

社会科学系のための

統計学概論

福原文雄著

培風館

社会科学系のための

統計学概論

福原文雄著

培風館

福 原 文 雄 略 歴
ふく ほら ぶん お

1970年 東京大学大学院経済学研究科
博士課程修了

現 職 青山学院大学経済学部教授

主 要 著 書

計量経済分析の方法(マグローヒル好学社, 1981)

© 福原文雄 1984

昭和59年1月20日 初版発行

昭和59年4月25日 初版第2刷発行

社会科学系のための
統計学概論

著 者 福原文雄
発行者 山本健二

発 行 所 株式会社 培風館

東京都千代田区九段南 4-3-12・郵便番号 102
電話 (03) 262-5256(代表)・振替東京 4-44725

定 価 ￥1380.

中央印刷・土曜製本

ISBN4-563-00843-5

C3033 ￥1380E

本書の内容の一部あるいは全部を無断で複製すると、著作権
および出版権侵害となることがありますので御注意ください。

は し が き

80年代にはいって社会全体が情報化への傾斜を急速に強めている。社会のあらゆる分野で大量の情報が蓄積され、そして意志決定のほとんどあらゆる局面で、この情報のより効果的な活用が求められている。近時、さまざまな計画や管理にかかわる意志決定に、科学的基礎を与える方法論の体系として統計学の重要性が一そう広く認識されてきているのは、けだし当然の成行きであるといえよう。

このような趨勢は、固有科学と統計学が直接結びつきたいわゆる計量科学——計量心理学、計量経済学、計量社会学、計量医学、計量行動学、計量政治学、計量国語学、等——の急速な進歩によっても窺い知られる。

疑いもなく統計学は、いまや社会科学系の人たちにとっても不可欠の学問となってきた。社会科学系学部に統計学が必修科目としておかれてきているのはこのためである。

このような社会的背景の下で優れた統計学の本が数多く出されている。正確かつ簡潔に統計学の理論が述べられているそれらの本は、しかし数学的な訓練をあまりつんでいない社会科学系の人たちにとっては、必ずしも適当ではないようにみえる。一方、文科系の人たちのために書かれた、統計的なものの考え方の理解を主目的とした叙述中心の本も少なくないが、しかしこれもまた社会科学系の人が使うのに必ずしも適当とは思われない。ことばにたよって統計的なものの考え方を理解したとしても、それではより上の統計学あるいはその応用分野へ進むことは困難ではあるまいか。

本書はこのような点にふかく考慮して書かれたものである。以下に本書の特色といったものを2, 3挙げてみよう。

まず本書では、記号の正しい使用にとくに留意した。統計学を初めて学ぶ人は、まず統計学によく使われる基本的な記号を正確にとらえ、これに慣れるよう努めるべきである。不慣れな記号にはじめて出合うとき、誰も多少の抵抗を感じるに違いない。しかしそれは最初のうちだけである。やがて記号の正しい使用がいかに理解を正確かつ容易ならしめるかを知るであろう。

読者が、たとえば統計量と統計値が異なった記号により区別して扱われていることの意味をふかく理解し得たならば、すでにより上の学習に進む準備ができたと考えてよいくらいである。

なお、記号の正しい使用は、数学的な難解さを意味するものではまったくない。数学に関するかぎり、高校初年度程度の力があれば本書を読みこなすには十分であろう。

つぎに、理論的に理解を容易ならしめるため本文中に適当な例題を多く配した。読者はその解を参考にしてあとの練習問題に取組んでほしい。著者の経験によれば、与えられた式にデータをあてはめるだけといった問題をいくら多く試みても、時間を費やす割にはそれほど理論的理解を深めることに役立たないようである。そこで本書では、適当な問題を適当な量だけしぼって練習問題とした。その中には、単に数値的解を求めるというのではなく、統計的なものの考え方を養うに役だつような問題も選ばれている。

本書は上に述べた意図にそうよう、統計学のもっとも基礎的理論をできるだけ体系的に示したものであるが、最終章の回帰分析における回帰モデルだけは唯一の例外をなしている。

周知のように、回帰モデルはとくに社会科学系の人たちにとってきわめて応用範囲の広い統計的手法のひとつである。回帰モデルの統計推論は、主として計量経済学の中で発展させられてきたものであるが、まさにその事が示すように、それはひとつの応用的理論である。回帰モデルは、統計学の基礎的理論体系に包括される回帰関数とは、理論的に別のものであるが、その形式的な類似性の故に両者はよく混同して論じられているようなので、本書の最後の部分をかりてその対比を試みたものである。

ともあれ本書が、主として社会科学系の学生諸君の統計学の教科書として、

あるいは従事している仕事の関係から統計学の基礎を身につけたいと考えている人たちの参考書として、役立ちうるならば幸いである。

本書はきびしい時間的制約の中でまとめられたものであるが、そんなことより何より著者自身の非才と不勉強の故に、至らない点が多いのではないかと恐れている。本書に目を通される専門の先生方、あるいは統計学の基礎的理解を求めて本書をお読み下さる一般の方々や学生諸君から、ご教示やご叱正を頂けるならば、それは著者にとって大へん有難いことである。

本書の数表の作成について、青山学院大学情報科学研究センター助手、本郷茂氏からご援助を頂いた。また例題の計算の一部を青山学院大学理工学部経営工学科助手、高梨公孝氏に負っている。記して感謝の意を表わしたい。

本書が、原稿の提出の遅れにもかかわらず、非常にはやく陽の目を見ることができたのは、偏に培風館編集部、松本和宣氏のお力添えによるものであり、心からお礼申し上げる次第である。

1983年11月

著 者

目 次

1 序 論	1
2 度数分布	5
2.1 要因, 変数	5
2.2 度数分布	6
2.3 累積度数分布	12
2.4 結合度数分布と周辺度数分布	13
3 分布の特性を表わす代表値	16
3.1 分布の位置の代表値	16
(i) 平均(値)	16
(ii) メディアン(中央値)	19
(iii) モード(最頻値)	20
(iv) 幾何平均	20
3.2 分布のちらばりの代表値	20
(i) 平均偏差	21
(ii) 標準偏差, 分散	21
(iii) レインジ	23
3.3 規 準 化	23
3.4 共 分 散	24
3.5 相 関 係 数	28
補註 1 和記号	30
4 確 率	32
4.1 試行と事象	32
4.2 確率の定義	33
(i) 算術的確率	33

(ii) 頻度的確率	34
(iii) 公理主義的確率	34
4.3 条件付確率と事象の独立	36
4.4 ベイズの定理	39
補註 2 集合と集合の演算	40
補註 3 順列と組合せ	43
5 確率分布	46
5.1 確率変数と確率分布	46
(i) 離散確率変数と確率分布	46
(ii) 連続確率変数と確率分布	48
5.2 期待値	50
(i) 離散確率変数の期待値	50
(ii) 連続確率変数の期待値	51
5.3 結合分布と周辺分布	51
5.4 条件付分布と回帰関数	55
補註 4 期待値の記号 E	55
6 主要な確率分布	58
6.1 離散型一様分布と乱数	58
6.2 ベルヌイ分布	59
6.3 二項分布	61
6.4 ポアソン分布	64
6.5 (連続型)一様分布(矩形分布)	67
6.6 正規分布	69
6.7 標準正規分布	71
(i) 標準正規分布とその応用	71
(ii) チェビシエフの不等式との比較	73
6.8 二項分布の正規近似	74
補註 5 チェビシエフの不等式の証明	76
7 標本分布	79
7.1 母集団, 標本	79
7.2 標本分布	82
7.3 標本平均の分布	83
7.4 標本分布としての二項分布	87
7.5 極限分布	88
7.6 標本比率の分布	93
補註 6 無作為標本の抽出について	94

8 統計的推定	97
8.1 母数の推定	97
8.2 推定量の特性	98
(i) 不偏性	98
(ii) 一致性	99
(iii) 有効性	100
(iv) 充足性(十分性)	101
8.3 最 尤 法	101
8.4 区 間 推 定	105
(i) 区間推定の意義	105
(ii) 母平均の信頼区間——分散既知の場合	106
(iii) 母平均の信頼区間——分散未知の場合	108
(iv) 母集団比率の区間推定	110
9 統計的検定	114
9.1 ネイマン・ピアソンの仮説検定	114
(i) 帰無仮説と対立仮説	114
(ii) 第1種過誤と第2種過誤	115
(iii) 棄却域の定め方	118
9.2 平均に関する片側検定	120
(i) 母分散既知の場合——右片側検定	120
(ii) 母分散既知の場合——左片側検定	122
(iii) 母分散未知の場合—— t -分布による検定	123
9.3 作用特性関数と検定力関数	124
9.4 平均に関する両側検定	128
(i) 母分散既知の場合の尤度比検定	128
(ii) 母分散未知の場合の尤度比検定	130
9.5 比率に関する検定	131
9.6 平均および比率の差に関する検定	135
(i) 標本平均の差の分布	136
(ii) 平均の差の片側検定	136
(iii) 平均の差の両側検定	138
(iv) 比率の差の検定	138
10 回 帰 分 析	142
——回帰関数と回帰モデル——	
10.1 線形回帰関数	142
(i) 回帰関数	142
(ii) 重相関係数	143
(iii) 推定	144

10.2	線形回帰モデル	145
(i)	線形回帰モデル	145
(ii)	係数母数の推定	146
(iii)	分布母数の推定と係数母数の推定量の分散	148
(iv)	決定係数	149
(v)	例題	150
10.3	係数母数 β の区間推定と仮説検定	152
(i)	β の区間推定	153
(ii)	β に関する仮説検定	153
(iii)	例題	154
付 表		157
1.	ポアソン分布表	158
2.	標準正規分布表	162
3.	t -分布表	166
4.	二項係数表	167
5.	一様乱数表	168
演習問題略解		171
索 引		185

1

序 章

科学はすべて何らかの客観的規準の上に立っている。それでは確率論(という科学)によって立つ規準は何であろうか。

インダス文明を築いた古代インドの住民は紀元前3千年頃すでに正6面体のサイコロを用いて賭に興じ、また側面に1から4までの数を表示した正4角柱をころがして神意を問うていたという。その数の現われ方の中にひそむ神秘性をかれらは経験的に知っていたのである。しかしなぜそのようなことが存在するのかは、かれらはもちろん現在のわれわれも知ってはいない。ともあれサイコロやカードの数の現われ方を支配するある法則——偶然の法則あるいは確率法則の存在を意識し、それを踏まえて問題に接近しはじめたのは16世紀から17世紀にかけてのことである。

16世紀中葉自ら賭博師でもあったカルダン(G. Cardan)がサイコロの背後にある「偶然」について議論を展開している。つづいて17世紀のはじめにガリレオ(G. Galileo)もまたあるサイコロ遊びの問題を正確な分析によって解いている。古代インドの住民が神意として畏懼した神秘性はかくして偶然の法則として明確に把握され、そして同じ世紀のフランスで2人の科学者、パスカル(B. Pascal)とフェルマ(P. Fermat)によって偶然の法則に関する科学としてまとめられる。このいわゆる古典確率論の先駆をなす成果は、その後多くのすぐれた科学者たちによってさまざまな角度から洗練されて行き、19世紀の初頭にいたってラプラス(S. de Laplace)により「確率の解析的理論」として集大成された。古典確率論は20世紀の30年代、コルモゴロフ(A. N. Kolmogorov)による近代確率論の誕生まで確率論の基調をなしていた。

ある現象が確率法則にしたがう場合、原因と結果に1対1の対応がないという意味でその生起に不確実性を伴っている。したがって少量の観測(観察もしくは計測)からその背後にある法則を正しく把握することは困難である。「大量観測」だけがそれを可能にする。大量観測の中から統計的な規則正しさが見出され、その規則正しさを生みだしている確率法則が帰納される。統計学は大量観測の結果認められた統計的な規則正しさを踏まえて、観測の対象となった集団に作用している確率法則の性質を追求する学問であるといえることができる。

大量観測の中にのみ統計的規則正しさ、ジュースミルヒ(J. P. Süßmilch)のいわゆる神の秩序は存在する。しかしその発見は、それを意識して探究することなしにはなされるものではない。いろいろな現象について大量観測がなされた歴史は古い。キケロ(M. T. Cicero)によれば紀元前にすでにローマで財政や軍事に関する統計が作成されていたという。ずっと降って15世紀初頭にはヴェニスで人口調査が行われている。しかしはっきりした目的意識をもって、大量観測の結果である「死亡表」から人間の出生・死亡の中に秘められた統計的規則正しさを発見したのはイギリスのグラント(J. Graunt)である。1592年のペストを契機に毎年ロンドンの年間の死亡数等が記録されることになった。このいわゆる死亡表は、しかしグラント以前にはほとんど有効に利用されることはなかったのである。グラントは死亡表をこまかく分析してそこにさまざまな「秩序」を見出している(「死亡表に関する…自然のおよび政治的諸観察」)。たとえばロンドンでは出生は女子13人に対して男子14人であるが、地方では女子14人に対して男子15人である、といった規則性。そしてかれはこのような秩序が人間生活のあらゆる面に存在するであろうと結論している。

この大数の安定性を確率の概念と結びつけて説明したのはベルギーのケトレー(A. Quetelet)である。そしてこのとき統計学は科学としての地位を保証されたとされている。ケトレーの方法は当時イギリスで盛んになっていた生物学と結びついて大いに発展する。主な推進者はガルトン(F. Galton)とその後継者カール・ピアソン(K. Pearson)である。現在あらゆる分野で広く応用されている回帰や相関の理論などがかれらによって精力的に研究され、統計学はひとつの体系を整えた。ピアソンによって確立されたいわゆる大標本理論までを通常記述統計学(descriptive statistics)とよんでいる。

記述統計学の目的は大量観測の結果に基づいて対象となる集団の特性を記述

することにある。記述統計学では観測結果としての統計資料(データ)から求められた代表値はそのまま集団の特性を表わすことになる。したがってここでは観測資料が十分多くあることが暗黙の前提になっている。

一般に対象となる集団について常に十分多くの観測資料が得られるとは限らない。限られた観測資料しか得られないような統計的実験も少なくない。したがってこのような状況への対応を統計学が指向するのはいわば必然的な成行きといえるであろう。この端緒となったのがアイルランドの化学技師ゴーセツ(W. S. Gosset)(筆名ステューデント: Student)が1907年に発見した t -分布である。これこそ近代数理統計学の先駆をなす業績である。ゴーセツの理論はなお不完全であったが、後にこれを数学的に完全なものとし、さらに多方面に独創的な研究を行って数理統計学(mathematical statistics)の基礎を確立したのはフィッシャー(R. A. Fisher)である。さらにエゴン・ピアソン(E. S. Pearson)やネイマン(J. Neyman)らの貢献によって数理統計学はその体系を確立して行く。

記述統計学では集団の特性が統計資料から直接求められるのに対して、数理統計学では、後に述べるように、母集団と標本が明確に区別されており、標本として得られた統計資料から求められるのは母集団の特性の推定値である。そして推定値は標本ごとに異なる。とすればどのような標本によるどのような推定値がよりよく母集団の特性を表わすかを評価することが必要になる。推定値全体を代表するものは推定量(広い意味での統計量)であるが、数理統計学の中心的課題は、母集団と標本を区別することにより必然的に生じてくる「統計量の評価」を行うことにある、ということができよう。

この数理統計学の発達は確率論の発展と密接にかかわっている。確率論は数理統計学の重要な基礎をなしており、その発展が数理統計学の進歩を促したが、同時に数理統計学からの刺激が確率論の内容を充実させて行ったことも事実である。実際、1930年代以降の数理統計学の発展はまことに目ざましいものがあるが、それはさきにふれたコルモゴロフの近代確率論(測度論的確率論)に大きく負っているといえることができるのである。

本書では数理統計学の基礎的な概念と方法が第4章以下で述べられる。

ところで統計的分析の場において、われわれはつねに厳密に数理統計学の立

場に立たねばならないかといえは、必ずしもそうでないこともある。実際には記述統計学の立場で、得られたデータから度数分布や必要な代表値を求め、それをそのまま集団の特性とみなすことで足りる場合も少なくない。対象となる集団について記述統計学が与える断定的な情報はそれなりに有用であるし、また一般に受け容れやすいのである。またさきに述べたように、その多くをカール・ピアソンに負う記述統計学の方法は、形式的にも理論的にも数理統計学の方法と密接に関連している。そこでまずこの後の2章において記述統計学の基礎を概観し、それから数理統計学に進むことにする。

2

度数分布

2.1 要因, 変数

集団に属する個体について分析の目的に応じて予め定められた観測項目を要因とよぶことにしよう. 要因は質的(定性的)要因(qualitative factor)と量的(定量的)要因(quantitative factor)に分けられる. そして質的要因を観測して得られたデータを質的データ, 量的要因のそれを量的データとよんでいる. 表 1 はある大学学部の学生について実施した調査結果の一部である. 表には 3 つの要因についての観測値が掲げられている. ここでは住居が質的要因である. 質的要因は, 表に示すようにその各属性(カテゴリーともいう)が標識によって表わされる. 質的要因にはその標識がこの例のように単なる名称(あるいは名義)を表わすものと, 順位を表わすものがある. 学業成績を A: 優, B: 良, C: 可, D: 不可 とするのは後者の例である. 質的要因のカテゴリーを表わす標識として 1, 2, 3, ... の数字を用いることがあるが, この場合の数字に数値としての意味はない.

これに対して量的要因は通常そのまま変数として扱われる. たとえば, 表 1 の例では「通学時間」(という要因)を変数 Y で表わす, 等. 観測値 $y_1(75)$, $y_2(60)$, $y_3(25)$, ... は変数 Y のとる値である.

変数は離散(型)変数(discrete variable)と連続(型)変数(continuous variable)に区別される. 離散変数とは表 1 の「家族数」のように実数値としてとびとびの値しかとり得ない変数である. 交通事故による死亡者数などもこの例である. これに対して連続変数とは家計の消費支出, 物の長さや重さ, 気温, 区画された土地の面積, 等のように, 少なくとも観念的には連続した実数値をとる得ると考えられる変数のことである.

表 1 学生 60 人の生活調査資料

番号	住居	家族数 人	通学時間 分	番号	住居	家族数 人	通学時間 分
1	A	4	75	31	C	4	70
2	A	5	60	32	C	5	35
3	C	4	25	33	C	3	20
4	A	4	55	34	C	4	45
5	A	3	70	35	A	3	60
6	C	7	35	36	A	4	75
7	D	3	40	37	B	4	35
8	C	4	60	38	A	3	80
9	B	7	25	39	A	4	85
10	C	2	50	40	B	4	15
11	A	5	50	41	A	6	50
12	A	4	45	42	B	4	45
13	B	4	30	43	B	5	45
14	B	6	50	44	D	5	30
15	C	4	45	45	A	5	60
16	B	6	40	46	D	4	40
17	C	3	40	47	D	7	45
18	A	4	60	48	C	4	45
19	A	3	50	49	C	4	55
20	B	6	10	50	D	5	15
21	A	4	60	51	B	3	30
22	B	5	45	52	A	4	55
23	A	5	55	53	A	5	30
24	C	3	25	54	B	5	35
25	D	6	30	55	A	6	30
26	B	5	35	56	B	2	30
27	C	5	25	57	A	6	60
28	A	5	70	58	A	4	35
29	B	5	30	59	B	4	30
30	B	3	35	60	B	4	35

A：自宅，B：アパート，C：下宿，D：間借。

2.2 度数分布

度数分布表の作成は，多くの場合統計的分析のもっとも基礎的な作業といえることができる。表 1 の「家族数」を例にとって度数分布を考えてみよう。家族数を変数 X で表わせれば X のとる値はこの例では 2 から 7 までである。 X がとる値に対応する観測個体の数を度数 (frequency)，度数を観測個体の総数で割ったものを相対度数 (relative frequency) という。変数 X のとる値とその度