ＳＰＳＳ／ＰＣ＋による解析手法　そのＩ

－探索的データ解析について－

藤　田　剛　志・小野寺　孝　義

はじめに

探索的データ解析(Exploratory Data Analysis) とは、調査および実験の結果、収集されたデータに含まれる特徴や構造を、字義通り、探索的に探り出すことを目的として、J.W.Tukeyを中心としたデータ解析学派と呼ばれる人々によって開発された一連のデータ解析の手法を指している。

近年、コンピュータの高性能化、低価格化および統計パッケージソフトウェアの発達により、以前に比べ手軽にデータ解析を行うことが可能になった。このことは、統計学にあまり精通していない人々にとって、研究方法の幅を広げる福音をもたらした。と同時に、どのようなデータであろうとも、コンピュータによって統計処理を行えば信頼できる結論を得ることができると安易に考える弊害が生じるようになった。

調査あるいは実験によって収集されたデータには歪みが含まれるものである。誤ったあるいは不十分な調査や実験ではもちろんのこと、適切な調査や実験を行ったとしても、収集されるデータには何らかの歪みが含まれる。このような歪みを含んだデータでは、どのように高度な統計手法を用いても、信頼できる結論を得ることはできない。それ故に、データの分析を開始する前に、データを注意深く探索することが重要になる。

探索的データ解析には、異常なデータを発見したり、異常なデータに対する頑健な(robust)推定量を求めたり、データ分布の左右の歪み方の違いや他の標準的分布との類似性を積極的に検出していくための様々な技法がある。データが得られたならば、すぐに分析を開始するのではなく、データを探索的に解析することによって、仮説の設定、変更、検定手法の再検討を行うことが大切なのである。

本稿では、パソコン用の汎用統計パッケージソフトウェアであるSPSS／PC＋を用いた探索的データ解析の主要な技法を解説することによって、探索的データ解析がどのように役立つかを検討することにする。

1．SPSS／PC＋によるプログラム書式

(1) SPSSについて

SPSSとは1968年にスタンフォード大学で開発された社会科学のための統計解析パッケージである。SPSSの名称はStatistical Package for Social Sciencesの頭文字をとったものである。1978年には、その名称はSuperior Performing Software Systemに変更され、社会科学だけではなく様々な研究分野で利用できるプログラム群を備えたパッケージとなった。日本では1973年から京都大学で稼働をはじめ、以後大型計算機センターを持つ全国の大学で利用されるようになった。10版を数えた1983年からは、SPSS^{x}と呼ばれるようになった。現在の4版に至り、再び、元のSPSSの名称に戻っている。

SPSS／PC＋とはパソコン用のSPSSである。

最初は、IBM PC／XT AT 用として開発された。その後、Macintosh や NEC の PC9801用などに移植された。今年になって Windows版も発売された。現在、IBM PC／AT や Macintosh用では4版、PC9801用では3版、Windows版では５版を利用することができる。

本稿では、SPSS／PC＋ 3.0J 版（J は日本語版の意味）を用いている。SPSS／PC＋3.0J 版で利用することができる主な解析手法としては、

分散分析、クラスター分析、相関係数の算出と検定（順位相関を含む）、クロス表作成とカイ自乗検定、記述統計、判別分析、探索的データ解析、因子分析と主成分分析、対数線形モデル、多変量分散分析（正準相関分析を含む）、グループ別平均値比較、ノンパラメトリック検定、ｔ検定、多重比較ｔ検定、単回帰と重回帰分析（残差分析を含む）、非線形回帰分析、信頼性係数解析、各種の時系列解析、多重回答分析

が含まれる。これらの代表的な解析手法には、さらに、下位の解析、検定が包含されている。こうした豊富な解析手法に加えて、SPSS／PC＋は簡単な図形出力、レポート作成機能を備えている。

(2) 探索的データ解析のプログラム書式

SPSS／PC＋による探索的データ解析は、サブプログラム EXAMINE によって実行される。EXAMINE は、収集されたデータの分布の特徴を視覚的に表示し、正規性や分散の等質性を検定し、頑健な推定量を求めることができる。このような機能の他に、EXAMINE には、記述統計量を求めるサブプログラム DESCRIPTIVES と頻度分布を求める FREQUENCIES の機能が包含されている。

EXAMINE のプログラム書式を図１に示す。

```
EXAMINE  VARIABLES=変数リスト
           [[BY 変数リスト]  [変数名  BY  変数名]]
[/COMPARE={GROUPS** }]
          {VARIABLES}
[/SCALE=[{PLOTWISE**}]
         {UNIFORM}
[/ID={$CASENUM**}]
     {変数名    }
[/FREQUENCIES [FROM(初期値)] [BY(増分)]]
[/PERCENTILES=[{HAVERAGE}]  [NONE]]
               {WAVERAGE}
               {ROUND   }
               {AEMPIRICAL}
               {EMPIRICAL}
[/PLOT=[STEMLEAF**] [BOXPLOT] [NPPLOT]
       [SPREADLEVEL(値)] [HISTOGRAM] [{ALL }]]
                                     {NONE}
[/STATISTICS=[DESCRIPTIVES**]
       [EXTREME({5})] [{ALL }]]
                {N}    {NONE}
[/MESTIMATOR=[{NONE**}]
              {ALL}
       [HUBER({1.339})]
              {C    }
       [ANDREW{1.34})]
              {C    }
       [HAMPEL({1.7, 3.4, 8.5})]
               {a,   b,   c  }
       [TUKEY{4.685})]]
             {C    }
[/MISSING={LISTWISE**}    [INCLUDE]]
          {REPORT  }
          {PAIRWISE}
```

{ }はオプションを表し、その中から一つを選ぶことを示している。

[ ]は、そのサブコマンドあるいはキーワードが省略可能であることを表す。

変数リストとある場合は、１つ以上の変数名を並べることができる。変数名とある場合は、変数名を１つだけ記入する。

**はサブコマンドが省略されたときのデフォルトを意味する。

図１　EXAMINEのコマンドと一般書式

VARIABLESサブコマンドは、解析の対象となる変数リストと変数の値の組み合わせに基づいて構成されるセルを指定する。唯一必須のサブコマンドである。COMPAREは、箱型図の表示法を制御するサブコマンドである。SCALEサブコマンドは、箱型図、幹葉表示、ヒストグラムにおいて、各セルとも同じ目盛りを用いて表示するかどうかを制御するサブコマンドである。FREQUENCIESサブコマンドは、度数分布表の出力とその出力書式を制御する。PERCENTILESサブコマンドはパーセンタイルの計算方法と区切り点を指定するためのものである。PLOTサブコマンドはプロット出力を制御し、幹葉表示、ヒストグラム、箱型図、Levene統計量付きの散布度対水準プロット、関連統計量付きの正規及び傾向化除去確率プロットが利用できる。基本的な記述統計量とM推定量の出力は、それぞれSTATISTICSサブコマンドとMESTIMATORサブコマンドによって制御される。

(3)仮想データ

表1は、年齢、最高血圧(mmHg)、血液中の総コレステロール(mg／100ml)に関する女性60人の仮想データを示したものである。年齢をAGE、最高血圧値をBPH、総コレステロールをCHOLという変数名で表す。以下、このデータを用いて、探索的データ解析を行っていくことにしよう。

表1　仮想データ（女性60名）

番号	年齢	最高血圧	総コレステロール	番号	年齢	最高血圧	総コレステロール	番号	年齢	最高血圧	総コレステロール
1	49	144	136	21	63	151	225	41	75	138	259
2	56	131	261	22	33	129	121	42	70	183	172
3	65	128	257	23	34	139	159	43	59	151	190
4	45	126	242	24	45	115	258	44	52	96	123
5	51	132	258	25	71	132	284	45	55	161	286
6	77	134	142	26	53	172	236	46	38	125	153
7	82	149	119	27	62	157	169	47	62	152	220
8	72	148	243	28	46	137	111	48	48	171	213
9	70	145	142	29	36	128	132	49	41	124	139
10	47	122	171	30	47	127	169	50	58	122	121
11	57	130	217	31	32	165	197	51	62	125	399
12	44	142	252	32	34	98	129	52	34	123	136
13	32	130	147	33	45	99	98	53	44	119	84
14	65	143	261	34	62	122	273	54	75	115	184
15	59	150	149	35	58	114	173	55	36	103	119
16	69	153	209	36	65	135	208	56	78	126	157
17	72	147	174	37	58	109	165	57	63	136	218
18	62	203	294	38	71	108	188	58	53	102	92
19	36	116	176	39	60	160	269	59	39	134	178
20	60	140	176	40	80	126	128	60	47	122	142

変数名　番号＝ID、年齢＝AGE、最高血圧値(mmHg)＝BPH
総コレステロール(mg/100ml)＝CHOL

2.データの表示方法

データから何らかの構造を探り出すためには、データの分布状態を視覚的に吟味することが有効である。SPSS／PC＋の探索的データ解析は、ヒストグラム(histogram)、幹葉表示(stem-and-leaf display)、箱型図（boxplot）によって、データの分布状態を表示することができる。

(1)ヒストグラム

データの分布を表示する方法の中で、最も古典的で、誰もがまず思い浮かべるのはヒストグラムであろう。SPSS／PC＋を用いて、最高血圧値(BPH)のヒストグラムを表示するには、次のように指定する。

```
EXAMINE  VARIABLES=BPH
        /PLOT=HISTOGRAM.
```

```
Frequency    Bin Center

    3.00          95    ***
    4.00         105    ****
    5.00         115    *****
   15.00         125    ***************
   12.00         135    ************
    8.00         145    ********
    6.00         155    ******
    3.00         165    ***
    2.00         175    **
    1.00         185    *
    1.00   Extremes    *

Bin width :      10
Each star:        1 case(s)
```

図2　最高血圧値のヒストグラム

その結果、図2のような、ヒストグラムが表示される。Frequency という見出しのついた最初の列には、それぞれの階級内に含まれるデータ数が示されている。Bin Center と記載された2番目の列には、階級の代表として階級値が示されている。他の値よりはるかに大きい、あるいははるかに小さいデータには Extremes というラベルが付けられる。Bin width は、各階級の幅を表している。この場合10mmHg の幅で階級が設定されている。この図から、最高血圧が120～130mmHg の人が多いことが分かる。

ヒストグラムは、データを簡潔に表示する一つの手段である。しかし、次のような問題点が指摘されている。

① データの数が少なくとも50、なるべくならば100以上ないと、分布のパターンを見ることができない。

② 同一のデータでも、階級の分け方によって、分布のパターンが異なる可能性がある。

③ 階級内のデータの偏りや外れ値などのデータの動向が覆い隠される可能性がある。

このような欠点を持つヒストグラムに代わって、用いられるようになったのが次に示す幹葉表示である。

(2)幹葉表示

幹葉表示は、ヒストグラムのようにデータ分布の形状を表示すると同時に、ヒストグラムでは表示することができないデータのすべての数値情報を図示することができる。いわば、表とグラフの長所を組み合わせたデータの表示方法であるといえる。SPSS／PC＋を用いて、上の例の最高血圧値の幹葉表示を図示するためには、次のように指定する。

```
EXAMINE  VARIABLES=BPH
        /PLOT=STEMLEAF.
```

その結果、図3のような幹葉表示が図示される。ヒストグラムと同様に、各行の長さは、特定の階級に含まれるデータの数に対応している。しかし、ヒストグラムのようにすべてのデータ

```
Frequency     Stem &  Leaf

     3.00       9   .  689
     4.00      10   .  2389
     5.00      11   .  45569
    15.00      12   .  222234556667889
    12.00      13   .  001224456789
     8.00      14   .  02345789
     6.00      15   .  011237
     3.00      16   .  015
     2.00      17   .  12
     1.00      18   .  3
     1.00 Extremes    (203)

Stem width:     10
Each leaf:       1 case(s)
```

図3　最高血圧値の幹葉表示

を＊で表す代わりに、幹葉表示では、実際のデータに対応する数値で表示されている。すなわち、幹葉表示では、データを幹(stem)と呼ばれる先行桁と葉(leaf)と呼ばれる後行桁の二つの要素に分けて表示する。幹はデータが表示のどの行に含まれるのかを決定するのものであり、葉は個々のデータを同定するために適切な幹のそばに書き込まれている。たとえば、図3では、最高血圧値125というデータは '12' という幹と '5' という葉に分けて表示される。

幹の数が少なく、分布の形状がはっきりしていない場合には、各幹の間隔を狭くし、分割して表示することが有効である。たとえば、図4の総コレステロールの幹葉表示を見てみよう。この図では、100の位で幹が設定されており、それが2行に分けて表示されている。幹を2分したときには、0から4の葉を持つデータは＊で、5から9の葉を持つデータは．で示された幹に表示される。また、葉の部分が2桁以上になる場合には、最初の1桁だけを残して切り捨

```
Frequency     Stem &  Leaf

     3.00       0   .  899
    17.00       1   *  11122222333344444
    17.00       1   .  55566677777778899
    10.00       2   *  0011122344
    12.00       2   .  555556667889
     1.00 Extremes    (399)

Stem width:     100
Each leaf:       1 case(s)
```

図4　総コレステロールの幹葉表示

てられる。この例では、一の位の値が切り捨てられている。したがって、総コレステロール121というデータは、'1＊' という幹と '2' という葉に分けて表示される。幹を5つに分けるときには、＊（0と1の葉に対して）、t（2と3に対して）、f（4と5に対して）、s（6と7に対して）、．（8と9に対して）によって、幹を区別する。

幹葉表示は個々のデータの値が図示されているために、よりよい階級設定の足がかりを提供することができるという長所を持っている。しかし、データ数が300を越える場合には、スペース的に表示が困難になるという欠点がある。

(3)箱型図

ヒストグラム、幹葉表示は分布の形状を把握することに主眼を置いたものである。一方、箱型図は分布の主要特性ならびに分布の裾の状態を視覚的に、容易に把握するために考案された。元々は、箱ヒゲ図(box-and-whisker plot)と呼ばれていたが、現在は箱型図と短縮されている。データが多すぎて、完全な幹葉表示を報告することができない場合、要約統計量（中央値、上・下ヒンジ、最大値、最小値）を報告するには、箱型図は特に便利な手段となる。

通常、分布の状態を示す尺度として、平均値、標準偏差が条件反射のように用いられている。しかし、平均値、標準偏差の統計量は、分布の裾にある少数のデータの影響を大きく受ける。たとえば、次のようなデータで考えてみよう。

{20　15　30　25　40　35　10　45　50　30}

これらのデータの平均値は23.0で、標準偏差は12.9となる。しかし、サンプリングによっては、最大値が50の代わりに、150となるかもしれない。この場合、平均値は40.0、標準偏差は40.1となる。この例は多少極端であるが、平均値と標準偏差が1つの異常なデータによって、影響されやすいことを示している。

そこで、箱型図では、分布の裾にある少数のデータの影響を受けることの少ない中央値(median)とヒンジ(hinge)と呼ばれる統計量を用いてデータを図示するよう工夫されている。図5に示すような箱型図を作成するためには、次の手順を踏む。

① データを小さい方から大きい方に順に並べ換える。

② 中央値を求める。中央値は、もしデータ数nが奇数であるならば(n+1)／2番目の値、もしnが偶数ならばn／2番目とn／2+1番目の値の平均値となる。

③ データ全体を、nが奇数であれば、{最小値から中央値}、{中央値から最大値}の二つに（中央値は重複する）、偶数であれば、{最小値からn／2番目の値}、{n／2+1の値から最大値}の二つに分けて、それぞれの中央値を求める。小さい方を下ヒンジ(lower hinge)、大きい方を上ヒンジ(upper hinge)という。

④ 上下のヒンジの差をとってヒンジ散布度(hinge spread)を求める。

⑤ 下ヒンジから上ヒンジまでの区域を箱で表し、箱の中に中央値を示す＊を入れる。

⑥ 次の式で定義される二つの内境界点(inner fence)を求める。

下ヒンジ－1.5×ヒンジ散布度

上ヒンジ＋1.5×ヒンジ散布度

内境界点に最も近い内側のデータを隣接値(adjacent value)と呼び、ヒンジから隣接値までヒゲ（┴）を伸ばす。

⑦ 次の式で定義される二つの外境界点(outer fence)を求める。

下ヒンジ－3×ヒンジ散布度

上ヒンジ＋3×ヒンジ散布度

⑧ 内境界点よりも外側にあり、かつ外境界点の内側にあるデータを離れ値（outside value)と呼び、(O)で表示する。外境界点の外側にあるデータを飛び離れ値(far out value)と呼び、(E)で表示する。離れ値と飛び離れ値は、あわせて外れ値(outlier)と呼ばれている。

SPSS／PC＋で、仮想データの最高血圧に関する箱型図を表示するには、次のように指定する。

```
EXAMINE　VARIABLES=BPH
　／PLOT=BOXPLOT.
```

その結果、図5に示す箱型図が表示される。この場合、中央値は131.5で、上ヒンジが147.5、下ヒンジが122となる。ヒンジ散布度は147.5－122＝25.5となる。内境界点を求めると、83.75(＝122－1.5×25.5)と185.75(＝147.5＋1.5×25.5)になる。したがって、83.75より低いデータ、あるいは185.75より大きなデータが離れ値となる。一方、45.5(＝122－3×25.5)以下のデータ、あるいは224(＝147.5＋3×25.5)以上のデータは、飛び離れ値となる。最高血圧のデータでは、番号18の203の値が離れ値と判断され、箱型図に(O)で表示されている。

一般に、想定している母集団が正規分布をしているとき、データ数が少なくても20以上あれば、箱型図は次のような特徴を持つ。

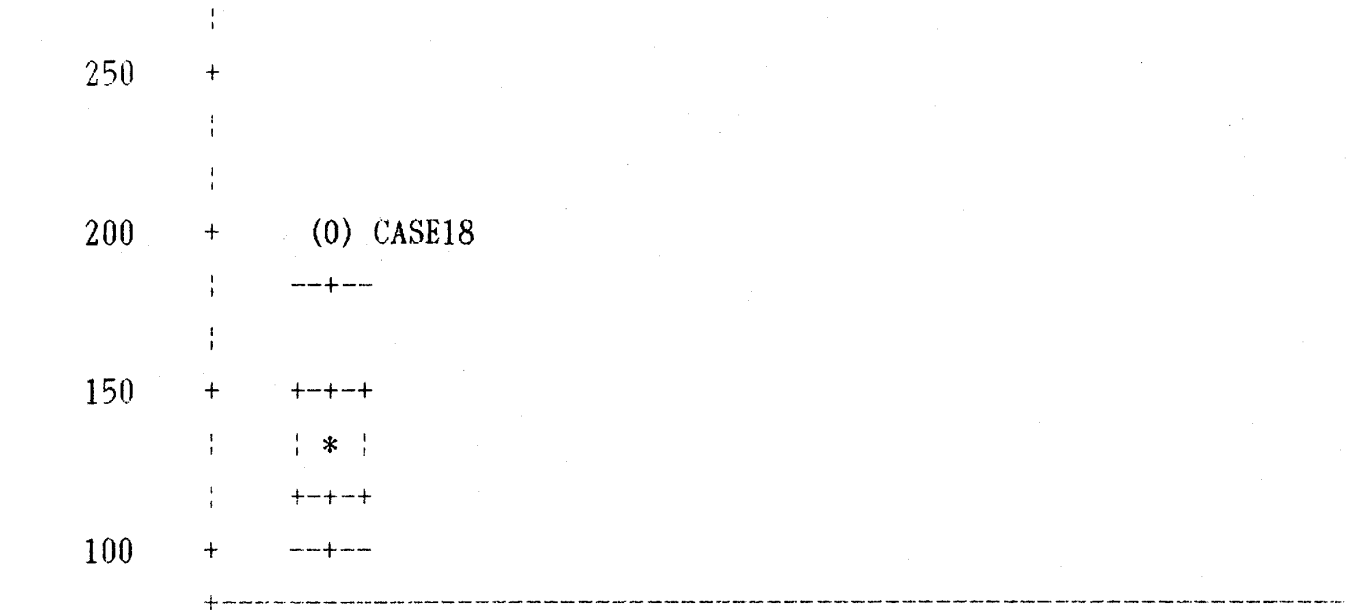
① 中央の＊は箱のほぼ中央にある。

② 箱の両側のヒゲはほぼ同じ長さになる。

③ 内境界点の外側に出るデータは、全体の2％程度である。

④ データが外境界点の外側に出ることは、ほとんどありえない。

この特徴から外れた箱型図が得られたときには、



Variables BPH

N of Cases 60.00

Symbol Key: * - Median (0) - Outlier (E) - Extreme

図5 最高血圧値の箱型図

異常値が含まれているか、母集団が正規分布しているという仮定を疑う必要がある。この場合、上の特徴に照らすと、最高血圧の仮想データには、異常値が含まれているとはいえない。

箱型図は、分布の概略を位置、散布度、歪み、裾の状態といった観点で知ることができる他に、外れ値の存否を検討したりするのに有効である。ただし、データ数や密度が直観的に把握できないということや、多峰性の分布の場合には不適当な印象を与えるなどの問題点がある。それ故に、箱型図は幹葉表示と併用して利用することが望まれる。

(4)平行箱型図

上に説明した箱型図が、特にその効果を発揮するのは、複数のデータ群を比較する場合である。箱型図を用いたデータ群の比較は、各データ群ごとの箱型図を平行に並べることによって行う。このような図示を平行箱型図(parallel boxplots)、または平行図式図(parallel schematic plots)という。平行箱型図の特徴は、その簡潔さ、すなわち、多数のデータ群を同時に比較できる点にある。さらに、平行箱型図では、個々の箱型図に基づいて、様々な観点(分布の位置、散布度、歪み、尖り、範囲、さらに外れ値の状況など)からデータ群の分布を比較することができるので、一般によく用いられている平均値、標準偏差などで分布を比較する場合に比べ、より豊富な情報を提供してくれる。

たとえば、表1のデータを用いて、年齢による最高血圧のデータ分布の違いを平行箱型図で比較してみよう。年齢段階は、40歳以下、41～50歳、51～60歳、61～70歳、71歳以上の5つに分けることにする。この場合、SPSS/PC+で、

平行箱型図を表示するには、次のように指定する。

```
RECODE  AGE(LO THRU 40=1)
    (41 THRU 50=2) (51 THRU 60=3)
    (61 THRU 70=4)  (71 THRU HI =5).
EXAMINE  VARIABLES=BPH BY AGE
                ／PLOT=BOXPLOT.
```

RECODE コマンドは、ある変数の値に新しい値を対応させ再コード化するためのコマンドである。このコマンドを用いて、年齢（AGE）のデータを上に述べた５つの年齢段階に再コード化した。VARIABLES＝BPH BY AGE の指定は、再コード化された AGE の値に基づいて最高血圧(BPH)のデータを細分し、５つのデータ群（セルと呼ばれている）を構成するためのものである。PLOT＝BOXPLOT と指定することにより、図６の平行箱型図が表示される。

３.等分散性の検定とベキ変換

t 検定や分散分析などの統計手法では、標本が正規分布に従う母集団から抽出されたものであること、等しい分散を有することといった仮説を前提にしている。したがって、こうした手法を用いてデータ分析を行う前には、標本の正規性、あるいは等分散性に関する仮説を検証することが必要である。

等分散性の仮説を検定するために、EXAMINE には、Levene 検定が用意されている。この検定では、まず、各データのセル平均からの

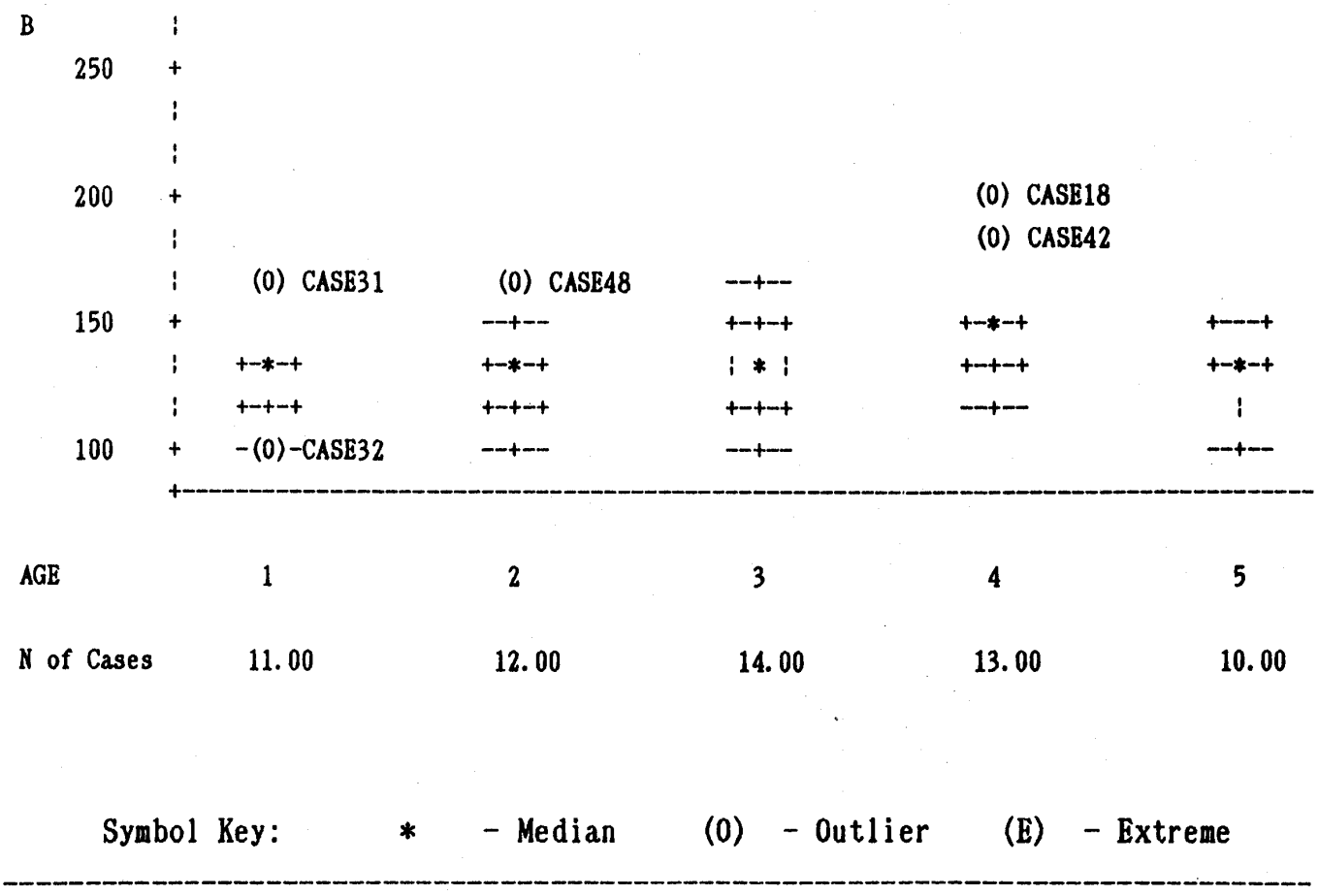


図６　最高血圧値の平行箱型図

絶対的な差が計算され、次に、これの差についての一元配置の分散分析が行われる。

もし、分散が等しいという帰無仮説が棄却されているにも関わらず、等分散性の仮説を前提とする統計手法を用いようとするならば、データの変換を検討しなければならない。データの変換とは、たとえば、データxを対数変換してlogxとするといった操作のことをいう。データを変換するには、通常、ベキ変換が用いられる。

データを適切にベキ変換するための指標を提供する手段として、EXAMINEには、散布度対水準プロット(spread versus level plot)が用意されている。散布度対水準プロットとは、各データ群の中央値とヒンジ散布度の自然対数を作図したものである。ヒンジ散布度は、四分位範囲(IQR)にほぼ相当する。

Levene検定と散布度対水準プロットを求めるためには、PLOTサブコマンドで、SPREADLEVELのキーワードを指定する。たとえば、図6に示したデータ群について、Levene検定付きの散布度対水準プロットを出力させるためには、次のように指定する。

```
RECODE AGE(LO THRU 40=1)
    (41 THRU 50=2) (51 THRU 60=3)
    (61 THRU 70=4) (71 THRU HI =5).
EXAMINE VARIABLES= BPH BY AGE
        /PLOT=SPREADLEVEL.
```

その結果、図7のような出力が得られる。Slopeは、図中のプロットに当てはめられた直線の傾きを表している。1からこの直線の傾きを引くことによって、データを適切に変換するためのベキ乗が示される。すなわち、あてはめられた直線の傾きをb、ベキ乗をpとすると、

$$p = 1 - b$$

として、p次ベキ変換を行えばよい。図7の、Power for transformationには、p＝1－bの値が記入されている。ベキ変換は、簡潔さと解釈のしやすさを考慮して、常に、最も近い1／2の倍数のベキ乗が選択される。最も広く使われているベキ変換を表2に示す。この場合は、p＝0.064に最も近い、p＝0をベキ乗として用いる。すなわち、各データxをlogxに変換する。

表2　よく用いられるベキ変換

ベキ乗	変換
3	立方
2	平方
1	変化せず
1/2	平方根
0	対数
-1/2	平方根の逆数
－1	逆数

データの変換が満足いくものであるか否かは、キーワードSPREADLEVELの後に、データ変換のためのベキ乗pを(p)で指定し、再度、散布度対水準プロットを出力して判断する。図8に、SPREADLEVEL(0)の結果を示す。図7のプロットよりも、図8の直線の傾きは0に近くなり、変換がうまくいったことを示している。

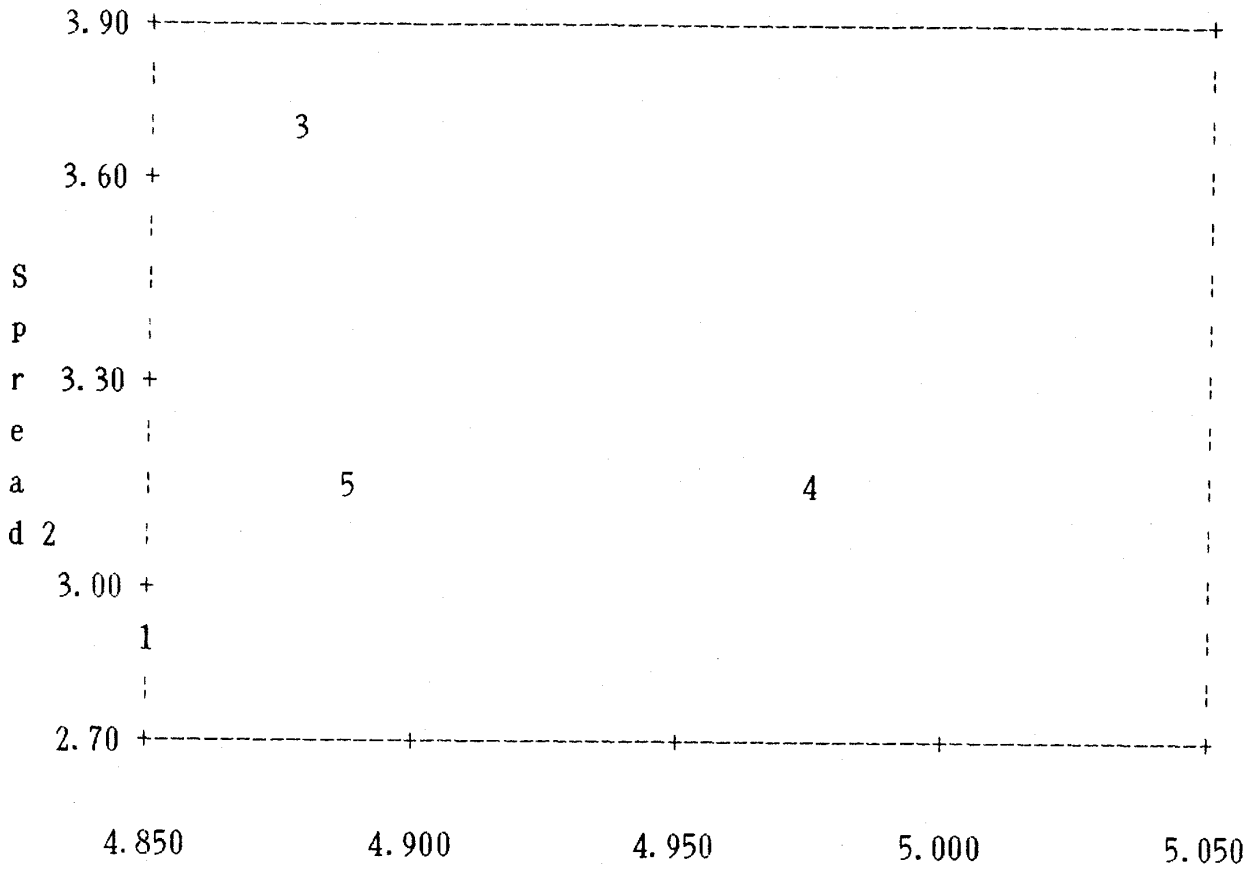
4.正規性の検定

EXAMINEで、データが正規分布からもたらされたものであるという仮説を検定するためには、正規確率プロット（normal probability plot）を用いる。正規確率プロットでは、各データは正規分布からの予測値と対になって表示される。正規分布からの予測値は、標本のデータ数とデータの大きさの順位に基づいて求められる。その作成手順を簡単に示すと、次のようになる。

標本数nのデータ、{X_1、X_2、……、Xn}を小さい方から大きい方に順に並べ換え、{$X_{(1)} \leqq X_{(2)} \leqq \cdots\cdots X_{(i)} \leqq \cdots\cdots X_{(n)}$} とする。

$P_i=(i-0.5)/n$とおくと、i番目の$X_{(i)}$より小さいデータは全体の$100P_i$パーセントになる。$X_{(i)}$をデータの$100P_i$パーセンタイル、P_iを累

Dependent variable: BPH
Factor variables: AGE



---

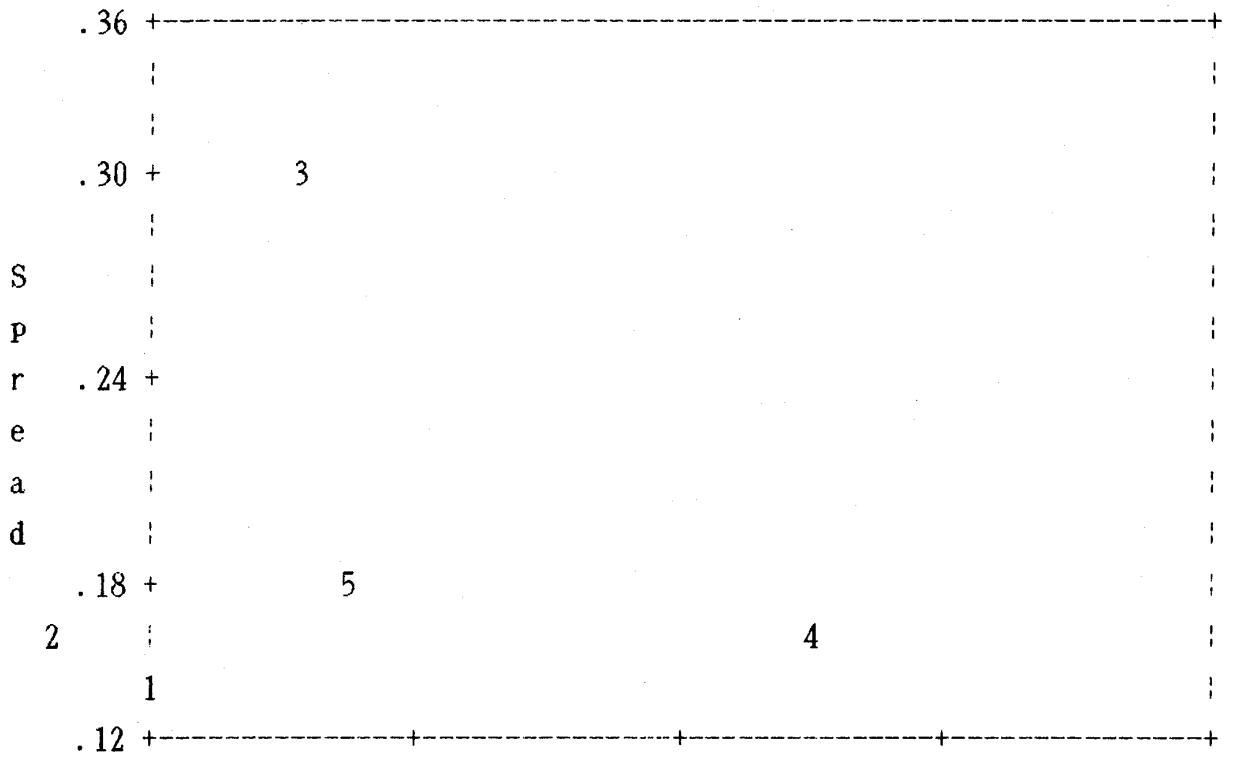
* Plot of LN of Spread vs LN of Level.

Slope = .936 Power for transformation = .064

Test of homogeneity of variance		df1	df2	Significance
Levene Statistic	.8701	4	55	.4878

図7 散布度対水準プロット

Dependent variable: BPH
Factor variables: AGE



* Plotted data transformed using P = 0

Slope = .003

Test of homogeneity of variance		df1	df2	Significance
Levene Statistic	.7542	4	55	.5596

図8 対数変換後の散布度対水準プロット

積比率と呼ぶ。正規確率プロットは、累積比率P_iに対応する標準正規分布のZ得点、Z_iを求め、$X_{(i)}$とZ_iを対にしてプロットする。

データ分布が正規分布に適合していれば、プロットした点はほぼ直線上に並ぶ。また、データの分布が左右対称であれば、Z＝0の線とプロットした点を結んだ折れ線との交点に関して、プロットした点は対称に位置する。

傾向化除去確率プロット(detrended normal plot)は、正規確率プロットの直線から実際のデータの偏差を示したものである。標本が正規母集団からのものであれば、各点はゼロを通る水平線の周辺に集まり、何らかのパターンが存在することはない。顕著なパターンが存在すれば、それは正規性からの逸脱を示唆している。

このように正規確率プロットおよび傾向化除去確率プロットは、データ分布の正規性をチェックするための視覚的情報を提供する。さらに、正規性を統計的に検定したいときには、Shapiro－Wilks 検定と Lilliefors 検定を用いる。ただしデータ数が50を越える場合には、Shapiro－Wilks 検定は実行されない。

これらの情報を EXAMINE で得るためには、PLOT サブコマンドの後に、NPPLOT のキーワードを指定する。たとえば、仮想データの最高血圧についてこれらの情報を求めるには、次のように指定する。

```
EXAMINE VARIABLES=BPH
  /PLOT=NPPLOT.
```

その結果、図９に示す出力が得られる。この図から、最高血圧のデータは、多少の歪みはみられるが、ほぼ正規分布していることが分かる。そのことは、Lilliefors 検定によっても保証されている。

```
    BPH
Valid cases:        60.0   Missing cases:        .0   Percent missing:       .0
    BPH

 2.40 +------------------------------+   1.20 +------------------------------+
      ¦                          *   ¦        ¦                          *   ¦
 1.60 +                      * *     ¦    .80 +                              ¦
      ¦                    ***       ¦        ¦                              ¦
  .80 +                   **         ¦    .40 +                      *       ¦
      ¦                 **           ¦        ¦     *              *         ¦
 -.00 +               **             ¦   -.00 +      * ****     ** *        ¦
      ¦              **               ¦        ¦      ***  *******           ¦
 -.80 +             **                ¦   -.40 +                              ¦
      ¦            **                 ¦        ¦                              ¦
-1.60 +       ***                     ¦   -.80 +                              ¦
      ¦       *                       ¦        ¦                              ¦
-2.40 +      *                        ¦  -1.20 +                              ¦
      +--+-------+------+-------+---+          +--+-------+------+-------+---+
         80      120    160     200               80      120    160     200
              Normal Plot                              Detrended Normal Plot

Page  4                          SPSS/PC+                              11/20/92

                        Statistic        df       Significance

K-S (Lilliefors)           .0791         60          > .2000
```

図９　正規確率プロット

5.頑健な推定による位置の推定

代表値として、通常最もよく利用されるのは平均値である。しかし、前述のように、平均値は外れ値によって大きく影響されるという欠点を持っている。この欠点を補うために、考案された位置に関する推定量の一つが調整平均(trimmed mean)である。

外れ値によって影響されにくい、つまり頑健な推定値を得るには、一般に、中央からデータが遠ざかるに従って重みづけを小さくし、その上で、平均を求める方法が採用されている。たとえば、体操の採点やジャンプの飛行審査などで用いられているように、最高点と最低点を除外して中央の残りの点の平均点を求める方法が上げられる。調整平均は、このような考えに基づいて求められる平均値である。すなわち、両裾の何パーセントかのデータの重みを0とし、その内側に入る残りのデータの重みを1として求められた平均値を調整平均という。最大値、最小値からa%のデータを除いた残りのデータについての平均値を、特にa%調整平均という。たとえば、20%調整平均の場合、最大値からの20%と最小値からの20%のデータが取り除かれる。推定値は、データの中央部分にあたる60%の値にのみ基づいて求められる。

調整平均の利点は、中央値と同様に、極端な値に影響されない推定ができることにある。しかし、中央に位置するたった一つあるいは二つの値に基づく中央値とは違って、もっと多くの真ん中にある値に基づいているので、調整平均は中央値に比べてデータをより有効に活用しているといえる。EXAMINEでは、デフォルトによって出力される記述統計量の中に、5%調整平均(5% Trim)が表示される。

調整平均では、中央値から遠く離れたデータはかなり厳しく扱われ、まとめて除外される。それに代わる寛大な方法は、他の値から遠く離れたデータを計算に含めるが、中心に近いデータに比べてより小さな重みづけを与える方法である。このようにして求められる推定量をM推定量(M-estimator)とよぶ。

各データに対して重みづけを与えるための方略には、数多くの方略があるので、それだけ多くの異なるM推定量が存在する。一般に使われているM推定量は、分布の中心からの距離が大きくなればなるほど重みづけが小さくなるように、重みづけが割り当てられている。SPS

Mean	134.4000	Std Err	2.6941	Min	96.0000	Skewness	.6841
Median	131.5000	Variance	435.4983	Max	203.0000	S E Skew	.3087
5% Trim	133.5741	Std Dev	20.8686	Range	107.0000	Kurtosis	1.1365
				IQR	25.7500	S E Kurt	.6085

M-Estimators

Huber	(1.339)	132.9568	Tukey	(4.685)	131.8310
Hampel	(1.700, 3.400, 8.500)	132.9371	Andrew	(1.340 * pi)	131.7995

図10　最高血圧値の記述統計量とM推定量

S／PC＋のEXAMINEでは、Huber、Tukey、Hampel、Andrewの4つのM推定量が利用できる。

表1の最高血圧のデータについて、M推定量を求めるには、次のように指定する。

```
EXAMINE  VARIABLES=BPH
  /MESTIMATORS.
```

図10は、上の実行結果の出力から、記述統計量とM推定量の部分を表したものである。デフォルト指定により記述統計量として、平均値(Mean)、中央値(Median)、5%調整平均(5% Trim)、標準誤差(Std Err)、分散(Variance)、標準偏差(Std Dev)、最小値(Min)、最大値(Max)、レンジ(Range)、四分位範囲(IQR)、尖度(Skewness)、尖度の標準誤差(S E Skew)、歪度(Kurtosis)、歪度の標準誤差(S E Kurt)が表示される。

MESTIMATORSのサブコマンドによって、4つのM推定量が表示されている。これら4つの推定量は、いずれも、平均値134.4より小さい値となっている。最高血圧のデータは、分布が正の方向に多少歪んでいるために、平均値が値の大きなデータによって影響されていることが分かる。

おわりに

SPSS／PC＋による探索的データ解析は、データ分析の予備的なステップとして、有益な種々の情報を提供する。すなわち、データの分布形状や要約統計量、等分散性や正規性に関する検定、データをベキ変換するための指標、さらには、頑健な推定量である。これらの情報を視覚的に表示するのが、SPSS／PC＋による探索的データ解析の特徴である。結果の視覚的表示は、データの多様な側面を同時に、直観的に読みとるのに役立つので、データ解析において重要な位置を占めるであろう。データをベキ変換することは、データの単純な関係を求めたり、複数のデータ群の比較を容易にすることに役立つ。それ故に、どのようなベキ変換を行うべきかに関する情報は、データ解析を行う際に貴重な情報となる。また、t検定や分散分析のように、母集団の等分散性、正規性を前提とする検定を行う場合には、データの等分散性や正規性に関する情報は検定手法を吟味する際に役立つ。さらに、データ収集の際のコントロールが必ずしも十分ではない探索的段階では、種々の要因によって外れ値が生じたり、一部のデータが異常であったりすることが多い。このような問題を避けるためには、一部の異常なデータが解析の結果に大きな影響を及ぼさないこと、すなわち頑健な推定量を求めることが必要である。

引用および参考文献

海保博之　編著 1985　心理・教育データの解析法10講　基礎編　福村出版.

海保博之　編著 1986　心理・教育データの解析法10講　応用編　福村出版.

中里溥志　1989　実験データのグラフ表示　サイエンティスト社.

坂本元子、丹後俊郎　1987　栄養情報の統計解析　朝倉書店.

芝祐順、渡部洋、石塚智一　編　1984　統計用語辞典　新曜社.

SPSS Inc. 1991　SPSS Statistical Algorithms 2nd ed. Chicago,SPSS Inc.

SPSS JAPAN Inc. 1991　SPSS／PC+ Base Manual V3.0J（SPSS／PC+ NEC PC−9800対応版マニュアル）.

竹内啓　編集委員代表　1989　統計学辞典　東洋経済新報社.

垂水共之、西脇二一、石田千代子、小野寺孝義 1990　新版 SPSS^{X} II　解析編1　東洋経済新報社.

渡部洋、鈴木則夫、山田文康、大塚雄作　1985　探索的データ解析入門－データの構造を探る－朝倉書店.

渡部　洋、大塚雄作、鈴木則夫、山田文康　1984　行動科学データ解析のための探索的方法　行動計量学　第12巻　第1号　59−80頁.

山本嘉一郎、吉村英、竹村和久　1991　パソコンSPSS　基礎編　東洋経済新報社.