1981年8月　天　気　Vol. 28, No. 8

551. 501. 45

ヒストグラムの誤差と描き方

第2報　系統的誤差・描き方*

菊地原　英和**

目　次

1. 概　要

第2節以降では，まず各種系統的誤差の特徴，性質等を概説した後，一山型分布の代表としてえらんだ正規分布等について，頂点低下率，面積誤差率等により，分布曲線の高さの低下，度数折線の面積誤差の大きさを評価し，第1報で提案した階級数の式（第3節参照）がこれら誤差から見ても問題が無いことを示す．次に，気象データの2，3のヒストグラムの実例について，第1報のノモグラムを利用した標本誤差の評価及び各種系統的誤差の検討を行ない，併せて階級数の式をチェックし，ヒストグラムの描き方を考察する．続いて一山型以外の主な分布型について，ヒストグラム作成上の条件を考察し，最後に，以上を総合してヒストグラムの描き方を提案する．

なお，引用等の便宜上，式，図，表の番号は第1報からの一連番号を採用した（(28)式，第20図，第6表までが第1報）．

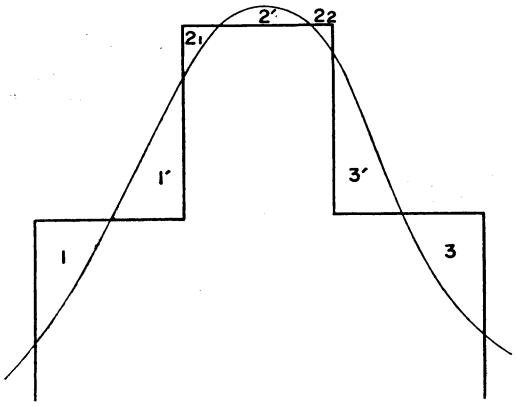
2. 分布曲線とヒストグラム，度数折線の関係

階級区分を決めたとき，データから描かれる実測度数（k で表わす）のヒストグラムが，母集団の分布曲線 $f(x)$ に対して持つ誤差は，期待度数（k_0 で表わす）のヒストグラムと $f(x)$ との差異を示す系統的誤差と，期待度数 k_0 のまわりの実測度数 k の標本変動による偶然誤差，つまり階級別度数の標本誤差の和で表わされ，両者を独立なものとして別々に扱うことができる．本節以降しばらくは系統的誤差を扱うので，標本誤差は無いと仮定して議論する．従って次の式を仮定する．

$$k=k_0=Np_0 \qquad (29)$$

* On the errors contained in a histogram and reasonable expression for histograms. Part 2 Systematic errors and a practical method of drawing.

** Hidekazu Kikuchihara, 気象大学校.



第21図　分布曲線とヒストグラムの関係.
〔面積〕 $1=1'$, $2_1+2_2=2'$, $3=3'$

$$p_0=\int_x^{x+d} f(x)dx=k/N=k'd \qquad (30)$$

ただし，N は標本の大きさ，d は階級幅，x は階級の左端であり，k' は単位幅当りの相対度数で，次の式で与えられる.

$$k'=k/Nd \qquad (31)$$

k の代りにこの規準化度数 k' を縦軸にとってヒストグラムを描けば，各階級について，分布曲線の面積とヒストグラムの面積は相等しい ((30)式). この関係を第21図に示す. もちろん，このヒストグラムの全面積は1に等しい. ヒストグラムの高さ k' は，$f(x)$を階級内で積分平均したものであって，この階級内平均化によって，① 階級幅 d よりスケールの小さい $f(x)$の構造はヒストグラムでは消滅する. ② ヒストグラムは分布曲線に比べて，峰は低く谷は浅くなる (度数折線も同じ).

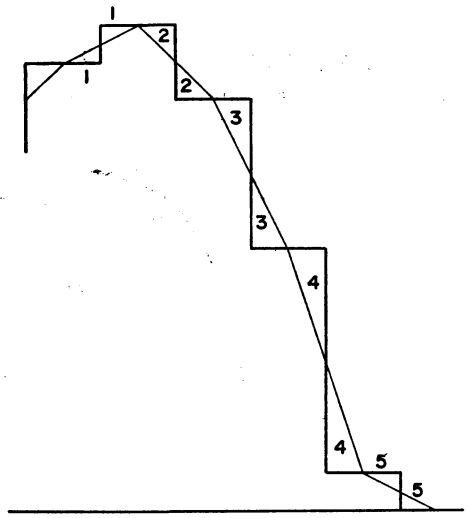
第22図には，ヒストグラムとその柱の頂上中央の点を順次結んだ度数折線との面積的関係を示す. 度数折線はヒストグラムの柱の肩を削って，度数がより小さい隣の階級の凹所を埋めた形になっていることがわかる. ヒストグラムと $f(x)$ が階級内で等面積であるから，これは次のことを意味する.

③ 度数折線と $f(x)$ は，階級内の面積が一致しない.

両者の面積の差を「面積誤差」と呼ぶことにし，ΔS で表わせば，第22図から次の式で与えられる.

$$\Delta S=\frac{d}{8}\{(k_L'-k')+(k'_R-k')\}=\frac{d}{4}\left(\frac{k_L'+k_R'}{2}-k'\right) \qquad (32)$$

ここで k_L', k_R' は，左隣，右隣の階級の規準化度数である. なお，度数折線全体の面積は，$f(x)$ と同じく1である.



第22図　ヒストグラムと度数折線の関係.
(同じ番号の三角形は合同である.)

系統的誤差には以上3種のほかに④モード，谷の位置の見掛け上の移動がある. これらには階級幅はもちろん，階級境界位置も影響する. 各系統的誤差の性質，大きさ等は次節以降で順次取り上げるが，度数折線とヒストグラムは，面積誤差の有無の点でちがうほかは全く同じである.

3. 頂点低下率

後の引用の都合上，第1報で提案した階級数の式(28)をここで再録しておく. この式は，中心部にモードを持ち，両裾で0に漸近する一山型の母集団分布を対象としている.

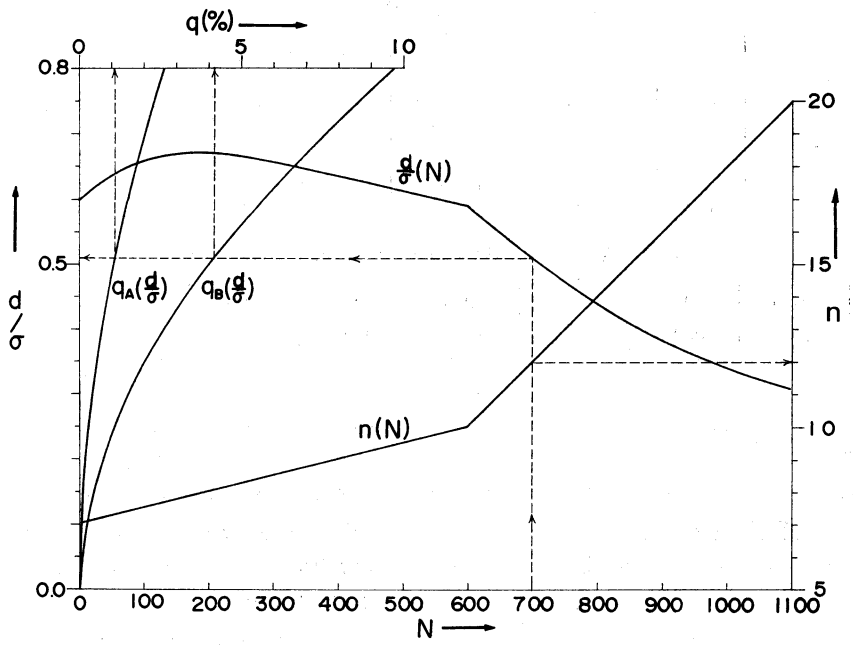
$$\left.\begin{aligned} n&=0.0055N+6.73 \quad (50\leqq N\leqq 600)\\ n&=0.020N-2.00 \quad (600\leqq N\leqq 1100)\\ u&=20 \qquad\qquad\quad (1100\leqq N)\end{aligned}\right\} \qquad (28)$$

この式の系統的誤差から見た妥当性はまだ検討してなかったので，頂点低下についてこの節で検討する. (28) 式の対象が一山型分布であるから谷が浅くなる問題は扱わないが，階級内平均化の影響は頂点低下と同等である.

標本誤差の場合と同様に，頂点低下についても，低下量ではなく相対誤差，つまり次の頂点低下率 q で評価するほうが適切である.

$$q=(f(\tilde{x})-k_M')/f(\tilde{x}) \qquad (33)$$

ここで k_M' はモード階級の規準化度数，つまりヒスト



第23図　正規分布の頂点低下率 q および実用式の N, n, d/σ の関係.

グラムの高さで，$\tilde{x}$ はモードである.

一山型分布の代表として，正規分布について，階級数に（28）式を採用したときの頂点低下率を求める図を作成したのが第23図で，右側に階級数 n，左側には標準偏差 σ で階級幅を割った値 d/σ，上側には q が目盛ってあり，$N=700$ のときを破線で例示してある．この場合 $q_A=1.1\%$，$q_B=4.2\%$ と読み取れるが，これは，モード階級内のモードの位置によって，階級中央のときの 1.1 %から端のときの 4.2% まで低下率が変化することを示す.

d/σ は図が示すように，$N\fallingdotseq180$ のとき極大値 0.67 に達し，ここで $1.8\%\leqq q\leqq 7.0\%$ となる．$N=1100$ 以上は図に描いてないが，d/σ は再び増加に転ずるが，d/σ が再び 0.67 に達するのは $N\fallingdotseq10^8$ の所で，実質上このような大きな N は考えないでよいから，結局，次のように結論できる.

階級数に(28)式を使ったとき，正規分布の頂点低下率は 7 %以下であって，標本誤差よりも小さい.

この数値例から見て，特に尖りや歪みが大きい分布や特殊な分布を除いた一般の一山型分布では，この誤差は高々10%程度と推測される．つまり，(28) 式は，頂点低下率から見ても問題は無いと言える.

4. 面積誤差の性質と大きさ

階級内の度数折線と分布曲線の面積の差である面積誤差 ΔS を表わす式（32）を，既に第 2 節で導いた．この式から，次の定性的な性質が直ちに知られる.

（1）度数折線には，分布曲線の峰を削って谷を埋める形の面積誤差があり，誤差の大きさは，両隣の平均度数とその階級の度数の差に比例する．それ故，誤差の大きさは峰と谷で大きく，斜面部と裾で小さい.

（2）度数折線の屈折が上方に凸の階級で負，上方に凹の階級で正の面積誤差があり，屈折しない階級では面積誤差が無い.

次に，階級数の式（28）の対象である通常の一山型分布について見れば，一番問題なのは，この誤差が最も大きいモード階級での面積減少が，相対誤差でどの程度かということである．この量を「頂上部面積減少率」と呼び，δ_M で表わすことにすれば，（32）式から，δ_M は次の式で与えられる.

$$\delta_M=\frac{-\Delta S_M}{k_M' d}=\frac{1}{4k_M}\left(k_M-\frac{k_L+k_R}{2}\right) \qquad (34)$$

ここで，添字 M はモード階級を，添字 L, R はその左隣，右隣の階級を表わし，k' は規準化度数，k は通常の度数である．さらに（29）式を使って母集団確率 p_0 で表わせば

第7表　正規分布の頂上部面積減少率 δ_M (%).

頂点の位置	階級幅 (d/σ)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
階級の中央	0.13	0.50	1.1	1.9	2.9	4.0	5.2	6.5	7.9	9.2
階級の端	0.13	0.50	1.1	1.8	2.7	3.7	4.7	5.7	6.6	7.5

第8表　両隣とモード階級の度数比 $\bar{k}/k_M$ と δ_M の関係.

$\bar{k}/k_M$	0.0	0.2	0.4	0.5	0.6	0.7	0.8	0.9	1.0
δ_M	25.0%	20.0%	15.0%	12.5%	10.0%	7.5%	5.0%	2.5%	0.0%

$$\delta_M=\frac{1}{4p_{0M}}\left(p_{0M}-\frac{p_{0L}+p_{0R}}{2}\right) \qquad (35)$$

となるから，δ_M は，確率密度 $f(x)$，階級幅 d 及びモード階級内のモードの位置で決まる．

量的な検討のため，$f(x)$は正規分布とし，区間$[0,x]$の確率積分

$$\Phi(x)=\int_0^x f(t)dt=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt \qquad (36)$$

を使えば，(35) 式は次のように書ける．

$$\left.\begin{aligned}\delta_M&=\frac{1}{8\Phi\left(\frac{d}{2}\right)}\left\{3\Phi\left(\frac{d}{2}\right)-\Phi\left(\frac{3d}{2}\right)\right\}\\&\text{頂点が階級の中央のとき}\\\delta_M&=\frac{1}{8\Phi(d)}\{2\Phi(d)-\Phi(2d)\}\\&\text{頂点が階級の端のとき}\end{aligned}\right\} \qquad (37)$$

一般の正規分布 $N(\mu,\sigma^2)$ のときは，d/σ をこの式の d と考えればよい．階級幅 d の種々の値について，$\Phi(x)$の表を利用して計算した δ_M の値は第7表のようになった．

表の数値の大きさは正規分布のものであるが，数値が示す傾向は，δ_M の一般的な性質を示している．このことも含めて，結果を箇条書きにまとめると，次のようである．

(3) 頂上部面積減少率 δ_M は，通常の一山型分布では階級幅 d と共に増加する．著しく歪んだ分布などでは必ずしもこうならない．

(4) 頂上附近で左右対称な分布のとき，モードが階級の中央のとき δ_M は最大で，端に寄るほど小さい．しかしこの変化は，頂点低下率のように顕著ではない．

(5) 正規分布では，普通使われる階級幅のとき高々数%で，頂点低下率と同程度であり，偶然誤差に比べてずっと小さい．階級数の実用式 (28) を使ったときは，δ_M は約1～5%の範囲にある．

さらに，気象データの実例の検討から知られた次の性質を追加しておく（後記5.3. 節）．

(6) δ_M の大きな値は，尖った分布よりも，むしろ歪みの大きい分布で現われやすい．このようなとき，実例でも10%を越す．

この（6）に関連して，歪み，尖りの影響を少し述べる．

1）モードの一方側が急落した分布

著しく歪んだ分布で比較的よく現われる型で，モードの他の側は割合平坦に近い．いま，右側で急落しているとすれば，$k_L\fallingdotseq k_M$ と置ける．この条件を使えば (34) 式は

$$\delta_M\fallingdotseq(1-k_R/k_M)/8=(1-k_R'/k_M')/8 \qquad (38)$$

となる．δ_M の最大値は $k_R'=0$ のときの12.5%である．$k_R\ll k_M$ ならば δ_M は10%を越す．

2）著しく尖った分布

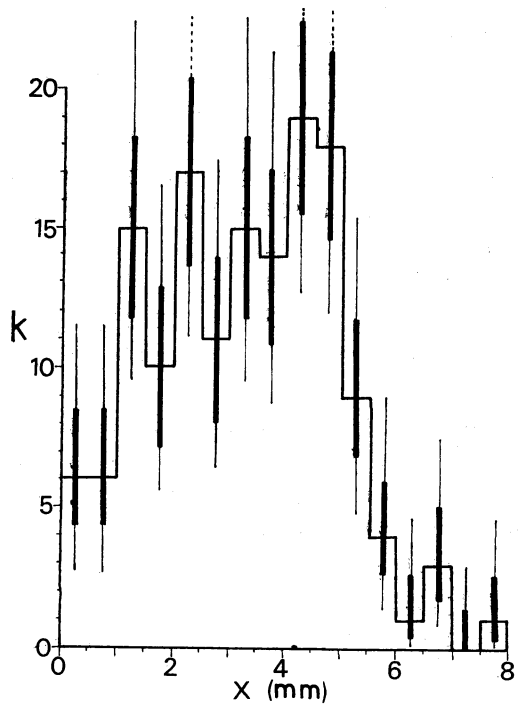
左右の階級の平均度数を使えば，(34) 式は

$$\left.\begin{aligned}\delta_M&=\frac{1}{4}\left(1-\frac{\bar{k}}{k_M}\right)=\frac{1}{4}\left(1-\frac{\bar{k}'}{k_M'}\right)\\\bar{k}&=(k_L+k_R)/2,\quad \bar{k}'=(k_L'+k_R')/2\end{aligned}\right\} \qquad (39)$$

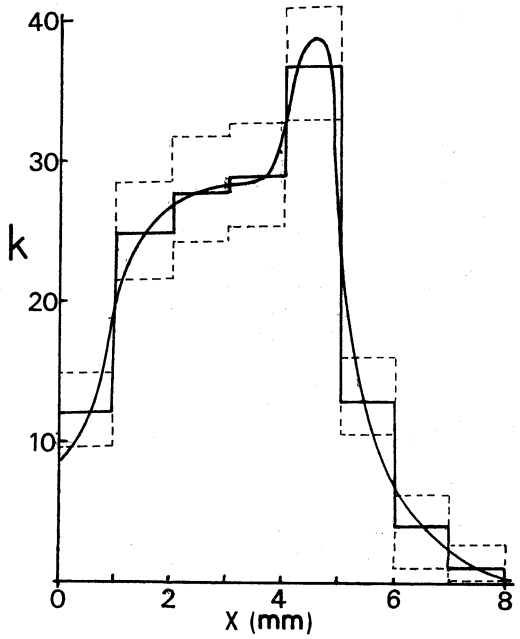
と書ける．比 $\bar{k}/k_M$ を種々に与えて δ_M を計算すると，第8表のようになる．

表で見ると，両隣の平均度数がモード階級の6割以下でないと，δ_M が10%以上にはならないから，尖りのために δ_M が10%以上にはなりにくいと考えられる．以上が性質（6）の理由である．

以上を総合して，階級数の式 (28) を使ったときの δ_M は，大きくても10%程度で，(28) 式は面積誤差からも問題がないことがわかる．



第24図　東京，4月の日蒸発量のヒストグラム（1953～1957年）．
$N=149$，$n=16$，$d=0.5$ mm



第25図　東京，4月の日蒸発量のヒストグラムと標本分布曲線．
$N=149$，$n=8$，$d=1$ mm

5. 実例による考察

ここでは，気象データによる，不規則なヒストグラムと深い谷があるヒストグラム各1例を取り上げ，これまで述べた偶然誤差及び系統的誤差を調べ，併せてまだ取り上げていない系統的誤差を考察する．材料はかなり以前に筆者が集めておいたヒストグラムから選んだので，統計期間が古いが，議論に影響は無い．ヒストグラムと同時に，度数折線も検討する．

5.1. 不規則なヒストグラム

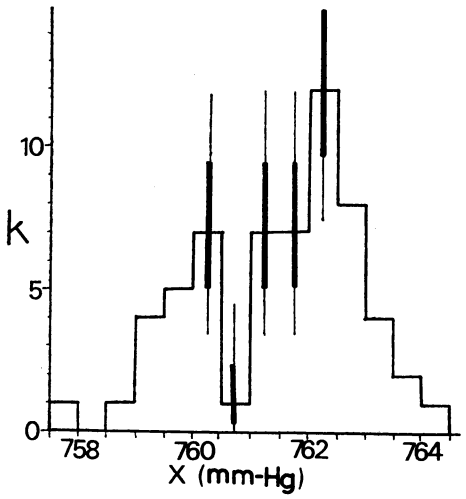
第24図は，東京の4月の日蒸発量を5年間集めて，$d=0.5$ mm の階級幅で描いたヒストグラムで，欠測1日を除いて，$N=149$，階級数 $n=16$ である．

図のようにヒストグラムには $x=1$～4 mm の所に，かなり顕著な凹凸があり，不規則な形状を示している．

母集団分布を一山型と仮定したとき，第1報で定義した度数逆転が生じている現象，換言すれば，隣接する階級の度数を比較したとき，モード階級から遠い階級の度数のほうが大きい現象を，「見掛け上の度数逆転」と定義すれば（混乱のおそれが無い場合は「度数逆転」と略称する），度数逆転は $x=1$～4 mm で3か所，右裾で1か所起こっている．この標本の場合，実用式（28）による階級数は $n=8$ となり，使われた階級数が多過ぎるので，この逆転は標本変動によって起こっている可能性が強い．この点を確かめるために，第4～7図から，各階級の母集団度数の信頼係数 $\alpha=90\%$ 及び $\alpha=50\%$ の信頼区間 $[k_L, k_U]$ を求め，第24図に細い縦線と太い縦線で表示した．これを見ると，逆転が起こっている4か所全部で，狭い方の $\alpha=50\%$ の信頼区間が両隣の階級で重なっている．つまり，図の度数逆転は標本変動で容易に起こりうる．それ故，図の度数逆転は平滑化したほうがよい．

階級数の過大が原因と思われる第24図の標本分布の不規則性を緩和するために，2階級づつ合併して，$d=1$ mm，$n=8$ 階級で描いたヒストグラムが第25図である．この図では度数逆転は解消し，第24図よりずっと単純な一山型の分布になっている．

第25図には，$\alpha=50\%$ の信頼区間の上，下限を，破線のヒストグラムで示してある．第24図の太い縦線の範囲と比較すれば，かなり幅が狭くなり，ヒストグラムの信頼性が増していることがわかる．この図には後記第8節



第26図　東京，4月の月平均海面気圧のヒストグラム（1890～1949年）．
$N=60$，$n=14$，$d=0.5$ mm-Hg

で述べる方法で描いた標本分布曲線も描いてある．ただし，標本が小さく，50%信頼区間の幅が広いうえに，モードの右側で度数が急落しているので，描き方にはかなり主観が入り，母集団分布の推定としては信頼度が小さい．この程度のデータ数では，むしろヒストグラムのままのほうがよい．

5.2. 深い谷があるヒストグラム

第26図は，東京，4月の月平均海面気圧のヒストグラムで，階級幅 $d=0.5$ mm-Hg，標本の大きさは $N=60$ と非常に小さく，階級数は $n=14$ で，実用式による階級数の約2倍である．当然大きな標本誤差が見込まれるが，その割には形が規則的である．ただ，760.5～761.0 mm-Hg の所に深い谷があり，これが母集団分布の谷の存在を意味するのか，標本変動による見掛け上の谷かの判定が必要である．このために，図には，谷の近くの階級について，第24図と同様な $\alpha=90\%$及び50%の信頼区間が記入してあり，谷と両隣の階級の信頼区間は，$\alpha=50\%$では十分離れ，$\alpha=90\%$では一部が重なっている．以下に，この関係から，母集団での谷の実在の有無を客観的に判定する方法を述べる．

2組の独立な標本の標本比率を $p_A=k_A/N_A$，$p_B=k_B/N_B$，$p_A>p_B$ とし，母比率 p_{0A} の信頼係数 $\alpha=1-2\beta$ の信頼区間の下限を p_{0AL}，母比率 p_{0B} の同じ信頼係数の信頼区間の上限を p_{0BU} とするとき，標本の大きさの関係が $N_A \gg N_B$ 又は $N_A \ll N_B$ のいずれでもないときには，仮説「$p_{0A}=p_{0B}$」は，$p_{0AL}>p_{0BU}$ のとき近似的，に危険率 β^2 で棄却されることが知られている（参考文献（2）p.134）．ヒストグラムの2つの階級間では，$N_A=N_B=N$ で，k_A，$k_B \ll N$ ならば一応独立と見なせるから，これを近似的に適用できる．$\alpha=50\%$では危険率は $\beta^2=0.25^2=0.625$，$\alpha=90\%$では $\beta^2=0.05^2=0.0025$ であるから，結局次のように言える．

ヒストグラムの2つの階級間で，$\alpha=50\%$（90%）の信頼区間に重なった部分が無ければ，危険率約6%(0.25%）で，両階級の母集団度数（母集団比率）には有意の差がある．

これを問題の谷に適用すれば，谷，左隣，右隣の母集団度数 k_0，k_{0L}，k_{0R} の間に，$k_{0L}>k_0<k_{0R}$ という関係があることが，危険率約6%で言える．つまり，同じ危険率で次のように結論できる．

第26図に見られる深い谷は，母集団分布の谷を反映したものである．結局二山型分布ということである．このようなときは，一般的に言えばデータの均質性の吟味が必要になるが，主題から外れるのでここでは立ち入らない．

5.3. 境界位置，階級幅によるヒストグラム，度数折線の差異

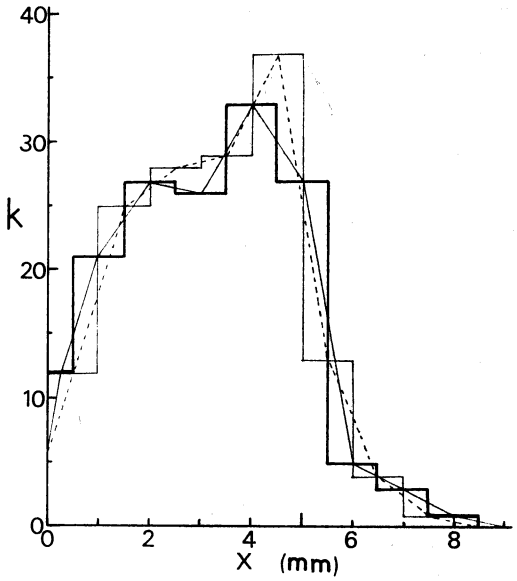
階級を2つずつ合併して階級幅を2倍にした階級別度数は，合併する階級の組み合わせを変えれば2通りでき，それらは，階級幅 d は同じで境界位置が $d/2$ ずれている．第24図の階級別度数を合併して作った2通りのヒストグラムと度数折線を重ね書きしたのが第27図で，その一方は第25図と同じである．また，第26図の階級別度数から同様に作成したのが第28図である．各図の2つの度数分布は，相互の境界位置のずれが最大になっているから，同じ標本，同じ階級数のヒストグラム，度数折線が，境界位置によってどの程度変わるかの限界を，ほぼ表わしている．図から，境界位置の影響として，次のことが挙げられる．

（1）モードの位置が最大で $d/2$ 程度変わる．

（2）分布の大勢は変わらないが，峰の高さ，谷の深さがかなり変動する．

この（2）について，峰の高さの差異は両方の平均値を基準として，第27図で11%，第28図の主峰で5%，第2の峰では14%とかなり大きい．その原因として考えられる次の性質も追加しておく．

（3）標本分布が不規則なほど境界位置の影響が大きい．



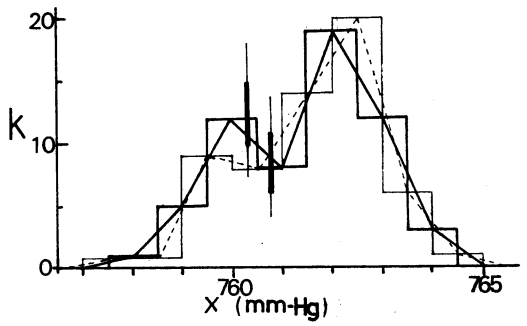
第27図　境界が $d/2$ ずれたヒストグラム，度数折線の比較（日蒸発量）．

次に，階級幅を倍にしたための変化を，もとの幅の第24図，第26図と比較して見ると，標本誤差の減少は別として，第26図で見られる深い谷が第28図では非常に浅くなっている．谷が浅い方はもちろん，深い方の太線のヒストグラムでも，谷と左の階級の $\alpha=50\%$ の信頼区間が重なり，危険率約6％の検定で有意な差は無い．つまり，階級幅より狭い谷が，階級内での平均化で消えた例である．このことについては次の節で考察する．

最後に，度数折線の「ヒストグラムに対する頂上部面積減少率」について述べる．第27，28図の4つの実例について求めた値は第9表の通りである．

表の値は，第27図実線の場合を除きどれも10％を越え，第4節で述べた，分布曲線に対する頂上部面積減少率の一般的な大きさに比べて，かなり大きい．この理由は，ヒストグラムに標本誤差が無いことを仮定しても，次のように説明がつく．

第27図破線で減少率が大きい理由は，第4節でも述べた，分布の歪みによる右隣の階級の度数急減が原因である．また，第28図の場合は，階級数が少なく二山型のために，両隣の度数が共に急減しているためである．仮にこの分布で，左側の第2の峰が無かった，つまり一山型とすれば，全体のデータを $n=6$ くらいの階級に分けたことになり，階級数が著しく少ない場合に相当する．



第28図　境界が $d/2$ ずれたヒストグラム，度数折線の比較（月平均海面気圧）．

第9表　ヒストグラムに対する頂上部面積減少率．

図の種類	第27図 実線	第27図 破線	第28図 実線	第28図 破線
δ_t	4.9%	10.8%	11.8%	12.5%

逆に，第28図の面積減少率が12％程度であることから，一山型分布で実用式の階級数を使えば，分布の尖りが原因で頂上部面積減少率が10％を越える可能性は小さいと見られ，第4節の結果と一致する．

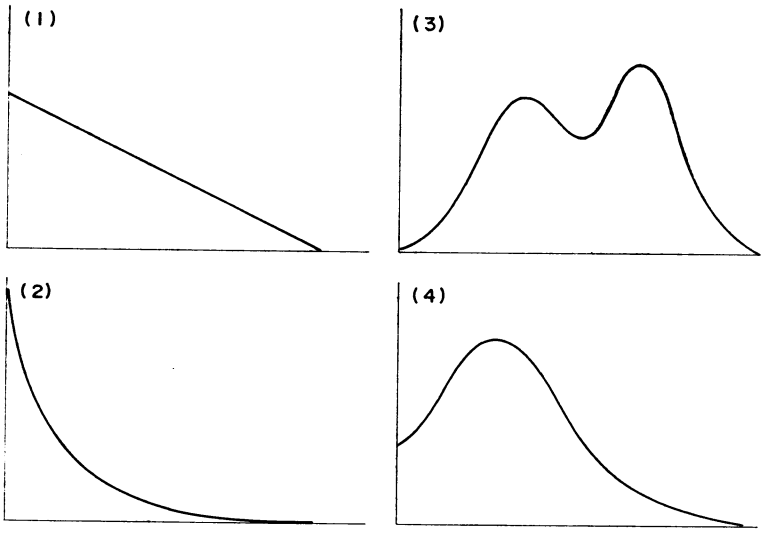
6. 分布曲線型とヒストグラムの誤差

第1報，第2報を通じて，前節までは通常最も多く見られる分布曲線型である，両裾で0に漸近する一山型分布に対象をほぼ限定して，等間隔階級区分のヒストグラム及び度数折線の各種誤差について調べたが，ほぼその作業を終えたので，本節では一般の分布曲線型について，ヒストグラムの誤差を調べる．

誤差のうち標本誤差は，既述のように標本の大きさ N と階級別度数 k できまり，分布曲線型には直接関係しないから，次のことが共通の原則になる．

ヒストグラムの標本誤差を小さくするには，度数 k が小さい階級（特に k が10未満の階級）をなるべく作らないように階級区分する（第1報4節参照）．つまり等間隔区分に限定せず，分布曲線型に応じて階級幅を変えることも必要になる．

系統的誤差のうち，分布曲線型による影響を特に考える必要があるのは，階級内平均化による分布の不明確化である．つまり，階級内平均化で分布曲線型の特徴が失われないように階級幅をえらぶことが原則である．以上のことを，第29図に挙げた代表的な分布曲線型について，具体的に考察する．



第29図　一山型以外の主な分布曲線型（両側有界分布を除く）.
(1) 単調減少・勾配一定型　(2) 単調減少・勾配漸減型　(3) 二山型
(4) 切断型.

(1) 単調減少，勾配一定型（第29図の (1)）

標本誤差が無いと仮定すると，階級幅の広狭にかかわらず，ヒストグラムの頂上中点は分布曲線（直線）上にくる．つまり，度数折線が分布曲線と一致し，系統的誤差が無い．それ故，標本誤差を考慮して次のようにすればよい．

Nが大きいときは等間隔，Nが小さいときは度数が大きい領域で階級幅を狭く．度数が小さい領域で階級幅を広くとる．

(2) 単調減少，勾配漸減型（第29図 (2)）

前の分布に比べて，左端のモードの背が高く，右に裾が長く延びた型で，その極端な例は，日降水量など逆J字型とかL字型とか呼ばれている．系統的誤差の特徴は

1）モード階級での頂点低下率が大きい．

2）分布曲線が上方に凹のため，ヒストグラムの頂上中点は分布曲線より上方にずれる．

このために，階級幅が広いと，モード付近の分布型の推定が著しく困難，不正確になる．これを例示したのが第30図で，このヒストグラムに対応する分布曲線は，図の曲線Ⅰ，Ⅱ，Ⅲなど，どのようにも描ける．階級幅は図の1/5程度に狭くする必要がある．

一方，標本誤差から見ると，モード付近は階級幅を小さくしても，元々度数が多いから支障は無いが，右裾の部分は期待度数が著しく小さくなり，標本誤差が増大して分布が不安定になる．分布が長く延びた裾の部分では，逆に通常より階級幅を大きくとって期待度数を増やす必要があり，この部分は分布曲線の勾配変化も小さく，幅を広げても上記2）の系統的誤差からも支障は無い．結局，等間隔区分は不適当で，次のように区分するのが合理的である．

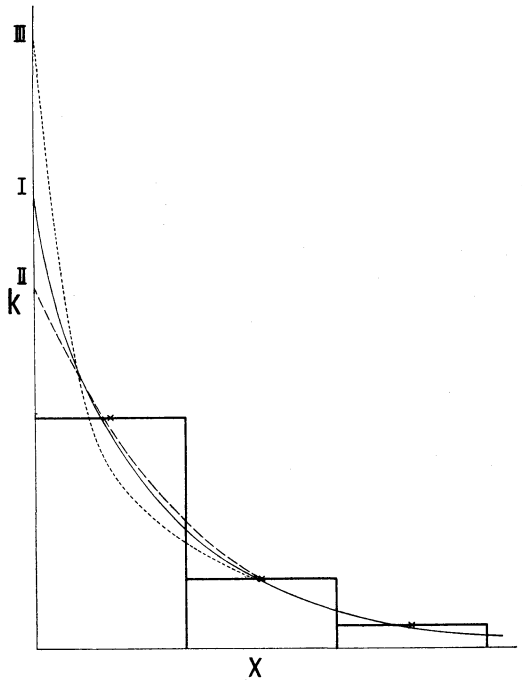
データが多いモード付近は特に細かく区分し，データが少ない裾の部分は特に階級幅を広げる．（データの実情に応じて，3～数個の領域ごとに階級幅を決める．）

(3) 二山型（第29図の (3)）

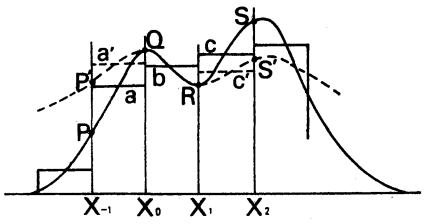
ヒストグラムの階級幅が広くて，二山型分布の2つの峰が同じ階級に入れば，もちろん谷は消滅し，この分布型の特徴が失なわれる．階級内平均化のために，ヒストグラムの谷は分布曲線の谷より必ず浅くなるが，その浅くなり方は階級幅 d が広いほど著しく，境界の位置にも関係する．このことは前節の実例でも示したが，階級幅は，標本誤差による凹凸と見誤らない程度の谷が残るように小さくとる必要がある．以下この点を調べる．

1）$d=D$ の場合

2つの峰から中間の谷までの距離のうち，小さい方を D とし，まず階級幅 d が D に等しい場合を考察する．谷の中心が階級の中央と一致するときは，ヒストグラムにかなりの谷が残るのは明らかであるから，最も谷が残りにくいように，「峰から谷まで」が1つの階級になっ

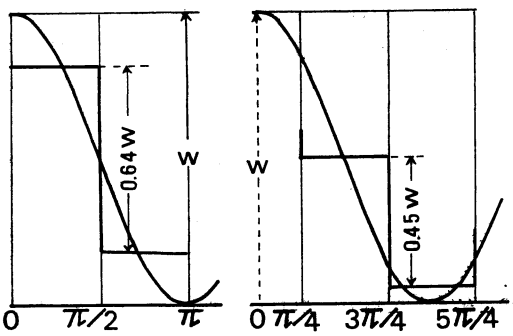


第30図　同じヒストグラムに対応する分布曲線（勾配漸減型）.



第31図　二山型分布の谷とヒストグラムの谷の有無（$d=D$ の場合）.

実線のヒストグラム a, b, c は実線の曲線 PQ, QR, RS に対応，破線のヒストグラム a', c' は破線の曲線 $P'Q$, RS' に対応.



第32図　峰から谷までの Cosine 近似（$d=D/2$ の場合）.

ている場合を第31図に示す．峰を X_0 谷を X_1 とし，両隣の階級の端を図のように X_{-1} 及び X_2 とする．また点 X_i の分布曲線の高さを $h_i=f(x_i)$，$(i=-1\sim1)$，3つの階級のヒストグラムの高さを左から H_-, H_0, H_+ と書くと，分布曲線の形により次の4通りの場合がある.

① $h_{-1}<h_1$, $h_0<h_2$ のとき（曲線 PQRS）

左の峰から谷側の QR よりも外側 QP の方が分布曲線の落ち込みが激しいから，ヒストグラムは実線 b, a で表わされ，$H_-<H_0$ となる．同様に RQ より RS の方が分布曲線の上がり方が激しいから，ヒストグラムは実線 b, c で表わされ，$H_0<H_+$ となる．結局，$H_-<H_0<H_+$ となり，ヒストグラムは実線 a, b, c で表わされるように谷が消える.

（注）厳密にはヒストグラムの高さは境界での分布曲線の高さだけでは決まらない．上記の条件 $h_{-1}<h_1$, $h_0<h_2$ は近似的な表現である．以下についても同様.

② $h_{-1}>h_1$, $h_0<h_2$ のとき（曲線 P'QRS）

同様の考察からヒストグラムは図の $a'bc$，つまり，$H_->H_0<H_+$ となり，浅い谷が残る.

③ $h_{-1}<h_1$, $h_0>h_2$ のとき（曲線PQRS'）

ヒストグラムは abc', $H_-<H_0>H_+$ となり階級$[X_0, X_1]$ の谷は消えるが，X_2 を左端とする階級での $f(x)$ の平均の高さが図の c' より高いか低いかによって，谷が消える場合と，階級 $[X_1, X_2]$ が浅い谷になる場合とがある.

④ $h_{-1}>h_1$, $h_0>h_2$ のとき（曲線 P'QRS'）

ヒストグラムは $a'bc'$, $H_->H_0>H_+$ となるが，前項と全く同じ理由で，谷が消えるときと $[X_1, X_2]$ が浅い谷になるときがある．前項とは左の峰の位置が1階級ちがっている.

2）$d=D/2$ の場合

第31図の左の峰から谷の少し先までを，三角関数 $\cos\theta$ で近似し，$d=D/2$ の階級で区分したヒストグラムを示したのが第32図である．横軸は位相角 θ で表現し，左図は境界が峰，谷と一致しているとき，右図は，峰，谷が階級の中央にあるときで，後者は谷の所のヒストグラムが最も低くなる場合である.

左図のとき，谷を含む階級と左隣の階級のヒストグラ

ムの高さの差は，分布曲線の谷の深さ W の64%，右の図では45%である．つまり左図の位置では，分布曲線の谷の深さの64%は少なくとも残り，もし峰の左側の落ち込みが谷側よりも少なければ，$[-\pi/2, 0]$ の階級のヒストグラムの高さは更に高くなり，谷の深さも64%より深くなる．右図で，峰の左側の分布曲線の落ちこみ方が谷側と同程度で，$[-\pi/4, 0]$ まで同じ cosin 近似ができるとすれば，峰を含む階級のヒストグラムは斜面部の階級より更に $0.45W$ 高くなり，谷の深さは $0.9W$ となる．$0.45W$ は，峰の外側がいかに急落しても確保される下限値であるから，これと $0.9W$ を勘案すれば，峰の外側が急落している殊殊な形を除いた通常の二山型分布では，峰の階級は斜面部の階級よりある程度ヒストグラムが高く，ほぼ左図と同程度の谷は確保されていると見てよいであろう．一般の境界位置はこの2つの図の中間にあるから，谷の深さの残り方もこの2つと同程度と見てよい．

以上1），2）の考察から結論をまとめれば，次のようになる．

二山型分布の峰から谷までの距離（小さい方）D の階級幅のヒストグラムでは，谷が全く消滅する危険がある．$D/2$ の階級幅ならば，分布曲線の少なくとも6割程度の深さの谷がヒストグラムに残る．

なお，第5.3.節の実例でものべた，峰，谷の位置の $\pm d/2$ のずれが，この項の検討例でも見られる．

（4）切断型（第29図の（4））

通常の一山型分布に従う変量の，一定限界値以上のデータだけを集めれば，それに対応する母集団分布はこのようになる．原理的には，分布曲線を延長して一山型の分布を再現したときの面積の増加率を S（>1）とすれば，$N'=NS$ に相当する（28）式の階級数で，一山型全体を区分すればよいことになる．簡単な1例を挙げよう．

（例）正規変量の平均値からの正偏差の分布

負偏差のデータは捨てるとして，正偏差のデータ数を $N=300$ とする．母集団分布は正規分布を平均値で2等分した右半分で，面積増加率 S は2.0である．$N=300$ の階級数を直接（28）式から求めると，$n=8.4$ つまり8階級となるが，上の方式では $N'=NS=600$ に相当する階級数として $n=10.0$ が得られるから，その半分の5階級でよい．

一般の場合に上のような原理的な階級数の決め方をするのはわずらわしいから，次のことを考慮して階級数を決めればよいであろう．

切断型分布では，（28）式より少なめの階級数でよい．

7. ヒストグラムの描き方

各種誤差の検討は前節までで終わり，この節では，これまでの検討に基づいて，母集団分布の推定を目的とした，ヒストグラムの描き方の標準的手法を提案する．

度数折線ではなくヒストグラムを対象とした理由は，前者は分布曲線に対して，後者の誤差に加えて面積誤差があり，分布曲線の推定に不適当だからである．ただし補助的な利用はあり（後記第8節），ヒストグラムの頂上中点を結んで容易に得られる．以下では手法を箇条書きし，必要な場合は * 印以下に補足説明し，末尾に主に関係がある節番号を〔 〕で示した．

（1）できるだけ多数のデータを集める．

* 階級数の実用式（28）ではヒストグラム作成の最低条件を $N=50$，$n=7$ としたが，これは大きな標本誤差と系統的誤差を覚悟してのことで，分布の大勢が推定できれば上々である．標本誤差の点から，ヒストグラムが十分信頼できる中心誤差率10%（90%信頼区間の片側）の確保には N が600以上が必要であり，分布型にもよるが，系統的誤差から十分安心できる理想的な階級数 $n=20$ には $N=1100$ 個が必要である．つまりこのくらいまでは，N が大きいほどよい〔第1報7節〕．

（2）母集団分布の型について知識を持っており，等間隔区分が著しく不適当なときは，分布型に応じた階級区分をする〔第6節〕．これに該当しないときは（3）による．

* 一部の領域で標準の幅 d と異なる階級幅 d' を使ったときは，その領域のヒストグラムの高さは，度数 k に d/d' を乗じて，幅 d の度数に換算した度数で描く．数種の階級幅を採用したときも同様にする．

（3）階級区分の標準的な方法は下記による．

1）標本の大きさ N から実用式（28）で階級数 n を決める．

2）標本の範囲を n で割った値を，きりのよい数値に丸めた値を階級幅 d とする．

3）境界位置もなるべくきりのよい数値を採用し，等間隔に階級区分する．

（4）ただし，次の事項に留意する．

1）上記（3）の1）による階級数 n が10未満のときは，とりあえず階級数を $2n$ としてヒストグラム

を描く

2）各階級の度数が著しくアンバランスのときは，度数が多い所はさらに分割し，度数が少ない所は合併する．

＊ 1）は狭い谷など母集団分布の細かい構造の消滅を防ぐためである．

（5）上記（2)～(4）の階級区分でヒストグラムを描く．

＊ 縦軸は基準化度数 k' でなく，通常の度数 k でよい〔第2節〕が，不等間隔のときの均質化は必要である（上記（2）の＊参照）．

（6）各階級の度数の信頼限界を，第4～7図及び(8)式で求め，ヒストグラムに記入する〔第1報3節，第2報5節〕．

（7）階級数を $2n$ としたとき，ヒストグラムに度数逆転（第24図）や深い谷（第26図）などの不規則性があれば，記入されている信頼限界によって，母集団分布の特徴か標本変動による見掛け上のものかを判定する．

（8）前項で標本分布の不規則性が標本変動によると判定されたときは，2階級ずつ合併して，階級数 n で改めてヒストグラムを描き，信頼限界をつける．

（9）第（7）項で度数逆転や谷などが母集団分布によると判定されたときは，その部分を除いて2階級づつ合併してヒストグラムを描き，信頼区間をつけるか，又は $2n$ 階級のヒストグラムをそのまま採用する．

＊ 母集団分布による度数逆転や谷があり，他の部分に標本変動による顕著な不規則もあるときは，前者を除いて2階級合併，他の部分の不規則性が小さいときは $2n$ 階級のままとするのが良い．

(10) 特に標本分布の不規則性が著しく，第（8），(9)項の2階級合併の2通りの組み合わせ方で結果が著しく違う場合は，その両方（階級境界が $d/2$ ずれた2つのヒストグラム）を重ね書きする（第27，28図参照）．

＊ 母集団分布は両方のヒストグラムを総合して推測する．

(11) 第（3）項で階級数 n を決めるとき，1個，2個などごく少数のデータの値が他と著しく飛び離れているときは，そのデータを除いて階級数，階級幅を決めて階級区分し，飛び離れた値の部分は幅を広げた階級で図に追加する〔第1報8節〕．

＊ ヒストグラムを平滑化して標本分布曲線を描くときは，後記第8節による．

以上が各種誤差の性質と大きさを考慮した，母集団分布推定のためのヒストグラムの描き方の標準手法である．これは，ヒストグラムを描くときには，いつでもこのようにせよという意味ではなく，このように描くのが，最も合理的であるという意味の提案である．それ故，各条項の内容と意味を知った上で，場合に応じて一部の作業を省略することは，一向さしつかえが無い．例えば，描いたヒストグラムの形が十分規則的なときには，ヒストグラムに信頼限界をつけたり，標本誤差を評価したりする必要は通常無い．その反面，ヒストグラムが不規則だったり特異なときは，このことが必要であり有効である．

8. 標本分布曲線の描き方

一般に，標本誤差が大きくてヒストグラムの凹凸が激しかったり，著しく歪んでいて隣接階級で度数が急落しているようなときは，これを平滑して標本分布曲線を描いても，見掛け上の形が良くなるだけで，曲線に客観性が乏しく信頼できない．反対にヒストグラムが非常に規則的なときには，強いて平滑曲線を描かなくても，母集団分布の推測はでき，ヒストグラムの方が客観的で良い．それ故，ここではヒストグラムがあまり不規則でなく，多少の凹凸と階段をならして標本分布曲線を描く場合，それも等間隔階級で，度数の急落も無いとして描き方を述べる．

（1）等間隔の階級別度数でヒストグラムを描き，次に各柱の頂上中点を結んだ度数折線を重ね書きする．折線は両端の度数0の階級まで延ばしておく．

（2）第21図に示したような分布曲線とヒストグラムの関係を念頭に置き，ヒストグラムと各階級内の面積が変わらないように，平滑化曲線を描く．

この場合，最も描きやすい斜面部から描きはじめ，中心部と裾の方へ延ばしてゆく．斜面部では度数折線がそのまま，又はわずかな修正で利用できることが多い．

（3）分布の中心のモードの所は，両側面の曲線の傾向を延長して，曲線の頂点はヒストグラムより多少高く，かつ等面積の条件が保たれるように，推定曲線を描く．

（注1）モード階級の隣りの階級で度数が急落している歪みや尖りが大きい分布では，モード階級の分布曲線の形やモードの高さの推定が非常に不正確になる．このような分布を対象から除いた理由である．

（注2）モード階級と両隣の階級の度数から，モード階級内の位置を決める計算式があるが，この3階級

にわたって，分布曲線がモードを中心とした左右対称形であることを前提としており，この条件は成立たないことが多いので，使わない方がよい．

（4）分布の裾の所はヒストグラムより分布曲線の方が広がっているのが常態であるから，適当に曲線を延ばしておく．この裾の部分は，もともと信頼が置ける曲線は描けないからである．

9. まとめ及びあとがき

この報告（第1報，第2報）は，従来まともな研究が少なく，総合報告的なものも無い標題の問題を取り上げ，各種誤差をできるだけ総合的・量的に評価し，ヒストグラム，度数折線の合理的な描き方を示すことによって，従来常識的で概ねあいまいな方法で処理されていたこの作業を，できるだけ明確な根拠で行ない得るようにすることを目的としたものである．報告した内容は，この日常的な問題を，概ね初歩的，常識的な手法で検討したもので，結果も大勢において従来の常識とかけ離れたものではなく，理論的な成果は特に無い．しかし目的としたデータ処理技術の立場からは，母集団分布に対する各種誤差の性質と影響が，量的な面も含めて総合的に整理されたことのほか，部分的にも，下記の点等で一応の成果が得られたと思われる．

（1）階級別度数の標本誤差の大小を表現する，信頼係数50%，90%の信頼区間を求めるノモグラムを作成し（第1報第4～7図），これによってヒストグラムの不規則な凹凸，谷などの母集団分布での実在性の有無が，あまり手数をかけずに客観的に判定できることを確認した．このノモグラムは，母集団度数から標本度数の変動範囲を求めるにも近似的に使える．

（2）標本の大きさ N から等間隔階級区分の階級数 n を求めるための，根拠があいまいで誤差から見ても不適切な従来の式に代えて，誤差に関して根拠を持った階級数の実用式（28）を提出し，その適用範囲及び性質を示した．

（3）各種誤差の性質，大きさを考慮したヒストグラムの描き方の標準的な手法を提出した．

以上，主な成果をまとめたが，これで問題がすべて明確になったわけでは勿論ない．今回採用した誤差検討の手法にしても，例えば度数逆転の評価法（第1報6.3.節）など，かなりあいまいな点があることを承知のうえで採用した部分もある．また，特殊な分布型の場合の誤差やヒストグラムの描き方についても，ひと通りの検討を行なったが，十分とは言えない．これらのに点ついては，今後の課題としたい．

なお，創価大学情報科学研究所 池田貞雄 教授，統計数理研究所 平野勝臣の両氏には，文献の所在，入手等について多大の御援助をいただいた．終りに臨み，厚く感謝の意を表明する．

文　献

菊地原英和，1981：ヒストグラムの誤差と描き方　第1報 標本誤差，天気，28，395-411．

小河原正己他，1957：統計公式および図表とその使い方，気象学講座，19，地人書館．