＜解説＞

統計処理で陥りやすい問題点

高　木　廣　文

看　護　研　究

第14巻　第3号　別刷

昭和56年7月15日　発行

医学書院

＜解説＞

統計処理で陥りやすい問題点

高木　廣文*

近年コンピュータの普及により，データを統計的に解析することが著しく容易になり，統計解析の理論を知らなくとも，各種複雑な分析が可能になってきた。それに伴い，ある程度の統計的な解析がなされていないと，科学的な研究とは認め難いというような風潮もあるように思う。そのため，特定の研究を行なう場合，得られたデータを統計的に処理することが大前提となり，このため一つの落し穴に研究者は陥りやすくなる。すなわち，研究者自身が自分の集めたデータの特性を良く把握しないうちに，統計的な解析を行なうということが起こりうる。たとえば，血圧やコレステロール値などの生理学的検査結果を，単に平均値や分散のみで示して満足することがよくある。しかし，血糖値などの測定値をヒストグラムなどで図示し，その分布をながめることも，実はきわめて重要なことなのである。このような手法は記述統計学 descriptive statistics とよばれているが，実は研究の初期の段階ではきわめて有用であり，カルテや調査票からの誤記入やデータのパンチミスなどを発見するためにも，必ず行なうべきものである。また，研究対象となった患者群の職業，年齢構成，性，居住地，病院への来院理由などは，研究の基礎資料として把握されている必要がある。これらの点が押さえられていれば，その後の推測統計学 inductive statistics の手法を用いる際の助けにもなるし，結果の解釈でも大きな誤りを犯すこともなくなるはずである。

上述のように，実際に統計的分析を行なう場合，知らずに陥っているかもしれない問題点について，以下にいくつか述べてみたい。

1. 母集団と標本

ある地域や特定の疾患をもつ人々を調査しようとする場合，対象となる集団を母集団 population とよぶ。通常は対象となる人口が多くなるに従い，全員を調査することはできなくなり，母集団の中の一部のものを調査対象にする。このように，実際に対象として選定されたものを標本 sample という。さて，統計的解析を行なう場合，標本がどのように抽出されたかが第一に問題となる。標本抽出法は各種あるが，基本的にはすべて完全に無作為 random であることを要求する。言葉をかえれば，人為的でなく，全く「でたらめ」にとることが前提となっているともいえる。仮に研究者のもっているデータが無作為抽出によって得られた標本によるものでなければ，その後の解析は手順どおりに行なったとしても，何の意味もないただの数字の羅列を得るだけであるといっても過言ではない。しかし，実際問題としては，それほど完全な無作為抽出が行なえないことが，医学や看護学の領域では多いはずである。

特定の地域住民の健康調査などを行なう場合，

* 聖路加看護大学

住民台帳などをもとにして，調査研究計画の段階で乱数表などにより，標本を無作為に選ぶことは可能である。実際の調査の場で，不在や調査拒否などによる欠落がある程度は起こりうるものの，それが極端に高率でなければ，ほぼ無作為標本を得たものとして扱ってかまわないだろう。これが特定の疾患を有するような患者群を研究対象とする場合は，どのようになるであろうか。たとえば，糖尿病患者の合併症と治療薬との関係を研究するような場合を考えてみよう。母集団は日本中の糖尿病患者ということになる（世界中でもよいが)。このような中から500例なり1000例なりの患者を無作為に選ぶとしたら大変なことになろう。そもそも全患者を正確に把握することが大変な作業である。これには日本中の医療機関の協力が必要であり，不可能ではないにしてもその労力は膨大なものとなろう。このような協力体制が得られ，かつ無作為に標本となる患者のデータが得られた場合のみ，統計的解析が意味のあるものとなる可能性をもつということは特に注意すべきであろう。すなわち，通常行なわれているように，特定の医療機関を受診した患者だけを扱った研究は，何かしら上述の点を満たしていないものと考えるべきであろう。それではこのような特定の医療機関のみの研究は全く無意味かというと，一概にそうともいいきれない。

重要な点は，研究者が扱っているデータが，母集団の各種特性を十分に代表するものであることが保証されているか否かである。基本的な点として，男女別の構成比や年齢構成が母集団に比べ著しい偏りがないか，また職業の分布が特定の職種に偏っていないかなどが吟味されねばならない。もしも母集団における各種の基本的特性が未知の場合には，このような点を比較することは不可能であるから，単にこれらの点については記述的方法によって提示するだけでも，疫学的には価値のあるものとなろう。なぜならば，同様の報告が増加するにつれ，各結果を相互に比較することにより，対象とする母集団の特性を全く正確とはいかないまでも，その輪郭を浮かび上がらせることは可能かもしれないからである。

標本の代表性については，このほかに地域分布に関するものがある。近隣の住民のみが受診する医療機関なのか，そうではなく遠距離の人でもわざわざ来るのかなど，各医療施設によって異なっているはずであり，特定の地域を代表することができるか否かなどを，研究の初期の段階で慎重に考慮すべきである。

結局，研究者が自分のデータについて熟知していることを要求されるということである。統計的な処理により，ある有意な結果が得られたからといって，それをすぐ鵜飲みにすべきではない。重要な結果であればあるほど，自己のデータに対する客観的な見方が要求されてくるのである。

患者-対照研究 case-control study などの場合，上述のことは対照群の選定の場合にもあてはまるのはいうまでもないことであろう。信頼性の高い，代表性のあるデータさえ得られれば，すでに研究の90%は終了したといっても過言ではないだろう。

2. 分布の代表値

Average という言葉がある。日常的には平均値という意味で用いられる場合が多いように思う。しかし，統計学では「代表値」という意味で用いられる。何に対しての代表かというと，ある変数の分布を代表するものという意味であり，この値のまわりの人数なり確率なりが最も大きいということも含んでいる。代表値は分布上の特定の変数の値を示すために，位置 location の尺度ともよばれ，平均値 mean，中央値 median，最頻値 mode の三つが重要である。このうち一般には平均値が最もよく使用されている。この理由としては，多くの統計的手法が平均値と分散 variance に基づいていることにもよる。通常は変数の分布，もしくは誤差の分布として図1に示すような一峰性（山が一つ)で平均値に関して対称な正規分布 normal distribution を母集団の分布として仮定している。このため，平均値，中央値，最頻値の三者の値は完全に一致し，統計的処理のしやすい平均値を用いることは意味がある。しかし，生理学的なデータ（血圧値や血糖値など）は正規分布というよりも，一般に右側の裾野を長く引いた分布になることが知られている。図2にこのような分布形を示

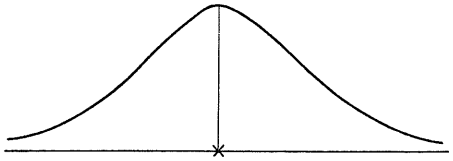

図 1 正規分布（対称な分布）

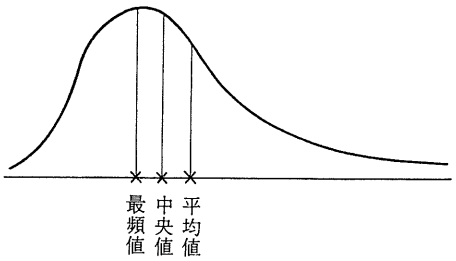

図 2 右裾の長い分布

したが，この例としては対数正規分布がある。個人所得の分布などが対数正規分布に従うことが知られている。このような分布は下限が決まっているが，上限は下限に比べ限りがないといえるような場合に，よくみられるものである。たとえば，所得の場合下限は0に決まっているが，上限は数十億となってもかまわない。このような場合，平均値は最も多くの人の所得を示す最頻値に比べ，より右側の高いほうに傾くことになる。このようなときに，平均値を分布の代表値として用いるのは問題があり，最頻値か中央値を用いたほうがより実態を示すことになる。例として11人の収縮期血圧を調べたとしよう。測定結果は（110, 112, 112, 115, 120, 120, 125, 125, 130, 135, 190）であったとする。この平均値を計算すると126.7となり，11の内8つの測定値よりも高く，右に偏っていることがわかる。中央値は120であり，数値上の差はあまりないが，より実情に近い値となっている。実際のデータ分析においては，190という測定値が問題となる。これは唯一他の値からとび離れたもので，はずれ値 outlier とよばれるものであるが[1]，これをそのまま含めて分析するか，もしくは取り除くかが問題となる。一般的に，他の値に比べ極端に高い数値をとることで，何らかの異常を疑うべきであり，他の標本とは異質のものとして，以後の解析からはずすことを考慮すべきであろう。実際，190という測定値を除外して計算すると平均値は120.4となり，中央値120とほぼ一致し，どちらを用いても大差のないことになる。

はずれ値を分析に含めるか否かは研究者の目的にもよるが，その存在を知らずに統計的処理を行なうと，結果の解釈で思いもよらない誤りを犯す可能性がある。この点からも，また特性の分布を知るという意味でも，ヒストグラム等をまず描いてみるということがきわめて重要なのである。もしも分布の右裾が長く，対数正規分布があてはまるような場合には，測定値のまま分析に用いるのではなく，対数をとるなどの変換を行なうことを考えてもよいだろう[1]。また図3に示すような，山が二つあるような分布が得られるかもしれない。一般にこのような二峰性を示す分布は，異質な集団を知らずに一まとめにした場合などに起こりうるもので，適当な基準により集団を二つのグループに分割し，別々に解析を行なうほうがよい。これは集団分割とよばれる方法で，たとえば，特定の疾患について高血圧群と正常群に二分して統計的分析を行なうなどが考えられる。

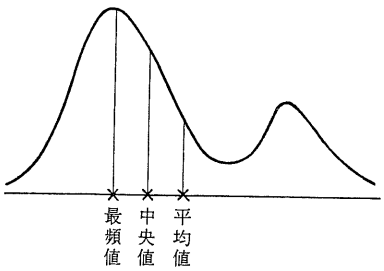

図 3 二峰性分布

上述のように，単に平均値のみを計算し，これを分布の代表として用いることが，ときにはいかに誤った解釈をもたらすかが理解されたと思う。統計的処理はまず一つ一つの変数の分布を描くことから始まるということを理解すべきなのである[1]。

3. 標本数の問題

表1のデータについて考えてみよう。二つのグループの空腹時血糖値の平均値と標準偏差を示したもので，A群とB群の標本数はともに30例である。いま二群の平均値の差を検定することにしよ

表 1　A群とB群の空腹時血糖値

	標本数	平均値	標準偏差
A　群	30	118mg/dl	31mg/dl
B　群	30	131mg/dl	33mg/dl

$t=1.546$

う。通常の t 検定では等分散性（母集団での母分散が等しい）を仮定するので，まず等分散の検定を行なう。これは $F=1.13$ となり，自由度(29,29)の F 分布の上側2.5%値（2.10）と比べると小さく，かつその逆数よりも大きいので，等分散を仮定することができる。さて，t 検定に用いる式は，二群の標本数を m, n, 平均値を $\bar{x}$, $\bar{y}$, 標本分散を S_x^2, S_y^2 とすると，

$$t=\frac{|\bar{x}-\bar{y}|}{\sqrt{\frac{m\,S_x^2+n\,S_y^2}{m+n-2}\cdot\frac{m+n}{mn}}}$$

となり，自由度（$m+n-2$）の t 分布に t が従うことで検定できる。実際に計算すると $t=1.546$ となり，自由度58の上側2.5%の t 値2.01より小さいので，危険率5%で有意差なしとなる。この場合，A群とB群で空腹時血糖値の平均値に差がないという結論になるだろう。ところで，この検定では両側検定を用いたが，これは両群の平均値について，一方が他方に比べ，より高いとか低いとかの事前の情報がないためである。もしB群のほうが何か特別の疾患を有していたり，特定の処理を行なった群であり，そのためB群での平均値が高いことが事前に期待されるものならば，片側検定でよいことになる（しかしもしB群の平均値が低ければ検定することに何の意味もない）。表1がそのような状況での結果だとすると，上側5%の t 値1.672と比べるべきであろう。しかし，計算結果はこの値よりも小さいので，この場合も有意ではない。

どちらの場合でも，両群の平均値に差は認められないものとして，研究は終わるわけであるが，実際の平均値の差13 mg/dl は無視してもよいであろうか。かりにこの差が母集団での本当の差を示しているものなら，研究者は誤った結論を下したことになる。

それではこの結果が倍の標本数，つまり各群60例ずつのものであったとすればどうなるであろうか。計算すると，$t=2.205$ となり，これは自由度118の上側2.5%点の値1.98より大きく，有意差ありと判定される。この場合，研究者はA群とB群の空腹時血糖値には有意な差があるものと結論するだろう。

表 2　特性Aの有無とある疾患の発生の有無

特性A ＼ 発症	あ り	な し	計
あ り	6 (a)	24 (c)	30
な し	2 (b)	28 (d)	30
計	8	52	60

$p=0.127$

このように標本数の影響は重大である。ほんのわずかの差を統計的にみつけだそうとすれば，標本数はそれだけ大きくとる必要がある。一般的にいえば，少数例の研究でみつけられるような大きな差はそれほどないと考えるべきであり，標本数は多いにこしたことはないのである。

つぎに表2のような場合について考えてみよう。これはAという特性の有無によって，各群30例ずつ抽出し，ある疾患の発症状況を調べた結果である。特性Aをもつ者は，Aをもたない者に比べ，疾患を発症する危険が高いものとして事前に考えられていたとする。見込比 odds ratio(ad/bc で推定）は3.5であり，この予想が正しいことがわかった。しかし，これに対して Fisher の直接確率 exact test を計算すると，確率 $p=0.127$ (片側検定）となり，有意とはならない（連続性の補正済み χ^2 値は1.298)。結局，この場合でも研究者は見込比3.5という危険因子をみつけたにもかかわらず，結果を意味のないものとして棄てざるをえないことになろう。しかし，表3のように

表 3　標本数を二倍にした場合

特性A ＼ 発症	あ り	な し	計
あ り	12	48	60
な し	4	56	60
計	16	104	120

$p=0.029$

その標本数が各々60であった場合には，直接確率は0.029となり，有意なものとして特性Aを危険因子と同定できる。このように，標本数の影響を軽くみることはできない。標本数を決めるための方法もあり，その点については他の参考文献をみていただきたい[2,3]。

上記二例でみたように，標本数は多くとったほうが，統計的に有意な差をみつけやすいことが理解されたと思う。しかし，各種の条件により，少数の標本しか得られないような状況にある場合は，どのようにしたらよいであろうか。一つには得られた結果がかりに統計的に有意でなくとも，それはそれとして結果を提示し，さらに有意差を発見できる可能性のある標本数を計算することも重要かもしれない。かりに数百の標本で可能ならば，実際にそれはやってみる価値はあるように思う。数万のレベルになると，これは実現不可能であろう。それにしても，そのような情報は，次回の研究を行なう際や，他の研究者にとっても有用なものとなるはずである。

4. 相関と因果

一般に統計的解析により，ある事象の因果関係を解明することは不可能であるといわれる。このことは，疫学的にある疾患の原因を追求したとしても，ある要因をその原因として同定することは不可能であることを意味している。これは統計解析が基本的には，調査に基づくデータを対象とし，実験のように各種条件をコントロールし，純粋に一つの要因と特定の結果との関係を検討できるものと異なるからである。しかし，統計的手法にはそのような限界があるにせよ，特定の因果関係について推論することはできるのである。注意すべき点は，どのような分析による結果でも，それは因果関係を立証するものではなく，示唆するものとして扱わねばならないことである。

通常，二つの変数の関係を表すためには，相関係数 correlation coefficient を用いる。いま二つの変数 x と y があるとしよう。x の増加につれ y も増加し，あるいは x の減少につれ y も減少するような場合，x と y には正の相関があるといい，逆に x の増加（減少）に伴い，y が減少（増加）すれば，x と y には負の相関があるという。このような関係が全くみられなければ，x と y は無相関であるといわれる。

二つの変数（要因）に因果関係がある場合，必ず相関関係があるはずであるが，相関関係があるからといって因果関係があるとは限らない。しかし，相関関係があることは最低条件として必要なわけであり，統計的方法では相関を基礎に置くので，以下に相関についての問題を若干述べたい。

相関係数は二つの変数 x と y の標準偏差を S_x, S_y とし，共分散を S_{xy} としたとき，

$$\gamma=\frac{S_{xy}}{S_x\,S_y}$$

によって求めることができる。相関係数 γ は1から－1までの値をとり，絶対値が1に近いほど二変数の相関は強く，0に近いほど弱いことになる。しかし，相関係数は二つの変数の直線的な関係を記述するものであるという点は注意が必要である。もしも，データを散布図上に表したとき，図4のような直線ではなく，曲線関係を示すような場合には，相関係数をそのまま用いるのは正しくない。データの値を測定値のまま用いるのではなく，大きさにより順位をつけ，その順位から相関係数を計算するか，変数の値の対数をとったり二乗したりする変換を行なうなどが必要となるかもしれない。このような曲線関係を生のデータから直接みつけることは，ほとんど不可能であり，やはり相関係数を計算する前に図4のような散布図を描き，特別なパターンが存在するかどうか調べるべきだろう[1]。

通常，相関係数の計算とともに，その有意性の

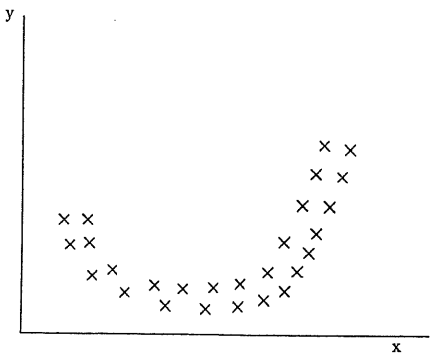

図4 曲線相関

検定を行なうわけであるが，これについての問題はないだろうか。相関係数の検定は，標本数を n としたとき，

$$t=|r|\sqrt{\frac{n-2}{1-r^2}}$$

が，自由度 $(n-2)$ の t 分布に従うことで行なわれる。たとえば，標本数が30である二変数の相関係数が0.1であったとすると，$t\fallingdotseq 0.532$ となり，これは明らかに有意でない。しかし，標本数が500の場合，同じ相関係数が得られれば，$t\fallingdotseq 2.243$ となり，これは5％の危険率で有意である。ここでも前述のように，標本数の多さが有意性に影響することがわかる。ところで，相関係数が有意であるという意味は何なのであろうか。これは母集団における相関が0ではないということを示すもので，実際の目的にもよるが，極端に小さな相関関係を発見しても，あまり意味がないともいえる。相関係数の二乗は一方の変数が他方の変数の変動を，どのくらい説明できるかという尺度になり，これを決定係数などともよぶ。相関係数が0.1の場合，その二乗は0.01になり，わずか1％の説明力があるにすぎない。このような点を考慮すると，ただ単に有意なだけでは，実質的な意義をもたない場合もあることが理解できると思う。相関係数の数値が大きいこともまた重要な点なのである。

ところで相関関係は用いるデータにより，直接相関と間接相関とに区別して考えねばならない。直接相関は個人別のデータによって得られるもので，血圧と体重の関係を調べたりする場合がそれであり，一方間接相関は県別の癌死亡率と食物消費量の関係をみる場合のように，データが個人ではなく集団を単位とするものから得た場合である。それゆえ，間接相関は直接相関に比べ，その結果を直接的に個人のレベルまで普遍化することができない。地域別のデータでは正の相関を示すにもかかわらず，特定の地域で個人別に調べたところ，負の相関を示すこともありうる[4,5]。これは図5に示したような関係である。

地域別のデータによって，変数間に相関関係が示唆された場合，最終的には個人を単位とする研究により，直接相関の存在を確認することが望ましい。しかし，間接相関が全く意味のないということではなく，それ自身研究上の指針となりうるので軽んじるべきではない。

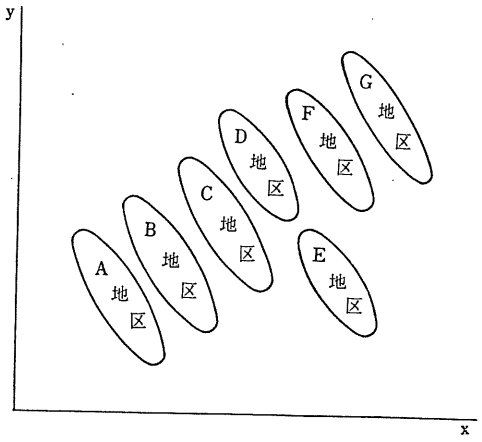

図5 間接相関と直接相関
各地区内では負相関であるが，地区別データでは正相関を示す。

一方，直接相関にもいくつかの問題が起こりうる。これは主にデータの取り方や分析の仕方によるものである。例として，高校生の数学と英語の成績について調べたとする。その結果，両者の成績にはほとんど相関がなく，弱い負の相関がみられただけであった。これは数学と英語の能力が互いに独立なことを示すわけであるが，常識的には変な気がするであろう。そこで，これを男女別に相関係数を計算し直したところ，両者ともに高い相関を示した。これは図6に示したような状態であり，数学は男生徒が一般に成績が良く，英語は女生徒で良いということを考慮していなかったことによることがわかった[5]。このように，ある特

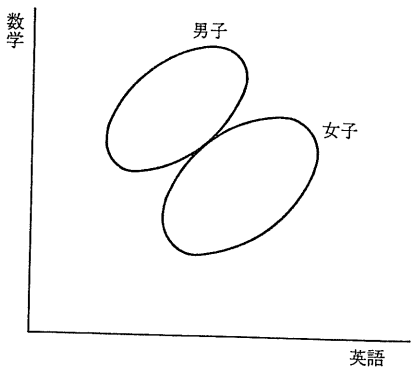

図6 層別化を要する例

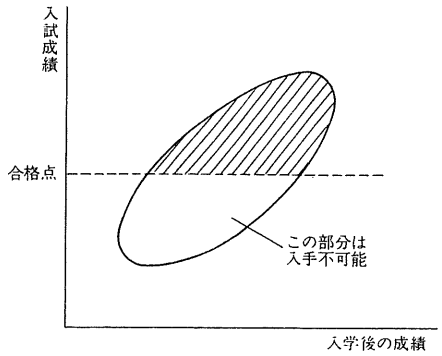

図 7　打切りによる影響

性による影響をさけるためには，あらかじめ，集団をその特性によって分け，各々について分析する必要がある。このような集団の分割の仕方を層別化 stratification とよぶ。

もう一つは，打切りデータ truncated data の問題である。入学試験の成績と入学後の成績の相関を考えてみよう。入学後の成績のデータは，入学試験の成績が合格点以上の者だけについて得られるのであり，図7で斜線で示した部分の相関をみることになる[5]。このため，一般に相関は小さなものとなることが予想される。そのため，かりに入試の成績と入学後の成績に相関があまりないからといって，入試は無用であるとはいえないことになる。もしも入試を行ない，成績と関係なく入学させ，その後の成績との関係を調べれば，高い相関が得られる可能性も十分にありうるのである。

このように，相関の有無を調べ，因果関係に迫まるには，いくつかの注意すべき点がある。この点からも，研究者が自分の所有するデータが，どのような性質のものなのかを熟知している必要がある。

5. おわりに

いくつかの問題について簡単に触れたが，実際に統計的な処理を行なう場合には，その場その場で何かしら問題となることが多いように思う。それらのすべてについて言及することは不可能であるが，基本的にはデータに対しての理解の深さが問題になるように思う。データは集めたが何をしてよいか悩んだり，どのような分析法を用いればよいか見当がつかないようでは，やはり困るのである。

統計学は dirty data からすばらしい結果を生みだす魔法の道具ではなく，データなりの結果を与えるだけであることを認識しておくことも必要かもしれない。

参考文献

1) Hartwig, F. & Dearing, B. E. 著；柳井晴夫，高木廣文訳：探索的データ解析の方法，朝倉書店，1981.
2) 柳井晴夫，高木廣文：臨床試験における標本数の定め方，医学のあゆみ，第102巻，第13号，881-886，1977.
3) Fleiss, J. L. 著；佐久間昭訳：計数データの統計学，東京大学出版会，1975.
4) 山本俊一，他：がんの疫学統計の方法に関する研究，厚生省「がんの疫学」班分担研究報告，1980.
5) 柳井晴夫，高根芳雄：多変量解析法，朝倉書店，1977.