2008 年度前期　統計学で考える　第１１回

熱い夏，寒さの夏―クラスター分析

環境問題の中で，最近とくに話題になっているのが「地球温暖化」です．このような気象の問題を考えるには，天気図と実際の気候との関係を知り，特徴的な気候があらわれる代表的な天気図のパターンを導くことが有効です．そのためには，天気図という複雑なものを，代表的なパターンに分類する必要があります．このために用いられるのが「クラスター分析」という統計的分類手法です．今回の講義では，まずクラスター分析について簡単に説明し，これを使って極端な冷夏と猛暑になった 1993, 94, 95 年の天気図から代表的天気図パターンを導き出した研究を紹介します．

多変量データと散布図

第７回の講義で,「データの分布」について説明しました．「(測定対象や現象が) 分布する」とは,「ある測定対象や現象から得られる数量が大小ばらばらである」という意味です．例えば,「日本人男性の身長」は分布する，ということができます．この例での「身長」のように，大小いろいろな値になる数量のことを変量といいます．統計データ解析とは，一言でいえば，分布している変量から情報を引きだす手法ということができます．

「統計学で考える」の前半では,「身長」「試験の点数」といったひとつの変量について，標本から変量の度数分布についての推定・検定をする，といった手法を説明しました．しかし，世の中には２つ以上の変量で表現されるデータもたくさんあります．例えば試験の点数の場合でも，一人の人の成績は数学，英語,... といった複数の科目の点数（変量）の組み合わせで評価されます．このように，ひとつの個体（人など）が複数の変量の組み合わせで表されているデータを多変量データといい，多変量データの分布を取り扱う統計手法を多変量解析といいます．

ひとつの変量の分布を目に見えるように表現するために，ヒストグラムを用いることを第４回の講義で説明しました．これに対して，多変量データの分布を目に見えるように表現するのに用いられるのが散布図です．

表 1 は，日本のいくつかの都市の緯度と年平均気温を表しています．このデータは，各都市が緯度と気温の２つの変量で表されている多変量データです．このデータの分布を目に見えるように，緯度と気温の２つの変量をそれぞれ横軸・縦軸とし，各都市を対応する緯度・気温の位置に配置します．例えば，札幌市は北緯 43.05 度，年平均気温 8.0 ℃ですから，横軸 43.05，縦軸 8.0 の位置に印をつけます．このようにして個体（ここでは都市）を配置した，図 1 のような図を散布図といいます．この場合は変量が２つなので，散布図は横軸縦軸でできる平面になります．変量が３つ以上になると軸も３つ以上になりますが，この場合も紙の上に描けないだけで，理屈には違いはありません．図 1 の散布図を見ると，一見して各都市がほぼ直線に沿って並んでおり,「緯度が高（低）いと気温が低（高）い」という傾向があることがわかります．このような変量の関係を知る統計手法は相関分析・回帰分析と言われており，この講義では扱いませんが「統計データ解析Ｂ」で取り扱います．

地名	緯度（度）	気温（℃）
札幌	43.05	8.0
青森	40.82	9.6
秋田	39.72	11.0
仙台	38.27	11.9
福島	37.75	12.5
宇都宮	36.55	12.9
水戸	36.38	13.2
東京	35.68	15.3
新潟	37.92	13.1
長野	36.67	11.4
静岡	34.97	16.0
名古屋	35.17	14.9
大阪	34.68	16.2
鳥取	35.48	14.4
広島	34.40	15.0
高知	33.55	16.3
福岡	33.92	16.0
鹿児島	31.57	17.3
那覇	26.20	22.0

表 1: 日本の年の緯度と気温

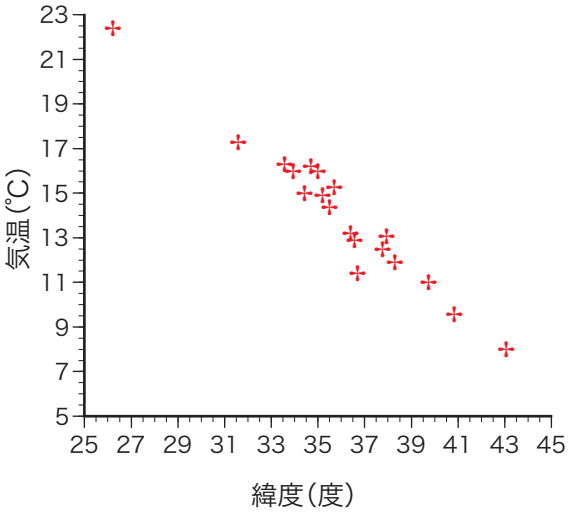

図 1: 散布図：緯度と気温の関係

クラスター分析

これに対して，表 2 は，「食品 1〜8」について，100 グラムあたりの「熱量」と「ナトリウムの含有量」を表したものです[1]．表 2 のデータを散布図に表すと図 2 のようになります．この散布図では，個体（ここでは食品）が直線に沿ってならんでいるといった様子は見られませんから，「熱量が高いとナトリウムの含有量が少ない」などといった傾向を見いだすことはできません．そのかわりに，1〜8 の食品がおおまかに 3 つの塊（クラスター）に分類できそうに見えます．各クラスターに入っている食品は，散布図上で近い位置にある，すなわち「似た性質の食品」と考えることが出来ます．つまり，1〜8 の食品は，おおまかにいってどんな性質の食品に分類できるかがわかります．このような分類を行う統計手法がクラスター分析です．

k-means 法

クラスター分析は，クラスターへの「もっとも良い」分類がされるような，クラスター間の境界を求めます．「良い」分類とはどういう分類でしょうか？ クラスターへの分類とは，クラスター内にある個体（＝散布図上の点）をクラスターの重心で代表して表現しようという考えだといえます．このとき，クラスター内の個体と重心とのへだたりが小さいほど，すなわち各個体がクラスター内でより「かたまって」いるほど，重心はより良い代表であると言えます．つまり，クラスター内の個体の分散が小さいほど，良いクラスターだというわけです．重心の位置は，各変量（＝各軸）について，それぞれクラスター内で平均を求めれば得られます．そして，各個体についてそれが属するクラスターの重心との距離の 2

[1] この表は説明のために作成したもので，実際の食品とは関係ありません．

食品	熱量 (kcal)	ナトリウム含有量 (mg)
1	6	5
2	20	5
3	4	2
4	10	3
5	18	3
6	8	9
7	12	11
8	13	8

表 2: 食品の熱量とナトリウム含有量

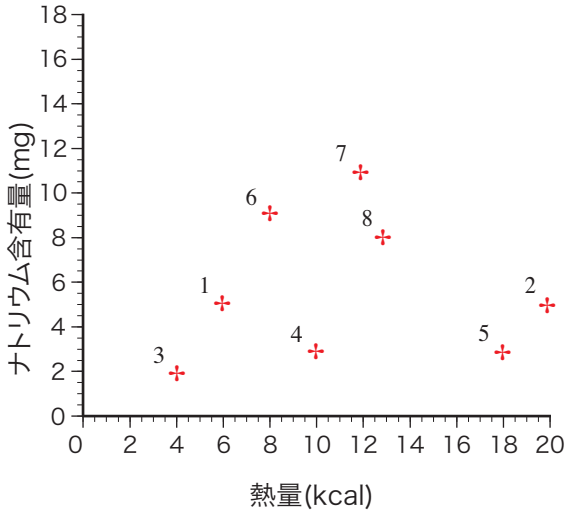

図 2: 散布図

乗を求め，それを全部の個体について合計したものが小さいほど，「良い」分類であるということになります．

表 2・図 2 のような簡単な例なら，「良い」分類をするのはそう難しくはありません．しかし，たくさんの変量で表されるたくさんの個体があるときには，クラスターの個数を決めておいたとしても，その個数のクラスターへの分け方は無数にあります．ですから，クラスターへの一番「良い」分類を求めるのは大変難しい問題となります．この問題を解くにはさまざまな方法がありますが，ここでは代表的な ***k*-means 法**を，表 2・図 2 を例に説明します．*k*-means 法では，いくつのクラスターに分けるかはあらかじめ決めておきます．そして，まず適当にクラスターに分類し，これをクラスターへの一番「良い」分類に徐々に近づけてゆきます．

ステップ１　クラスターの重心の「候補」を適当に配置します．そして，各個体はそれぞれ一番近い「重心候補」のクラスターに属するとして，各個体をクラスターにいったん分類します．

図 3 の例では 3 つのクラスターに分類するとし，最初の「重心候補」を，（熱量，含有量）= (8, 4)［★１］, (10, 12)［★２］, (16, 8)［★３］としました．その結果，クラスター (I) には 1, 3, 4 番の個体が，クラスター (II) には 6, 7 番の個体が，クラスター (III) には 2, 5, 8 番の個体がそれぞれ分類されます．

ステップ２　分類された各クラスターについて，あらためてそのクラスター内の各個体の重心の位置を求めて，ステップ 1 で決めた「重心候補」を移動します．1, 3, 4 番の個体の重心は，熱量 = (6 + 4 + 10)/3 = 6.3，含有量 = (5 + 2 + 3)/3 = 3.3 となります．同様に 6, 7 番の個体の重心は熱量 = 10，含有量 = 10，2, 5, 8 番の個体の重心は熱量 = 17，含有量 = 5.3 となります．図 4 のように，これらの位置に★１，★２，★３がそれぞれ移動します．

ステップ３　各個体はステップ２で求めた重心のうち一番近いもののクラスターに属するとして，分類をやり直します．

図 4 で，８番の個体と新しい重心★２の距離との２乗は $(13 - 10)^2 + (8 - 10)^2 = 13$，★３との距離の２乗は $(13 - 17)^2 + (8 - 5.3)^2 = 23.3$ ですから，８番の個体は★２のクラスターに属することになり，図 4 の点線のようにクラスターの境界線が変わります．

ステップ４　ステップ２，３を，重心の位置に変化がなくなるまで繰り返します．

このようにして，一番「良い」分類に逐次近づけてゆきます．

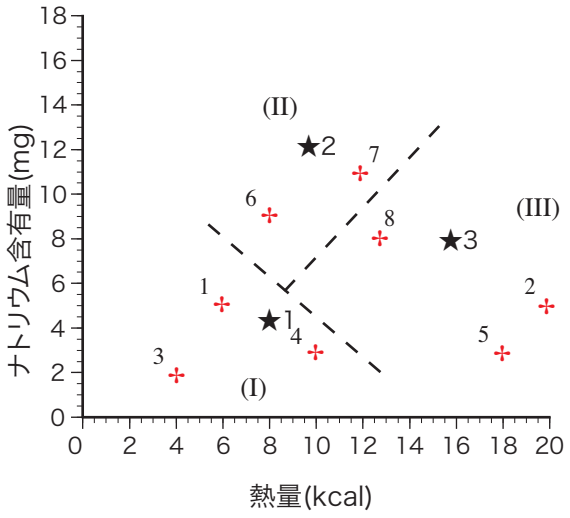

図 3: 初期「重心候補」

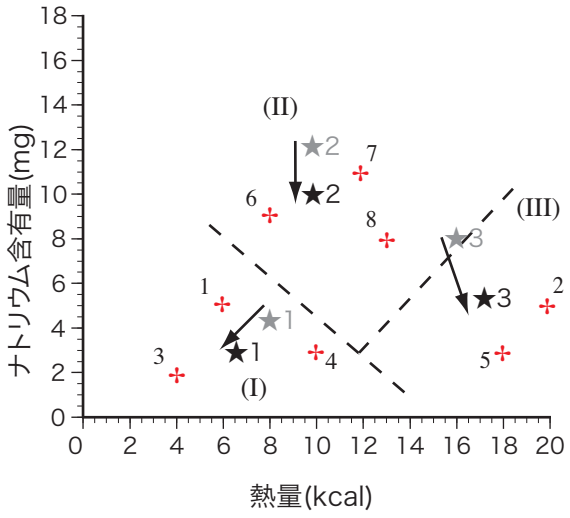

図 4: 重心の移動

天気図の分類

今回の講義では，ある期間の毎日の天気図を，似た天気図をまとめることで「代表的な天気図のパターン」に分類し，天候の違いが「代表的な天気図のパターン」とどう関係あるかを調べた研究を紹介します．［講義では，研究論文[2]から取り出した図を配付しました．学外公開用の講義録では無断複製できませんので，この講義録には図 5-7，表 3 は収録されていません．上の論文を参照しながら以下をお読みください．］．

1993, 94, 95 年の夏は，極端な天候が続きました．1993 年は極端な冷夏で大冷害が発生し，米の緊急輸入が行われました．94 年は正反対で，全国各地で猛烈な暑さとなりました．さらに 95 年は，北日本は冷夏，西日本は猛暑と地域格差が大きな夏となりました．そこで，これらの各夏の，毎日の天気図をクラスター分析によって分類し，「代表的なパターン」をとりだすことによって，天気図と気候の関係を調べることにしました．

クラスター分析を用いるには，天気図を表 1 や表 2 のような数字の組にする必要があります．そこで，図 5［上記論文の図 2］のように天気図上に 40 地点の位置を定め，この 40 地点での気圧の天気図全体の平均からのへだたりを求めます．この 40 個の変量の組をクラスター分析の対象とします．

このままでは，40 次元の散布図上でのクラスター分析をする必要があります．しかし，各天気図間であまり変化しないような変量を調べても，どの天気図でもその変量はほとんど同じなわけですから，分類にはほとんど無関係です．そこで，40 個の変量を組み合わせて，各天気図間でなるべく大きな変化をするような（分散が大きな）新しい変量をつくります．この手法を主成分分析[3]といい，この研究ではそのような新しい変量を 8 個作っています．

[2]村木千恵, 大瀧慈, 水田正弘,「主要点解析法による極東夏期天気図の分類」, 応用統計学, **27,** 1, 17–31 (1998).
[3]「統計データ解析 B」でとりあげます．

図 6 ［上記論文の図 7 (a)］は，地上天気図についての新たな 8 個の変量を *k*-means 法によって 5 つのクラスターに分類し，各クラスターの重心に対応する代表的天気図パターンを表したものです．クラスターの数を決定するには，クラスター数を 1 から順に増やしてゆき，重心に対応する天気図のパターンが大きく変化しなくなった時点でのクラスター数を採用しています．同様に，図 7 ［上記論文の図 7 (b)］は 500hPa 高層天気図を 4 つのクラスターに分類したものです．表 3 ［上記論文の表 3］は，各年の 7 月，8 月について，各パターンに分類される天気図が現れた日数を求めたものです．とくに高層天気図について，各年で顕著な違いが現れており，代表的天気図パターンと気候とに明確な関連があることがわかります．