OR メモランダム

このコラムは，ORにかかわる概念，知識(手法，原理)，それらの図解，よい教材や問題，実学ORの実施経験，そこから得られた智恵やアドバイス，失敗談と教訓，新しい視点，視座，フレームワーク，未だ解けていない問題，面白い研究テーマなどを，“新鮮に”，しかも“コンパクトに”表現し，提示していただくものです．ユニークなアイディア，フレッシュな見方，発想，だれかと意見をたたかわせたい問題提起など，ふるってご投稿ください．(原稿は，刷り上がり，半ページから3ページに納まるようにお書きください．簡単に！ 加筆訂正をお願いする場合があります)

相関があるかを見つける簡便法

上田　太一郎

結　論　筆者が主張したいこと，つまり結論から述べます．

相関係数をrとして，

$$r^2 > 4/(n+2)\quad (n：データ数)$$

が成立すれば相関があるとする

ということです．

1．データマイニングの1つの方法 —相関を見つける—

このごろ「データマイニング」ということばをよくききます．データウエアハウス（データの倉庫）に貯えられているデータをマイニング（採掘）してなにか役に立つ情報，知見，仮説等を得る方法といえるでしょう．データマイニングに統計手法は重要な役割をするものと考えています．たとえばPOSデータから，ある商品と別の商品との売上げの関連を見つけることを考えます．すなわち，ある商品の売上げ高が増すと，別の商品の売上げ高が増す（あるいは減る）傾向にあることを見つけたいとします．この指標には周知のように相関係数があります．

2．相関があるかを見つける簡便法

相関があるか検定するには統計量$t=r(n-2)^{1/2}/(1-r^2)^{1/2}$が自由度$n-2$の$t$分布に従うことを用います．求めた相関係数$r$が$t$分布表の，たとえば5％有意点の値より大なら仮説$r=0$を棄却し，危険率5％で相関があると判定します．この検定法は電卓とt分布表があれば可能ですが，1％で仮説が棄却されるが，5％では棄却されないとか危険率によって結論が異なることがあります（この点は筆者はいつもしっくりこないところです）．

そこで，ここではt分布表が不要で，かつ紙とエンピツでできる簡単な方法を提案します．

簡便法

$r^2 > 4/(n+2)$　(n：データ数) が成立すれば相関がある

3．どのようにして$r^2>4/(n+2)$を導いたのか

では，どのようにして$r^2>4/(n+2)$を導いたのか説明します．

回帰分析の重要なテーマに変数選択問題があり，いろいろな選択規準が提案されています．たとえば佐和[1]，芳賀ほか[2]があります．いずれも，重相関係数をデータ数と説明変数の個数で調整した規準です．筆者もデータ解析用言語S([3]，[4])を用いて試行錯誤の末，自分なりに選択規準を作ってみました(Sを使用すると意外に簡単に求まりました)．それは，Rを重相関係数として

$$R_u=1-(1-R^2)\{(n+k+1)/(n-k-1)\}\quad (1)$$

を最大にする変数の組合せがよい，というものです．ここで，kは説明変数の個数です．(1)の選択規準をいろいろなケースに適用してみましたが，上記佐和の規準，芳賀ほかの規準を用いた場合と同様な結果が得られています．

さて，単回帰式は重回帰式の特殊なケースです．そこで，(1)を単回帰式に適用すると，$R^2=r^2$，$k=1$です．また$R_u>0$が成立しないと意味がありません．したがって，$1-(1-r^2)(n+2)/(n-2)>0$です．これを解くと，

$r^2>4/(n+2)$　となります．

うえだ たいちろう　三菱電機（株）情報システム製作所
〒247 鎌倉市上町屋325

4．適用例

4.1 いんちきなサイコロか

2つのサイコロを同時に投げ，その目の出方に相関があるか調べてみます（もちろん相関はないはずです）．相関があれば，いんちきなサイコロと疑いたくなります．1つのサイコロの目は3，2，3，4，5，4，同時にもう1つのサイコロの目は1，3，1，4，6，3と出ました．相関係数は0.7035です．かなり大きな値のようで，相関がありそうかなと心配です．t検定するとt値＝1.980，自由度＝4の5％点は2.132となり，帰無仮説：相関係数＝0が採択されます．簡便法では$r^2=0.7035^2=0.4949<4/(6+2)(=0.5)$となって相関があるとはいえないことになります．つまり，いんちきなサイコロとはいえないことになります．

4.2 入社時に学力優秀デキルと限らず？［5］

ソフトウエア製作の生産性は個人差が大きく，またシステムの構築はシステムエンジニアリングの良否がシステムの出来不出来に関係するといわれています．したがって，新人の採用にあたっては，入社時に将来の人材であるシステムエンジニアを見つけたいものです．そこで，入社時に将来の人材（システムエンジニア）を見つけようとして，入社時の英・数・国の成績と数年後の能力評価データに相関があるかどうか調べました（データ数127）．たとえば，数学の得点とプロジェクト管理能力評価とに相関があれば，採用に当っては数学の得点を重視することになります．数学とプロジェクト管理能力評価とでは，相関係数r＝0.132でした．簡便法では$0.132^2=0.017<4/(127+2)=0.031$となり，相関があるとはいえません．$t$検定でも$t$値＝1.489，5％点＝1.979となり，帰無仮説が採択されます．つまり，相関があるとはいえないわけです．

4.3 百貨店の焼き立てパンショップの調査データから評価項目の相関を見つける

表1は96年9月28日の日経流通新聞に載った首都圏在住の20代以上の女性489人に聞いた調査データです（同新聞では％で表わしていましたが，ここでは絶対数に直しました）．

相関係数行列（**表2**）からパンショップにとって貴重な仮説が得られます．まず，人気のあるショップにするには，菓子パン，ペストリーがおいしく（相関係数0.757），パンの種類が豊富（同0.635）で，価格が妥当（同0.670）なことです．人気があれば，認知され（同0.895），購入経験も多くなる（同0.951）ようです．

同様にして，相関係数行列を見ると，認知度を高めるようにするには，菓子パン，ペストリーがおいしく（同0.607），価格が妥当な（同0.662）ことです．ここで，興味あるのはフランスパン，クロワッサンがおいしいのと店の雰囲気に高級感があるのは，認知度を高めるのに貢献しないどころか逆（相関係数が負，各々−0.576，−0.499）に低くすると言えます．これはさらにデータをとり，慎重に結論を出すべきことですが，注意すべき仮説です．

購入経験を高めるようにするには，菓子パン，ペストリーがおいしく（同0.738），パンの種類が豊富で，価格が妥当（同0.748）であることです．店の雰囲気に高級感があるのと購入経験とは逆の相関（−0.524）になっているのも注意すべきことです．

5．簡便法とt統計量との関係

$t=r(n-2)^{1/2}/(1-r^2)^{1/2}$から$r^2$を求め，$r^2>4/(n+2)$に代入すると$t$の絶対値＞2となります．つまり，簡便法は「$t$統計量を用いた検定で，$t$の絶対値＞2が成立するとき相関があると判定する」と同値になります．相関の有無の判定にはどちらを用いてもよいことになります．

表3はデータ数に対応した$4/(n+2)$（仮に判定点と呼びます），自由度，$t=2$のときの危険率を一覧表にしたものです．これを見ると自由度が小さいときは危険率は比較的大きく，自由度が60くらいでほぼ5％になり，自由度が増すごとに少しずつ減少していることがわかります．自由度が100でもあまり変わりなく5％弱となっています．

6．おわりに

相関係数は周知のように線形な関係を表わすモノサシです．したがって非線形な（2次式とか円）関係のとき，たとえば円の場合には0になるときがあります．また，本来相関がないのに1つの外れ値（測定ミス，入力ミス等による）のために相関があるような場合も出てきます．いずれの場合もグラフを描くのがよいといわれています．

このようなことに注意しながら簡便法を使ってみてはいかがでしょうか．

表1 女性が選ぶ百貨店の焼き立てパンショップ調査データ

	人気度	認知度	購入経験	食パン，イギリスパンがおいしい	フランスパン，クロワッサンがおいしい	菓子パン，ペストリーがおいしい	パンの種類が豊富	その店の素材や作り方が好き	価格が妥当	店の雰囲気に高級感がある
アンデルセン	196	411	328	88	122	235	230	88	112	39
サンジェルマン	166	391	323	108	112	191	200	68	156	34
木村屋総本店	156	465	386	44	29	249	88	83	137	24
ポンパドウル	147	342	279	122	200	156	191	68	108	29
フォション	112	249	166	171	235	142	98	117	24	259
ドンク	98	284	225	93	171	122	103	64	108	29
ヴィ・ド・フランス	88	230	181	44	156	176	205	54	166	20
ジョアン	78	181	127	122	210	147	196	108	93	78
ホテルオークラ	54	298	137	230	88	54	39	39	24	137
ぼるとがる	54	225	147	93	117	127	122	49	103	15
ダロワイヨ	34	132	59	73	156	166	98	73	24	249
ルノートル	34	108	73	88	137	147	93	54	34	117
プチモンド	24	210	83	39	44	98	78	20	68	20
ビゴの店	20	78	29	64	298	93	112	171	49	78
エディアール	15	68	29	73	235	108	73	147	0	235

人（複数回答）

まず，表1から相関係数行列（表2）を求めます．簡便法で相関があると判定したものにアンダーラインをつけました．

表2 相関係数行列

	人気度	認知度	購入経験	食パン，イギリスパンがおいしい	フランスパン，クロワッサンがおいしい	菓子パン，ペストリーがおいしい	パンの種類が豊富	その店の素材や作り方が好き	価格が妥当	店の雰囲気に高級感がある
人気度	1									
認知度	0.895	1								
購入経験	0.951	0.963	1							
食パン，イギリスパンがおいしい	0.104	0.151	0.021	1						
フランスパン，クロワッサンがおいしい	-0.252	-0.576	-0.450	0.114	1					
菓子パン，ペストリーがおいしい	0.757	0.607	0.738	-0.392	-0.306	1				
パンの種類が豊富	0.635	0.355	0.488	-0.166	0.132	0.578	1			
その店の素材や作り方が好き	-0.093	-0.360	-0.237	-0.064	0.786	-0.015	0.032	1		
価格妥当	0.670	0.662	0.748	-0.332	-0.356	0.610	0.696	-0.317	1	
店の雰囲気に高級感がある	-0.391	-0.499	-0.524	0.339	0.419	-0.262	-0.459	0.412	-0.817	1

参考文献

[1]佐和隆光：「計量経済学の基礎」(1970)，東洋経済新報社，178－184.

[2]芳賀敏郎，竹内啓，奥野忠一：“重回帰分析における変数選択の新しい規準”(1976)，「品質」，Vol. 6, No. 2.

[3]渋谷政昭，柴田里程：「Sによるデータ解析」(1992)，共立出版.

[4]上田太一郎：“S－PLUSの有効利用”(1994)，「オペレーションズ・リサーチ」Vol. 39, Nol. 11.

[5]上田太一郎，小林整功：“システムエンジニアリング企業における入社試験結果と数年後の能力評価との相関について”(1993)，情報処理学会第47回(平成5年度後期) 全国大会予稿集.

表3 データ数と判定点，危険率

データ数	判定点	自由度	$t=2$のときの危険率
3	0.8000	1	0.2952
4	0.6667	2	0.1835
5	0.5714	3	0.1393
6	0.5000	4	0.1161
7	0.4444	5	0.1019
8	0.4000	6	0.0924
9	0.3636	7	0.0856
10	0.3333	8	0.0805
11	0.3077	9	0.0766
12	0.2857	10	0.0734
13	0.2667	11	0.0708
14	0.2500	12	0.0687
15	0.2353	13	0.0668
16	0.2222	14	0.0653
17	0.2105	15	0.0639
18	0.2000	16	0.0628
19	0.1905	17	0.0617
20	0.1818	18	0.0608
30	0.1250	28	0.0553
40	0.0952	38	0.0527
50	0.0769	48	0.0512
60	0.0645	58	0.0502
70	0.0556	68	0.0495
80	0.0488	78	0.0490
90	0.0435	88	0.0486
100	0.0392	98	0.0483