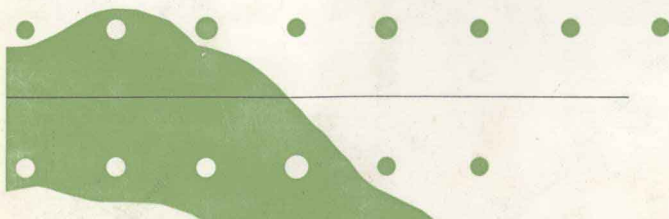
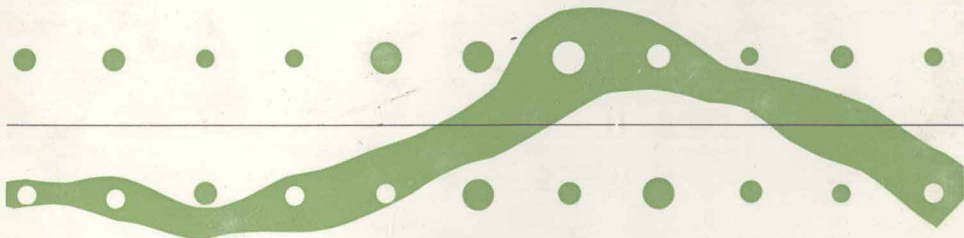
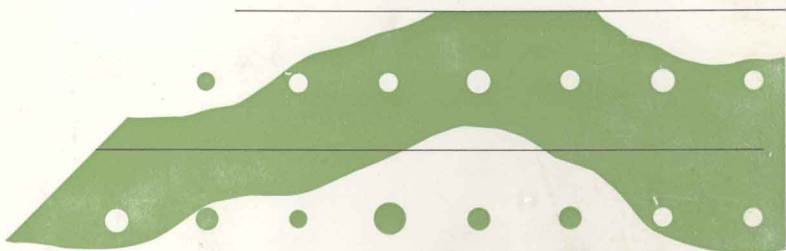


環境統計学

情報処理の考え方

鈴木栄一著



環境統計学

情報処理の考え方

鈴木栄一著

地人書館

すずき いいち
鈴木栄一

1924年福島県に生まれる。中央气象台，気象研究所などを
を経て現在青山学院大学教授。理学博士。

<著書>

『実例統計学演習』(共著) <青林書院，1961> / 『気象統計学』 <地人書館，1968> / 『PROGRAMING / FORTRAN 入門』 <地人書館，1969> / 『統計数値表』(共編著) <日本規格協会，1972> / 『情報科学の基礎』(共著) <青学出版，1973> / 『計量地理学への招待』(共編著) <青学出版，1974> / 『例解入門 推測統計法』(共著) <朝倉書店，1974>

環境統計学 <情報処理の考え方>

©昭和54年2月20日

昭和54年10月15日

第1版第1刷発行

第1版第2刷発行

著 者 鈴木 栄 一

発 行 者 中 田 威 夫



発 行 所 株式会社 地 人 書 館

〒112 東京都文京区後楽 1-1-10

電話 03 (815)4422 (代表)

振替口座 東京 6-1532番

定価はカバーに表示してあります。

本書の内容の一部あるいは全部を無断で複写複製（コピー）することは，法律で認められた場合を除き，著作者および出版社の権利の侵害となりますので，その場合には予め小社あて許諾を求めてください。

序

最近十数年間に、我々を取りまく社会環境は大きく変貌したといわれている。

全国的高速道路網が整備拡充され、新幹線計画が具体化し、都市及びその周辺での開発が進展し、コンピュータは日常化され、都心での超高層ビル、マンションが林立し、自動車や冷暖房が大部分の家庭に用意されている。これにともない、大気汚染、振動公害、騒音被害といった多様な災害に強い関心がもたれ、調和のとれた福祉社会への要望が各方面から出される様になった。

多くの公的、私的機関や大学内で環境科学に関する調査研究が急速に進展するところで、必然的に起こる課題は環境計測データの統計処理をどうすればよいかである。そこで、必ずしも統計学の教科書にある諸方法がそのまま適用できない場面が多くなり、手法の開発を必要とする状況がでてきた。

たとえば、新幹線トンネル工事をしている周辺地域で、樹木（植生）の変化があらわれたとき、工事自体による地下水変化なのか、寡雨多日照によるものか、総合的判断をすべき場面では、従来の統計手法はそのまま適用されない。また大規模な開発や新施設設置にともなう環境変化の予測や将来おこりうる状況の推測といっても、まったく経験のないことなので、統計的考察に限度があり、そこで新手法の開発が必要となっている。

ここでは、筆者が環境データの収集や統計処理に参画したり、関係したりし

た各種委員会のなかで得られた知見、同時に断片的ではあるが、調べた内外の諸文献のなかから、統計的手法のうち、多少とも役立ちそうなもの、および将来の新手法開発の基盤になりそうなものを少しまとめた形で記述してみた。

したがって、完全な体系的記述ではなく、統計学の教科書的にみれば少し型破りであるが、実際問題を処理する上で、十分参考になると考えている。

筆者は昨年夏、アメリカのコロラド大学とボルダーのNCAR (National Center for Atmospheric Research) を短期研究留学で訪問し、多くの教訓を得て帰国した。

アメリカでは自然保護を重視し、自然景観をそこなわない開発（クリーク設置や道路づくりなど）に思い切った投資が行われ、自動車王国であるアメリカでも広大な地域のあちこちに乗り入れ禁止区域が明示されている。また、環境科学関係コンサルタントの調査結果や意見が十分にとり入れられている。

アメリカでの環境データの統計処理は素朴な手法ながら種々工夫されていた。この点は日本と大差ないので、今後ここにのべられた手法と同様に、開発が進むであろう。

したがって環境統計学（仮にこのようなタイトルを用いた）で取り上げられるべき処理手法も、現在はかなり流動的なものであると考えられる。

本書は当初、環境情報科学センターから刊行されたものであるが、その後、種々の事情から、地人書館より同じタイトルの書名で出すこととなり、内容的に多少修正することになった。

この点については、快よく本書の刊行を御了承頂いた地人書館の中田威夫社長、ならびに内容の不備な箇所を丹念に点検し修正して頂いた同社の上條 宰取締役に心から感謝したい。

同時に、本書の初版刊行時から終始御協力された諏訪寿夫氏に厚く感謝する次第である。

昭和54年3月

著 者

目 次

序

第1章 測定値統計と回帰

1.1 平均値，中央値とパーセント点	1
1.2 降水現象の相関分析	6
1.3 回帰分析における説明変数選択	9

第2章 極値と計画値

2.1 極値記録のあらわれ方	15
2.2 関与条件と極値	19
2.3 計画値	25

第3章 情報理論

3.1 情報とはなにか	31
3.2 事象系と情報量	32
3.3 事象の結合と伝達情報量	35
3.4 分布，相関と情報量	39

第4章 特殊な統計処理

4.1 欠測値の処理方法	45
4.2 乱数の利用法	47

第5章 予測の方法

5.1 現代の予測技法	53
5.2 降水の予測	56
5.3 都市交通の予測	58
5.4 情報理論による予測と判断	61

第6章 判別分析と予測

6.1 判別の基本的考え方	67
6.2 判別式を導くためのモデル	70
6.3 多変量2項分布とは	73
6.4 多変量2項分布における判別指標の線型化	77
6.5 線型判別指標の応用例	79

第7章 都市・交通データの処理

7.1 都市計画のシステム	87
7.2 交通データの統計処理	90

第8章 大気汚染データの処理

8.1 汚染と要因との関係分析	95
8.2 汚染濃度と気象要因の時系列分析	98

第9章 環境データの多変量解析

9.1 多変量解析の考え方と方法	103
9.2 農業生産物の成分分析と因子分析	104

9.3 名古屋地域の因子分析	109
9.4 気象現象の多変量解析	114
9.5 崩壊現象の判別分析	121

第10章 数値分類とパターン

10.1 数値分類とは	125
10.2 数値分類の手法	128
10.3 数値分類の応用例	130
10.4 農業立地とパターン	138
10.5 連続パターンの表現法	140

第11章 ダミー変数と要因

11.1 ダミー変数とは	145
11.2 特殊ダミー変数による要因表現	147
11.3 ダミー変数と農業問題	149

第12章 数量化手法 (1)

12.1 数量化1類の考え方	153
12.2 数量化1類の幾何学的考察	156
12.3 数量化1類の応用例	159
12.4 数量化2類の考え方	170
12.5 数量化2類の応用例	174

第13章 数量化手法 (2)

13.1 数量化3類とは	183
13.2 数量化3類の手法と計算手順	186
13.3 定性的分割表における数量化	187
13.4 数量化3類の変形手法と応用	188

13.5 数量化4類とは	190
13.6 数量化3, 4類の適用面と問題.....	192
結び	195
参考文献	197
事 項 索 引	

第1章 測定値統計と回帰

1.1 平均値，中央値とパーセント点

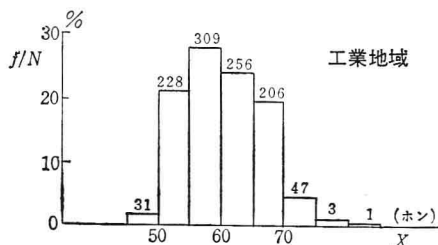
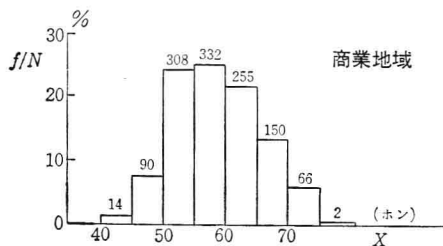
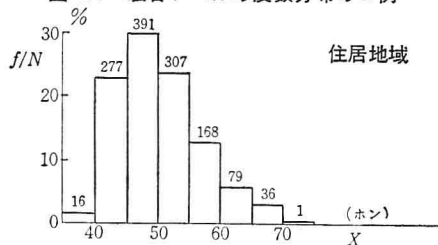
都市の騒音レベルをホン単位で計測し，騒音の環境基準をきめるのに，よく中央値と90パーセント点の上端と下端が用いられている。

図1.1は東京都の3地域における騒音レベルの度数分布で，各地域の総数 N のうち，各階級に入る度数 f を調べ，相対度数分布図にまとめたものである。

いずれも正規分布より少しくかたよっている。この場合の問題は，いわゆる連続変量 X （したがって確率密度関数 probability density function，略して *p.d.f.* が存在する）の分布における代表値パーセント点 (percentage point) の安定性である。 X の分布関数 $F(x) \equiv P(X \leq x)$ が与えられたとき， $F(x)=0.50$ に対応する x の値が中央値である。ここで安定性とは，データ（標本）のとり方をかえてもあまり変化しないとき，安定しているという意味で用いられるものとする。分りやすくするため，仮定として図1.1のような騒音レベル X は正規分布にしたがい，その *p.d.f.* が

$$f(x; \mu, \sigma) = dF(x)/dx = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (1.1)$$

図 1.1 騒音レベルの度数分布の1例



〔注〕 縦軸は相対度数 f/N の%を示す。

〔資料〕 東京都公害研究所の調査による。

で与えられるとする (分布が Gamma 分布でも同様な考察が可能である)。

n コの標本変量 (sample variate) $x_1, x_2, \dots, x_n$ が与えられたとき、つぎの定理が知られている。

(定理 1) Cramér-Rao の不等式
確率変数 X の $p.d.f.$ を $f(x; \theta)$ とし、母数 (parameter) θ の不偏推定を $\tilde{\theta}$ とするとき、 $f(x; \theta)$ に関する適当な正則条件のもとで、 $\tilde{\theta}$ の分散 $V(\tilde{\theta})$ に対し、つぎの不等式が成り立つ。

$$V(\tilde{\theta}) \geq 1/n \int_{-\infty}^{+\infty} \left\{ \frac{\partial \log f(x; \theta)}{\partial \theta} \right\}^2 \times f(x; \theta) dx \quad (1.2)$$

(定理 2)

確率変数 X の $p.d.f.$ を $f(x; \theta)$

とし、 θ の最尤推定量を $\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ とすると、 $f(x; \theta)$ に関する若干の正則条件下で

$(\hat{\theta} - \theta)$ は $n \gg 0$ のとき漸近的に正規分布

$$N(0, 1/n \int_{-\infty}^{+\infty} \left\{ \frac{\partial \log f(x; \theta)}{\partial \theta} \right\}^2 f(x; \theta) dx)$$

にしたがう。

これら 2 つの定理において $\theta = \mu$ とし、正規分布の場合に適用すると、つぎのようになる。

標本平均 (sample mean) $\bar{X}$, 標本中央値 $X_{(m)}$ (sample median) とともに不偏

推定で、 $n \gg 0$ のとき、

$\bar{X}$ は正規分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$, $X_{(m)}$ は正規分布 $N\left(\mu, \frac{\pi}{2} \frac{\sigma^2}{n}\right)$ にそれぞれしたがる。

つまり、 $X_{(m)}$ の分散は $\bar{X}$ の分散の $\pi/2=1.57$ 倍でいつも大きい。平たくいえば、データ (標本) のとり方を変えたときのちがいを示すならば $X_{(m)}$ の方が $\bar{X}$ より大きく、それだけ安定性を欠くことになる。(定理 1) における Cramér-Rao の不等式で、 $\bar{\theta}=X_{(m)}$ とすると明らかに不等号 $>$ が成り立つが、 $\bar{\theta}=\bar{\theta}=\bar{X}$ とすると $n \gg 0$ のとき等号 $=$ が成り立つ。

従来の騒音統計で、なぜ中央値をとるのか不明確であるが、正規分布をもとのデータに適用する限り、平均値でよいはずである (代表値はむしろ標本平均でよい)。次にパーセント点、とくに 90 パーセント点の上端、下端を

$$\int_{-\infty}^L f(x; \theta) dx = \int_U^{+\infty} f(x; \theta) dx = 0.05 \quad (1.3)$$

で定義したとき、下端 L 、上端 U を n コの標本 $x_1, x_2, \dots, x_n$ から推定したときの安定性を考察する。

標本変量を大きさの順にならべ

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(i)} \leq x_{(i+1)} \leq \dots \leq x_{(n)}$$

としたものを順序標本 (ordered sample), $x_{(i)}$ を i 番目の順序統計量 (order statistic) という。このとき L は $n/20$ にもっとも近い i に対する $X_{(i)}$, U は $19n/20$ にもっとも近い i に対する $X_{(i)}$ で推定するものとしたとき、 $X_{(i)}$ が標本のとり方を変えてもあまり変化しない (ばらつかない) とき、安定していることになるを考えよう。

一般に連続な X の $p.d.f.$ を $f(x)$ とすると $X_{(i)}$ の分布の $p.d.f.$ は

$$\begin{aligned} g(x_{(i)}) &= \frac{n!}{(n-i)!(i-1)!} \left(\int_{-\infty}^{x_{(i)}} f(x) dx \right)^{i-1} \\ &\quad \times \left(\int_{x_{(i)}}^{+\infty} f(x) dx \right)^{n-i} f(x_{(i)}) \end{aligned} \quad (1.4)$$

で与えられ、これより $X_{(i)}$ のばらつきを示す分散を $V(X_{(i)})$ と書くと、結局、

$$V(X_{(i)}) = E(X_{(i)}^2) - E(X_{(i)})^2 \quad (1.5)$$

ただし,

$$\left. \begin{aligned} E(X_{(i)}^2) &= \frac{n!}{(i-1)!(n-i)!} \int_{-\infty}^{+\infty} x^2 f(x) \\ &\quad \times \{\Phi(x)\}^{i-1} \{1-\Phi(x)\}^{n-i} dx \\ E(X_{(i)}) &= \frac{n!}{(i-1)!(n-i)!} \int_{-\infty}^{+\infty} x f(x) \\ &\quad \times \{\Phi(x)\}^{i-1} \{1-\Phi(x)\}^{n-i} dx \\ \Phi(x) &= \int_{-\infty}^x f(x) dx \end{aligned} \right\} \quad (1.6)$$

となる. この $V(X_{(i)})$ の数値計算は面倒で, 今のところ X が規準正規分布の場合, $n=2\sim 50$ に対し, 数表が作成されているにすぎない (E. D. Teichroew (1956) が $n=2\sim 20$ に対し, 計算したものを鈴木・平川 (1969) が検討し, 数値積分方式をかえてさらに $n=2\sim 50$ に対する $V(X_{(i)})$ とその平方根である標準偏差 $D(X_{(i)})$ の値を正確にもとめ, 数表化した).

この場合, 90パーセントを内部とする左右いずれも5%ずつの上端, 下端について考察するのであるから, 正規分布の対称性により上端, 下端とも同じ標準偏差をもつはずであるから, ここでは下端の標準偏差 $D(X_{(i)})$ だけとりあげて, その数値の一部を下に示す.

表 1.1 順序統計量 $X_{(i)}$ の標準偏差 $D(X_{(i)})$ の計算例

n	i	$D(X_{(i)})$	n	i	$D(X_{(i)})$
20	1	0.5250682	40	2	0.3534787
				3	0.3042629
25	1	0.5084396	45	1	0.4704831
	2	0.3831173		2	0.3469120
30	1	0.4958204		3	0.2978173
	2	0.3709130	50	1	0.4644485
35	1	0.4850688		2	0.3412921
	2	0.3613124		3	0.2923283
40	1	0.4774849		4	0.2645392

たとえば $n=30$ のとき $n/20=1.5$ であるから, 下端を $X_L = \{X_{(1)} + X_{(2)}\} / 2$ にすれば,

$$\begin{aligned}
 V(X_L) &= E(X_L^2) - E(X_L)^2 \\
 &= \frac{1}{4} \{E(X_{(1)}^2) + 2E(X_{(1)}X_{(2)}) + E(X_{(2)}^2)\} \\
 &\quad - \frac{1}{4} \{E(X_{(1)})^2 + 2E(X_{(1)})E(X_{(2)}) + E(X_{(2)})^2\} \\
 &= \frac{1}{4} [D(X_{(1)})^2 + D(X_{(2)})^2 + 2\{E(X_{(1)}X_{(2)}) \\
 &\quad - E(X_{(1)})E(X_{(2)})\}] \doteq 0.18632
 \end{aligned} \tag{1.7}$$

$$\therefore D(X_L) \doteq 0.43176$$

となる. これは X が規準正規分布 $N(0, 1^2)$ にしたがう場合であるから, 標本からパーセント点の推定値が得られたら, その分散をもとめるのに実際の標本分散不偏推定 $\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$ をこれにかけて修正する必要がある. $n > 50$ のとき, 表 1.1 で示された値より小さくなることは明らかで, つぎのようにすればよい.

正規分布 $N(\mu, \sigma^2)$ からの大きさ n の標本 $x_1, x_2, \dots, x_n$ による標本平均 $\bar{x}$, 標本分散 s^2 とすると,

$$\bar{x} = \sum_{i=1}^n x_i / n, \quad s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n \tag{1.8}$$

と書かれ,

母集団分散 σ^2 に対する信頼係数 $1-2\alpha$ の信頼限界は,

$$[\sigma_1^2 = ns^2 / \chi_{n-1}^2(\alpha), \quad \sigma_2^2 = ns^2 / \chi_{n-1}^2(1-\alpha)]$$

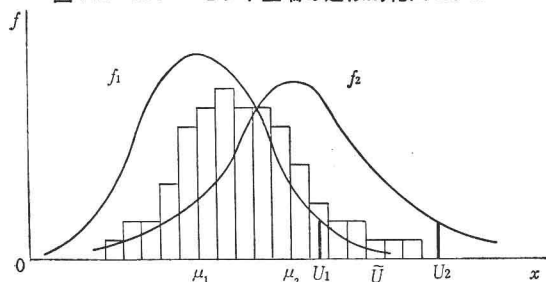
ただし, $\chi_{n-1}^2(\alpha)$, $\chi_{n-1}^2(1-\alpha)$ はそれぞれ自由度 $= n-1$ の χ^2 -分布における $100\alpha\%$ 点, $100(1-\alpha)\%$ 点である.

また母集団平均 μ に対する信頼係数 $1-2\alpha$ の信頼限界は,

$$[\mu_1 = \bar{x} - s \cdot t_{n-1}(\alpha) / \sqrt{n-1}, \quad \mu_2 = \bar{x} + s \cdot t_{n-1}(\alpha) / \sqrt{n-1}]$$

ただし $t_{n-1}(\alpha)$ は自由度 $= n-1$ の t -分布における $100(1-2\alpha)\%$ 点である.

図 1.2 90 パーセント上端の近似的範囲モデル



そこで図 1.2 に示されるように、 $n > 30$ のときは大標本的近似が許されるところとて、

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left\{ -\frac{(x-\mu_i)^2}{2\sigma_i^2} \right\} \quad (i=1, 2) \quad (1.9)$$

とおくと、90 パーセント上端が、

$$[U_1 = \mu_1 + 1.645\sigma_1, U_2 = \mu_2 + 1.645\sigma_2]$$

に入る確率は標本平均と標本分散が独立に分布するからだいたい $(1-2\alpha)^2$ となる。

したがって、この範囲がなるべく狭く、かつ $(1-2\alpha)^2$ がなるべく大きくなっているとき、90 パーセント上端の標本推定 $\tilde{U} = \bar{x} + 1.645 s_1$ が安定していると考えてよいであろう。

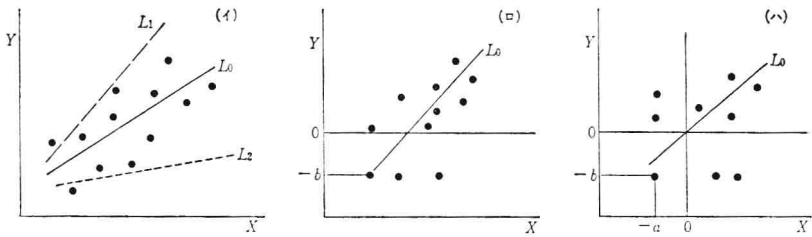
1.2 降水現象の相関分析

ここではどの統計学の書物にもある相関係数の説明を省略する。

2 地点間の降水の相互関係、1 地点の降水と他の連続な物理量との相互関係を量的に把握する場合、0.0 ミリと無降水（－）の区別をどうするかがしばしば問題となる。

一般に 2 つの変量 X, Y の相関関係を分析する場合、図 1.3(i)のごとく、 X, Y とも連続な物理量であれば、得られた n 組のデータ (X_i, Y_i) ($i=1, 2, \dots, n$) を方眼紙上に記入し、まず直線をひく。そして X, Y の関係を示す直線としては L_1, L_2 のように記入された n コの点全体からズレたものでなく、 L_0 のように n コの点全体の中央部に直線をひくのが常識的である。この常識

図 1.3 相関分析における 3 コの散点図モデル



的直線の式は各点から L_0 までの距離の平方和を最小化することによってきめられる（最小 2 乗法では Y 方向の距離の平方和 S^2 を最小にする）。

そこで、 X が連続な物理量で、 Y が降水現象の場合は、図 1.3(II) のように、無降水に $-b$ ($b > 0$) を与え、直線 L_0 をひいたとき、 L_0 からの距離平方和が最小になるように b をきめることが与えられる。この考え方は得られた資料の組が表 1.2 のごとくなっているとき、 X 、 Y の相関係数 2 乗 $\rho^2(X, Y, b)$ を最

表 1.2 変量 X 、 Y の対応モデル

X	x_1	$x_2 \cdots \cdots x_{n_1}$	$x_{n_1+1} \cdots \cdots x_{n_1+n_2}$
Y	$-b$	$-b \cdots \cdots -b$	$y_1 \cdots \cdots y_{n_2}$

大化することにより達せられる。ただし、 $\rho^2(X, Y, b)$ はつぎの形で示される。

$$\rho^2(X, Y, b) = \text{Cov}^2(X, Y, b) / s_x^2 \cdot s_y^2(b) \quad (1.10)$$

ただし、 $\text{Cov.}(X, Y, b)$ は X 、 Y の共分散、 s_x^2 、 $s_y^2(b)$ はそれぞれ X 、 Y の分散である（ X 、 Y の共分散、 Y の分散はともに無降水に与えられる b の値のきめ方によりかわりうるので b の関数となるところから、それぞれ $\text{Cov.}(X, Y, b)$ 、 $s_y^2(b)$ と書かれる）。

そこで、 $\partial \rho^2(X, Y, b) / \partial b = 0$ 、 $\partial^2 \rho^2(X, Y, b) / \partial b^2 \leq 0$ を満足するように b をきめる式をつくるとつぎのようになる（途中の計算式省略）。

$$b = \frac{CQ(D - B \cdot C) + (B - M)Q}{P(D - B \cdot C) + (M \cdot C - D)} \quad (1.11)$$

ただし、 $n_1 + n_2 = n$ 、 $P = n_1/n$ 、 $Q = n_2/n$

$$M = \sum_{i=1}^{n_1} x_i / n, \quad B = \sum_{i=1}^{n_1} x_i / n_1, \quad C = \sum_{i=1}^{n_2} y_i / n_2, \quad D = \sum_{i=1}^{n_2} x_{n_1+i} y_i / n_2$$

さらに, X, Y とも降水現象でその散布図が図 1.3(イ)のごとくなっている場合は, それぞれの無降水に $-a, -b (a, b > 0)$ を与え, X, Y の相関係数の 2 乗 $\rho^2(X, Y, a, b)$ を最大にするよう

$$\left. \begin{aligned} \partial \rho^2(\quad) / \partial a &= 0, \quad \partial \rho^2(\quad) / \partial b = 0, \\ \frac{\partial^2 \rho^2(\quad)}{\partial a^2} &< 0, \\ \frac{\partial^2 \rho^2(\quad)}{\partial a^2} \frac{\partial^2 \rho^2(\quad)}{\partial b^2} - \left(\frac{\partial^2 \rho^2(\quad)}{\partial a \partial b} \right)^2 &> 0 \end{aligned} \right\} \quad (1.12)$$

の条件を満たすように a, b をもとめる 2 元連立一次方程式をつくり, それをとくことができる。

最後に, $X_1, \dots, X_p$ を $p (> 2)$ の連続物理量, Y を降水現象とするとき, Y と $X_1, \dots, X_p$ との重相関係数 2 乗 $R^2(Y: X_1, \dots, X_p; b)$ を最大にするよう b をきめることができる。

この目的のためには p 次元空間での直交変換

$$(X_1, \dots, X_p) \rightarrow (\zeta_1, \dots, \zeta_p)$$

$$(\text{Cov. } (\zeta_i, \zeta_j) = 0, i \neq j)$$

によって統計的に独立な変数 $(\zeta_1, \dots, \zeta_p)$ におきかえたとき, つぎの関係が満足されることを利用すればよい。

$$\left. \begin{aligned} R^2(Y: X_1, \dots, X_p; b) &= r^2(Y, \zeta_1; b) + \dots + r^2(Y, \zeta_p; b) \end{aligned} \right\} \quad (1.13)$$

ただし, $r(Y, \zeta_i; b)$ は Y と ζ_i との単相関係数

つまり, $\partial R^2(Y: X_1, \dots, X_p; b) / \partial b = 0$ から導かれる方程式の正根をもとめればよい。この場合の直交変換としては, Schmidt の直交化法 ($X_1, X_2, \dots, X_p$ の順に直交化する方法) と成分分析的直交化法 (X の順序に無関係な直交一次変換の方法) があるが, 後者の方が一般性があってよい。後者ではよく知られたつぎの関係が成り立つ。

$$\left. \begin{aligned} E(\zeta_i) &= 0, \quad \text{Var. } (\zeta_i) = \lambda_i, \\ \text{Cov. } (\zeta_i, \zeta_j) &= \lambda_i \delta_{ij}, \quad \sum_{i=1}^p \lambda_i = p \end{aligned} \right\} \quad (1.14)$$