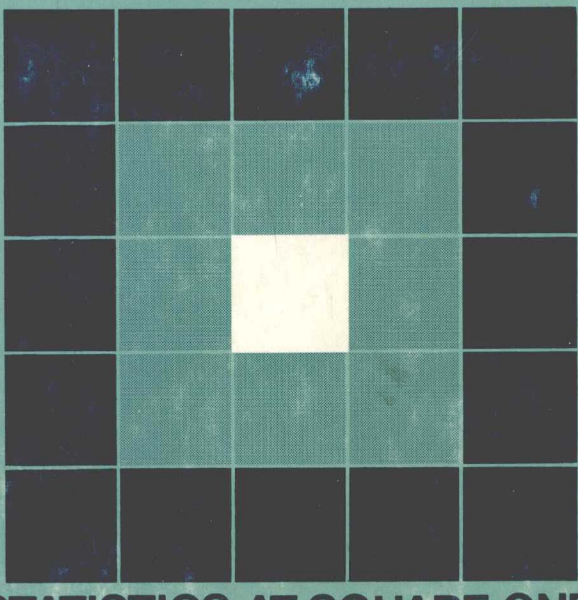
医学・薬学・生物学のための

統計処理

T. D. V. Swinscow 著

西村昂三 監訳　大島邦夫 訳



STATISTICS AT SQUARE ONE

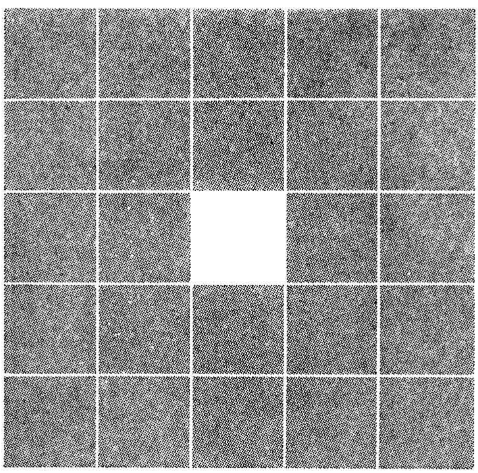
共立出版株式会社

医学・薬学・生物学のための

統計処理

T. D. V. Swinscow 著

西村昂三 監訳　大島邦夫 訳



STATISTICS AT SQUARE ONE

監訳者・訳者 紹介

西村昂三（にしむら・こうぞう）

1929年２月16日 京都生まれ
1952年 京都府立医科大学卒業
1955～58年 テネシー大学医学部小児科レジデント
1958年 テネシー大学医学部小児科大学院修了（M. S. Ped）
1958～60年 ボストン小児病院フェロー
米国小児科専門医
現在，聖路加国際病院小児科医長・同看護大学教授・東京医科大学小児科客員教授・医博

大島邦夫（おおしま・くにお）

1947年４月25日 東京生まれ
1971年 東京理科大学理工学部数学科卒業
1978年 米国ヒューストン大学博士課程修了
現在，東京理科大学理工学部数学科助手・Ph. D.
東京小児白血病治療共同研究委員会統計担当委員

医学・薬学・生物学のための
統計処理

定価 1600 円

1982年10月１日 初版１刷発行

監訳者 西村昂三
訳 者 大島邦夫 © 1982

発 行 **共立出版株式会社**／南條正男
東京都文京区小日向４－６－19
電話東京（03）947局2511番（代表）
郵便番号 112/振替口座 東京1-57035番

印 刷 藤本綜合印刷

製 本 関山製本

社団法人
自然科学書協会
会員

検印廃止
NDC 417
ISBN 4-320-01080-9

Printed in Japan

序

医学上の多くの所見は，言葉によって最も自然に言い表わせる．たとえば，「患者の脈拍は，記録できなかった」といった具合である．

これに対し，他には具体的な数字で表現する場合がある．たとえば，「患者の脈拍数は，1分間80であった」などである．

統計処理の目的は，数量的データを付け加えることにより，データ全体をよりわかりやすくすることにある．この本の目的は，そういった統計処理の方法を，実例を用いて解説することである．また数学の知識は基本的なもので十分である．実例は主に臨床医学から採用したが，他の生物学の分野にも適用できる配慮をしてある．

統計処理に精通していることは，たとえそれが基礎的なものであっても，次の二つの理由で有益である：

それは，第1に，研究者は自分のデータに関して論議することで，データをより一層深く理解できるのと同様に，自分で統計学的手段によりデータを処理することは，より明瞭なデータ全体の意味内容を研究者自身に示唆するからである．

第2には，この本に書かれているような比較的，基礎的な問題について聞ける統計学の専門家が身近にいないことが多いし，いたとしても，彼らは自分の仕事で忙しいからである．

医学および，他の多くの生物学上の問題を解析するにあたって，複雑で高度な統計処理は，ほとんど必要ではない．むしろ研究者は，真に比較可能な無作為に抽出されたデータを得るための研究計画に，より一層の工夫を施さねばならない．そしてこの無作為抽出されたデータに対する統計処理は，可能な限り単純にしておくのが望ましい．

実際，単純な統計処理では有意な結果を得られなかったが，より精巧で複雑な統計処理により，有意な結果を得たという場合には，特に注意深い考察が必要である．この有意な結果は，複雑な処理のために，単純な処理よりもより多くの要因を使用したことによって得られたものかもしれないし，これはまた，実際には重要でない有意差の存在によるものかもしれないのである．

統計学上での有意な結果が，必ずしもわれわれの分野においては有意とはならないことを認識しておくべきである．

結論としては，まず，完全で厳密な研究計画が第１で，次に，単純な統計処理が，概して，実際に応用する上で最も意義ある結果を生むということである．

ここで，統計処理を行なう場合，どのような電卓を選んだらよいかに触れておく．

この本で述べる計算内容には，ごく普通の小型電卓で十分である．少々高価な電卓には，普通のキーの他に，２乗，平方根，符号変換などのキーが付いている．普通の電卓で十分に用はたりるが，２乗，平方根などのキーの付いた電卓は，計算をはるかに楽にするばかりか，メモリー付きであるならば，計算途中の結果を紙に書きとめる手間をも省ける．

電卓の長所は，ボタン一つで，２乗，あるいは平方根を求め

られることであるが，短所は 8 桁以上の数表示ができないことである．標準偏差を計算する上で，この制限を取り除く処理方法は，第 13 章で示している．

大量のデータについて統計計算を行なっているときには，おうおうにして邪魔の入るものである．注意力が散慢になったり，来客があったり，電話がかかったりして，エラーを起こしてしまう．機械は完全無欠であるが，その使用者は人間なのである．したがって，すべての計算は十分な用心をしながら検算すべきである．検算を容易にするために，私は長い計算の途中の結果を記しておくノートを作っている——たとえ，その仕事が，縦に長く並べた数を合計するような簡単なことであってもである．そして私は，2 度目に計算をするときには，最初のときには違った順序でやってゆくようにしている．

本書は，*British Medical Journal* に 1976 年に連載された記事を基に，加筆・修正されたものである．出版にあたって，P. Armitage 教授から多くの有益な助言をいただいたことに，特に感謝している．また Dr. A. W. F. Edwards, Dr. I. D. Hill, Dr. D. Middleton, それに，R. G. Record 教授からは，連載された記事の読者として，多くの御批評を頂いたことに感謝している．無論，本文中に誤りがあれば，その責任は，すべて私にある．

T. D. V. Swinscow（スウィンスコゥ）

目　次

1

表作成と平均値

統計処理ことはじめ

あるデータに対して，統計処理——それが最も単純なものであっても——を行なう前には，データを何らかの表にするべきである．たとえば，約 30 例ぐらいの比較的少ないデータであれば，小さい順に並べてみることである．

例をあげて説明すると，ある市民病院小児科医のグリーン医師は，その地域の子供の尿に含まれる鉛の含有量を調査することにした．まず，ある通りに面した家に住んでいる 1 歳以上 16 歳未満の小供 15 名の尿中鉛濃度（μmol/24 h）を調べて，小さい順にその値を並べ 表 1.1 を作った．この表は，ときおり，**配列**（array）と呼ばれる．

表 1.1——15 名の子供の尿中鉛濃度（μmol/24 h）

0.1, 0.4, 0.6, 0.8, 1.1, 1.2, 1.3, 1.5, 1.7, 1.9, 1.9, 2.0, 2.2, 2.6, 3.2

このような表を作る利点は，まず，それぞれのデータとの前後関係が見ただけでわかるばかりか，一様に増加しているか，特筆すべきギャップがあるかなど，単純な考察で不自然な点が見つけられること，さらには，最小値が 0.1 で最大値が 3.2 の

値域（range）にすべてのデータが納まっていることがわかる，ということである．そして，グリーン医師は，0.1 から 3.2 まで，出てきた値を順次加えて 22.5 を求め，次に，これを 15 名で割り，1.5（μmol/24 h）を算出した．つまり**平均値**（mean, average）を求めたのである．

この計算を，数学記号を用いて示すと

$$\bar{x} = \frac{\sum x}{n}$$

となる．ただし，$\bar{x}$（エックス・バーと読む）は平均値，x は各観測値，n は人数，$\sum$ はギリシャ文字でシグマと呼び，ここでは，順次 n 個まで加えるという意味で使われており，合計を求めることである．

度数分布とは

しかしながら，グリーン医師の当面の目的に対して，15 名という人数から得られる結果は少なすぎるので，彼は，調査対象を，その通りを中心にひらけている小工場地域に広げることにした．そして，1 歳以上 16 歳未満の子供 140 名について尿中鉛濃度を検査し，このデータの扱いを容易にするために，最も適した型である度数分布表を作成することにした*)．

ここで，**度数分布表**（table of frequency distribution）を作るということは，この例においては，単に鉛の濃度により子供の数を区分けすることである．つまり尿中鉛濃度に対して，それに対応する小供の数を順次記録することである．

---

*) 訳者注：一般的にいえば，**度数分布**（frequency distribution）とは，度数（データの個体数）で表わされたデータの分布状態のことである．

具体的には，まずグリーン医師は，度数分布表を作るにあたって，適当な濃度の**階級**（class）と，その境界値を定めなければならないのである．しかし，もしその階級が小さければ，子供の数は各々ばらばらにそれぞれの階級に入ってしまって扱いにくいし，大きすぎれば，子供の数が一つの階級に入りすぎて単純になり，知りたい結果が得られないことになる．また，階級幅を定める場合は，実際のデータがその階級に簡単に割り振りできるようにすることと，各階級が重複しないようにすることが大切である．

さて，グリーン医師が，子供140名の尿中鉛濃度を調べた結果，その取りうる値の範囲は，0.1 から 4.2（μmol/24 h）であった．ここで，もし各階級幅を 0.5（μmol/24 h）として分けると，その取りうる範囲をすべて含むには九つの階級が必要であり，もし 0.25 にすると 18 階級，そして，0.2 にすると 22 階級がそれぞれ必要になる．一般的には，10～20 階級に分割することは容認される．また階級幅をすべて等間隔に取れば，それぞれの階級の**級中値**（half-way points）を簡単に見つけることができる．

このことは，平均値や標準偏差を求めるときに実際，必要である．たとえば，1.8 と 2.0 の級中値は 1.9 であるが，1.75 と 2.0 の級中値は，直ちには見つからないが，1.875 となる．大切なことを付け加えれば，最初は広すぎる階級幅を選ぶよりも，どちらかといえば，やや狭めのほうがよい．それは，後になって，必要とあらば小階級幅をまとめることで容易に広くできるからである．

以上を考慮して，グリーン医師は階級幅 0.2 を採用して 22 の階級に分割することにした．

その度数分布表は 図 1.1 に示されているが，その読み方を説明すれば，それぞれの階級幅に入る尿中鉛濃度の値 1 個に対して縦線を 1 本引き，4 番目の値までは縦線を 1 本ずつ付け加え，5 番目の値がその階級に入るときは，今まで引いた縦線 4 本に重なるように横線を 1 本引く*).

グリーン医師は，図 1.1 の度数分布表では表が大きくなりすぎたので，階級幅を 図 1.1 で取ったものの 2 倍（0.4 μmol

鉛濃度 (μmol/24h)		
0-	I	1
0·2 -	I	1
0·4 -	III	3
0·6 -	IIII	4
0·8 -	卌	5
1·0 -	卌	5
1·2 -	卌 II	7
1·4 -	卌 IIII	9
1·6 -	卌 卌 I	11
1·8 -	卌 卌 II	12
2·0 -	卌 卌 卌	15
2·2 -	卌 卌 III	13
2·4 -	卌 卌	10
2·6 -	卌 IIII	9
2·8 -	卌 IIII	9
3·0 -	卌 II	7
3·2 -	卌	5
3·4 -	卌 I	6
3·6 -	III	3
3·8 -	IIII	4
4·0 -		0
4·2 -	I	1
4·4		0
合計		140

図 1.1——度数分布表作成のためのデータ表

*) 訳者注：欧米では，一般的にこのような方法で，5 個を一単位としてかぞえるが，日本においては，「正」という漢字が，この操作を代用している．わざわざ直訳したのは，外国の慣習を知る一つの機会と思われたからである．

/24 h) にして，2番目の表を作ることにし，これが妥当であるかどうかを調べてみた.

（1） 人数の分布は，階級値の増加にともない一様に変化しているので，階級を合併することが，合併以前の階級における人数の相対的変化を不明瞭にはしない.

（2） 各値が，生理的なばらつき，検体の採取，検査室の計量などにおいて誤差を含んでいるので，各値は見かけより厳密とはいえない.

以上二つの理由から，グリーン医師は，階級幅を2倍にして表 1.2 を，改めて作成した.

ここで，一つの問題が提起されるのであるが，1歳以上 16 歳未満の子供 140 名を一集団と考えて，これを基礎として計算することは妥当なのであろうか. このことは，後節の“同類，異類の比較と計算”のところで詳しく述べる.

表 1.2——子供の尿中鉛濃度

鉛の濃度（μmol/24 h）	人数
0 －	2
0.4－	7
0.8－	10
1.2－	16
1.6－	23
2.0－	28
2.4－	19
2.8－	16
3.2－	11
3.6－	7
4.0－	1
4.4	
合　計	140

度数分布から平均値を求める方法

グリーン医師は調査範囲を拡張し，140 名の子供の尿中鉛濃度を調べて，その結果を 表 1.3 にまとめた．表 1.3 の 1 番目と 2 番目の列（column）は 表 1.2 と同じである．

平均値を求めるためには無論，度数分布表を作る以前の生データを順次加えて 140 で割ればよいのであるが，度数分布表を用いた簡便計算処理で求めた結果が，表 1.3 に示されている．その処理について説明しよう．

まず，平均値を求めるために，次の二つの仮定を定めた：

（1） 尿中鉛濃度は，現実には，測定機器の精度により測定値の連続性（continuity）には制限が加えられるが，最小値 0 から，最大値 4.4 までの，任意の値を取りうるものとする．（連続変量と離散変量については第 2 章の中の“連続変量と離散変量”の節で詳しく述べる．）[*]

（2） 任意の階級に含まれる値（たとえば，0.4 以上 0.8 未満の間には 7 個あるが）は，一様に分布しているものとする．

仮定（2）を考慮すると，ある階級における級中値はその階級の両端に近い値よりも，その階級を良く代表している．たとえば，0.4 以上 0.8 未満の間には 7 個の値があるが，7×0.6（級中値）は，7×0.4，7×0.79，または 7×（級中値以外の数）よりも，その階級においては，真の 7 個の値を合計したものとして，一般的には，より良い概算である．この理由により，そ

---

[*] 訳者注：身長や体重のように，測定または計量されるものを**連続変量**と呼び，人口とか，ある事実の起こる回数などを**離散変量**という．尿中鉛濃度は，前者である．

れぞれの階級の級中値を列（3）に示した．それぞれの階

列（4）は，列（3）×列（2）の値，つまり，それぞれの階級の級中値に，その階級の子供の人数を掛けたものである．列（4）の下に記されている 305.6 は尿中鉛濃度の合計である．平均値を求めるためには，合計を 140 で割ればよい．

もしこの計算をメモリー付電卓で行なうならば，列（4）は，書く必要はない．尿中鉛濃度の掛け算を行ない，その結果を一つ一つメモリーに入れ，最後に，各値はメモリーから自動的に引き出され，その合計を求めることができる．そしてまた，列（1）の階級幅を以前に述べたように，注意深く選んでおけば，級中値は明白で，列（3）は必要なくなり，メモリー付電卓を用いれば，表 1.2 から直接合計を求めることができる．

表 1.3——子供 140 名の尿中鉛濃度の平均値の計算

(1) 鉛濃度 μmol/24 h	(2) 人数	(3) 列(1)の級中値	(4) 列(3)×列(2)
0 —	2	0.2	0.4
0.4—	7	0.6	4.2
0.8—	10	1.0	10.0
1.2—	16	1.4	22.4
1.6—	23	1.8	41.4
2.0—	28	2.2	61.6
2.4—	19	2.6	49.4
2.8—	16	3.0	48.0
3.2—	11	3.4	37.4
3.6—	7	3.8	26.6
4.0—	1	4.2	4.2
4.4			
合計	140		305.6

鉛濃度の平均値＝305.6/140＝2.18 μmol/24 h

同類，異類の比較と計算

前節で1歳以上 16 歳未満の子供 140 名を一集団（同類）と考えて計算することが妥当であるかどうかが，問題になった．彼らが子供である限りにおいては，この調査対象は明らかに同集団に属しているし，多くの共通点がある．しかしこのことはしばしば注意を要することがある．なぜならば，子供といっても男女の別があり，かつ1歳になったばかりの子供もいれば 16 歳くらいという子供もいる．このような相違は，しばしば隠されてしまうが，しかし十分に考慮すべきである．

たとえば，1973 年中期のイギリス統計局の発表によれば，バーミンガム病院地区の人口は，5,163,200 名で，サウスウエスタン病院地区の人口は，3,246,200 名であった．前者に対する女子の割合は 50.49% であり，後者は 51.81% であった．その差は 1.31% である．この事実は，婦人が比較的かかりやすい病気，または，かかりにくい病気に対して，二つの地区の全体を比較しようという試みについて，すでに比較することは無価値であると結論するのに十分であるかもしれない．

さらに，バーミンガム地区の人口の 35% は 45 歳以上であり，サウスウエスタン地区の人口の 41% が 45 歳以上であった．多くの病気に対する感染のしやすさは年齢により異なるので，この差も同様に，この二地区の比較に対しては十分考慮されねばならない事項である．

つまり，この二地区の総人口形態には，いろいろと重要な相違点が隠されているのである．

上記と同様に，子供の尿中鉛濃度を調査するときも，子供の