

第 1 章

多変量解析のマップ

この章では、資料のタイプ（量的データ、質的データなど）から解説を始めます。そのあと、

資料のタイプ別（量的データ、質的データ、
量的データと質的データの混合）

分析の目的別（予測、判別、要約）

に分けて、分析法とそのあらましについて、適に具体例を挙げながら解説していきます。

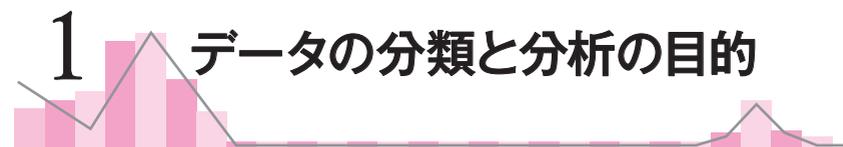
この章で、多変量解析の全体像をつかんでほしいと思います。

目的変数 説明変数	量的データ (予測)	質的データ (判別)
量的データ	単回帰分析 重回帰分析	判別分析 ロジスティック分析
質的データ	数量化 I 類	数量化 II 類

データの要約

主成分分析、因子分析、数量化 III 類

1 データの分類と分析の目的



この章では、まず多変量解析で扱う資料の具体例を示します。次に、その資料から、どういう視点で何を導き出すかを実例に沿って述べてみましょう。

多変量の資料を持っているが、その資料をどう分析してよいかわからないという人は多くいると思います。この章を読むことで、資料に見合った多変量解析の仕方をつかんでください。

まずは、多変量とそうでない資料の違いについて。

次の資料は、AからFまでの6人の算数の小テストの結果を表しています。

	A	B	C	D	E	F
算数	4	7	3	2	8	6

1人に対して、1つの点数しかないので多変量の資料ではありません。「算数の小テストの結果」のように、資料の観測項目や調査項目を**変数(変数)**といいます。上の表に国語の小テストの結果を加えた資料、

	A	B	C	D	E	F
算数	4	7	3	2	8	6
国語	3	8	7	8	3	5

は、調査項目、つまり変数が2個になるので、多変量の資料です。このように変数が複数ある資料を**多変量の資料**といいます。

表から、例えばBは算数が7点、国語が8点です。このような（人、算数の点数、国語の点数）の組が資料の構成要素になっています。このよう

な構成要素のことを**個体**といいます。上の表は、2変量からなる個体が6個集まって資料を構成しています。個体の個数を資料のサイズといいます。上の資料の**サイズ**は6です。

AからFの6人について、さらに詳しい多変量の資料があります。

各人について、身長、体重、性別、算数、国語、理科、社会、勉強好き、試験の可否を表にまとめてあります。この資料は、9個の変数を持つ資料です。ちなみにこの可否とは、算数、国語、理科、社会の小テストのしばらく後に行なわれた、入学試験の可否のことで、算数、国語、理科、社会の小テストの結果から直接判定できるものではないとします。

	身長	体重	性別	算数	国語	理科	社会	勉強好き	試験可否
A	162	58	男	4	3	4	8	好き	合格
B	170	70	男	7	8	6	3	嫌い	不合格
C	185	75	女	3	7	5	8	大嫌い	不合格
D	158	45	女	2	8	3	5	大好き	合格
E	165	52	女	8	3	7	4	嫌い	不合格
F	171	87	男	6	5	7	6	好き	合格

ここで変数の種類について説明しておきましょう。

変数には大きく分けて、量的データと質的データがあります。

数値で表されたデータを**量的データ**といい、そうでないデータを質的データといいます。

上では、身長、体重、算数、国語、理科、社会が量的データで、性別、勉強好き、可否が**質的データ**です。

量的データの中でも、長さや重さなどのように絶対的な0を持つような尺度を**比率尺度**といいます。質量、長さ、時間など物理的な量の多くは比率尺度です。一方、温度（摂氏or華氏）のように指数の差が意味を持つものを**間隔尺度**といいます。

質的データで扱う変数のことを**アイテム**や**項目**といいます。性別、勉強好き、合否はアイテムです。アイテムの欄の中に書き込まれた数字や記号のことを**カテゴリー**といいます。「男」、「大好き」、「合格」はカテゴリーです。

質的データは数値では表されませんが、決め事によって数値で表すこともできます。

例えば、性別の欄では男性を1、女性を0で、合否の欄では合格を1、不合格を0で表すことにします。

また、勉強好きの欄では、

大嫌い→0

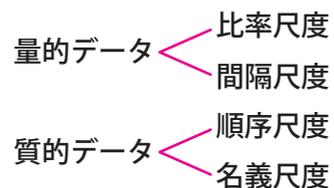
嫌い→1

好きでも嫌いでもない→2 好き→3 大好き→4

と表すことにします。すると上の資料はすべて数値で表すことができます。それでも、質的データを表す数値それ自体に意味はありませんから、性別、勉強好き、合否が質的データであることに変わりはありません。質的データを表す数値は区別のために与えられた単なる記号であって、性別の欄では男性を1、女性を2という決まりで書きこんでも構わないわけです。

ただ、「勉強好き」の数値に関しては、少し数学的に意味があるといえるでしょう。「勉強好き」の数値が大きければ大きいほど、勉強が好きであることを示しているからです。このような尺度を**順序尺度**といいます。一方、性別、合否のように内容を区別するためだけに用いられる尺度を**名義尺度**といいます。

まとめると次のようになります。



この表では上にいけばいくほど、数字が表す数値としての意味が強くなっていることがわかるでしょう。

分類自体にあまりナーバスになる必要はありません。

温度も、摂氏は水の融点（氷が解ける温度）を0℃、沸点（水が沸騰する温度）を100℃として等間隔に目盛を振ったものであることから間隔尺度と考えられますが、摂氏に273℃を足せば絶対温度となり、ボイル・シャルルの法則を満たすことなどから比率尺度であるとも考えられます。同じ温度で2つの尺度となるのもおかしな話です。

比率尺度、間隔尺度、順序尺度、名義尺度の順に数値の意味が強くなり、その順に数学の理論に乗せやすくなるということを捉えておけばよいのです。

この表を用いて、多変量解析における問題意識・目的を解説していきます。

多変量解析の目的は大きく分けて、

予測、判別、要約

の3つです。それぞれの場合にどのような分析方法を用いるかをおおざっぱに説明していきましょう。

その前に、予測と判別という用語の違いについてコメントしておきます。

算数の点数から理科の点数を予想するのが**予測**、算数の点数から合否を予想するのが**判別**です。予測の場合は数値を予想し、判別の場合はカテゴリーを予想します。つまり、量的データを予想することを予測、質的データを予想することを判別、と多変量解析では使い分けています。