はしがき

この書物は，大学において初めて統計学や統計的データ分析を学ぼうとする学生諸氏のために書かれたテキストである．“わかりやすく” を念頭に，具体例や例題，図表を組み入れて，統計学からの見方・考え方を説明し，学生を含め多くの読者が統計学の基本を学び，その面白さや意義を理解できるように工夫したつもりである．

統計分析，とりわけデータ分析は今日では文科系，理科系の学問の諸分野ばかりではなく，社会・経済では幅広く用いられている．研究，調査，ビジネス，リスク評価や政策立案などにかかわる様々な分野では，それぞれに固有な問題設定や解決へのアプローチとともに，統計データを用いる分析を進め，問題の理解を深め結論を導くという共通する方法を用いることが多いが，それは統計学の内容そのものと重なる．

統計学の方法も日々に発展を重ねているが，統計学の一部が高校数学の教科書で議論されていることから，文科系の学生諸氏の中には厄介な分野と敬遠しがちになると耳にすることがある．そうした学生諸氏も卒業してから金融ビジネス・業界マーケティング，あるいは工場の現場・医療機関・公的機関などの仕事の中でデータ分析に遭遇せざるを得ない機会も日々に増大しているのである．

本書では実例により全体を鳥瞰する第 1 章，統計データの整理と記述のための基礎事項を扱う第 2 章～第 4 章，統計学で必要となる確率の知識をまとめた第 5 章～第 8 章，統計的推測の基礎事項をまとめた第 9 章～第 12 章，社会・経済・時系列データを扱った第 13 章，第 14 章，各章の内容の理解を助ける事項をまとめた付録，により構成されている．特に，第 1 章は統計学がどのように役立つのかをいくつかの実例を通して説明しており，初めて統計学を学ぶ読者にとって問題意識を高める上で助けになるだろう．付録 1 では，統計計算ソフトウェアとして ‘R’ と ‘エクセル’ についての解

説を与えたが，本書で扱われるデータ分析の方法を実装するときに役立つだろう．また付録 2 では，微分・積分や行列・行列式など統計的方法を学ぶ上で役立つ数学の基礎知識をまとめておいたが，本書での数理的記述を理解する上で助けとなるよう配慮したつもりである．

各章は基礎的事項に加えて発展的事項を含めたが，文科系の学生諸氏にとっての統計学入門，データ分析入門では，まずは基礎的事項および付録で説明されている計算機を利用したデータ分析の基礎的操作に習熟することが当面の目標である．さらに統計分析やデータ分析の勉学をすすめたい諸氏には，各章の発展的事項が参考となるだろう．なお，いくつかの章には章末問題を用意したが，問題演習を行うことにより内容の理解が容易になるはずである．章末問題の略解の PDF ファイルをインターネットで容易に検索できるグーグルのサイト（https://sites.google.com/site/ktatsuya77/）に置いたので，読者は必要に応じて自由にダウンロードして利用されたい．

なお，本書は東京大学経済学部 2 年生向けの基礎科目「統計」，法学部の科目「統計学」の講義の中から生まれたことを付け加えておく．受講生たちによる講義への質問，TA（teaching assistant）を務めてくれた院生からのコメントは大いに参考となった．また，東京大学出版会の大矢宗樹氏には，丁寧に原稿を読んでいただき，図表と数式の多い本書の刊行にあたってご尽力を頂いた．これらの方々にこの場をかりて感謝の意を表したい．本書を通して多くの読者が統計学の有用性についての面白みや理解が深まることを期待したい．

2016 年 9 月

久保川達也
国 友 直 人

第 1 章

統計学とその役割

我々の日常生活は様々なデータに囲まれていて，データから必要な情報を引き出して生活している．例えば，人間ドックから得られるデータは日頃の食生活を反映しており，健康的に生きるための改善がなされる．病院では様々なデータをとって病気の原因を突き止めていく．天候や気温などの気象データから，今日着ていく服を決めたり，傘を持っていくかを判断する．また旅行会社・家電メーカー・農家にとっても，気象データに基づいた長期・短期の天気予報は経営を維持していく上で重要である．その他，為替レートや株価などの金融データ，景気指標，GDP（国内総生産），失業率データ，家計調査データなど，様々なデータや統計指標（データを加工したもの）が利用されている．

この章では，統計で何ができるのかについていくつかの例を紹介し，統計とはどのようなことをする学問であるのかを説明する．

1.1 データは語る

例 1.1　分布から実態を探る　4 月の初めに予備校でのクラス分けのために，500 人の受講生について数学の実力試験を行ったところ，得点が図 1.1 のように分布していることがわかった．この得点分布の中心は 50 点付近にありそうで，実際，平均点を計算すると 48.6 点となる．得点分布の全体をながめてみると，35 点から 45 点をとった

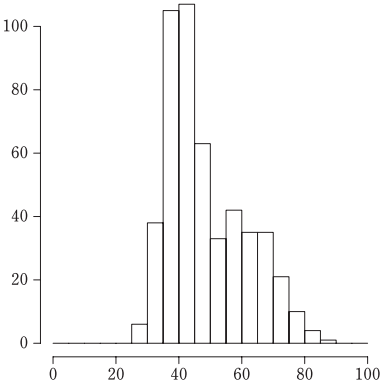

図 1.1　全体の分布

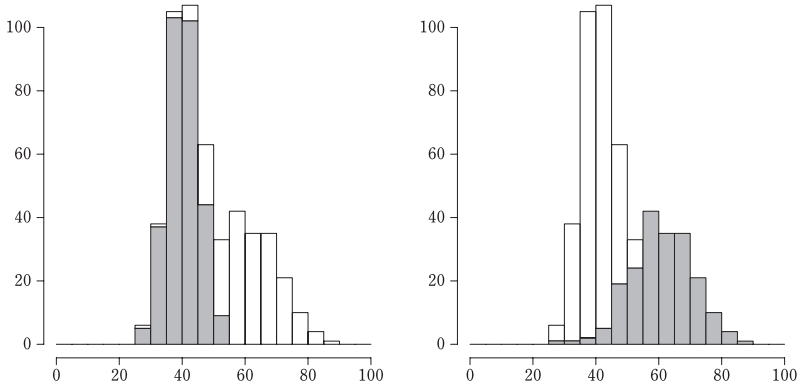

図 1.2 （左）現役生の分布，（右）浪人生の分布

学生が多いものの，より高い得点をとった人数も多く，全体的に右に歪んだ形をしていることがわかる．この歪みの理由として，現役高校生と浪人生の成績が混在していることが容易に想像がつく．実際，現役生と浪人生に分けて分布を描いてみると図 1.2 となる．この左の図が現役生，右の図が浪人生であり，現役生の分布と浪人生の分布を組み合わせることによって受験生全体の分布が合成されている．現役生の平均は 40.3 点でその周りに分布しているのに対して，浪人生の平均は 60 点で現役生の得点より散らばりが大きいことがわかる．

このように，データの全体の分布を描いてみると，分布の全体的な特徴を捉えることができ，その特徴が何に由来するのかを探っていくと，データの背後にある姿を捉えることができる．

図 1.3 は所得の分布を描いたものである．全体を眺めると右に歪んだ分布をしている．200 万円を中心に分布している集団，600 万円を中心に分布している集団，1,000 万円を中心に分布している集団，また業種によってはも

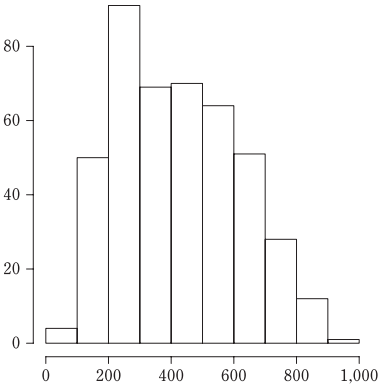

図 1.3 所得分布

っと高所得の範囲で分布しているかもしれない．こうした様々な集団の所得分布の合成として全体の所得分布が描かれることになる．したがって，これを性別，年齢別，業種別，役職別に分けて分布を描いてみると，社会の実態が多少垣間見られるかもしれない．

例 1.2　2 つの変数の関係を測る　ある中学校の生徒 200 人の教科別成績に関して，数学と理科の得点が以下の表で与えられる．

生徒	1	2	3	···	200
数学 x	50	58	43	···	52
理科 y	40	38	35	···	42

これらを 2 次元のデータと見なして $(50, 40), (58, 38), (43, 35), \ldots (52, 42)$ を x-y 平面にプロットしたのが図 1.4 の左の図である．また，同様にして数学と社会の得点をプロットしたのが図 1.4 の右の図である．左図を眺めると，数学の得点と理科の得点の間には正の比例関係があり，数学の得点が高ければ理科の得点も高いように分布していることがわかる．一方，数学と社会の得点の間には，このような関係をみることはできない．このように 2 次元データを x-y 平面にプロットすることにより 2 つの変数の間の関係を捉えることができる．しかし，図を眺めて判断するというのでは解析する人によ

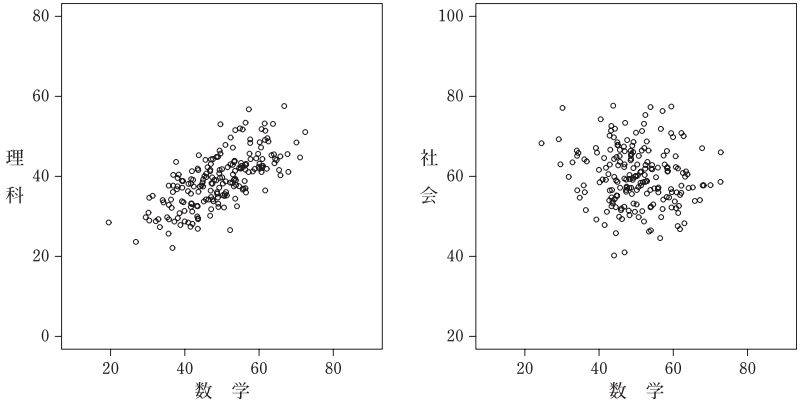

図 1.4　2 次元データのプロット：(左) 数学 – 理科，(右) 数学 – 社会

って判断が分かれてしまい，また解析する人の恣意性が入りこんでしまう危険性もある．そこで，2 つの変数の関係を測る客観的な統計指標が必要になる．それが相関係数である．具体的な定義は 4.1 節で扱うことになるが，上の例では，数学と理科の得点の間の相関係数は 0.74，数学と社会の得点の間の相関係数は −0.12 となる．−0.12 は相関関係があると考えるのか，無いと考えるか，どちらに判断したらよいかについては，統計的推測の検定法を用いる必要がある．検定方法は確率に基づいた推測統計法の一つであり，本書の後半で学ぶことになる．

例 1.3　関係の有無を問う　喫煙と肺ガンの間の関係など，2 つのカテゴリーの間の関係を調べたい場合がある．上の例 1.2 では数学と理科の得点の関係を調べたが，得点は 0 から 100 までの値を取りうる変数である．これに対して，喫煙については，喫煙しているか否かの 2 値（0 - 1 データ）のみとり，肺ガンについても，肺ガンか否かの 2 値しかとらない．例えば，全体で 106 人調査し，肺ガンの患者 63 人のうち，喫煙していた人が 60 人，喫煙していない人が 3 人，また健常者 43 人のうち喫煙していた人が 32 人，喫煙していない人が 11 人であったとする．このようなデータを整理するのに次の分割表を用いると便利である．

カテゴリー	肺ガンの患者	健常者	計
喫煙していた	60	32	92
喫煙していない	3	11	14
計	63	43	106

この場合，喫煙と肺ガンとの因果関係はどのように測ることができるのだろうか．喫煙していた人と喫煙していない人について，（肺ガン患者数）÷（健常者数）を計算すると，

$$\text{喫煙していた人} = \frac{\text{肺ガン患者数}}{\text{健常者数}} = \frac{60}{32} \fallingdotseq 1.88,$$
$$\text{喫煙していない人} = \frac{\text{肺ガン患者数}}{\text{健常者数}} = \frac{3}{11} \fallingdotseq 0.27$$

となり，喫煙していた人の方が肺ガンになるリスクが高いことがわかる．しかし，このような数値が高いから因果関係があると示唆できても，どの程度高ければ因果関係があると断言してよいだろうか．そのために用意されている統計手法がカイ 2 乗検定という方法である．詳しくは後半の推測統計において学習することになるが，考え方を簡単に説明しておこう．仮に因果関係がないと仮定すると，分割表データは次のようになる．

カテゴリー	肺ガンの患者	健常者	計
喫煙していた	55	37	92
喫煙していない	8	6	14
計	63	43	106

そこで両方の分割表データの差の 2 乗 $(60-55)^2$, $(32-37)^2$, $(3-8)^2$, $(11-6)^2$ を計算し，それらの和なり加重平均が大きければ因果関係があると考えて良いであろう．カイ 2 乗検定統計量は，それらのある種の加重平均のことで Q という記号で表される．Q の値が $\chi^2_{1,0.05} = 3.841$ を越えていれば，有意水準 5% で有意であるといい，誤り確率が 5% で因果関係があると判断する．この手法をカイ 2 乗検定という．いまの場合，$Q = 31$ となり，3.841 より大きいので，喫煙と肺ガンとの間には因果関係があると判断される．

かつて，妊娠中の母親が睡眠薬を服用したときに奇形の乳児が生まれるケースが多かったため，睡眠薬との因果関係が疑われたことがあった．以下の分割表データはそのときに用いられた実際のデータである．

カテゴリー	正常な乳児	異常な乳児	計
睡眠薬の服用	2	90	92
睡眠薬の非服用	186	22	208
計	188	112	300

このデータにカイ 2 乗検定を適用すると，因果関係があることが立証される．

例 1.4　因果関係を直線で表す　次の表は，200 人の親子（父親と息子）の身長のデータである．

番号	1	2	3	4	5	···	200
父親の身長	161	172	170	170	168	···	169
息子の身長	175	175	173	181	174	···	171

父親の身長を x 軸に，息子の身長を y 軸にとって x-y 平面にプロットすると図 1.5（左）になる．この図を眺めると，息子の身長は父親の身長に比例して高くなることがわかる．そこで，父親の身長で息子の身長を説明する直線 $y = a + bx$ を引くことができないか考えてみる．図 1.5（右）のように様々な直線の引き方があるが，一体どのように引くのがよいだろうか．

ここで大事な点は，父親の身長で息子の身長を説明する直線 $y = a + bx$ とは，父親の身長から息子の身長への因果の方向を考えている点である．そこで，例えば $(x, y) = (170, 165)$ なる点について，その $x = 170$ に対応する直線上の点は $(170, a + 170b)$ であるから，この 2 つの点の長さの 2 乗 $(a + 170b - 165)^2$ を考える．図 1.6（左）のように，このような長さの 2 乗を全てのデータに関して和をとり，それを最小にするように a, b の値を決める．これを最小 2 乗法 (least squares method) という．親子の身長のデータでは，最小 2 乗法により得られる直線は

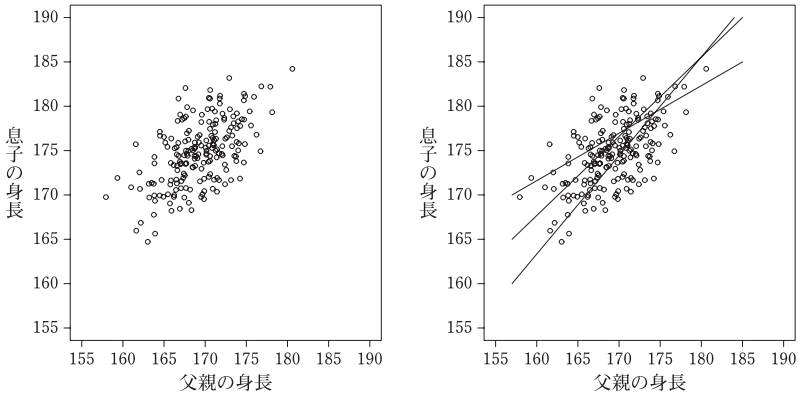

図 1.5　父親と息子の身長の関係

$$y = 80 + 0.56x$$

となり，図 1.6（右）で描かれる．この直線を回帰直線という．

回帰直線は，父親の身長 x に基づいて $y = 80 + 0.56x$ なる直線で息子の身長 y を説明できるので，例えば，父親の身長が $x = 168$ のときには息子の身長は $80 + 0.56 \times 168 = 174.08$ と予測できるので，まだ子どもが幼少であっても大人になると 174 cm 程度になると予想することができる．回帰直線に基づいたデータの分析を回帰分析 (regression analysis) といい，イギリスの遺伝学者・人類学者のフランシス・ゴルトンにより提唱されたとされている．ここでは父親の身長だけを利用しているが，母親の身長 z も利用可能であれば，父母両方のデータを用いて息子の身長を説明する直線 $y = a + bx + cz$ を考えることができ，この方が y に対する説明力が高くなるように思われる．これを重回帰分析という．

重回帰分析の面白い例がイアン・エアーズ著『その数学が戦略を決める』の中で取り上げられている．それは，ある地域のワインの価格は，冬の降水量，育成期の平均気温，収穫期の降水量で決まり，ワインの価格は

（ワインの価格）
$= a + b \times$（冬の降雨量）$+ c \times$（育成期平均気温）$- d \times$（収穫期降雨量）

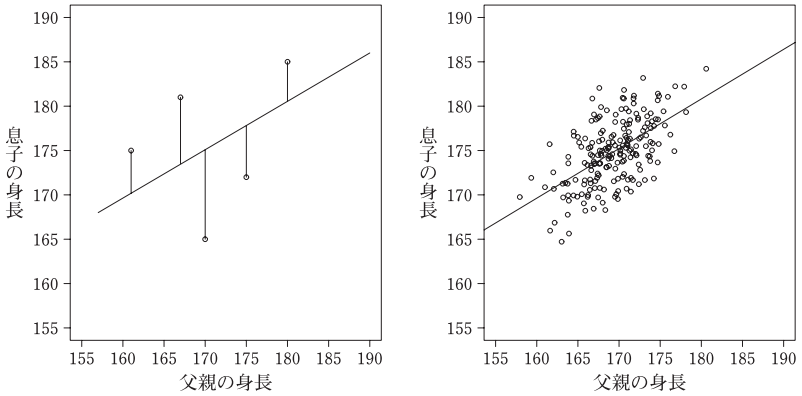

図 1.6　最小 2 乗法による回帰直線

なる形の式から予測できるというものである．ここで a, b, c, d の値は重回帰分析から求めることができる．冬の降雨量がわかっていて，春から夏にかけての天気の長期予報が出ていれば，それらの情報を利用して今年のワインの価格を 3 月や 4 月の時点でも予測できることになる．実際には様々な要因が影響するので，ワインの予測価格がピンポイントで当たることはありえないが，投資するか否かの戦略を考える際，予測式などの統計情報に基づいた戦略は，何も考えずに例年通りの投資を行うことに比べて，はるかに‘賢い’決定を与えることになるであろう．

回帰 $y = a + bx$ を求めてみたとき，仮に x の係数 b が 0 となるときには，x から y への因果関係はないと考えられるであろう．しかし，$b \neq 0$ であるからといって因果関係があると判断してよいであろうか．因果関係の有無を統計的に判断するには，統計的検定法を用いる必要がある．本書の後半で，確率に基づいた推測統計を学ぶ中で，検定の考え方を理解することになる．

例 1.5　2 値選択への因果関係を調べる　宅地の坪当たりの価格 (x_i) と購入したか否か (y_i) に関して 20 件のデータが観測されているとしよう．ここで，y_i は購入したか否かを表すデータなので，

$$y_i = \begin{cases} 1 & \text{購入したとき} \\ 0 & \text{購入しなかったとき} \end{cases}$$

という 0 - 1 の 2 値データの形をとることになる．宅地の価格と購入結果に関する因果関係の有無を解析する問題なので回帰分析を用いることが考えられるが，y_i が 0 - 1 の 2 値のみをとるので，通常の回帰分析を当てはめるのは好ましくない．そこで，次のように $y_i = 1$ となる確率に対して回帰式を当てはめる．

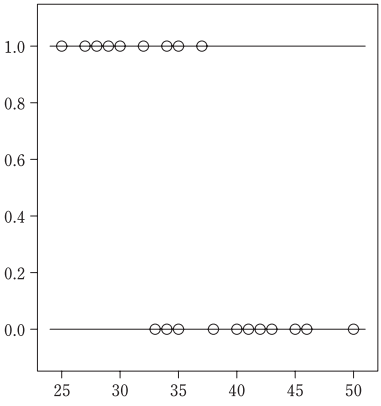

図 1.7　2 値データのプロット

$$\log\left\{\frac{P(y_i=1)}{1-P(y_i=1)}\right\}=a+bx_i,\quad i=1,\ldots,100$$

これをロジスティック回帰モデルという．この他，正規分布の分布関数を当てはめるプロビットモデルなどもある．いずれにしても，このようなデータを解析するには，単なるデータの加工程度の方法では不十分で，確率を導入し確率モデル・統計モデルを作って，より相応しいデータ解析の方法を用いる必要がある．

このように 2 値選択への因果関係を調べる問題は，既婚女性が仕事に就く要因の分析，殺虫剤の濃度と害虫の死滅との関係など様々な応用例があり，ロジスティック回帰などの統計手法が利用されている．

1.2　統計の役割

前節の例 1.3 では，「肺ガンと喫煙との因果関係を調べたい」という目的でデータがとられ，分割表という形でデータを整理し，カイ 2 乗検定を用いて因果関係があると判断する過程が述べられている．このように，統計は，まず分析目的があり，その目的に応じてデータの収集がなされ，データの整理・加工・推論などの統計分析を通して，現象の理解・予測・意思決定がなされる一連の過程をいう．

$$\text{分析目的}\Rightarrow\text{データの収集}\Rightarrow\text{統計分析}\left(\begin{array}{c}\text{記述統計}\\ \text{推測統計}\end{array}\right)\Rightarrow\begin{array}{c}\text{現象の理解}\\ \text{予測}\\ \text{意思決定}\end{array}$$

■データの収集

統計学の研究の大部分は統計分析に注がれているので，統計分析が統計学だと思われがちであるが，実際に最も大事なのはデータを収集する部分である．これは標本抽出 (sampling) と呼ばれ，全体を反映するデータを収集する必要がある．その際に注意しなければならない点は，⑴偏りのないデータ，⑵十分な数のデータ，を収集することである．1936 年米国大統領選挙