統計入門

〔第12回〕

パソコン利用の統計の実際（2）

佐伯圭一郎* 高木廣文** 中井里史*** 新田裕史****

はじめに

前回では，データの入力から統計解析の段階におけるパソコンの利用をいくつかのソフトウェアに例をとって解説した。今回は，解析結果の出力の段階および実際の調査研究のいくつかの場面におけるパソコンの利用について紹介する。

I. 解析結果の出力

1. グラフ作成ソフトウェア

統計解析をパソコン上で行っても，その結果を手作業で図や表に仕上げていたのでは，パソコンを利用するメリットが半減してしまう。グラフを作成するにしても，データを修正したり設定を変更したりするたびに手で描いていては大変な手間である。また，データの解析の途中でもデータをグラフ化することにより，得られる情報は大きくデータ解析とグラフ化は切り放せない関係にある。

HALBAU には多数の解析においてグラフ化して表示・出力する機能が用意されている。ロータスにおいては，各種のグラフが用意され細かい指定が可能である。ただし，最終的な出力を紙に印刷した場合，研究論文用としては，まだ満足のいく水準に達していない。そのため，場合によってはグラフ作成のための別のソフトウェアを利用するのも一つの方法である。

図1に HALBAU のグラフ，**図2**に Lotus 1-2-3の例，**図3**は Lotus 社のグラフ作成ソフトウェアである Lotus Freelance の出力例を示す。

図3のように，グラフ作成専用のソフトウェアを用いれば一応のレベルの図を作成することができることが理解できよう。

2. さまざまなグラフ

統計で用いるさまざまなグラフのうち，データの分布をグラフ化したヒストグラム，2変数の関係を図示した散布図ですでに本連載で解説されている。データ分布を示すグラフとして箱ヒゲ図（box-and-whisker）も最近は一般的になっている（**図4**）。

また，視覚に訴えるさまざまな種類のグラフがソフトウェアによっては用意されている。棒グラフや円グラフに立体的な修飾を加えただけのグラフから，棒グラフで棒の代わりに絵を用いる絵グラフなどは研究論文で用いることは稀であるが，プレゼンテーションにおいては有用であろう。地図グラフや顔グラフなど，さまざまな種類のグラフ表現が存在する，詳細は各種文献を参照していただきたい。

グラフの作成における一般的な注意点であるが，グラフは視覚的に受ける印象が優先されるため，たとえば棒グラフで下の部分を省き，差を大きく見せるといった操作が行われやすい。グラフ作成における誤解を招きやすい例を含めて統計解析結果の示しかた一般について，文献1）に平易に述べられている。

3. 作　表

作表においても，解析結果の数値を再度ワープロで新しく入力しなおすことは煩雑であり，数値

* 獨協医科大学公衆衛生学
〔〒321-02 栃木県下都賀郡壬生町北小林 880〕
** 文部省統計数理研究所
*** 東京大学医学部保健学科疫学
**** 国立環境研究所

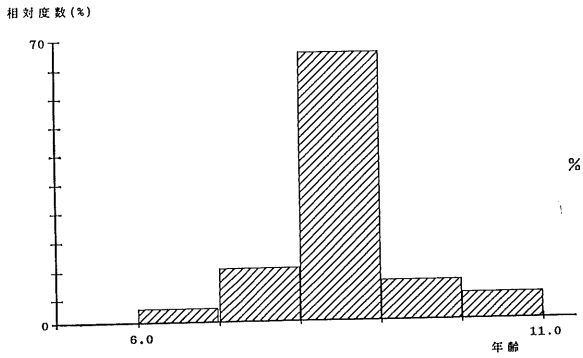

図 1 HALBAU によるヒストグラムの例
(年齢のヒストグラム)

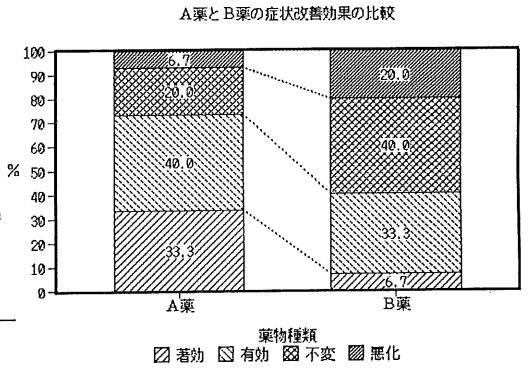

図 2 Lotus 1-2-3 によるグラフの例
(症状改善の内訳)

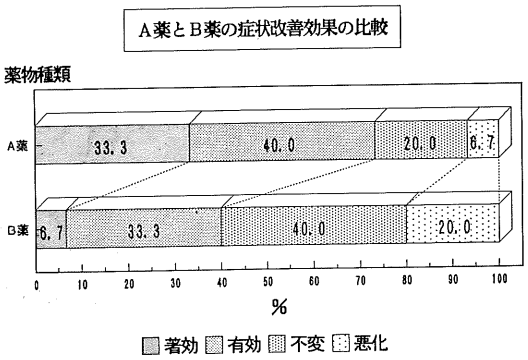

図 3 Lotus Freelance によるグラフの例 (図2と同じグラフ)

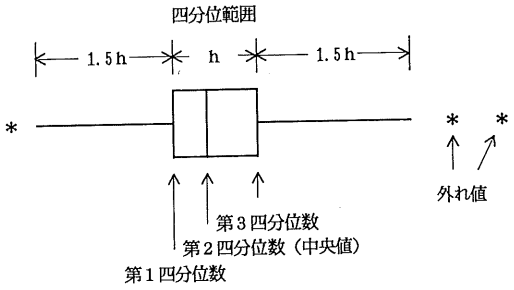

図 4 箱ヒゲ図の例

の入力の誤りの発生の可能性もある。Lotus 1-2-3では，罫線も引け簡易なワープロ程度には印刷することもできる。HALBAU の場合には，結果の出力メニューで，プリンターではなく，ファイルに文書として出力することができるので，出力ファイルを日本語ワープロで加工して，タイトルなどをつけることにより作表が行える。

なお，一般的な作図・作表のルールを守ることを心がけ，適切な情報を過不足なく盛り込むように，たとえば検定をつけ加える場合に有意かどうかを星印で示すだけといったことを避け，きちんと検定統計量と自由度くらいは書き加えるようにしていただきたい。

II. その他の計算機環境

パソコンよりも身近ではないが，一部の方は，大型のコンピュータやいわゆるワークステーションを使える環境にあるだろう。このような環境でも，使うソフトウェアは異なるであろうが，データの入力などに関する基本的な考え方はこれまでの場合と同じである。

大型計算機上で使用可能な統計パッケージとして，SAS や SPSS といったソフトウェアがよく使われている。ともに統計解析だけでなく，作図・作表，データベースといった機能も含んでいる。各種の操作を行うには，プログラム言語に類する命令を覚え，各自で目的に合わせた指示を作成しなければならない。統計解析はパソコンと同様に結果を検討しながら次の処理を実行するという対話型の処理も，ひとまとまりの処理をあらかじめプログラムに作成しておき，自動的に実行させるバッチ処理も可能である。機能や操作法は文献2, 3) などに詳しいが，非常に多種の高度な統計解析が高速に，また大規模なデータでも行える点が大型計算機の統計ソフトウェアの特長であり，ソフトウェア自体も長年に使用されているた

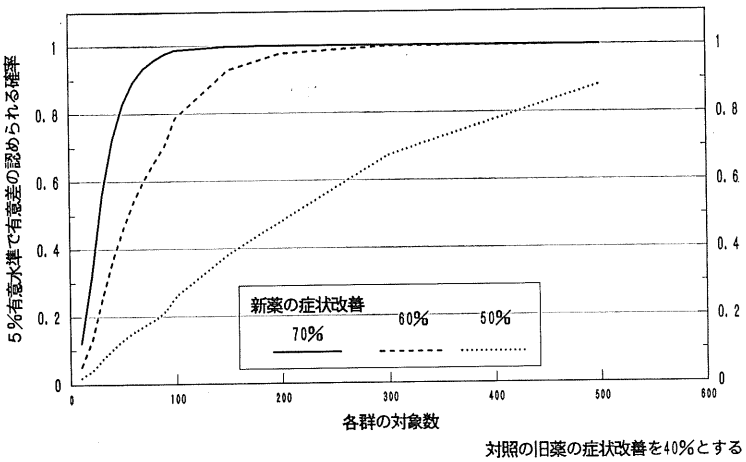

図 5 検出力曲線の例

め信頼性も高い。また，最近ではこれらのソフトウェアのパソコン版も高価ではあるが，国産のパソコンで利用できるようになっている。

数百から数千のケース数の場合には，現在の状況ではパソコンで十分処理が可能であり，コストの面でもパソコンを利用する方が有利であるが，データ数が莫大であったり大型計算機を利用しやすい環境にある場合にはこれらのソフトウェアを利用するということもあろう。大型計算機の統計ソフトウェアで統計解析を行う場合でもデータはパソコンで利用したデータを利用することができる。たとえば大型計算機で一般的な固定書式という形式にはHALBAUから変換できるようになっている。したがって，データの入力や管理，基礎統計をパソコンで行い，複雑で特殊な分析法を大型計算機で行うといった使い分けも可能である。

III. 調査研究におけるさまざまなパソコンの利用法

統計解析については，これまでに解説してきたが，統計解析以外の調査研究のさまざまな局面でパソコンを手助けとして利用することができる。もちろん，目的に応じたソフトウェアを使いこなし，場合によっては自らプログラミングを行う必要があるが，以下では簡単にいくつかの例を紹介するにとどめよう。後は必要に応じたソフトウェアの解説書，自身でプログラムを作成するのであれば文献4）などの入門書を参照されたい。

1. 対象数の検討

たとえば，新薬が従来の薬の効果を上回っているかどうかを研究する場合がある。すでに理解していることであるが，統計的検定は対象数に影響され，対象の数が多くなればなるほど“有意な”結果が得られやすくなる。現実には対象数はさまざまな制約から無制限に増やせないし，不必要に多くの対象を取り扱うことは浪費である。

本当に差が存在する場合，研究データからたとえば“有意水準5％で有意な差がみられた”という結果が導かれる確率を検出力といい，この確率は対象の数，ならびに当然であるが真の差に影響される。したがって，真の効果を仮定してやれば，どの位の対象数の場合には正しく有意差ありという結果を得る可能性がどの程度であるか計算できる。また，逆に“有意差がない”という結果が得られた場合にも，用いた対象数ではどの程度の差を検出できるかということを知ることで結果を考察する場合の助けになる。

図5に対象数（各群に同じ人数を振り分けるとする）によって，“5％の有意水準で有意差あり”という結果が得られる確率が変化するかを計算した例を示す。従来の薬物で症状改善のみられる確率を40％とし，新薬の症状改善を70％，60％，50％とした場合に有意水準5％でイェーツの補正をしたカイ2乗検定を行うと仮定している。

この例では，新薬の症状改善の真の値が70％の時には数十例で十分であるが，50％ではかなり多

くの対象を用いなければならないことが理解できよう。また，従来の薬に比べて差が小さい場合には，対象が少なければ統計的検定ではほとんど有意にはならないということがわかるであろう。

計算には真の効果を仮定することが必要であることなどから，あくまで参考として用いるのみであるが研究デザインの段階で重要な情報となる。

2. データベースとしての利用

1）マージ作業

一度限りの調査で終わる場合にはさほど重要ではないが，継続して対象を観察する形式の調査の場合には，各観察時点のデータを蓄積し，最終的にはまとめて解析を行わなければならない。各時点のデータをまとめて1つのデータにする作業もパソコンの利用が大きな助けになる。

各調査時点のデータを1つにまとめる（マージする）さいに，手作業で行っていたのでは対象が多くなると作業量は莫大なものになる。パソコンでデータベースのソフトウェアなどを用いて，各対象を同定してまとめる作業を自動的に行わせ，同定が不完全な部分のみを人手で行うことが効率的である。実際の調査では，個々の対象を同定する部分，たとえば氏名や生年月日，もしくは最初に与える ID などはデータの原本の段階から記入漏れや誤りが存在することがしばしばである。このように不完全なデータでもパソコンを利用してチェックを行うことにより，かなりの効率で各対象の同定が可能となる。

2）標本の抽出

また，大きな集団から一部の対象を抜き出して調査する場合，たとえば住民の台帳から特定の年齢のもののみをある人数抜き出して対象とする場合や，これまでの患者のデータから同じ疾患の患者のみを抜き出して分析を行いたい場合がある。また，複雑な作業としては，ある病気の患者1人に対して，性や年齢などでマッチングさせた対照を別の集団から選び出す場合がある。

これらの抽出法を用いる各種の研究デザインとそれに応じた統計解析の方法は，文献5, 6）や最近いくつか出版されている臨床疫学という分野に関する文献を参照していただきたい。実際の標本抽出の作業自体の大部分は，表計算ソフトウェアなどを利用してパソコン上で行うことができる。

3）情報の管理

パソコンで取り扱えるデータは，統計解析に用いるような種類のデータだけではない。文献のデータベースを自分で構築することや，CD-ROM やネットワークで提供されている医学文献のデータベースを有効に利用することができれば，研究に大いに役立つことであろう。

また，いわゆるパソコン通信において，医学・医療関係者を対象としたものもいくつか存在し，活発な活動が行われている。興味をもたれた方はそのようなネットワークに参加することも有意義であろう。

おわりに

繰り返しになるが，パソコンを使うことそれ自体は目的ではない。データの数や状況によっては従来の方法で充分な場合もある。パソコンを利用して行うことで充分なメリットが得られるかどうかは状況に応じて判断していただきたい。

また，本連載ではふれなかったが，実際にデータを得る段階で，データの誤差を少なくするための方法として質問紙の作成法や測定の精度管理といった問題も統計解析に影響を与える重要な問題である。統計学の理解に加えて，調査法に関する文献も各自で参照していただければ，研究の成果はさらに向上することであろう。

文　献

1) ダレル・ハフ：統計でウソをつく法，講談社ブルーバックス，1968
2) 市川伸一，大橋靖雄：SAS によるデータ解析入門，東京大学出版会，東京，1987
3) 三宅一郎，他：SPSS 統計パッケージI・II，東洋経済新報社，東京，1976，1977
4) 新田裕史，佐藤俊哉：パソコン BASIC 生物統計学入門，現代数学社，東京，1989
5) Anderson S, et al: Statistical Methods for Comparative Studies. John Wiley & Sons, 1980（重松・柳川監訳：疫学・臨床医学における比較研究の統計学，ソフトサイエンス社，東京，1982）
6) 吉村　功編著：毒性・薬効データの統計解析—事例研究によるアプローチ，サイエンティスト社，東京，1987