電卓シリーズ（4）

相関係数の計算と統計図のプログラム

吉井守正（鉱床部）

今回は　多量なデータを処理する方法として筆者が考えたものについて述べ　その具体例として相関係数と回帰直線を求める計算とプロッタを使ってのデータの図化とくにヒストグラム・相関図・三角図について記す.

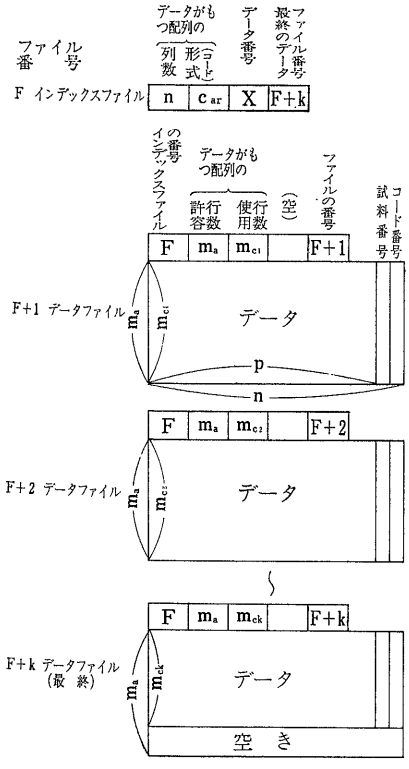
1. 多量なデータを処理するための工夫

データ数が多くなると　データの入出力に手間取るというほかに　さまざまな問題が派生して来る.　そこで実際に経験した事柄をもとに　多量なデータへの対応策をいくつかご紹介し　ご批判をあおぎたい.

1.1. 磁気テープを使う場合

1.1.1. ファイルの構成

データを磁気テープなどのいわゆる外部メモリーに貯えることによって　計算機本体のメモリー容量の何倍ものデータを処理する事ができる.　筆者が使用している



第1図　磁気テープのファイルの構成

YHP-20計算機にも昨年からカセットテープレコーダが付属し　多量なデータをさばくのに大いに役立っている.　そこで磁気テープの割り付けや使い方のコツなどつぎに示そう.

磁気テープは一般にはいくつかのファイルに区切って使われる.　さて　データの量がファイルの数個分にまたがるようなときには　どのように設計しておけば能率がよくしかも安全であろうか.　筆者が現在用いている方式はつぎのようなものである.

ファイルの構成と内容を第1図に示す.　データを収めるファイルをデータファイルと呼ぶ事にする.　データは一般に配列をもっているので　考えやすいようにファイルを数表の形で表現しよう.　ひとつのファイルに入れられるデータの配列は第1図に示すように最大 m_a 行 n 列の規模をもつものとする.　そのファイルの大きさが　その計算機のメモリー容量によってきめられることは申すまでもない.

データ中の1試料に属する要素の数　たとえば化学分析値の場合は成分数が　この方式では配列の列数 P に対応する.　各試料には試料番号と分類用のコード番号（1.3で述べる）が付けられるのでファイルの列数 n は $P+2$ 列となる.　これらはデータによって定まった値なので　メモリー容量すなわちデータレジスタ数を列数 n で割った値（の端数を切り捨てたもの）が　そのデータに対するファイルの許容行数 m_a となる.　これはそのファイルに収容可能な試料数と同じ意味である.

データの量に応じてデータファイルの数をあらかじめ算出して　ファイルを作っておかねばならない.　このとき　試料数が配列の行数に対応する事を念頭におけば必要なファイル数はすぐ計算できる.　なお　筆者の方式では　データファイルの番号はつねに一連でなければならない.　データをこれらのファイルに順次入力するためのプログラムも作られているが　それについては省略する.　データが入力されたあとのファイルの状況はデータを満載したファイルはその使用行数 m_c が許容行数 m_a と同じになるし　ファイルに空きがある（一般には最終のファイル）場合は　当然 $m_a > m_c$ となる.

これら一連のデータファイルの前に見出用のファイルを付けておく.　これをインデックスファイルと呼ぶ事

にする．　このファイルには　データ全体についての事項を記入しておく．　すなわち　データの列数・配列の形式を示すコード番号・データ全体の番号（データ名）・最終のデータファイル番号などである．　これとは別に各データファイルにも　おもにそのファイルに関係する事項を記入する部分を作っておく．　すなわちインデックスファイルの番号・データの許容行数と使用行数・そのデータファイルの番号などである．

計算を実行するときは　まず必ずインデックスファイルの内容を入力し　つぎにデータファイルの内容を必要に応じて順に入力する．　そして現在計算機に入力されているデータがどのファイルのものか　その配列規模がどの位かなどについては　上に述べた項目をチエックする行程をプログラムに付けておけば　ファイルの取り違いや配列の計算処理の誤りを防ぐ事ができる．

1.1.2. テープ走行時間の節約

磁気テープの欠点のひとつは　その走行時間がほかの計算処理時間に比べて圧倒的に長いという点である．筆者の経験では YHP-20 計算機に1000個の数値をテープから入力する場合　1回について約25秒かかる（ただし同社の新型機ではこれより数倍早いという話だが）．計算ループのかけ方によっては　ひとつの計算の中で同じファイルの内容が繰り返えし入力されるような事が起きるが　このような場合は　計算時間の大半がテープ走行時間に食われ　しかもそれが実用にならないほどの長時間に及ぶ事さえまれでない．

そこで　計算の中間結果をメモリーに足し込むなど適宜必要な処置をして　できるだけテープの走行回数を減らすように努める必要がある．　その具体例は　相関係数の計算（2.2）とヒストグラム（3.1.1）のところで述べる．

1.2. 数値が欠けているデータの取り扱い

実際の計算では　データの中の数値がところどころ欠けたものを　やむを得ず処理せねばならない事が起きる．たとえば化学分析データで　いくつかの試料の中である成分の分析値がないという場合がこれに当たる．　このようなときは分析値のあるものだけについて　ともかく処理しようとするのが普通であろう．

この場合　数値がないという事と数値が0という事は当然区別されねばならない．　そこでこの数値がない事を意味する“数値”をどう決めるかであるが　これは一般には　その計算に使われる数値の定義域の外にある数値を使えばよい．　筆者はプログラムによってまちまちだが−90000以下とか -10^{99} 以下の数値が入力されたら“数値なし”とする判断を付けている．　“数値なし”の印字は“……”とするようにもしている．

化学分析値などマイナス値がないものは　マイナスの値をこの意味に使える．　これを応用して $-n$ を入力すると“以下 n 個分空欄”の意味に使う事もできる．　そのようにしたプログラムも作ったが　結構便利である．

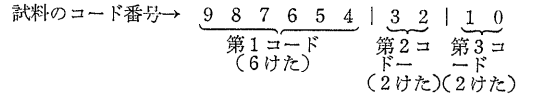
1.3. コード番号による試料の分類と選択

磁気テープに収められているデータの一部についてだけ計算をしたい場合　必要なものを選び出すために　数ファイルにまたがる量のデータを編集しなおすのは　大変やっかいである．　そこで　必要な試料だけを選び出す簡単な方法を考えてみよう．

筆者のやり方では　試料番号0の試料は計算から除くという判断が各プログラムに付けてある．　各ファイルにあるデータの数値を変更するのは訂正用のプログラムを使うと容易なので　番号を0に改めればよい．

データが分類可能なときは　入力の際に使用者が任意に分類用のコード番号を設計して　各試料ごとに付けておき　計算の際にこのコード番号に合ったものだけを処理するという方法ができる．　これについて説明しよう．

YHP-20計算機の場合は1個の数値が10けた以内に定められているので　筆者は各試料のコード番号を10けた以内で表現する事にした．　この10けたの数値を3個に区切り　3種類のサブコートに分ける．　それらはつぎのとおりである．



使用者は　この3種のコードに任意に意味付けをしてよい．　たとえば岩石試料なら　第1コードを産出地域（999999ヵ所に区分可能）　第2コードを時代（層準）（99に区分可能）　第3コードを　岩質（同じく）　という具合に定義して分類したらよいだろう．

計算の際には　各サブコートの何番を選ぶかを指定する．　もしそのサブコードにとくに着目しなくてよい場合は　0を指定する．　たとえば第1コードから第3コードへ順に　21，0，14　を指定すると　コード番号21××14をもつ試料が選ばれる（×はこのけたの数字に無関係の意味）．データ全体を取り扱いたいときは　各コードとも0を指定すればよい．

つぎに　試料を選択する行程のために　筆者が考えた方法を示そう．　試料のサブコードを d_1，d_2，d_3　使用者が指定したサブコードを S_1，S_2，S_3 とする．　このとき　判断の行程はつぎのとおりである．

1. もし $d_1 \neq 0$ で $d_1 \neq S_1$ ならば　計算から除く
2. もし $d_2 \neq 0$ で $d_2 \neq S_2$ ならば　計算から除く
3. もし $d_3 \neq 0$ で $d_3 \neq S_3$ ならば　計算から除く

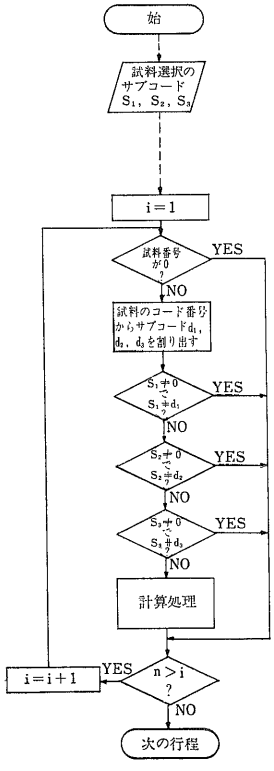
以上3段の判定で計算から除外されなかったものが計算のために選択された試料である．　行程としては単純明快である．

第1コードから第3コードまでを上のようなケタ数に区切ったのは　使いやすさを考えての事であり各サブコートとも1ケタで良いのなら YHP—20の場合1レジスタあたり最大10種のサブコードを作る事ができる．

使用者はこれを任意に組み合わせてよいので　サブコードのうちどれとどれに着目して組み合わせるかという数だけを考えてもサブコードの種類が増すとその数は飛躍的にふえる．

しかし上に述べた方法によれば判断行程の数はサブコードの数にしかならない．　行程をひとつにしてループ計算にすればプログラムも節約できる．　したがって計算時間さえ問題にしなければ　このようなやり方で十分多種の条件を重複させる事ができる．　試料選択の行程を第2図に示す．

上の行程を眺めていると計算機がなるべく仕事をすまいと努めているように見えるので　このやり方を“なまけもの法”とでも名付けたらどうだろう．



第2図　試料を選択する行程

2　相関係数と回帰直線

相関係数を求めるプログラムは　プログラミングの技術としては特筆すべきものはないのだが　さまざまな目的で利用されたという実績を考えて　ご紹介をする事にした．　さらに　その計算結果を評価する際に注意すべき点についても付け加えておいた．

筆者は統計学に対してとくに深い知識があるわけではないので　相関係数についての詳しい議論はさておきとくにプログラムと関係する要点だけを述べてみよう．

2.1.　相関係数とは

ひと口でいえば　相関係数はふたつの変量X，Yの関係の強さを示すものでX，Yを同時に観測して得られたn個の数値の組

$(x_1, y_1), (x_2, y_2), \cdots\cdots, (x_n, y_n)$

をXY平面上にプロットして得られた相関図の直線性の度合を表わす値である．　数式の上では　X，Yの共分散をXとYのそれぞれの標準偏差の積で割った値である．すなわち相関係数rは

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \cdots\cdots(1)$$

である．　ここに $\bar{x}$ は $(x_1, x_2, \cdots, x_n)$ の　また $\bar{y}$ は $(y_1, y_2, \cdots, y_n)$ のそれぞれ平均値である．

rは　$0 \leqq |r| \leqq 1$ の範囲にあり　$|r|$が1に近いほどX，Yは強い相関関係にあり　また$r>0$の場合は正（順）相関　$r<0$の場合は負（逆）相関という．

変量X，Yに属する数値の組に直線的関係が認められる場合は　最小二乗法により相関図に回帰直線を引く事ができる．　回帰直線は　任意の直線 AB を考え　AB に点 $P_1(x_1, y_2)$，$P_2(x_2, y_2)$，……，$P_n(x_n, y_n)$ を通り　y軸に平行な直線を引き　AB との交点をそれぞれ $P_1'(x_1, y_1')$，$P_2'(x_2, y_2')$，……，$P_n'(x_n, y_n')$ とするとき　y_i から y_i' までの長さの二乗の和　すなわち

$$L=\sum_{i=1}^{n}(y_i-y_i')^2 \quad \cdots\cdots(2)$$

を最小にするように引かされた直線の事である．

回帰直線の式を

$$y=ax+b \quad \cdots\cdots(3)$$

とする．　このとき勾配　aは

$$a=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \cdots\cdots(4)$$

切片 b はこの直線が点 $(\bar{x},\ \bar{y})$ を通ることから

$$b=\bar{y}-a\bar{x} \quad \cdots\cdots(5)$$

としてそれぞれ求められる．　これらの式をみちびく計算の詳細については　高橋ほか（1975）などにわかりやすく説明されているので　ここでは省略する．

2.2. 相関係数を求めるプログラム

まず r を求める計算は　(1)式を展開し

$$r=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right)}} \quad \cdots\cdots(6)$$

という形にして行なう．　r の値を求めるには(6)式を構成する

$$n,\ \sum_{i=1}^{n}x_i,\ \sum_{i=1}^{n}x_i^2,\ \sum_{i=1}^{n}y_i,\ \sum_{i=1}^{n}y_i^2,\ \sum_{i=1}^{n}x_iy_i \quad \cdots\cdots(7)$$

の各項の値をあらかじめ求める必要がある．

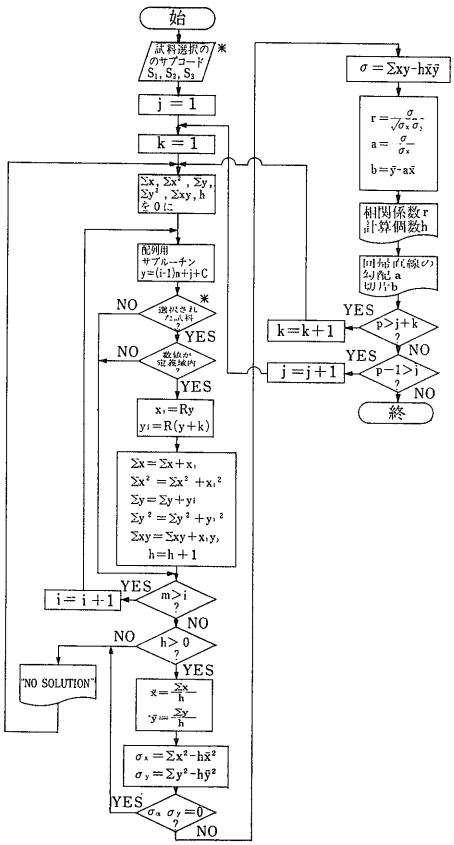
a についても同様に　(4)式を展開し

$$a=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2} \quad \cdots\cdots(8)$$

の式から求める．　(8)式を構成する項は(7)式に含まれているので　a は容易に算出できる．

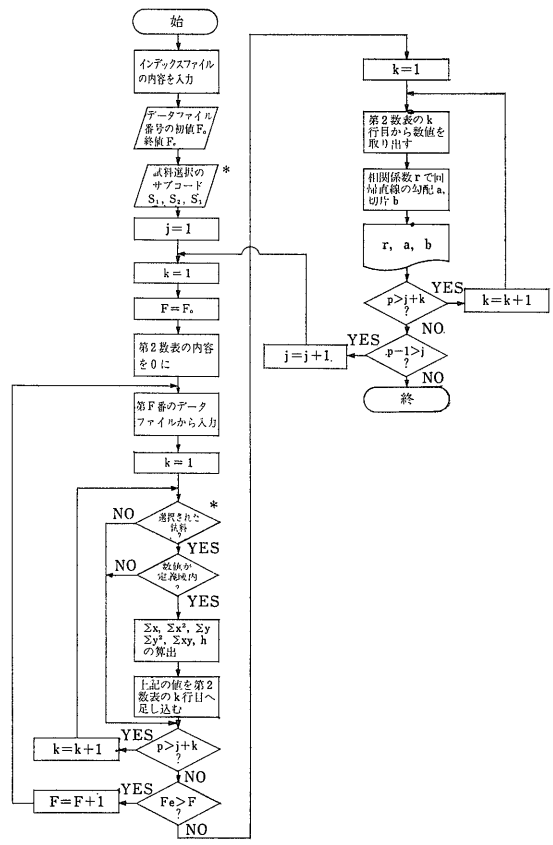
筆者はプログラムでは　n 行 p 列の配列をもつデータに対して　その列要素を変量 C_1, C_2, …, C_p としてそのすべての組について総当たりで相関係数を計算するようにしてある．　すなわちその組み合わせと順序はまず C_1 対 C_2, C_3, …, C_p　つぎに C_2 対 C_3, …, C_p そして最後は C_{p-1} 対 C_p である．　その行程を第3図に示す．

データが数ファイルに及ぶときは　テープ走行時間の節約のために C_j 対 C_{j+k}, C_{j+k+1}, ……, Cp についての計算を各ファイルごとに行ない．　(7)式の各数値をメリモーの第2数表に記入しあとまでめて値を取り出すようにする．　第2数表は k 行6列の配列をもち　各列には(7)式の各数値が対応する．　各行要素に対しては C_j 対 C_{j+k}, C_{j+k+1}, ……, Cp の計算結果が第 k 行に記



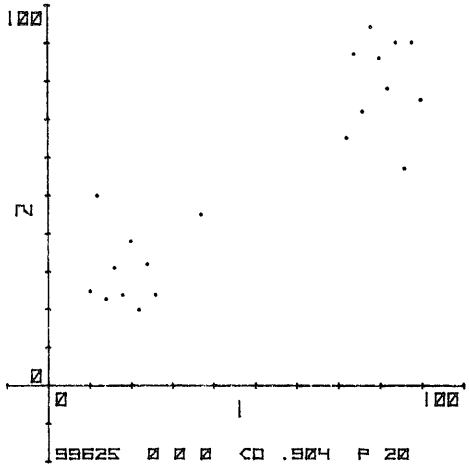
第3図　相関係数の計算行程

図には記さなかったが　回帰直線の a, b の値は $r\geqq 0.9$ であって使用者が必要とするときに限って出力する．　m は配列の行数　p は変量の（列）数　n は配列の列数　（$p+2=n$　試料番号とコード番号が加わるため）．　j は変量 c_j　k はこれと組み合わされる変量 c_{j+k} のそれぞれ列番号用の計数器．　配列用サブルーチンについては吉井(1977)を見よ．　＊第2図参照



第4図　相関係数の計算で磁気テープを使う場合の計算行程

第2数表は k 行6列の配列をもち　その操作は k ループで行なう．　合計3ケ所に k ループが配置されている．　f はファイル番号用の計数器．　計算の主要部については第3図を見よ．　＊第2図参照．



第5図　相関係数を求め場合に問題となる例
点の塊が2ケ所に別れて分布するとき　全体としては $r=0.90$ となるが　個々の塊では左下のものが $r=0.21$ 右上のものが $r=0.02$ である.

入される.　これを各ファイルについて繰り返して　足し込むと　最後のファイルの計算が終ったところで　第2数表には C_j 対 C_{j+k}, C_{j+k+1}, ……, C_p の相関係数を求めるのに必要な数値がそろう事になる.　これらの組に対して一挙に相関係数を計算し結果を出力する.　この行程を第4図に示す.

この方法によってもテープは　$p-1$ 往復する事になり能率の上ではまだ問題だが　それでも積和計算のループの中に　テープ走行を組み込むと実に $p(p-1)/2$ 回も往復する事になってしまう.

2.3. いくつかの問題点

実際の計算で考えておかねばならない点について　つぎにいくつか述べる.　数学的には $|r|$ が1に近い値ならば　そのふたつの変量 X, Y には直線的関係があるといえる.　だが　実際のデータでどのような場合にもこの結論を下してよいものか　少し検討しよう.

いまここに $r=0.90$ のデータがあり　その相関図が第5図に示されるようなものであったとする.　この図をみると点は離れた2ケ所に塊を作って分布している.　このようなパターンを示すときは必ず $|r|$ の値が1に近くなる.　なぜならば計算上は点が2個しかないデータ（$|r|=1$）に近い結果となるからである.　もし2ケ所の塊が何かの条件（地学的データなら地域・層準・岩質など）の差を反映したためにできたものだとしたら　統計上は別の標本として扱うべきものかもわからない.　個々の塊について計算すると　第5図の左下の塊りでは $r=0.21$　右上の塊では $r=0.02$ となり　相関関係は認められない.

ここに示した例は　データの分類を変えただけで相関係数が大きく変化する場合のある事を示している.　このような例はしばしばみられるので注意を要する.

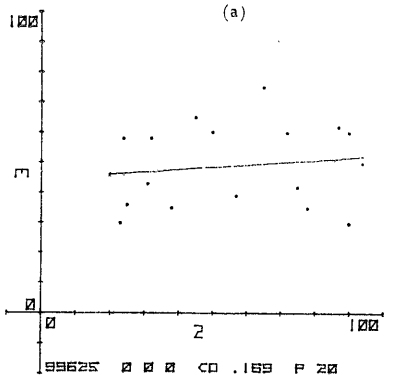
回帰直線を引く際にも問題がある.　第6図にその例を示す.　第6図(a)では一見回帰直線が順当に引かれているように見える.　しかし第6図(b)ではこの直線がまったく見当違いの方向に延びている.　このふたつの図は実は同じデータを X 軸と Y 軸を入れ替えて計算しただけなのである.　ではなぜ図の向きによって　このような大きな違いが出るのだろう.

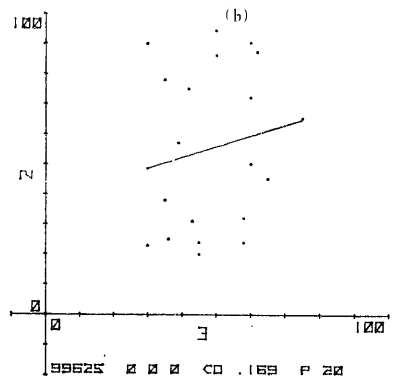
この場合　r が0.17と大変低い値にある点が重要なヒントになる.　r を求める(1)式と　a を求める(4)式は形がよく似ているので　(4)式の分母分子に

$$\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad \cdots\cdots(9)$$

を掛けて整理すると

$$a = r \cdot \frac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \quad \cdots\cdots(10)$$





第6図
回帰直線の問題を示す例

(a)　この図では相関図に一見正しく回帰直線が引かれているように見える.

(b)　この図は(a)図と同じデータを X 軸と Y 軸を入れ替えてプロットしただけである.　直線はここでは明らかに違った方向に引かれており　回帰直線とは認められない

となる．　これで回帰直線の勾配aが相関係数rの関数になっている事がわかる．　したがって$|r|$が小さくなると　回帰直線は点$(\bar{x}, \bar{y})$を中心に勾配がゆるやかになる方向に回転し　$r=0$ではX軸と平行になってしまうのである．

すなわち回帰直線は$|r|$が1に近い場合以外は　うかつに計算から求めて引いてはならないのである．　筆者のプログラムでも　$|r|\geqq 0.9$であって使用者が必要とする場合のほかは回帰直線の定数a，bを算出しないようにしている（第3図と第4図ではこの判定を省略した）．

rの値から相関関係を論じ　また回帰直線を引く場合は　いずれもふたつの変量の間に直線的な関係があるという事を前提としている．　だから理論式が一次式で表わされるときに観測値からその実験式をみちびくというような目的には　相関係数や回帰直線の考えが適合する．

しかしこれを化学成分間の相関関係の推定に応用するような場合は　上にあげたような種々の問題が起きる可能性がある．　したがって数値だけから安易に結論を急ぐのは大変危険である．

これらの検討は必ず相関図を作って行なうべきである．

3. 統　計　図

データが多いときは数表のままで検討するよりも　図に表わして眺める方が　そのデータのもつ性質をよく理解できる事は申すまでもない．　しかし作図は手間のかかる仕事だし　人によって得手不得手もあり　一般にわずらわしい．　そこで計算機にプロッタを接続して　必要な図を使用者の思うままに描かすと　能率的だし正確でもある．

では代表的な3種類の統計図すなわちヒストグラム・相関図・三角図を描くプログラムについて　目下筆者が実用化しているものをご紹介しよう．

これらのプログラムでのデータの取り扱い方は共通である．　まずこれについて簡単に述べておく．

計算機へのデータの入力は各プログラムの中で磁気テープから行なわれるか　または別に用意されたプログラムによってキーボードから行なわれる．　計算処理はメモリー内の数表（吉井　1977）の列要素に対して行なわれる．　その配列の列番号をここでは成分番号と呼ぶ事にする（化学分析値の各成分に対応するので）．

図には成分番号が意味する成分名を書く事もできる．これにはそれらを対応させるサブルーチンを作って　主プログラムに付け加えておく．　同様に図の標題を書く事もできる．　これらは使用者が適宜作るものとする．その詳細は省略する．

3.1. ヒストグラム

ひとつの変量Xに属するn個の数値　$x_1, x_2, \cdots, x_n$をいくつかの階級（区間）に分けて　各階級に属する数値の度数（個数）を求め　棒グラフにしたものをヒストグラム（度数柱状図）という．　XY座標のX軸に階級をY軸に度数をそれぞれ取るものとする．　度数百分率（＝度数×100/n）による表示もしばしば用いられる．階級の幅（X軸のきざみ）は　一般には階級ごとに異なっていてもよいのだが　ここでは各階級が同じ幅をもつ場合に限る事にしよう．

3.1.1. プログラム

まず作図条件の入力が必要である．　それにはつぎのようなものがある．

1. ***X*** **軸　側**
   - a　図示範囲の最小値 $X_{\min}$ と最大値 $X_{\max}$
   - b　階級の幅d（軸の目盛を兼ねる）

2. ***Y*** **軸　側**
   - a　度数の最大値 $f_{\max}$
   - b　軸の目盛

これらと図との関係を第7図に示す．

ヒストグラムのプログラムの行程を第8図に示す．

ヒストグラムを描くには　まず各階級ごとの度数を算出する必要がある．　一般には各階級の境界値を第7図のように$a_0, a_1, \cdots, a_k$とするとき$x_1, x_2, \cdots, x_n$の中から　$a_{l+1}>x_j\geqq a_l$となる数値の個数（これが度数）を求める．

プログラムでは　各階級をはさむふたつの境界値のうち値の大きい方をa_l　小さい方をa_sとして　もし

$$a_l>x_j\geqq a_s$$

ならば度数用の計数器f_iを1だけ足し上げるという操作を$x_1, x_2, \cdots, x_n$に対して繰り返す．　階級の移動は初値をそれぞれ

$$a_s=X_{\min} \qquad a_l=X_{\min}+d$$

として　以下順次

$$a_s=a_l$$
$$a_l=a_s+d$$

の操作で行なう．　これらによって度数$f_1, f_2, \cdots, f_k$の算出が行なわれる．

つぎにヒストグラムを描く行程について述べる．　いま座標内の任意の点(x_1, y_1)から別の点(x_2, y_2)まで直線を引く命令を

PLT$(x_1, y_1)\rightarrow(x_2, y_2)$

と表わす事にすると　$a_s=a_l$　$a_l=a_{l+1}$の階級では

PLT(a_s, f_i)→(a_i, f_i)

という命令になる．　プロッタのペンを下げたままにして　この操作を繰り返せばよく　各階級の境界値のたて線も自動的に引かれる．　ヒストグラムの右端にあるたて線は　これらのループから脱け出したところで（a_l=a_k のとき）

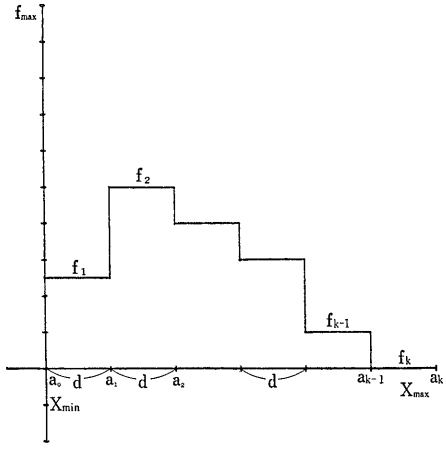
PLT(a_l, f_k)→(a_l, 0)

の命令で引かせる．

さて　ヒストグラムは原理的には度数 f_i を算出するごとに　その階級の分を描くようにすればよい．　しかしデータが磁気テープの数ファイルに及ぶ場合は　この方法では階級の数だけテープが往復し　その走行時間が長過ぎて実用にならないおそれがある．　そこで各ファイルのデータごとに度数 f_1, f_2, …, f_k を求めてメモリーに足し込み　あとで全体について一挙にヒストグラムを描くという方法をとるとよい．

この方法を積極的に利用するとヒストグラムの重ね描きができる．　すなわち作図条件が同じの複数のデータについて各度数を積算し　前のヒストグラムの上に積み重ねてゆくのである．　これとコード番号による試料の選択とを組み合わせると　ヒストグラムの合併などが容易となり　データの解析に役立つ．　この重ね描きの例を第9図に示す．

描かれた図の下には　データの番号・選択コード・作図条件・累積度数など必要事項を書き添える．　それらを同時にプリントアウトしておくと　図の検討や再生に便利である．　その例を第10図に示す．



第7図　ヒストグラム
a_0, a_1, …, a_k 階級の境界値　d 階級の幅　f_1, f_2, …, f_k 度数

3.1.2. 階級の幅と図形の変化

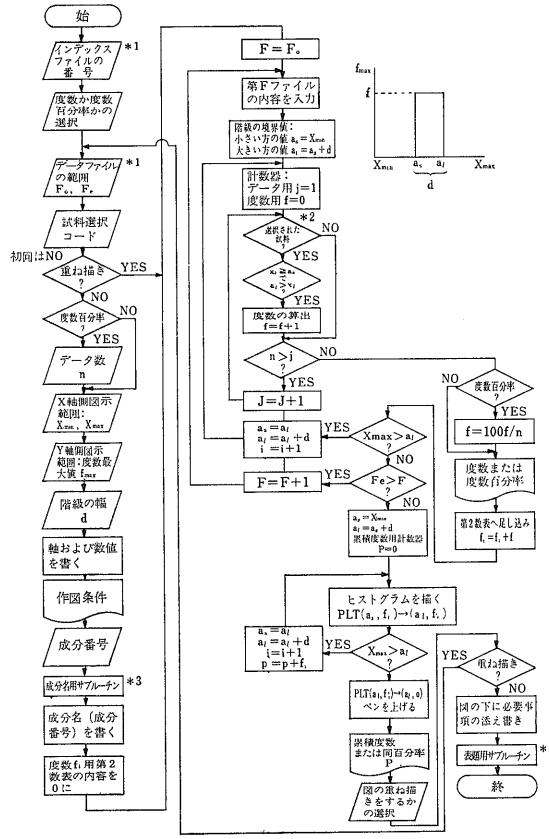
ヒストグラムの形は　階級の幅 d の大きさによってかなり違ってくる．　同じデータに対して d の値を　10, 15, 20, 30と変化させて得られた図形を第11図に示す．d の値は大き過ぎていけないのは申すまでもないが　小さ過ぎても　でこぼこが目立って全体の状況がわからなくなる．　適当な値がどの辺にあるかはデータによって異なるので試行錯誤せねばならない．　これを手作業で行なうのは大変な手間であるが　計算機に順次異なった d の値を入れては図を描かせ　もっとも説得力のある図形を選択するのは容易である．　その上度数の算出は正確だから　この辺は計算機の威力である．

3 2. 相　関　図

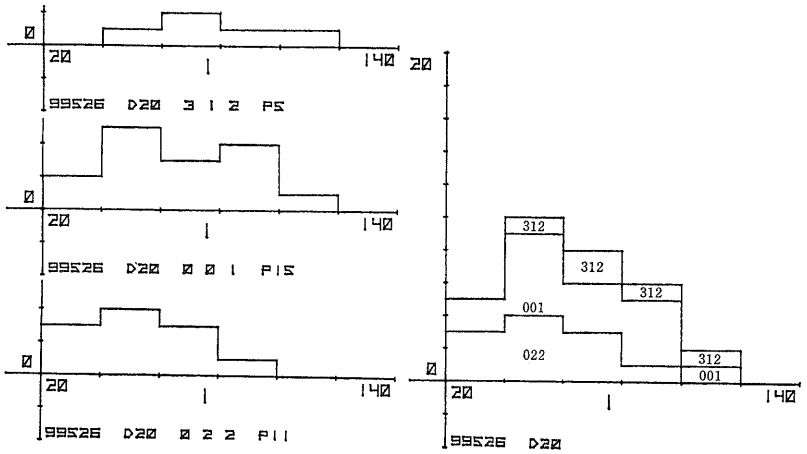
2種の変量 $X(x_1, x_2, \cdots, x_n)$ と $Y(y_1, y_2, \cdots, y_n)$ の対応する数値（x_i, y_i）の組の相関関係を調べるために XY 座標に表わしたものを相関図という．

3.2.1. プログラム

相関図の行程を第12図に示す．



第8図
ヒストグラムを描くプログラムの行程
*1. 磁気テープに入っているデータの入力
*2. コード番号による試料の選択
*3. 使用者が別途用意するサブルーチン．　成分名用のサブルーチンを付けないときは図に成分番号が書かれる．



第9図
ヒストグラムの重ね描き
ひとつのデータをコード番号によって3つに分けた例を図の左側に示す　図の下の数字は左からデータ番号・階級の幅・選択コード(3字)・累積度数を示す
ヒストグラムを　選択コード022　001　312 の順に下から積み重ねてできた図を右側に示す.

作図条件としてはX軸側Y軸側とも　つぎのものを入力する.

1. 図 示 範 囲
   a　最小値 X_{min}, Y_{min}
   b　最大値 X_{max}, Y_{max}

2. 軸 の 目 盛

データの処理でほかのプログラムと共通な点は省略する.　相関図の例を第13図に示す.

相関図の図示範囲は　使用者が任意に指定できるために　データの一部が図示範囲外に飛び出す場合もあり得る.　これを使用者が承知している場合はよいのだがもし図外となる数値が出たのを知らないでいると問題がおきる.　そこで図外になりプロットからはずされた値は別にプリントアウトしておく.　こうすれば　使用者への警告にもなり　図示範囲の検討をする上での参考にもなる.　このプリントアウトには作図条件そのほかの必要事項も印刷するその例を第13図（右）に示す.

数値の組 (x_i, y_i) をプロットする行程は

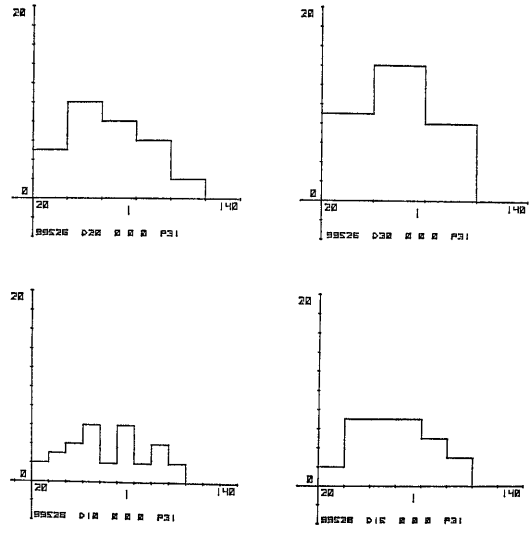
PLT(x_i, y_i)；ペンを上げる.

という命令で　これを繰り返せばよい.　上の命令のあとに停止 (STOP) 命令を付けて　1点打つごとに動作が停止し　プログラムを進めるキーを押すと次の1点を打ちまた動作が止まるという方法もできるようにしておく.　これを利用すると　打たれた点に使用者が番号そのほか必要なメモを　そのつど書き添える事ができて便利である.　図の重ね描きもできる.　相関図の場合は重ね描きと言っても点の追加に過ぎないので　プログラムもヒストグラムに比べてずっと簡単である.

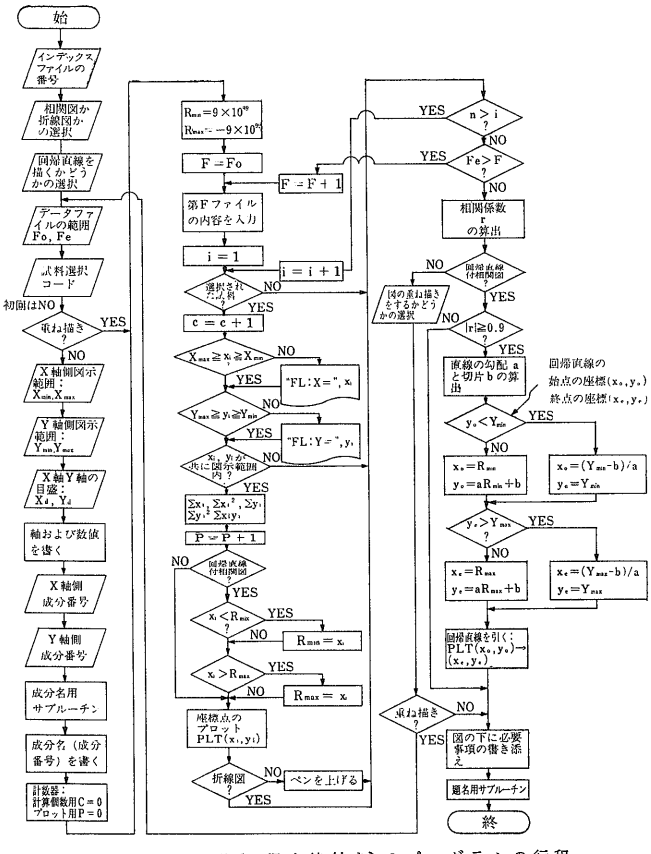
また　ペンを上げる命令を略して　ペンをおろしたま

```
JOB=
              99526
FILE:FRM=,TO=
                  5
                  8
SELECT=
                  0
                  0
                  0
X:FRM=           20
TO:X=,Y=        140
                 20
DIV:X=,Y=
                 10
                  2
CMP=              1
          2.000
          3.000
          4.000
          6.000
          2.000
          6.000
          2.000
          4.000
          2.000
          0.000
          0.000
          0.000
T=
         31.000
```

第10図
ヒストグラムの記録用プリントアウト
上からデータ番号(作業番号)・磁気テープのファイルの区間・選択コード・X軸側の図示範囲の最小値・両軸の最大値・軸のきざみ（X軸側は階級の幅を意味する）. 入力データの成分番号・各階の度数およびその合計（第11図左上のヒストグラムの記録).



第11図
階級の幅の違いによるヒストグラムの変化
同じデータでも階級の幅によってパターンが異なる.
階級の幅は　左下10　右下15　左上20　右上30

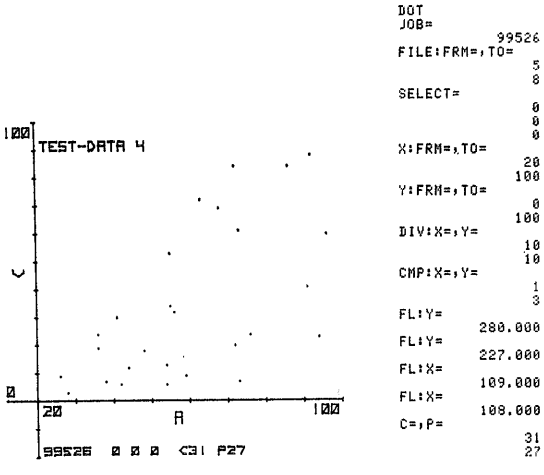


第12図　相関図(回帰直線付き)のプログラムの行程

ま上の操作をすると折線図が描ける．　なお　この場合一般には　片方のデータの組がある規則的な順序に並べられている必要があろう．

3.2.2. 回　帰　直　線

相関図に回帰直線を引く事もできる．　ただし筆者の



プログラムでは　相関係数の絶対値が0.9以上で　使用者が希望するとき　という条件を付けている．　そのわけは前の章(2.3)で説明した．回帰直線を引いた例を第14図に示す．

回帰直線の勾配 a と切片 b を求めるには相関係数の計算が必要なので　その行程を付け加えておく．　この計算は図示範囲内にある点の数値に対してだけ行なう．

回帰直線の線分は図中の点の分布する範囲にだけ引くように工夫するとよい．　それには図示範囲内にある点のX座標の最小値と最大値をそれぞれメモリー(R_{min} および R_{max})に記憶させる．　いま最小値を求める方法を考えてみよう．　初値は $R_{min}=9\times10^{99}$ (十分に大きい値)とし x_i が $x_i<R_{min}$ ならば R_{min} へ x_i の値を入れる．　この操作を $x_1, x_2, \cdots, x_n$ について行なった結果 R_{min} にはその最小値が入れられている．　最大値も同じような手順で　R_{max} に入れる事ができる．

回帰直線の式を

$$y=ax+b \quad (a\neq 0) \cdots\cdots (11)$$

とする．　いまこの線分の始点 (x_0, y_0) について考えてみると　その座標の値は

$$x_0=R_{min} \quad y_0=aR_{min}+b \cdots\cdots (12)$$

である．　このうち x_0 値はつねに図示範囲内にあるが y_0 は必ずしもそうとは限らない．いま仮りに y_0 の値が図の下方へ飛び出す($y_0<Y_{min}$)とすれば始点を上の直線上で移動して図の下限の境界線上にもって来る必要がある．　すなわちこの場合の始点は

$$x_0=\frac{Y_{min}-b}{a} \qquad y_0=Y_{min} \cdots\cdots (13)$$

とする．　$y_0>Y_{max}$ の場合は上の式の Y_{min} を Y_{max} に置き換える．　同様にして終点 (x_e, y_e) についても必要な処理をする．　回帰直線を引く命令は

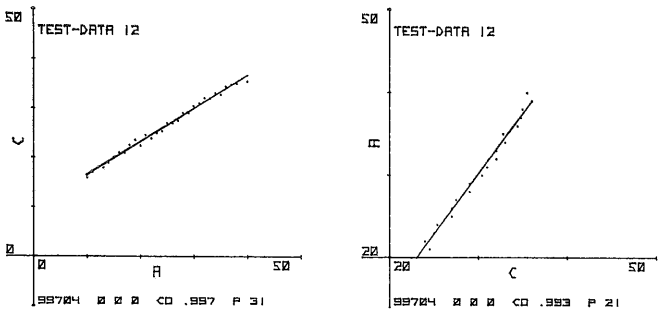
PLT(x_0, y_0)→(x_e, y_e)；ペンを上げる

となる．　始点が図の下方になる例を第14図(右)に示す．

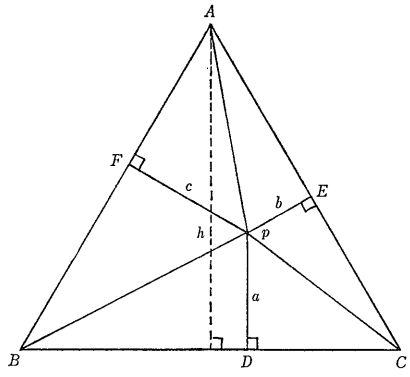
3.3. 三　　角　　図

3種の変量 $A(a_1, a_2, \cdots, a_n)$ $B(b_1, b_2, \cdots, b_n)$ $C(c_1, c_2, \cdots, c_n)$ の対応する数値の組 (a_i, b_i, c_i) の相互の比率を正三角形内の点に対応させて表わしたもの

第13図
相関図とそのプリントアウト
相関図の成分名(AとC)は成分名用サブルーチンによる．　図の下の数値でC23は計算数　P27はプロット数で両者に差があるのは図示範囲から4個の数値が飛び出した事を示す．
プリントアウトの成分番号(CMP)1と3は図のAとCの成分と対応する．　その下に図示範囲から出た数値4個が示され　それがどちらの軸で起きたかもわかる．



第14図 回帰直線付き相関図
右の図は左の図の成分を入れ換え 作図条件も変更したもので これにより回帰直線の始点が図外となった. この場合はX軸上に始点を求め直して回帰直線を引いている



第15図 三角図の原理
正三角形では $h=a+b+c$

が三角図である. 地学関係の研究ではしばしば使われて なじみの深い図でもある.

三角図は 正三角形内の任意の点Pから各辺におろした垂線の長さa, b, cの和がその正三角形の高さhに等しいという性質を利用したものである.

蛇足ながら これを証明しておこう. 証明法はさまざまあるので各自試みられたい. 筆者が考えたものをつぎに記す.

まず1辺の長さが2の正三角形ABCを考える. この内部に任意の点Pをとり Pから辺 BC, CA, AB におろした垂線の長さをそれぞれa, b, cとする. またPと頂点A, B, Cをそれぞれ直線で結ぶ. これらを第15図に示す. このとき 各辺の長さは

$$BC=CA=AB=2$$

だから $\triangle BCP$, $\triangle CAP$, $\triangle ABP$ の面積はそれぞれa, b, cとなる. また $\triangle ABC$ の高さはhだから $\triangle ABC$ 面積はhとなる. 図から明らかに

$$\triangle ABC=\triangle BCP+\triangle CAP+\triangle ABP$$

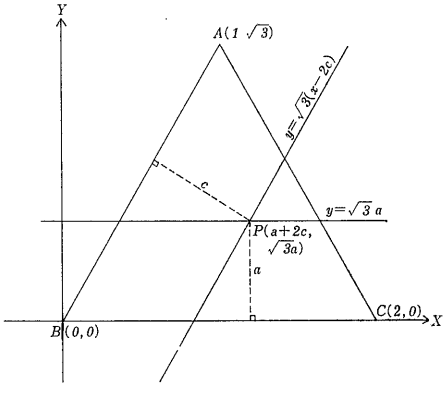
したがって

$$h=a+b+c$$

これで証明が終った.

3.3.1. プログラム

三角図では作図条件の指定はなく 3種の成分番号を入力するだけなので 使い方は簡単で プログラムも比較的短かい. ほかのプログラムと共通する点の説明は省略する.



第16図 三角図のプログラムでの正三角形の配置とプロットすべき点Pの求め方

ではXY座標上に正三角形をどのように置き プロットする点をどのように決めたら良いか述べよう. 筆者がいくつか検討したところでは この正三角形は頂点のひとつを座標の原点に置き 1辺をX軸上に置いてその長さを2とすると 式が簡単になるようだ. これを第16図に示す. すなわち各頂点の座標はA(1, $\sqrt{3}$) B(0, 0) C(2, 0)とする.

つぎに点Pの座標の求め方について述べる.

数値の組 a_i, b_i, c_i (ただし $a_i\geqq 0$, $b_i\geqq 0$, $c_i\geqq 0$) の和

$$S=a_i+b_i+c_i$$

を求め $S>0$ならば ($S=0$のときはあとで述べる) Sに対する a_i, b_i, c_i の割合を

$$a=\frac{a_i}{S},\quad b=\frac{b_i}{S},\quad c=\frac{c_i}{S}$$

とする. これらを第16図の正三角形の辺 BC, CA, AB からPまで距離に対応させる.

まずaはBCに平行な直線を表わし それが点A(1, $\sqrt{3}$)を通るから その方程式は

$$y=\sqrt{3}a$$

と書ける. 同様にcはABに平行な直線でC(2, 0)を通るので その方程式は

$$y=\sqrt{3}(x-2c)$$

となる. この2直線の交点がPであり その座標は

$$P(a+2c,\ \sqrt{3}a)$$

として求められる. この点のプロット命令は

PLT($a+2c$, $\sqrt{3}a$); ペンを上げる

である.

なお $S=0$のときは 分母が0となるため上の分数計算ができない. しかし $a=b=c$の関係には変わりないのでこの場合は便宜的に

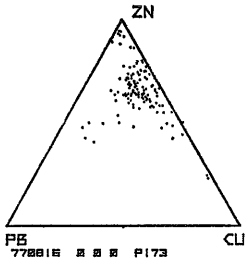
$$a = c = \frac{1}{3}$$

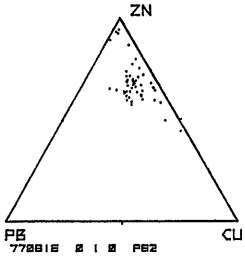
として取扱う．　$S=0$ の場合の対策をしておかないと　不条理演算となって計算が停止してしまうのは申すまでもない．　三角図のプログラムの行程を第17図に示す．　また三角図の例を第18図に示す．

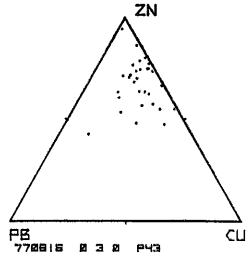
参　考　文　献

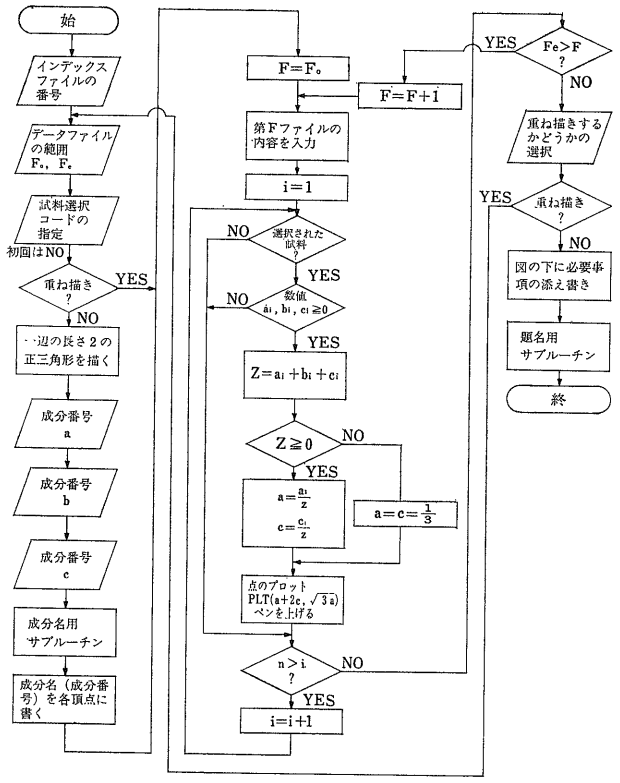
高橋宏一・脇本和昌・平野勝臣（1975）：統計学要論　206 p.　共立出版

吉井守正（1977）：卓上型電子計算機によるいくつかの計算例　その1　配列をもつデータの処理．地質ニュース　no. 275　p. 16—19









第17図　三　角　図　の　プ　ロ　グ　ラ　ム　の　行　程
成分番号 *a*, *b*, *c* の順で正三角形の上の頂点から左回りに　各成分の位置が定められる

第18図　三　角　図　の　例
ある岩石の化学分析値をもとに描いた例．　左上が全データ（選択コード000）左下(010)と右下(030)はそれを岩質ごとに表わしたものである．　コード番号を活用するとデータの解析が容易になる