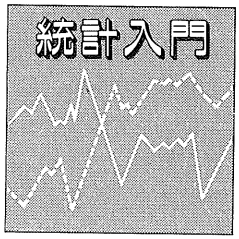

〔第11回〕

パソコン利用の統計の実際 (1)

佐伯圭一郎* 高木廣文** 中井里史*** 新田裕史****

はじめに

前回までの連載で，研究上必要な統計手法の大部分を学んできた。今回と次回では手持ちのデータから，統計解析，結果のレポートに至るまでの過程におけるパソコンの利用について概説する。

I. パソコンと統計ソフトウェア

最近のパソコン（パーソナル・コンピュータ）の普及はめざましく，読者の机上または周辺には1セットはパソコンが置かれているであろう。大部分が，例えば一太郎（ジャストシステム）といった日本語ワープロを利用しているであろうし，自分で利用しないにしても原稿の清書に使ってもらっているであろう。

パソコンを使用することの利点は，いつでも好きな時に，結果をみて適宜修正を加えながら作業することが容易に，自分でできるという点である。統計解析においても，プログラミングやコンピュータといった専門的な知識を必要とせず，適切な解析方法を選択することさえできれば，手元で十分な解析が行える。しかもデータを一度パソコンに入力すれば，分析結果を考慮しながら，条件を変えての分析，別な解析手法の適用も簡単に行えるという利点がある。また，解析結果のグラフ化や作表も，一貫してパソコン上で行える便利さがある。もちろん，各種の処理を行うためには，統計処理やそれぞれの目的に適したソフトウェアを使用しなければならない。

以下では，統計解析に用いるソフトウェアの実例として，統計パッケージ HALBAU（現代数学社），Lotus 1-2-3（ロータス）を取り上げる。紙面の制約上，すべての機能を解説することはできないので，詳しくは文献1，2などを参照してもらいたい。

HALBAU は，柳井らによって開発された統計パッケージであり，NEC PC-9801 シリーズおよび互換機上で動作する。基礎統計から多変量解析まで主要な統計手法のほとんどを網羅している。操作は，初心者にもわかりやすいメニュー選択方式を採用している。

Lotus 1-2-3 は，ロータス社製の表計算ソフトウェアのベストセラーで，国内外の主要なパソコン用に発売されている。統計処理の機能は基本的な統計量や度数分布，回帰分析などに限られるが，データの入力や加工といったデータの管理，簡単なグラフの作成などの機能が含まれ，非常に有用なソフトウェアといえる。表計算ソフトウェアのジャンルでは，アシストカルク（アシスト），エクセル（マイクロソフト）など，他のソフトウェアもほぼ同様の機能をもっている。

なお，これらソフトウェアは著作権を保護された著作物であり，違法なコピーを使用することなく，正規に購入しなければならない。

調査や実験から得られたデータをパソコン上で解析する場合，実際の作業は大きく三段階に分けられる。

1）データのパソコンへの入力

2）統計解析

* 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

3）作図・作表
の三段階である。

II. データの入力

1. データのコード化

データは量的データと質的データに区別できる。パソコンで取り扱う場合には，量的データは数値をそのまま入力すればよい。質的データの場合には，例えば“症状あり＝1”，“症状なし＝2”と数値に置き換える必要がある。この置き換えの作業をコード化という。表計算ソフトウェアや最後にふれる大型計算機の統計パッケージなどでは質的データを文字のままで入力することもできるが，取り扱いの容易さからコード化しておくことが一般的である。

質的データをコード化する際には，“著効＝1，有効＝2，不変＝3，悪化＝4”などと，1から順に整数を与える。順序データの場合には，コードも順につける必要がある。

あらかじめ選択肢の数が決まっている項目の場合には特に問題はないが，例えば“日常行っている運動は何ですか”といった自由回答形式の質問をした場合，コード化の手間は増加する。一つには，あらかじめ回答を抜き出した一覧表を作成し，コード化する方法がある。また，表計算ソフトウェアを使用する場合は，いったん文字のままで回答を入力，集計し，後にコード化する方法もとれる。

また，データの入力の形式でわかりづらいと思われるものは，選択肢を複数選べる複数回答の項目であろう。例えば“20の症状のなかで当てはまるものに丸をつけて下さい”といった形式である。この場合には，それぞれの症状について“あり・なし”といった変数20個にして入力することがよいであろう。

2. 欠損値の取り扱い

回答もれや検査をしなかったという理由で，ある変数の値が得られないケースがままある。この欠損値（欠測値）をどう入力するかということは，使用するソフトウェアで異なる。Lotus 1-2-3など表計算ソフトウェアの場合には，ワークシートの該当する部分を空白にする，つまり何も入力しないことで，欠損値を示す。HALBAUの場合には，マイナスの大きな値（標準では−99）によって，その値が欠損値であることを表わすように決められている。設定の変更も可能であるが，医学・生物学的なデータの場合には大きなマイナスの数値はほとんどなく，実際上は十分である。

3. 入力作業

実際に数値をパソコンに入力する作業（パンチするという）は，単純な作業でケース数に比例して作業量が増加してしまう。変数の数にもよるが，ケース数が数100例程度であれば，自分でデータの原本をみながら入力することができる。しかも，生のデータを自分でみることによって，なにかしら得られるものがあるだろう。

データが大量であれば，業者にパンチを依頼するのが一般的であろう。または，周囲にデータ入力を頼める人がいれば依頼することもできる。

その際に，アンケート用紙やカルテからそのまま入力する方法と，いったん集計用紙にコード化をすませた状態で転記して，それを機械的に入力する方法とがある。どちらも一長一短があるが，どちらにせよデータのコード化，入力については，第三者に依頼する場合には詳細な説明を添える必要がある。

4. データのチェック

正しいデータを用いなければ，統計解析の結果の信頼性は保てない。人間は誤りをおかすものであり，正確に入力したつもりでも，パンチ作業にあけるミスは必ず発生する。データに入ってくる誤りは測定，調査の段階から幾種類もあるが，入力の段階の誤りは，入念にチェックすることで避けることができる。

データの入力が終了したら，誤りなく入力されているかのチェックを必ず行う。画面もしくはプリントアウトと原本を突き合わせて，丁寧にチェックしていただきたい。また，度数分布や最大値，最小値などを求めること，変数間の関係に注目することなどにより，機械的なチェックを加えることもできる。

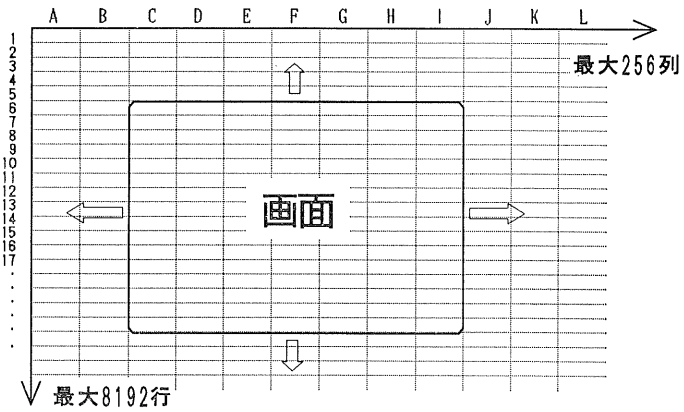

図 1　ワークシートの概念

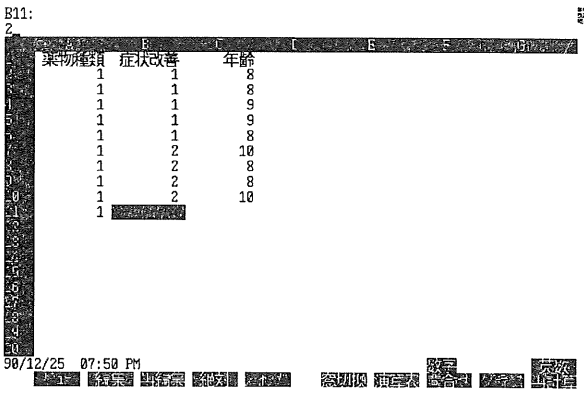

図 2　Lotus 1-2-3（データ入力中）

III. 統計解析の実際

1. Lotus 1-2-3

Lotus 1-2-3 を使用して，データの入力，加工，基礎統計までの手順を説明しよう。

1）データ入力

表計算ソフトウェアとは，**図 1** に概念を示すように，縦横に仕切られた巨大な集計用紙（ワークシート）をパソコン上に実現したものである。一つ一つのます目をセルとよび，一つのセルに一つの数値が入る。このワークシートの行方向がケース，列方向に変数が並ぶという形式が一般的である。また，それぞれのセルには列（英字），行（番号）によって番地が付けられている。

サンプルデータは，投与した薬物の種類と症状の改善効果，および患者の年齢について30例のデータとしよう。薬物の種類は（薬物A＝1，薬物B＝2），症状の改善は（著効＝1，有効＝2，不変＝3，悪化＝4）とコード化し，年齢（歳）はそのまま入力する。

図 2 に実際のデータ入力中の画面を示すが，一番上の行にはその列の変数の名前を入れ，その下に各ケースのデータを入力している。

2）基本的な統計量

ロータスで簡単に実行可能な統計解析は，最大・最小値，平均値や標準偏差，頻度・度数分布，回帰分析などである。最大・最小値，平均値や標準偏差などは**表 1** に示す統計関数という機能を使って実現される。

この表はロータスの統計関数の一部である。例

表 1 Lotus 1-2-3 の主要な統計関数

@ SUM	：合計を求める
@ COUNT	：データ数をカウントする
@ AVG	：平均値
@ MAX	：最大値
@ MIN	：最小値
@ VAR	：分散
@ STD	：標準偏差
@ RANK	：データの順位

えば年齢の平均値を求めるのであれば，@ AVG 関数を利用し，データの範囲は C2 から C31 であるので，年齢の列の下に @ AVG（C2..C31）と入力すればよい。

質的データを前に説明したルールでコード化している場合は，頻度は度数分布と同じ方法で求めることができる。ワークシートの任意の部分に階級値の上限の値を記入し，ロータスのメニューで頻度を求める命令を選べばよい。

3）データの加工

例のデータでは，症状改善効果は 4 段階の評価であるが，この変数を有効と無効の 2 段階に変更するといったデータ変容の作業も解析の作業の一段階として必要な場合がある。その場合でも新たにコード化しながらデータを入力する必要はない。ここでも関数を利用して，データの変容をワークシート上で行うことができる。

この場合は，新しい変数をデータの右端に追加し，1 番目のケースのセルに @ IF（B2<=2, 1, 2）という関数を入力する。詳しい意味は文献を参照して欲しいが，効果の値が 2 以下であれば，新しい変数の値を 1 に，2 を越える場合は 2 にするという意味である。この式を Lotus 1-2-3 の複写の機能を使って，全ケースについて求めれば新しい変数が作成される。

また，年齢を 5 歳刻みにまとめるといった，量的データをいくつかのカテゴリに分割してカテゴリカルデータに変換する場合も同様の方法が使える。もちろん，四則演算や他の関数を使って新しい変数を作成することも容易である。

クロス表を作成することや，t 検定のための t 値など検定統計量を計算することもできるが，検定で有意確率などを計算する機能はロータスには含まれておらず，これは以下で HALBAU で実例を示す。

表 2 HALBAU 基礎統計分析プログラムメニュー

1. 基本統計量の計算
2. 項目別カテゴリ度数の算出
3. 度数分布表とヒストグラムなどの作成
4. 幹葉表示の作成
5. クロス表の作成と検定
6. 四分表の併合
7. 相関係数の計算と検定
8. 相関図の作成
9. 3 変数以上の同時相関図の作成
10. 2 群の母平均値の差の検定
11. 一元配置分散分析＋α
12. 順位和検定，符号付順位和検定，符号検定
13. Kruskal-Wallis 検定＋α
14. 二元配置分散分析（母数モデル）
15. 5 数要約値と箱ヒゲ図の作成
16. 領域図の作成

2. HALBAU

1）データの準備

HALBAU の機能は大きく，

1）データおよびファイルの操作

2）基礎統計学パッケージ

3）多変量解析パッケージ

に分かれている。1）のデータおよびファイルの操作の機能の中に，データの入力機能が用意されているが，それは文献を参照していただくこととし，ここでは先に Lotus 1-2-3 で入力，チェック，加工をすませたデータを HALBAU で利用する方法を説明しよう。

Lotus 1-2-3 において，変数名の行の上に 1 行挿入し，左端からケースの数と変数の数を入力する。欠損値である空白に −99 を入力し，データ，変数名以外の部分は消去して保存しておく。

ロータスで保存したファイルは特殊な形式であり，そのままでは HALBAU で取り扱えない。HALBAU で取り扱うデータファイルの形式は，K3 フォーマットという形式である。ロータスの初期メニューでファイルの変換を選び，変換先の

表 3 “2群の母平均値の差の検定” 結果の出力

群変数名：薬物種類

変数名 群（群変数の値）	標本数	平均値	不偏 標準偏差	等分散の 検定	等分散の 場合	Welchの 検定
年齢	30	8.033	0.809	1.708	1.624	1.624
1 (1.0− 1.0)	15	8.267	0.884	(14, 14)	(28)	(26.21)
2 (2.0− 2.0)	15	7.800	0.676	(0.3278)	(0.1155)	(0.1164)

( ) 内：上段は自由度，下段は有意確率

表 4 “クロス表の作成と検定” 結果の出力

項目名（縦−横）：症状改善—薬物種類			（無効標本数 0）
	1.（%）	2.（%）	3.（%）
1.	5 (33.3)	1 (6.7)	6 (20.0)
2.	6 (40.0)	5 (33.3)	11 (36.7)
3.	3 (20.0)	6 (40.0)	9 (30.0)
4.	1 (6.7)	3 (20.0)	4 (13.3)
合計	15(100.0)	15(100.0)	30(100.0)

カイ2乗値（自由度） 4.75758 (3) 有意確率 0.190434

クラメールの関連係数 0.398228

表 5 “順位和検定” 結果の出力 [Wilcoxon の順位和検定（U検定）]

群変数名：薬物種類（確率は両側確率）

変数名 群（群変数の範囲）	標本数	順位和	U統計量	数値表の 有意確率	正規近似値	正規近似値 の有意確率
症状改善	30					
1 (1.0− 1.0)	15	184.00	64.00	……	2.1038	0.03540
2 (2.0− 2.0)	15	281.00	161.00			

形式を K3 フォーマットとすれば HALBAU で使用可能なデータとなる。

2）HALBAU による統計処理

① 基礎統計

HALBAU の基礎統計学パッケージには，以下の統計解析が含まれている（**表 2**）。

2群の母平均値の差の検定（t 検定）で投薬群別の年齢の差をみてみよう。群別変数に，薬物の種類を指定し，分析に用いる変数に年齢を指定すると，以下のような出力が得られる。等分散性の検定ならびに等分散の場合，不等分散の場合の検定結果がまとめて出力されている。統計結果の意味については，第6回を参照して欲しい（**表 3**）。

薬物と症状改善効果のクロス表は，クロス表の作成を実行すれば統計量をあわせて以下のように出力される（**表 4**）。

この場合は，症状改善効果は順序統計量なので，薬物の種類との関連を探るには，ウィルコクソン（Wilcoxon）の順位和検定が適している。これも基礎統計に含まれ，次のような出力が得られる（**表 5**）。

② 多変量解析

多変量解析は以下の分析が用意されている（**表 6**）。

表 6 HALBAU 多変量解析プログラムメニュー

1. 重回帰分析と数量化理論1類
2. 主成分分析
3. 正準相関分析と数量化理論3類（双対尺度法）
4. 判別分析
5. 数量化理論2類
6. 非線形モデルと生命表解析
7. 因子分析
8. クラスター分析

解析の実行も基礎統計と同様に，メニューで選択するだけで容易であるが，各解析の意味，前提条件などを理解したうえで分析を行いたい。

おわりに

パソコンはいわゆるブラックボックスとして操作して結構だが，解析の目的，なんのためにその解析方法を選んだのかという意識とデータの質は自身で確保してもらいたい。

次回では，パソコンを利用した解析作業の三番目の段階である作図，作表へのパソコンの利用を中心に，大型計算機の利用や調査の計画についても解説しよう。

文　献

1) 高木廣文，佐伯圭一郎，中井里史：HALBAU によるデータ解析入門，現代数学社，1989.
2) 真鍋龍太郎，他：文科系のコンピュータ/応用篇，岩波書店，1988.

*　　*　　*