東女医大誌　第78巻　第8・9号
頁　437～442　平成20年9月

総　説

医学研究における統計学（2）

医学研究における統計学的処理の実際

東京女子医科大学総合研究所研究部

清水（シミズ）　悟（サトル）

（受理　平成20年5月12日）

Statistics in Medical Research (2)
Present Status of Statistical Processing in Medical Studies

Satoru SHIMIZU

Department of Medical Research Institute, Tokyo Women's Medical University

We present an individual case of required statistical analysis in medical studies. We explain in detail, the method for creating a database and also describe how to produce the original data which serves as resource for research. Points concerning which care should be taken concerning the input of data are also described.

Reference is made to the statistical analysis design for statistical work, pointing out that the relation between the purpose variable and the explanation variable constitutes the design. The statistical analysis design should decide on the method for statistical analysis, and is also a method for the research purpose.

In addition, software for statistical analysis software has been described including the features and cost of the software and the required conditions when choosing statistical analysis software. Furthermore, the relation between statistical inspection and logic deployment such as the value of statistical inspection required for logic deployment of a scientific paper is described.

Key words: statistical analysis, medical studies

はじめに

本稿では著者らの経験した医学研究における統計処理の実際から医学研究に有効な統計処理法について解説する．まず「統計学とは？」という素朴な質問には「データの集約」という回答をするのが常であるが，本質的には与えられたデータから「何が言えるか？」と言うことに他ならない．以後，その点を底流に解説を進める．

わが国では5年ごとに国勢調査が行われているが，その調査結果でマスコミの話題に上るのは人口の増減である．元来，国勢調査は国家行政の基礎資料となるもので，明治5年に人口および諸々の調査が行われて以来，初期は不規則ではあったが大正9年から5年ごとに，国勢調査が行われ今日に至っている．人口は有病率，死亡率，罹患率等医学研究に必要な指標などを算出する場合に必要不可欠な基礎情報であるが，国勢調査が行われない年の人口は出生届，死亡届および転入出届からの推計人口を用いることになる．

前述の調査では合計（sum），平均（average），割合（proportion），比（ratio），率（rate）などで集団の持つ情報を集約する記述統計学（descriptive statistics）の方法が用いられる．そのような統計処理に対し，実験などの結果を検討する際は，サンプル（標本）から母集団の持つ情報を推定する推測統計学（inductive statistics，stochastics）という方法がある．前者の方法は対象集団（母集団）をすべて調べる全数調査であり，後者の方法は母集団を代表する標本を調べ，母集団の持つ情報を知ろうとする標本調査である．

医学研究においては標本から母集団の持つ法則・真理を探ろうとする場合が多いことから，本稿では医

	A	B	C	D	E	F
1	Pno	Name	Sex	Age	Weight	test_A
2	94567	Taro	male	23	54.3	positive
3	2345	Hanako	female	28		positive
4	23567	Gonta	male	65	83.5	negative
:	:	:	:	:	:	:
99	3965	Joshii	female	43	66.6	negative
100	43256	Tago	male	74	62.5	positive

図1 Excelによるデータの作成

学研究に用いられる推測統計学の方法を解説する.

1. 情報収集とデータ・ベース化

臨床データおよび実験結果が得られたという時点から解説をはじめるが，統計処理の過程で直面する問題の多くが研究計画の段階で検討されていなければならない事項に起因するものが目立つ．それらの中でも大半は何を問題にしているのか？何と何を比較するのか？という基本的なことが確認されていないことに由来する場合が多い．そのような状況は理論モデル[1]が不明確な状態である.「とりあえずデータがあるので，何か出して欲しい！」などは序の口で,「このデータの有意差を計算して！」中には「p（ピー）を出して！」など，具体的な対象のない漠然とした要求には統計学は答えられないのである.

よく「データはどのくらい集めたらいいのか？」という質問があるが，基本的に統計学ではデータ数はgiveされるものであって，その過多過少を評価することはあり得ない．ただ母集団から，正確な母集団の情報を持った標本を抽出する場合の基本原理に関する研究分野があり，統計学とは若干ジャンルが異なり,統計学より哲学や科学論に近い領域であるが，それでも統計学で扱う課題も多く含んでいるため，近接領域であることは間違いない．そのような意味で前述の必要サンプル数の質問には「基本的に正規分布する程度あればいい」「2群間で平均値の比較が必要で，2群の平均値の差が仮定できるなら望ましい標本数を求める計算式はあります」というような回答ができる.

推測統計学ではデータ（標本の情報）から一般的推論（母集団の情報）を目的とする．そのためには推論のためのそれなりの手順を踏んでおかなければならない．最も重要な手順は標本が母集団から無作為的に抽出された無作為標本(random sample)であるかどうかである．そして，その標本が母集団を代表するに足る集団かどうかが論文の考察では最初に議論されるべきことになる.

データの統計処理を目的とする場合，データは手元のパーソナル・コンピュータでデータ・ベース化(入力)しておくのが後の統計解析上,効率的である.とりわけMicrosoft Excelを使うとデータのretrieve作業も楽であり，統計解析ソフトの多くがExcelのデータシートを直接扱うことが出来るので都合がよい．その場合，図1のようなデータ形式で後述の①～⑥の要件を満たしていると統計解析がスムーズに進められる.

データをどのように作成するかは研究者個人の好みで構わないが，統計解析をパソコンのソフトを使って行う場合，図1のようなデータ形式が好ましい．基本的な構造として…

① 1行(1レコード)が対象者1人(1件)のデータで占められている.

② ひとつの列は1変数（項目）である.

③ 1行目は各列の変数名(半角の英数)を入れる.

④ 欠測値の場合は空白であること.

⑤ カテゴリー変数の場合，入力値は半角英数であること.

⑥ ワークシートにはデータのみがあり，集計結果（平均等）などがないこと.

以上の6条件が満たされていれば，統計解析ソフトで統計処理がほぼ可能となるはずである．時々統計解析の途中，結果の異変に気づき，データを調べてみると，データの最終行に続けて，平均だとか標準偏差などを計算してあるものがあったりする．またデータ入力の時，0(ゼロ)とO(オー)，1(数字)とl(Lの小文字)の間違いなどがよくあるが，昔のコンピュータでは読み込み時に厳密に変数の型（例

表　目的変数と説明変数が数値，カテゴリーの種類別統計解析方法

		目的変数	
		数値変数	カテゴリー変数
説明変数	数値変数	相関係数 回帰分析，重回帰分析 分散分析（一般線形モデル） 因子分析	ロジスティック回帰分析 数量化理論１類～４類
	カテゴリー変数	t 検定 分散分析（ANOVA） ロジスティック回帰分析	分割表と独立性の検定

えば数値変数か文字変数かなど)が定義されていて，文字と数値の間違いがあれば，たちどころにエラー表示を出して処理を中止していた．最近のソフトは文字も数値も何でも区別なく“貪欲に”読み込んでしまうようで，結果を注意して見ていないと間違ったまま結果を出してしまうことがある．カテゴリー変数の場合の間違いは空白が余計に付加されているようなことが多い．「female」と「female　」の違いに気づく方はまず少ないと思える．データ入力は注意深く行っていても，どこかに間違いはあるもので，集計結果を見ながらデータを retrieve することが重要である．

2. 統計解析デザイン

どのようにデータに対して統計処理を行うかは，目的によって自ずと定まるが，「何でもいいから，結果を出したい！」という希望には同情はするが，目的が明確にならなければ方法が定まらないというのは普遍の真理である．「何を見たいの？」「何を証明したいの？」「何を明らかにしたいの？」という問いに答えられなければ，どのように統計処理を進めるのかは決められないのである．データを持ち込まれて，統計解析を求められることが多い立場から経験的にいうと，研究計画の段階で統計解析方法について検討するか，統計学の専門的知識を持っている方に相談しておくのが良い研究に繋がると言える．

以上のことを前提に推測統計学での方法は，目的変数（従属変数）と説明変数（独立変数）が，まず明らかになっている必要がある．例え２群間で平均値を比較する課題でも，２群に分ける条件（因子）は存在するはずで，具体的には性別，検査の陽性陰性などで２群に分けていることになる．その場合，何かの平均値（平均値を計算できるのであるから，当然数値データである）が目的変数であり，条件（因子）が説明変数ということになる．ちなみに多変量解析という言葉があるが，単に，目的変数に対して説明変数が複数である場合を指し，高度な統計解析手法を意味するものではない．

目的変数と説明変数とも，数値データの場合もあり，カテゴリー[注1]・データである場合もあるが，目的変数が数値であるかカテゴリーであるかは統計解析の方法を選ぶ時に大きな判断材料となる．ちなみに目的変数，説明変数とも数値変数なら，相関，回帰および重回帰分析など，比較的よく知られた方法が用いられる．厳密には単に数値データと雖も，「数量化と標識」という視点から変数の性質を区分する[2]ことも必要となる場合もあるが，詳細は専門書[2]を参照されたい．

目的変数および説明変数がそれぞれ数値変数，カテゴリー変数である場合，主な統計処理法を表に示した．表に示した以外にも解析方法は沢山あるが，ここで示した方法のみでも医学研究に必要なデータ解析は大部分できてしまう．このように目的変数と説明変数，数値変数とカテゴリー変数との組み合わせから統計方法を決定できる．その際注意すべきこととして，数値変数が百分率(例：43.5% など)の場合では，そのままでは統計処理が出来ない，あるいは“不向き”である．百分率で表される変数とは，ある属性のモノが全体に占める割合を示しており，二項分布する変数である．統計処理方法（ことに統計的検定）の多くは正規分布[注2]を前提として成立しているので，二項分布するような変数は，それなりの方法を取らねばならない．正規分布を示さない変

注１　category．哲学的には範疇と訳すべきであるが，観測結果を分類した，いくつかの基本概念を指す．本論においては属性，定性などという言葉に置き換えても論旨の上で差し支えない．

数であっても，便宜的に正規分布へ近似させる方法[注3]がある．また医学で扱うデータは指数分布するものが多く，ことに血中の物質濃度などは測定された数値を対数に変換してから統計解析を行うと，よりよい結果が得られる場合が多い．

3. 統計解析ソフト

実際にデータの統計解析ではパソコン用のアプリケーション・ソフトを使うことになるが，同じデータでもソフトによって多少違う結果が出てくることがある．そういう場合に，統計解析ソフトについて詳細な日本語の解説書があるかどうか，ソフトを選ぶ際に重要なポイントになる．市販の統計解析ソフトの多くがパソコンへのインストール方法のみ記述した解説書を添付するだけで，実際の処理過程（計算方法）が不明確な場合が多いが，市販本として解説書や事例集が出版されている場合もある．

そこで，学術誌や関係する論文を参考にして，多く使用されている統計解析ソフトを選ぶのが最適である．言わば統計解析用のアプリケーション・ソフトはブラック・ボックスのようなもので，中で何をやっているのか解らない場合が多いので，定評のあるソフトを選ぶ方が無難と言える．

表計算ソフトExcelを使って統計解析を行うことは可能であるが，Excelでは統計処理のために関数を用意していて，統計処理＝関数の使い方という定式が成り立つようである．それでも，Excelの関数を使いこなして，統計処理をするのは，結構大変な作業になる．理由は操作が複雑で，手順をスクリプト化することは出来るが，そのスクリプトを使って他のデータを同じように処理するのは，難しい点からあまりお薦めできない．Excelのアドイン・ソフトとして，統計解析用のものが販売されている．これはマウス操作だけで，統計解析が出来るようになっており，結構便利に使えるが，あまり学術論文での使用例は少ないようである．

手順の[注4]“軌跡”であるExcelのスクリプトとは違って，SASなどは“手順”をプログラムすることによって，同じような処理を反復的に適用することが可能である．それによって作ったプログラムは“知的財産”であり，他のデータ解析に応用できるというのも強みである．パソコンの統計解析ソフトを選ぶ場合，まず自分の周囲を見回して，よく使われているソフトを採用するのが最も堅実である．また，そのソフトに習熟した方を見つけたら，是非とも友達になるべきである．

統計解析用のアプリケーション・ソフトで最も古くからあって，現在も使われているものにSPSS[注5]がある．当初は汎用機用のコンピュータ・ソフトであったが，日本語による解説書[4)5)]もあって，非常に良く使われたソフトである．現在ではパソコン版[注6]が登場して以来，操作性とグラフ作成機能が非常に便利になり，価格も10万円程度なので，個人で買うにはちょっと高い気もするが，医学関係で使用する特殊な統計処理法も含まれている点から医学研究で有効な統計処理アプリケーション・ソフトと言える．

医学関係の欧文論文検索システムであるmedline，pubmedなどで，統計処理に使われている統計ソフトを検索してみると，SAS，STATA，STATISTICA，SPSSなどが多く使われているようである．そのうちで，日本語による解説書が多くあるのはSAS，SPSSとSTATISTICAでSTATAは英文の解説書は非常に多いが，日本語のものが見当たらない．価格は数万円から数十万円程度である．

欲を言えばキリがないが，統計解析の事例や解説が日本語で多くあり，無料で使える統計解析ソフトはないのか？問われれば，答えは“あーる”である．R（アールと呼ぶ）というソフトがある．GNU[注7]プロジェクトで開発された統計解析用のソフトウェアであるが，基本的に無料で配布されている．また解説は事例を掲載したWebサイトも多くあり，改定作

注2　データのヒストグラム（histgram）を描いたとき，中央に頻度が高いことを示す山があり，左右に対称な分布を正規分布という．数量的にはデータの歪度（skewness）と尖度（kurtosis）がともに0に近いものである．

注3　百分率（%）の値をpとし，$\theta=\sin^{-1}\sqrt{p}$に変換すると，正規分布に近似させることができる．ただしpは$-1\leqq p\leqq 1$の範囲にあること．

注4　SAS（Statistical Analysis System），米国SAS Institute社製のコンピュータ・ソフトウェア．1960年代から統計解析ソフトウェアとして開発がはじまり，現在多くの分野で使われている．

注5　SPSS（Statistical Package for the Social Science），1960年代にスタンフォード大学で開発がはじまり，その後シカゴ大学で改良が重ねられた統計解析用ソフトウェア．現在，SPSSはパソコンでも稼動するシステムとして，広く使用されている．

注6　対応OSはWindowsのみである．医学用として，Dr.SPSSとして販売されている．

注7　Linux等無料で使えるフリーソフトを開発するプロジェクトであり，そのプロジェクトで作られたソフトウェアは基本的に無料で配布される．

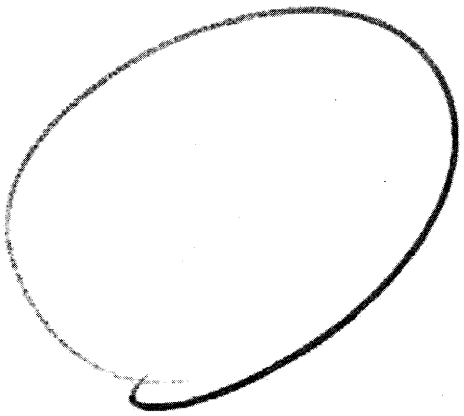
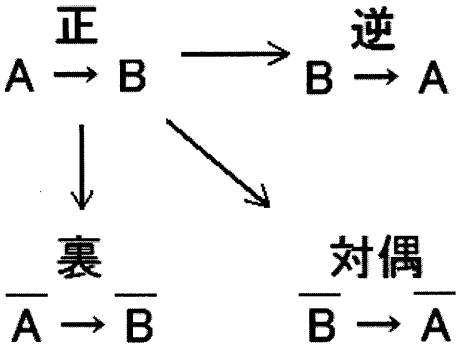


図 2



A → B: AならばBである

$\overline{A} \rightarrow \overline{B}$: AでないならばBではない

図3 アリストテレスの形式論理

業も進んでいるようで，今後，学術誌や学会で広く使われていくようであれば，統計解析ソフトとして定着するように思える．

統計解析ソフトとして何を使うかということを考えたとき，学会や学術誌でよく使われているものを選ぶのが間違いない．論文の中で，定評のあるソフトの，どういう解析法を用いたか記述すれば，読者に明確に伝わりやすいからである．

4. 論理と統計的検定

本稿の冒頭でも述べたが，統計解析を「p（ピー）」を出すものと理解している方が時々いるが，統計的検定は判断に迷う場合に役立つもので，決して論文の論理を左右するものではないことを強調したい．図2を見て，三角形や四角形に見えることはないはずで，どう見ても○（マル）あるいは円である．円の公式を当てはめて，本当に円になっているか検証するのは無駄である．元々データは論理の根拠として示すか，データから論理を引き出すものである．グラフで増加傾向にあるのか？減少傾向にあるのか？見れば判るものを検定する必要はない．そういう判定が難しい場合に統計的検定が必要となる．

統計解析用のソフトウェアを使う場合，統計的検定の結果，検定で使う統計量の値がより大きな値を示す残り確率であるｐ値を，結構，正確に算出（例えばｐ＝0.0014など）するためか，そのまま論文に引用しているのを多く見かける．統計学の原理主義的観点からすれば，そういう場合のｐ値の提示は誤解を招く恐れがある．というのは，統計量とともに算出された残りの確率であるｐ値は，検定の条件として当初設定した判断基準より小さいか？どうか？が問われるのであるから，ｐ＝0.0014という数値を提示していると，非常に厳しい判断基準を用いたものと誤解される．必要なのは検定にあたって，自分はどういう判断基準で，判断したかということが重要なのである．同時に，ここでいうｐ値はあくまでも，判断のための材料であって，比較のためのものではない．ｐ値が0.05未満だったら，帰無仮説を棄却するという時の判断に用いるのであって，ｐ＝0.001なので，ｐ＝0.01よりも“大きな”有意な差を認めたことにはならない．つまり比較の材料ではないのである．

統計的検定である因子によって2群に分けられたA群とB群があって，A群B群の分散が等しいという条件のもとで，それぞれの平均値を比較するため，平均値に差があるか？という検定をする場合，検定方法はｔ検定を選ぶことになるが，その検定結果が，「2群間には有意な差がある」と判断し，その場合，判断が間違いである確率が5％未満（ｐ＜0.05）であった．さて，その結果から論文の中で，「A群はB群よりも平均値が大きい（小さい）」と断言できるだろうか？答えは否である．この場合の統計的検定で判断されたのは「A群とB群の平均値に差がある」という点だけであって，「A群はB群よりも平均値が大きい（小さい）」という点は検定していない．重要なことは，何を検定して，判断したか？ということである．図3にアリストテレスの形式論理の模式を示すが，「AならばBである（A→B）」という命題を証明（判断）した場合，常に成り立つのは対偶命題のみである．逆命題も裏命題も，場合によっては成り立つが常に成り立つ訳ではない．

形式論理では「人間なら2本足である」という論理が真であるなら，同時に成り立つのは，その命題

の対偶である「2本足でないなら人間ではない」という論理である．逆の論理である「2本足なら人間である」また裏の論理である「人間でないなら2本足ではない」は成り立たない．

つまり「差があるか？」を検定したのであって，「大きい（小さい）か？」を検定した訳ではない．「A群の平均値はB群の平均値より大きい（小さい）か？」ということを判断するにはt検定法で，片側検定という方法で可能である．

おわりに

以上まとめると統計処理の上で最も重要な点は，データから「何を見たいのか」「何を証明したいのか」「何を明らかにしたいのか」目的を明確にしておくことが必要である．それは研究目的を明確にすることに他ならない．

次稿では具体的に平均値あるいは分散を使って集団群を比較する方法を解説する予定である．

文　献

1) **池田　央**：「調査と測定，社会科学・行動科学のための数学入門4」，新曜社，東京（1980）
2) **水野哲夫**：「統計の基礎と実際」，光生館，東京（1970）
3) **稲葉三男，北川敏男**：「統計学通論」，共立出版，東京（1960）
4) **三宅一郎，山本嘉一郎**：「SPSS統計パッケージI基礎編」，東洋経済新報社，東京（1976）
5) **三宅一郎，中野嘉弘，水野欽司ほか**：「SPSS統計パッケージII解析編」，東洋経済新報社，東京（1977）
6) **田栗正章，藤越康祝，柳井晴夫ほか**：「やさしい統計入門」，講談社，東京（2007）