新しいヒストグラム

川崎医科大学　数学教室
仮　谷　太　一
（昭和54年9月27日受理）

A New Type of Histogram

Taichi KARIYA
Department of Mathematics, Kawasaki Medical School
Kurashiki 701-01, *Japan*
(*Received on Sept.* 27, 1979)

連続型変量母集団からの無作為標本データにもとづく統計的な解析は，ヒストグラムの作成から始めるのが普通であるが，現在広く用いられているヒストグラムは，階級の個数・階級の区切り方により，かなり大きく変容することは周知の事実である。階級の個数・階級の区切り方には多分に経験的知識が必要であり，特にとび離れた値（outliers）がある場合などには，人により，同じデータからひどくイメージの違ったヒストグラムの得られることがある。このようなことは，初学者に大きな不安を与え，統計的な方法に対する不信の原因ともなりかねないので，ここに標本データから一意的に決まる新しいヒストグラムを提案する。なお，この新しいヒストグラムは，その構成方法が簡明で，一意的であるから，電子計算機による自動描画が可能であり，正規乱数データにもとづく等積ヒストグラムが例示される。

Abstract

The first step to statistical analysis based on random sample data taken from some continuous population is usually begun by drawing the graphical representation or histogram. In this representation the number of classes and class boundaries are essential. However, the number and boundary values ot classes are decided by personal empirical knowledge in many cases. So from the same data, we are oblidged to get several histograms which differ considerably in an immediate visual impression, which causes beginners another problem. In this article we propose a new type of histogram. This one is as simple as the histogram commonly used in the drawing rule but determined uniquely by data. We can easily draw the new histogram and also automatically by computer. New histograms based on normally distributed data will be illustrated.

§1. ヒストグラムの問題点

連続型変量データを要約し，データ全体としての分布・構造を視覚的に把握することは，データ解析の第一歩であるが，そのためにはまず，データの分布している変量の範囲をいくつかの階級にわけて度数分布表を作らねばならない。ところが階級の個数・階級の境界値の定め方

・最初および最後の階級のとり方などについては，多分に経験的な知識を必要とすることが少なくない。従って同じデータから，要約する人の知識・経験の深浅によって，かなりイメージの違ったヒストグラムが得られることになる。殊に生物学や医学データによく現われる outliers（とび離れた値）がある場合には，この傾向は一層顕著である。もっともこういった不安定さは，偶然変動を伴う事象の理解に，必らずしも有害であるとは言えないという意見もあるかも知れないが，初学者にとっては，大きな不安のみなもととなり，思わぬ誤解や統計的方法に対する不信の原因ともなりかねない。

ところで階級の個数 m については，データの個数を n とするとき，従来 Sturges の公式

$$m = 1 + \log_2 n \tag{1}$$

による値が，一応の目安とされてきたが，n が 500 を越えない範囲ならば，この値を用いることで特に問題はないであろう。データ数が 500 を越えるときは，もう少し大きな値のほうが適当のように考えられる。

慣用されてきたヒストグラムで，問題となるのは，前に述べたように，最初の階級をどこから始めるべきかということであり，特に outliers があるときの最初および最後の階級の定め方である。これらについて簡明な一般的ルールを決めることは恐らく困難であろう。

なおこれまで度数分布表は，統計的推測の立場から，ヒストグラムを描くための前処理であったばかりでなく，標本データを要約する特性値としての平均値・分散などを計算する基礎表の役割をもっていた。しかし現在では，卓上電子計算機の普及により，平均値や分散は，直接生データから計算されるようになったので，そのための基礎表としての役割は不必要になったと言っても過言ではない。

ヒストグラムを描いて標本データ全体の構造を掴み，また相互に比較する目的のためだけならば，ヒストグラムの構成手順が簡明で，経験的知識を必要としない，しかもデータから一意的に決まるよい描画法を積極的に採用すべきであろう。

§2. 等積ヒストグラムの作成手順

データ全体の構造を視覚化するための基礎的描画法においては，グラフ化の手順が直観的でわかり易いことが極めて大切である。与えられたデータの構造・特性を，ある原理に照らして最もよく表現するようなグラフを探究することも，時と場合により必要となるが，その場合にはグラフの構成手順が原理的であるため，直観的理解に苦しむことになろう。ここでは慣用されてきたヒストグラムと同じように，グラフの構成手順が，直観的でわかり易い等積ヒストグラムを提案する。

ここに提案する等積ヒストグラムは，従来のヒストグラムが横軸を多くの場合等分割していたのに対し，縦軸を等分割するものであり，その構成手順が直観的でわかり易く，しかも経験的知識を必要とせず一意的に決まるという長所をもっている。

次に，大きさ n のデータにもとづいて，等積ヒストグラムを作成する手順を述べよう。

1）n個のデータを大きさの順に（小から大へ）並べ替え，等しいものはまとめて，$x_1<x_2<\cdots<x_{n'}$ とする。一般に $n'\leqq n$ である。また x_i に等しいデータの個数を f_i とする。$\sum_{i=1}^{n'} f_i=n$.

2）x における累積相対度数 $S_n(x)$ を

$$S_n(x)=(x \text{ を越えないデータの個数})/n \tag{2}$$

で定義するとき

$$S_n(x_i)=(f_1+f_2+\cdots f_i)/n \tag{3}$$

となる。点 $(x_i,\ \{S_n(x_{i-1})+S_n(x_i)\}/2)$ $i=1,\ 2,\ \cdots,\ n'$ を順次線分で結ぶとき得られるグラフを，累積相対度数折れ線グラフと呼ぶことにする。この折れ線グラフは，従来のいわゆる累積相対度数折れ線グラフを多少修正したものになっているが，これは左右の対称性を保つための工夫であり，階段関数 $y=S_n(x)$ の蹴あげの中点を結びつけたものになっている。

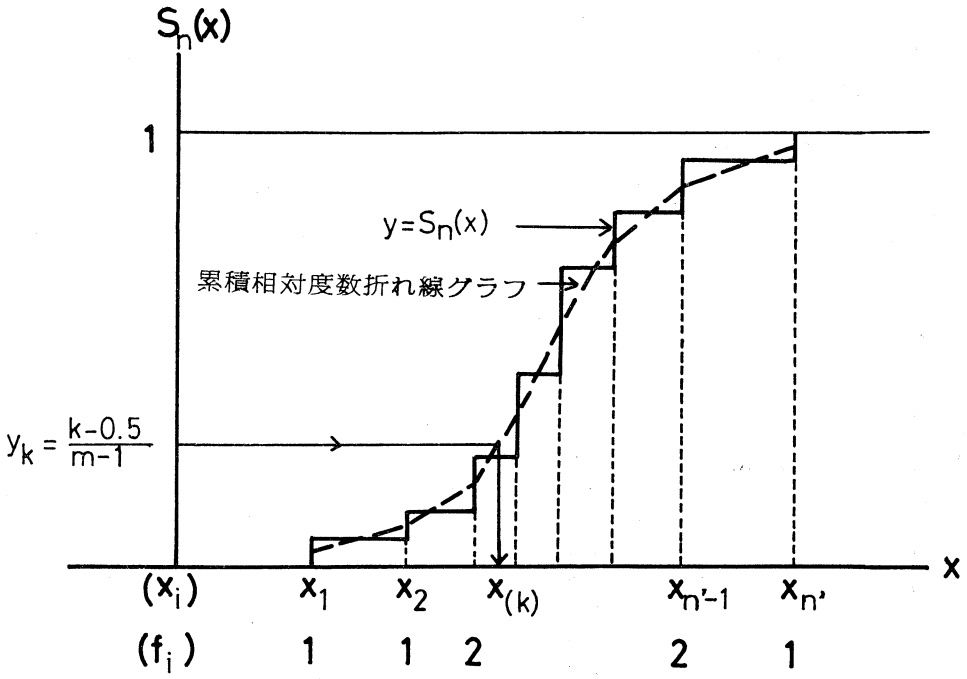


図1　累積相対度数折れ線グラフと $x_{(k)}$

3）階級の個数 m を次式により定める。

$$m=[e(l_n n-1)+0.5] \tag{4}$$

ここに，e は自然対数の底であり，$e=2.718\cdots$，また［ ］はガウスの記号である。n のいくつかの値に対する m の値を示すと表1のようになる。

表　1

n	m
25	6
52	8
110	10
1000	16
10000	22

4）累積相対度数が $y_k=(k-0.5)/(m-1)$ $(k=1,\ 2,\ \cdots,\ m-1)$ に対応する x の値 $x_{(k)}$ を，累積相対度数折

れ線グラフを用いて算定する<図1>。すなわち，次のように計算する。

$\bar{S}_n(x_{j-1}) < y_k \leqq \bar{S}_n(x_j)$ のとき

$$x_{(k)} = x_j - (x_j - x_{j-1})(\bar{S}_n(x_j) - y_k) / \{\bar{S}_n(x_j) - \bar{S}_n(x_{j-1})\} \tag{4}$$

$k=1, 2, \cdots, m-1$

ただし $\bar{S}_n(x_j) = \{S_n(x_{j-1}) + S_n(x_j)\}/2$ $j=2, 3, \cdots, m-1$, $\bar{S}_n(x_1) = S_n(x_1)/2$

なお，両端の $x_{(0)}$, $x_{(m)}$ については，いろいろの工夫があるが，次に示すのはその1例である。c, c' は適当な正の定数とする。

$$x_{(0)} = x_{(1)} - c(x_{(2)} - x_{(1)}) \tag{5}$$

$$x_{(m)} = x_{(m-1)} + c'(x_{(m-1)} - x_{(m-2)}) \tag{6}$$

5）区間 $[x_{(k-1)}, x_{(k)}]$ $k=1, 2, \cdots, m$ の上の柱の高さ h_k を次のように計算する。

$$h_k = \left(\frac{1}{m-1}\right) \Big/ (x_{(k)} - x_{(k-1)}) \quad k=2, 3, \cdots, m-1 \tag{7}$$

なお，h_1, h_m は(5)(6)に示した $x_{(0)}$, $x_{(m)}$ を用いる場合

$$h_1 = \{0.5/(m-1) - n_L/n\} / (x_{(1)} - x_{(0)}) \tag{8}$$

$$h_m = \{0.5/(m-1) - n_U/n\} / (x_{(m)} - x_{(m-1)}) \tag{9}$$

ただし n_L は $x_{(0)}$ 以下のデータ数，n_U は $x_{(m)}$ より大きいデータ数

6）$x_{(k)}(k=0, 1, \cdots, m)$, $y_k(k=1, 2, \cdots, m)$ を用いて等積ヒストグラムを描く<図2>。

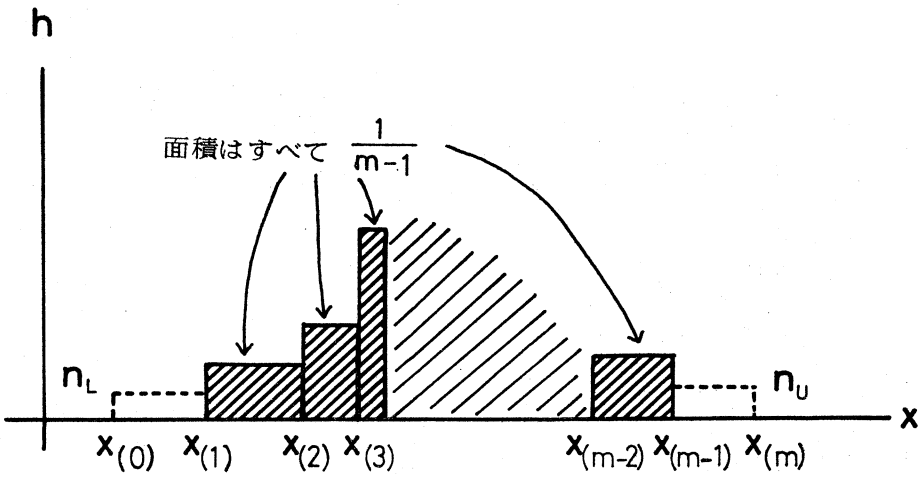


図2 等積ヒストグラム

7）ヒストグラムに含まれないデータ数 n_L, n_U をヒストグラムの両側にそれぞれ記入する。

注意：(1)縦軸の目盛を適当に決め，ヒストグラムが全体としてやや横に広い長方形内におさまるようにする。

(2) 分布のすそに，とび離れて位置する outliers は，7）に示した個数に含まれる。

§3. 例　　題

簡単な例について，等積ヒストグラムの描き方を示すことにしよう。

次のデータが与えられたとき，上の手順に従って等積ヒストグラムを描け。

$n=32$

31, 32, 48, 61, 68, 47, 49, 54, 53, 44, 59, 53, 39
68, 56, 55, 48, 39, 42, 51, 47, 35, 44, 46, 50, 50
63, 40, 53, 54, 50, 56

1）データを大きさの順に（小から大へ）並べ替え，等しいものはまとめて整理すると $n'=21$，各 x_i に対する度数 f_i（表2の第3列）が得られる。

2）この度数 f_i から累積度数を求め，$S_n(x_i)$，$\bar{S}_n(x_i)=\{S_n(x_{i-1})+S_n(x_i)\}/2$ を計算する＜表2＞

3）$m=[2.718\cdots\times(ln\ 32-1)+0.5]=7$

4）図3の縦軸にそうて $y_k=(k-0.5)/(m-1)$（$k=1, 2, \cdots, m-1$）を目盛り，矢印の方向に進めば $x_{(1)}, x_{(2)}, \cdots, x_{(6)}$ が得られる＜表2の$x_{(k)}$の列＞。例えば $x_{(3)}$ は式(4)により次の

表－2　計　算　表　$n=32$, $n'=21$, $m=7$

i	x_i $(i=1\sim n')$	度　数 f_i	累積度数	累積相対度数 $S_n(x_i)$	$\frac{S_n(x_{i-1})+S_n(x_i)}{2}$	$x_{(k)}$	h_k
						(26.8889)	
1	31	1	1	0.0313	0.0156		
2	32	1	2	0.0625	0.0469		0.0097
3	35	1	3	0.0938	0.0781		
4	39	2	5	0.1563	0.1250	35.4444	
5	40	1	6	0.1875	0.1719		0.0194
6	42	1	7	0.2188	0.2031		
7	44	2	9	0.2813	0.2500	44.0000	
8	46	1	10	0.3125	0.2969		0.0394
9	47	2	12	0.3750	0.3438		
10	48	2	14	0.4375	0.4063	48.2222	
11	49	1	15	0.4688	0.4531		
12	50	3	18	0.5625	0.5156		0.0566
13	51	1	19	0.5938	0.5781		
14	53	3	22	0.6875	0.6406	51.1667	0.0476
15	54	2	24	0.7500	0.7188		
16	55	1	25	0.7813	0.7656	54.6667	
17	56	2	27	0.8438	0.8125		
18	59	1	28	0.8750	0.8594		0.0208
19	61	1	29	0.9063	0.8906		
20	63	1	30	0.9375	0.9219	62.6667	
21	68	2	32	1.0000	0.9688		0.0104
						(70.6667)	

ようになる。

$\bar{S}_n(x_{10}) < y_3 \leqq \bar{S}_n(x_{11})$　であるから

$x_{(3)} = x_{11} - (x_{11} - x_{10})(\bar{S}_n(x_{11}) - y_3)/(\bar{S}_n(x_{11}) - \bar{S}_n(x_{10}))$

$= 49 - 1\,(0.4531 - 2.5/6)/(0.4531 - 0.4063)$

$= 49 - 0.7778 = 48.2222$

なお式(5), (6)より $c = c' = 1$ とすれば

$x_{(0)} = 2x_{(1)} - x_{(2)} = 2 \times 35.4444 - 44.0000 = 26.8888$

$x_{(7)} = 2x_{(6)} - x_{(5)} = 2 \times 62.6667 - 54.6667 = 70.6667$

5）両端を除き，各区間の柱の面積 $1/(m-1)$ に等しくなるように，高さ h_2, h_3, …, h_{m-1} を計算する。例えば h_4 は式(7)より次のようになる。

$$h_4 = \left(\frac{1}{m-1}\right) \Big/ (x_{(4)} - x_{(3)}) = 0.16667/(51.1666 - 48.2222) = 0.0566$$

$x_{(0)}$ 以下，$x_{(7)}$ より大きいデータはないから $n_L = 0$, $n_U = 0$。従って式(8)(9)より

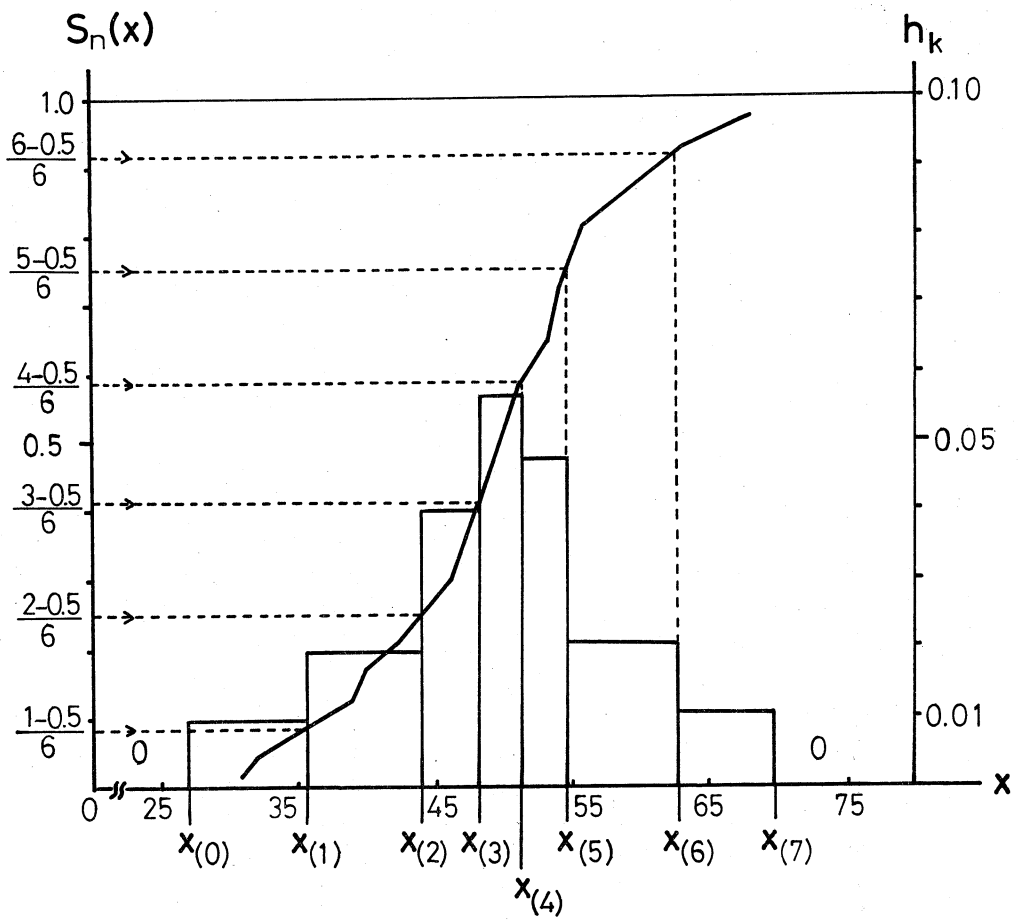


図 3　累積相対度数折れ線グラフから等積ヒストグラムへ

$h_1=0.08333/(35.4444-26.8889)=0.0097$

$h_7=0.0833/(70.6667-62.6667)=0.0104$

6）$x_{(k)}$，h_k を用いて，やや横に広いグラフになるよう縦座標の目盛を定めて，等積ヒストグラムを描く＜図3＞。

7）ヒストグラムの両側に n_L，n_U を記入すれば等積ヒストグラムは完成する。

§4. 正規乱数データにもとづく等積ヒストグラム

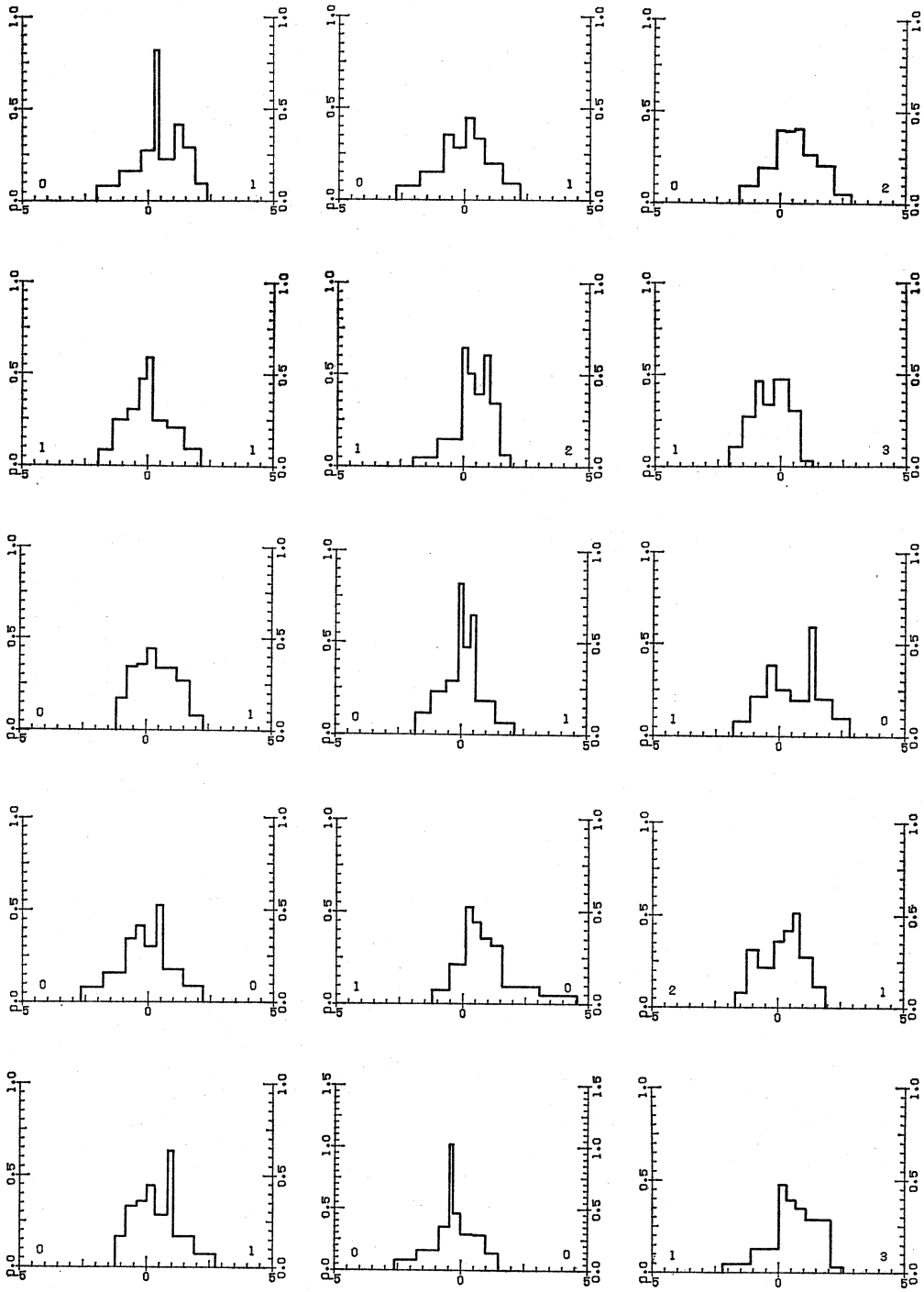
次には，大きさ $n=50$ および $n=100$ の正規乱数データにもとづく等積ヒストグラムを，図4，図5に示す。正規乱数としては，Lehmer's random number generator

$X_{n+1}\equiv 630360016 * X_n \pmod{2^{31}-1}$　$X_0=715827883$ による一様乱数を Box-Müller の方法で変換したものを，それぞれ50個ずつ，100個ずつに区切って使用し，Computer に自動描画させたものである。かなりよく正規分布の相を表現していると思われる。

謝　　辞

Computer により，正規乱数を発生させ，等積ヒストグラムを実際に描かせる仕事は，川崎学園コンピュータ・センター谷口和夫氏にお願いした。ここに記して感謝の意を表する次第である。

図4　大きさ $n=50$ の正規乱数データにもとづく等積ヒストグラム



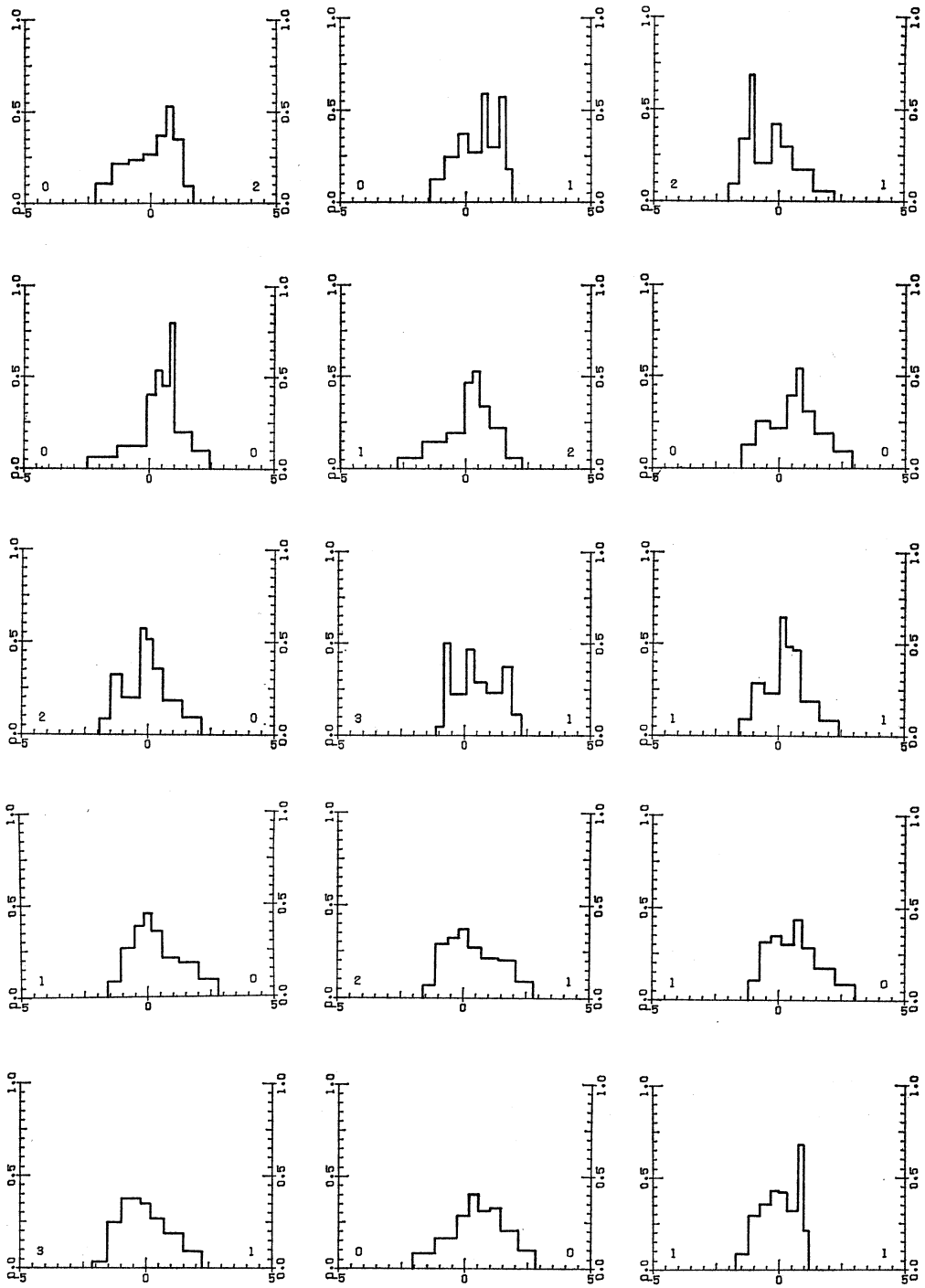
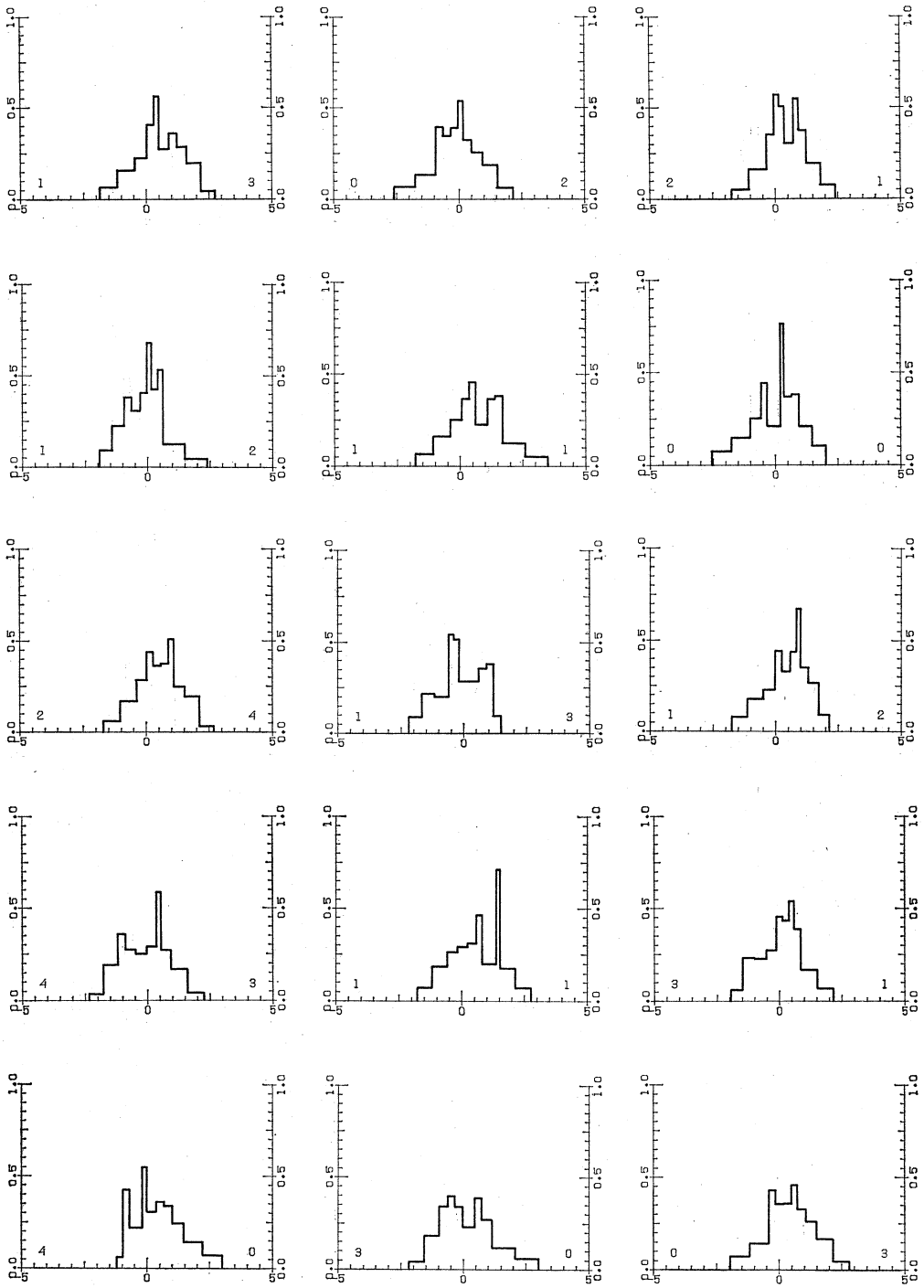
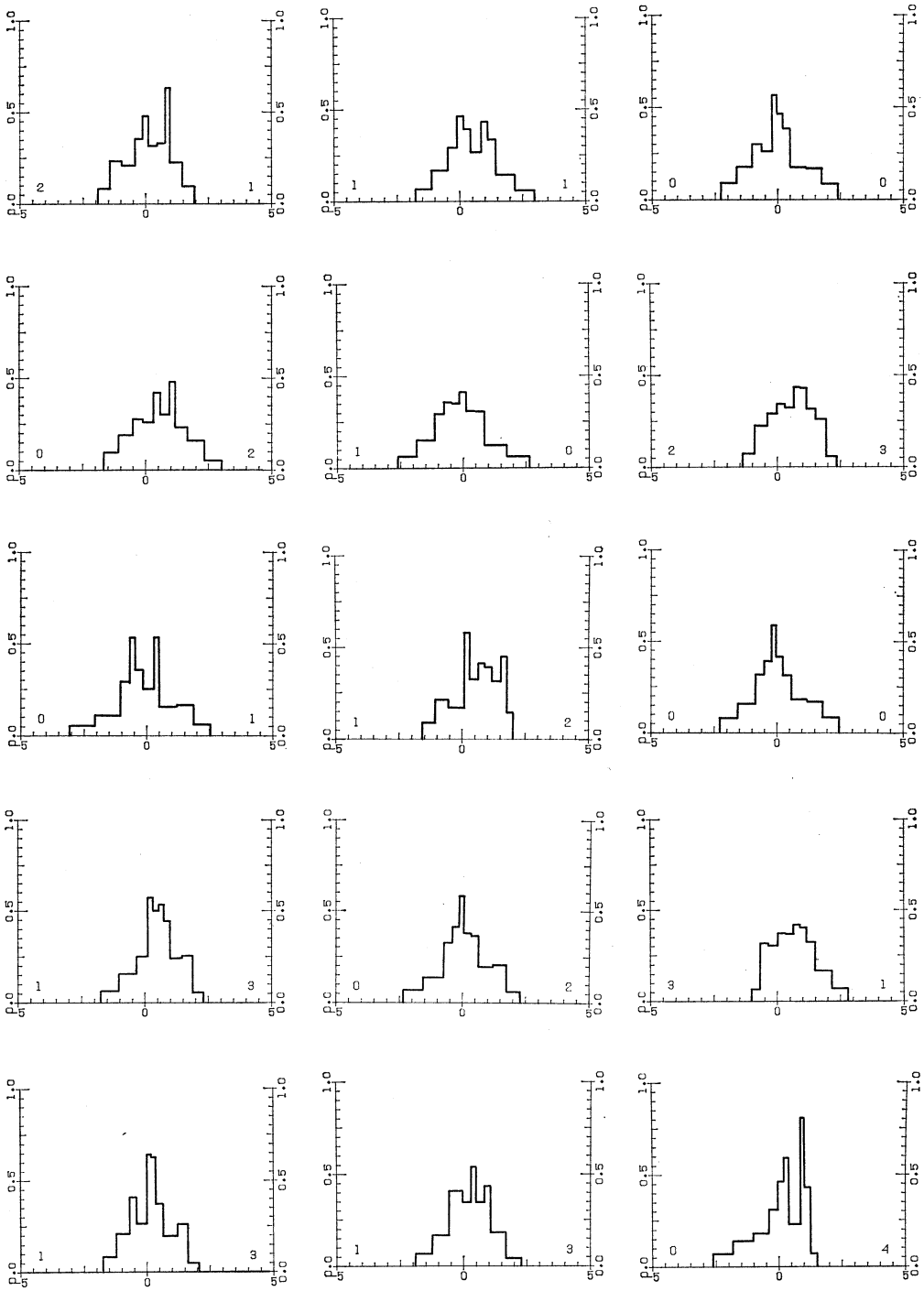

0 2
0 1
2 1
0 0
1 2
0 0
2 0
3 1
1 1
1 0
2 1
1 0
3 1
0 0
1 1

図5 大きさ $n=100$ の正規乱数データにもとづく等積ヒストグラム




2 1
1 1
0 0
0 2
1 0
2 3
0 1
1 2
0 0
1 3
0 2
3 1
1 3
1 3
0 4