9.2 解説2：多変数データの統計分析

本書に掲載のソフトウェアプロジェクトの実績データは、図表9-2-1のような、欠損値を含む<個体×変数>型データであり、1つの個体（プロジェクト）に対して多数の変数が記録されている。本節では、このような多変数の関連の分析に適した統計解析手法について紹介する。なお、図表9-2-1のデータはあくまでも本説明のために作成した架空のデータであることにご留意いただきたい。また、紙面の都合上、詳細は割愛するが、図表9-2-1のようなデータセットを作成するにあたって、各プロジェクトにおける変数の定義を可能な限りそろえる必要がある。例えば、工期（月数）については、各プロジェクトにおける1人月あたりの時間数（例えば1人月=160人時など）の定義をそろえる必要があるし、規模（FP）については、FPの計測手法が同種のプロジェクトのみを用いることが望ましい。

図表9-2-1 ●<個体×変数>型データの例

プロジェクトID	業種	アーキテクチャ	主開発言語	規模(FP)	工期(月数)	平均要員数	外部委託率	生産性(FP÷人時)
1	銀行	C/S	PL/I	556	15	15	0.5	0.0225
2		スタンドアロン	C	80	8	6	0	0.0970
3	銀行	C/S	COBOL	77	6	1	0	0.1016
4	銀行		Visual Basic	255		6	0.25	0.1203
5	製造業	スタンドアロン		349	4	3	0.25	0.1273
6	銀行	混合	C++		4	11	0.5	0.1835
7	銀行	C/S	COBOL	375	2	2	0	0.2022
8	公共	混合	Java	271	12		0	0.1551

9.2.1 目的変数と説明変数を決める

まず、分析に際し、興味のある変数(目的変数)を1つ決める。例えば、図表9-2-1において、目的変数として「生産性」を選ぶこととする。次に、目的変数との関連を調べたい変数（説明変数）を選ぶ。ここでは、説明変数として「業種」から「外部委託率」までの7変数を選んだとする。この場合の興味は、それぞれの説明変数は生産性との関連があるのか、その関連の強さはどの程度か、またその関連は統計的に有意といえるのか、といったことである。

9.2.2 説明変数の尺度をそろえる

目的変数である生産性は、値を足したりかけたりすることが可能な**量的データ**である。一方、説明変数には、業種、アーキテクチャといった値の演算が不可能な**質的データ**と、規模や工期といった量的データが混在しているため、統一的に扱うことが困難である。そこで、全ての説明変数を質的データに揃えてから分析することとする。ここでは、規模、工期、平均要員数、外部委託率について、値の大きさに応じて下位25%のデータ、上位25%のデータ、それ以外の中位のデータの3つのカテゴリに分けることとする。規模（FP）についてこの処理を行った例を図表9-2-2に示す。このようにして全ての説明変数を質的データにそろえたものが図表9-2-3である。

図表9-2-2 ●量的データから質的データへの変換

プロジェクトID	規模(FP)
1	556
2	80
3	77
4	255
5	349
6	
7	375
8	271

規模の降順に並び替える

プロジェクトID	規模(FP)	
1	556	上位25%
7	375	
5	349	中位
8	271	
4	255	
2	80	下位25%
3	77	
6		

IDの昇順に並び替える

プロジェクトID	規模(FP)
1	上位
2	下位
3	下位
4	中位
5	中位
6	
7	上位
8	中位

図表 9-2-3 ●変換後の説明変数と目的変数

説明変数							目的変数
業種	アーキテクチャ	主開発言語	規模 (FP)	工期 (月数)	平均要員数	外部委託率	生産性 (FP÷人時)
銀行	C/S	PL/I	上位	上位	上位	上位	0.0225
	スタンドアロン	C	下位	中位	中位	下位	0.0970
銀行	C/S	COBOL	下位	中位	下位	下位	0.1016
銀行		Visual Basic	中位		中位	中位	0.1203
製造業	スタンドアロン		中位	下位	中位	中位	0.1273
銀行	混合	C++		下位	上位	上位	0.1835
銀行	C/S	COBOL	上位	下位	下位	下位	0.2022
公共	混合	Java	中位	上位		下位	0.1551

9.2.3 分散分析を行い、質的データと量的データの関連を調べる

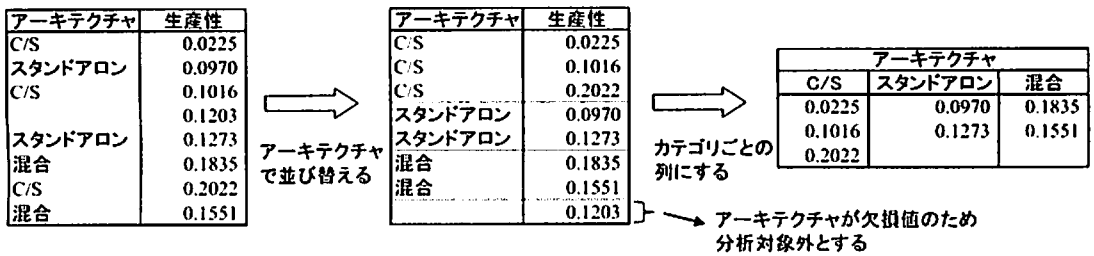
図表 9-2-3 のように、目的変数が量的データであり、説明変数が質的データである場合、**分散分析**が適用できる。分散分析にはいくつか種類があるが、ここでは最も簡便な一元配置分散分析に限定して話を進める。(一元配置の) 分散分析は、ある質的データ (例えばアーキテクチャ) のカテゴリ (C/S、スタンドアロン、混合) の間で、ある量的データ (生産性など) の平均値に差があるかどうかを検定する手法である [2]。分散分析では、質的データ (説明変数) と量的データ (目的変数) の間に関連があるか否か (すなわち、カテゴリ間で平均値に差があるか否か) を統計的に確かめるための p 値という指標と、その関連の大きさを表す**寄与率** (効果量、相関比とも呼ばれる) という指標が算出できる。

分散分析は Excel や SPSS のようなソフトウェアで行うことができる。例えば、アーキテクチャと生産性について分散分析する場合、図表 9-2-4 に示すように、アーキテクチャのカテゴリごとにプロジェクトを分類した表へと作り直すことで、Excel の分析ツールにかけることが可能となる。寄与率は Excel による分散分析の結果から、以下の式で求めることができる。

$$\text{寄与率} = \frac{\text{グループ間変動}}{\text{グループ内変動} + \text{グループ間変動}}$$

寄与率は 0 から 1 の実数値をとり、1 に近いほどデータ間の関連が強い (大きい) ことを示す。

図表 9-2-4 ●分散分析に用いる表の作り方の例



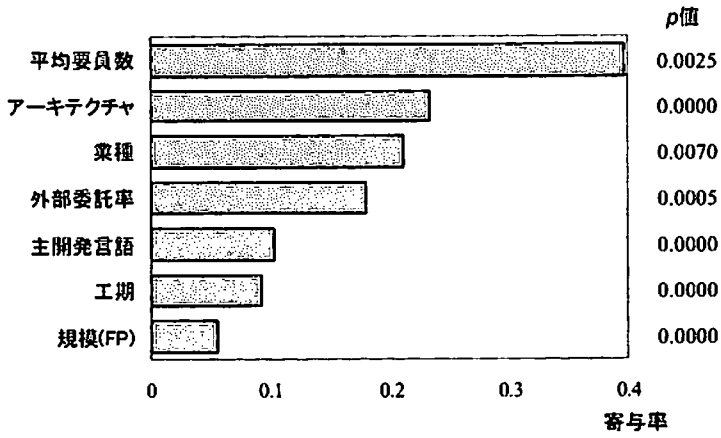
9.2.4 分析の事例

分散分析を試行した事例を紹介する。分析に使用したデータセットは 2005 年度版データ白書 2005 の第 7 章、図表 7-1-2 の定義に基づき、1009 件のプロジェクトから絞り込んだ 211 件の新規開発プロジェクトである。工数、工期、外部委託率、生産性の定義もデータ白書 2005 に準じる。平均要員数については、開発総工数÷工期により求めた値を用いている。

生産性に対する各プロジェクト特性の寄与率と p 値を図表 9-2-5 に示す。図より、この試行では、平均要員数、アーキテクチャ、業種、外部委託率、種開発言語、工期、規模 (FP) の順に寄与率が高いことが分かる。この 7 変数全てが有意水準 5%で生産性に寄与している (すなわち p 値≦ 0.05) といえる。一般に、プロジェクトの数が少なかったりカテゴリが細かすぎる場合は、p 値が大きくなり、統計的に有意でなくなる場合がある。

このような場合は、プロジェクト数を増やしたり、いくつかのカテゴリを一つにまとめる（例えば、「C」と「C++」という2つのカテゴリを「C/C++」として一つにまとめるなど）ことにより、1カテゴリあたりのプロジェクト数を増やすことが必要となる。

図表 9-2-5 ●分散分析の試行結果（生産性に対する各変数の寄与率とp値）



なお、この結果は、例えば平均要員数が生産性と大きな関連を持つことは示されているが、平均要員数の大小の違いによって生産性がどの程度変わってくるのかについては、さらなる分析が必要となる。例えば、平均要員数の「上位」「中位」「下位」のカテゴリごとに生産性の分布を表す箱ひげ図を描いたり、2つのカテゴリ間で平均値の差の検定（t検定）を行うことが必要となる。

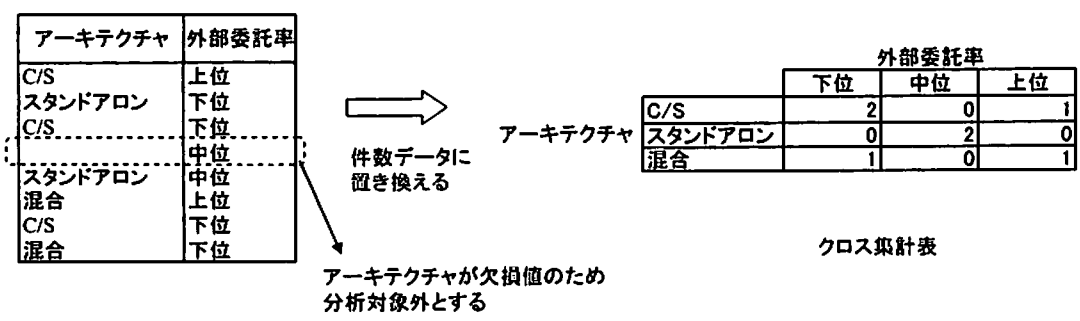
9.2.5 カイ二乗検定を行い、質的データ間の関連を調べる

分散分析により、各説明変数と目的変数との関連を知ることができたが、説明変数同士もまた互いに関連を持っている可能性がある。例えば、アーキテクチャと主開発言語との間には何らかの関連があるかもしれない。ところが、規模や工期といった量的データ間の関連の大きさや関連の有無は**相関係数**や**無相関検定**などにより調べることができるが、アーキテクチャや言語といった量的データには適用できない。このような場合には、**χ^2（カイ二乗）検定**が適用できる。

χ^2検定は、独立性の検定とも呼ばれ、質的データ間に関連があるか否か（独立であるか否か）を検定するための手法である［2］。χ^2検定では、質的データ間の関連が統計的に有意であるか否かを確かめるためのp値と、その関連の大きさを表す**クラメールのV**（Cramer's V）という指標が算出できる。クラメールのVは、質的データにおける相関係数とも位置づけられる。

χ^2検定を行うには、図表9-2-6に示すようにクロス集計表を作成する必要がある。図中のクロス集計表の各マスの値は、該当するプロジェクトの個数（ケース数）である。χ^2検定はExcelの分析ツールには含まれていないが、群馬大学の青木繁伸先生のWebページ［1］で行うことができ、p値、クラメールのVともに算出できる。クラメールのVは0から1の値をとり、1に近いほど変数間の関連が大きいことを表す。

図表 9-2-6 ●χ^2検定に用いるクロス集計表の作り方の例



χ^2 検定を試行した事例を紹介する。図表 9-2-7 は、図表 9-2-2 と同様のデータセットを用いて χ^2 検定を行った結果である。表の各マスの上段の数値がクラメールの V であり、下段の数値が p 値である（ここでは p 値をパーセント表示している）。有意水準 5%で関連がある（p 値≦ 5%）と認められた部分についてはクラメールの V の値を太字で示した。なお、クラメールの V の値が大きいからといって統計的に有意となるとは限らない。有意となるか否かは、プロジェクトの数やカテゴリの細かさなどにも影響を受けるためである。

図表 9-2-7 より、この分析の試行では、平均要員数は他の 6 つの変数と関連が認められることや、規模と工期は関連が強いことなどが分かる。ただし、分散分析と同様、より詳細に関連を分析するためには、2 変数に絞って散布図や箱ひげ図を描いたり、t 検定などを行うことが必要となる。

図表 9-2-7 ●χ^2 検定の試行結果（クラメールの V と p 値）

	外部委託率	平均要員数	工期	規模(FP)	業種	アーキテクチャ
平均要員数	0.35 0%					
工期	0.22 10%	0.26 0%				
規模(FP)	0.23 7%	0.47 0%	0.53 0%			
業種	0.31 13%	0.26 1%	0.18 32%	0.18 33%		
アーキテクチャ	0.30 4%	0.44 0%	0.24 0%	0.17 12%	0.20 13%	
主開発言語	0.31 16%	0.41 0%	0.28 0%	0.27 1%	0.28 1%	0.46 0%

9.2.6 まとめ

ここでは、質的データと量的データの関連を調べるための分散分析と、質的データ間の関連を調べるための χ^2 検定について概説し、それぞれ手法の試行の事例を紹介した。紙面の都合上、各手法を適用する際の前提条件については述べていない点に注意されたい。例えば、（一元配置）分散分析は、各カテゴリの分散が等しいことを前提にしており、分散の均一性についても予め検定しておく方が望ましい。詳しくは、検定に関する図書や Web サイト（[2] など）を参照していただけたらと思う。

参考情報

[1] 青木繁伸：統計電卓（CGI）独立性の検定（カイ二乗検定）, http：//aoki2.si.gunma-u.ac.jp/calculator/chi_sq_test.html
[2] 青木繁伸：統計学自習ノート, http：//aoki2.si.gunma-u.ac.jp/lecture/