統計数理 (1996)
第44巻 第2号 149-162

特集「統計ソフトウェアの新展開」
[研究ノート]

SAS/INSIGHTによるデータ分析の教育

東京大学* 浜田知久馬
SAS Institute** 岸本淳司

(1996年7月 受付)

1. はじめに

最近のコンピュータ技術の変化の速さは驚くべきものである．ハードウエア・ソフトウエアの双方が半年間隔でモデルチェンジ・バージョンアップを繰り返している．10年前，ハードディスクが20メガバイト用意されていれば，ずいぶん贅沢な感じがしたが，いまではオーダーが2桁かわり，ギガの単位でハードディスクの容量を評価するようになった．ハードウエアの進歩に応じ，急速にソフトウエアも量的かつ質的に変化し，今ではフロッピーディスク1枚で提供されるようなソフトウエアはほとんどなくなり，随分使い勝手のよいソフトウエアが増えてきた．統計解析ソフトウエアについてもユーザーフレンドリーなものが現在普及しつつある．これに伴い統計解析の方法論・教育法についても変化が要求されている．一昔前，正確な統計計算を電卓を使って行うことが重要な課題であった．したがって多くの実務家向けの統計学の教科書では，計算方法の記述に紙面の大部分が割かれ，電卓を使って計算ができるように，あるいはフォートランなどでプログラム化ができるように配慮されていた．ところが最近ではEXCEL等の表計算ソフトでも重回帰分析等が可能であり，また最新の統計ソフトウエアであるVisual Stat, STATVIEW, JMPを用いれば，マウスだけの操作で主成分分析やロジスティック回帰などの統計手法を実行することが可能である．コンピュータ技術の発展によって数式が苦手な実務家に対しても統計的な方法論の利用の道が開かれ，統計学の恩恵を享受することができる母集団が大幅に拡大したといえる．反面，高度な統計手法の内容・前提条件・計算方法をよく理解しないでも計算できるため，誤用も増えている．医学系のメジャージャーナルであるLANCETに投稿されてきた論文191について，統計的な側面についてレビューした結果，半数の論文は問題がなかったが，25%の論文は統計について改訂が必要であり，25%の論文は統計が原因で論文をリジェクトせざるを得なかったことが報告されている (Gore et al. (1992))．著者が経験した中でも，ある医師が200症例の胃癌患者について死亡のリスク因子を解析するにあたり，30数種類の予後因子を同時にモデルに含め，変数選択を行わなかったためリジェクトされ，相談を受けた例がある．また統計学の1つの特徴は結果を計量的に表現することであるが，残念ながら数学の専門教育を受けていない多くの実務家には，複雑な統計量の意味と使い方を正確に理解することは容易ではない．このため実務家の中には，“検定症候群(Statistical Significance Syndrome)”に陥る人がでてくる．統計量の意味をよく考えず，統計解析では検定で有意差がつくかどうかだけを問題とする人達である．

複雑な多変量解析の手法が容易に計算できる環境が整いつつあるが，実務家のためにこれらの手法を誤りなく適切に使用し，結果を解釈する新しい教育の方法論が求められている．実務家は固有技術に裏打ちされたデータに対する直観を持っている．データの内容をよく理解して

* 医学部 薬剤疫学教室：〒113 東京都文京区本郷7-3-1.
** 〒104 東京都中央区勝どき1-13-1 イヌイビル・カチドキ8F.

ない統計学の専門家より，彼らははるかに個別の生データを評価する能力では勝っている．彼らの感覚がデータ解析の中で生かされるような統計ソフトウエア，直観的に統計手法を理解できるための教育の方法論はないものか？ 本稿では著者の東京大学医学部での教育の経験に基づいて，SAS/INSIGHT を用いた視覚的なデータ解析，直感的な統計学の教育の可能性，検定症候群の患者に対する処方箋を追求する．

2. SAS/INSIGHT とは？

SAS/INSIGHT は SAS のプロダクトの1つであり，グラフィカルユーザーインターフェースを駆使して，探索的な解析を行うためのツールである．複数のウインドウ上の様々なグラフを連動させながらマウスだけの操作でメニューを選び，対話的に解析を進めることができる．これまでエディター上にプログラムを記述しなければ解析できなかった SAS のイメージを根本的に変えるものである．特徴を以下に述べる（SAS Institute (1993, 1994)）．

1) EXCEL に類似したスプレッド・シートによってデータの表示・変更・追加・変数の変換を容易に行うことができる．
2) グラフィカルな機能に優れ，各種グラフ（ヒストグラム，箱髭図，層別箱髭図，モザイク図，散布図，ラインプロット，3次元回転プロット）を作成することができる．
3) データシート，グラフが連動して変化する（ブラッシングの機能がある）．
4) Generalized linear model（ポアソン回帰，ロジスティック回帰，重回帰分析）を用いた解析を対話的に行うことができる．
5) グラフ上の観測値の値の確認，除去がダイナミックに実行できる．
6) データの分布形についてパラメトリック（正規，対数正規，指数，ワイブル）およびノンパラメトリックなあてはめを視覚的に行い，また特定の分布があてはまるかを Kolmogorov の検定によって評価することができる．
7) % 点を含めた各種要約統計量を計算する（PROC UNIVARIATE の機能に対応）．
8) 変数間の相関係数を計算し，主成分分析を実行する．
9) 多項式回帰，Kernel 法，スプライン法による曲線のあてはめを視覚的に行う．

3. 東大医学部における学生実習の紹介

東京大学医学部では学生6人を1組として，1週間のコースの情報処理実習を行っている．この実習では乳癌についてのリスク因子を探索することを目的とし，必要な文献で情報を検索し，実際の乳癌患者のカルテから必要なデータを抜き出し，データベース化し，SAS を用いて解析を行い，擬似的に論文を作成させるものである．実習全体の目的は臨床研究とそのレポートを作成するための一連の流れを学生に体験させることである．この中で統計解析の実習が1日とられており，著者の浜田が所属する薬剤疫学講座が担当している．この実習では，1人に1台ずつコンピュータ端末を提供し，乳癌患者の死亡についてのリスク因子を WINDOWS 版 SAS/INSIGHT（リリース 6.11）を用いて統計解析する．学生が入力したデータであるため入力ミスがある可能性が高く，複雑な解析を行う前にデータのクリーニングも行わせている．実習全体の流れは次のようになる．

・WINDOWS 版 SAS の使い方の実習（画面間の移動・大きさの変更，メニューの選択）
・乳癌データの SAS データセット化

・SAS/INSIGHT による野球データの解析
・SAS/INSIGHT による乳癌患者のリスク因子の解析

乳癌データの解析は学生自身に自由に行わせるが，本題に入る前に 1995 年度のプロ野球選手の一軍打者の成績のデータを用いて，SAS/INSIGHT の使い方と統計解析の手順を教官が誘導しながら実習する．例題としてプロ野球データを選択した理由は，計量的なデータが多く，しかも変数の内容について説明しなくても多くの学生が理解しているためである．

データ解析の一般的な手順は次のように考えることができる．

1) 外れ値，異常値のチェック，分布形のチェック，1変数ごとのデータの要約（箱髭図，ヒストグラムを描く，平均，標準偏差（SD）などの要約統計量を計算する）
2) 2変数間の関連の解析（散布図，層別箱髭図，クロス集計表を作成する，相関係数を計算する．群間で違いがあるかを検定する）
3) 多変量解析（重回帰分析，ロジスティック回帰，Cox 回帰，主成分分析，因子分析）

この実習の重要な目的は，3）の解析に進む前に，1），2）の解析を十分に行う必要があること，特に外れ値，異常値はデータにつきものであり，これらのチェックを最初に行う必要があることを学生に認識させることである．データの入力ミス，特に桁間違い等については，データ解析の初期の段階で修正する必要があるし，外れ値はデータ解析の結果に大きな影響を与えるので，生じた原因についてよく調べる必要がある．また分布については，多くの統計手法が正規分布や等分散性を仮定するので，これらの点を確認しておく必要がある．もう1つの目的は，得られた結果に応じて，次にどのような解析を行うべきか（どのようなグラフを描くべきか）学生が考える習慣を付けさせる点にある．具体的にはプロ野球選手データの年俸（年俸）に影響を与える因子を評価することを解析の目的として，次の手順で実習する．

1) 年俸のヒストグラムの作成と分布形の吟味
2) 年俸の箱髭図の作成と外れている個体の同定
3) 年俸の要約統計量の計算
4) 他の変数についてのヒストグラムの作成
5) 打点，ホームラン数，安打数，年俸等の散布図行列の作成，最大の選手の同定
6) 三次元回転プロットの作成
7) 12 球団で層別した箱髭図の作成と，球団ごとの高年俸の選手の同定
8) 重回帰分析による年俸に影響を与える因子の評価

SAS/INSIGHT ではマウスだけの操作でこの一連の解析を対話的に行うことができる．

4. SAS/INSIGHT によるプロ野球データの解析

実習の流れに沿って SAS/INSIGHT による野球データの解析を紹介する．図 1 に SAS/INSIGHT を起動し，データを読み込んだ画面を示した．画面は EXCEL のようなスプレッドシート形式であり，図の左上をみると 318 人について 29 変数あることがわかる．1995 年度に一軍の出場経験があった打者が，12 球団併せて 318 人存在した（日刊スポーツクラブ (1996)）．このうち出場試合数が少ない，いわゆる一軍半の打者は打率等が不安定なので，通算出場試合数が 100 試合以下の選手を除いた 210 人について解析する．このためには編集→オブザーベーション→計算からの除外と，順次メニューを選んでいき，GAME_C（通算出場試合数）＜＝100 を指定する．このような指定を行うと，解析対象となる観測値の左端にマークが付く．ここではチー

318 \ 29	名義変数 TEAM	名義変数 POSITION	ラベル 名義変数 NAME	間隔変数 AGE	間隔変数 YEAR
▲ 59	読売	内野手	落合	43	18
60	読売	内野手	吉岡	25	7
▲ 61	読売	内野手	福王	32	11
62	読売	内野手	大森	29	7
▲ 63	読売	内野手	長嶋	30	9
▲ 64	読売	内野手	緒方	28	10
65	読売	内野手	佐々木	18	0
66	読売	内野手	出口	25	7
▲ 67	読売	外野手	吉村	33	15
▲ 68	読売	外野手	マック	33	2
▲ 69	読売	外野手	岸川	31	13
70	読売	外野手	後藤	27	9
▲ 71	読売	外野手	井上	30	12
▲ 72	読売	外野手	松井	22	4
▲ 73	読売	外野手	広沢	34	12
74	横浜	捕手	大塚	30	12
◆ 75	横浜	捕手	谷繁	26	8
76	横浜	捕手	葉室	30	8

図1. SAS/INSIGHT のデータシート.

ム別にマークを変えており，巨人の選手は▲，横浜の選手は◆というように表示されている．このように SAS/INSIGHT では，ある個体を除いたり，ある条件を満たす個体のみを，解析対象とすることが簡単に指定できる．

1) 年俸のヒストグラムの作成

メニューから解析→棒グラフを指定し，変数に SALARY（推定年俸：単位 万円）を指定することによって図2のヒストグラムが出力される．ヒストグラムをみると年俸が3億6000万円を超える選手も存在するものの，全体としては1億円以下の選手が圧倒的に多く，大きく右にスソを引いた分布であることが分かる．図では左から2番目の棒をクリックすることによって黒く反転させている．SAS/INSIGHT ではウィンドウが連動しており，全ての画面で SALARY がこの範囲にある選手が反転する．

2) 年俸の箱髭図の作成

メニューから解析→箱髭図を指定し，変数に SALARY を指定することによって図3が出力される．この箱髭図では平均 ±SD が菱形で示されている．この図をみるとメディアンに比べて平均がかなり高くなっていることがわかる．また図1に示したように，NAME をラベル変数にしてあるため，観測値をクリックすると，対応する NAME の値が出力される．年俸の高い選手はマックと落合であることがわかる．実際の実習では，年俸の高い選手名を上から10番目まで調べさせている．また観測値をダブルクリックするとウィンドウが開き，その個体についての

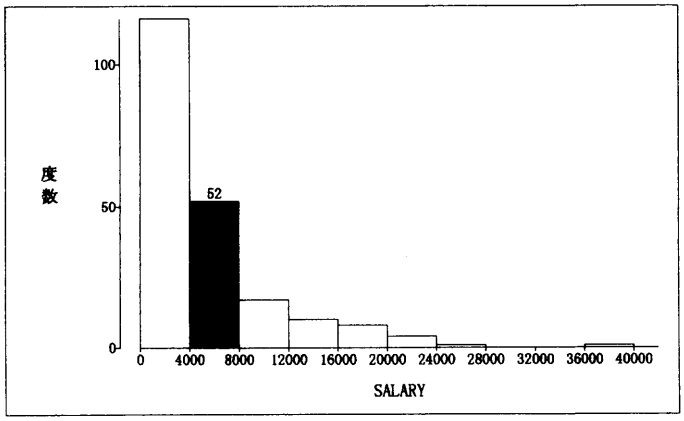

図2. 年俸のヒストグラム.

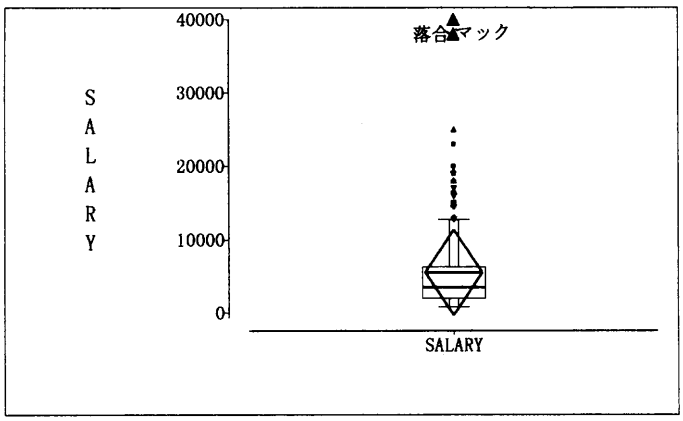

図3. 年俸の箱髭図.

全変数の情報が出力される.

3) 年俸の要約統計量の計算

メニューから解析→分布を指定し,変数にSALARYを指定することによって,図4が出力される.年俸の平均値5521万円に対してSDが5865万円であり,平均を上回るバラツキがあることがわかる.歪度2.7779,尖度10.2636であるから,右に大きくスソを引いていることがわかる.メジアンは平均より2000万円以上低い3500万円,モードは1600万円である.高年俸の選手(最高4億円)に引っ張られて平均は高くなっているが,プロ野球選手とはいえ,必ずしもうらやむような高給取りばかりではなく,多くの選手の年俸は5000万円以下である.

4) その他の変数についてのヒストグラムの作成

この後,実習ではその他の変数についても,グラフを描かせて,分布形を吟味させ,外れ値,異常値がどの個体のものであるかを同定させている.また年俸の高い個体がどのような特徴を持っているかを調べるため,同時に複数の変数のヒストグラムを描かせる.

メニューから解析→棒グラフを指定し,SALARYを含めて95年度の成績等を示す12変数を指定することによって図5の集合ヒストグラムが出力される.複数の変数の分布形について一度に把握することができる.AGE, YEARはほぼ正規分布に近く,GAME95(95年度出場試合数),DASUU95(95年度打数)は一様分布に近い,これに対し残りの95年度の成績を表す

SALARY

モーメント			
N	210.0000	重みの合計	210.0000
平均値	5520.9286	合計	1159395.00
標準偏差	5864.5946	分散	34393469.6
歪度	2.7779	尖度	10.2636
無修正平方和	1.359E+10	修正済平方和	7.188E+09
変動係数	106.2248	標準誤差	404.6955

パーセンタイル値			
100% 最大値	40000.0000	99.0%	25000.0000
75% Q3	6300.0000	97.5%	20000.0000
50% メジアン	3500.0000	95.0%	18000.0000
25% Q1	2000.0000	90.0%	13000.0000
0% 最小値	840.0000	10.0%	1400.0000
範囲	39160.0000	5.0%	1230.0000
Q3-Q1	4300.0000	2.5%	1050.0000
モード	1600.0000	1.0%	900.0000

図4. 年俸の要約統計量.

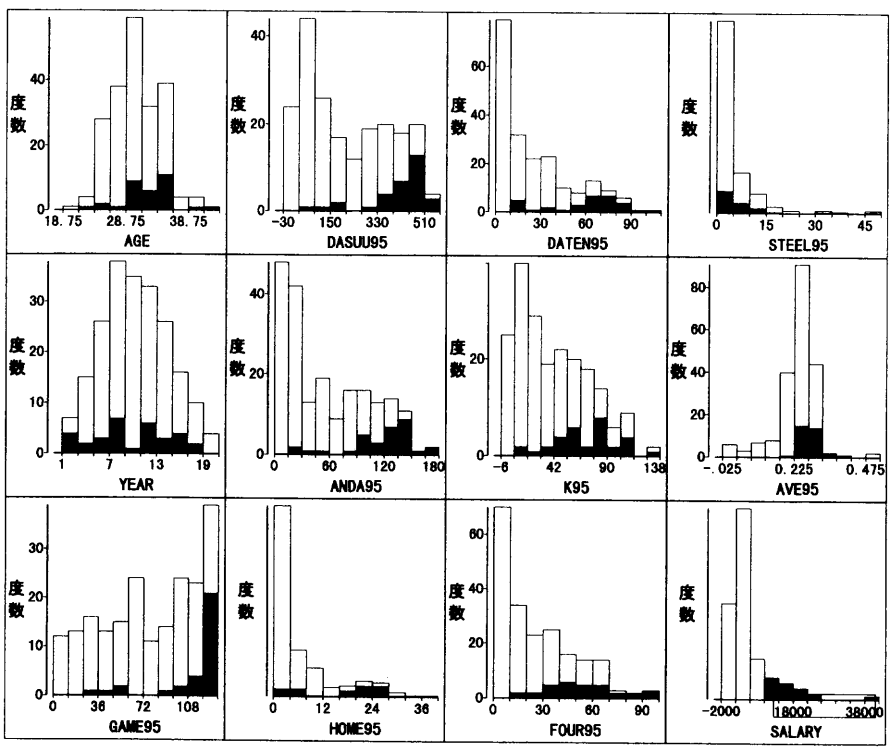

図5. 集合ヒストグラム.

変数は，AVE95 (95年度打率) を除いて右にすそを引いた分布であることがわかる．図では更に右下にあるSALARYのヒストグラムで，年俸の高い選手をドラッグさせ黒く反転させている．AGE, YEARについては全体の分布と年俸の高いサブグループの分布がほぼ等しいが，残りの変数については多少の例外はあるものの，年俸の高い選手は基本的には95年度の成績もよ

いことが読みとれる．

5) 散布図行列の作成

実習では，より多くの変数について散布図を作成し，変数間の関連，特徴的な個体を調べているが，ここでは紙面の都合上 DATEN95 (95 年度打点)，DATEN_C (通算打点)，SALARY の 3 変数について散布図行列を作成する．

メニューから解析→散布図を指定し，DATEN95, DATEN_C, SALARY を X 軸と Y 軸の変数に両方指定する．結果は図 6 のようになる．DATEN95 と DATEN_C の間には顕著な関連は認められないが，DATEN95 と SALARY, DATEN_C と SALARY の間には，正の相関が認められる．また外れ値と思われる値が見られる．ここでは編集→オブザーベーション→ラベルありと，順次メニューを選んでいき，SALARY＞＝23000 を指定することにより高年俸の選手の名前を出力させている．DATEN_C と SALARY の散布図（下段中央）では落合が相関を上げる方向，3 人の外人選手が相関を下げる方向で外れていることが判る．落合は通算打点，年俸とも抜群であるのに対し，外人選手は大リーグを経験してから来日するため，日本での通算成績はそれほどではないが，高年俸になる．特にマック，ジャクソンは来日 2 年目であり，1 年目の成績がそのまま通算成績になっている．

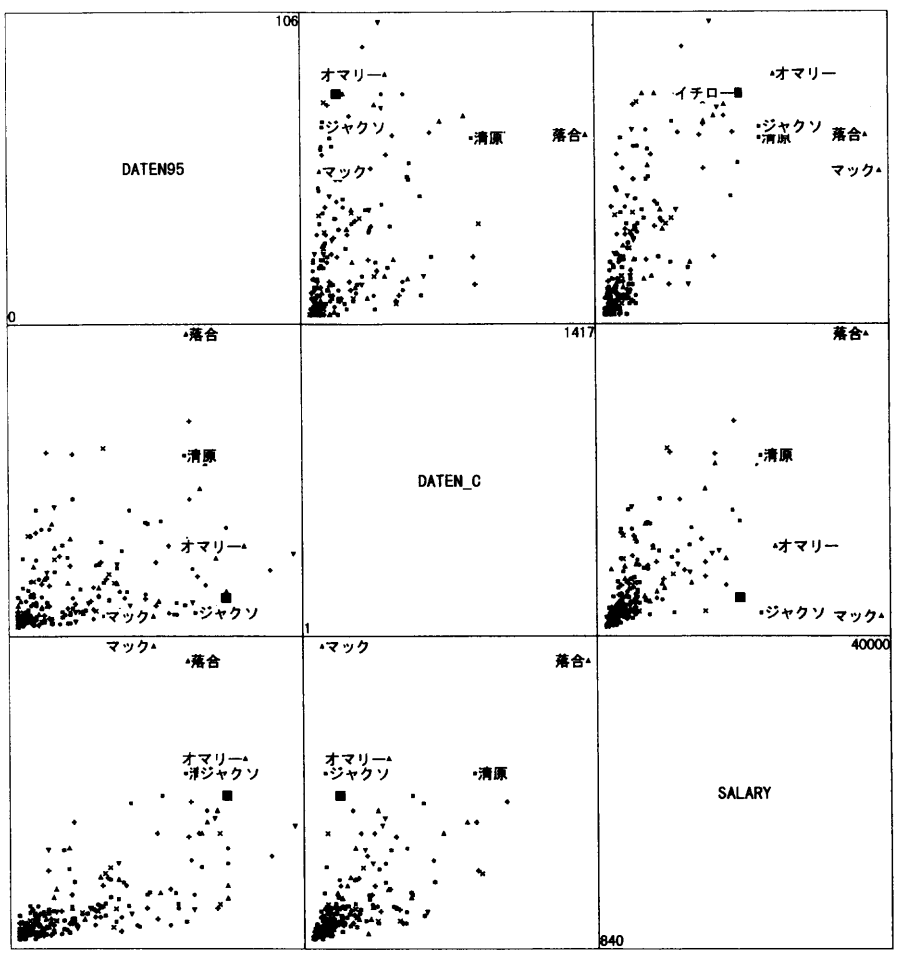

図 6. 散布図行列．

6) 3次元回転プロットの作成

メニューから解析→回転プロットを指定し，X, Y, Z 軸の変数としてそれぞれDATEN95, DATEN_C, SALARY を指定する．結果は図7のようになる．画面上ではプロットを回転させることによって立体的に3変数間の関連を把握することができる．紙面上ではうまく表現できないのが残念である．3次元空間上でも落合とマックが異なった方向で外れていることが確認できる．落合は通算成績と年俸の双方が抜群，マックは年俸は高いが通算成績は低い．

7) 層別箱髭図の作成

どのチームの選手が高いかを調べるため，チームで層別した年俸の箱髭図を作成してみる．メニューから解析→箱髭図を指定し，SALARY を Y 軸，TEAM を X 軸の変数に指定する．結果を図8に示す．図1と同様に平均 ±SD が菱形で示されている．読売の平均年俸が圧倒的に

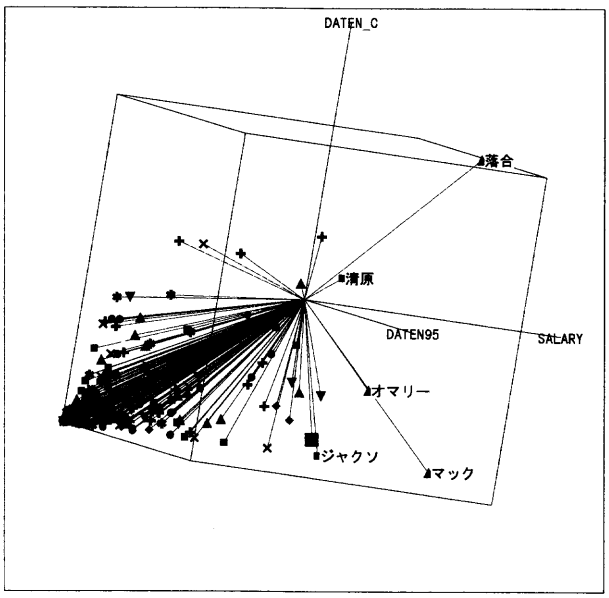

図7. 3次元回転プロット．

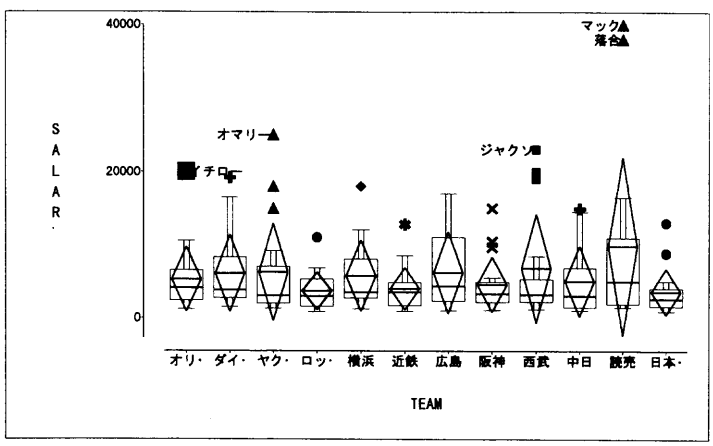

図8. 層別箱髭図．

高いことがわかる．図には示されていないがクリックすると実際の値が判る．読売の平均年俸は9795万円，これは最も低い日本ハムの3586万円と比べて6000万円以上，2番目に年俸の高い西武の6761万円と比べても3000万以上も高い．しかしながらマック（4億円）と落合（3億8000万円）の2つの外れ値を含んでおり，この2人だけで実に7億8000万円のお金を払っている．読売のメジアンは4950万円とほぼ平均9795万円の半分であり，読売はメジアンでも12球団トップであるものの，2位の広島が4350万円であり，外人と外様の選手を除いた生え抜きの選手に限れば，それほど高い年俸を貰っているわけではない．メジアンで比較する限りは，チーム間で大きな年俸の差は認められない．

8)　年俸を目的変数とした重回帰分析

実習では，次に年俸に影響を与える因子を重回帰分析によって評価する．メニューから解析→回帰を指定し目的変数としてSALARYを指定し，説明変数の方は，今まで得られた結果と，野球に関する知識から，学生に自由に選ばせている．ここでは，1つの例として，過去の実績を表す変数としてDATEN_C，現在の能力を表す変数としてDATEN95，そしてTEAMをモ

当てはめの要約

応答変数の平均	5632.3775	寄与率	0.6432
誤差の標準偏差	3650.9575	自由度調整済み寄与率	0.6188

タイプ III 検定

変動因	自由度	平方和	平均平方	F 統計量	P 値 (F)
TEAM	11	304063566	27642142.4	2.0738	0.0240
DATEN95	1	1.494E+09	1.494E+09	112.0991	0.0001
DATEN_C	1	924771062	924771062	69.3778	0.0001

パラメータ推定値

変数	TEAM	自由度	推定値	標準誤差	t 統計量	P 値 (t)	トレランス	VIF
INTERCEPT		1	-455.5570	946.6438	-0.4812	0.6309	.	0
TEAM	オリック	1	795.0119	1240.4959	0.6409	0.5224	0.5027	1.9892
	ダイエー	1	648.8478	1311.2026	0.4948	0.6213	0.5258	1.9018
	ヤクルト	1	1104.6541	1297.1668	0.8516	0.3955	0.5372	1.8613
	ロッテ	1	-122.7631	1258.3861	-0.0976	0.9224	0.5129	1.9497
	横浜	1	-34.5637	1342.8062	-0.0257	0.9795	0.5669	1.7639
	近鉄	1	422.5542	1222.2461	0.3457	0.7299	0.4736	2.1113
	広島	1	877.5299	1339.8830	0.6549	0.5133	0.5694	1.7562
	阪神	1	413.2431	1275.0750	0.3241	0.7462	0.5261	1.9007
	西武	1	2729.9335	1229.2918	2.2207	0.0276	0.4890	2.0451
	中日	1	495.7756	1242.3847	0.3991	0.6903	0.5012	1.9952
	読売	1	4555.5439	1360.5211	3.3484	0.0010	0.5523	1.8107
	日本ハム	0	0	.	.	.	.	.
DATEN95		1	120.2835	11.3607	10.5877	0.0001	0.8138	1.2288
DATEN_C		1	11.2403	1.3495	8.3293	0.0001	0.8085	1.2368

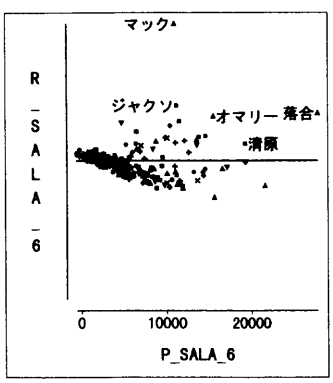

図9.　重回帰分析の結果1．

デルに取り込むことにする．結果を図9に示す．

モデル全体での寄与率は64%であり，3要因とも5%の水準で有意となる．TEAMについては日本ハムを基準として各チームがどの程度，年俸が高いかをダミー変数によってモデル化している．読売の年俸が最大で，日本ハムに比べて4556万円高いことがわかる．DATEN95, DATEN_Cのパラメータ推定値が120，11であり，95年の打点が1点増えるごとに120万円，通算打点が1点増える度に11万円，年俸が増加することがわかる．また一番下に示されているのが，横軸にモデルによる予測値，縦軸に残差をとったプロットである．

残差が正の方向で大きい（成績で予測されるよりは年俸が高い）選手はマックとジャクソンであり，特にマックは残差が2億円を超えている．この2人は先に述べたように来日2年目であり日本での実績が少ないので，他の選手と同じ土俵で年俸を比較することには無理がある．またこのクラスの大リーガーは毎年1人来日するかしないかであり，かなり特殊なケースでもある．そこでこの2選手を除いて解析をやり直してみる．グラフ上で観測値をクリックしてハイライトした後で，編集→オブザーベーション→計算からの除外を行うことによって取り除く

当てはめの要約			
応答変数の平均	5376.2624	寄与率	0.7438
誤差の標準偏差	2763.0480	自由度調整済み寄与率	0.7261

タイプ III 検定					
変動因	自由度	平方和	平均平方	F 統計量	P 値 (F)
TEAM	11	91673869.6	8333988.15	1.0916	0.3700
DATEN95	1	1.043E+09	1.043E+09	136.5954	0.0001
DATEN_C	1	1.240E+09	1.240E+09	162.4085	0.0001

パラメータ推定値								
変数	TEAM	自由度	推定値	標準誤差	t 統計量	P 値 (t)	トレランス	VIF
INTERCEPT		1	−247.5171	716.8391	−0.3453	0.7303	.	0
TEAM	オリック	1	783.8377	938.8113	0.8349	0.4048	0.5032	1.9872
	ダイエー	1	308.5817	992.7596	0.3108	0.7563	0.5258	1.9019
	ヤクルト	1	1065.4952	981.7087	1.0853	0.2792	0.5377	1.8598
	ロッテ	1	−296.7727	952.4767	−0.3116	0.7557	0.5133	1.9484
	横浜	1	71.0898	1016.3155	0.0699	0.9443	0.5673	1.7628
	近鉄	1	132.4891	925.3762	0.1432	0.8863	0.4738	2.1106
	広島	1	894.9131	1014.0366	0.8825	0.3786	0.5698	1.7550
	阪神	1	266.2194	965.0644	0.2759	0.7830	0.5265	1.8994
	西武	1	1884.4084	943.0199	1.9983	0.0471	0.4987	2.0050
	中日	1	369.9604	940.2997	0.3934	0.6944	0.5016	1.9935
	読売	1	1939.1237	1054.9707	1.8381	0.0676	0.5640	1.7732
	日本ハム	0	0	.	.	.	.	.
DATEN95		1	102.5529	8.7747	11.6874	0.0001	0.7976	1.2537
DATEN_C		1	13.1924	1.0352	12.7440	0.0001	0.7898	1.2661

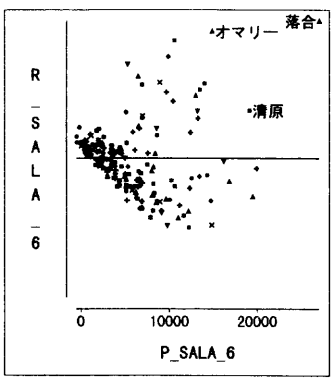

図10. 重回帰分析の結果2．

ことができる．結果は図10のようになる．寄与率は74%と10%増大する．またTEAMのp値は0.3700と有意でなくなる．読売と西武は大リーガーには破格の高給を用意するが，全体として12球団間で給与体系に大きな差はないといえる．

さて一通りの解析はしてみたが，もちろん解析が終了したわけではない．例えば次のような点を考慮してモデルは改善されるべきである．

- 層別解析：外人選手と日本人選手の年俸を同じモデルで考えるには無理があるかもしれない．
- 変数の追加：このモデルでは立浪（中日），川相（巨人）の残差が正の方向で大きくなる．すなわち1，2番タイプの巧守好走の選手の評価が低い．この点を改善するためには盗塁数あるいは守備能力を表す変数をモデルに含めるべきかもしれない．
- 変数変換：年俸は大きく右にスソを引いた分布であるが，対数変換を行うことによってほぼ正規分布に近くなる．したがって対数変換後の年俸（変換前の年俸について各因子が等比的に影響を及ぼすモデル）について重回帰分析を行う方が，より適切であるかもしれない．

このようなモデル改善のためのテクニックが，SAS/INSIGHTでは簡単に実行できる．実習では時間の許す限りモデルのあてはめとその診断を繰り返し，より良いモデルを探索させる．その後乳癌のリスク因子については，データクリーニングを含め，学生自らが考えながら実習する．

5. SAS/INSIGHTによる教育の利点と問題点

前節で述べたような手順で学生実習を行い，学生にはたいへん好評である．SAS/INSIGHTによる実習の利点について，学生の印象をふまえてまとめてみた．

1) 統計学の印象を“数式を使って難しそう”から，“おもしろそう”に変えることができる．

2) プログラミングを行わずにマウスだけの操作で解析ができるため，習得が速い．野球データでSAS/INSIGHTの使い方を理解させると，本題の乳癌データの解析については，学生は試行錯誤しながら，ある程度までは自分自身で解析することが可能である．

3) 複雑な統計解析を行う前に，様々なグラフを描かせることによって，データの視覚的な吟味の重要性，データ解析の手順を理解させることができる．

4) SAS/INSIGHTではブラッシング，変数ラベルの表示等が容易にできるため，観測値の特定が簡単である．この機能を利用することにより，外れ値，異常値について検討することの重要性を認識させることができる．学生は卒業後，臨床データを扱うことになる．外れ値がどの患者のものかを特定し，他の測定値を調べ，どのような原因で外れ値が生じたかを探求することが必要である．

5) グラフ化，変数の除去，変数の変換等が簡単に実行でき，これらのテクニックを駆使して，コンピュータと対話しながら探索的な解析を行う重要性と楽しさを体験することにより，疑問・仮説を持ったときに，どのような解析をすべきか考える習慣を付けることができる．

6) 統計学の教育

著者達は，次の例を用いて偏回帰係数の意味を実習させている．最初に，野球選手の年俸を目的変数として，プロ入りしてからの年数を説明変数として単回帰分析を行う．回帰係数は正で有意になる．これに対し，通算打点を説明変数として加えると，今度は年数の回帰係数は負で有意となる．この結果の意味について学生に考えさせる．10年で1000打点を記録した選手

と，20年で1000打点に到達した選手では，どちらの年俸を上げるべきか？ 当然10年で1000打点の選手の方が1年当たりの成績は良いので，年俸は高くすべきである．偏回帰係数は他の変数の値をそろえたときのある変数の影響を示すものである．

また乳癌のリスク因子の解析の中では，共変量による調整を行った解析として，層別解析と数学モデルによる解析の双方を実習している．複数の要因で層別した箱髭図等で交絡，交互作用の影響を認識した上で，モデルによる解析で交絡の影響，交互作用を定量的に評価する．統計学的に重要な概念を，実際のデータに基づいて，視覚と統計量の双方を照らし合わせながら学ぶことは，より高レベルの統計教育の手段としてたいへん有効である．このようにSAS/INSIGHTによる実習は，単に記述統計のみならず，統計量，統計学的概念を理解させる上でも有効である．

逆にこのような教育の問題点をまとめてみた．

1) 講義形式では効果的な教育を行うことができず，実習が必須となる．操作法の細かい点を正確に実行させるためにはテューターが必要である．東大の実習では6人の学生に対して，教官が2〜3人ついている．またSASのインストールが可能な高性能パソコンが必要であり，人的資源とコンピュータ環境についてかなりの投資が必要である．
2) 安易に観測値の除去等を行うことができるため，後知恵的な解析を奨励することにつながりかねない．探索的な解析と確証的な解析の違い等は別に機会を設けて教育する必要がある．
3) 良いテキストを作成するのが困難である．画面ごとに細かく説明していくと，すぐに膨大な量になってしまうし，なかなかSAS/INSIGHTの魅力を文章によって表現することは困難である．将来的にはビデオテキスト等の作成を検討する必要があるだろう．
4) 解析に夢中になって，長時間ディスプレイを見続けるため，学生の中には視覚に変調をきたすものがいる．また視力が極端に低い学生は実習が困難である．定期的に休憩はとるようにしているが，個人差が大きい．

本稿ではSAS/INSIGHTを用いた統計実習を紹介したが，著者がSAS/INSIGHTを用いている理由は，著者の所属する環境で利用しやすかったためであり，同様の教育はJMP，STATVIEW等を用いても可能であると考えられる．また今後もより使い易い統計パッケージは生まれてくることが当然予想される．

これまで統計学の教育は講義形式に演習をおりまぜながら，一通りのことを教育するのがオーソドックスな方法であった．しかし生物系の学生の統計学の講義に対する印象を聞くと，必要性をよく理解していなかったため，ほとんど覚えていない，あるいは数式をたくさん見せられたため統計嫌いになったというのが多い．もちろん数式を用いた統計教育を否定するわけではないが，生物系の学生には，物理・化学系と比較して，より実証的色彩を深めた教育が必要であり，著者の経験では，実際のデータに触れる卒業研究前は，医薬系の学生は統計学の必要性がよく理解できていないため，体系だった教育は非効率的である．むしろ学生に興味が持てるデータを提供し，統計パッケージを用いて何ができるのかを体験させ，統計学に対するアレルギーを持たせないことの方がより重要である．さらに偏回帰係数の説明の例で示したように，より複雑な統計量の意味・概念についても，工夫によっては実習させることは可能であり，インターラクティブなソフトウエアによる実習は，新しい統計教育の手段として，魅力いっぱいである．

参 考 文 献

Gore, S. M., Jones, G. and Thompson, S. G. (1992). The Lancet's statistical review process: areas for improvement by authors, *Lancet*, **340**, 100-102.

日刊スポーツクラブ (1996). 『96 プロ野球選手写真名鑑』, 日刊スポーツ出版社, 東京.

SAS Institute (1993). *SAS/INSIGHT® User's Guide Version 6 Second Edition*, SAS Institute, Cary, North Carolina.

SAS Institute (1994). *SAS/INSIGHT® Software Changes and Enhancement Release 6.10*, SAS Institute, Cary, North Carolina.

 Proceedings of the Institute of Statistical Mathematics Vol. 44, No. 2, 149-162 (1996)

Education Using SAS/INSIGHT

Chikuma Hamada
(Faculty of Medicine, University of Tokyo)
Junji Kishimoto
(SAS Institute Japan Ltd.)

SAS/INSIGHT is an interactive software for data exploration and analysis. With it, we can explore data through a variety of interactive graphs (histogram, box-plot, scatter plot, rotating plot and so on) and analyses (multiple regression, logistic regression and principal component analysis) linked across multiple windows. In this software, all operations we want to perform are listed in menus. We can choose it with the mouse.

We have been teaching the medical students in University of Tokyo for several years using SAS/INSIGHT. The detail of statistical course using computer systems is introduced in this article. According to our experience, the possibilities of new methodology for education of statistics is discussed.

Key words: SAS/INSIGHT, education, exploratory analysis, interactive analysis.