焦点／研究の基本的諸側面

＜解　説＞

変数の分類と統計的分析

高　木　廣　文

看　護　研　究

第17巻　第4号　別刷

1984年10月15日　発行

医学書院

＜解　説＞

変数の分類と統計的分析

高木　廣文*

1. はじめに

大型電子計算機の進歩に伴い，多くの研究分野でデータ解析のために統計学のかなり高度な手法が用いられるようになった。現在では，マイコン・パソコンなどのソフトの開発などにより，全く統計学を知らなくても，データをコンピュータに入力するだけで，手軽に豊富な出力結果を得ることさえできるようになった。一面では便利になったものの，多くの手法のうち，どの分析法を使用すれば，自分のもつデータや研究目的に適したものか迷うこともあるだろう。また，出力された結果の解釈も，数字に不慣れな者にとっては，きわめて困難な場合もあるだろう。

適用すべき分析方法の選択や結果の解釈は，自分のもつデータの性質やその限界さえ理解していれば，大きく誤ることは少ないものである。「データを理解する」ということは，研究者にとって当然のことであるが，これが意外に難しいことも事実である。

また，どの分析方法を用いるべきかは，データを収集する以前に考慮されてしかるべきものであろう。そのためにも，自分がどのようなデータを調査対象から得ようとしているのかを，研究の計画段階でよく吟味すべきであろう。

ここでは，上記のような視点から，特に「変数」のもつ意味とその測定方法による分類，および研究目的による分類，また実際のデータのもつ特性による使用できる方法などについて説明することにしよう。ただし，各手法についての説明はここでは行なわないので，それらの点に関しては適当な成書[1,2,3]を参考にしてほしい。

2. 変数について

血圧は人により，また同一人であっても日により時間により，微妙にその値が異なるものである。血圧を知るには血圧計によって，対象となるものの測定を行なう必要がある。我々は対象の血圧については，測定が終了するまで知ることはできない。いいかえると「血圧」というものは，測定により初めてその値が確定するものである。このように，その値が対象や各測定により一定ではなく，その都度変化するものは「変数」とよばれる。

サイコロの目は1, 2, ····, 6の6個あることを知っていても，サイコロ投げをした場合，どの目が出るかは，やってみるまで判らない。すなわち，サイコロの出目を観察することにより，初めてその値が確定する。したがって，「サイコロ投げの出目」は変数である。ところで，正しく作られたサイコロならば，すべての目の出る可能性は等しいことが予想される。統計学では，その可能性は「確率」とよばれている。したがって，サイコロの出目は1, 2, ····, 6についてすべて等しい確率を

* 聖路加看護大学

もち，その値は1/6となる。このように，変数の取りうる値に対して特定の確率が付与されている場合，変数は「確率変数」とよばれる。通常の統計的手法では，変数といえば確率変数を指す場合が多い。

ところで，変数の各値が取りうる確率の状態は，変数の「分布」とよばれる。一般に，統計学では「正規分布」を変数の分布として仮定することが多い。また，データから特定の目的のために計算される量(平均値，分散など)は「統計量」とよばれる。統計量も1つの変数であり，データが母集団から無作為に抽出された場合，その計算方法や目的に応じてt分布，カイ二乗分布，F分布などを仮定し，統計的仮説検定を行なうことが多い。

3. 変数の測定尺度による分類

変数は，その値がどのような尺度によって測定されているかによって，大きく2分類される。すなわち，「量的変数」と「質的変数」の2つである。

量的変数は「定量的変数」ともよばれるもので，変数が血圧値や体重のように，ある「物差し」によって測定されているものである。測定された数値には意味があり，大小の比較ができる。また，対象についての測定値は「計量データ」ともよばれることがある。

計量データは，その数値のもつ特性により，さらに2分類して考えることができる。

仮に，A子さんの収縮期血圧の測定値が120 mmHg，K子さんが160 mmHgである場合，K子さんはA子さんに比べ40 mmHg血圧が高いということができるし，その比をとって4/3倍であるともいえる。このように，数値間の差のみでなく，比にも意味がある場合，「比尺度」による変数とよばれる。比尺度は測定が絶対零点(原点)をもつ場合である。

これに対して，特定の科目の学力を調べるために行なう試験成績はどうであろうか。統計学の試験で，A子さんは20点，K子さんは80点であったとしよう。2人の得点差は60点であるとはいえるが，K子さんはA子さんより4倍の学力があるとはいえない。これは，試験の成績が仮に0点であっても，それが直ちに統計学の学力が0であることを示すものではないからである。このことは，0という得点が必ずしも測定の原点ではないことによる結果である。このように，数値の差は意味があるが，その比に意味のないものは，「間隔尺度」による変数とよばれる。

質的変数は「定性的変数」ともよばれ，血液型やものの好き嫌いなどのように，ある特性に対する反応の有無によって与えられるものである。血液型の場合，A，B，O，ABの4つのうち1つに対して対象のデータが与えられる。このような場合，A子さんがA型，K子さんがAB型であっても，2人のデータの差をとって，

$$AB-A=A\ (B-1)$$

とか，その比をとって，

$$AB/A=B$$

などとすることができないことは明らかであろう。このような変数は「名義尺度」による変数とよばれる。また，変数とよぶかわりに「項目」，そして血液型のA，B，O，ABなどの細目を「カテゴリ」とよぶことがある。そのため，質的変数のデータは，カテゴリカルデータともよばれることがある。さらに，計量データに対して「非計量データ」もしくは「計数データ」ともよばれる。ところで，飲酒や喫煙に関して，質問紙などを用いて調査を行なうことがある。そのような場合，カテゴリとして「1. 好き」，「2. 普通」，「3. 嫌い」などとして，どれか1つに丸をつけさせることが多い。そのようにして得られたデータは，単にカテゴリの番号を示すだけなので，その差や比をとることはできないが，好みに関しては大小関係，すなわち順序を付けることができる。このような場合「序数尺度」，もしくは「順序尺度」による変数とよばれることがある。

上記の変数の分類を図1に示した。このように，変数はその特性により分類することができるが，これは単なる区分ではなく，この分類に応じて使用できる統計的方法も限定されてくるのである。それについては後述するが，研究を行なう場合には，常に自己のもつデータがどのような特性をもつものかを考えておく必要がある。なお，序数尺度によるデータを，目的によっては間隔尺度によるものとして分析を行なうことがよくある。

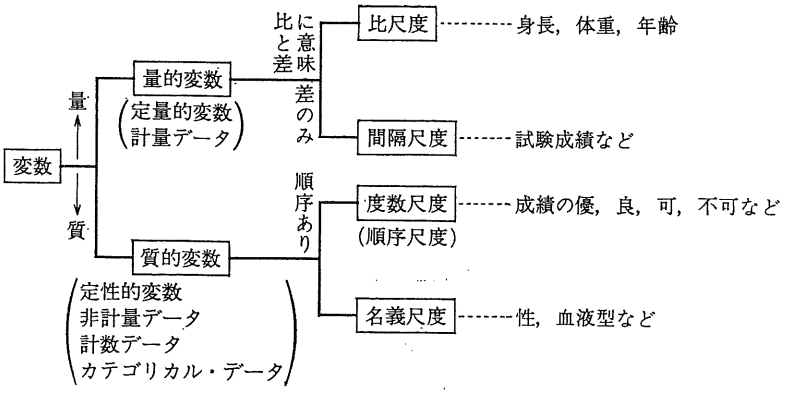

図1　変数の測定尺度による分類

4. 基準変数と説明変数

ある変数の数値の動き(変動)が他の変数の変動と関係する場合，一方の変数の値を用いることで，他方の変数の数値を予測することができる。そのような場合，予測のために用いる変数は「説明変数」，予測すべき変数は「基準変数」とよばれる。

数理統計学では基準変数，説明変数のかわりに「従属変数」，「独立変数」とよぶことが多く，社会学や経済学などでは「内生変数」「外生変数」などとよぶことがある。また，基準変数は「外的基準」，説明変数は「内的基準」ともよばれる。このように，呼称が異なっても，実際には同一の内容を示している。その用語の関係を表1に示した。

表1　変数の目的別分類と呼称

予測・識別の基準となる変数	予測・識別をするために用いる変数
基準変数	説明変数
従属変数	独立変数
内生変数	外生変数
外的基準	内的基準

分析方法によっては，回帰分析のように，基準変数と説明変数の両方が存在する場合と，基準変数を特別に設定しないで，変数間の相関関係について分析をする場合がある。これは，研究目的により異なるものであるが，基準変数のない場合に，ある場合のような分析をすることは不可能である。これは，あたりまえのことであるが，目的に応じた変数の設定が研究においては重要である。

基準変数と説明変数が量的変数の場合，予測という点から回帰分析を行なうのが一般的である。基準変数と説明変数が共に1つであれば単回帰分析，説明変数が多数あれば重回帰分析を行なうことになる。また，基準変数も多数あれば正準相関分析を用いることができる。同様に，説明変数が質的変数として与えられていれば，数量化理論Ⅰ類を用いることもできる。

基準変数が質的変数の場合，説明変数が1つで量的変数であれば，平均値の差の検定(t検定)や分散分析(F検定)などを行なうことができる。また，説明変数が1つで質的変数であれば，一様性の検定(カイ二乗検定)を行なうことができる。

説明変数が量的変数で多数ある場合，識別問題として判別分析を用いることができる。また，説明変数が多数の質的変数からなる場合，数量化理論Ⅱ類などを用いることができる。なおこのような場合，基準変数はグループや病気の診断名などを示すカテゴリからなるものである。

基準変数がない場合，変数間の関係(構造)を調べたり，いくつかの変数を合成して新しい変数を作り，特定の目的のための指標となる尺度とすることもある。量的変数の場合，各変数間の相関係数などをもとに，主成分分析や因子分析などを用いることが多い。また，質的変数の場合には，2変数間の関連を調べるには，クロス表などをもとにして，独立性の検定(カイ二乗検定)や関連係数などを求めることが多い。さらに，変数が多数あれば，数量化理論Ⅲ類などを用いればよい。

上記のように，基準変数の有無と各変数の測定

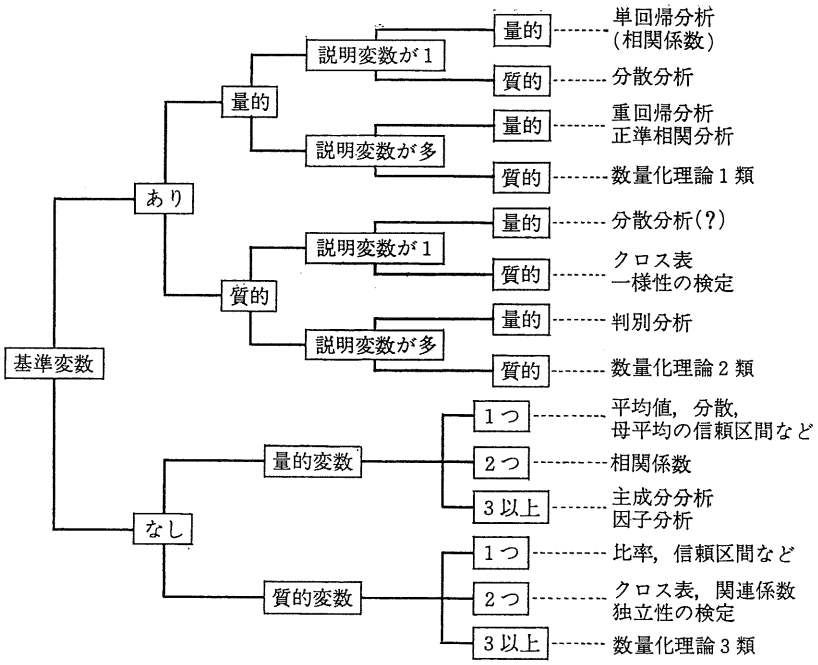

図2　変数の特性による分析方法の分類

尺度の質により，どのような分析方法を用いればよいかは，ほぼ決まってくるともいえよう。図2にその関係を示したが，すべての方法を図中に示しているわけではないことに注意してほしい。

次に，実際の研究における変数の設定と測定の問題，さらに分析結果の解釈などにおける問題点などについても，順に触れていくことにしよう。

5．変数の設定・測定と分析方法

高年齢者の循環器系疾患の予防のため，「収縮期血圧」と「塩分摂取量」の関係を調べ，減塩指導により，どのくらい「収縮期血圧」が低下するかを研究するものとしよう。対象者の年齢や他の疾患の既往などの影響は考慮すべき問題であるが，ここでは話を簡単にするために，上記の2変数のみについて考えよう。

収縮期血圧は血圧計によって簡単に測定できるが，データとしてどのように保存するかは，その後の分析方法との関係で時に問題となる。測定値をそのままの数値でデータとして収集するのならば，特に問題となるようなことはない。しかし，時として収縮期血圧値を適当なカテゴリにより区分してしまうことがある。すなわち，140 mmHg未満を「1．正常」，140-160 mmHg未満を「2．境界型」，160 mmHg以上を「3．高血圧症」などとすることがある。このような場合，測定が比尺度で行なわれたのにもかかわらず，データとしては序数尺度によるものとなってしまう。このようなカテゴリ化を便利なものと考えることがあるかもしれないが，後述するように，実際の分析では，必ずしもよいものとはいえない場合が多い。

同様に，塩分摂取量についても，各対象に対して食物調査などにより，摂取量を量的に測定したものを，わざわざカテゴリ化することさえある。

このようなカテゴリ化は，データの収集段階では避けるべきものである。なぜならば，量的なものから質的なものへの変換は常に可能であるが，その逆は一般に不可能だからである。わかりやすくいえば，かりにA子さんの収縮期血圧値が120 mmHgである場合，それを1としてデータに記入したとすれば，もとの値がその後の分析で必要になっても，決して1という値からは導くことができないだろう。しかし，120という値から1というカテゴリを得ることは簡単にできる。

最も重要なことは，質的変数に適用できる方法はすべて量的変数にも適用できるが，その逆は必

ずしもいえないということである。たとえば，量的変数では平均値や分散を求めることは常にできるが，質的変数では単に各カテゴリの人数を数えることができるにすぎない。さらに，より高度な分析方法にもこのことがいえるのである。結局，量的変数をわざわざ質的変数にかえてデータとすることは，研究者が自ら適用できる方法の範囲を狭めているとしかいえないことになるだろう。

2変数が量的変数の場合，収縮期血圧を基準変数に，塩分摂取量を説明変数として回帰分析を行なうことができる。それにより，2変数間の関係を回帰式として量的に把握することが可能である。また，その結果を用いて，塩分摂取量をどの程度にすれば，許容できる範囲内に血圧値を設定できるかを検討することもできるだろう。また，ピアソンの積率相関係数などを求め，その関係の強さを量的に示したり，検定もできる。

2変数が質的変数の場合，クロス表を作成し，関連の強さを調べることになる。すなわち，関連係数を求めたり，独立性の検定をカイ二乗検定などで行なうことになるだろう。しかしこの場合，2変数の関連の有無は調べられるが，どのような量に塩分摂取量を設定すればよいのか，という情報を得ることはできない。

それでは，収縮期血圧が量的変数で，塩分摂取量が質的変数である場合は，どのような方法が考えられるであろうか。このような場合には，一元配置分散分析を行なうことがよくあるだろう。すなわち，塩分摂取量の各カテゴリを因子の水準を示すものとして，分散分析を行なうものである。しかし，仮に結果が有意なものであったとしても，カテゴリの設定が主観的になされているので，減塩指導に用いるためには，やや説得力に欠くものと考えられる。

収縮期血圧が質的変数で，塩分摂取量が量的変数の場合にも，形式的には一元配置分散分析を用いることができる。しかしその場合，塩分摂取量を基準変数として扱っていることになる。仮に，分析の結果が有意であっても，その解釈に問題が生じるだろう。すなわち，普通ならば原因として塩分摂取を考えているのに，逆に血圧を原因にし，塩分摂取を結果としてみることになる。このような方法では，いったい何を調べているのか，よほど注意して結果を解釈しないと，大きな誤りを犯すことになるだろう。

このように，変数の測定値をデータとしてどのような尺度が適切なものかは，研究の進展に大きく影響するものである。本質的に質的なものでないかぎり，量的にデータを把握することが，データ収集における基本である。

6. 変数の示す内容と結果の解釈

収縮期血圧と塩分摂取量の関係を調べる場合，血圧の測定は比較的簡単であるが，塩分摂取量について食物調査により，量的にデータを把握することは，財源や人的資源の面で負担が多すぎるだろう。そのような場合，質問紙によっていくつかの項目に答えさせ，その得点の合計によって，塩分摂取に関する指標を作成することがある。すなわち，「減塩醤油を使っていますか」，「減塩味噌を使っていますか」などの質問に「1. 使わない 2. たまに使う 3. よく使う」のような選択枝から1つを選ばせ，それぞれに1，2，3の得点を与えるようにする。そのような質問を10項目ぐらいまとめて，総合点により，塩分摂取を示す1つの変数とする。実際には，各項目は序数尺度からなるものであるが，この場合には，あたかも間隔尺度によるもののように扱うことになる。しかし，多くの項目の合計点を間隔尺度として用いても，それほど大きく誤ることはないようである。ただし，項目やカテゴリの設定は，ある意味ではかなり主観的になされるので，因子分析などにより，設定した項目間に共通性が高いことを確認しておくことが必要である。なお，このような操作は尺度化とよばれるもので，塩分摂取に関するような指標は「塩分尺度」などとよばれることが多い。

上記のように，質問紙により塩分尺度のほかにも，「糖分尺度」や「脂肪分尺度」などを設定することができる。そのような尺度を説明変数に，血圧値のデータを基準変数として(重)回帰分析を行なうことができる。また，血圧値のデータが量的でなく，「1. 高血圧症患者」，「2. 健常者」のようなグループを示す質的変数を基準変数に用いるのであれば，平均値の差の検定，分散分析，さらに

判別分析なども用いることができる。

それでは仮に，収縮期血圧と塩分尺度に有意な関係(相関)が認められた場合，それはどのようなことを示すものと考えればよいのだろうか。「塩分摂取(量)は収縮期血圧に影響を及ぼすことが検証された」とするのは，明らかに間違いである。「検証」というのはいいすぎであるし，最も重要な点は「塩分尺度」は必ずしも対象者の「塩分摂取量」を正しく示すわけではないということである。この辺りのことを厳密に考えるものはあまり多くはないが，質問紙によって測られたものは，常に対象となるものの「意識」を反映するものであることを考えねばならない。すなわち，意識としては塩分を少ししか摂取していないと本人が考えていても，実際にはきわめて多量に摂取している場合もありうるし，その逆の場合もある。そのようなことから，仮に変数間に有意な関係が認められても，安易に結論を急ぐべきではないし，有意でなくとも，それは単に質問紙の作り方が悪かっただけなのかもしれない。この場合では，設定した塩分尺度により，収縮期血圧の予測がある程度可能であり，その点から，その尺度得点が高いものには，本人の意識として塩分摂取量を少なくするように指導するのに用いるのがよいだろう。何とも歯切れの悪い言い方であるが，実際の量を測定していないのだから，結果を過小評価するぐらいでちょうどよいのである。

実際の研究では，質問紙を用いて何かを調べることが多いと考えられるので，そのような場合の注意すべき点を若干述べてみた。重要な点は，研究者が設定した仮説と分析結果が一致するようにみえる場合，結論を急がずに，もう一度その結果の意味するところを再考することであるといえるだろう。

7. おわりに

これまで，簡単にではあるが変数のもつ意味や，測定の尺度による分類，分析目的による分類などについて述べてきた。また，測定尺度による分析方法の違い，結果の解釈などにも，少しではあるが触れてみた。多くの内容を短い紙面で説明しようとしたため，表現に不十分な点があると思うが，その意味するところだけは理解してもらえるものと思う。すなわち，研究者は自己のもつ変数（データ）に関して，十分な理解をもつべきであるという点である。変数の性質，測定尺度，意味する内容などを正確に把握していれば，分析方法の誤用や結果の解釈で大きく誤ることは少ないものである。

参考図書

1) 高木廣文：ナースのための統計学，医学書院，1984.
2) 柳井晴夫，高根芳雄：多変量解析法，朝倉書店，1977.
3) 豊川裕之，柳井晴夫編著：医学・保健学の例題による統計学，現代数学社，1982.