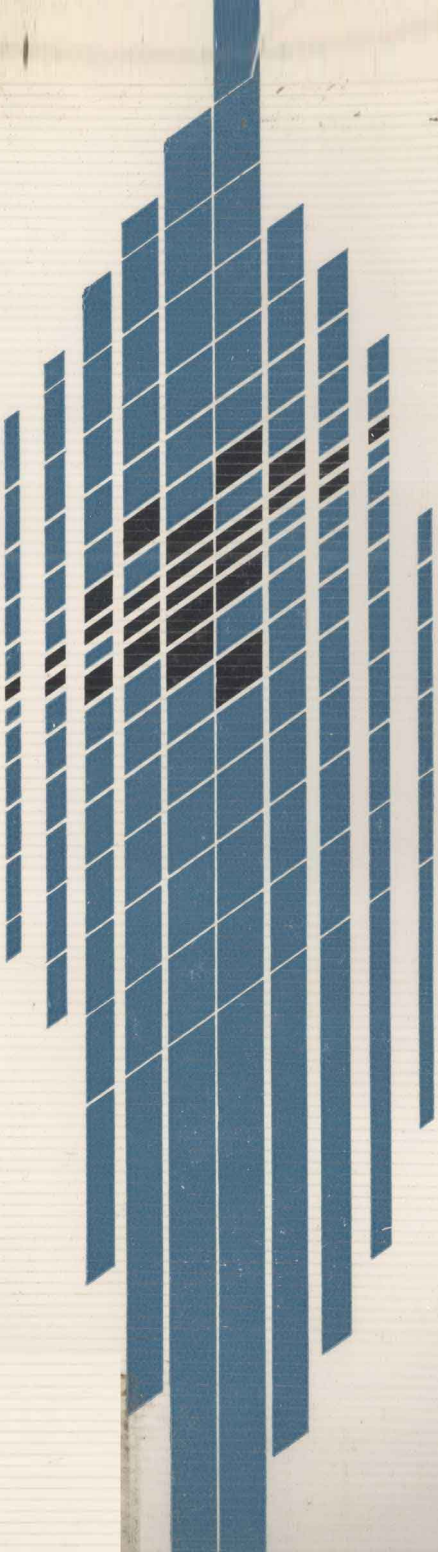

現代人の統計

多元的データ分析の基礎

林 知己夫 編

7

駒澤 勉 著

朝倉書

多元的データ分析の基礎

駒澤　勉

現代人の統計
7

朝倉書店

著者略歴

1935年　東京に生れる
1959年　早稲田大学第一理工学部卒業
現　在　文部省統計数理研究所研究室長
　　医学博士

現代人の統計 7
多元的データ分析の基礎　　定価 2000 円

1978年2月10日　初版第1刷
1979年6月20日　第2刷(訂正版)

著　者　駒澤勉
発行者　朝倉鏻造
発行所　株式会社 朝倉書店
東京都新宿区新小川町2-10
郵便番号　162
電話 03(260) 0141 (代)
振替口座 東京 6-8673 番

〈検印省略〉

政弘印刷・渡辺製本

3341-111807-0032

はじめに

われわれが統計的データ解析によって現象を把握，解明するためには，その現象を予測や分類問題に置き直して分析することである．一般に現象を記述しているデータは多元的なものであり，多次元空間上にその現象のデータ構造が形づくられている．その構造は個々の個体(人やもの)が持っている固有の特性(性，年齢，学歴，職業，…)の要素(男，42歳，大学卒，公務員，…)とか，また(血清総コレステロール値，最大血圧値，最小血圧値，心電図，…)の要素(190 mg/dl，146 mmHg，82 mmHg，正常，…)などから構成され，各特性要素は，個体または特性に複雑に交錯し，反応し合い，社会現象や医学現象を表現している．このような個体と特性間の多元関係がおりなす現象の構造を解明してゆくには，われわれが容易に判断できる空間，すなわち，数量の大小関係や平面的な位置関係で予測や分類が判断できるせいぜい一次元ないし，二次元の最も理解しやすい空間を見つけ出して分析することである．もし，単独の特性データの平均値や分布の型で取り扱っている現象が把握できる場合にはそれだけで解釈をすればよいわけである．どうしても多次元的な関連で分析を進めなければならないとき，現象から得た調査データや計測データをもとにそれらの多次元の情報をほとんど損失することなく最小次元にデータ構造を移して分析する多変量解析法，質的データの数量化理論が必要である．

本書は現象を多元的データにより分析するための基礎について，高校卒の数学知識があれば文科系・理科系を問わず理解できる．文部省統計数理研究所付属統計職員養成所の公開講座で担当した‘数量化理論’や‘線型計算法’のテキスト，他官公庁の研修機関や大学で，この種の講義をした内容や実際にデータ解析してきた長年の経験をもとに，多次元データ解析の実践での中心部の考え

方とその適用の仕方を記述した．従来のこの種の書と多少異なる構成法で話を進めている．意識的に「林の数量化理論」を念頭に置いて，量的データと質的データの場合を解析上で呼応させながら，また，多次元解析で重要な線型計算の話も数学から多少遠のいた読者のために載せた．1章はわれわれ専門の者がなにげなく用いている数学や統計記号の話（これら記号を見ただけで拒絶反応を起さないよう）とデータの幾何学的な読み取り方など分析法に入る前のこころえである．2章から4章が分析法の中心で，2章は予測のための分析，数量化第Ⅰ類，3章は判別・分類のための分析，数量化第Ⅱ類，4章は成分分析，数量化第Ⅲ類から成っている．5章は線型計算の基礎を見なおしておこうという読者のために，6章では 2～4 章までの幾つか実際にすぐに役立つコンピュータ・プログラムを，さらに，付として実例にもとづいた分析の進め方を記述してある．

最近，行動計量科学ということばを多く耳にする．私の恩師である林知己夫先生のことばをかりれば――行動する現象をデータにもとづいて解析することは，理論と実際とが一体となってはじめて有効なものとなりこれには体験や統計的方法は言うまでもなく重要なものであるがこれら以外の知識，洞察力，知恵，勘のよさといった，いわゆる分析者のセンスをも必要とするのであって，きわめて人間的なところが重要である――といわれている．このことばは理論と実際を常に一体と考え学問研究をしていられる先生ならではである．われわれ，それほど体験や知識を集積していないものにとってはデータ解析の核心的理論の基礎をしっかりと身につけその上で適用分野での数多くの実データにもとづく解析の繰り返し試行経験を重ねるとともに，多くの適用上の試練を受け普遍化しなければならない．この点で，多次元データ解析を行なおうとする読者の糧となれば幸いである．終りに，日ごろ御懇篤なる御指導，またこの度は本書の御校閲を賜った林知己夫先生に深謝し，あわせてデータ計算処理に御援助下さった研究室の平野秀子嬢，さらに朝倉書店の編集部の方々には大変御面倒をおかけし，ともに深く謝意を表します．

1978年1月

駒澤　勉

目　　次

1 序　説

1.0 はじめに

われわれ，特に行動計量科学分野の者にとって，現象を把握し，解明するために，現象に関する各種の調査データや測定データをもとに分析する統計数理的なデータ解析法はなくてはならぬものである．また，今日の情報科学の時代にとって実際の現象解明問題に利用できる一般的な予測や分類の方法論とそれを適用してデータ解析するためのコンピュータは切り離して考えることはできない．ところで，われわれが扱おうとしている現象は一般に多次元空間上にその現象の構造を形造っている．その構造は，個々の個体が持っている固有の特性要素から構成され，各種特性要素(多変量，または多元的パターン)は，個体間(または特性間)に複雑に交錯し，反応し合い，お互い多元的な関係で結び合っている．このような個体と特性間の多元的関係がおりなす現象の構造を解明してゆくにはわれわれが容易に判断できる空間，すなわち，一次元ないし，せいぜい二次元空間のわかりやすい空間に移して分析する必要がある．そのようなデータ解析の方法論に多変量解析法，ならびに数量化理論がある．

これらの解析法の分類にはいくつか考えられるが，ここでは一つの分類の考え方として，外的基準(目的変量)が‘ある’，‘ない’に大きく分けて，基本的解析法とその処理計算手順を記述する．外的基準が‘ある’とか，‘ない’というのは，現象を知るために調査または測定データ中に予測や分類する目的の情報が‘ある’か，‘ない’かということである．たとえば，m 個の測定結果 $(X_1, X_2, \cdots, X_m)$，X_1 は年齢，X_2 は心電図所見，…，X_m は最大血圧値，を得たとき，血清総コレステロール値を他の特性変量(年齢，心電図所見，…，最大血圧値)で予測したいとき測定結果中に血清総コレステロール値があれば数量

の外的基準が‘ある’という．また疾患鑑別をしたいとき測定結果中に目的に応じた疾患所見があれば質の外的基準が‘ある’という．すなわち，現象解明の目的—基準—，実測値の血清総コレステロール値，疾患所見等を外的基準と呼ぶのである．

外的基準の‘ある’場合の分析法は問題を客観的に解析処理し，分析効果も科学的明確に示される．一方，外的基準の‘ない’場合は現象が形造っている構造空間のうち，最も情報量を有する空間を客観的な解析法で見つけ出すが，その解釈づけに当っては基準のないため主観的になり，適切な現象分析に至るまで試行錯誤の解析・分析が必要である．

1.1　基本的分析法

多次元の基本的解析法とその考え方について少し記述しておこう．表1で見るように，主に分析は外的基準の‘ある’ときその外的基準が‘数量’であれば予測問題を取り扱い，‘定性的なもの’であれば判別・分類問題を取り扱う客観的な分析法である．また，外的基準の‘ない’ときは，解析結果を用いて主観的に分類する問題を取り扱う分析法である．

表1　基本的な多次元データ分析法の分類表

外的基準がある場合（客観的）	それが数量である場合（相関係数を最大にすること，または，予測誤差を最小にすること）	→予測のための分析法（回帰分析，数量化Ⅰ類など）
	それが定性的なものである場合（判別的中率・相関比を最大にすること）	→判別，分類のための分析法（線型判別関数，数量化Ⅱ類など）
外的基準がない場合（主観的）	・総合合成量の分散を最大にしたり特性項目への反応パターンにもとづき相関係数を最大にすること	→成分分析や，分類のための分析法（主成分分析，因子分析，主座標分析，数量化Ⅲ類，数量化Ⅳ類など）
	・2つの対象間の関係表現にもとづく解析など	

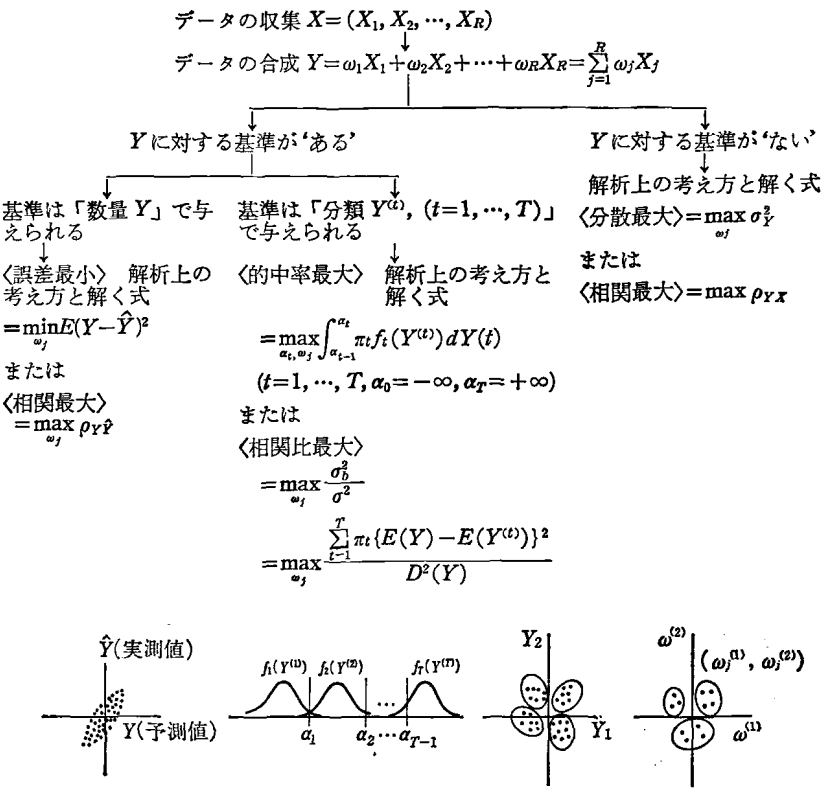

図 1 データ解析の流れ図

1.2 記号とデータの対応

理論家は一般的表現について便利で簡略なので記号表現を多く用いる．たとえば，平均値を $\bar{x}=(1/n)\sum_{i=1}^{n} x_i$ で記する．もし，x が身長であれば $\bar{x}$ は身長の平均値を意味し，体重であれば，$\bar{x}$ は体重の平均値とする．そのとき，個体数(サンプル数，または対象者数)を記号 n で，x_i は第 i 番号の個体の身長の値とする．記号 $\sum_{i=1}^{n}$ は，第1番目から第 n 番目までの和(加算)を意味する．ところで，データが身長だけでなく体重，バスト，…，とたくさんあったときの記号表現はどうしているだろうか？ データが3種類(3変量)ぐらいであれば身長に x，体重に y，バストに z と3種の英字を対応づければこと足りるが，10, 20

種類と多種類(多変量)のデータのとき，それに対応するだけの英字を用意してもよいがかえって煩雑でわかり難くくなってしまう．そこで，データの種類がどんな多くとも1種類の英字xで表現する記号の工夫をする．たとえば，平均値$\bar{x}$を身長平均$\bar{x}_1$，体重平均$\bar{x}_2$，バスト平均$\bar{x}_3$，…，など$\bar{x}$に添字を付けて各特性変量の記号化をする．m種類の特性データがあったとすれば，それらの平均値を$\bar{x}_j$, $j=1,2,\cdots,m$と記号表現すればよい．このときの平均値の表現は$\bar{x}_j=(1/n)\sum_{i=1}^{n}x_{ij}$, $j=1,2,\cdots,m$となり平均値$\bar{x}$の定義式にjを付加することで多種類のデータを英字xだけで表現ができる．x_{ij}は第i番目の個体の第j変量の値を表わしている．$j=3$が体重に対応しているとすれば$x_{20,3}$は第20番目の個体(サンプル)の体重値である．

次に，データのかたまりの表現についてふれてみよう．一般に，平均値$\bar{x}_j$, 個体値x_{ij}など値を表現するには小文字がよく用いられるのに対して，データのかたまり，たとえば第j変量$(x_{1j},x_{2j},x_{3j},\cdots,x_{nj})$を記号表現するのに大文字を用いて$X_j$と表現する．この表示法をベクトル表現という．次に，変量X_jのかたまり$(X_1,X_2,\cdots,X_m)$を記号表現するには大文字Xと記せばよい．

表 2

特性 / 個体	X_1(身長)	X_2(バスト)	X_3(ウエスト)	X_4(ヒップ)	X_5(座高)	X_6(体重)
1 駒沢	168	98	88	103	90	69
2 田中	165	97	95	105	92	78
3 〃	163	95	79	95	93	66
4 〃	179	96	79	100	99	70
5 〃	156	83	74	85	87	56
6 〃	162	85	74	87	89	55
7 〃	166	93	76	95	90	63
8 〃	152	80	62	85	84	44
9 〃	157	78	57	82	83	42
10 〃	153	85	58	87	84	45

$$X=(X_1, X_2, \cdots, X_m)=\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2m} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nm} \end{bmatrix}$$

または,

$$X=\{x_{ij}\};\ i=1,2,\cdots,n, j=1,2,\cdots,m$$

この X をマトリックス(行列)表示と呼び，x_{ij} をマトリックス要素という．マトリックス要素 x_{ij} がデータから成るとき X をデータ行列といい，一般に行側に特性，列側に個体を配置する．

1.3 データの幾何学的な読み取り方

データが作りだす構造を空間に配置し，理解しやすい方向から観察することは分析上きわめて重要である．たとえば，2特性，3個体の簡単なデータ行列で空間の図を見てみよう．

個体(対象) \ 特性	1 身長	2 体重
1 吉田	x_{11}	x_{12}
2 田中	x_{21}	x_{22}
3 池田	x_{31}	x_{32}

$$X=\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}=\begin{bmatrix} 153\,\text{cm} & 65\,\text{kg} \\ 168 & 78 \\ 172 & 82 \end{bmatrix}$$

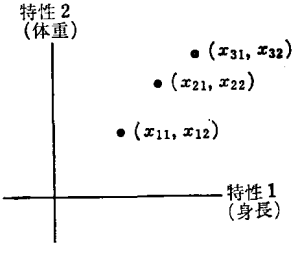

図 2 個体の図

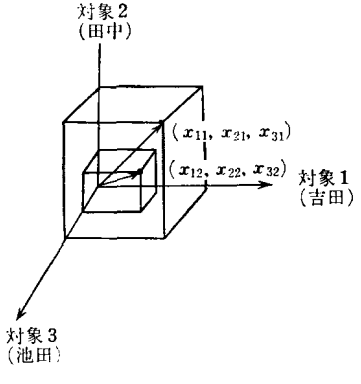

図 3 特性の図

図2は個体が直交座標の第1特性（身長）軸と第2特性（体重）軸に3個体を散布した図である．特性空間に個体がどう位置づけられ，全個体がどんな形(構造)になっているかを読み取ることができる．個体数が大量であれば星雲のごとく見えるだろう．図3は三次元の個体空間に特性がどう位置づけられ，特性がどういう方向を示しているか観測することができる．すなわち，特性ベクトル(たとえば，身長ベクトル，体重ベクトル)がどうなっているか観測できる．

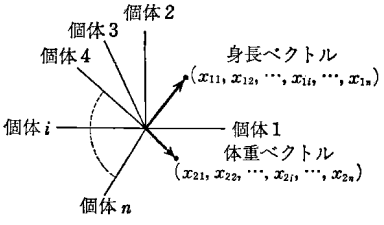

図 4　n 次元での特性の図

一般に，個体数が n であれば n 次元空間上で各特性ベクトルがどうであるかを観測していることになる．

図2で個体数が多くなると図5のごとく，特性空間における個体の構造(星雲状の形)がより浮き彫りになる．

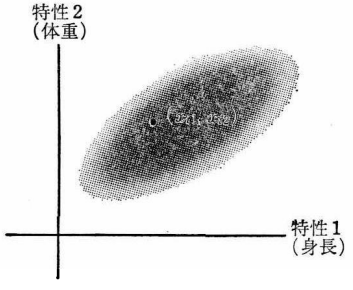

図 5　対象数が多いときの図

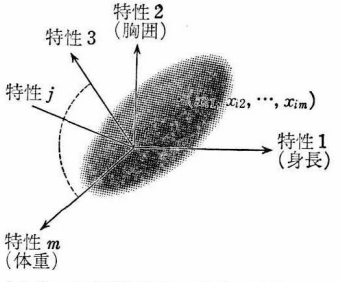

図 6　対象数が多いときの図

はじめにもいったように，実際にデータが作りだす構造がわれわれの目に写るのは三次元(立体構造)までである．そこで，一次元(直線)，二次元(平面)，三次元(立体)以内で最も理解しえる次元に m 次元の構造を射影して観察することである．時としては，もとの特性項目（身長，体重など）の座標軸 (X_1, X_2) から見るのでなく，たとえば，図7に示すように新しい座標軸

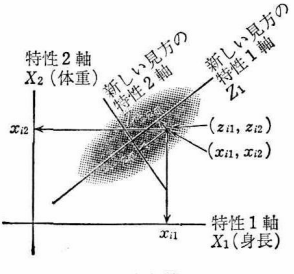

図 7

(Z_1, Z_2) を考え，同じ空間上の座標点を (x_{i1}, x_{i2}) でなく (z_{i1}, z_{i2}) で分析することもある．

この場合，新しい座標軸を z_{i1}，$i=1, 2, \cdots, n$ のバラツキが最大になるようきめてデータ解析する分析法に4章で述べる主成分分析がある．たとえば，新しい座標での Z_1 が大きい値を示せば体格が良く，Z_2 の値が大きければ体力があり，小さい値を示せば Z_1 では体格が悪く，Z_2 では体力がないと解釈(名づけ)できるならば，この時，Z_1 軸は体格軸，Z_2 軸は体力軸と見なし，新座標軸でデータ構造を分析しようとするわけである．

1.4 データの代表値とベクトル演算

ベクトル表現のしかた，基本的なベクトル演算の定義，またその幾何学的意味を理解しておくことは多元的統計データ分析において有用なことである．これを簡単に述べておく．

たとえば，2つの特性(身長，体重)について n 個のサンプルの観測値

身長：$x_1, x_2, x_3, \cdots, x_n$

体重：$y_1, y_2, y_3, \cdots, y_n$

が得られたとき，各々の分散や相関係数をベクトル表現で考えてみよう．

$$\langle\text{身長ベクトル}\rangle = X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix}, \quad \langle\text{体重ベクトル}\rangle = Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

また，ベクトル要素がすべて平均値，$\bar{x} = (1/n)\sum_{i=1}^{n} x_i$，$\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ からなる2つのベクトルを導入する．

$$\langle\text{平均ベクトル}\rangle = \bar{X} = \left.\begin{bmatrix} \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix}\right\} n\text{個}, \quad \bar{Y} = \left.\begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}\right\} n\text{個}$$

$$\langle\text{偏差ベクトル}\rangle = X' = X - \bar{X} = \begin{bmatrix} x_1-\bar{x} \\ x_2-\bar{x} \\ \vdots \\ x_i-\bar{x} \\ \vdots \\ x_n-\bar{x} \end{bmatrix}, \quad Y' = Y - \bar{Y} = \begin{bmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \vdots \\ y_i-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{bmatrix}$$

$$\langle\text{ベクトル内積}\rangle = \begin{cases} (X', X') = (x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2 \\ (Y', Y') = (y_1-\bar{y})^2 + (y_2-\bar{y})^2 + \cdots + (y_n-\bar{y})^2 \\ (X', Y') = (x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \\ \qquad \cdots + (x_n-\bar{x})(y_n-\bar{y}) \end{cases}$$

これらを利用し分散，相関係数を表現してみると

$$\langle\text{身長 } X \text{ の分散}\rangle = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}(X', X')$$

$$= \frac{1}{n}(X-\bar{X}, X-\bar{X})$$

$$\langle\text{体重 } Y \text{ の分散}\rangle = \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n}(Y', Y')$$

$$= \frac{1}{n}(Y-\bar{Y}, Y-\bar{Y})$$

$$\langle\text{身長 } X \text{ と体重 } Y \text{ の共分散}\rangle = \sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{n}(X', Y') = \frac{1}{n}(X-\bar{X}, Y-\bar{Y})$$

$$\langle\text{身長 } X \text{ と体重 } Y \text{ の相関係数}\rangle = \rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}}$$

$$= \frac{(X-\bar{X}, Y-\bar{Y})}{\sqrt{(X-\bar{X}, X-\bar{X}) \cdot (Y-\bar{Y}, Y-\bar{Y})}}$$

で表現できる．

ところで，幾何学的には偏差ベクトル X' の長さは統計量の標準偏差の $\sqrt{n}$ 倍に対応し，2つのベクトルのなす角の余弦 $\cos\theta$ が相関係数に対応する．すなわち，

$$\cos\theta = \frac{(X', Y')}{\sqrt{(X', X') \cdot (Y', Y')}} = \rho_{XY}$$

となり，$-1 \leqq \cos\theta = \rho_{XY} \leqq 1$ という関係も出て来る．

1.5 記号と統計量の対応

前にも述べたが一般に記号表現は複雑なものを簡単に表わすことができるので便利である．平均値を $\bar{x}$，m や μ で表わし，分散を s^2 や σ^2 で表わしたりする．英文字を使用したときはサンプルでの統計量，ギリシャ文字を使用したときは母集団での統計量の話にと，専門家の間では暗黙の了解のうちに用いられている．この辺のことは明確な約束がなされていないむきがあるので統計関係の著書を読んだとき，読者にとって混同しやすく，サンプルなのか母集団での話なのか適当に解釈してしまうことが多い．

時として，統計家はサンプルにも母集団にも通用する記号で平均値と分散を表現する．平均値は期待値 Expectation の略 E で，分散を偏差 Deviation の略 D をもって表現する．たとえば，変量 X の平均値を $E(X)$，分散を $D^2(X)$ と記号表現する．変量 X が

1） 離散的データで $X=(x_1, x_2, \cdots, x_n)$ であれば

$$\langle \text{変量 } X \text{ の平均値} \rangle = E(X) = \frac{x_1+x_2+\cdots+x_n}{n}$$

2） 連続量のデータで変量 X の確率密度関数 $f(x)$ がわかっていれば

$$E(X)=\int_{-\infty}^{\infty} xf(x)\,dx \quad \text{ただし} \quad \int_{-\infty}^{\infty} f(x)\,dx=1$$

となる．

次の $E(X-E(X))^2$ は何を意味しているか考えてみよう．これは，たとえば，変量 X が身長とすると，身長の平均を各サンプルの身長の値から引いて2乗したものの平均値と理解すればよいのである．いいかえれば，偏差の2乗平均，すなわち分散を意味していることになる．

$$D^2(X)=E(X-E(X))^2$$

記号 $E(X)$ は括弧内の量をたし合わせる性質から $E(X)$ に関する次のような演算式が成り立つ．

1） いくつかの変量 $X_1, X_2, \cdots, X_m$ の和の平均値は

$$E(X_1+X_2+\cdots+X_m)=E(X_1)+E(X_2)+\cdots+E(X_m)$$