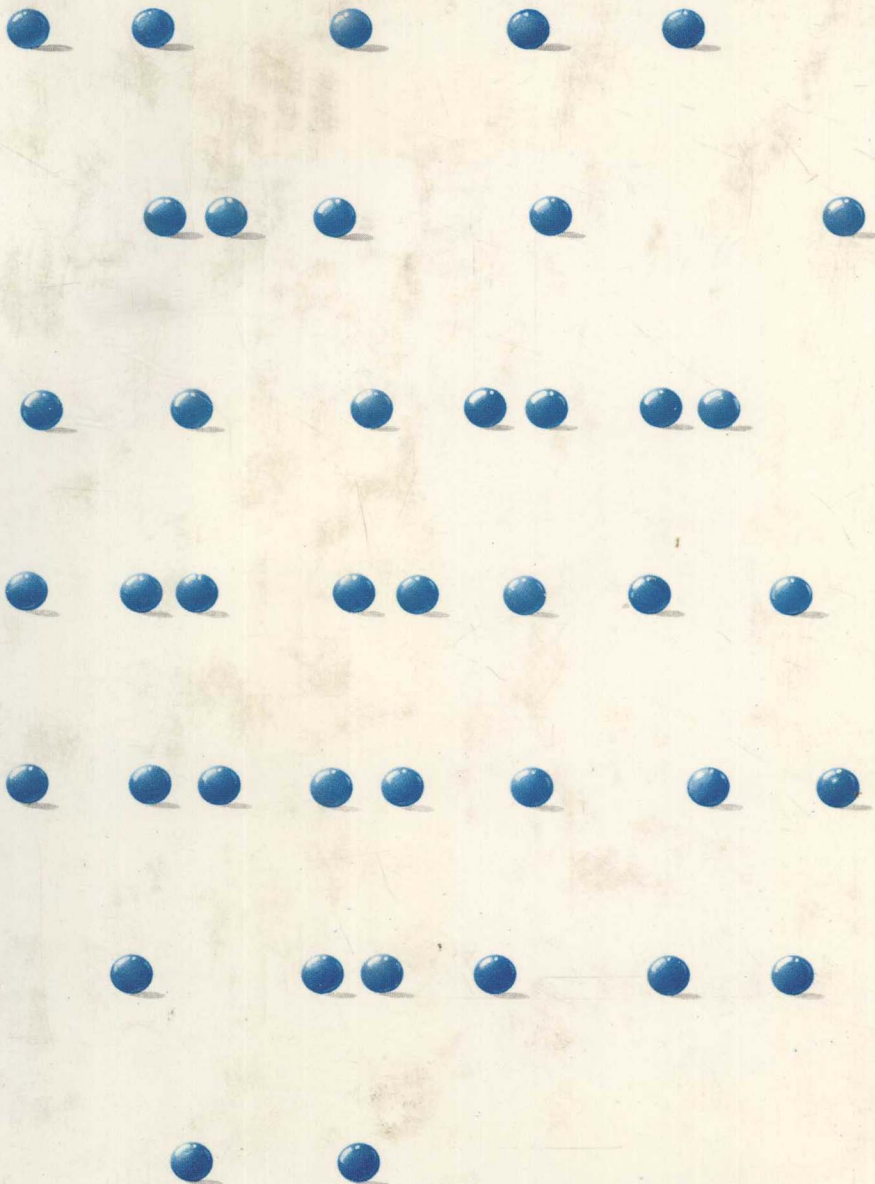


基本統計学

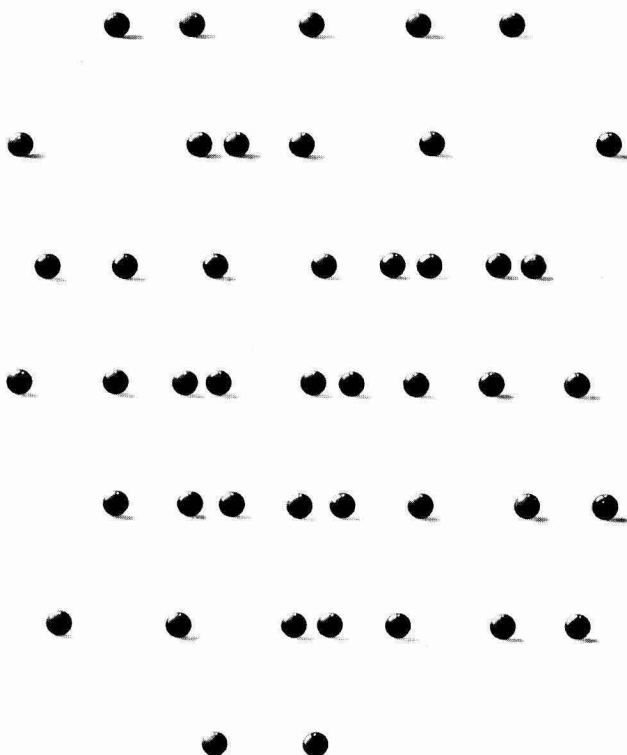
広島大学教授 理学博士 岡本 雅典
統計数理研究所室長 理学博士 鈴木義一郎 共著
統計数理研究所室長 理学博士 杉山 高一



実教出版株式会社

基本統計学

広島大学教授 理学博士 岡本 雅典
統計数理研究所室長 理学博士 鈴木義一郎 共著
統計数理研究所室長 理学博士 杉山 高一



著者略歴

おかもと まさ のり
岡本 雅典

理学博士

昭和 26 年東京理科大学理学部数学科卒業。
昭和 29 年同大学研究科修了。現在、広島
大学経済学部教授。

すず き だ いちろう
鈴木義一郎

理学博士

昭和 35 年東北大学理学部数学科卒業。昭
和 37 年同大学大学院修士課程修了。現在、
文部省統計数理研究所 第三研究部 第二研
究室室長。

すぎ やま たか かず
杉山 高一

理学博士

昭和 38 年東京理科大学理学部数学科卒業。
昭和 40 年同大学大学院修士課程修了。現
在、文部省 統計数理研究所 研究指導普及
室室長。

基 本 統 計 学

NDC 417

1977 年 3 月 15 日 第 1 刷発行

1980 年 4 月 10 日 第 6 刷発行

著 者 岡本 雅典・鈴木義一郎
杉山 高一

発 行 者 宇 野 豊 蔵

印 刷 中 央 印 刷 株 式 会 社

製 本 株式会社 若林製本工場

発 行 所 実 教 出 版 株 式 会 社

東京都千代田区五番町五番地 〒 102

電話東京(263)0111(大代表)振替東京 4-183260

© M. B. OKAMOTO, G. SUZUKI,
T. SUGIYAMA 1977

定価はカバーに表示してあります

3041—2173—3205

ま え が き

本書は、大学1～2年生のための半年あるいは1年間の統計学の講義を念頭において、できる限り軽い数学的「装備」で、統計学の基本的概念を理解できるよう配慮した。昨今の大学生には、少数ではあるが非常に意欲的に勉強しようとする人達と、大多数の“生活エンジョイ派”とがいる。彼らは同時に統計学の講義を受けにくるが、統計学全体の視野と基本的な考え方を知りたがっている点では、両者とも共通している。したがって、本書では基本的概念の理解に重点をおいたが、より意欲的な学生のためにやや高度の理論に属すると思われる内容については、*印を付したり、小さい活字で併記しておいた。これらの箇所を除くなどして適宜配慮を加えて頂けるならば、半年間の講義用テキストとして、看護学校や医療短大等でも利用可能なように工夫した。さらに、極力実際的な例題を用いるよう配慮したので、一般社会人の独習用参考書としてもご利用頂けるだろう。

統計学の面白味は、実際のデータを如何に料理するかにある。したがって、データの料理法のうま味を体得できる機会を、読者自らが作られるよう希望する。これが、統計データの縮約化を第1章にもってきた理由である。確率および確率分布のところでは、種々の分布が実際問題のどのような現象に現われるかについて詳しく述べた。標本分布という基本概念がなかなかわかりにくいと思われるので、電子計算機内で発生させた乱数をもとに、標本平均の標本分布を実験的に求めるプロセスを示しておいた。プログラミングの演習にもなるので、機会があれば各自同様のことを試みるとよい。統計的検定の論理が「石橋をたたいて渡る」方式であるため、統計的方法の効用をかなり消極的にしている面は否定できない。この間の事情について、正しい認識を与えるよう強調した。

回帰分析の考え方、ノンパラメトリックな推論形式についても平易に解説

ii まえがき

し，従来の初年級向け教科書にはない特徴を与えるよう工夫した．実験計画法の章でも，その基本的考え方の説明に重点をおいたが，簡単な2因子実験の解析程度までは理解できるよう配慮した．問題は随所に入れてあるので，こまめに解かれることを希望する．略解を付したが，特に難解と思われる問には*印を付し，解答もやや詳細に記述した．

本書は3人の執筆者の十分な意見調整のもとに書かれた共著である．2ないし3章分を各人が分担執筆し，終りに全般的調整を行ったが，なおアンバランスの個所があるかも知れない．最後に，企画その他に御尽力いただいた実教出版株式会社 橋本正之氏に深く感謝する．

昭和52年1月 著者一同

目 次

1 章 データの縮約化

§ 1. データの縮約化序論	1
§ 2. データの性格とグラフ化	2
2.1 データの性格 (2)	
2.2 データのグラフ化 (7)	
§ 3. データの縮約値	10
3.1 モードとメディアン (10)	
3.2 平均値 (12)	
§ 4. データの散らばり	15
4.1 レインジと四分位偏差 (15)	
4.2 平均偏差と標準偏差 (17)	
§ 5*. 多元データの縮約	18
5.1 多次元データ解析の眼目 (18)	
5.2 散布図と相関表 (20)	
5.3 回帰直線 (21)	
5.4 相関係数 (24)	
5.5 分割表と関連係数 (27)	

2 章 確 率

§ 1. 確率の概念	29
1.1 経験的な確率の定義 (29)	
1.2 標本空間と確率の定義 (30)	
1.3 和事象と積事象 (31)	
1.4* 確率の公理的定義 (32)	
§ 2. 確率の性質	33

§3. 条件つき確率と乗法公式	34
§4. 事象の独立	38
§5. ベイズの定理	39

3 章 確率変数と確率分布

§1. 確率変数	43
§2. 離散型分布	44
2.1 二項分布 (44)	
2.2 超幾何分布 (47)	
2.3 ポアソン分布 (49)	
2.4 幾何分布 (51)	
2.5 負の二項分布 (52)	
§3. 連続型分布	53
3.1 分布関数と確率密度関数 (53)	
3.2 正規分布 (55)	
3.3 一様分布 (56)	
3.4 指数分布 (57)	
§4. 期待値と分散	58
§5*. 同時分布	61
§6*. 和の分布	64
§7*. 確率変数列の極限法則	66

4 章 標 本 分 布

§1. 全数調査と標本調査	71
§2. 標本分布	74
2.1 乱数の発生と標本実験 (74)	
2.2 標本分布 (76)	
2.3 χ^2 -分布 (78)	

§ 3*. t -分布と F -分布	80
3.1 t -分布 (80)	
3.2 F -分布 (81)	

5 章 推 定

§ 1. 統計的推論の考え方	83
§ 2. 的中率と推定幅	84
§ 3. 信頼区間	86
3.1 平均に対する信頼区間 (86)	
3.2 比率に対する信頼区間 (88)	
§ 4. 標本平均に基づく母平均の区間推定	89
4.1 正規分布の平均に対する区間推定 (89)	
4.2 出現比率に対する区間推定 (90)	
4.3 信頼区間を求める一般原理 (91)	
§ 5*. 母集団特性量と標本特性量	92
§ 6*. 不偏推定と一致推定	95
6.1 不偏推定 (95)	
6.2 推定量の効率と一致性 (96)	
§ 7*. 十分統計量	98
7.1 十分統計量とは (98)	
7.2 ハルモス-サヴェッジの判定条件 (98)	
7.3 十分統計量による不偏推定の改良 (99)	
§ 8*. 最尤法	100
8.1 最尤推定とは (100)	
8.2 最尤推定量と十分統計量 (101)	

6 章 検 定

§ 1. 統計的仮説検定の考え方	103
------------------	-----

1.1	仮説検定とは (103)	
1.2	帰無仮説を棄却する基準 (105)	
1.3	検定方式の棄却域と検定力 (106)	
§ 2.	標本平均に基づく仮説検定	107
2.1	分散既知の場合の平均の検定 (107)	
2.2	分散未知の場合の平均の検定 (108)	
2.3	2組の平均の差異の検定 (110)	
§ 3.	比率の検定	111
§ 4*	一般の場合の検定方式	114
§ 5*	一様最強力検定	115
5.1	最強力検定とは (115)	
5.2	ネイマン-ピアソンの基本定理 (116)	
5.3	基本定理の応用 (117)	
§ 6.	カイ2乗検定	119
6.1	多項分布の検定 (119)	
6.2	適合度検定 (120)	
6.3	独立性の検定 (120)	

7 章 正規分布を仮定できない場合の検定法

§ 1.	分布の位置のずれに敏感な検定	122
1.1	ウイルコクソンの検定 (123)	
1.2	コルモゴロフ-スミルノフ検定 (126)	
§ 2.	分布の散らばりの違いに敏感な検定	129
2.1	アンサリー-ブラットレイ検定 (129)	
2.2	カイパー検定 (132)	
§ 3.	ラページ検定	134

8章 回 帰 分 析

§1. 直線回帰	137
§2. 回帰モデルについての条件	140
§3*. 重回帰モデル	141
§4*. 重相関係数	144
§5*. 変数選択の問題	146

9章 実 験 計 画 法

§1. 実験計画法の考え方	150
1.1 実験計画法とは (150)	
1.2 実験計画法における三原則 (151)	
§2. 完全無作為化法	153
§3*. 乱塊法	158
§4*. 要因実験	161
4.1 要因効果 (161)	
4.2 母数模型と変量模型 (163)	

おわりに	166
参考文献	168
解 答	170
付 表	184
索 引	194

1 章

データの縮約化

§ 1. データの縮約化序論

データを集める。これは一体何のためだろう。データ（標本）に基づき、的確な判断を行いたいからに他ならない。採択可能な選択肢の中から、最良と考えられる判断を1つ選び出す、その際に利用されるのがデータである。無目的に集められるデータは、役に立たないと考えたほうが正しい。少なくともデータを集めようとする段階では、当面の“問題”が何であるかについて、明確な認識がなされていなければならない。そして、当該問題の解決に役立つようなデータを集めるよう心掛けるべきである。

平均の値を求める。これは一体何のためだろう。データのもつ情報を、ひとまとめにして表現したいからに他ならない。つまり、ある的確な判断を行うためにデータが集められる。生のデータそのままでは繁雑にすぎること、判断の便に供しやすいスタイルにデータが縮約される。そのような縮約値の代表格のものとして、平均値が登場する。したがって、平均の意味がわかるということは、当該データのおよその情報を読み取ることができるということである。一般に“データを読む”ということは、そう画一的にできるものではないから、平均の意味を理解することはそうやさしいことでもない。ところが平均の値を導出するのは、大抵の場合同じ手続きを踏めばよい。すなわち、“データの総和をデータの個数で割り算する”という演算である。これは、誰がやっても簡単にできることである。このように平均は親しみやすいだけに、ときには思わぬ誤解を生じさせる危険もあることを忘れてはならない。

標準偏差を求めたりする。これは一体何のためだろう。平均がデータのよい縮約値であるとはいえ、それだけでデータの全貌を的確に代表し得るもの

2 1章 データの縮約化

とは限らない。一般にデータは、平均値のまわりで、ある種の広がりをもってばらついているものである。そのようなばらつきの程度を示すものさし、それが標準偏差ということになる。通常我々が手にする多くのデータでは、平均値からプラス・マイナス標準偏差の範囲にデータの66%が、さらに2倍の幅をとった範囲では95%程度のデータが入っている、といった“経験則”が成立している。したがって、標準偏差の小さなデータの平均値は、データをよく代表し得るものと考えられるが、そうでない場合では、平均がデータの情報をよく吸収しているとはいいきれない。結局、平均値に標準偏差の値も付与しておけば、データのもつ情報はかなり保存されるものと考えられる。

データを図式化して表現することもある。これは一体何のためだろう。データのもつ情報を、視覚的に把えてみたいからに他ならない。ヒストグラム、散布図などから得られる情報は、一片の数値などでは表現し得ない部分を含んでいる。ただ、図式化の方法が必ずしも一通りには決まらないだけに、うまく図式化を行わないと、かえって誤った印象を与えることにもなりかねない。グラフとは、データと目盛の関数であることを留意しなければならない。

§2. データの性格とグラフ化

2.1 データの性格

§1で述べたように、生のデータを適当に縮約すれば、何がしかの情報を失うことになる。情報損失を最小限に食い止めるためにも、当該データの性格を熟知しておくことが重要である。

データというと、すぐ数値のようなものを考えがちになるが、そうでないタイプのものも結構ある。たとえば、交通事故を起こした原因を集積したデータ、公害の発生源を集めたデータ（表1.1）などは、発生件数という数値が対応はしていても、データの本質的な部分ではない。第一、この種のデータの平均を求めてみよといわれても、求めようがないことに気付かれるであろう。

表 1.1 公害発生種類別苦情件数

発生種類	昭和 45 年度	昭和 49 年度
大 気 汚 染	12 911 (20.4)	12 145 (15.4)
水 質 汚 濁	8 913 (14.1)	14 496 (18.3)
騒音・振動	22 568 (35.6)	24 195 (30.6)
悪 臭	14 997 (23.6)	17 140 (21.7)
そ の 他	4 044 (6.3)	11 039 (14.0)
総 数	63 433	79 015

(総理府公害等調査委員会)

表 1.2 結婚式や葬式は盛大にやるべきか

調 査 年 度	よ く な い	し か た が な い	身 分 相 応 に	盛 大 に	そ の 他	計 % 実数
昭和 28 年	31	5	48	8	8	100 (2 254)
昭和 33 年	48	5	38	6	3	100 (1 449)
昭和 38 年	35	4	52	6	3	100 (2 698)
昭和 43 年	37	4	47	7	5	100 (3 033)
昭和 48 年	30	9	53	5	3	100 (3 055)

(統計数理研究所・日本人の国民性調査)

表 1.2 の結婚式や葬式に対する考え方を集計したデータの場合も、ほぼ同じ性格であるが、前のデータと少し異なっている。それは、反応カテゴリーの間に、ある一定の順序関係が存在していることである。いずれのタイプも、**質的データ**とか、**カテゴリカルデータ**とか呼ばれているものであるが、特に表 1.1 のデータのようなタイプを名

目尺度のデータ、表 1.2 のようにカテゴリー間に順序関係があるタイプを順序尺度のデータと名づける。

出生児数ごとに分類した夫婦組数の分布(表 1.3)の場合はどうか。分類基準となっているカテゴリーの間に順序があるばかりでなく、0 子、1 子、2 子、……という数値が等間隔になっていることが、前のデータのタイプとは

表 1.3 出生児数の割合

出生児数	夫婦組数	%
0 子	1 093	11.7
1	1 891	20.2
2	3 884	41.5
3	1 745	18.7
4	468	5.0
5	135	1.4
6 子以上	54	0.6
不 明	85	0.9
合 計	9 355	100

(人口問題研究所・第 6 次出生力調査)

4 1章 データの縮約化

異なっている。観測された値が等間隔の数値をとることから、観測値同志を加えた値が意味をもつ。足し算が許されるということから、平均値が意味をもつ。実際、表1.3のデータの平均は1.9のように算出される。これより、夫婦は平均して2人弱の子供をもっていることがわかる。ところで5段階評価による成績分布について、5が何個あり、4は何個といった表現を用いるが、これも同種のデータと考えてよいか。昔だったら、甲乙丙とか、優良可のような記号を用いたりしていたことからわかるように、この種のデータは順序尺度のデータではあっても、数値間の間隔が一定とは考えるべきでない。数値間の差が意味をもち、したがって、平均値を求めることが許されるようなタイプを **区間尺度** のデータと名づける。

表1.4や表1.5のような、東京での日最高気温のデータや降雨量のデータなども、やはり区間尺度のデータである。特に降雨量のデータの場合には、観測値間の差が意味をもつばかりか割合を考えることもできる。たとえば、1日で20mm降った日は10mmしか降らなかった日より2倍の雨が降ったことがわかる。ところが、セ氏20度の日はセ氏10度の日より2倍暖かいということにはならないから、気温のようなデータで数値間の割合を考えてみても意味がない。両者の違いは、降雨量の場合の“0”は絶対的な意味をもち、負の観測値というものとは考えられないが、気温の場合の“0”は相対的なもので負の値も観測されるという点にある。0の意味が絶対的で、観測値間での割合も意味をもつとき、**比率尺度** のデータと呼ぶ。比率尺度のデータについては、幾何平均のようなものも求められる。

質的なデータではなく、数値により記録されるタイプのものは **量的データ** と呼ばれる。ただし5段階評価での成績分布のように、数値が分類のための記号の役割しか果たしていない場合は、量的データといわない。量的データでも、出生児数の分布のように離散的な値しかとらないタイプを **計数データ**、気温とか降雨量のデータのように連続的な値をとり得るタイプを **計量データ** と名づける。最高気温や降雨量のデータをみると、観測値が適当に丸められているためにとんでもない端数はできていないが、本質的には連続的にどんな値でもとり得るものである。計量データ（カテゴリー数が多い計数デー

表 1.4 東京地区の日最高気温

	50年 1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
1日	5.6	10.6	9.0	14.9	24.1	24.6	23.3	32.0	33.2	22.5	17.1	10.8
2	4.2	7.7	8.8	13.4	22.3	25.0	25.0	32.8	32.9	21.7	17.1	11.1
3	8.6	8.2	10.3	15.8	19.8	24.0	26.2	33.5	31.6	18.8	17.3	12.8
4	6.3	5.2	10.8	16.9	22.7	26.0	21.5	33.2	32.3	26.4	17.3	15.3
5	5.8	13.2	7.8	17.9	28.3	22.5	23.9	35.5	30.4	22.8	16.7	14.0
6	8.7	10.3	9.8	23.0	17.2	27.7	27.4	34.9	32.5	27.9	12.9	12.6
7	10.4	10.5	12.2	21.0	20.6	22.2	31.5	27.3	32.2	23.6	23.4	13.5
8	6.0	10.0	14.3	15.8	22.1	26.1	31.6	30.1	31.9	24.1	20.3	10.6
9	12.4	9.2	18.1	21.6	22.9	27.0	28.7	28.9	31.7	25.6	18.0	6.4
10	10.3	8.5	11.1	22.0	21.7	26.9	24.5	28.2	30.6	25.0	15.0	9.3
11	9.1	8.6	15.3	20.8	21.9	19.9	31.1	30.6	31.2	22.1	13.4	7.5
12	9.0	10.7	13.6	18.1	23.1	22.1	28.0	31.5	32.2	19.4	16.1	11.0
13	8.0	10.3	12.1	17.4	24.3	23.7	22.1	32.9	32.6	17.9	15.9	9.1
14	10.0	11.0	9.4	19.0	22.4	22.4	26.3	32.9	28.3	20.9	15.8	7.8
15	11.5	8.3	8.2	22.0	25.0	23.0	31.3	33.4	30.0	20.3	16.8	10.6
16	7.8	6.6	10.5	19.4	19.9	26.4	31.2	31.5	34.1	16.9	25.1	12.0
17	8.4	6.0	10.6	18.1	21.6	27.9	31.9	31.3	33.2	17.9	21.2	10.2
18	9.2	10.9	13.1	17.8	22.6	28.0	35.6	32.8	30.0	15.8	18.6	10.1
19	8.2	9.7	10.9	18.5	19.8	29.5	29.7	31.3	30.5	18.0	14.1	11.0
20	9.3	9.0	9.0	21.4	22.1	25.3	29.3	33.5	31.4	23.3	17.5	10.0
21	9.1	5.2	18.6	16.0	22.9	26.8	30.4	34.3	32.2	23.2	10.9	9.0
22	7.0	8.2	17.8	11.1	22.4	22.3	29.6	29.7	31.5	20.8	13.0	12.9
23	10.2	8.9	14.5	11.6	24.4	27.6	31.5	29.2	23.6	19.4	12.9	11.5
24	10.4	10.1	9.2	11.1	24.9	25.5	32.3	32.8	21.3	17.9	13.1	11.8
25	17.6	15.0	10.3	14.8	22.3	21.4	32.2	29.1	25.5	16.3	14.2	10.4
26	7.5	10.2	8.2	21.4	21.4	27.2	32.3	27.2	27.3	17.6	15.9	11.4
27	6.5	7.3	12.4	21.7	22.5	27.1	31.5	30.9	25.0	19.0	12.0	9.8
28	11.1	10.4	14.8	23.7	24.8	22.4	31.3	31.4	26.0	15.1	17.5	9.5
29	7.1	—	17.4	22.4	26.1	26.3	32.4	31.5	21.1	24.7	15.9	9.0
30	8.4	—	18.1	24.0	26.8	24.1	33.8	31.6	21.3	19.1	14.3	11.2
31	9.3	—	16.5	—	21.3	—	34.0	32.1	—	11.9	—	9.7
平均	8.8	9.3	12.3	18.4	22.7	25.0	29.4	31.5	29.6	20.5	16.2	10.7
平年値	9.4	9.9	12.8	18.4	22.6	25.3	29.2	31.0	27.0	21.1	16.4	11.8

(日本気象協会)

表 1.5 東京地区の降雨量 (単位ミリメートル)

	50年 1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
1日	4.5					1.5				0.0	0.0	
2	4.5	0.0				0.0				1.0	0.0	
3					17.0	1.5	2.5			12.5		
4		18.5			12.0		93.0			0.0		
5			10.0	0.0		8.0	7.5	0.0	3.0	63.0	0.0	5.0
6			12.0	1.0	3.0	0.5	37.5	2.0			33.5	1.0
7		32.5				27.5		0.5		0.0	88.5	6.5
8	10.0	0.0		18.5	0.0	0.0	0.5	0.0	0.0	42.5		19.5
9	0.0				0.0					0.0	0.0	19.5
10			10.0	0.0	0.0	1.0	9.0	0.0		0.0		0.0
11				0.0	0.0	9.5	0.0			0.0	0.0	0.0
12						1.0	5.5			9.5	0.0	0.0
13			0.5	0.5			4.0			1.5		
14		0.0	0.0	2.0	0.0		0.0				5.0	
15		8.0	1.0		0.0			0.0	1.5		58.5	
16	20.0				24.5			0.5			0.5	
17				0.0	18.5		0.0	15.5		0.5		
18				0.0		0.5		3.5	0.0	59.0	0.0	
19					63.0	0.5		2.0	0.0	29.0	29.5	
20		0.0	9.0		10.5	5.5			0.0	0.0		
21		16.0	62.5	54.0		1.5	16.0				3.5	
22	28.0			30.5		29.5	0.0	0.0	0.0		19.0	0.0
23					0.0	2.5	0.0	2.5	32.0		2.5	
24			0.0	11.0		1.5	0.0		31.5	7.0	0.0	
25			0.0	34.5		4.5		0.0		1.0		
26	0.0		0.0	0.5	0.0	5.0						
27	0.5	0.0		0.0	2.5	0.0	0.0				4.0	0.0
28	0.0					4.5			0.0	1.0	0.5	
29		—		1.5		0.0			31.5	0.5		0.0
30		—	3.0	7.5		0.5				18.0		0.0
31		—	0.0	—	0.0	—			—	27.0	—	
計	67.5	75.0	108.0	161.5	151.0	106.5	175.5	26.5	99.5	273.0	245.0	51.5
平年値	48.8	64.6	98.0	121.9	145.4	192.3	140.0	152.8	182.1	203.3	96.2	57.9

(注: 0.0は記録計には出ない微小雨量で, 無記入が雨なしの日)
(日本気象協会)

タ)を、十数個程度のクラスにまとめて扱おうとすると、どのようなクラス分けが適当であるかを考える必要がでてくる。この問題は、考えようによってはとてつもなく難しい問題の1つなのである。

2.2 データのグラフ化

データをグラフ化し、視覚を援用した情報の吸収法が、有効なことが多い。ただ、図は式より親しみやすいと考えられているせいか、安易にグラフ化されてしまうために、時として思わぬ誤解を生むような図が表示される場合もある。

図1.1は、表1.1の公害発生種類別の昭和49年度のデータを、円グラフと、棒グラフで図示している。名目尺度のデータであるから、件数の多いも

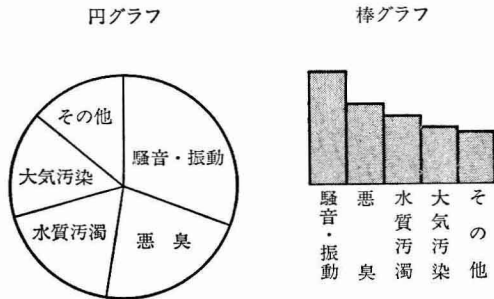


図 1.1 公害発生種類別苦情件数 (昭和49年)

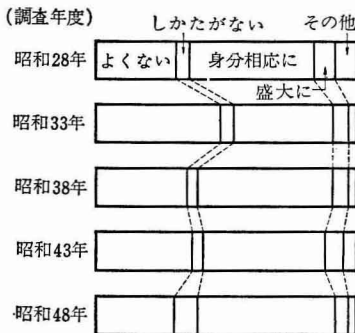


図 1.2 “結婚式や葬式は盛大に”の帯グラフ

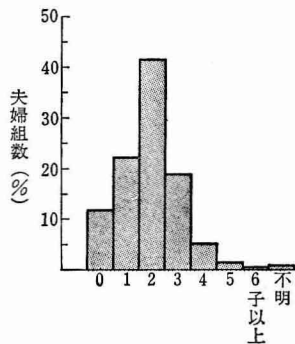


図 1.3 出生児数の分布