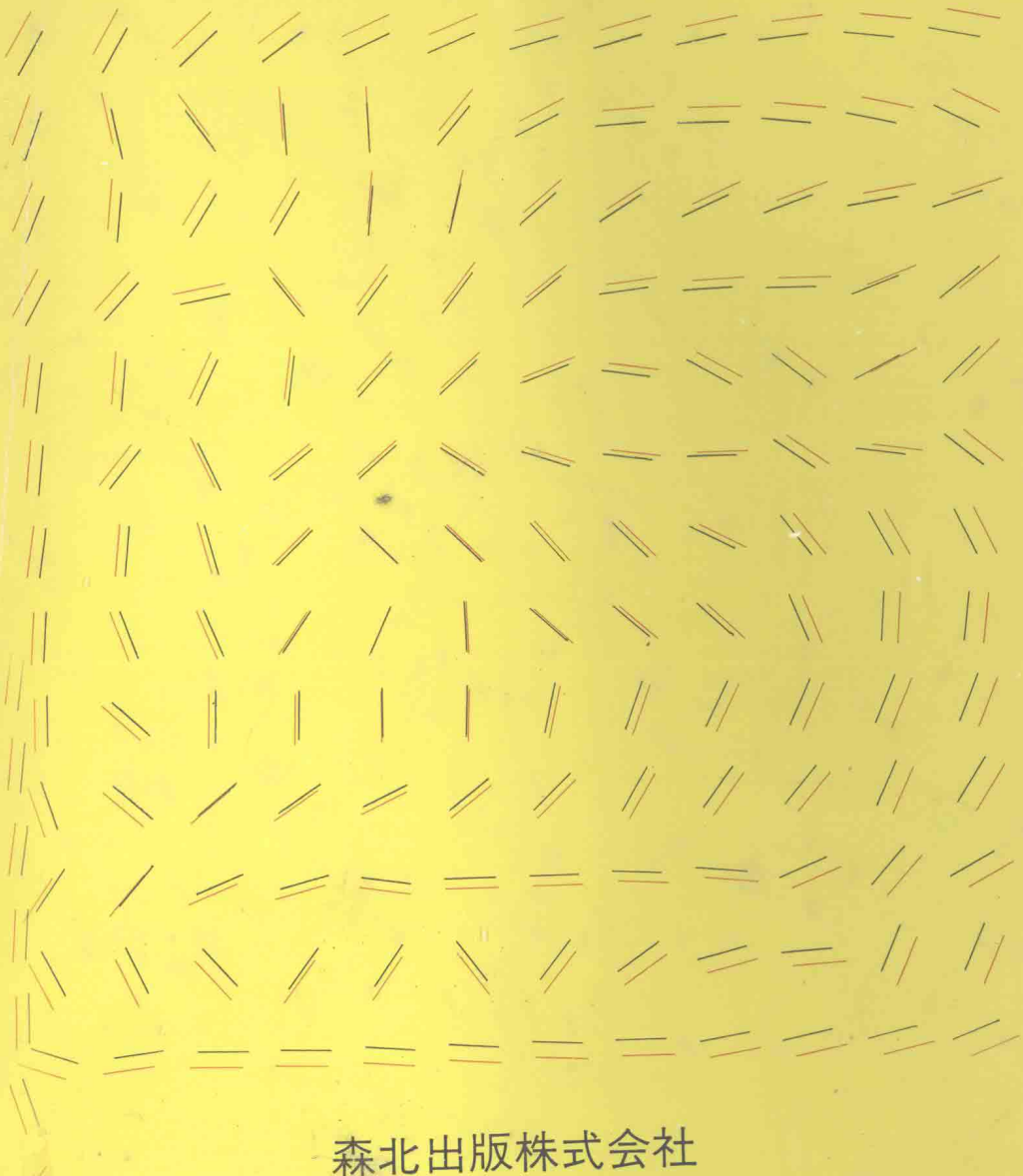
身近なデータによる統計解析入門

脇本和昌 著

岡山大学教授・理学博士



森北出版株式会社

身近なデータによる

統計解析入門

岡山大学教授・理学博士

脇本和昌著

森北出版株式会社

著者略歴

脇本　和昌
1936年　岡山県に生まれる
1959年　岡山大学卒業
現　在　岡山大学教授
理学博士
おもな著書　乱数の知識（森北出版）
標本抽出論入門（槇書店）

身近なデータによる
統計解析入門

1973年 1月 30日　第1版第1刷発行
1980年 8月 20日　第1版第10刷発行

定価はカバー・ケースに表示してあります。

著者との協議により検印は廃止します。

著者　脇本　和昌
発行者　森北　肇
印刷者　小笠原秀雄

発行所　森北出版　株式会社
東京都千代田区富士見 1-4-11
電話 東京 (265) 8341 (代表)
振替 東京 1-34757 郵便番号 102

日本書籍出版協会・自然科学書協会・工学書協会　会員

落丁・乱丁本はお取替えいたします　印刷　秀好堂印刷／製本　石毛製本

1041－0903－8409

Printed in Japan

まえがき

情報化社会といわれる昨今，コンピュータを含めて統計解析の知識は，誰れでも必要欠くことのできないものとなってきています．いきなりむづかしい書物を読み，無味乾燥で難解な数式をいじりまわして，あたかもこれらの知識が備わっているかのように錯覚している人をよく見かけますが，統計解析（コンピュータによるものも含めて）の主な目的は，決してそのような難解な数式の運用ではなくて，われわれの身近かに存在するさまざまな現象やデータを単に人間の直観によるだけでなく，数学的道具を用いてより科学的に把握し，人間生活や社会生活に広く役立てていくことにあります．

この書物は，こうした統計解析の本質を少しでも理解していただければと考えて書かれたもので，統計解析そのものがわれわれの生活の中で，あるいは，われわれをとりまく社会・自然の中でどのように適用され，応用されているかを，身近な，しかも具体的な例をひきながら，その理論的な背景をできるだけやさしく解説したものです．したがって，特に高度の数学的知識が必要であるということでは決してなく，誰れでも知っている常識的知識さえあれば，十分に読みこなせ，しかもその応用技術を修得できうるように配慮してあります．しかしながら，本書を読まれる人たちが，更に，一歩すすんで高度の理論や数式を必要とされるような場合には，この書物を手がかりにどんどん専門的な書物へと読み進まれるよう期待します．簡単で容易だと思われる現象が，よくよく考えてみると非常に内容的に高度な数学的理論を生み出す材料であることも少なくないのです．

この書物では，以上述べたほかに次のような点にも留意しました．

・初心者からよく発せられる質問のうち，主なものについてはそのままの形で本文に採用し，理解を助けるように解答を与えている．

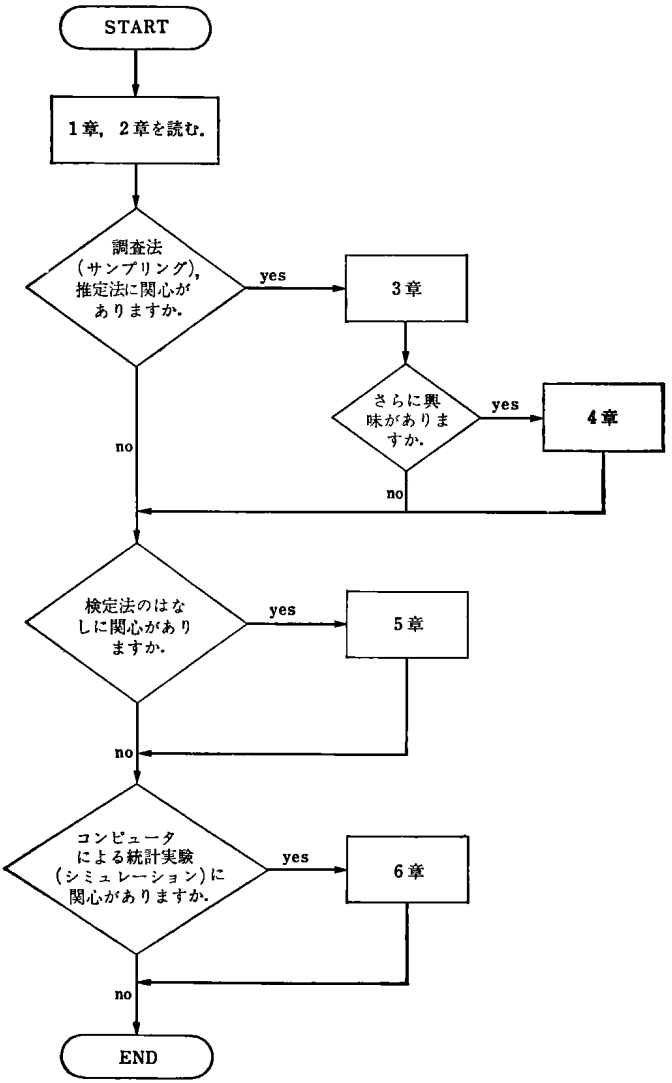
・初心者にはいくらかやっかいな数式，証明などは，本文中ではできるだけ省略したが，特に数学的厳密さを要求する人のために，巻末の補注にまとめて載せ，参考になるよう意を用いた．

最後に，この書物を書くにあたり，いろいろお力添えいただいた岡山理科大学一村稔氏に対し，また終始お世話いただいた柳沢茂八，渡辺侃治氏をはじめ森北出版社の方々に厚くお礼申し上げます．

1973 年 1 月

著　者

読者のためのフローチャート



目　　　次

第5章　統計的仮説検定の考え方と方法

第6章　確率モデルとシミュレーション

——コンピュータの統計的利用法——

第1章　データの記述と尺度化の方法

統計データを集める際に，人間の身長とか体重，製品の長さや重さのように数値化できるものと，意見とか好みのように数値化が困難なものがある．数値化されないとデータとして把握することはむづかしく，抽象的なものをうまく数値化することは，統計データの解析以前の問題として大切なことである．たとえば，眼がどのくらいよく見えるかという抽象的な事柄を，視力表を用いて 1.0 とか 1.2 とか数値化しているが，これは実にうまい方法だと思われる．数値化する場合に肝心なことは，その得られた数値によってその数値のつけられた対象物が，客観的に把握できなければならない．背がどれだけ高いかを，物指しを用いて何 cm にするかといった数値化の問題は，明らかに客観性を持っているが，リンゴがどのくらい好きかということを数値化する場合，どのようにして客観性を持つ数値が得られるか，その方法はきわめてむづかしい．それゆえ，一方ではこの数値化の方法の研究が進められることも大切であるが，この章では 1.6 節で簡単な数値化の例を示す以外はすでに数値化された統計データを取り扱うことにする．

1.1　データの代表値

人間とか物とかまたその集団に対して，統計データが得られたとき，そのデータの代表値として重要なものの一つに，よく使われる**平均値**（mean value）または**総計値**（total value）がある．たとえば，ある学校の一つのクラスの成績の総計点，平均点，ある町の全世帯の総収入，平均収入，ある地域における農作物の農薬の残留量の平均値，総計値などがそれにあたる．いまデータ N 個の値を $x_1, x_2, \cdots\cdots, x_N$ とするとき総計値，平均値はそれぞれ次のように表わされる．

$$(1.1) \qquad T = x_1 + x_2 + \cdots\cdots + x_N = \sum_{i=1}^{N} x_i \quad \text{（総計値）}$$

$$(1.2)\qquad \bar{x}=\frac{1}{N}(x_1+x_2+\cdots\cdots+x_N)=\frac{1}{N}\sum_{i=1}^{N}x_i \quad (\text{平均値})$$

データの数 N が大きいときは，データを適当な階級にまとめて表にして，さらに図に表わすとデータの様子がよくわかる．

例 1　次に示すデータは昭和 46 年度の某県全市における 161 の幼稚園の園児数である．

147	149	44	92	40	28	144	27	500	117	38	41
180	61	40	165	86	22	212	112	190	80	16	51
192	78	28	361	12	15	248	31	153	47	37	32
198	298	48	292	19	42	94	48	164	52	222	90
192	57	38	163	284	97	144	48	262	9	77	126
232	61	39	199	89	130	27	41	365	43	16	140
163	109	107	152	281	93	72	19	459	43	49	80
96	374	264	65	87	51	14	105	271	58	52	
118	66	85	101	37	44	180	66	186	35	11	
103	107	210	281	191	99	175	80	83	68	43	
278	61	116	145	71	182	74	241	218	23	27	
212	77	70	182	50	78	350	83	209	72	35	
248	66	245	267	29	40	45	65	166	19	35	
94	12	193	188	43	59	146	34	243	28	166	

次の表はこのデータを，幅 20 (人)からなる階級にわけて，その階級に属する幼稚園の数（度数）と，**階級値**（階級の中央値）を示したものである．

表 1.1　園児数にもとづく幼稚園の度数と累積度数

階級(人)	階級値	度　数	累積度数	階級(人)	階級値	度　数	累積度数
1～ 20	10.5	11	11	261～280	270.5	5	150
21～ 40	30.5	23	34	281～300	290.5	5	155
41～ 60	50.5	23	57	301～320	310.5	0	155
61～ 80	70.5	21	78	321～340	330.5	0	155
81～100	90.5	14	92	341～360	350.5	1	156
101～120	110.5	10	102	361～380	370.5	3	159
121～140	130.5	3	105	381～400	390.5	0	159
141～160	150.5	8	113	401～420	410.5	0	159
161～180	170.5	9	122	421～440	430.5	0	159
181～200	190.5	11	133	441～460	450.5	1	160
201～220	210.5	5	138	461～480	470.5	0	160
221～240	230.5	2	140	481～500	490.5	1	161
241～260	250.5	5	145	合　計		161	

このようにデータのとる値とその度数が示してある表を**度数分布表**といい，これに対応する柱状のグラフ（図 1.1）を**ヒストグラム**という．また表 1.1 の一番右の列に示すように，その階級までの度数を全部加えたものを**累積度数**といい，累積度数の示してある表を**累積度数分布表**という．

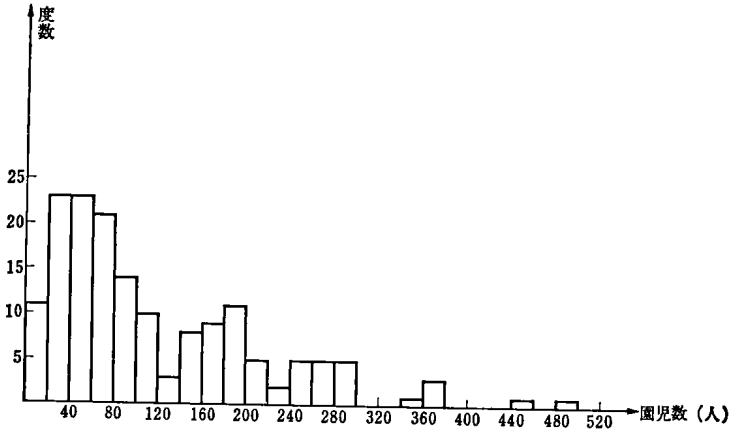


図 1.1 表 1.1 に対応するヒストグラム

さて，この場合の平均値は，次のようにして算出する．園児の数が 1 人～20 人の階級に属する幼稚園の数は 11 あって，その中には園児 12 人の幼稚園もあれば 19 人の幼稚園もある．しかし計算の便宜上 11 の幼稚園全部が階級値の 10.5 人と考える．以下の階級においても同様に考えて平均値を

(*1.3*) $\bar{x}=(10.5\text{人}\times 11+30.5\text{人}\times 23+\cdots\cdots+490.5\text{人}\times 1)\div 161\fallingdotseq 118\text{人}$

とする．

一般に N 個のデータを k 個の階級にわけ，階級値 $m_1, m_2, \cdots\cdots, m_k$ がそれぞれ $f_1, f_2, \cdots\cdots, f_k (f_1+f_2+\cdots\cdots f_k=N)$ の度数をもつとき，平均値を次のように定義する．

平均値を次のように定義する．

$$(1.4)\quad \bar{x}=\frac{1}{N}(m_1\times f_1+m_2\times f_2+\cdots\cdots+m_k\times f_k)$$

$$=\frac{1}{N}\sum_{i=1}^{k}m_i f_i$$

この式（*1.4*）で示す平均値は完全に式（*1.2*）で示す $\bar{x}=\frac{1}{N}(x_1+\cdots\cdots+x_N)$ とは一致しないが非常に近い値を示し，データ処理上さしつかえないことが確かめられる．たとえば，例1で式（*1.2*）による平均値は 117.9 で，式（*1.4*）による平均値は 118.2 であり，その差は非常に小さい．

平均値（または総計値）は，データの代表値のうちで大切なものの一つであるが，どんな場合でも平均値（または総計値）で一つの集団を特徴づけることは好ましくない．たとえば，農作物の被害とか，台風の被害，農薬の残留量などの人体に対する害については，当然，過去に得られた資料の最大値でその対策をたてるべきである．このように最大値もまた大切な代表値の一つである．

データ $x_1, x_2, \cdots\cdots, x_N$ の**最大値**は，資料を次のように大きさの順に並べたとき，

（*1.5*）　$x_{(1)} \leqq x_{(2)} \leqq \cdots\cdots \leqq x_{(N)}$

$x_{(N)}$ のことであり，**最小値**は $x_{(1)}$ のことである．

最大値の用いられる例は，公害データのほかにもいろいろある．走り幅とび，走り高とびなどの競技でも，6回飛んだときの記録を順序づけして，

$$x_{(1)} \leqq x_{(2)} \leqq \cdots\cdots \leqq x_{(6)}$$

とすると，その人の評価として最大値 $x_{(6)}$ が代表値となる．

極言すれば，1回だけ良い記録をだせば，ほかは何でもよいということになる．この評価法が良いか悪いかについては，いろいろ議論があり，むしろ，平均値を用いた方が選手のためによいのではないかという意見もあろう．

式（*1.5*）において，ちょうど中央にくる値のことを**中央値**（median）—または**中位数**とも呼ばれる—という．データ個数 N が奇数であるときは $\frac{N+1}{2}$ 番目の値が中央値で，N が偶数のときは $\frac{N}{2}$ 番目の値と $\frac{N}{2}+1$ 番目の値の算術平均を中央値とする．たとえば $N=5$ のとき

$$x_{(1)} \leqq x_{(2)} \leqq x_{(3)} \leqq x_{(4)} \leqq x_{(5)}$$

において，中央値は $\frac{N+1}{2}=\frac{5+1}{2}=3$ 番目の値 $x_{(3)}$ である．

$N=6$ のとき

$$x_{(1)} \leqq x_{(2)} \leqq x_{(3)} \leqq x_{(4)} \leqq x_{(5)} \leqq x_{(6)}$$

において，中央値は $\frac{1}{2}(x_{(3)}+x_{(4)})$ である．

この中央値が上手に利用されたおもしろい話（例 2）を紹介しよう．

例 2

25 頭の象の群をひきいる大サーカス団があり，あるとき興行のため海を渡ることになり，船に積む関係で 25 頭の象の全体の重さを測る必要が生じた．しかし，考えてみると，25 頭の象全部を 1 頭ずつ測るのは大変な労力である．ところが，このサーカス団の中に非常に機転のきく象使いがいて，次のような方法により，ただ 1 頭の象の重さだけ測って，全体の象の重さを算出した．

まず，25 頭の象を大きさの順に 1 列に並べ，大きい方からも小さい方からも 13 番目にあたる花子という名前の象の重さ（中央値）を測った．そして，その重さを 25 倍したものを全体の象の重さと考えた．花子という名の象の重さは，3.85 トンであったので，25 頭全部の重さを 3.85トン × 25 = 95.25トン と算出した．もちろん，この方法で得られた値は 25 頭の象全部を測定して得られる値とは正確には一致しないが，非常に近い値となることは直観的にもうなづけよう．この誤差は船に積む重量の算出としてはさしつかえないものである．

この算出方法がどの程度正確に全体の象の体重を推定するかを実際の例で示してみよう．あいにく象のデータを持ち合わせていないので，かわり

表 1.2　各クラスの体重　(単位: kg)

クラス A (34 名)	59.0	66.5	60.5	63.5	65.5	56.0	51.5	56.0	57.5	57.5	56.0	56.0
	53.5	54.0	59.5	70.5	54.5	55.0	47.5	57.0	61.0	53.0	52.5	53.0
	57.0	47.0	47.5	63.0	53.0	51.5	48.5	46.5	53.0	57.5		
クラス B (33 名)	50.5	57.0	60.0	55.0	51.0	53.5	51.5	56.5	52.0	60.5	62.0	56.5
	50.0	61.0	57.0	49.0	62.0	51.5	62.0	50.0	54.0	60.0	56.0	56.5
	59.0	67.0	45.5	51.0	60.0	50.0	54.0	51.5	51.5			
クラス C (30 名)	63.5	61.5	66.0	49.5	54.5	63.0	64.0	59.0	52.5	53.0	48.0	62.0
	48.0	55.5	62.0	53.5	57.5	57.0	49.0	74.5	60.0	50.5	58.0	59.5
	54.0	54.0	58.0	56.0	50.5	56.0						
クラス D (32 名)	57.0	53.5	57.5	54.5	57.0	62.5	50.0	50.5	50.0	56.0	60.5	67.0
	54.0	59.0	53.0	58.0	63.0	52.0	52.5	66.5	57.0	56.0	59.0	51.5
	57.0	55.0	53.5	54.5	54.5	54.0	66.0	58.5				

に人間の体重の場合を考えてみた．

表 1.2 に示すデータはある大学の K 学部の 1 年生の 4 つのクラスの男子学生の体重の実測値を示す．

この 4 つのクラスで象使いの算出の方法を用いて，各クラスとも，ただ 1 人だけの体重を測ってクラス全体の体重の総計値を計算してみる．

クラス A の中央値は偶数人なので，大きい方から 17 番目か 18 番目の学生の体重の平均値であるが，そのどちらか 1 人を測ることにする．この場合，どちらの学生の体重を測っても 56.0 kg となるので 56.0 kg × 34 = 1 904.0 kg をこのクラスの体重の総計値とみなす．実際，全部の学生の体重を測って合計すると資料より 1 901.0 kg となり，その差はわずか 3 kg である．

他のクラスについても同様に計算してみると，

クラス B：大きい方から 17 番目の学生の体重（中央値）を測る．その学生の体重は 55.0 kg，よつて 55.0 kg × 33 = 1 815.0 kg を総計値とみなす．

正確な総計値 = 1 824.5 kg

クラス C：大きい方から 15 番目，または 16 番目の学生を測る．15 番目の学生を測ったとき，体重は 57.0 kg，よって 57.0 kg × 30 = 1 710.0 kg を総計値とみなす．16 番目の学生を測ったとき，体重は 56.0 kg，よって 56.0 kg × 30 = 1 680.0 kg を総計値とみなす．

正確な総計値 = 1 710.0 kg

クラス D：大きい方から 16 番目，または 17 番目の学生を測る．どちらを測っても 56.0 kg，よって 56.0 kg × 32 = 1 792.0 kg を総計値とみなす．

正確な総計値 = 1 810.5 kg

これらの例において，中央値はその集団の平均値と非常に近い値であることがわかるであろう．ところが，いつでも中央値は平均値に近い値を示すかというと決してそうではない．例えば，[例 1] の幼稚園の園児数について考えてみると，データから中央値は 85 人と算出されるが，式 (*1.3*) に示すとおり平均値は 118 人で中央値とは大きく異なっている．すなわち，データの構造によって中央値は平均値に近くなったり，離れたりする．中央値と平均値が近い値を示す場合はヒストグラムが平均値をはさんで左右対称に近い場合で，両者の値が離れるのはヒストグラムが左右非対称の場合であることは直観的にもわかるであろう．

また，次のようなところにも中央値は利用されている．

例 3

特定の技術がいつ開発されるか? 1990 年の世界の総人口はいくらか? などの将来予測に関して最近話題になっている“デルファイ手法”というのがある．これは，いずれも甲乙のつけ難い専門家のグループ（n 人としよう）に同じ質問をして，n 人の解答の中央値でもってその質問に対する予測値とする．たとえば，アメリカ合衆国のランド研究所で 150 人の宇宙開発の専門家に対しておこなわれたものであるが,「人間の火星上陸と帰還はいつか?」という質問に対して，その解答の中央値は 1985 年となっている．よって予測値として 1985 年を採用する．この場合，人間の火星上陸が実現する時点（真の値）は，1985 年とはかなり異なるかも知れない．しかし，この 150 人の専門家のグループの半数の人の解答値よりも，中央値のほうが真の値に近いことは容易にわかる．中央値のもつこのような性質を利用したのが“デルファイ手法”である．

度数の最も多いデータの値のことを**最頻値** (mode) または流行値ともいい，流行，需要などのデータにおいて代表値として適当なことが多い．その他にもいろいろな代表値があり，次の例に示すものもその 1 つである．

例 4

体操競技では，1 人の演技に対して，5 人の審査員がそれぞれ点数をつける．そして，その演技者の評価として，5 人のデータのうち最大値と最小値を除いた残り 3 つの値の平均点で．得点を決めている．すなわち

$$x_{(1)} \leqq x_{(2)} \leqq x_{(3)} \leqq x_{(4)} \leqq x_{(5)}$$

のとき代表値として次の値を用いる．

$$\frac{1}{3}(x_{(2)} + x_{(3)} + x_{(4)})$$

この例における代表値は，もちろん 5 人全部の点数を加えて 5 で割った値とは異なっている．この代表値の性質について，理論的にも興味をもつ人もあるが，要するに目的とするところは，1 部の審査員が特定の演技者に対して極端によい点をつけたり，悪い点をつけたりすることを避けるためであろう．この方法を，学校の生徒の 1 年間の成績の評価に使ったらどうであろうか．かりに，1 年間に 6 回テストをしていれば，そのうちの 1

番点数のよいものと最も点数の悪いものを除いて，残りの4つの値を平均する．理由は，実力はありながら風邪をひいたとか，その他の原因で極端に悪い点をとる場合があったり，また，問題にやまをかけていて，たまたまそれがあたり，1度だけ良い点をとる場合も考えられるからである．

このように，データの代表値は，その目的に応じて，平均値・総計値・中央値・最大値などのうちどれを用いるかを考えることが大切であることを強調したい．

1.2 データの“ばらつき”の尺度

——分散，標準偏差——

いま A 商店と B 商店でそれぞれ 10 個の卵を買い，重さを測ったら次のようであった．

A 商店　65 g，60 g，73 g，62 g，55 g，68 g，57 g，67 g，60 g，68 g

B 商店　63 g，65 g，61 g，62 g，63 g，60 g，64 g，66 g，64 g，66 g

このデータから卵の重さの平均値は

A 商店；　63.5 g

B 商店；　63.4 g

となるので両商店とも同じ種類の卵であるというと，とんだ間違いである．よくデータを見ると，A 商店の卵はかなり重さがまちまちで，B 商店の卵は重さがほぼそろっていることがわかる．このように平均値だけでそのデータを表わすと“ばらつき”の異なっているデータを判別することができないので，平均値からの“ばらつき”の度合を次のように考える．

平均値とそれぞれの値の差の2乗の和を考え，それを平均する意味で総個数 10 で割った値をこのデータの分散と呼び，分散の平方根を標準偏差と呼んでいる．この例では卵の重さのばらつきが大きいほど，分散・標準偏差の値は大きくなることは容易にわかろう．それでは A 商店，B 商店のデータをもとに実際に分散・標準偏差を求めてみると次のようになる．簡単のために単位は省略する．

$$\text{A 商店：}\quad \text{分散} = \frac{1}{10}\{(65-63.5)^2+(60-63.5)^2+\cdots\cdots+(68-63.5)^2\}$$

$$= 28.65$$

$$\text{標準偏差} = \sqrt{28.65} = 5.35$$

$$\text{B 商店:}\quad \text{分散} = \frac{1}{10}\{(63-63.4)^2+(65-63.4)^2+\cdots\cdots+(66-63.4)^2\}$$

$$=3.64$$

$$\text{標準偏差}=\sqrt{3.64}=1.91$$

一般的に N 個のデータ

$$x_1, x_2, \cdots\cdots, x_N$$

があるとき，このデータの**分散** (variance)，**標準偏差** (standard deviation) をそれぞれ次のように定義し，記号は σ^2，σ を用いて表わす．σ はシグマと読む．

$$(1.6)\quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \quad \text{(分散)}$$

$$(1.7)\quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \quad \text{(標準偏差)}$$

ただし，$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ とする．

また次のような例を考えてみよう．

2人の薬剤師 A, B がいる．A は熟練者で B は未熟練者である．この2人が 10 g の薬の山を 10 枚の袋に 1 g ずつふりわける仕事をしている．もちろん人間のすることで正確に 1 g ずつ全部の袋にふりわけることは不可能で，どの袋も 1 g から多少ずれているのが普通である．表 1.3, 図 1.2 は一定時間内に A と B が作った 100 袋について度数分布表とヒストグラムを示したものである．このとき両者の分散，標準偏差を求める．

前の卵の例と考え方は同じで，式(*1.4*)で示す平均値と階級値の差の2乗と度数の積の和を考え，それを平均する意味で総数で割った値を，度数分布表で表わされたデータの分散と呼び，その平方根を標準偏差と呼んでいる．A，B両者の分散，標準偏差を求めると次のようになる．