平成14年度修士（情報科学）論文・発表要旨

多次元統計データ解析の理論と応用

猪股　力

2003年2月14日

1　はじめに

多変量解析とは, いくつかの項目の間の関連性を統計的に分析し, 現象を要約して簡潔な表現を与えたり, 現象の背後にひそむ構造を浮き彫りにしたり, ある項目を他のいろいろな要因から予測 (説明) したりする手法である. 多変量解析にはさまざまな手法がある. 本論文では, さまざまな多変量解析を紹介し, さらには, 多変量解析を行ううえで必要となる多変量分布や検定についても紹介する. 最後に多変量解析を用いて「2年目のジンクス」を統計的に処理し, 検証する.

2　計算例

2.1　2年目のジンクス

日本のプロ野球において, 1年目に活躍した選手は2年目には活躍できないことを「2年目のジンクス」と言う. このジンクスは本当に存在するのだろうか. 過去に新人王をとった選手の成績を元に考える. 表1は, 1964年からの新人王を獲得した選手のその年の成績である. 表2は新人王を獲得した次の年の成績を表している. 表3は2002年度のセ・パ両リーグの成績上位投手の成績である. これらをデータとして分析を行う.

2.2　分析手法

分析するにあたって**正準相関分析**, **主成分分析**を用いる.

正準相関分析とは, 2つの変量群の間の相関を分析する分析手法である. 両変量群の相関を最もよく表すような新しい変量に変換し, その変量によって2つの変量群の相関について考える.

これを用いて, 1年目の成績と2年目の成績の相関について考える.

それぞれの群において正準相関分析を行い, 1年目と2年目の成績の間の相関関係を見る.

主成分分析とは, ある問題に関していくつかの要因が考えられるとき, それをできるだけ情報の損失なしになるべく少数の主成分で表現する手法である. ここでは, 1年目と2年目の成績の総合的な指標を得ることを目的に用いる.

新人王獲得選手を次のような2つの群に分割する.
G_1 群:堀内, 安田, 斉藤, 津田, 野茂, 松坂
G_2 群:上記以外の選手

この2つの群の分割の規準は, 新人王獲得翌年度に何らかのタイトルをとった選手を G_1 群, 1つのタイトルも獲得できなかった選手を G_2 群とする.

2.3 母平均ベクトルの検定

G_1 群と G_2 群の2年目の成績について, 有意の差があるかどうかを検定する. まず, 2つの群の分散共分散行列についての検定を行う.

$$\text{帰無仮説 } H_0 : \Sigma_{(1)} = \Sigma_{(2)}$$
$$\text{対立仮説 } H_1 : \Sigma_{(1)} \neq \Sigma_{(2)}$$

仮説 H_0 の下で
$W = |\hat{\Sigma}_{(1)}|^{\frac{n_1}{2}} |\hat{\Sigma}_{(2)}|^{\frac{n_2}{2}} / |\hat{\Sigma}|^{\frac{n}{2}}, \quad n = n_1 + n_2$
とおくと

$$\chi_w{}^2 = -2 \log_e W$$

が近似的に自由度 $p(p+1)/2$ の χ^2 分布に従うことを用いる.
G_1 群, G_2 群の母平均ベクトルを $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ とする. このとき, 2つの母平均ベクトルについて

$$\text{帰無仮説 } H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$
$$\text{対立仮説 } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

の検定を行う. 仮説 H_0 の下で
$T^2 = \dfrac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$
が自由度 $(p, n_1 + n_2 - p - 1)$ の T^2 分布に従うことを用いる. ここで $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2$ は標本平均ベクトル, $\boldsymbol{S}$ は標本分散共分散行列を示す.

検定の結果, 帰無仮説は棄却され2つの群に有意の差があることがわかった. また, 1年目の成績に関しては有意の差はなかった.

2.4 正準相関分析結果

G_1 群, G_2 群それぞれにおいて, 1年目と2年目の成績を用いて正準相関分析を行った. 変量は, 1年目の成績からは投球回数, 四死球, 自責点, 奪三振, 2年目の成績からは勝利数, 勝率を用いた.

G_1 群における結果で**冗長性指数**をみると, 1年目の成績に対して44%, 2年目の成績に対して81%であった. これが表すことは1年目から2年目を予測 (説明) できる割合は44%であり, 2年目から1年目を予測する割合は81%と高いことを意味している.

この**冗長性指数**を G_2 群についてみると, 1年目の成績から2年目の成績を予測できる割合は7%程度, 2年目の成績から1年目の成績を予測できる割合は10%程度ということで, 共

にかなり低いことわかる. つまり, G_2 群の選手たちは1年目と2年目の成績が大きく離れていることが考えられ, そこに2年目のジンクスが存在することが考えられる.

冗長性指数とは, 一方の変量群から他方の変量群をどのくらい予測できるかを示す指標である.

また, 正準相関係数がゼロであるかどうかの検定については, 帰無仮説は棄却された.

2.5 主成分分析結果

新人王獲得選手の成績を, 表3の2002年度セ・パ上位投手成績に加えて分析を行った.

第1主成分の係数は, 投手にとってマイナスの変量である四死球と自責点は負の値になっていて, それ以外の変量は正の値で大きなばらつきがないことから, 第1主成分は総合的な指標であると解釈される.

第2主成分は, 勝率の係数が大きいことから勝率を表す指標と考えられる. これらを用いて主成分得点を計算し, その得点を比較し検証する.

第1主成分, つまり総合的な指標に関しては大きく得点が落ちた選手は少ないが, 得点が上がった選手は少ない. 近年の選手では藪選手, 川上投手が上がっているほかは得点が上がった選手はいない. この2人の得点の上がり方を昔の選手と比べてみても, 藪選手は0.3程度, 川上選手が0.01程度に対して, 斉藤明夫投手は3程度上がっている. 他に上がっている選手と比べても, 藪, 川上両投手の上がり方は小さいことがわかる. これは近年では2年目に大きく成績を伸ばすことができないことを表していると考えられる.

3 結論

以上の結果をふまえると, 昔の選手は1年目, 2年目とも比較的変わらない成績を残していた. ここ十数年の選手たちの成績が落ちていることから, 2年目のジンクスというものは近年の現象であると考えられる.

参考文献

[1] 田中豊・脇本和昌: 多変量統計解析法, 現代数学社, 1983.

[2] 道家暎幸・今田恒久: 多変量解析序論, 東海大学出版会, 2001.

データ提供

[3] プロ野球記録博物館 HP.

[4] 日本プロ野球記録年間, ベースボール・レコード・ブック, ベースボールマガジン社.