データ解析に見るグラフ

新村　秀一

1. はじめに

地球が豪々と音をたてて回っていると言った人がいたが，最近のデータ解析も目まぐるしく変化してきているようである．昔は統計解析という言葉がよく使われたが，最近ではデータ解析と呼ばれるようになってきた．そこに時代の変遷を見る想いである．たとえば，探索的データ解析[5]のように，単に統計手法にとどまらず図式表示を重視する傾向が強まっている．

この影響は統計パッケージにも及んでいる．そこで代表的な統計パッケージＳＡＳを中心としてその一端を紹介したい．この他にも，チャーノフ図に代表されるような多くのグラフ手法が統計に用いられているが割愛する．

2. ＳＡＳとは

本稿では，ＳＡＳを紹介するのが目的ではないし，紙面も少ないので巻末の文献[1]～[3]を参考にしてほしい．当初，ＳＡＳはデータ管理機能とプログラミング機能を含む統計パッケージとして開発されたが，その後グラフィック・時系列解析・品質管理・ＯＲなどを追加して，エンド・ユーザー言語に変身してきている．

本稿の執筆を牧野先生から勧めていただいたさいに，文献[6]にＳＡＳのグラフ処理が数多くとりあげられており，その観点から読者に紹介してほしいとのことであった．しかし，調査の時間も十分にないので，文献名の紹介にとどめ後日の宿題としたい．

しんむら　しゅういち
住商コンピューターサービス㈱
〒101 千代田区東神田２―５―15

3. 統計手法とグラフ

従来の統計書は，個別手法のアルゴリズムの紹介に主眼を置いたものが多かった．しかし，現実の分析対象に向かって問題解決を迫られた読者には，一連のデータ解析の手順を示す必要がある．また，統計手法とグラフ表現を併用する必要もある．**表1**は，そのような主張から執筆した「統計処理エッセンシャル」からの表である．

すなわち，統計手法は大雑把に分けると予測に関するものと変数の分布を調べる手法とに別れる．後者に関して，対象とする変数の数と変数が数値か文字かによって分類したものがこの表である．

1個の数値変数のデータに関しては，従来平均値・標準偏差・最大値・範囲・変動係数等の要約統計量が用いられてきた．これらは数値として客観的に把握できるが，棒グラフのような図式表現を用いれば視覚的かつ具体的に把握できる．

2変数の要約統計量としては，相関係数がある．それを補うグラフ表現としては，2変数の散布図が用いられる．Anscombe[7]は，**図1**に示す平均値・分散・共分散のまったく等しい2変数の有名なデータを示した．相関係数が意味のあるのは(a)の場合であり，(b)は曲線相関を，(c)は異常値を，(d)は右の孤立点1個に大きな影響を

表1　変数の分布を調べる手法[2]

	数値変数	カテゴリー変数
1変数	基礎統計量 バーチャート	単純集計 バーチャート
2変数	相関分析 散布図	2重クロス集計
多変数	主成分分析	多重クロス集計 数量化Ⅲ類

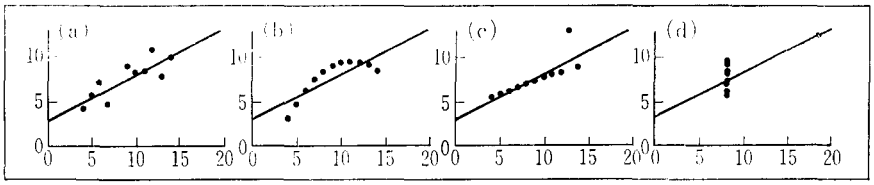


図 1 Anscombe(1973)のデータ[3][7]

受けているが，このような状況はカテゴリーデータを用いたりした場合におこりやすい．このようにグラフは，要約統計量のもつ欠点を積極的に補足してくれる．

3変数以上のデータは，主成分分析を用いて少ない次元に投影してデータの分布が把握できる．

データ解析においては，原表の様式・変数名・カテゴリーの決定等の前準備の後，データをコンピュータに入力するわけであるが，入力ミスや異常値の発見そしてデータ編集等を行なわなければいけない．後先が逆になるが，この入力ミスや異常値の発見に，1変数の要約統計量のグラフ表示である幹葉表示・箱ヒゲ図などが多く用いられている．また，最近の統計パッケージでは，正規確率プロット図も簡単に出力される．

4. 回帰分析とグラフ

4.1 回帰分析とグラフ

回帰分析に関して，そのアルゴリズムあるいは分散分析表が射影子あるいはピタゴラスの定理と結びつけて考えれば深い理解が得られることはよく知られている．

4.2 残差のプロット

ここでは，回帰分析のモデル・ビルディングで，残差のプロットが重要な役割を果たしていることを紹介したい．

J.P. SALL は，アメリカの1790年から1970年までの10年毎の人口データ(**図 2**)を用いて，良いモデルの探索を述べている[1]．

最初は，このデータに次の単回帰モデルを適用した．

人口＝$b0$＋$b1$*年

この結果，**図 3** のように満足な統計量が得られた．ここで満足していては，データ解析としては失敗である．

このモデルの残差を描くと次の残差プロットになる．

これから，モデルに2次の項が必要なことが示唆される．

人口＝$b0$＋$b1$*年＋$B2$*(年)**2

この多項回帰モデルをデータに当てはめると，**図 5** の残差プロットになる．

1940年と1950年の2点の残差は，戦後の景気後退による外れ値である．このため，この2点を1とし残りの点

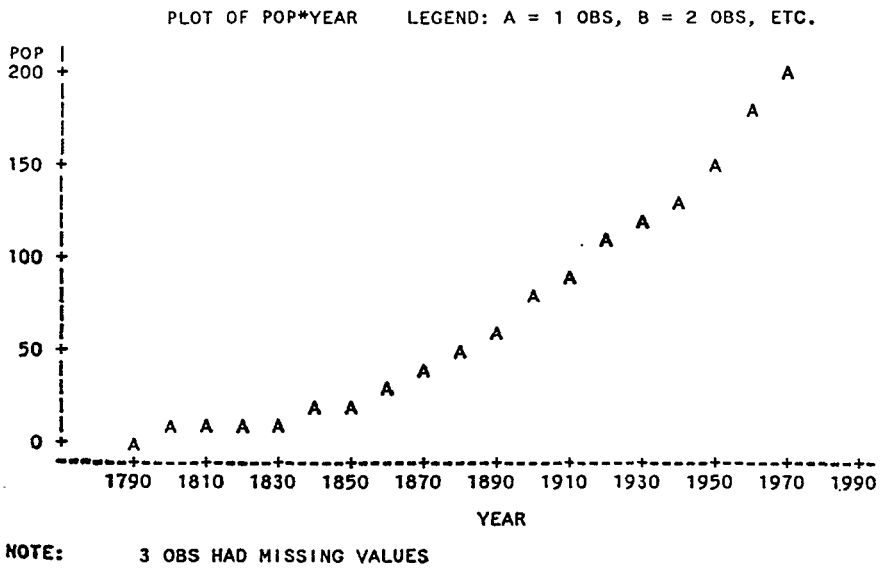


図 2 アメリカの人口データ[1]

DEP VARIABLE: POP

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	66336.469	66336.469	201.873	0.0001
ERROR	17	5586.293	328.605		
C TOTAL	18	71922.762			

ROOT MSE	18.127478	R-SQUARE	0.9223
DEP MEAN	69.767474	ADJ R-SQ	0.9178
C.V.	25.98271		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR HO: PARAMETER=0	PROB > |T|
INTERCEP	1	-1958.366	142.805	-13.714	0.0001
YEAR	1	1.078795	0.075928	14.208	0.0001

図 3　単回帰モデルの分析結果[1]

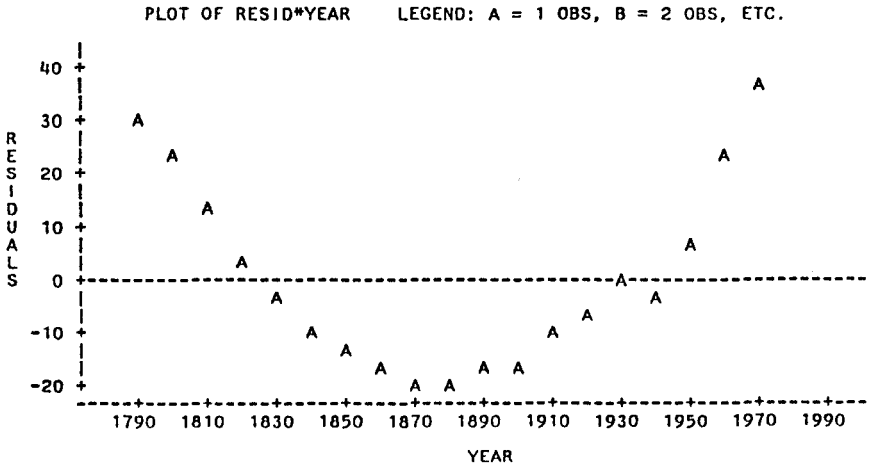


図 4　単回帰による残差プロット[1]

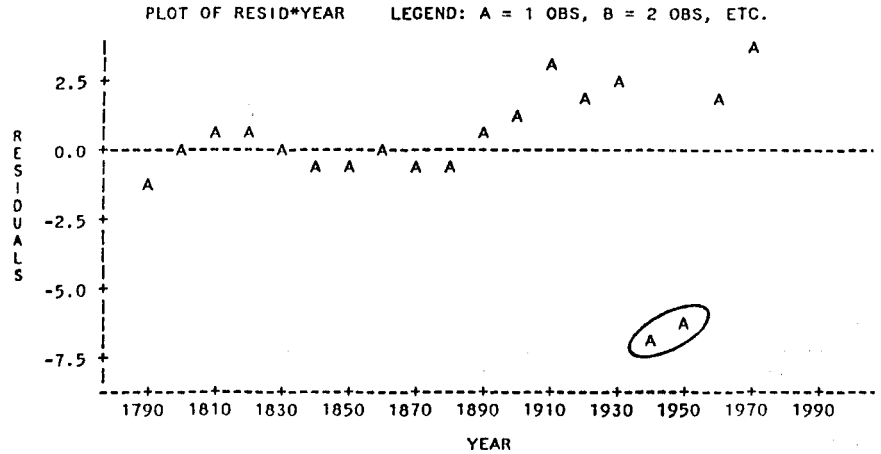


図 5　多項回帰モデルの残差プロット[1]

DEP VARIABLE: POP

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	3	71909.581	23969.860	27277.636	0.0001
ERROR	15	13.181051	0.878737		
C TOTAL	18	71922.762			

ROOT MSE	0.937410	R-SQUARE	0.9998
DEP MEAN	69.767474	ADJ R-SQ	0.9998
C.V.	1.34362		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR HO: PARAMETER=0	PROB > |T|
INTERCEP	1	20982.754	288.247	72.794	0.0001
YEAR	1	-23.366377	0.307114	-76.084	0.0001
YEARSQ	1	0.006506678	.00008175416	79.588	0.0001
DUMMY	1	-8.741519	0.779307	-11.217	0.0001

図 6　ダミー変数を用いたモデルの分析結果[1]

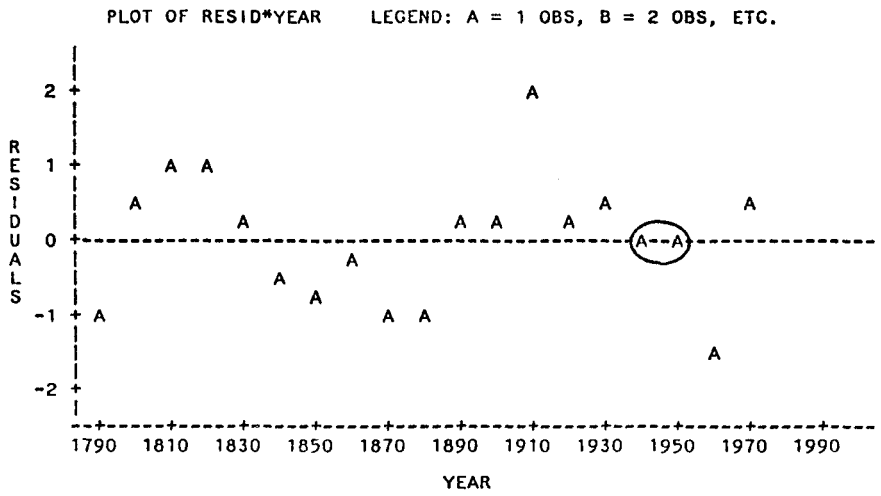


図 7　ダミー変数を用いたモデルの残差プロット[1]

を0とするダミー変数をモデルに導入する.

人口$=b0+b1$*年$+B2$*(年)**2$+B3$*DUMMY

これによって，次の分散分析表と残差のプロットが描かれる．この2点は，870万人（回帰係数−8.742）の落ち込みを示しており，回帰係数は1％で棄却される.

4.3　残差と非線形回帰分析

この後，このデータにロジスティックモデル・分割モデル・誤差の絶対値を最小化するＬＡＶ回帰分析・繰り返し重み付き回帰分析を非線形回帰分析ＮＬＩＮで説明している．ここで重要なのは，これらの一連の回帰モデルが，次の重み付き誤差平方和によって統一できることである．また，ＬＡＶ回帰分析はＬＰを用いて，通常の回帰分析は2次計画法の問題になる.

$$SSE=\sum_i w_i e_i^2$$

4.4　多重共線性

回帰分析において，残差の検討の重要性を述べた．この他の問題として，回帰分析の変数選択がある．この時に問題になるのは，多重共線性である．多重共線性とは，説明変数の間に高い相関がある場合に，回帰係数や統計量に悪影響をおよぼす．**図8**はこの多重共線性を説明する概念図である．上図は，2個の説明変数 MAXPULSE と RUNPULSE の間に高い相関があることを示しており，Z軸方向は目的変数を表わしている.

4.5　変数選択問題における図式表示

4.2でみたように時系列データであれば，残差の時系列プロットを描けばよかった．一般のデータでは，残差

の検討として偏回帰プロットや残差の分布を箱ヒゲ図・正規確率プロット・幹葉図等を用いて検討できる．志村［4］は，偏回帰プロットを用いた変数選択を論じている．筆者も別の角度からこの問題に関して意見をもっているが割愛する．

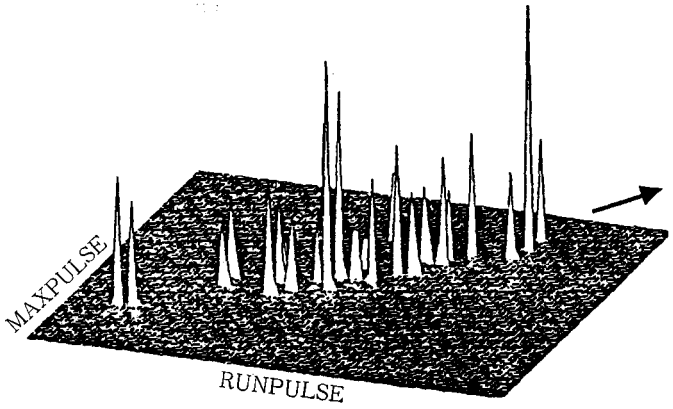
5. 品質管理

日本における品質管理の成功は，わかりやすさ使いやすさを追及した大衆運動だったことにあるのではないかと想う．そして，グラフがその手助けをしている．

SAS/QC の責任者である Rodriguez［8］も日本的な品質管理に注目し，数回にわたり調査のため訪日している．図9は，彼が開発している SAS/QC の箱型管理図である．また，SAS/GRAPH を用いて図10のようなパレート図や特性要因図が描ける．

6. SASのレポート機能

SASは，プログラム言語としてみても，PL/1やコボルに比べて10倍以上の威力がある．OSは別として，ソフトウェアを1つ選ぶとすれば「All in one system」のSASを選べば間違いがないであろう．図は身長と体重のデータを読み込んで年齢と性別毎にそ



共線性のないデータの3次元表示

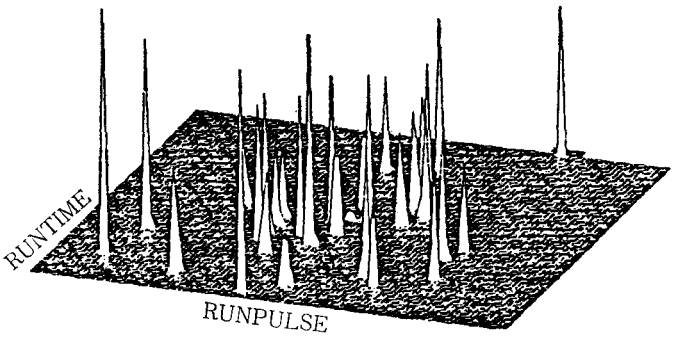


図 8 多重共線性のあるデータの模式図［1］

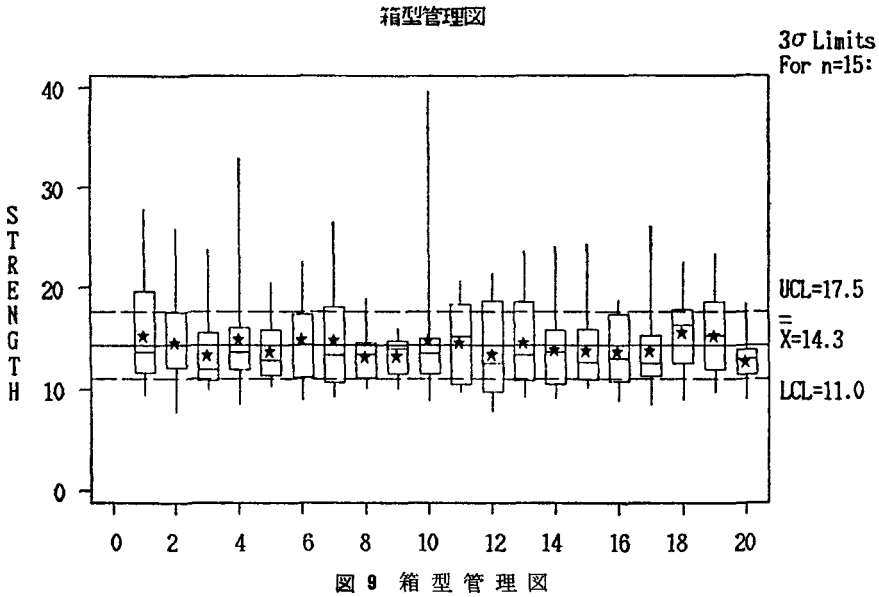
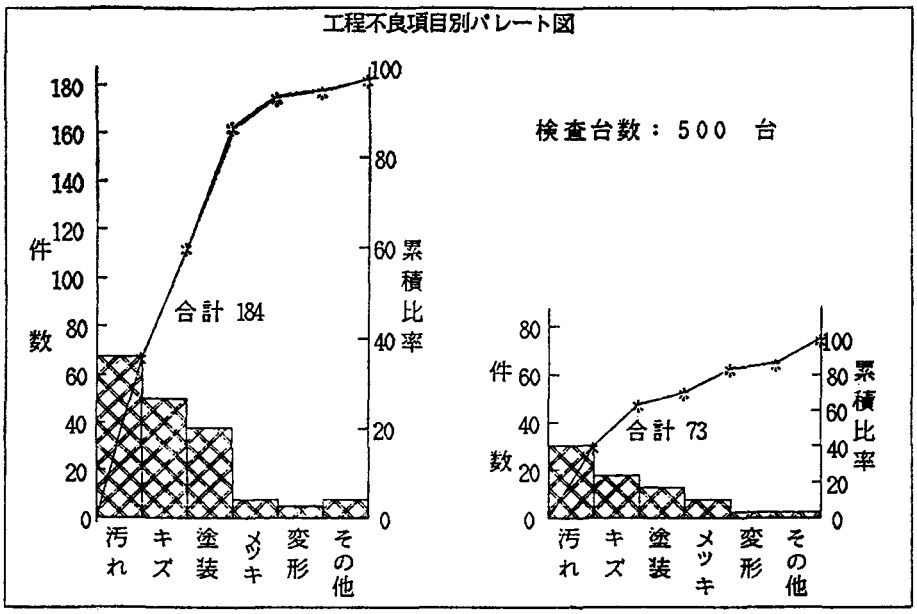


図 9 箱 型 管 理 図

A．パレート図



B．特性要因図

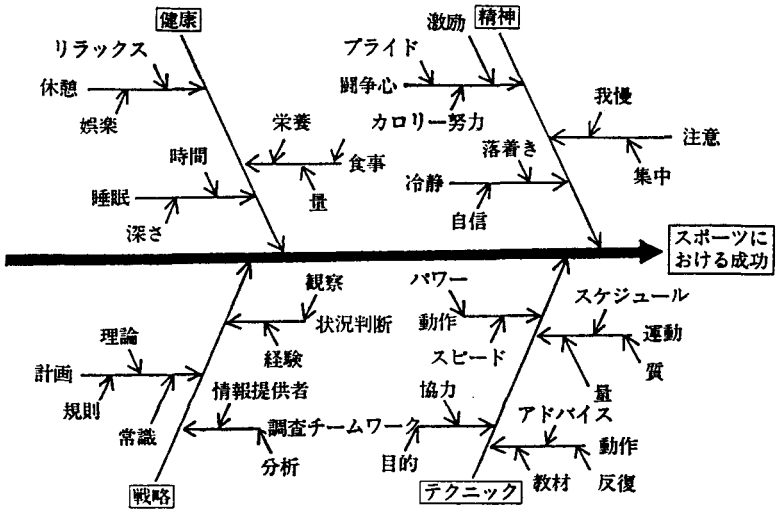


図 10　パレート図や特性要因図

の平均図を計算し，肩幅と胴の長さに反映してレポートしたものである．ほんの数行でプログラムできる．

7. 終わりに

日本においては，大学教育において正規の統計あるいはデータ解析がとりあげられていないが，企業をはじめとして広く使われている．これに大きく貢献したのはソフトウェアの普及であろう．一方，品質管理の成功はグラフ等の助けを借りたわかりやすさに加え，最近ではソフトウェアの普及や統計手法の取り込みが見られる．諸先輩方には，生意気とお叱りを受けそうだが，ORも一皮むけた伊達ものになる必要があるのかも知れないと考える今日この頃です．

文　献

[1] J. P. Sall (新村訳) (1986)，SASによる回帰分

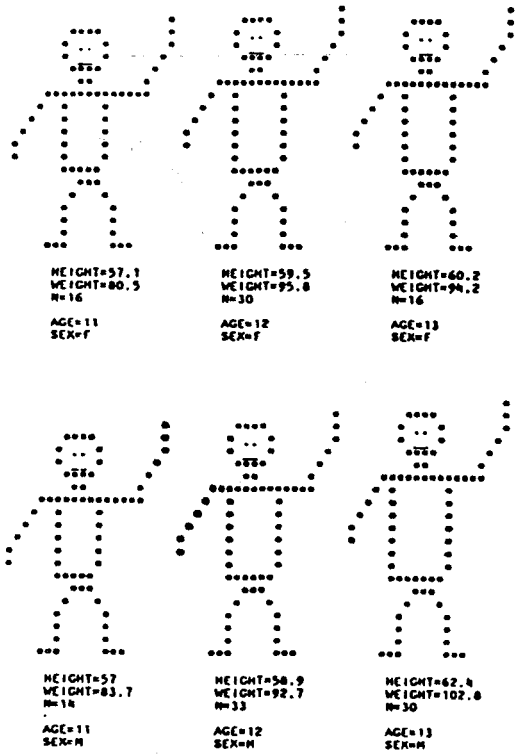


図11　レポート機能によるグラフ表示

析の実践，朝倉書店

[2] 高森　寛，新村秀一(1987)，統計処理エッセンシャル，丸善

[3] 市川伸一，大橋靖雄(1987)，ＳＡＳによるデータ解析入門，東大出版会

[4] 志村健一，大谷部恵子，吉澤　正(1986)，散布図による変数選択，ＪＳＱＣ第16回年次大会研究発表要旨，65-68

[5] J.W.Tukey(1977). "Exploratory Data Analysis", Massachusetts, Addison-Wesley

[6] S. H. C. duToit, A. G. W. Steyn, R. H. Stumpf, "Graphical Exploratory Data Analysis", Springer-Verlag.

[7] F.J.Anscombe(1973). "Graphs in statistical analysis", The Amer. Statist. 27 : 17-22

[8] R.N.Rodriguez (大橋靖雄訳) (1986). "統計的品質管理におけるグラフィック手法のデータ解析への応用",「品質」, 16, [4], 55-65.

[9] 新村秀一 (1987). "体験に基づく汎用統計パッケージの紹介",「品質」, 17, [3], 14-21