統計数理　第36巻　第2号（1988）

公開講演会要旨

グラフによる統計分析

統計数理研究所　馬　場　康　維

（昭和63年11月4日，統計数理研究所 講堂）

1. はじめに

統計分析を行なう際，グラフは有用である．情報の整理や表現の手段として，あるいは思考や推論の補助として，分析の様々な局面でグラフが用いられている．近年，パーソナル・コンピュータの普及によって比較的簡単にグラフが描けるようになったことが一つの誘引ともなり，統計分析のための様々なグラフが考案されている．

数量的な表現に比して，グラフ表現は多くの情報を同時に表現でき分析者の直感に訴えるという特徴を持っており，これが最大の利点である．しかし，同時に，グラフを読む際に誤解する危険もはらんでいる．グラフの持つこの特徴は，したがって，逆に欠点になることもある．このような観点から，グラフ表現は多くの情報を含み，直感に訴えるものを持ち，客観的な情報を与えるものが望ましいことになる．

様々なグラフの分類が考えられるが，こういった観点から統計分析に用いられるグラフを分類するならば

（1）　表現力
（2）　描き方の一意性
（3）　数値情報との対応
（4）　統計的な概念との結びつき
（5）　推測的か記述的か

などの項目が分類の尺度になるであろう．

ここでは，様々なグラフの中から顔形グラフ，順位グラフ，推測のためのグラフを紹介する．上記の分類からすると，顔形グラフは表現力はあるが描き方に一意性のないグラフであり，順位グラフは表現力と一意性を持ったグラフである．

2. グラフ表現の例

【顔形グラフ】

顔の形，目の大きさ，口の大きさなど顔の特徴を表わすものに各変数を割り当てることによって多次元データを表現する方法の一つに，チャーノフ（Charnoff (1973)）の顔形グラフがある．図1には，表1のデータをもとにして描いた顔形グラフを示した．顔の各部と変数は以下の様な関係を持たせてある．

（1）　額の広さ：　　広いほど1住宅当たりの敷地面積が広い
（2）　あごの大きさ：　大きいほど貯蓄高が高い
（3）　口の大きさ：　　大きいほど消費支出が多い

表1.　12都道府県のデータ

顔のパラメータ / 区分 / 都道府県	顔上半分の離心率	顔下半分の離心率	口の幅	目の幅	眉の傾き
	1住宅当たり敷地面積[1] (m^2)	1世帯当たり貯蓄現在高[2] (千円)	消費支出[3] (千円)	住宅敷地価格[4] (千円/3.3 m^2)	男性の実労働時間[5] (時間)
北海道	264	5082	273.6	118.1	208
山形	382	4687	293.5	117.5	203
茨城	440	6821	281.7	164.3	205
東京	153	8206	309.2	895.9	191
石川	278	7278	275.0	213.7	200
愛知	245	8456	287.5	346.4	201
大阪	126	7190	267.6	551.3	197
鳥取	293	6913	282.4	155.6	204
広島	211	6597	291.3	288.5	201
高知	179	5461	231.1	279.5	201
福岡	251	5322	265.5	201.3	198
鹿児島	284	4025	242.8	145.8	204

総務庁統計局：昭和62年11月発行の社会生活統計指標より．
1）昭和58年，2）昭和59年，3）昭和60年非農家1世帯当たり月平均，4）昭和60年，
5）昭和60年月平均実労働時間．

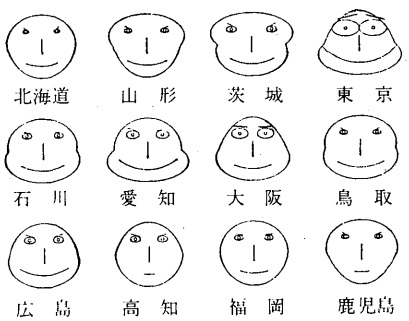


図1.　12都道府県の顔形グラフ（馬場（1988）参照）．

(4)　目の大きさ：　　大きいほど住宅敷地価格が高い
(5)　眉の傾き：　　上がるほど実労働時間が多い

各都道府県の特徴は一目瞭然であろう．

このグラフは変数と顔のパラメータの割り当てが適切であれば非常に分かりやすいグラフになるが，変数の割り当て方によっては全く違った印象のグラフになる．また，変数の割り当て方に任意性のあるのが欠点である．

【順位グラフ】

食べ物の味，絵画の良さなど計器による測定の困難なものを感覚によって評価しようという場合に，順位がしばしば用いられる．ここでは，順位データのグラフ表現法の一つである順位グラフを紹介する．

表2．6つの街の順位づけ

判定者＼アイテム	新宿	渋谷	原宿	池袋	有楽町	上野	性別
1	4	1	2	5	3	6	2
2	5	1	3	4	2	6	2
3	2	3	1	5	6	4	1
4	2	5	6	1	3	4	1
5	6	1	4	5	2	3	1
6	2	1	6	4	3	5	1
7	2	1	3	5	4	6	1
8	3	4	2	6	1	5	1
9	1	3	2	4	5	6	1
10	3	4	5	6	2	1	1
11	2	1	6	5	3	4	1
12	6	1	5	3	2	4	2
13	3	5	6	4	1	2	2
14	4	2	3	6	1	5	1
15	4	3	2	5	1	6	2
16	1	3	2	6	4	5	2
17	3	2	1	5	4	6	1
18	1	2	3	4	5	6	1
19	5	2	4	6	3	1	2
20	4	3	1	6	2	5	2

数値は順位を示す．性別欄の数値は1：男，2：女である．

次の質問は1987年に統計数理研究所で行なわれた公開講座の出席者に対するアンケートの一部である．

> 次の6つの街に，好きな順に順番をつけて下さい
> 渋谷，新宿，原宿，有楽町，池袋，上野

回答者の中からランダムに抽出した20人のデータを表2に示した．この表をもとに男・女別に描いた順位グラフを図2に示した．

このグラフは次の様な意味を持つ．円周上の目盛は順位に対応する方向を表わしている．各街に対応する直線はアイテムベクトルと呼ぶが，方向が平均順位を表わし，長さが評価の一致度を表わしている．このグラフから，男・女によって評価の異なる街と異ならない街とがあることが読み取れる．

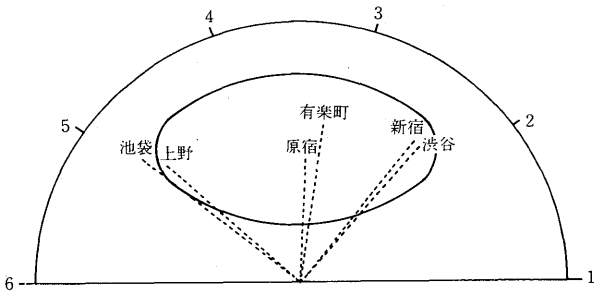
アイテムベクトルは以下の様にして描く．一般的な形で示すとして，順位をつけられるもの（これをアイテムという）の数を k，順位をつける人（以下，判定者という）の数を n とする．f_{ij} をアイテム i につけられた順位 j の頻度とし，

$$p_{ij}=f_{ij}/n$$

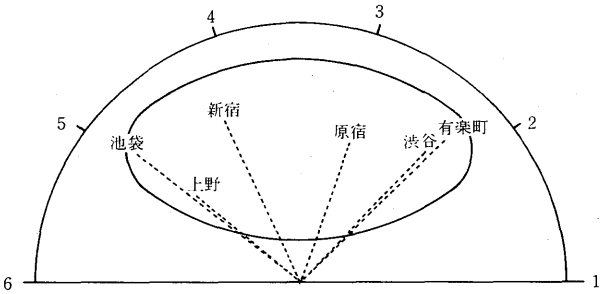
とする．

$$\theta_j=\pi(j-1)/(k-1)$$

を順位 j に対応する角度とし，ベクトル



(a)



(b)

図2. 6つの街の順位づけ. (a) 男: 12人, (b) 女: 8人

$$x_i=\left(\sum_{j=1}^{k} p_{ij}\cos\theta_j,\ \sum_{j=1}^{k} p_{ij}\sin\theta_j\right)$$

を作る. これが, アイテムベクトルである.

ここで, どのアイテムにも区別するほどの差がないと仮定しよう. すると, 一つのアイテムに与えられる順位はランダムに順位をつけたものになるであろう. 一つのアイテムに対してランダムに順位を割り当てた場合に得られるアイテムベクトルの終点の分布は二次元の正規分布で近似できることが知られている (Baba (1986)). 図2にはアイテムベクトルがこの外側に位置する確率が0.05になる楕円が示してある. 一つの街につけられた順位はランダムであるという仮説を検定するとすれば, この楕円は有意水準5%の棄却楕円ということになる.

【グラフを用いた推測】

グラフを用いて, 質的な変数からある種の数量を推測するという方法について述べよう. 一つの例として, 性別, 年齢, 塩からいものを食べるかどうか, 油っこいものを食べるかどうかという質問に対する答から血圧を推測するという場合を考える.

18人の人の答のパターン(以下, 反応パターン)と血圧を調べた結果が表3に示してある. 各質問(アイテム)に対する答(カテゴリー)に図3のような線分を対応させる(詳細は馬場(1988)を参照のこと). 図4の折れ線は各個体の反応パターンに対応する線分をつないだものである. たとえば, *a, a, a, a, a* という記号のついた折れ線は (塩からいものを食べる, 60-70代, 男, 労働: らく, 油っこいものを食べる) という反応パターンを表わしている. 円周上の目盛は血圧を表わす.

原点と折れ線の終点を結んで円周まで延ばした点が血圧の推測値となる. 上記のパターンか

表3．アイテム・カテゴリーへの反応パターンおよび目的変量の推測値と実測値

アイテム	性		年齢			油っこいもの		塩からいもの		労働		最大血圧値（mmHg）		
カテゴリー / 個体	男	女	20〜39	40〜59	60〜79	食べない	食べる	食べない	食べる	らく	きつい	実測値	推*測値	推**測値
1	∨		∨				∨	∨			∨	124	122.8	122.7
2		∨		∨			∨		∨		∨	154	142.9	146.1
3		∨		∨		∨		∨		∨		130	129.0	128.8
4	∨			∨			∨		∨		∨	168	152.8	154.2
5		∨	∨			∨		∨		∨		134	124.9	125.0
6	∨				∨	∨			∨	∨		180	181.0	180.9
7	∨		∨				∨	∨			∨	114	122.8	122.7
8		∨			∨		∨		∨	∨		172	175.1	174.7
9		∨		∨			∨	∨		∨		120	131.0	130.4
10	∨				∨	∨			∨	∨		176	181.0	180.9
11		∨		∨		∨			∨		∨	108	140.7	144.3
12	∨				∨	∨		∨		∨		162	156.6	153.0
13		∨	∨			∨		∨			∨	123	113.1	112.9
14	∨				∨		∨		∨	∨		185	182.7	182.7
15	∨		∨			∨		∨			∨	110	120.9	120.9
16		∨		∨		∨			∨		∨	168	140.7	144.3
17		∨			∨		∨		∨	∨		172	175.1	174.7
18	∨		∨				∨	∨			∨	122	122.8	122.7

*：グラフを用いた推測値，**：数量化Ⅰ類による推測値（馬場，脇本（1983）参照）．

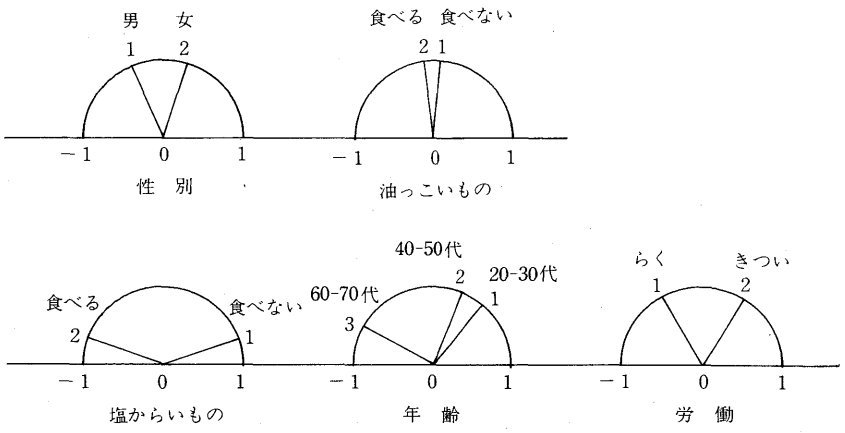


図3．アイテム・カテゴリーに与えられる角度．アイテムはレンジの大きい順に左下から右上に並べてある（馬場，脇本（1983）参照）．

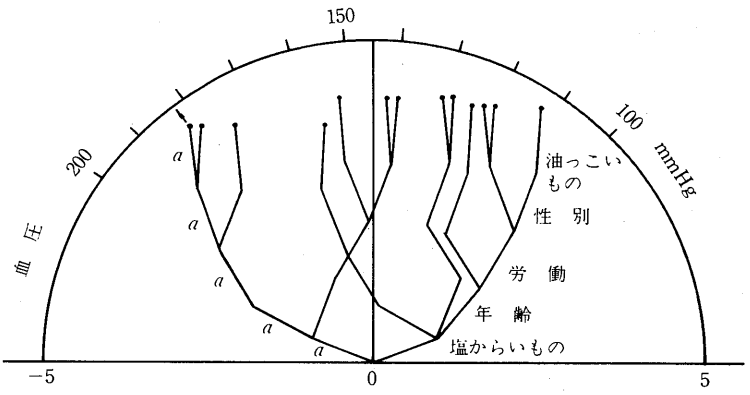


図4. 数量化の結果のグラフ表現．円周上の目盛は最大血圧値を表わす．矢印は経路 a で表わされる反応パターンから予測される最大血圧の値（馬場，脇本（1983）参照）．

ら推測される血圧は182である．推測という観点から見ると，この方法は林の数量化第I類とほぼ同じ結果を与えるが推測に用いた個体の反応パターンも読み取れるという利点を持っている．数値的な比較のためにこの方法と数量化理論による結果を表3に併記しておいた．

3. おわりに

近年，パーソナル・コンピュータの発達と普及にはめざましいものがあり，それとともに使いやすい各種のソフトフェアや周辺機器が開発され，コンピュータ・グラフィクスも利用されやすくなって来ている．また，大型計算機を用いた計算をするにしても以前はカードによるバッチ処理が主体であったのが，今や，スクリーンエディタを用いたTSS処理がほとんどである．

こういったコンピュータ環境が誘引となって，統計分析の方法は様変わりして来ているように見える．コンピュータによる統計分析というと大型計算機のバッチ処理が主体であった時代には一つの手法を試すにしてもかなりのターンアラウンドタイムを要した．したがって，多くの場合，データやプログラムのほんの僅かな修正にもかなりの時間を割かねばならず，これは試行錯誤を繰り返すことをためらわすには十分であった．しかし，最近のコンピュータ環境は分析における試行錯誤を容易にしている．一つの手法，一つのモデルにこだわるよりもデータの構造を様々な角度から分析することが可能になって来ている．

こういう状況の下では，グラフの持つ役割は重要である．コンピュータとの会話の窓口として，データの構造を見るための手段として，あるいは分析者の成果を他者に伝えるための言語としてますます需要が増大していくに違いない．

参考文献

Baba, Y. (1986). Graphical analysis of rank data, *Behaviormetrika,* **19**, 1-15.

馬場康維（1988）．グラフ解析（第3章），『パソコンによるデータ解析（村上，田村編）』，朝倉書店．

馬場康維，脇本和昌（1983）．ベクトル変換を用いた数量化法，統計数理研究所彙報，第30巻2号，67-75.

Charnoff, H. (1973). The use of faces to represent points in k-dimensional space graphically, *JASA,* **68**, 361-368.