A を推定するには，もう少し工夫が必要です．(8) 式を，各科目について並べてみましょう．

$$\begin{aligned}
1 &= \{a[\text{数理}](\text{数学})\}^2 + \{a[\text{言語}](\text{数学})\}^2 + d^2(\text{数学}) \\
1 &= \{a[\text{数理}](\text{英語})\}^2 + \{a[\text{言語}](\text{英語})\}^2 + d^2(\text{英語}) \\
1 &= \{a[\text{数理}](\text{理科})\}^2 + \{a[\text{言語}](\text{理科})\}^2 + d^2(\text{理科}) \\
1 &= \{a[\text{数理}](\text{国語})\}^2 + \{a[\text{言語}](\text{国語})\}^2 + d^2(\text{国語})
\end{aligned} \tag{10}$$

また，異なる科目の成績どうしの共分散を考えてみましょう．仮定から，異なる共通因子得点の間の共分散が 0 で，異なる科目どうしの独自因子得点の間の共分散も 0 なので，例えば数学の成績と英語の成績の共分散 $Cov(z(\text{数学}), z(\text{英語}))$ を求めると，これらの項が打ち消され，

$$Cov(z(\text{数学}), z(\text{英語})) = a[\text{数理}](\text{数学})a[\text{数理}](\text{英語}) + a[\text{言語}](\text{数学})a[\text{言語}](\text{英語}) \tag{11}$$

となります．どの科目の組み合わせについてもこのようになります．

(10) 式，(11) 式を，すべての科目，すべての変量について組み合わせると，各科目の成績の分散共分散行列（ここでは，成績が標準化されているので相関行列）を R とするとき

$$R = AA' + D \ \text{ すなわち } \ AA' = R - D \tag{12}$$

$$D = \begin{pmatrix} d^2(\text{数学}) & & & 0 \\ & d^2(\text{英語}) & & \\ & & d^2(\text{理科}) & \\ 0 & & & d^2(\text{国語}) \end{pmatrix} \tag{13}$$

という関係があることが導かれます[2]．

因子負荷行列 A を求めるには，(12) 式を解かなければなりません．もし $R - D$ がわかっていれば，この講義の第５－７回の「主成分分析」の項で触れた，「対角化」の手法を使って解くことができます．しかし，実際には D は未知ですから，反復法によって近似解を求める必要があります．

[2] くりかえしますが，A' は行列 A の転置行列をさします

これは，$R-D$ の固有値を $\lambda_1, \lambda_2, \ldots, \lambda_p$，対応する固有ベクトルを $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p$ とするとき，$R-D$ が

$$
\begin{aligned}
R - D &= \left(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p\right)
\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}
\begin{pmatrix} \boldsymbol{b}_1' \\ \boldsymbol{b}_2' \\ \vdots \\ \boldsymbol{b}_p' \end{pmatrix} \\
&= \left(\sqrt{\lambda_1}\boldsymbol{b}_1 \ldots \sqrt{\lambda_p}\boldsymbol{b}_p\right)
\begin{pmatrix} \sqrt{\lambda_1}\boldsymbol{b}_1' \\ \vdots \\ \sqrt{\lambda_p}\boldsymbol{b}_p' \end{pmatrix}
\end{aligned}
\tag{14}
$$

と表せることを利用します．もしも $R-D$ が，(12) 式のように，因子負荷行列 A とその転置 A' の積で表せるのなら，(14) 式の２行目が，それぞれ A と A' に対応します．このとき，(3) 式のように，因子負荷行列の列の数は因子数と同じですから，(14) 式の固有値のうち，0 でないものは因子数と同じ数だけしかなく，他は 0 になるはずです．そこで，D に適当な初期値を入れて固有値を求め，大きいほうから因子数だけ残して，他は 0 にしてから，(14) 式の２行目で「仮の」A と A' を求め，それらから $D = R - AA'$ によって新しい D を求め，... という計算を繰り返して，A と D を求めます．