電子計算機による統計教育の一方法

二　宮　理　憙*

(1973 年 11 月　受付)

On the Teaching of Statistics by Using of Computer

Satoki Ninomija

In this paper we discuss on the method that had been used for the teaching of statistics in our university from some years. By this method our students can understand more easy many statistical concepts: sampling distributions, variance of sampling and statistical variance in Monte Carlo methods.

Aoyama Gakuin University

ま え が き

筆者はこれまで約10年間，青山学院大学の経済学部，経営学部および理工学部で統計学の講義をおこなってきている．以下で述べる方法はここでの経験によって作られたもので，現在実際に用いられている電子計算機（主に IBM 7040）を使っての統計教育の一つの方法である．この方法は，現在では，理工学部経営工学科２年次生を対象におこなわれているが，この講義は統計学の一般論を教えるため「統計学入門」(本間鶴千代著）をテキスト（以下テキストとあるのはこの本である）として用いており，そこに載っているいろいろな定理，分布論などに関する課題を計算機のシミュレーションを用いて，確めたり，あるいは具体化して理解するように指導している．さらにはプログラム技術と演習の言語の統計的性質の理解を深めるために言語処理の問題も加えてある．

計算機を用いての演習は，今のところ次のような 14 項目を用意しいる．

① ヒストグラム，平均，分散
② 一様分布
③ 正規分布
④ 分散の計算における誤差
⑤ チェビシェフの定理
⑥ 中心極限定理
⑦ χ^2 分布
⑧ t 分布
⑨ F 分布
⑩ Cauchy 分布
⑪ ２次元正規乱数
⑫ 文字の出現頻度の計算
⑬ 単語の出現頻度の計算
⑭ π の計算

これらの課題を学生に，いわゆる宿題として与え，レポートとして提出させている．

各項目の詳細については以下の節で述べる．

統計的方法は便利なためか，ときとしてあやまって用いられているようである．これは数式

---

* 青山学院大学理工学部

の変形だけで統計学をマスターしたと考えているからではなかろうか．この方法のねらいは数式の変形だけでなく，定理や理論のもつ真の意味を学生に理解させるためである．さらにこの方法は解析的な能力があまりない学生にも適応でき，良い成績を上げている．したがって，この方法は文科系，社会科系の学生とか，高校生にも向くものと思われる．

なお，プログラムは，学生に作らせるのが原則であるが，プログラムのあまりに上手でない学生のために使いやすいサブルーチンの型にまとめてあり，それを使わせる場合もある．

1. ヒストグラム，平均，分散

統計ではある一つの集団を全体としてどのような性質を持っているかについて調べるのが大

例 ある学校で100人の生徒の体重を測って次の資料を得た．これを用いてヒストグラムをつくれ．

51.0 (kg)	48.5	53.7	49.5	49.4	44.7	45.0	53.5	41.5	56.8
48.5	42.4	52.0	49.7	59.0	55.0	53.2	47.6	64.0	52.3
47.3	57.0	53.5	38.0	55.2	50.1	48.5	54.2	52.4	46.0
57.0	53.5	46.4	55.0	45.1	53.6	54.0	53.7	41.7	55.3
48.7	48.4	58.5	46.5	39.5	53.8	46.0	46.0	52.5	56.5
45.0	47.6	48.6	52.4	52.4	49.4	54.0	48.2	43.0	57.0
44.2	46.6	44.4	51.0	40.0	65.3	49.0	42.2	45.2	52.0
55.0	49.8	48.7	52.5	57.4	49.0	52.4	54.0	50.5	52.0
48.0	50.4	46.2	46.8	40.0	57.0	49.4	58.0	45.5	49.0
55.2	46.0	43.0	47.5	53.2	43.8	49.0	61.0	38.2	44.6

図 1-1

```
      SUBROUTINE HIST(N,A,D,X,AA,AB)
      DIMENSION X(N),AKUGI(300),MCOUNT(300),STAR(100)
      DATA STAR/100*1H*/
       S=0
      S2=0
      DO 100 I=1;N
      S=S+X(I)
      S2=S2+X(I)*X(I)
  100 CONTINUE
      FN=N
      AVR=S/FN
      VAR=S2/FN- AVR*AVR
      ABW=SQRT(VAR)
      WRITE(6,110) N,AVR,ABW
      AMAX = A
      AMIN = A
      DO 10 I=1,N
      XXX = X(I)
      IF(XXX.LT.AA.OR.XXX.GT.AB) GO TO 10
      AMAX = AMAX1(AMAX,XXX)
      AMIN = AMIN1(AMIN,XXX)
 10   CONTINUE
      II = ALOG10(D)
      IF(AMIN.GE.A) GOTO 20
      C = 0
      IF(AMIN. LT.0.0)C=-0.99999999
      MIN = AMIN*10.**(-II)+C
      AMIN = MIN*10**II
 20   MAX = AMAX*10.**(-II)+.999999
      AMAX = MAX*10**II
      II = (AMAX-AMIN)/D+1.00000001
      IF(II.GT.300) GO TO 90
      III = II+1
      DO 30 I=1,II
      MCOUNT(I) = 0
 30   AKUGI(I) = AMIN+FLOAT(I-1)*D
      MCOUNT(III) = 0
      DO 50 I=1,N
      DO 45 J=1,II
      IF(X(I).GT.AKUGI(J)) GO TO 45
      MCOUNT(J) = MCOUNT(J)+1
      GO TO 50
 45   CONTINUE
      MCOUNT(III) = MCOUNT(III)+1
 50   CONTINUE
      WRITE(6,3000) (I,I=10,100,10)
      L = MCOUNT(1)
      IF(L.LE.0) GO TO 55
      LL = L
      IF(L.GT.100) LL = 100
      WRITE(6,6000) L,(STAR(J),J=1,LL)
 55   DO 80 I=2,II
      L = MCOUNT(I)
      IF(L.LE.0) GO TO 60
      LL = L
      IF(L.GT.100) LL= 100
      WRITE(6,4000) AKUGI(I-1),AKUGI(I),L,(STAR(J),J=1,LL)
      GO TO 80
 60   WRITE(6,4000) AKUGI(I-1),AKUGI(I),L
 80   CONTINUE
      L = MCOUNT(III)
      IF(L.LE.0) RETURN
      LL = L
      IF(L.GT.100) LL = 100
      WRITE(6,5000) L,(STAR(J),J=1,LL)
      RETURN
 90   WRITE(6,9999) II
      WRITE(6,9998)   AMIN,AMAX
      STOP
  110  FORMAT (1H1,20X, I8,5X, 5HAVR =,  E15.7, 5X, 5HABW =, E15.7)
 3000  FORMAT(1H0,26X,1H0,10(7X,I3)/
     *     1H ,3X,5HHAN-I,8X,4HKAZU,1X,10(1H+,9(1H-)),1H+)
 4000  FORMAT(1H ,1H(,F6.2,6H **** ,F6.2,1H),1X,I4,2H I,100A1)
 5000  FORMAT(1H ,1H(,5X,8HMUGENDAI,5X,1H),1X,I4,2H I,100A1)
 6000  FORMAT(1H ,2H( ,16HMAINASU-MUGENDAI,3H ) ,I4,2H I,100A1)
 9998  FORMAT(1H0,10X,4HMIN=,E15.7/1H0,10X,4HMAX=,E15.7)
 9999  FORMAT(1H1,10X,3HII=,I8)
      END
```

P-1 ヒストグラム，平均値と標準偏差のプログラム

N ： データの個数

A ： ヒストグラムの区間の端点

D ： ヒストグラムの区間の巾

X ： データ

AA： ヒストグラムの最小値

AB： ヒストグラムの最大値

このプログラムがCALLされると平均値と標準偏差を計算して印刷し，つづいて区間AAとABとの間にAを区間の端点として区間の巾をDとしたヒストグラムを画くプログラムである．

切であるという観点から与えられたデータのヒストグラムを作るプログラムを作るという課題を与えている．テキストの 75 頁の例題（図 1-1）のヒストグラムを作るのはあまり困難ではないが，一般にヒストグラムを作るにはどのような条件が必要かを考えてゆくと非常に複雑な問題になってくる．学生に対して

①級の数（あまり少くても，あまり多すぎてもだめであり，また，用いるデータの個数にも依存する.）

②区間の巾（はんぱな数は用いない．一般には等間隔を用いるが，両端の区間については必らずしもそうでない.）

③区間の端（はんぱな数は用いない.）

この三つの条件を満すようなプログラムを作れという課題となっているが，計算機のメモリーの大きさなど考慮すると非常に困難な問題となるので，P-1 のようなプログラムを作るように指導した.（あまり平たんになりすぎている場合は区間の巾を大きくしてやり直す）

集団の性質を見るには，適切なヒストグラムがよいのではあるが，そのヒストグラムはまだ多くの数字を含みすぎている.（例えば，区間のはしとか度数）．そこで，これをもう少し簡単な量で現わすことはできないであろうか，と考えられよう．その一つの方法として平均値というものがある．平均値というものは集団の特性を示す一つの尺度であり，非常に簡単な尺度である．それがためにいろいろと使って便利でもあるが，また危険性をも含んでいる．なぜなら，いろいろな個々の特性全部をただ一つの数値にまとめてしまったからである（例えば平均値が同じで分散の異なるヒストグラムを与える)．そのような欠点を補なうために分散というものがあり，これは集団のばらつきの状態を示す一つの尺度である．以上の観点から平均と分散（および標準偏差）を計算するプログラムをもサブルーチンの型にまとめさせ，いつでも使えるように準備させた（2 以後の課題に対しては，必要な際いつでも使うように指導している).

2. 一 様 分 布

分布論に関する課題を処理するために，いろいろな分布を作る一番の基礎となる一様分布の乱数をまず作らせた．念のためここでは，図 2-1 の順序で，いろいろな分布の擬似乱数を作らせている．

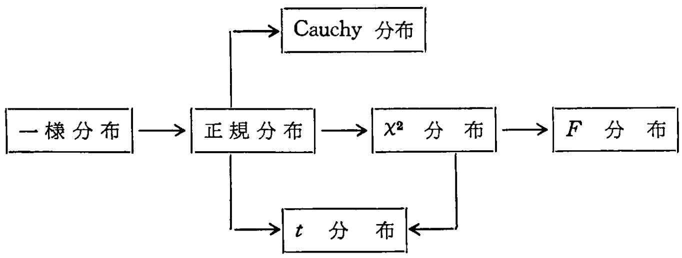


図 2-1 分 布 の 関 係

一様分布を基にして，いろいろな乱数を作る．それぞれの分布は矢印の関係で結ばれている．

一様分布の乱数を作る方法は，いろいろあるが，計算機の計算時間（プログラムの簡単さと関係する）を考慮にいれ，その一様乱数としての性質はいろいろと議論されている[3)]が，混合合同法を用いた．すなわち

$$x_{n+1} = a\,x_n + b \pmod{p} \qquad (2,1)$$

ここで，a, b, p, x_0 は定数である．

乱数としての性質をよくするための a, b, p, x_0 に関していろいろな条件がわかっているが[3)]，学生にはそれを教えないで数回の実験をさせた．提出させたレポートは勿論，実験結果はヒストグラム，平均，標準偏差を計算機で書かすようにしてある．このような教育方法を始めた初期の頃，用いる計算機がないため，学生の自宅の電話番号の下 4 桁を用いたこともあったが，

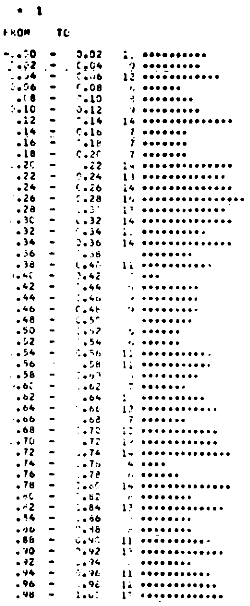


図 2-2　一　様　分　布

(2-1) に従って計算した一様乱数の1例. 区間の巾が小さいため一様分布から離れているように見えるが，モンテカルロ法を用いるには，このような乱数が出現する場合もある.

これは少しかたよっているようである.

作った乱数に対しては当然のことながら，乱数としての検定が問題となるが，ここではいわゆる検定に対しては何ら特別の考慮をしてはいない．一様乱数であれば，よくながめて，規則がみいだせないならばよいという立場をとっている.「教育とはすべてについて説明することではない」という立場である.

3. 正　規　分　布

x_i を区間 $[0,1]$ の一様乱数とするとき，

$$Y=\sum_{i=1}^{12} x_i - 6 \tag{3,1}$$

なる Y は，ほぼ $N(0,1)$ に従う正規乱数となる．解析的には，12 個でなく，無限個の和でなくてはならないが，12 個でも充分に近似できることを中心極限定理の項で示すようにしてある．実際には，このような乱数 Y を 500 個作り，平均，標準偏差，ヒストグラムを書かせた．しかしこのようにして作った正規分布はどこまでも擬似正規乱数の正規分布であり，真の正規分布でない（これより後で述べるいろいろな分布でも同様であるが）ことに注意を与え，特にこのような分布の端の方の性質は不用意に用いてはならないと教えてある．(図は図6を参照)

4. 分散の計算での誤差

計算機での計算では，数式としては同意でもその Algorithm が異なれば，その計算結果は異なったものとなりうることを充分に理解しておく必要がある．これは計算機が有限桁のメモリーしかもたず，いわゆる計算においては体 (Körper) ではなく一種の疑似体での演算をしていることを理解させるためである．この題目はそのための一つの材料である.

分散とは

$$\sigma_x{}^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2=V_1 \tag{4,1}$$

ここで　$\bar{x}=\dfrac{1}{N}\sum\limits_{i=1}^{N}x_i$

で与えられるが，簡単な代数的な計算により

	単精度	2倍精度
x_0	0.10000000E 01	0.10000000000000D 01
x_1	0.10000001E 01	0.10000001043081 3D 01
x_2	0.10000002E 01	0.10000001937151 0D 01
x_3	0.10000003E 01	0.10000002980232 2D 01
x_4	0.10000004E 01	0.10000004023313 5D 01
x_5	0.10000005E 01	0.10000005066394 8D 01
x_6	0.10000006E 01	0.10000005960464 5D 01
x_7	0.10000007E 01	0.10000007003545 8D 01
x_8	0.10000008E 01	0.10000008046627 0D 01
x_9	0.10000009E 01	0.10000008940696 7D 01
$\bar{x}$	0.10000004E 01	0.10000004500150 7D 01
V_2	0.44703484E-07	0.82600593032111 7D-13
V_1	0.83399952E-13	0.82325257722004 6D-13

図 4-1　分 散 の 計 算

V_1(4-1) と V_2(4-2) の式でこのような大きな差がでるのは (4-2) における x_i^2 の計算で 15 桁となる数が単精度の場合には，後の方で捨てられてしまうからである．従って，倍精度で計算をすれば大きな差はでない．それでも少しの差のあるのは 10 進法の数を 2 進法に変換して計算し，さらにそれを 10 進法に変換しているからである．(2倍精度で x_1 は 0.100000010430813D01 となっているのは単精度の数を 2 倍精度に変換した場合に起ったものである.) この計算は IBM 7040 でおこなった．これは 2 進法で 36 桁のメモリーをもっている．一般的にこのような計算誤差は式，計算機，およびそのコンパイラーに関係するので注意する必要がある.

$$\sigma_x^2=\frac{1}{N}\sum_{i=1}^{N}x_i^2-\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2=V_2 \tag{4,2}$$

となることがわかる．ところがある x_i に対してはこの (4-1) と (4-2) が等しくならない．学生には x_i として 1.0000000，1.0000001，1.0000002，$\cdots$，1.0000009 を与え，(4,1) と (4,2) で計算させる．計算の際，単精度と倍精度の二つの方法で実験させた．さらに (4,1) と (4,2) とでそのように異なる結果がでるような x_i の集合はどのような集合であるかを検討させた．もちろんこの答の 1 つは相対精度が小さい場合となるが，相対精度が同じように小さくともデータが異なれば異なった精度の分散となり，さらに同一のデータに対してすら，和の順序を変えれば異なった結果を得るなど，「計算」(実は擬似体）の複雑な性質の一端を味わわせた.

さらに最小二乗法を用いて $Y=aX+b$ の a, b を求める際，あるいは相関係数を求める場合などにも，このような現象がおこることを確かめさせ，統計学での計算で注意するべき点の多いことを喚起させた.

5.　チェビシェフの定理

チェビシェフの定理（テキスト 14 頁）を解析的に証明することはそれほど困難なものではない．積分学や確率論の初等的な知識だけで充分であろう．しかし，その証明ができるということと，この定理を本当に理解しているということには大きな差があるように思える．チェビシェフの定理とは；X は平均 μ（有限），分散 σ^2（有限）をもつ確率変数とすれば，任意の $k>0$ に対して

k	$1/k^2$ の値	実験結果
1	1.0	0.31
1.5	0.44	0.13
2.0	0.25	0.05
2.5	0.16	0.02
3.0	0.11	0.00
3.5	0.08	0.00

図 5-1　チェビシェフの定理

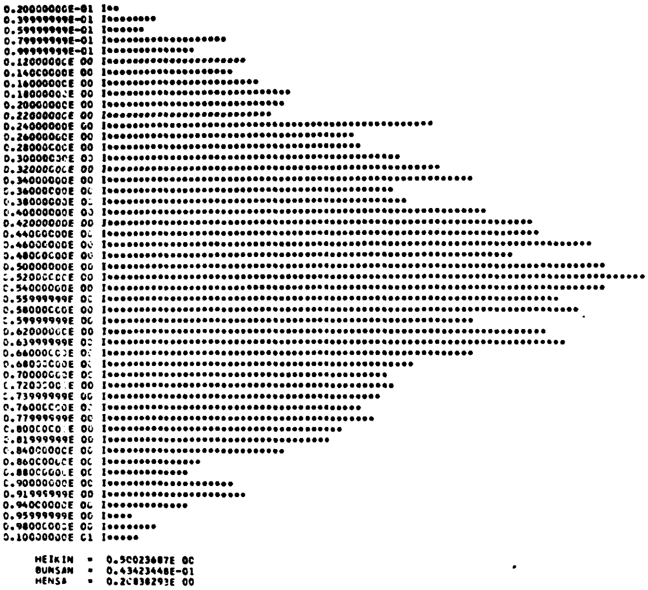
(5,1) 式において X を正規分布をする乱数とし，k を 1,0 から 0.5 づつ増して 3.5 まで変えて (5,1) の成立つことを確めた.

$$p(|X-\mu|>k\sigma)\leqq\frac{1}{k^2} \tag{5,1}$$

が成立することである．この定理の理解は，まず任意の k とはどんな意味か，確率変数とは，確率とはなどの疑問を解く例題ともなるように，いろいろな分布に従う乱数（確率変数）に対して，いろいろな k について実験して上の不等式が成り立つことを確めさせた．さらにこの不等式は一般的ではあるが能率が悪いことをも理解させる（正規分布ならば，その標準偏差で押えた方が良いこと).

6.　中 心 極 限 定 理

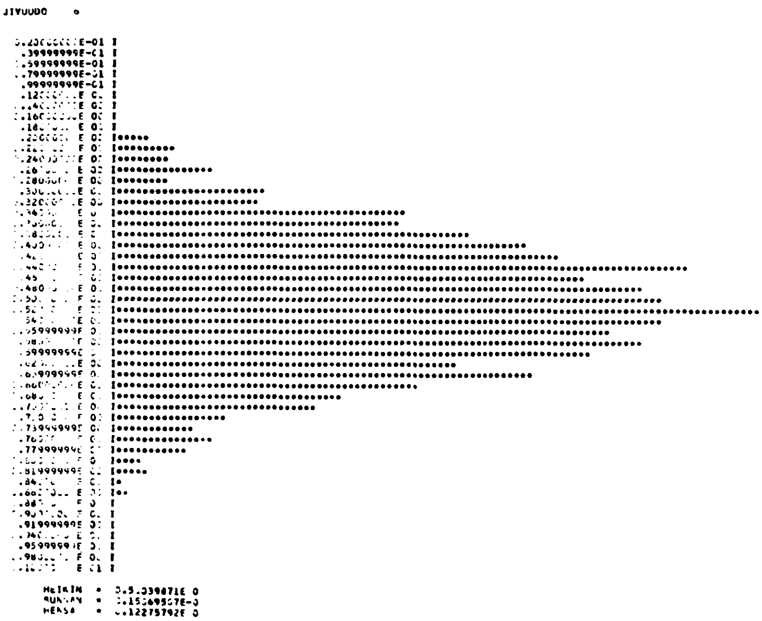
中心極限定理（テキスト 56 頁）も，解析的に証明するにはそれほどの困難をともなわない



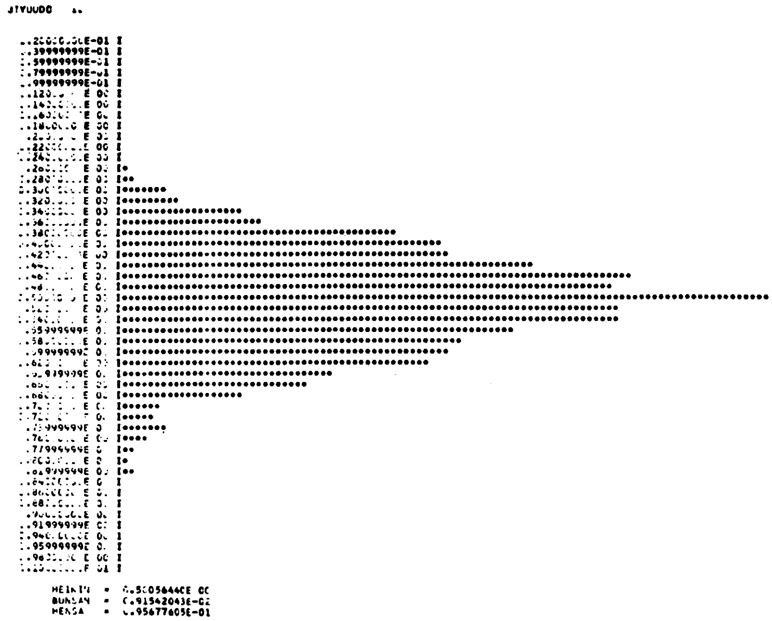
(6-1) 2個の和

(6-2) 3個の和

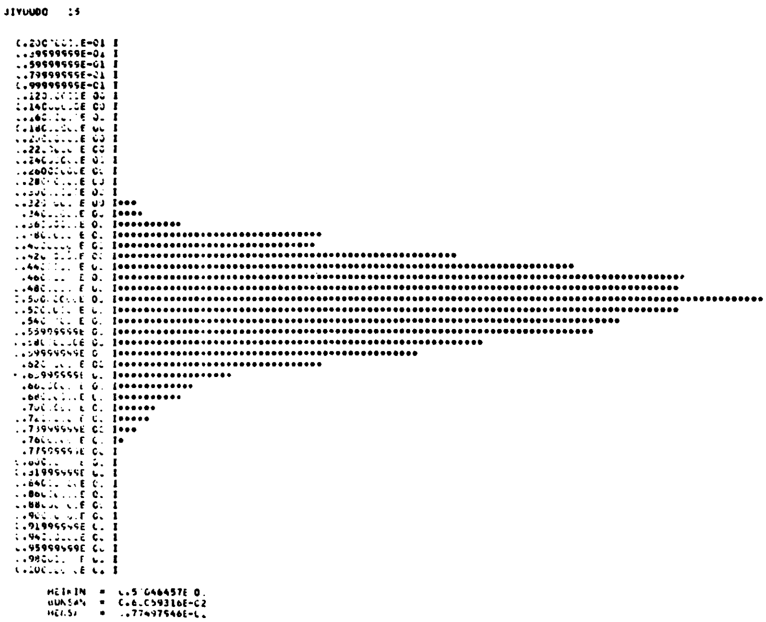
(6-3)　　5個の和



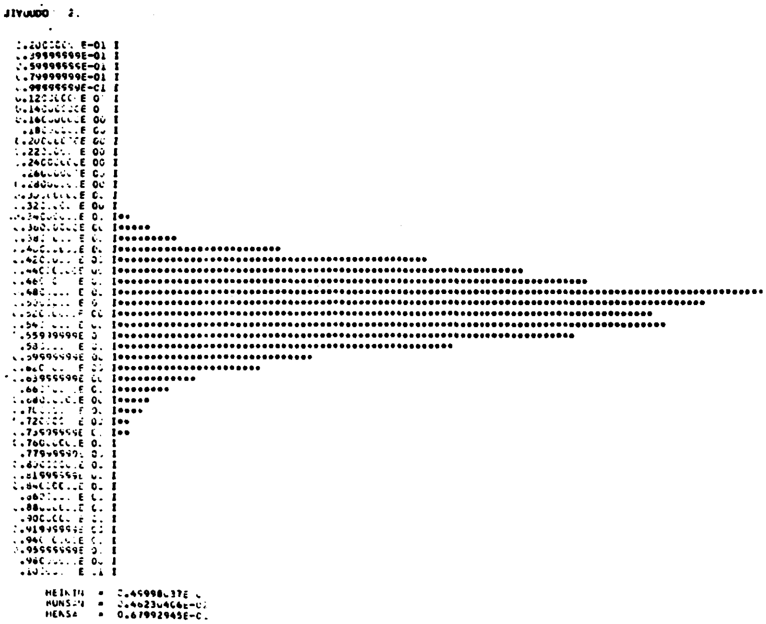
(6-4)　　6個の和



(6-5) 10 個の和



(6-6) 15 個の和



(6-7)　20 個の和

図 6-1 中 心 極 限 定 理

n 個の一様乱数の平均値を 500 組作り，そのヒストグラムを画かせたものである．n が大きくなるにつれて，標本平均が解析的な平均に近づき，標準偏差はしだいに小さくなり，平均値の回りに分布が集ってくることがわかる．

KOSU	HEIKIN	BUNSAN	HENSA	C
1	0.522181	0.083204	0.288452	0.288675
2	0.508971	0.043388	0.208298	0.204124
3	0.504462	0.028209	0.167956	0.166667
4	0.505182	0.020947	0.144731	0.144338
5	0.509202	0.017186	0.131094	0.129099
6	0.506506	0.014424	0.120101	0.117851
7	0.505374	0.012846	0.113341	0.109109
8	0.504730	0.011195	0.105904	0.102062
9	0.504730	0.009748	0.098732	0.096225
10	0.502720	0.009179	0.095805	0.091287

図 6-8 正 規 分 布 へ の 近 似 状 態

[0,1] の一様乱数を KOSU 個加えた場合の平均値 (HEIKIN) と分散 (BUNSAN) (および標準偏差 (HENSA)) とその理論値 (C) との関係を示したのもである．偏差とその理論値との差は N が大きくなれば充分に小さくなるといえよう．

(なかにはこのような証明さえ手こづる学生もいるが)，しかし，その意味を充分に理解し，統計的解析に上手に用いられるようになるのには大きな困難があるように思われる．それはこの一様乱数の分布論における極限操作は，n が案外小さい (n=6, 7 ぐらいで) 場合でも n の充分に大きな場合 ($n\to\infty$) の状態に近いことがあるからである．学生のなかには，これの証明の実際の分布の変化の具合を適格に把握しているのは少いようである．実験に用いたのは一様乱数 (X_i) の集合であり，$\bar{X}=(X_1+X_2+\cdots+X_n)/n$ のヒストグラムを n を 1, 2, 3, $\cdots$, 20 と変えて作らせた．n=1 では一様乱数そのものであり，n=2 では図 (6-1) に示すように三角分布，n=3 で少しつりがね型，そして n=6 ぐらいからもう正規分布とみなしてよいような分布になることがわかる．学生に対しては，n によって分布のばらつき具合と平均値の安定の具合を検討させた．

また，任意の分布をもつ観測データのグループについても，このようなことを試みさせ，標

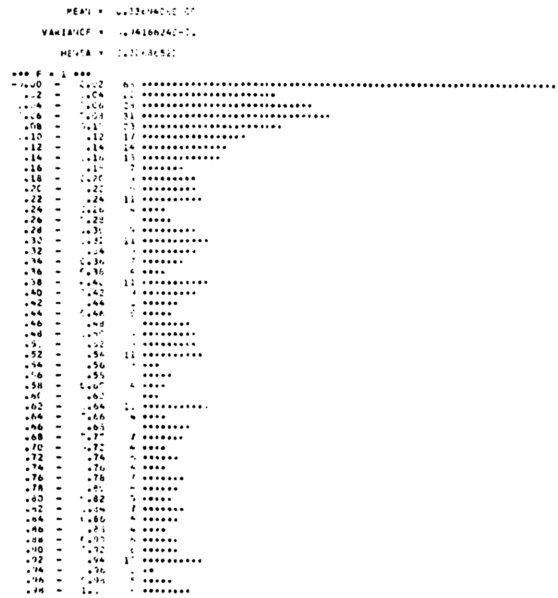
本平均の期待値は，母集団平均に等しく，分散は母集団分散の $1/n$ に等しい正規分布になる様子を実際にヒストグラムを書くことにより確めさせた．

7.　χ^2 分　布

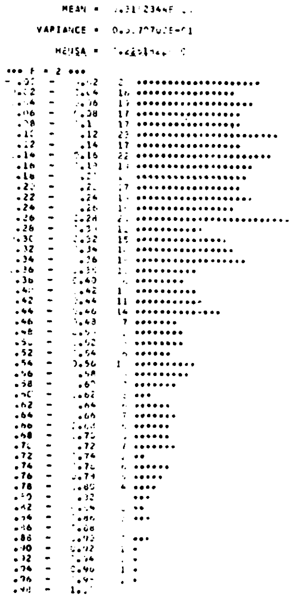
Y_i を $N(0,1)$ の正規分布に従う乱数とするとき

$$K=\sum_{i=1}^{N} Y_i^2$$

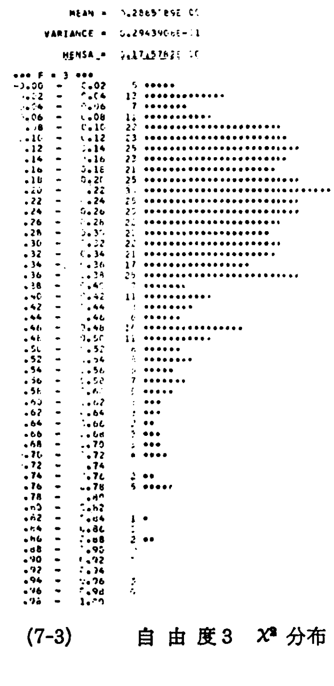
は自由度 N の χ^2 分布に従うことは，テキストにも証明が載っているが，これをシミュレーシ



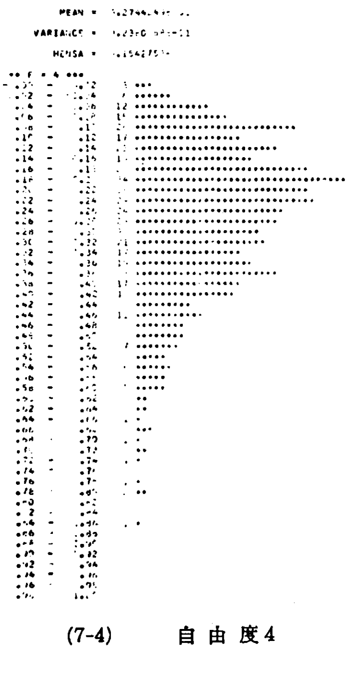
(7-1)　　自 由 度1



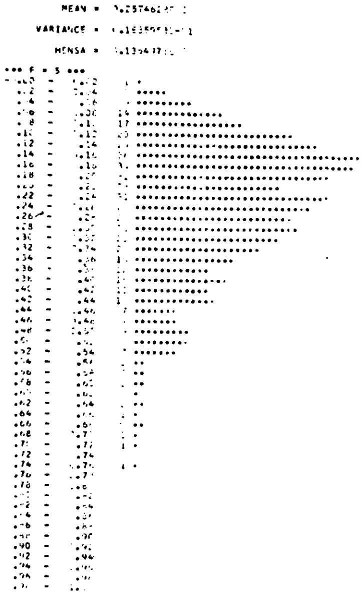
(7-2)　　自 由 度2



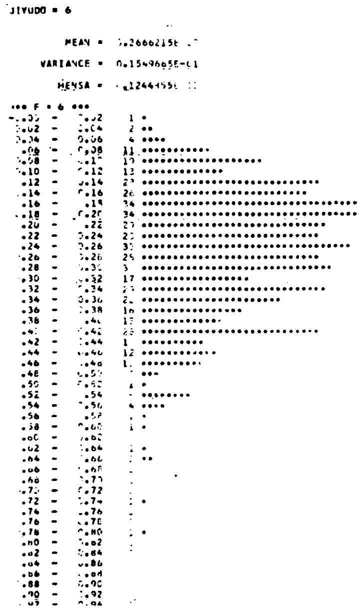
(7-3)　　自 由 度3　χ^2 分布



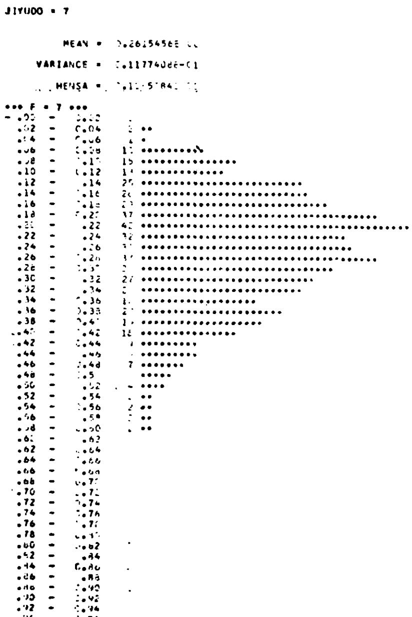
(7-4)　　自 由 度4



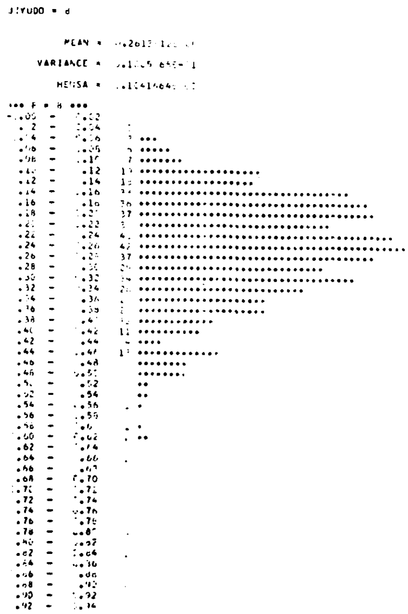
(7-5)　　自 由 度5



(7-6)　　自 由 度6

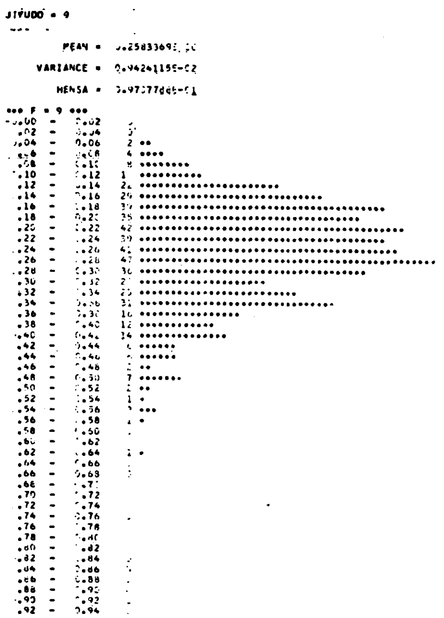


(7-7)　　自 由 度7

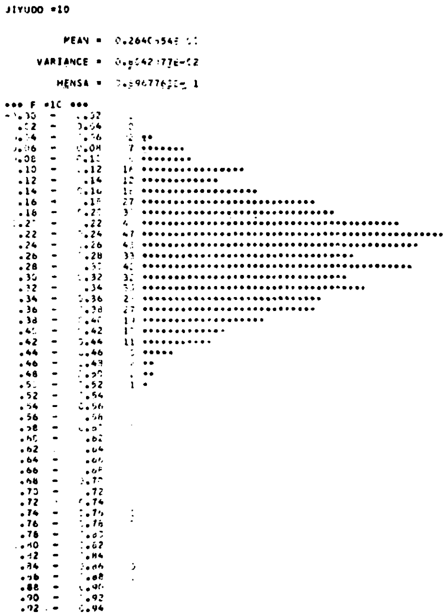


(7-8)　　自 由 度8

χ^2 分 布



(7-9) 自 由 度9



(7-10) 自 由 度10

図 7 χ^2 分 布

自由度が大きくなるにつれて凸凹の少ない安定した分布に近づいていることがわかる．このようなことは解析的に導くのは困難なことであろう．(自由度1のグラフの先（10.0 以上）の所は印刷してない.)

ョンで確めさせた．N が 1, 2, …と大きくなるに従ってその分布がどのようなものに近づいてゆくかということと，その安定性について調べさせた．テキストには，自由度による分布のグラフは載っているが，これだけでは安定性はわからない．(もちろん，式を解析的に解けばある程度わかるが)．あまり解析的な能力を用いないでこのような分布の安定性を検討できるのはこの方法の特徴といえよう．これにより，検定になぜ χ^2 分布が用いられるかが理解できる．

8. t 分 布

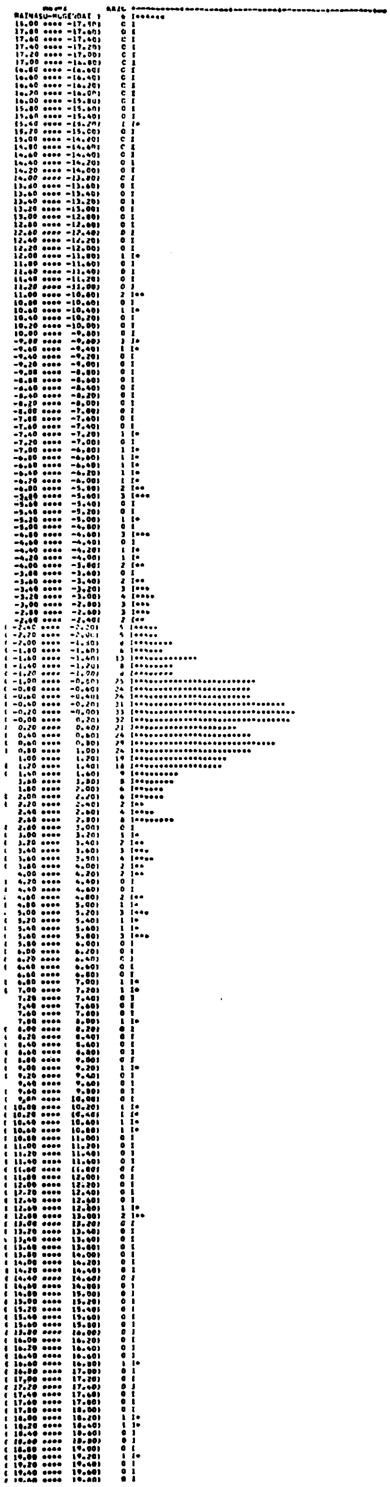
Y_i は $N(0,1)$ に従い，K_i は自由度 N の χ^2 分布に従い，かつ両者が独立なるとき

$$T_i=\frac{Y_i}{\sqrt{\dfrac{K_i}{N}}}$$

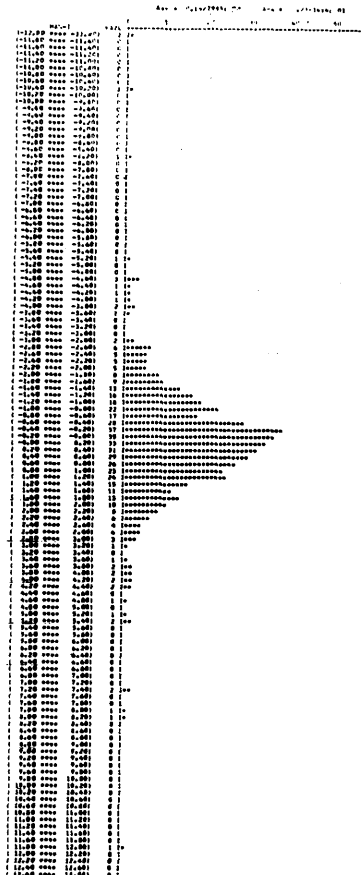
は自由度 N の t 分布に従う．(テキスト 63 頁)．これをシミュレーションで確めさせた．自由度 N により分布の変化の仕方はテキスト 63 頁に載っているが，実験の結果から，学生に対して，自由度1とか2の t 分布の様子などから自由度の小さい t 分布の不安定なことが，手にとるように示されたようである．(従って自由度の小さい場合の検定は意味のないこと)．また推定論，検定論などで，なぜ t 分布が用いられ，さらに大標本の場合，正規分布が用いられるかが理解できた（t 分布は正規分布に近づいてゆくので)．

9. F 分 布

KN, KM が独立でそれぞれ自由度 N, M の χ^2 分布に従うとき $\dfrac{KN_i}{N}\Big/\dfrac{KM_i}{M}$ は自由度 N, M の F 分布に従う．N と M をいろいろ変えて分布の型がどう変るかということと，その安定性がどうなるかについて実験させた．どのような N, M に対して実験すれば能率的にそのようなことがわかるかは学生自身に考えさせた．結果は当然のことであるが，図9からわか

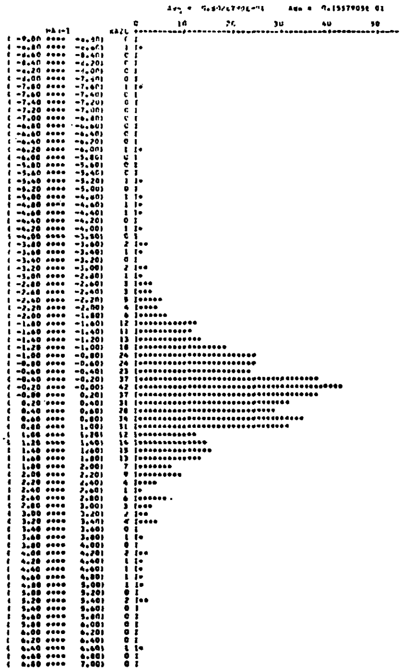


(8-1) 自 由 度 1

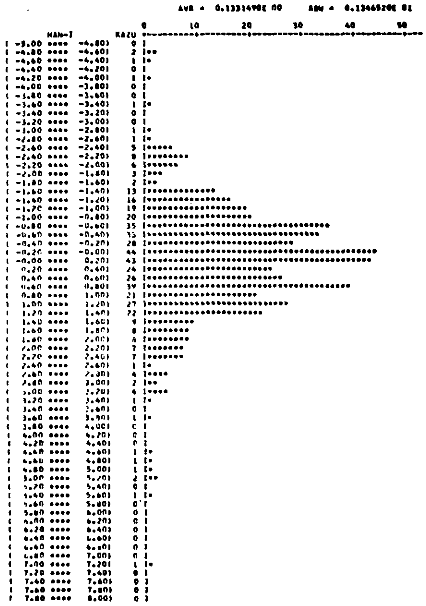


(8-2) 自 由 度 2

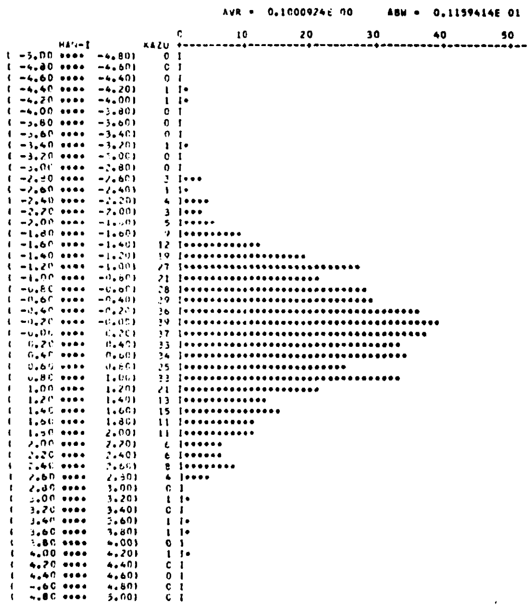
t 分 布



(8-3) 自 由 度 3

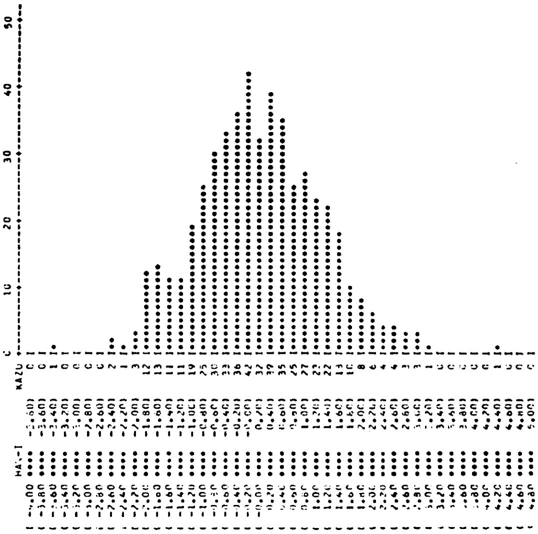


(8-4) 自 由 度 4

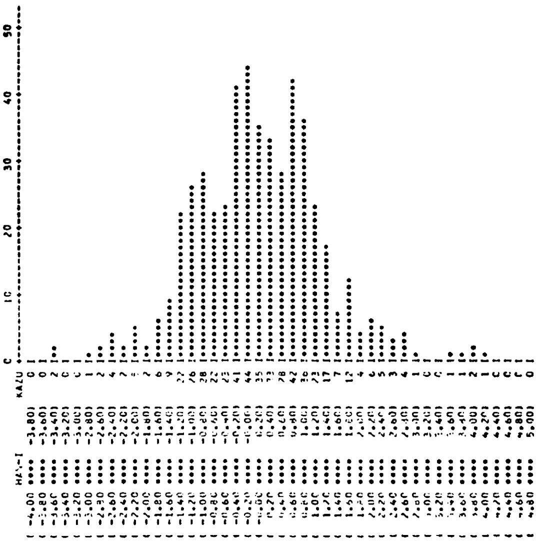


(8-5) 自 由 度 5

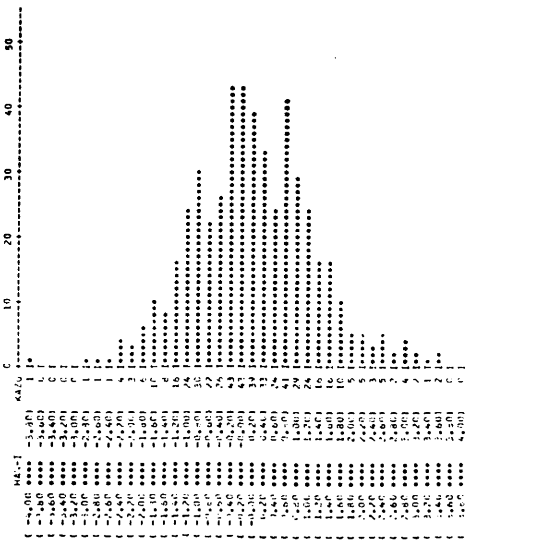
t 分 布



(8-6) 自由度 6

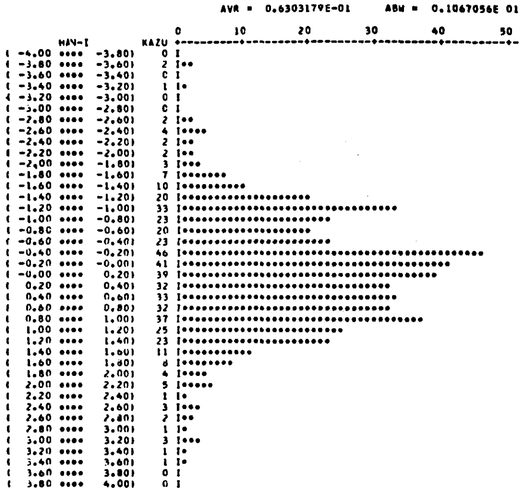


(8-7) 自由度 7

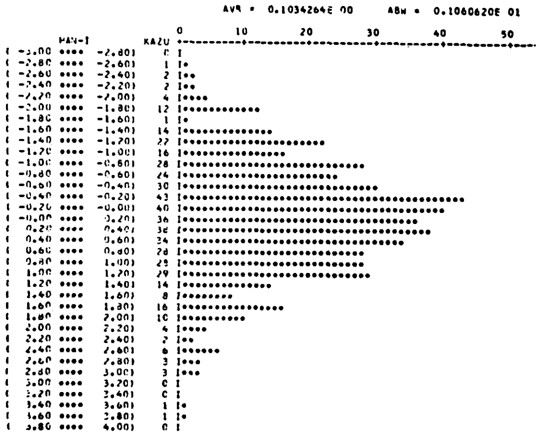


(8-8) 自由度 8

t 分布



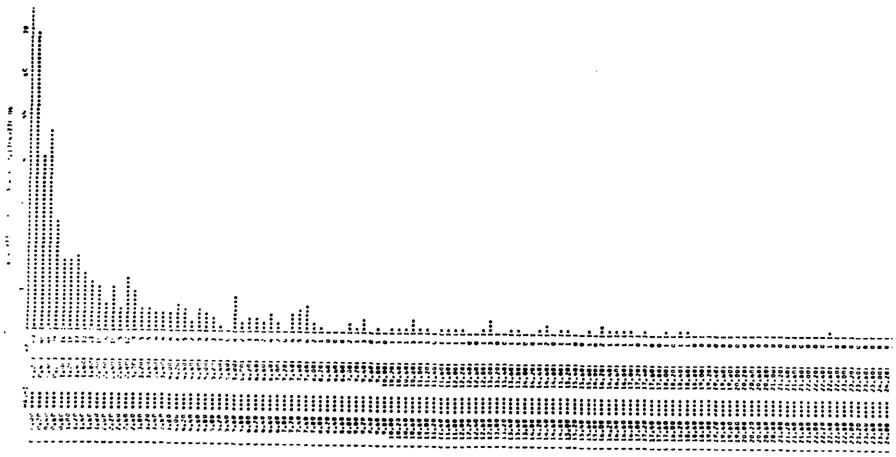
(8-9) 自 由 度 9



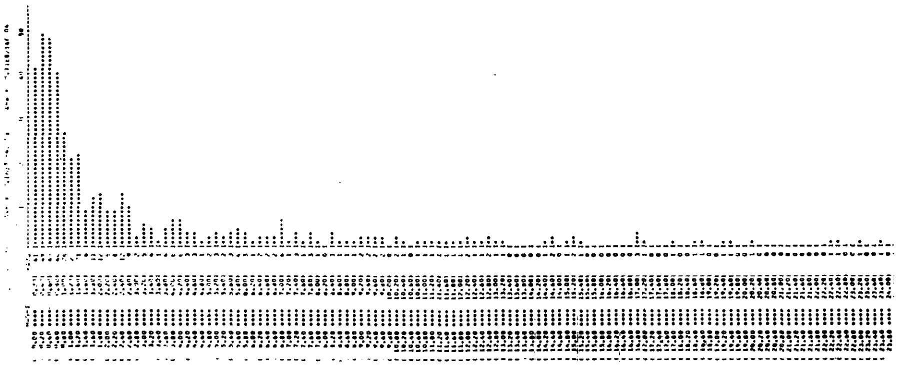
(8-10) 自 由 度 10

図 8 t 分 布

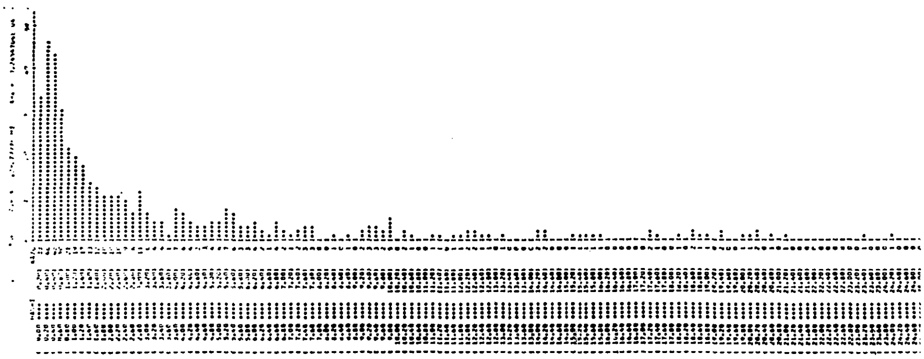
分布の自由度と分布のばらつきの程度がこれらの図から読みとられる．自由度1の分布では特にばらつきが大きいことに注意する必要がある．また自由度が大きくなると正規分布に近づいていることがわかる．



自由度 (1,2)



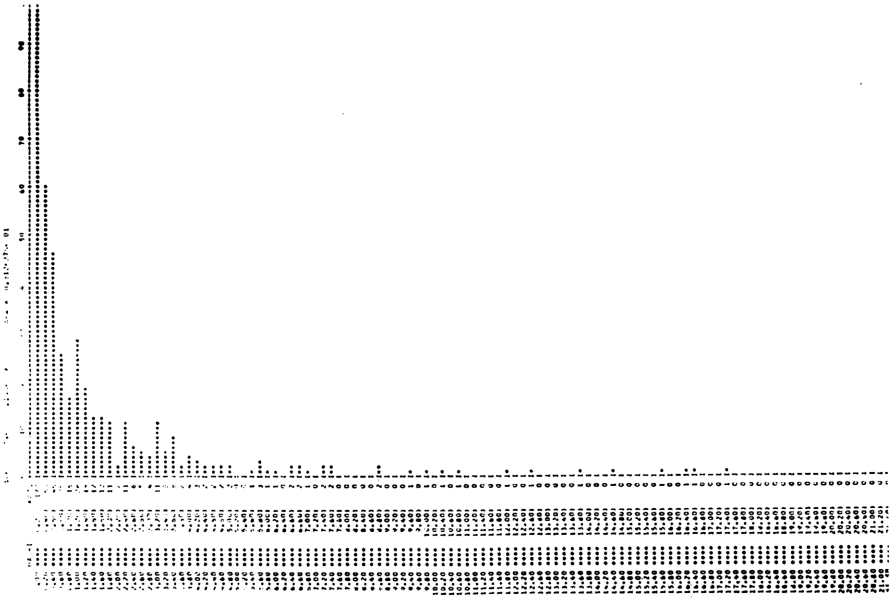
自由度 (1,5)



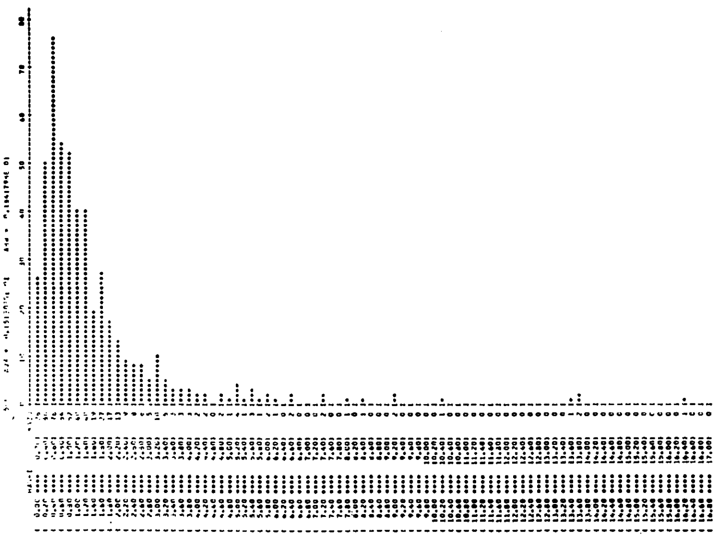
自由度 (1,10)

F 分布

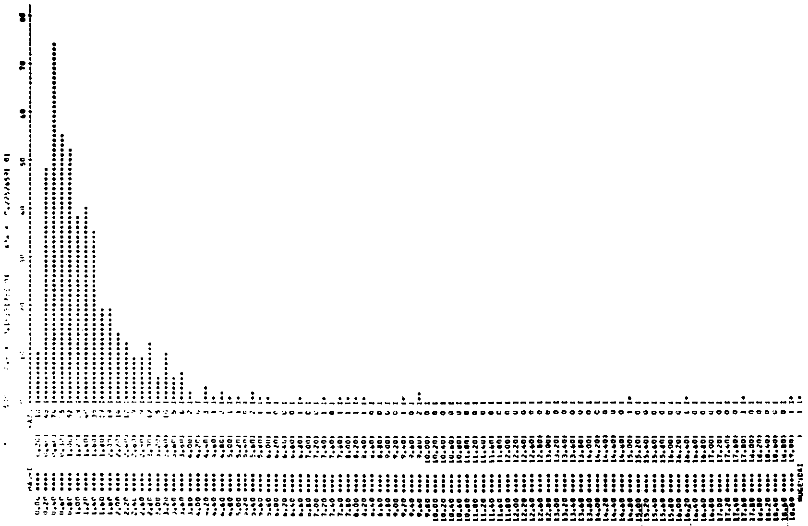
F 分布



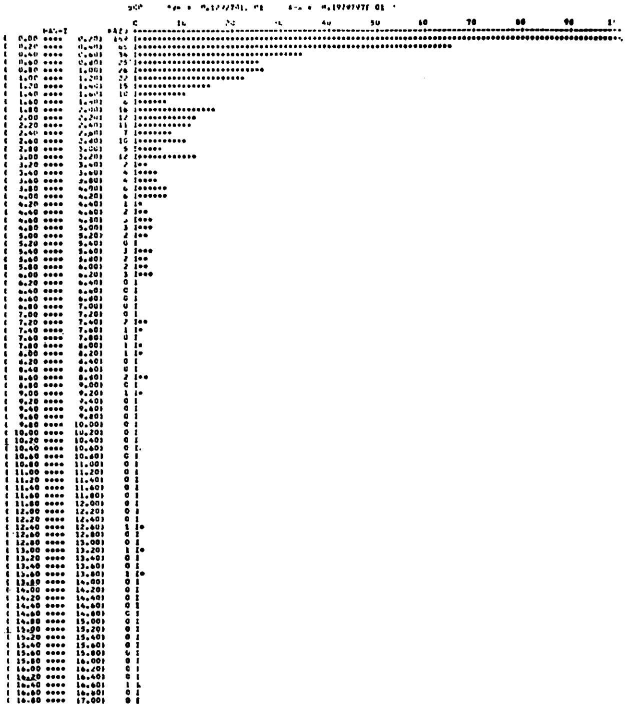
自由度 (5,1)



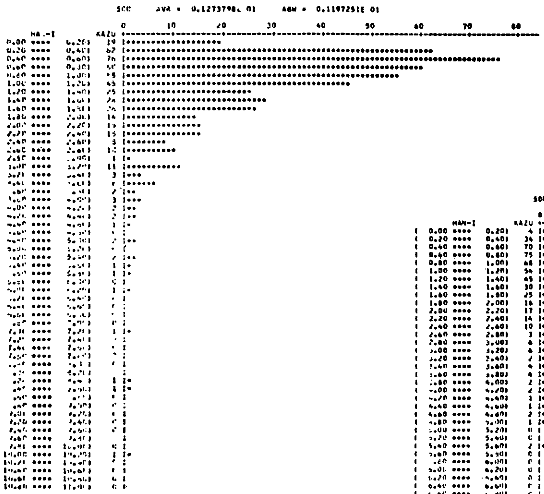
自由度 (5,5)



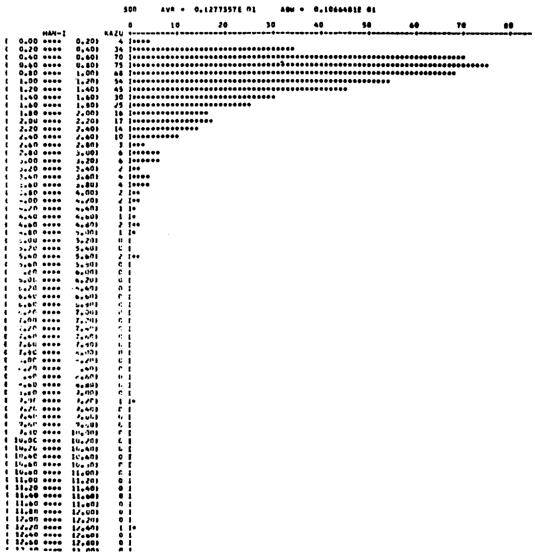
自由度 (5,10)



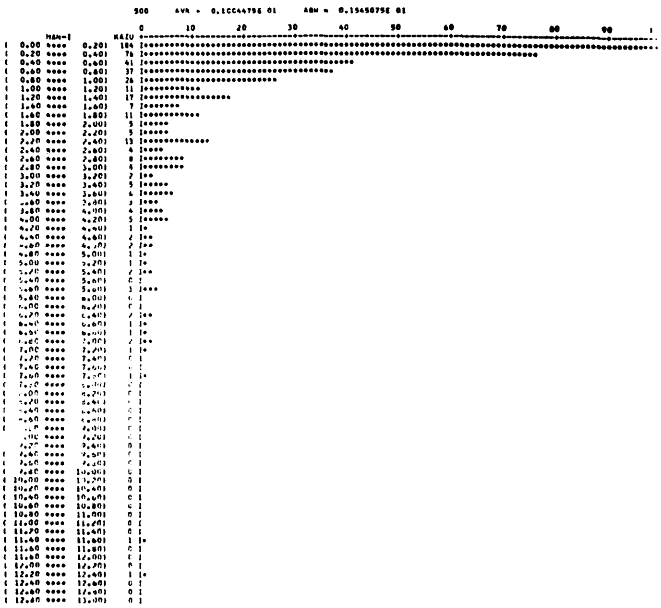
自 由 度 (10,1)



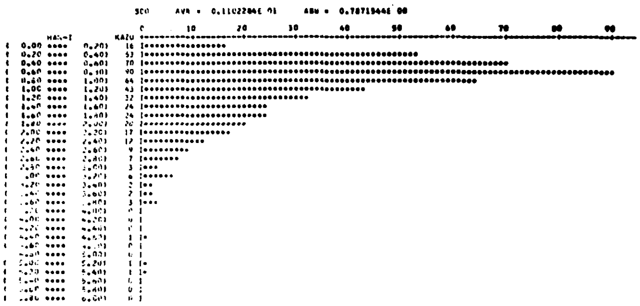
自 由 度 (10,5)



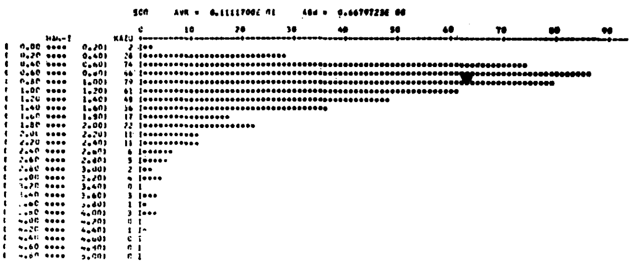
F 分 布 (9-10,10) 自 由 度 (10,10)



自 由 度 (20,1)



自 由 度 (20,5)



自 由 度 (20,10)

図 9 *F* 分 布

自由度はそれぞれ（分母の自由度，分子の自由度）で示してある．((1,2)，(1,5)，(1,10) などのグラフは先の方が印刷してない）自由度による分布の型の変化はテキストに載っているが，シミュレーションによるこの方法では，その安定性が一目でわかるのが特徴である．

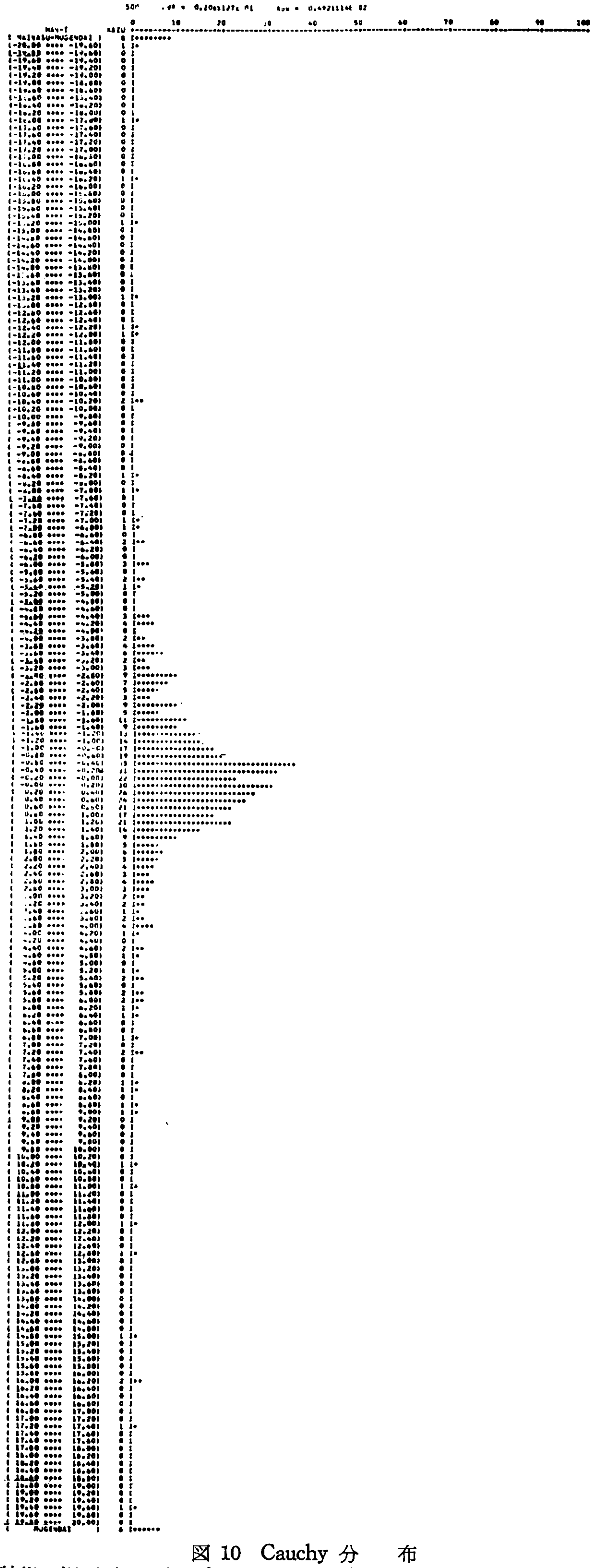


図 10 Cauchy 分　布

この分布の特徴は裾が長いことである．−20.0 より小さいものと ＋20.0 より大きいものはまとめてそれぞれ印刷してある．式の上からは，このような裾の広がりを想像することはなかなか困難であろう．また乱数を変えるたびに分布の（−20〜−5）と（5〜20）の間の 1 個づつであるデータは変わってくる．これからしてもこの分布の不安定なことがよくうかがえる．

るように，分母の自由度が大きくなると裾が短かくなり，それが小さくなると長くなる．また分子の自由度が大きくなるとその分布の mode が大きく現われ，それが小さいと次第になくなり“0”の所から単調に減少してゆくことがわかる．また分母の自由度が大きい程，分母の変数が0の近くの値をとる確率が小さくなるので，分布の安定性が増してくることもわかる．実験計画法などで，F分布が用いられている理由がわかってくる．

10. Cauchy 分布，その他の分布

$Y1_i$ を $N(0,1)$ の正規分布，$Y2_i$ を $N(0,1)$ の正規分布に従うそれぞれ独立な乱数とするとき

$$C_i = Y\,1_i / Y\,2_i$$

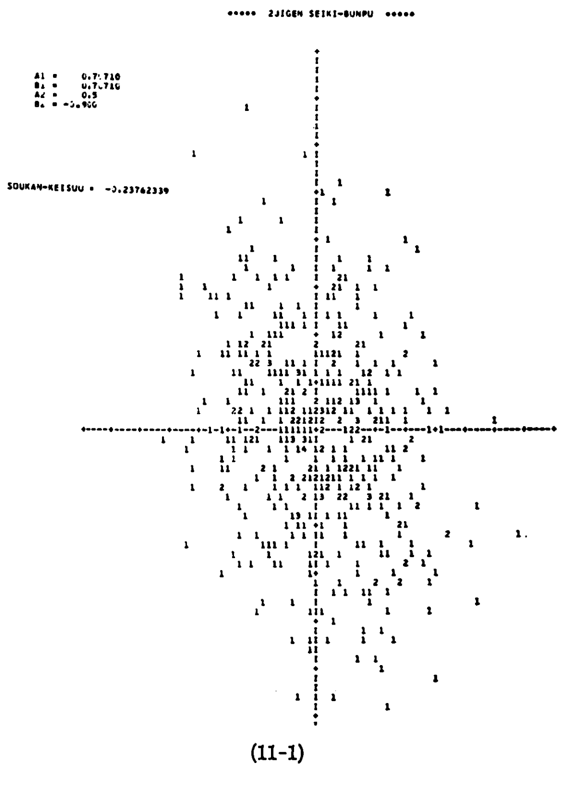
は Cauchy 分布に従うことはテキスト 70 頁の問題にもあることだが，これを実際に解析的に示すのはそれほど簡単な問題ではない．そこでこの問題をシミュレーションで確めさせた．実験結果は 500 個の C_i であるから，平均，分散をも求められているがこれはどこまでも 500 個のデータに関するシミュレーションの結果であり，解析的には平均，分散の存在しないことを説明し，シミュレーションの適応の危険性および限界について考えるように指導した．さらに，解析的にはその分布が簡単には求められないような場合にもこのシミュレーションの方法が有効であることを強調した，しかし，その結果の使用法をあやまれば危険であることをも教えている．

11. 2 次元正規乱数

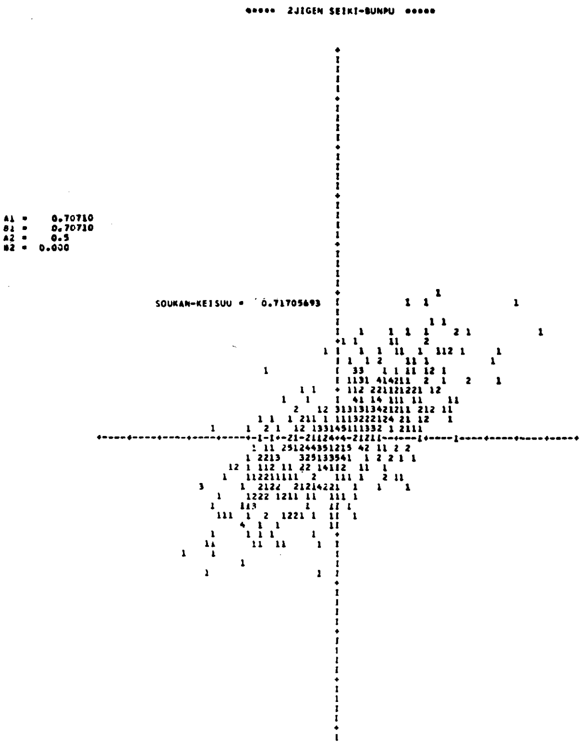
3) で作った正規乱数をもとにして，次のようにして2次元正規乱数を作った．

x_{1i} および x_{2i} は $N(0,1)$ に従う正規乱数とするとき，その線型結合

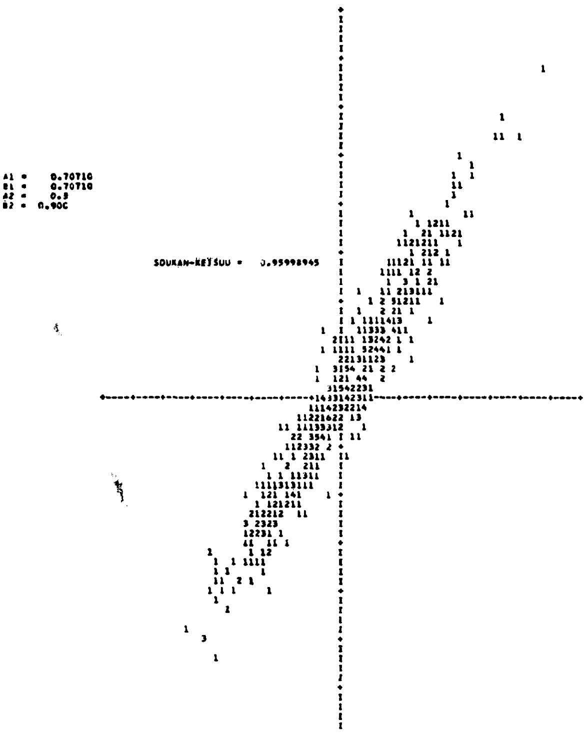
$$\begin{cases} y_{1i} = a_1\,x_{1i} + b_1\,x_{2i} \\ y_{2i} = a_2\,x_{1i} + b_2\,x_{2i} \end{cases} \qquad (11,1)$$



(11-1)



(11-2)



(11-3)

図 11　2 次 元 正 規 乱 数

(11-1) 式によって作った2次元正規乱数の例（標本の個数500個），それぞれの数値はその場所の頻度を示す母相関係数はそれぞれ前から −0.27, 0.71, 0.96, 標本相関係数はそれぞれ前から −0.23, 0.71, 0.95 である.

*** RESULT TABLE (1) ***

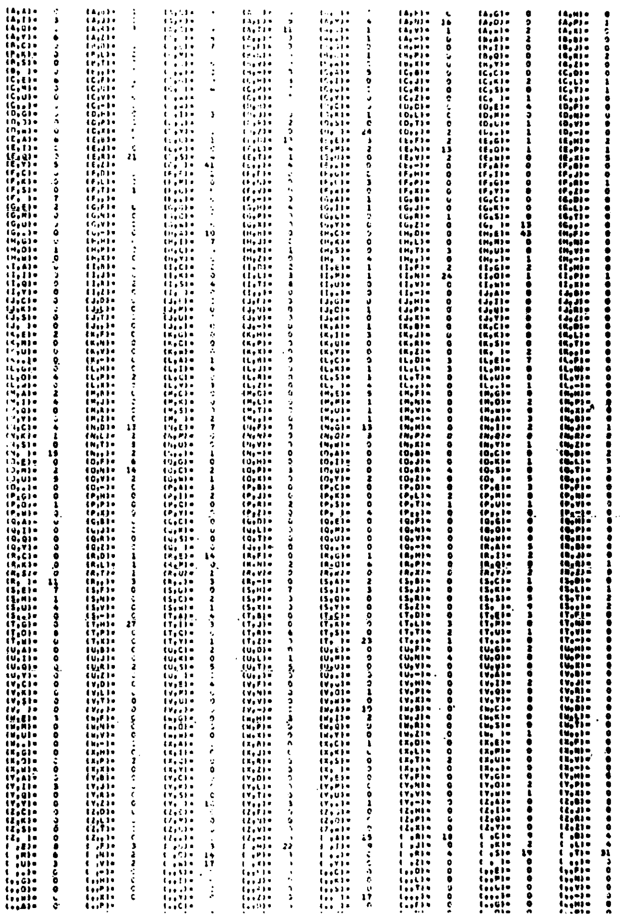
A	...	17
B	...	8
C	...	7
D	...	1C
E	...	23
F	...	3
G	...	2
H	...	7
I	...	15
J	...	2
K	...	4
L	...	1
M	...	13
N	...	19
O	...	14
P	...	9
Q	...	[illegible]
R	...	16
S	...	1C
T	...	19
U	...	11
V	...	1
W	...	1
X	...	[illegible]
Y	...	3
Z	...	[illegible]
1	...	15
2	...	5
3	...	9
4	...	1
5	...	9
6	...	1
7	...	7
8	...	2
9	...	5
C	...	26
(	...	4
)	...	3
•	...	7
/	...	11
+	...	1
-	...	4
*	...	4
,	...	1C
.	...	7
	...	44

(12-1)

*** RESULT TABLE (2) ***

	...	44
C	...	26
E	...	23
N	...	19
T	...	19
A	...	17
R	...	16
I	...	15
1	...	15
O	...	14
M	...	13
/	...	11
U	...	11
S	...	1C
X	...	1C
,	...	[illegible]
D	...	10
3	...	9
5	...	9
P	...	9
B	...	8
•	...	7
C	...	7
7	...	7
.	...	7
H	...	7
2	...	5
9	...	5
*	...	4
K	...	4
(	...	4
-	...	4
Y	...	3
)	...	3
F	...	3
8	...	2
G	...	2
J	...	2
+	...	1
W	...	1
6	...	1
L	...	1
V	...	1
4	...	1
Q	...	C
Z	...	C

(12-2)



(12-3)

図 12　文　字　の　頻　度

ある英語の文章に出現する文字の頻度を計算した結果，12-1 はアルファベット順，12-2 は頻度の順．英語においてはたいていの場合この順となる．図 12-3 は 2 文字づつの現われる頻度である．TH の組が多い．

(y_{1i}, y_{2i}) は2次元正規乱数となる．ここで a_i, b_i $(i=1,2)$ は定数である．このときの母相関係数 ρ は

$$\rho = \frac{a_1 a_2 + b_1 b_2}{\sqrt{(a_1^2 + b_1^2)(a_2^2 + b_2^2)}} \tag{11,2}$$

となる．学生には (a_i, b_i) をいろいろ変えることによって，いろいろな相関係数をもつ2次元正規乱数を作るように要求しているが，学生は式 (11,2) をもとにして (a_i, b_i) を適当に与えて図のような結果を得ている．

この2次元正規乱数のヒストグラムによって学生は直感的に2次元分布の特徴をとらえ，その相関係数との関係をも理解することができた．

さらに母相関係数は (11,2) で与えられるが実際のデータによる相関係

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x}_i)^2 \Sigma(y_i - \bar{y})^2}} \tag{11,3}$$

をいろいろなデータに対して計算させ，この r と ρ との差は r の大きさ（$-1 \leq r \leq 1$）と，サンプルの大きさに関係することを実験で示させ，ρ の推定値の導入を試みた．

12. 文字の頻度の計算

計算機は数の計算をできることは全ての学生が良く理解していることであるが，文字と単語――いわば非数値なもの――を扱うこともでき，言語の統計的性質，マルコフ連鎖などを理解させるのがこの題目の目的である．

ここでは任意の文章を読み込んで，そこに出てくる一つ一つの文字の頻度を計算して，例えば，英語ではEとかTの出現頻度が大きく，ZとかQは少ないことを理解させる．

また2文字の組の出現頻度をも計算し，タイプライターなどの文字の配列が，合理的なものに近いことを教える．さらに3重，4重の文字列の出現頻度をも計算して，一つの言語における文字列のマルコフ性を調べて，しいてはマルコフ過程へと問題を拡大してゆく予定であったが，今回の講義では時間不足のため行なっていない．

学生はそれぞれ異なった文章を入力として解析しているが，同一の言語であれば文字の頻度に一つの統計的傾向あることをみいだしている．

13. 単語の出現頻度の計算

12. で述べた文字の頻度の計算は非数値的な計算といっても非常に簡単なものであったが，単語の頻度計算となると，それほど簡単なプログラムで処理できそうもない問題となる．

「データ処理においては，一般にデータに対して前処理をすればするほど，それ以後のプログラム（仕事）は簡単になる 4), 5)．あるいは結果の正確度が増加する」という事実をもとにして

1) 入力データを FORTRAN で制御しやすい型に穿孔してから．5)
2) 入力データを一般の文章のように穿孔してから

と二通りのプログラムを作るように指導したが，1) のプログラムは文献 2) にのっているため，プログラム作りがあまり上手でない学生は 2) にのっているプログラムをそのままか，あるいわ少し変型して用いていた．

出現した単語に対して ABC 順（辞書順）のソートと出現頻度順のソートを行ない印刷させる．出現頻度順に印刷した単語については，出現頻度の多いものから 20% まで（一般区間）に含まれる単語は，その言語個有の性質からくる単語であり，英語では THE, A, IS などがある．次に 20%～30%（主題区間）に含まれる単語は，入力した文章の性質を現わしていて，いわば入力した文章の主な論点と深いかかわりがある．

そこでこの主題区間に含まれる単語をみれば，その文章は何について述べているかがだいた

*** ALPHABET-SORT ***

NO	TANGO	HINO	SIYORITSU	NO	TANGO	HINO	SIYORITSU	NO	TANGO	HINO	SIYORITSU
1	A	19	0.022	119	FINANCIAL	1	0.001	237	POSTURE	1	0.001
2	ABDICATE	1	0.001	120	FIRST	2	0.002	238	PREPARE	1	0.001
3	ABILITY	1	0.001	121	FOR	7	0.008	239	PRESENTED	1	0.001
4	ABOUT	1	0.001	122	FORM	1	0.001	240	PRESIDENT	1	0.001
5	ACM	24	0.028	123	FOUND	1	0.001	241	PRIMARY	1	0.001
6	ACTIVITIES	2	0.002	124	FRIENDLINESS	1	0.001	242	PROBIDES	1	0.001
7	ACTIVITY	3	0.003	125	FROM	5	0.006	243	PROBLEMS	1	0.001
8	ADDITION	2	0.002	126	FUNCTION	1	0.001	244	PROCEEDS	1	0.001
9	ADVERTISING	2	0.002	127	FUTURE	2	0.002	245	PROFESSION	1	0.001
10	ALL	3	0.003	128	GIVE	1	0.001	246	PROFESSIONAL	2	0.002
11	ALSO	1	0.001	129	GOVERMENT	1	0.001	247	PROGRAM	1	0.001
12	ALTHOUGH	2	0.002	130	GREAT	1	0.001	248	PROGRAMS	1	0.001
13	AMICUS	1	0.001	131	GRIND	2	0.002	249	PROVIDE	2	0.002
14	AN	10	0.012	132	GROUPS	1	0.001	250	PROVIDES	1	0.001
15	AND	25	0.029	133	HAD	1	0.001	251	PROVISION	1	0.001
16	ANOTHER	1	0.001	134	HAPPENS	1	0.001	252	PUBLIC	2	0.002
17	ANSWER	2	0.002	135	HARDWARE	1	0.001	253	PURELY	1	0.001
18	ANTHONY	1	0.001	136	HAS	6	0.007	254	PURITY	1	0.001
19	ANY	5	0.006	137	HAVE	6	0.007	255	PURPOSE	1	0.001
20	APPROACH	1	0.001	138	HEARD	1	0.001	256	PURPOSES	1	0.001
21	APPROPRIATE	3	0.003	139	HERE	2	0.002	257	QUESTION	1	0.001
22	ARE	3	0.003	140	HOPEFULLY	1	0.001	258	QUITE	1	0.001
23	AREAS	1	0.001	141	HOPERFULLY	1	0.001	259	RAISED	1	0.001
24	ARGUMENTS	1	0.001	142	I	5	0.006	260	RALSTON	1	0.001
25	AS	12	0.014	143	IF	5	0.006	261	RANGED	1	0.001
26	ASIDE	1	0.001	144	IMPORTANCE	1	0.001	262	RECEIVED	1	0.001
27	ASPECTS	1	0.001	145	IMPORTANT	5	0.006	263	RECEIVES	1	0.001
28	AT	2	0.002	146	IN	25	0.029	264	RECENT	1	0.001

(13-1)

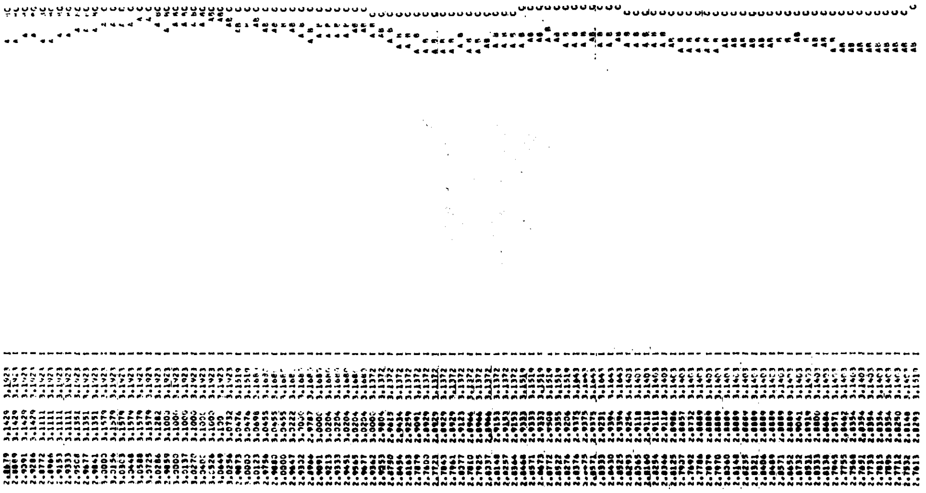
*** FREQUENCY-SORT ***

NO.	HINO	TANGO	SIYORITSU	RUISEKI	NO.	HINO	TANGO	SIYORITSU	RUISEKI	NO.	HINO	TANGO	SIYORITSU	RUISEKI
1	48	THE	0.055	0.055	119	2	JOINT	0.002	0.724	237	1	POSTURE	0.001	0.868
2	33	OF	0.038	0.093	120	2	JOURNALS	0.002	0.726	238	1	PREPARE	0.001	0.869
3	28	TO	0.032	0.125	121	2	USEFUL	0.002	0.728	239	1	PRESENTED	0.001	0.870
4	25	IN	0.029	0.154	122	2	WHAT	0.002	0.731	240	1	PRESIDENT	0.001	0.871
5	25	AND	0.029	0.183	123	2	LARGE	0.002	0.733	241	1	PRIMARY	0.001	0.872
6	24	ACM	0.028	0.211	124	2	LET	0.002	0.735	242	1	PROBIDES	0.001	0.873
7	19	A	0.022	0.232	125	2	LONG	0.002	0.738	243	1	PROBLEMS	0.001	0.875
8	14	BE	0.016	0.249	126	2	WITHIN	0.002	0.740	244	1	PROCEEDS	0.001	0.876
9	13	IT	0.015	0.264	127	1	ABILITY	0.001	0.741	245	1	PROFESSION	0.001	0.877
10	12	INDUSTRY	0.014	0.277	128	1	GIVE	0.001	0.742	246	1	DON	0.001	0.878
11	12	IS	0.014	0.291	129	1	GOVERMENT	0.001	0.743	247	1	PROGRAM	0.001	0.879
12	12	AS	0.014	0.305	130	1	GREAT	0.001	0.745	248	1	PROGRAMS	0.001	0.880
13	11	SUCH	0.013	0.318	131	1	BUYERS	0.001	0.746	249	1	BEHOLDEN	0.001	0.881
14	10	AN	0.012	0.329	132	1	GROUPS	0.001	0.747	250	1	PROVIDES	0.001	0.883
15	10	NOT	0.012	0.341	133	1	HAD	0.001	0.748	251	1	PROVISION	0.001	0.884
16	9	COMPUTER	0.010	0.351	134	1	HAPPENS	0.001	0.749	252	1	DOUBTS	0.001	0.885
17	9	S	0.009	0.360	135	1	HARDWARE	0.001	0.750	253	1	PURELY	0.001	0.886
18	8	THAT	0.009	0.369	136	1	AREAS	0.001	0.751	254	1	PURITY	0.001	0.887
19	7	FOR	0.008	0.377	137	1	CAESAR	0.001	0.753	255	1	PURPOSE	0.001	0.888
20	7	THIS	0.008	0.386	138	1	HEARD	0.001	0.754	256	1	PURPOSES	0.001	0.890
21	7	SHOULD	0.008	0.394	139	1	ARGUMENTS	0.001	0.755	257	1	QUESTION	0.001	0.891
22	7	WHICH	0.008	0.402	140	1	HOPEFULLY	0.001	0.756	258	1	QUITE	0.001	0.892
23	6	CORPORATE	0.007	0.409	141	1	HOPERFULLY	0.001	0.757	259	1	RAISED	0.001	0.893
24	6	HAS	0.007	0.415	142	1	ABOUT	0.001	0.758	260	1	RALSTON	0.001	0.894
25	6	HAVE	0.007	0.422	143	1	CASES	0.001	0.759	261	1	RANGED	0.001	0.895
26	6	ONE	0.007	0.429	144	1	IMPORTANCE	0.001	0.761	262	1	RECEIVED	0.001	0.896
27	5	OR	0.006	0.435	145	1	CHAPTER	0.001	0.762	263	1	RECEIVES	0.001	0.898
28	5	OUR	0.006	0.441	146	1	CINCERNS	0.001	0.763	264	1	RECENT	0.001	0.899
29	5	ANY	0.006	0.446	147	1	INANY	0.001	0.764	265	1	DURING	0.001	0.900

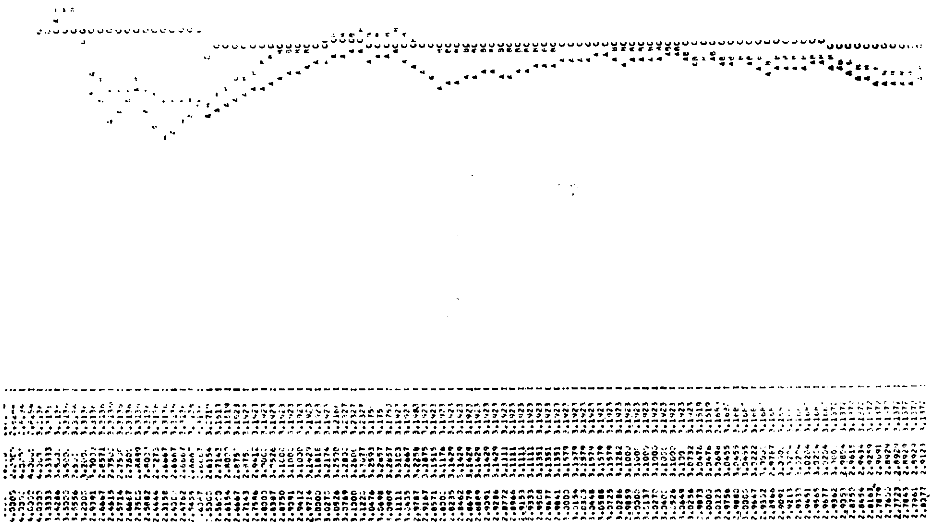
(13-2)

図 13 単語の出現頻度

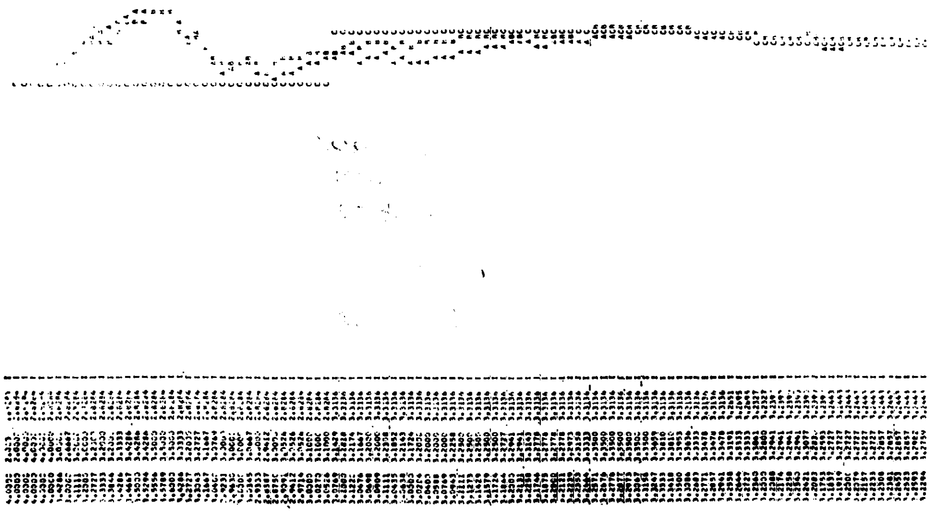
ある英語の文章に出現する単語の頻度を辞書順に並べたものの一部 (13-1) と頻度順に並べたものの一部 (13-2). 累積の 0.20 程度までは, 英語ではどんな文章でも THE, OF, TO, IN, AND などが現われ, 次の 0.20～0.40 までには, その文章固有に用いられている単語が現われる.



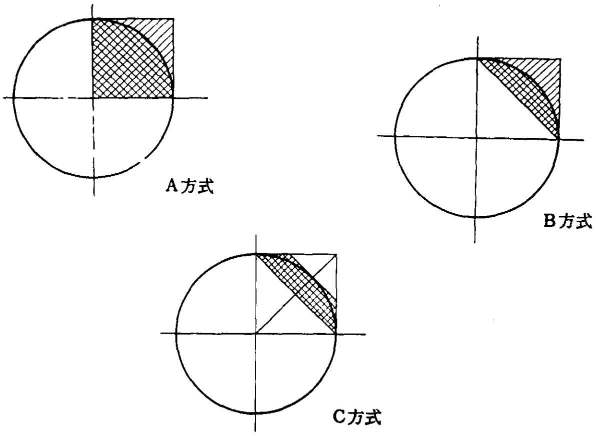
(14-1)



(14-2)



(14-3)



(14-4)

図 14 π の 計 算

A, B, C の方式でそれぞれを実験して，500 回毎に (14-1) 式により π の近似値を求め印刷したもの，A 方式の結果は A と印刷してあり，他も同様である．C 方式が最良ではないことに注意する必要がある．

いわかるが，学生への課題としては，主題区間に含まれている単語を二個以上含んでいる文章だけを出力するプログラムを作らせた．

14. π の 計 算

モンテカルロ法を用いた簡単な例として π の値を，一様乱数を用いて計算するという課題である．出力にあまりこらなければ，このプログラムは非常に簡単なため，多くの学生が興味をもって行なった．

モンテカルロ法で π の値を求めるのは図 14 に示すように正方形 $\alpha\beta\gamma\delta$ と扇形 $\alpha\beta\gamma$ の両軸の比を求めればよい．比を求めるには，X_n, Y_n をそれぞれ区間 [0, 1] の間の一様乱数とする．この乱数を N 個発生させ，

$$X_i^2 + Y_i^2 \leq 1 \tag{14, 1}$$

なる式を満したものを M 個とすれば，π は近似的に

$$\frac{M}{N} * 4 \tag{14, 2}$$

で与えられる．

この課題はシミュレーションを理解するのに役立つであろう．

ABCの方法を比較するとCの分散が一番小さいはづであるが[2]，しかし乱数の性質により，学生の実験結果では，Aの分散の方が小さいということになっている．このような所にもシミュレーションの限界をみいだせるように指導している．(もちろんこれはシミュレーションには目的に合った乱数を用いなければならないといってしまえばよいであろうが，目的とは何か，乱数とは何か，学生自身がいろいろと考えていたようである)．

15. 考 案

学生に宿題を与えた場合，その宿題が手のかかる面到なものであればあるほど，学生は他人のプログラムを借りるとか，協同で行なうということになろう．ここで与えた 14 の課題も，実際にプログラムを作り，計算機に通して，結果を出し，その考案を書くということになると，

決して簡単なものではない．前にも述べたように，他人のプログラムを借りるなどが行なわれるのは当然のことである（プログラムの完成には平均5回のデバックが必要である）．そこで学生に対しては，いわば積極的に協同で行なうように指導してある．（しかし，データだけは同じのを用いないようにと）．

統計学とはデータが変っても，そこの統計的性質は変らないことを認識することである．例えば π の計算においては，a，b，p，x_0 をいろいろ変えれば，いろいろと変った π の近似値を得ることができるが，2.8～3.5 の間にあることをよく理解したようである．これは一様，正規，χ^2，F，t，Cauchy の各分布についても，同じであり（a，b，p，x_0）を変えれば少しづつ異なってはいるが，統計的には同じような分布を得ることができる．また学生は課題が完成したとき，それを友人などとその結果をみせ合っているが，これも統計的ばらつきを理解させる上で有意義なことであろう．

この教育方式を始めた頃は，とにかく統計学を普及させようという立場であったことは事実である．もちろんこの立場は今も変わるべきでもない．しかし現在ではシミュレーションとか，多変量問題など統計学を用いるには充分な注意をせねばならない場合に対しての危険性を示すような簡単な例題が多く，いわば統計学の誤解を解く，あるいは統計学の危険性を示すことによって統計学を理解させようとしているのである．

この方法の開発にあたって，多くの方々，とくに統計数理研究所 第3研究部 第2研究室長 牧野都治博士，第5研究部 第1研究室長 清水良一博士よりいろいろと助言をえました．ここに謝意致します．

文　献

1. 本間鶴千代： 統計数学入門，森北出版 1970.
2. 竹内 啓： モンテカルロシミュレーションの統計的方法，経営科学，1973-3.
3. 山内二郎，森口繁一，一松 信： 電子計算機のための数値計算法 I，培風館，1965.
4. 樋口伊佐夫，二宮理憙： 電子計算機一基礎と応用，新光閣書房，1967.
5. 二宮理憙： 初級 FORTRAN，税務経理協会，1970.