日本統計学会誌
第21巻，第3号（増刊号），1992年
333頁～351頁

統計的グラフィクスの最近の発展†
——顧客の創造と視覚表現の諸法——

後藤昌司*1　白旗慎吾*2　垂水共之*3　馬場康維*4
安田嘉純*5　松原義弘*1　余田明夫*1　脇本和昌*3

Present Development of Statistical Graphics: Creation of Clients and Methods for Visual Representation

Masashi Goto*1　Shingo Shirahata*2　Tomoyuki Tarumi*3　Yasumasa Baba*4
Yoshizumi Yasuda*5　Yoshihiro Matsubara*1　Akio Yoden*1　Kazumasa Wakimoto*3

統計的グラフィクスは，データの記述・探索の枠を超えて，モデルの構成とあてはめ，さらには推測と決定にまで適用範囲を広げ，データの解釈の全過程で利用されるようになってきている．統計的グラフィクスの研究・開発では，その実際的活用が必須であるが，「特殊」を普遍化するよりもその「特殊」により磨きをかけることが重要である．本稿では統計的グラフィクスを研究する立場よりは，それを活用・改善し，開発する立場に重きをおき，若干の省察を試みる．最初に，データ解析における用途別諸法として，データ省察，多変量データの順序づけ，判別解析，生存時間データの解析における幾分特殊な統計的グラフィクスを概説する．統計的データ解析において，グラフィクスを用いることの利点は，定性的，定量的を問わず，多くの情報を直観的に把握できることにある．しかし，それと同時に，客観性に欠けるグラフィクスを用いると，様々な誤った解釈をするという危険もはらんでいる．データの視覚表現が，データのもつ情報を直観的に把握することを可能にし，かつ客観的な解析に耐えるようにするには，視覚表現が直観のどのような側面に訴えるかを明らかにし，それを統計量として捉えることも重要である．ここでは，統計量の視覚による解釈として，ノンパラメトリック諸法とグラフィクスとの対応について若干の考察を加える．

統計的グラフィクスの魅力は，その方法論を研究・開発・適用を通して深耕させるか，あるいは顧客を「六方」に拡大してそれらの顧客の支援をいかに獲得するかに負うている．本稿の最後では，統計的グラフィクス・ソフトウェアの最近の動向を述べ，統計的グラッフィクスの研究・開発と顧客の創造のあり方を論じる．

1. 序

統計学は不確実な要素を伴うデータの収集，解析，解釈を問題にする．それは応用科学であり，その宿命として社会環境の変化に大きく影響される．というよりも，社会環境の変化に即応しなければ存立していくことができない．計測技術と通信技術の発展はデータの収集と問題

*1 塩野義製薬・解析センター，〒564 吹田市泉町 1-22-41
*2 大阪大学・教養部，〒560 豊中市待兼山町 1-1
*3 岡山大学・教養部，〒700 岡山市津島中 2-1-1
*4 統計数理研究所，〒106 東京都港区南麻布 4-6-7
*5 千葉大学・工学部，〒260 千葉市弥生町 1-33

† 本論文は，日本統計学会60周年記念事業の一環として1991年7月26日神戸で行われた，記念講演会で発表したものの方法的背景をまとめたものである．

提起の形態を変えつつある．コンピュータは新しい解析方法，とくにデータ中心の解析法を生み出してきている．統計的グラフィクスは，データの記述・探索の枠を超えて，モデルの構成とあてはめ，さらには推測と決定にまで適用範囲を広げ，データの解釈の全過程で利用されるようになってきている．統計学の最近の研究動向を概観した米国数理科学委員会 BMS の報告書 (1989) には 12 個の研究分野について，統計学の理論と実践における最近の研究成果が要約され，統計学の将来の方向性が示唆されている．その一つがグラフィカル・データ解析と対話型統計計算である．そして，その最終節で，他の学問分野との協同研究がますます盛んになること，およびデータの収集と解析の両面での技術が今後も急速に発展することに呼応して，統計的方法論では標本再利用法とコンピュータ・グラフィクスの発展の期待されることが強調されている．

さらに，このような動向を実際に反映するかたちで，米国統計学会 ASA では“Journal of Computational and Graphical Statistics (JCGS)”の 1992 年の発刊が計画されている．その嚆文句は次の通りである：「JCGS では，計算手法とグラフィカル手法の総合報告，評論論文とともに原著論文を公表していくつもりである．掲載論文では，これらの分野におけるその時点の最高の知識が提示されなければならない．提示内容は正確で，明確で統計的素養のある読者が理解できるものでなければならない．研究論文では，数値解析手法や組み合わせ手法，プログラミング環境，シミュレーション，グラフィカル表示，グラフィカル手法，グラフィカル知覚などをとりあげた（統計学にとって価値のある）コンピューティングとグラフィクスにおける新しい研究を包含する任意の分野について論じることができる．興味深い特別のトピックスには，エキスパート・システム，数式処理，スーパ・コンピュータなどに関するものがある」(Eddy, 1990)．

統計的グラフィクスは，上記のように，発展が著明で，しかも今後にかけての貢献をより多大に期待されているが，「人間」が一人では人間になれないのと同じく，その大半の「魅力」を顧客に負うている．本稿では統計的グラフィクスを研究する立場よりはそれを活用・改善し，開発する立場に重きをおき，若干の省察を試みる．

統計的グラフィクスが統計的データ解析過程で演じる用途・役割やその発展の過程については，既に二三の報告で体系化して論じた（脇本他，1977：脇本他，1979：Goto，1981：Wakimoto & Goto，1987：Goto，1987：後藤他，1988：Goto et al.,1991)．ここでは，その発展の系譜を詳細に紹介する代りに「衝撃」を与えたと考えられる印象的な話題を点描する．

- オーディネーション諸法あるいは変換表現法：定型的な多変量解析諸法を背景にして数個の変量の線形変換あるいは極座標変換に基づいて，主として 2 次元座標のグラフィカル表現を与える方法も数多く提案されている．前者の典型が主成分プロットとバイプロットであり，最近ではコレスポンデンス解析法や多次元尺度構成法に対応したグラフィカル諸法も含まれる．後者の典型が Andrews プロット，脇本の星座グラフと連結ベクトル・グラフ，馬場の順位グラフなどである．
- 射影志向型諸法あるいは平滑化諸法：これは最近に盛りあがりをみせているコンピュータ集中型諸法であり，この中には射影追跡法，交替条件付期待値(ACE)，一般化加法モデルなどが含まれ，典型として「グランドツアー」(Asimov, 1985) が有名である．
- 定型的方法との連携諸法：この主題は保守的であるが，相当に根強い「人気」をもっている統計的グラフィクスの諸法を含んでいる．目標表示型諸法あるいは構造評価型諸法ともよぶことができる．とくに，比較，回帰，分類・判別を目標とするが，それらの診断も意図した諸法が多い．最近では，個体評価と集団評価の対応を活かすかたちで AID や回帰樹木法などが高く評価されている．

なお，このような諸法は海外，とくに米国を中心として研究・開発されており，その発展の基盤を提供している米国統計学会 ASA の 150 周年記念寄稿文のなかで，Haaland (1990) が統計的グラフィクス分科会の活動を要約している．そこでは 1986 年に当分科会が正式に発足したことに精力的に貢献した Dr. W. S. Cleveland の活動を印象深く紹介し，次の台詞で結んでいる：「統計的グラフィクスは，科学者の間に深く，また急速に浸透してきた．実際に，当分科会の歴史の大半は，他の科学分野に対する支援の労を中心にして展開されている．統計的方法論のなかで統計的グラフィクスほど世に受け入れられているものはない．実際に，統計的グラフィクスに携っている当事者は，ツールの発明が 10 %の成功にしかすぎず，残りの 90 %が当該のツールの採用にかかっている（Bill Cleveland の言より引用）と主張することができる」．因に，日本統計学会では，50 周年記念国際円卓会議が開かれ，その招待セッションの主題が "Graphical Methods in Statistics" であり，また『日本統計学会誌』にはこれまで 8 編の統計的グラフィクス関係の論文が掲載されているが，その殆どが筆者の一人（脇本）を中心とする研究者の手でなっている（Wakimoto & Taguri, 1974 : Taguri et al., 1976 : Wakimoto, 1977 : Wakimoto, 1981 : Mizuta & Kawaguchi, 1983 : Wakimoto & Shirahata, 1984 : Shirahata, 1990 : Goto et al., 1991）．

最近の学会・シンポジウムの特別セッション：統計的グラフィクス

第17回　日本行動計量学会大会 (1989)：データの視覚表現の諸法と実際（特別セッション．組織者：後藤昌司）．
第47回　ISI (1989) 招待論文セッション：Dynamic Presentation of Multivariate Data (Organized by Denby, L.)．
第 8 回「分類の理論と応用に関する研究会」シンポジウム (1990)：データの視覚化（共通テーマ）．
文部省科研費シンポジウム (1991)．統計的グラフィクスとその応用（組織者：脇本和昌・白旗慎吾）．

2. 統計的グラフィクスの用途別開発

2.1 データ省察用諸法

統計的グラフィクスが最初に脚光を浴びたのは 1960 年代の Dr. J. W. Tukey の主導による探索的データ解析の提唱と深く連動しているようである．そこでは，「データに語らせる」ことから，有意な仮説（モデル）あるいは生産的知見を導き出すことを志向しており，そこで提案された諸法はデータ省察の場で活用され，ますます発展・拡大している．とくに，幹葉表示とボックス・ウィスカー・プロット（箱ひげ図）はよく知られている．同様の主旨のもとに，確率プロッティング法も Wilk & Gnanadesikan (1968) を始めとしてその適用の拡大がはかられたが，この方法は，どちらかといえば，データ省察用ツールというよりは後続の解析へ切り口を与える予備的ツールであるというのが適切かもしれない．Q-Q プロッティング，P-P プロッティング，ガンマ・プロッティングなどがよく用いられている．

例えば，ボックス・ウィスカー・プロットやスキーマティック・プロットは，探索的データ解析（Tukey, 1970, 1977）の代表的道具として汎用されている．以降では，これらのプロットおよびその変法を総称してボックス・プロットと呼ぶ．現在，ボックス・プロットが組み込まれていない統計ソフトウェアはないといっても過言ではない．ただし，SAS, SPSS, Minitab などの代表的な統計ソフトウェアではボックスを構成する四分位点の定義が異なっている (Frigge et al., 1989)．いま，n 個の順序付き観測値を $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ とするとき，それらの統計ソフトウェアで用いられている四分位点の定義および外れ値の度合いを記号で区別して表すため，内壁（四分位点 $\pm k_1$ 四分位幅）と外壁（四分位点 $\pm k_2$ 四分位幅）の計算に用いる

表1　第1四分位点 Q_1 の定義

定義	第1四分位点	j と g の定義	k_1	k_2
1	$Q_1=(1-g)x_{(j)}+gx_{(j+1)}$	$j=[n/4],\ g=n/4-j$	1.0	1.5
2	$Q_1=x_{(j)}$	$j=[n/4+1/2]$		
3	$Q_1=\begin{cases}x_{(j)} & ,g=0\\ x_{(j+1)} & ,g>0\end{cases}$	$j=[n/4],\ g=n/4-j$		
4	$Q_1=(1-g)x_{(j)}+gx_{(j+1)}$	$j=[(n+1)/4],$ $g=(n+1)/4-j$		
5	$Q_1=\begin{cases}(x_{(j)}+x_{(j+1)})/2 & ,g=0\\ x_{(j+1)} & ,g>0\end{cases}$	$j=[n/4],\ g=n/4-j$		
6	$Q_1=(1-g)x_{(j)}+gx_{(j+1)}$	$j=[[(n+3)/2]/2],$ $g=[(n+3)/2]/2-j$	1.5	3.0
7	$Q_1=(1-g)x_{(j)}+gx_{(j+1)}$	$j=[n/4+5/12],$ $g=(n/4+5/12)-j$		
8	$Q_1=(1-g)x_{(j)}+gx_{(j+1)}$	$j=[n/4+1/2],$ $g=(n/4+1/2)-j$		

k_1 と k_2 の値を表1に与える．なお，定義1から定義5まではSASのオプション，定義6はTukey (1977)，定義7はHoaglin & Iglewicz (1987)，定義8はCleveland (1985) に依る．また，定義8で描かれるボックス・プロットは上下の外壁が標本のそれぞれ90％点と10％点に対応し，k_1 と k_2 に対応する値はない．表1の［•］はGauss記号を表す．これらの定義を適用して得られるボックス・プロットは，同一のデータに対する結果でも，様相の異なる印象を与える場合もある（松原，1989：余田他，1990）．ボックス・ウィスカー・プロットやスキーマティック・プロットは本来，データの分布の形状，とくに対称性と外れ値の点検を意図して提案されたものであり，実地に往々にみられる「比較」の場での適用は考慮されていない．すなわち，これらのプロットの適用では標本サイズが考慮されていない点に留意することが必要である．したがって，これらのプロットは不等標本サイズの比較表示に向いていない．McGill et al. (1978) はこの欠点の改良版としてノッチド・プロットを提案している．その改良点は，中央値の比較を行えるようにしたことである．ノッチド・プロットでは，標本サイズを箱のくびれのない部分の横幅，および中央値の信頼度をくびれの縦の長さに反映させている．すなわち，箱のくびれのない部分の横幅を標本サイズ n の平方根 $\sqrt{n}$ に比例させ，箱の左右のくびれの縦の長さを中央値の信頼区間に相当するように描く．中央値の信頼区間は［中央値−信頼幅，中央値＋信頼幅］であり，この信頼幅は

$$1.7\{1.25(Q_3-Q_1)/(1.35\sqrt{n})\}$$

で定義される．ここに，Q_1 と Q_3 はそれぞれ第1四分位点と第3四分位点を表す．さらに，上述のボックス・プロットは箱の部分に相当するデータの分布の密度を表さない．Benjamini (1988) はこの欠点を補う二つの変法として，データの密度情報を箱の側面に盛り込むヒスト・プロットとヴェイス・プロットを提案している．両プロットがこれまでのボックス・プロットと異なるのは矩形の箱の部分であり，それ以外の部分はスキーマティック・プロットと同一である．例えば，前述の観測値の分布状況をグラフィカル表示するときのヒスト・プロットでは，第1四分位値，中央値，第3四分位値での密度がそれぞれ第1八分位値と第3八分位値，第3八分

位値と第5八分位値，第5八分位値と第7八分位値の間の距離の逆数によって推定され，箱の上辺，中線，底辺の横幅をこれらの推定値に比例させた多角形が描かれる．したがって，ヒスト・プロットではスキーマティック・プロットでの箱が中線を共通の底辺としてもつ左右2辺の長さの等しい台形を重ね合わせた形におきかえられる．因に，ヒスト・プロットという呼称は，このプロットが第1八分位値，第3八分位値，第5八分位値，第7八分位値の各々を境界値とする5個の柱で構成されるヒストグラムとスキーマティック・プロットを合成したものになることに由来している．他方，ヴェイス・プロットでは，四分位範囲内で横幅を密度に比例させた壺様の形状が描かれる．この密度の推定には核推定法が用いられている．すなわち，四分位範囲内に存在するある点zでの密度が

$$f(z)=\frac{1}{hn}\sum_{i=1}^{n}W\left(\frac{x_{(i)}-z}{h}\right)$$

と推定される．ここに，hは核ウィンドウ幅を表し，密度軌跡の滑らかさを調節する．Wは対称な重み関数（例えば，余弦，正規分布，バイ・スクエア，ボックス・カーなど）である．例えば，ボックス・カーの場合に，Wは

$$W(u)=\begin{cases}1, & |u|\leqq 0.5\\ 0, & |u|>0.5\end{cases}$$

である．さらに，ヒスト・プロットとヴェイス・プロットのいずれにもデータの中央値に横幅に合わせて淡い影を付すことで信頼区間を呈示できる．

Bacon-Shone & Fung (1987) は単一変量データおよび多変量データにおける1個ないし複数個の外れ値を視覚的に検出するのに有用なQ-Qプロットを提案している．このQ-Qプロットでは，外れ値を1個から想定した$m(0<m<n)$個まで順に1個ずつ増した場合のm個のプロットのうち，直線からのずれの最も大きなものを吟味することによって外れ値の個数とその観測値を同定することができる．このとき，いずれのプロットでも直線からのずれが殆ど見られない場合には，この標本に外れ値が存在しないと結論づけることになる．この方式は，マスク効果（数個の外れ値が相互に近接して存在する場合に外れ値として同定されにくいこと）の影響を受けないという利点をもつ．

いま，$X=(x_1, x_2, \cdots, x_n)^T$はサイズ$n$の確率標本であり，$p\times 1$ベクトル$x_i$は互いに独立に多変量正規分布$N(\mu, \Sigma)$に従うと仮定する．ここに，$p$は変量数を表し，平均$\mu$と分散共分散行列$\Sigma$は未知である．標本平均ベクトルを$\bar{x}=(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$，標本散布行列を$S$で表す．ここに，$S_{jk}=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$, $j, k=1, 2, \cdots, p$であり，x_{ij}は第i個体の第j変数の観測値を表す．このとき，この標本にm個の外れ値$(x_{t_1}, x_{t_2}, \cdots, x_{t_m})$が存在すると想定し，その外れ値の添字ベクトルを$t=(t_1, t_2, \cdots, t_m)^T$で表す．ここに，$1\leqq t_i\leqq n$であり，$\{t_i\}$はすべて異なる．外れ値と想定した$m$個の観測値を除くすべての観測値を用いて求めた標本散布行列を$S_{(t)}$で表す．このとき，統計量

$$R_{(t)}=|S_{(t)}|/|S|$$

は，想定したm個の観測値が外れ値であるか否かを調べるのに有用である（Wilks, 1962）．ここに，$|S|$と$|S_{(t)}|$はそれぞれSと$S_{(t)}$の行列式を表す．$R_{(t)}$は$0\leqq R_{(t)}\leqq 1$の範囲にあり，$R_{(t)}$の値が小さいほど，外れ値と想定したm個の観測値がその標本から外れていることを示唆する．ところで，サイズnの標本から相異なるm個の観測値をとり出す場合に，そのとり出し方は${}_nC_m$通りある．標本から外れ値と想定した相異なるm個の観測値を除くすべての場合について$R_{(t)}$

を計算する．このとき，$R_{(t)}$ は ${}_nC_m$ 個求まる．${}_nC_m$ 個の $R_{(t)}$ のうち，最小値を $Z_{(t)}$ で表す．この $Z_{(t)}$ が $\{R_{(t)}\}$ での残りのすべての値より極端に小さければ，この $Z_{(t)}$ の計算時に除外した m 個の観測値 $(x_{t_1}, x_{t_2}, \cdots, x_{t_m})$ が外れ値である可能性が高いと考えられる．

外れ値が存在しないときの帰無分布に関して Bacon-Shone & Fung (1987) は

$$W_{(t)}=-\{n-(p+m+3)/2\}\log R_{(t)}$$

が近似的に自由度 pm の χ^2 分布に従うこと (Box, 1949) を利用し，その近似分布からの $W_{(t)}$ の期待分位点に対して $W_{(t)}$ をプロットすることにより外れ値が視覚的に検出できることを示唆している．

実際には，大きな $W_{(t)}$ (いいかえれば，小さな $R_{(t)}$) にしか興味がないので，想定した外れ値の個数が m であるとき，大きな方からせいぜい $n-m+1$ 個を吟味するだけで十分である．いま，1から $n-m+1$ まで降順に順序づけた $W_{(t)}$ を $W_{(t)}^{(k)}$ で表す $(k=1,\cdots, n-m+1)$．このとき，$W_{(t)}^{(k)}$ に対する大きい方から $n-m+1$ 個の期待分位点は

$$\chi_p^{2(k)},\ k=1, 2, \cdots, n-m+1$$

で近似することができる．ここに，$\chi_p^{2(k)}$ は累積 p 値

$$1-k/(n-m+2),\ k=1, 2, \cdots, n-m+1$$

に対応する自由度 p のカイ二乗分布の分位点である．この方法では

$$(\chi_p^{2(k)},\ W_{(t)}^{(k)}),\ k=1, 2, \cdots, n-m+1$$

すなわち X 軸に $W_{(t)}^{(k)}$ の期待分位点の近似値 $\chi_p^{2(k)}$，および Y 軸に順序付けた $W_{(t)}$ をプロットすることによって Q-Q プロットが作成される．

2.2　多変量データの順序づけ

多変量データに対する凸包はすべての観測値を含む最小の凸多面体を構成することで得られる．このとき，多変量観測値の同時あるいは周辺での極値はその標本に関する凸包の頂点として定義できる (Barnett, 1976)．これらの凸包の頂点をすべて捨て去り，残りの観測値について新たな凸包を構成することが凸包ピーリングと呼ばれる (Eddy, 1982)．これは単一変量での刈り込み概念の一般化になっている．これらの基本概念を組み合わせ，さらに標本の重心から観測値までの Euclid 距離，あるいは観測値の散布方向などを一つの基準とすることにより，多変量観測値のいくつかの順序づけが行える．

いま，p 次元の n 個の観測値が得られたとする．このとき，凸包ピーリング，および観測値の重心から Euclid 距離を利用した順序づけを次のように実行することができる．

①$i=1$, $n_0=0$ とおく．

②$m=n-\sum_{j=0}^{i-1} n_j$ 個の観測値をすべて含む最小凸包を構成する．この凸包の頂点に対応する観測値の集合を P_i，および集合サイズを n_i とする．

③P_i 内の観測値を全観測値の重心からの Euclid 距離を用いて $n_{i-1}+1$ 番目から順序付ける．

④P_i に含まれる観測値を捨てる：$i=i+1$ とする．

⑤以降，②から④を反復し，m が2以下のときに停止し，残された m 個の観測値を③と同様に順序づける．

上記の順序づけを利用することにより，極値，中央値，範囲，四分位点などの単一変量での順序統計量の概念をより高次に拡張することができる．とくに，2次元の場合にはボックス・ウィスカー・プロットと類似なグラフィカル表現が可能である．

いま，2変量確率変数Xのn個の観測値の確率標本を$x_1,\cdots,x_n$で表し，新たに次の集合を定義する．すなわち，全観測値集合をQ_n，上記の順序づき観測値を順位の低い順に捨て去り，そのたびに構成される凸包に含まれる観測値の集合を$Q_{n-1}, Q_{n-2},\cdots$で表す．これらの集合は包含関係

$$Q_n \supset Q_{n-1} \supset Q_{n-2} \supset \cdots$$

にある．このとき，任意の$k(1\leq k\leq n)$に対して，$\{x_i\}$がQ_kに含まれる割合は

$$G_n(Q_k)=\frac{1}{n}\sum_{i=1}^{n} I\{x_i \in Q_k\}$$

である．ここに，$I\{x_i \in Q_k\}$は$x_i \in Q_k$であるときに1，それ以外のときに0をとる指標関数である．したがって，$G_n(Q_k)$は階段関数であり，単一変量の場合の経験累積分布関数に相当する．また，ボックス・ウィスカー・プロットのヒゲの先端にあたるデータの最小値と最大値はそれぞれ$G_n^{-1}(1/n)$と$G_n^{-1}(1)$に対応する点と凸包，およびボックスの上辺と下辺を構成する第1四分位点と第3四分位点はそれぞれ$G_n^{-1}(0.25)$と$G_n^{-1}(0.75)$に対応する凸包に相当する．ここでは，このプロットを凸包プロットと呼ぶ．より一般的に，$G_n^{-1}(Q_k)$は2変量分布の標本確率等高線図を与える．凸包ピーリングに代えて，最小の角度をもつ凸包の頂点をピーリングの各ステップでとり除く点ピーリング（Eddy, 1981 : Friedman & Rafskey, 1981）による順序づけも示唆される．しかし，この順序づけでは標本確率等高線図は構成できない．

2.3 判別解析グラフィクス

通常の線形判別関数法では，判別方式は訓練標本に基づいて構成される．しかし，訓練標本に外れ値などが存在すると，それが判別係数など諸種の推定値に影響を及ぼすために，性能が良好でないことが危惧される．したがって，判別方式を構成する前に外れ値あるいは影響観測値を同定することが重要になる．これには訓練標本の個体の観測値を2次元平面上にプロットすることが有効である．さらに，個体の観測値をプロットすることで2群の分離状況や各群の観測値が視覚化されるため，判別結果を解釈する際にも有用であると期待される．そのような方法では，判別方式自体が平面上で表現できることが望ましい．

訓練標本の個体の観測値をプロットするのに有用な方法として，MV (Mean-Variance) グラフがある (Chang, 1987)．ここでは，2群の判別問題に限定し，第i群の観測値ベクトルがp変量正規分布$N(\mu_i, \Sigma_i)$に従うとする．このグラフでは2群の平均ベクトルの差異と共分散行列の差異がそれぞれ反映されるように2通りの方法で観測値を変換し，変換値の対が2次元平面上にプロットされる．いま，xを訓練標本に属する観測値ベクトルとするとき，第i群の平均ベクトルμ_iの推定値$\hat{\mu}_i$を用いて，$y=x-\hat{\mu}_2, d=\mu_1-\hat{\mu}_2$と定義する．このとき，$y$の標本平均ベクトルは第1群で$d$，第2群で0となる．

最初に，2群の平均ベクトルの分離を最大にするようにZ_1を構成する．$p\times 1$ベクトルg_1に対して，線形結合$g_1^T y$を考え，Σ_1とΣ_2の推定値$\hat{\Sigma}_1$と$\hat{\Sigma}_2$を用いて

$$Z_1=(g_1^T y-g_1^T d)^2/(g_1^T\hat{\Sigma}_1 g_1)-(g_1^T y)^2/(g_1^T\hat{\Sigma}_2 g_1)$$

と定義する．この式の第1(2)項は第1(2)群の平均までの$g_1^T y$のMahalanobis距離の平方である．通常，g_1はZ_1を最大にするように選定される．ところで，併合標本で$\Sigma_1=\Sigma_2=\Sigma$の推定値$\hat{\Sigma}$を求め$g_1=\hat{\Sigma}^{-1}d$を選定すると，これは周知の線形判別関数の係数ベクトルである．また，$\hat{\Sigma}$の代わりに，$0\leq t\leq 1$となるtに対して$t\hat{\Sigma}_1+(1-t)\hat{\Sigma}_2$を用いることもできる．ここに，$t$は許容線形判別関数法（Anderson & Bahadur, 1962：後藤，1973）のミニ・マックス方

式と同様に求められる．

次に，2群の共分散行列の分離を最大にするように Z_2 を構成する．最初に，$G_2^T d=0$ となる $p\times(p-1)$ 行列 G_2 を選定する．簡便に，G_2 はランク1の $p\times p$ 行列 dd^T の固有値0に対応する $(p-1)$ 個の固有ベクトルで構成することができる．このとき，Z_2 を

$$Z_2=(G_2^T y)^T(G_2^T\hat{\Sigma}_1 G_2)^{-1}(G_2^T y)-(G_2^T y)^T(G_2^T\hat{\Sigma}_2 G_2)^{-1}(G_2^T y)$$

で定義する．$G_2^T y$ の平均ベクトルは第1群と第2群のいずれに対しても0となり，$G_2^T y$ の共分散行列は第 i 群に対して $G_2^T\hat{\Sigma}_i G_2$ となる（$i=1,2$）から，上式の第1（2）項は第1（2）群に対する $G_2^T y$ と $G_2^T y$ の平均ベクトル0までのMahalanobis距離の平方である．

観測値を上記の手順で変換し，各個体について (Z_2, Z_1) をプロットすることでMVグラフが得られる．このグラフでは2群の標本平均ベクトルの差が Z_1 軸，共分散行列の差異が Z_2 軸に反映されるために，2群の共分散行列の違いが彫琢されるだけでなく，線形判別関数が適用できるか否かを点検するのにも利用できる．なお，Z_1+Z_2 は

$$(x-\hat{\mu}_1)^T\hat{\Sigma}_1^{-1}(x-\hat{\mu}_1)-(x-\hat{\mu}_2)^T\hat{\Sigma}_2^{-1}(x-\hat{\mu}_2)$$

で近似できるため，MVグラフ上で判別関数を直線 $Z_1+Z_2=c$ で与えることも可能である．ここに，c は直線の切片を表し，この判別関数で訓練標本を分類したときの過誤率が最小になるように選定される．

のみはむしデータ（Lubischew, 1962）にMVグラフを適用した結果を図1に示す．図1の「○」はCC（Chaectonema Concinna）群21例，「●」はCH（Chaectonema Heikertinger）群31例の個体を示す．このグラフの縦軸，すなわち Z_1 軸方向に注目すると $Z_1=0$ の上方向にCH群，下方向にCC群の個体が散布しており，平均差だけで2群が完全に分離されることがうかがわれる．他方，横軸，すなわち Z_2 軸方向へのちらばりが若干にCH群の方で大きいが，ほぼ等しく散布していることがわかる．ここでも2群の共分散行列の等しいことが支持される．図1中の直線は $Z_1+Z_2=c$ に基づいて構成した判別関数（判別境界）である．この場合には2

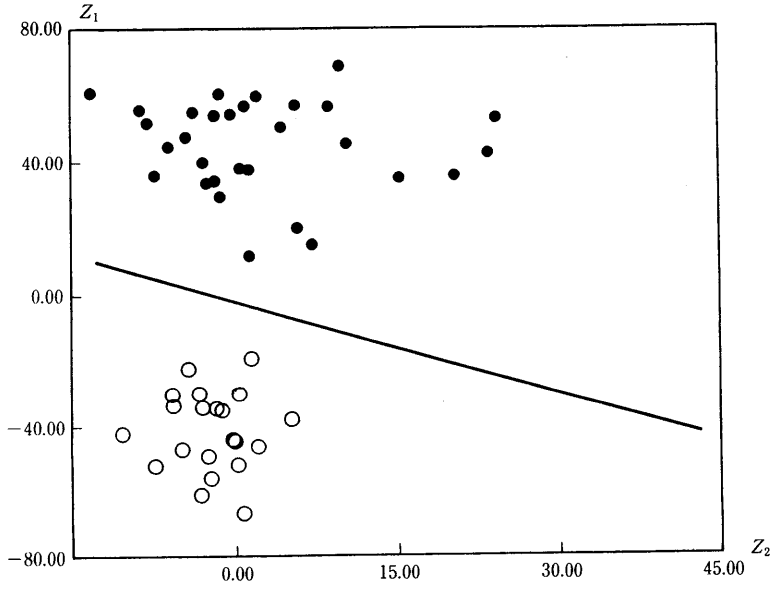

図1　「のみはむしデータ」に対するMVグラフ

群が完全に分離されているために c の値を一意に定めることは難しい．

MV グラフは訓練標本の個体の観測値をプロットするだけでなく，2群間の平均の差異と共分散行列の差異を同時に反映するように構成されるために外れ値の検出や等分散性の点検などに対して有用性の高いグラフであるが，このグラフの他に訓練標本の個体の観測値をプロットするグラフィカル接近法として星座グラフ（Wakimoto & Taguri, 1978）も提案されている．星座グラフは，p 個の各変数がその値の大きさに応じて 0 から π までの角度をもつベクトルで表され，それらの p 個のベクトルをつなぎ合わせてその最終点に星を描くことで作成される．このとき，各変数のベクトルの長さは，それら p 個の総和が 1 となる正の値をとるように与えられ，個体は星として半径 1 の半円内に表示される．この手法の難点はベクトルの長さを決める基準がないために再現性に乏しい点である．

判別方式の構成では，判別方式に付随するパラメータの推定が必要であるが，その推定に伴う標本変動が判別方式に反映されるため，得られた判別方式は標本変動による不確定要素を含むことになる．したがって，その判別方式に従って個体を割当てる際には，2者択一的な割当てを避け，不確定要素を考慮した判定保留領域を構成することが必要である．ここでは，不確定要素を定量的に表現したものを不確定度と呼び，不確定度に基づく判定保留領域の構成についてとりあげる．

個体の不確定度を表示する方法の一つとして「P_Δ プロット」を用いることができる（Critchley & Ford, 1985）．このグラフでは2個の母集団（群）に共分散行列の等しい正規分布が仮定され，個体の判別評点および不確定度が2次元平面上にプロットされる．さらに，このグラフ上で線形判別関数の信頼区間を表示できる．観測値ベクトル x を

$$x^*=QP(x-\mu_1)-\frac{1}{2}\delta$$

と変換する．ここに，P は $\Sigma^{-1}=P^TP$ を満たす正則行列，Q^T は第1列に $P(\mu_1-\mu_2)$ の大きさ1に規準化されたベクトルをもつ直交行列である．また，$\delta^T=(\Delta, 0, \cdots, 0)$ である．x^* は x が第1群からとられたときに $N(\delta/2, I)$，および第2群からとられたときに $N(-\delta/2, I)$ に従う．ここに，I は恒等行列である．x_1^* を x^* の第1成分とすれば，$l(x)=\Delta x_1^*$ が成りたつ．したがって，x_1^* は Δ で尺度化された x に対する判別評点である．ここで

$$a_i^2(x)=(x-\mu_i)^T\hat{\Sigma}^{-1}(x-\mu_i)\quad i=1, 2$$

とおくと

$$l(x)=\frac{1}{2}\{a_1^2(x)-a_2^2(x)\}$$

が成りたつ．このとき

$$(x^*-(-1)^{i+1}\delta/2)^T(x^*-(-1)^{i+1}\delta/2)=a_i^2(x)\quad i=1, 2$$

が成りたち，上式の左辺は x^* から第 i 群の重心までの Mahalanobis 距離の平方であるから，$a_i^2(x)$ および $l(x)$ は上記の変換に対して不変であることが示される．ところで，$\bar{x}_i$ の標本分布は $N(\mu_i, \Sigma/n_i)$ であるから，x_{ij} に変換を施して得られる x_{ij}^* の標本平均ベクトル $\bar{x}_i^*$ の標本分布は $N((-1)^{i+1}\delta, I)$ である．同様にして，標本共分散行列 $\hat{\Sigma}$ の標本分布は自由度 $n(=n_1+n_2-2)$，共分散行列 Σ の Wishart 分布 $W_p(n, \Sigma)$ に従うから，x_{ij}^* に基づく標本共分散行列 $\hat{\Sigma}$ の標本分布は $W_p(n, I)$ である．$a_i^2(x)$ で μ_i と Σ にそれぞれの推定量 $\hat{\mu}_i$ と $\hat{\Sigma}^{-1}$ を代入し，これを

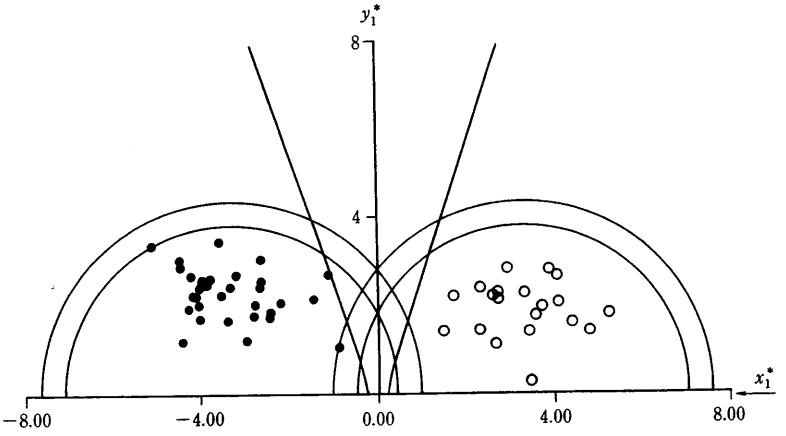

図2　「のみはむしデータ」に対する P_Δ プロット

$\hat{\alpha}_i^2(x)$ とおく．$\bar{x}_i^*$ と $\hat{\Sigma}^*$ の標本分布および $\hat{\alpha}_i^2(x)$ の不変性を利用すれば，x が与えられたときの $\hat{\alpha}_i^2(x)$ の分布を容易に導出できる．ここに，$\hat{\alpha}_i^2(x)$ は，自由度対 $(p, n-p+1)$，非心度 $n_i\Delta^2/4$ の非心F分布に従う．したがって，$\hat{l}(x)=\{\hat{\alpha}_1^2(x)-\hat{\alpha}_2^2(x)\}/2$ の分布は2個の非心F統計量の差の分布として表現できるため，$\hat{l}(x)$ の分散 $Var\{\hat{l}(x)\}$ を求めることができ，これを利用して $\hat{l}(x)$ の信頼区間を構成することができる．x_1^* を固定したときに $Var\{\hat{l}(x)\}$ は $x^{*T}x^*-x_1^{*2}$ に依存するため，$y_1^*=\sqrt{x^{*T}x^*-x_1^{*2}}$ とおけば，y_1^* が不確定度に相当する．したがって，一般に判別評点が同一であっても割当ての不確定度 y_1^* は個体によって異なる．この状況は (x_1^*, y_1^*) を2次元平面上にプロットすることで視覚的に表現できる．また，$\hat{l}(x)/Var\{\hat{l}(x)\}$ が近似的に正規分布に従うことを利用して判別関数の信頼区間を構成することができる．さらに，$\alpha_i^2(x)(i=1,2)$ が自由度 p のカイ二乗分布に従うことを利用して $100(1-\alpha)\%$ の許容領域を構成することができる．このとき，第 i 群の許容領域は，この平面の座標 $((-1)^{i+1}\Delta/2, 0)$ を中心とし，半径が自由度 p のカイ二乗分布の上側 $100(1-\alpha)\%$ 点の半円として表現される．

P_Δ プロットを「のみはむしデータ」に適用した結果を図2に示す．図2において，横軸は x_1^*，縦軸は y_1^* を示し，「○」と「●」はそれぞれCC群とCH群の個体を表す．線形判別関数は $l(x)=\Delta x_1^*$ で表せるため，直線 $x_1^*=0$，すなわち y_1^* 軸が判別境界となる．したがって，事前確率が等しいと見做した場合に x_1^* の符号によって個体が判別される．ここではCC群とCH群の個体がすべて正しく判別されている．また，2個の2重の半円はそれぞれの群の重心 $(-\Delta/2, 0)$ と $(\Delta/2, 0)$ を中心とし，内側の半円がデータの95％許容領域，および外側の半円がデータの99％許容領域である．この半円に基づいて外れ値あるいは影響観測値を同定することができる．このデータではいずれの群の個体も95％許容領域に含まれており，外れ値のないことが示唆される．また，y_1^* 軸に対称なV字型曲線は

$$\hat{l}(x)/Var\{\hat{l}(x)\}=\pm 1.96$$

を満たし，その曲線ではさまれた部分が判別関数 $l(x)$ の近似的な95％信頼区間を示す．この区間はゼロの判別境界を含んでいるために，この近似的な信頼区間内に落ちることが予測される対応個体は判定を保留すべきであろう．なお，このデータでは誤判別される個体はなく，判別関数の近似的な95％信頼区間内に落ちる個体もないことから，CC群とCH群は完全に分離されていることがうかがえる．

2.4 生存時間解析グラフィクス

生存時間分布は生存関数，ハザード関数，確率密度関数のいずれかで規定される．したがって，グラフィカル接近法もこれら3個の関数を基調にしたものが多い．一般に，分布の適合度や比較のグラフィカル表現には確率プロットが推奨されている（例えば，Wilk & Gnanadesikan,1968：田崎・後藤，1984）．確率プロットが累積分布関数に基づいていることと並行して，生存時間研究で適用されるハザード・プロット（Nelson, 1972）は累積ハザード関数に依拠して構成される．いま，死亡時間Tの確率密度関数を$f(t)$，分布関数を$F(t)$とすると，そのハザード関数と累積ハザード関数はそれぞれ

$$h(t)=f(t)/\{1-F(t)\}$$
$$H(t)=\int_0^t h(x)dx$$

で定義される．このとき，順序付き死亡時間$t_{(1)}\leq t_{(2)}\leq\cdots\leq t_{(m)}$が与えられているとすると，ハザード・プロットは$m$個の点$(H^*(t_{(i)}), t^*_{(i)})$のプロットからなる$(i=1,\cdots,m)$．これらの点がほぼ直線上に布置していれば，想定分布に適合していると判断される．ここに，$H^*(t_{(i)})$と$t_{(i)}^*$はそれぞれ$H(t_{(i)})$と$t_{(i)}$に対する想定分布に特有な関数による変換値を表す．

生存時間解析で頻用されている比例ハザード・モデル（Cox, 1972）の比例性や適合度に関する殆どのグラフィカル点検法は上記の累積ハザード関数あるいは対数累積ハザード関数のプロットに基づいている．比例ハザード・モデルでは共変量ベクトル$z=(z_1,\cdots,z_p)^T$をもつ個体の生存時間Tに対するハザード関数が

$$h(t:z)=h_0(t)\exp(\beta^T z)$$

で表される．ここに，$h_0(t)$は潜在基礎ハザード関数，βは未知の$p\times 1$回帰パラメータ・ベクトルである．このとき，潜在基礎ハザード関数$h_0(t)$にデータ適応型分布族，例えばベキ正規分布（Box & Cox, 1964 : Goto et al., 1983），対数ガンマ分布（Stacy, 1962 : Prentice, 1974），一般化z分布（Prentice, 1976）などを仮定した層化比例ハザードモデルのもとでの層間の比例性の点検には，分布を規定するパラメータと回帰パラメータβに関する階層仮説系列が構成できる．この結果として得られる分布の位置パラメータや尺度パラメータの推定値の層別プロットは層間の生存時間特性を解釈するための有用なグラフとなりうる（松原・後藤，1988）．

通常の回帰解析では，各種の回帰診断用グラフィクスが提案されている（例えば，Goto, 1981 : 松原・後藤，1984）．中途打ち切り観測値が含まれる場合には対応する個体の正確な応答がわからないために，それらのグラフィクスが利用できない．生存時間解析における外れ値や影響観測値の同定，あるいは（一般化）残差の視察用グラフィクスについては松原・後藤（1989）に要約されている．

3. 統計的グラフィクスの評価

3.1 統計的グラフィクスの認識と評価

統計的データ解析において，グラフィクスを用いることの利点は，定性的，定量的を問わず，多くの情報を直観的に把握できることにある．しかし，それと同時に，客観性に欠けるグラフィクスを用いると，様々な誤った解釈を与える危険もはらんでいる．統計的データ解析を行う者にとって，この統計的グラフィクスのもつ特徴の「直観的に」という点は長所であるとともに，短所ともなる両刃の剣である．したがって，多くの情報を含み，直観に訴えるという長所をもち，かつ客観性を備えていることが望ましいグラフィクスということになる．

Everitt (1987) は統計的グラフィクスのあり方と「芸術」の間に一線を画すことから，その統計的負荷に重きをおくことを強調し，3個の事例で主成分プロット，Andrews プロット，Chernoff の顔の比較・評価を試みている．その結果，表示力と解釈から主成分プロットの優越性を支持している．ただし，この評価には二三の疑問点がある．第1に，比較の対象になっている3方法の選択の問題である．一般に，多変量データのグラフィカル表現法には次元の縮小を意図するオーディネーション手法と記述的表現法がある．主成分プロットは前者，および Andrews プロットと Chernoff の顔は後者に含まれる．とくに，主成分プロットでは，2主成分への縮小の適切性を吟味することが必要である．このような目標・観点の違う手法を同一の土俵で評価することは公平性を欠く惧れがある．第2に，これらの3方法の背後にはそれぞれ提案者の意図と目標があり，評価指標もできるだけ客観的に多義に用意し，そのうえで査定することが必要である．例えば，Chernoff (1978) はそのような評価指標として16属性，および Gnanadesikan (1981) は「望ましい」の形容詞つきで7属性をあげている．第3に，統計的データ解析の過程では，統計的観点だけでなく術的側面に根ざす接近の姿勢も無視できない．とくに，多変量解析では結果の評価で「現象解釈による同定」(reification) が欠かせない．

Everitt (1987) は客観性について顔型グラフ，Andrews プロット，主成分プロットを評価の俎上にあげていたが，馬場 (1991) は上記3グラフのうち主成分プロットを星座グラフに代えた点で異なっている．馬場 (1991) は「グラフが共通認識や情報伝達の手段として用いられる場合には客観性が必要であるが，思考の節約や発見の手段として用いられる場合には必ずしも必要でない」と主張している．

3.2　統計量のグラフィカル解釈

データをグラフに表現し，そこからデータのもつ情報を直観的に把握しようという試みは相当に以前から行われている．ヒストグラムからはデータの形状に関する情報が得られ，星座グラフからは多変量データの全体像のみならず個体の様子を捉えることもできる．多変量データのグラフィカル表現として顔型グラフは一世を風靡し，大きな反響をよんだ．また，提案者が数理統計家としても一流と目される Dr. H. Chernoff であったことも幸いして，多くの適用がみられた．予想外の効果として，それまでの「統計的」といった形容詞の堅苦しさや統計家の「悪代官」的心象を払拭するのにも貢献したようである．これに触発された形で「木」表現法，「城」表現法，体型グラフ，GLYPH，星座グラフなどが次々と発表され，適用が試みられている．しかしながら，このようなグラフィカル表現は，例えば顔型グラフの場合に，とくに目や口の大きさや形のように人間がまず注目する部分にどの変量を割当てるかで印象が大きく異なることがある．グラフィカル表現はデータを直観的に把握することを容易にするが，表現するだけに留まるだけでは微妙な差異を問題にする場合に直観（願望）の衝突による水かけ論に終ることが危惧される．

データのグラフィカル表現が，データのもつ情報を直観的に把握することを可能にし，かつ客観的な解析に耐えるようにするには，表現が直観のどのような側面に訴えるかを明らかにし，それを統計量としても捉えることが重要であろう．逆に，統計量（その多くは直観により構成されている）が表す内容をグラフィカルに表現することも大事であろう．多くの場合に，グラフィカル表現で何らかの指針が得られても，その確認は定型的方法に頼ることが多いように思われるが，確認そのものもグラフに基づく統計量に頼ることが望ましい．

Fisher (1987) はグラフィカル手法による表示が後続あるいは前提としての統計解析（定型的方法）を示唆するが，それ以上のことは望めないとして，グラフィカル手法の中に統計量などの定型的方法の要素を組み入れることを提案している．その提案を確率プロッティングと比較，回帰，散布図（カイプロット）と連関の事例で具体化している．

確かに，グラフィカル（非定型的）手法と定型的手法の連携は大変に重要であるが，ある一つの目標（対立仮説）を規定してそれに応じた統計量を組み入れることは，そのグラフィカル表現法をその規定だけに絞られた用途に限定し，グラフィカル表現法がもつ本来の微妙な多様性や柔軟性を消してしまいそうである．そのような意味では，この提案は後続の統計解析への段差をなくすだけでなく段差を強調することもある．それにしても，カイプロットなどの独自的なノンパラメトリック諸法の提案が今後の一つの発展方向を与えることは確かであろう．そのような例として，白旗ら（Shirahata et al., 1985 : Shirahata, 1990 :白旗，1988）の研究も参考になる．

適合度検定：$X_1, \cdots, X_n$ を連続な分布関数Fからの確率標本と仮定する．このとき，帰無仮説

$$\mathrm{H}: F(x)=F_0(x)$$

の検定を考える．ここに，F_0 は指定された分布関数である．$Z_i=F_0(X_i)$ とおき，$Z_{(1)}<Z_{(2)}<\cdots<Z_{(n)}$ をその順序統計量とする．帰無仮説Hのもとでは $Z_{(1)}, \cdots, Z_{(n)}$ が単位区間での一様分布からの順序統計量であると考えられる．

点列 $P_0=O, P_1, \cdots, P_n$ とその間を結ぶ線分が，P_{i-1} から傾き $Z_{(i)}\pi$，長さ $\binom{n}{2}^{-1/2}$ の線分を引き，そのもう一方の端点を P_i とおくことによって作られる．ここに，O は原点である．さらに，線分 OP_n の中点に関し，この線分と対称な線分図が描かれ，その結果凸多角形が描かれる．これをDとする．帰無仮説が正しい場合には，Dがほぼ正多角形になると期待され，逆にDが正多角形から遠ければ帰無仮説が成立していないと考えられる（Shirahata et al., 1985）．

凸多角形の面積は正多角形のとき最大であることから，検定統計量がDの面積とされている．面積は

$$L=\binom{n}{2}^{-1}\sum_{i<j}\sin(Z_{(j)}-Z_{(i)})\pi=\binom{n}{2}^{-1}\sum_{i<j}\sin|Z_i-Z_j|\pi$$

と求められる．

対称性の検定：$X_1, \cdots, X_n$ を連続な分布関数 $F(x)$ からの確率標本と仮定し，$|X_1|, \cdots, |X_n|$ の間での $|X_i|$ の順位を R_i^+ とおく．帰無仮説

H：Fが原点を中心として対称である

の検定統計量としては

$$S=\sum_{i=1}^{n}a_n(R_i^+)\mathrm{sgn}(X_i)$$

が標準である．ここに，$a_n(0)=0\leq a_n(1)\leq\cdots\leq a_n(n)$ は与えられた定数である．ただし，分布の対称性が棄却された場合に，Sのみでは分布の対称性がどのように崩れているかがわからない．このとき，$|X_1|, \cdots, |X_n|$ の昇順に $X_1, \cdots, X_n$ を並べかえたものを $Y_1, \cdots, Y_n$ とおく．$Q_0=O$ とする．点 Q_k の横座標を $\sum_{i=1}^{k}a_n(i)$，縦座標をkとする$(k=1, \cdots, n)$．点 $O, Q_1, \cdots, Q_n$ をこの順に結び A_1 とおく．A_1 と縦軸に関し対称な線分図を A_2 とおく．A_1 と A_2 は対称性から最も遠い場合に対応する．次に，$P_0=O$ とし，点 P_k の横座標を $\sum_{i=1}^{k}a_n(i)\,\mathrm{sgn}(Y_i)$，縦座標を k にとり，$P_0, P_1, \cdots, P_n$ をこの順に結ぶ．この線分図をBとおく．この図から，分布の対称性が直

観的に判定でき，統計量SはP_nの横座標で表される．

k標本問題：$X_{i1}, \cdots, X_{in_i}(i=1, \cdots, k)$を分布関数$F_i(x)$からの確率標本とし，$N=\sum_{i=1}^{k} n_i$とおく．さらに，$R_{ij}$を全データにおける$X_{ij}$の順位とする．また，$i$番目に小さいデータが$F_i$から得られている場合には$Z_j=i$とおく．

帰無仮説をH：$F_1(x)=\cdots=F_k(x)$とし，対立仮説を

K_1：適当なi, j, xで$F_i(x)\neq F_j(x)$

とする．

$$Q_{ij}=\frac{\pi}{N+1}\left(R_{ij}-\frac{N+1}{2}\right)$$

とおく．F_iから得られたデータに対するグラフ$D_i(i=1, \cdots, k)$を次の手順で描く．

まず，$Q_{i1}, \cdots, Q_{in_i}$を昇順に並べかえ，$Q_{i(1)}, \cdots, Q_{i(n_i)}$とおく．点$P_{i0}=O, P_{i1}, \cdots, P_{in_i}$とその間を結ぶ線分が，$P_{i,j-1}$から傾き$Q_{i(j)}$，長さ1の線分を引き，その端点を$P_{ij}$とおくことによって作られる．この図と線分$OP_{in_i}$の中点に関して対称な線分を引く．この2つの線分図で多角形D_iを作る．

多角形$D_1, \cdots, D_k$はすべて凸多角形で，各辺の長さは1である．Hが正しい場合には各D_iはそれぞれ正$2n_i$多角形に近く，K_1のもとではそうではなくなる．したがって，視覚によりk個の標本が同一がどうかが判定できる．

この性質から，D_iの面積(多角形の面積は正多角形のとき最大)，頂点P_{in_j}の位置やOP_{in_j}の横軸となす角度などが検定統計量として考えられる．P_{in_j}の横座標をp_{ix}，縦座標をp_{iy}とする．また，D_iの面積をT_iとおく．これらの量は

$$p_{ix}=\sum_{j=1}^{n_i}\cos Q_{ij},\ p_{iy}=\sum_{j=1}^{n_i}\sin Q_{ij},$$

$$T_i=\sum_{j<l}\sin|Q_{ij}-Q_{il}|=\sum_{j<l}\sin\left|\frac{(R_{ij}-R_{il})\pi}{N+1}\right|$$

で与えられる．検定統計量として

$$S_1=\frac{2N}{\pi}-2\sum_{i=1}^{k} T_i/n_i,$$

$$S_2=\frac{2\pi^2}{\pi^2-8}\sum_{i=1}^{k}\frac{(p_{ix}-2n_i/\pi)^2}{n_i},\ S_3=2\sum_{i=1}^{k} p_{iy}^2/n_i,$$

$$S_4=S_2+S_3,\ S_5=\frac{8}{\pi^2}\sum_{i=1}^{k} n_i\{\tan^{-1}(p_{iy}/p_{ix})\}^2$$

などがある（Shirahata, 1990）．S_2とS_3は漸近的に互いに独立でともにHのもとでは自由度$k-1$のカイ2乗分布に従い，S_5はS_3と漸近的に同値となることが示されている．また，S_1は漸近的には自由度$2k-2$のカイ2乗分布に従う互いに独立な確率変数$Z_1, Z_2, \cdots$を用いて

$$S_1\approx\sum_{i=1}^{\infty}\frac{2}{(4i^2-1)\pi}Z_i$$

となることが示されている．白旗（1988）では，その漸近分布のパーセント点が計算されている．S_2 は尺度の違い，S_3 は位置母数の違いに敏感な検定であり，S_1 と S_4 はその両者を考えている．

4. 統計的グラフィクス・ソフトウェア

Lindley (1984) は英国統計学会の 150 周年を記念して寄稿した論文のなかで次の意味深な台詞をはいている：「統計学は基本的に寄生的であり，他の研究にのっかって生きていく．これは，統計学に対する蔑視ではない．すなわち，今日では多くの宿主は，かかえた寄生虫がいなくなると死亡する場合のあることが認められている．(寄生虫がいないと)食物を消化することのできない動物もいる．多くの分野の人々の努力によって，おそらくその宿主は死亡することはないであろうが，統計学なしでは確実に弱くなると考えられる．実験心理学はその一例である．生態学者はこの現象について二つの名前をつけている．すなわち，両方の種に好都合で義務的でない相互作用の原協力 (proto-operation) と両方の種に好都合で義務的な相互作用の相互扶助(mutualism)である．統計家は，統計学が後者の型に属することを望んでいる．しかしながら，統計学は前者であることが多い」．このことは統計学が顧客を創造することでしかその魅力を醸成できないようにも受けとれる．統計的グラフィクスについては，その研究・開発・適用・評価の過程がコンピュータ科学をはじめとする他分野の顧客との協力で進行することが多いだけに，このことはより正鵠を射ているように感じられる．このとき，賢い顧客との対応で留意しておきたいことは，同一水準での統計的グラフィクス・ソフトウェア（SGS）の開発で非生産的競争を演じないこと，便宜に走る「効率」主体の SGS の評価指標を「効果」におきかえること，使い込むほど味のある SGS，あるいは使い勝手のある（「使っていて技能水準・知識水準の向上する」）SGS の開発を心がけることであろう．他方，顧客は多いにこしたことはないが，ただ数が増えればよいのではない．快適さや満足度を追い求める風潮のなかで，わがままな「やぶユーザ」の出現には勇気をもって対峙することが必要であろう．

動的グラフィクスの起爆剤となったのは，スタンフォード大学線形加速器センターの PRIM 9 システムのプロジェクト・チームと AT & T ベル研究所のグループによるいくつかの研究であろう（後藤 他，1988）．これらの研究で提示された動的グラフィクスの考え方，および実践システムに組み込まれた動的グラフィクスの基本操作は，その後のパソコン版あるいはワークステーション版の動的グラフィクス・ソフトウェア（あるいはシステム）の基礎をなしている．そのようなソフトウェアとして，最近では MacSpin (Donoho et al., 1985)，JMP (Held et al., 1989)，SAS/INSIGHT (SAS Institute, 1990)，S (Becker & Chambers, 1984) などが著名である．これらに代表される著名なソフトウェアに組み込まれているたいていの動的グラフィクスは散布図と 3 次元回転を基調とし，しかもそれらはデータ解析過程の多様な利用形態に適合するように普遍化されている．

しかしながら，このようなソフトウェアの動的機能を操作するためには，利用者とシステムとの間で対話のできるシステム環境が必要とされる．現在の動的グラフィクスでは GUI (Graphical User Interface) が採用されており，利用者はマウスを使うことによって，アイコンやメニューを容易に選択・操作できる．パソコン用の動的グラフィクスに注目すると，多くのソフトウェアは GUI の完成（感性）度の高さから Apple Computer 社の Macintosh 上で開発されているようである．例えば，MacSpin と JMP の他に Data Desk, StatViewII, SuperANOVA, Exstatix, Fastat などのソフトウェアが提供されており，基本的な動的機能と Macintosh 独自の統一された操作法が用意されている (Best & Morganstein, 1991)．一方，ワークステーションでは，Plot Windows (Stuetzle, 1987) や ORION I (McDonald, 1988)

などの動的グラフィクスがSやSAS/INSIGHTとは異なる独自のGUI環境で公表されている．

最近の動的グラフィクスに関する斬新で先進的な研究の殆どはClevelandやBeckerなどに代表されるAT & Tベル研究所の一派，およびワシントン大学とカルフォルニア大学に属するBujaやAsimovらの一派の主導のもとに進められているといっても過言でない．Buja et al. (1988) では彼らの開発した，データを視覚化するためのコンピュータ・ソフトウェアの基本的構成要素とその機能が種々の例を用いて紹介されている．彼らの開発したソフトウェア自体は「データ・ビュワー」システムと呼ばれている（Hurley, 1988）．彼らの動的グラフィクスに対する思考の斬新さは射影と回転に集約されている．一般に，動的グラフィクスにおける図形の回転は3次元上で行われる．しかし，彼らは3次元またはより高次元のデータの2次元平面上への動的射影を行っている．とくに，データ解析の目標に応じて使い分けのできる3種の動的射影として，プロット補間，相関ツアー，グランド・ツアーを与えている．

いま，4次元データを $\{x^{(1)}, x^{(2)}, y^{(1)}, y^{(2)}\}$ で表す．このとき，プロット補間では

$$(x^{(1)}\cos t + x^{(2)}\sin t) \text{ 対 } (y^{(1)}\cos t + y^{(2)}\sin t)$$

がプロットされる．このプロットでは，t を0度から90度に変化させることによって，ディスプレイ画面上に生起するプロットが $(x^{(1)}, y^{(1)})$ プロットから $(x^{(2)}, y^{(2)})$ プロットに移動する．このプロットは実質的に4次元データの回転であり，$y^{(1)}=y^{(2)}$ のとき，3次元データの回転に対応する．

相関ツアーとは，x 変数の集合と y 変数の集合が与えられているとき，y 変数の線形結合を x 変数の線形結合に対してプロットする動的グラフィクスである．このプロット法は，線形結合が最適化こそされていないが，正準相関解析の考え方に基づいている．

グランド・ツアーでのプロット法は x 変数と y 変数を区別しない他は相関ツアーと同じであり，多次元空間における任意の射影走査を与える一種の射影追跡と見做せる．

5. 結びに代えて

統計的グラフィクスの目的は，観察者あるいは利用者が，データから便利な，あるいは説明できるようなかたちで結論を提示することでなく，データに関する有意な仮説あるいは知見をひき出しやすくすることである．そのため，統計的グラフィクスを「見る」ツールとして眺めるのでなく，「見ぬく」ツールとして用意するか，あるいは活用することが重要である．最近の動的グラフィクスの展開はそこに力点をおいているようにみえる．また，統計的グラフィクスの研究・開発では，その実際的活用が必須であるが，「特殊」を普遍化するよりもその「特殊」性により磨きをかけることが重要であろう．その意味では多彩な事例研究がより活発に奨励されてもよいと思われる．

最近のコンピュータとその周辺装置の急激な発展につれて，統計的データ解析も変容を迫られ，さらにその環境も逐次に更新されていく．そして，統計的データ解析も「統計的」の形容詞と無関係に統計的知識に疎い人たちによって行われ，統計的グラフィクスをはじめとする統計的ソフトウェアが誤用・乱用を問わず頻繁に使われていく．このとき，これらの人たちを含む顧客とうまくつき合うには相互の信頼関係にたつ共同参画型接近法を徹底して実践していくしか道はないように感じられる．我々が顧客のなかから「友」を選べるとすれば，研究・開発・実践の過程を長期にわたって一緒に歩み，「ものいわずとも」自然にわかってくれる人たちであろう．

謝辞：草稿の段階で吉村　功先生（名古屋大学）に貴重な助言や示唆を，また査読者の方々か

らも適切な助言を頂戴した．それらのご厚意に対して，ここに深く感謝の意を表します．

参 考 文 献

[1] Anderson, T. W. & Bahadur, R. R. (1962) . Classification into two multivariate normal distributions with different covariance matrices, *Ann. Math. Statist.*,**33**, 420-431.

[2] Asimov, D. (1985). The grand tour : A tool for viewing multidimensional data, SIAM J. *Sci. Statist. Comput.*, **6**, 128-143.

[3] Bacon-Shone, J. & Fung, W. K. (1987). A new graphical method for detecting single and multiple outliers in univariate and multivariate data, *Appl. Statist.*, **36**, 153-162.

[4] Barnett, V. (1976). The ordering of multivariate data, *J. Roy. Statist. Soc.*, **A139** (**3**), 318-355.

[5] Becker, R. A. & Chambers, J. M. (1984). S : *An Interactive Environment for Data Analysis and Graphics*. Wadsworth [渋谷政昭・柴田里程 共訳 (1987)．Sシステム I ：概要編．共立出版，渋谷政昭・柴田里程 共訳 (1987)．Sシステム II ：詳細編．共立出版].

[6] Benjamini, Y. (1988). Opening the box of a box-plot, *Amer. Statist.*, **42**, 257-262.

[7] Best, A. M. & Morganstein, D. (1991). Statistics programs designed for the Macintosh : Data Desk, Exstatix, Fastat, JMP, StatView II, and SuperANOVA, *Amer. Statist.*, **45** (**4**), 318-338.

[8] Board on Mathematical Sciences (BMS), Commission on Physical Sciences, Mathematics, and Resources, National Research Council, Washington, D. C. (1989). Statistical Sciences : Some research trends-statistics, *The IMS Bulletin*,**18** (**2**) 181-193.

[9] Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations, *J. Roy. Statist.Soc.*, **B26**, 211-252.

[10] Buja, A., Asimov,D., Hurley, C. & McDonald, J. A. (1988). Elements of a viewing pipeline for data analysis, *Dynamic Graphics for Statistics*, 277-308, ed. by Cleveland, W. S. & McGill,M.E.,Wadsworth.

[11] Chang,W. C. (1987).A graph for two training samples in a discriminant analysis, *Appl. Statist.*, **36** (**1**), 82-91.

[12] Chernoff,H. (1978). Graphical representation as a discipline. *Graphical Representation of Multivariate Data*, ed. by Wang, P. C., Academic Press, 1-11.

[13] Cleveland, W. S. (1985) . *The Elements of Graphing Data*. Wadsworth.

[14] Cox, D. R. (1972). Regression models and life tables, *J. Roy. Statist. Soc.*, **B34**, 187-202.

[15] Critchley, F. & Ford, I. (1985). Interval estimation in discrimination : The multivariate normal equal covariance case, *Biometrika*, **72**, 109-116.

[16] Donoho, A. W., Donoho, D. L. & Gasko, M. (1985). *MacSpin Graphical Data Analysis Software. D^2* Software.

[17] Eddy, W. F. (1981). Comment on the paper of Friedman & Rafsky (1981), *J. Amer. Statist. Assoc.*, **76**, 287-289.

[18] Eddy, W. F. (1982). Convex hull peeling, *COMPSTAT 1982-Part I : Proceedings in Computational Statistics*, 42-47, Physica-Verlag, Vienna.

[19] Eddy, W. F. (1990). Journal of Computational and Graphical Statistics to Begin Publication in 1992. *AMSTAT News*, **169**, November 1990.

[20] Everitt, B. S. (1987). Graphical displays of complex data : Scientific tools or simple art for art's sake. Invited paper at "Statistical Graphics" meeting of the 46th Session of the International Statistical Institute at Tokyo.

[21] Fisher, N. I. (1987). Graphical methods in statistics : Current and prospective views, Invited paper at "Statistical Graphics" meeting of the 46th Session of the International Statistical Institute at Tokyo.

[22] Friedman, J. H. & Rafsky, L. C. (1981). Graphics for the multivariate two-sample problem, *J. Amer. Statist. Assoc.*, **76** (**374**), 277-295.

[23] Frigge, M., Hoaglin, D. C. & Iglewicz, B. (1989). Some implementations of the boxplot, *Amer. Statist.*, **43**, 50-54.

[24] Gnanadesikan, R. (1981). Statistical graphics : capabilities, criteria, and calibration. Invited paper of the 43rd Session of the ISI at Buenos Aires, Argentina, I. P. 16. 2, 1-13.

[25] Goto, M. (1981). Some devices in graphical data analysis. Invited paper at the 43rd Session, of the ISI at Buenos Aires, Argentina, I. P. 16. 3, 1-16.

[26] Goto, M., Matsubara, Y. & Tsuchiya, Y. (1983). Power-normal distribution and its applications, *Res. Stat. Appl. Res., JUSE.*, **30** (3), 8-28.
[27] Goto, M. (1987). Statistical graphics: Discussion. *Proceedings of the 46th Session, of the ISI, at* Tokyo, 401-402.
[28] Goto, M., Matsubara, Y., Yoden, A., Tsuchiya, Y. & Wakimoto, K. (1991). Statistical graphics: A classified and selected bibliography, *J. Japan Statist. Soc.*, **21** (1), 37-61.
[29] Haaland, P. D. (1990). Historical sketch of the ASA section on statistical graphics, *Amer. Statist.*, **44** (2), 96-98.
[30] Held, G., Lehman, A. & Sall, J. (1989). Advance in graphical data analysis from SAS Institute. *Statistical Software Newsletter*, **15** (3), 85-90.
[31] Hoaglin, D. C. & Iglewicz, B. (1987). Use of boxplots for process evaluation, *J. Quality Technology*, **19**, 180-190.
[32] Hurley, C. (1988). A demonstration of the data viewer. Computer Science and Statistics, *Proceedings of the 20th Symposium on the Interface*, 108-114.
[33] Lindley, D. V. (1984). Prospect for the future: The next 50 years, *J. Roy. Statist. Soc.*, **A147**, 359-387.
[34] Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy, *Biometrics*, **18**, 455-477.
[35] McDonald, J. A. (1988). ORION I: Interactive graphics for data analysis. *Dynamic Graphics for Statistics*, 179-199, ed. by Cleveland, W. S. & McGill, M. E., Wadsworth.
[36] McGill, R., Tukey, J. W. & Larsen, W. A. (1978). Variations of box plots, *Amer. Statist.*, **32**,12-16.
[37] Mizuta, M. & Kawaguchi, M. (1983). A graphical dynamic representation of multidimensional data, *J. Japan Statist. Soc.*, **13** (1), 1-10.
[38] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics*, **14**, 945-966.
[39] Prentice, R. L. (1974). A log gamma model and its maximum likelihood estimation, *Biometrika*, 61, 539-544.
[40] Prentice, R. L. (1976). A generalization of the probit and logit methods for dose response curves, *Biometrics*, **32**, 761-768.
[41] SAS Institute, Inc. (1990). *SAS/INSIGHT User's Guide*, version 6 edition. SAS institute Inc.
[42] Shirahata, S., Wakimoto, K. & Tarumi, T. (1985). A goodness of fit test based on the linked line chart, Aust. *J. Statist.*, **27**, 163-171.
[43] Shirahata, S. (1990). Rank tests for the k-sample problem based on a linked line chart, *J. Japan Statist. Soc.*, **20**, 169-178.
[44] Stacy, E. W. (1962). A generalization of the gamma distribusion, *Ann. Math. Statist.*, **33**, 1187-1192.
[45] Stuetzle, W. (1987). Plot windows, *J. Amer. Statist. Assoc.*, **82**, 466-475.
[46] Taguri, M., Hiramatsu, M., Kittaka, T. & Wakimoto, K. (1976). Graphical representation of correlation analysis of ordered data by linked vector pattern, *J. Japan Statist. Soc.*, **6** (2), 17-25.
[47] Tukey, J. W. (1970). *Exploratory Data Analysis* (limited preliminary ed.). Addison-Wesley.
[48] Tukey, J. W. (1977), *Exploratory Data Analysis*. Addison-Wesley.
[49] Wakimoto, K. & Taguri, M. (1974). On the representation method of multiple correlation by pattern of connected vectors, *J. Japan Statist. Soc.*, **5** (1), 9-24 (in Japanese).
[50] Wakimoto, K. (1977). Tree graph method for visual representation of multidimensional data, *J. Japan Statist. Soc.*, **7**, 27-34 (in Japanese).
[51] Wakimoto, K. & Taguri,M. (1978). Constellation graphical methods for representation of multidimensional data, *Ann. Inst. Statist. Math.*, A**30**, 97-104.
[52] Wakimoto, K. (1981). K-multiple chart and its application to the test for homogeneity against ordered alternatives, *J. Japan Statist. Soc.*, **11** (1), 1-7.
[53] Wakimoto, K. & Shirahata, S. (1984). A coefficient of concordance based on the chart of linked lines, *J. Japan Statist. Soc.*, **14** (2), 189-197.
[54] Wakimoto, K. & Goto, M. (1987). Graphical representation and practical data analysis (Guest editorial), *Computational Statistics & Data Analysis*, **5** (2), 83-84.
[55] Wilk, M. B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, **55**, 1-17.
[56] Wilks, S. S. (1962). *Mathematical Statistics*. John Wiley & Sons.

[57] 後藤昌司 (1973). 多変量データの解析法. 科学情報社.
[58] 後藤昌司・松原義弘・脇本和昌 (1988). グラフィカル接近法の最近の発展. 行動計量学, **15** (**2**), 45-70.
[59] 白旗慎吾 (1988). カイ二乗確率変数の重み付き和の分布関数の計算. 計算機統計学, **1**, 37-44.
[60] 田崎武信・後藤昌司 (1984). データ解析における確率プロット法の用途. 行動計量学, **12** (**1**), 29-34.
[61] 馬場康維 (1991). 統計的グラフィックスの主観と客観. 日本統計学会 60 周年記念パネル討論「統計グラフィックス」予稿集, 28-31.
[62] 松原義弘・後藤昌司 (1984). 回帰分析におけるグラフィカル接近法. 行動計量学, **12** (**1**), 20-28.
[63] 松原義弘・後藤昌司 (1988). 生存時間解析における統計的モデルとその適用上の留意点. 癌生時研誌, 8, 93-97.
[64] 松原義弘 (1989). 正確なもりだくさん. 教育と情報, **381** (**12**), 44-45.
[65] 松原義弘・後藤昌司 (1989). 生存時間解析におけるグラフィカル表現. 応用統計学, **18**, 85-97.
[66] 余田明夫・松原義弘・後藤昌司 (1990). 統計的グラフィクスにおける卑近な方法：最近の発展と適用上の留意点. SHI-Preliminary Research NO. 187, 塩野義解析センター[松原義弘・余田明夫・後藤昌司(1989). 統計的グラフィクスにおける卑近な方法：最近の発展と適用上の留意点. 第 17 回日本行動計量学会大会発表論文抄録集, 119-122].
[67] 脇本和昌・後藤昌司・田栗正章・松原義弘 (1977). 多次元データのグラフ解析法. 応用統計学, **6** (2-3), 43-82.
[68] 脇本和昌・後藤昌司・松原義弘 (1979). 多変量グラフ解析法. 朝倉書店.