2012/03/29

藪友良

1章．統計学とは

統計学という言葉を聞いたことがある方は多くても、統計学とは何かを答えられる方は少ないでしょう。一言でいえば、統計学とはデータを収集し、それを分析する学問です。統計学が対象としうる事象は多岐に及び、日常生活の中でも統計学を活用した事例は数多く存在しています。

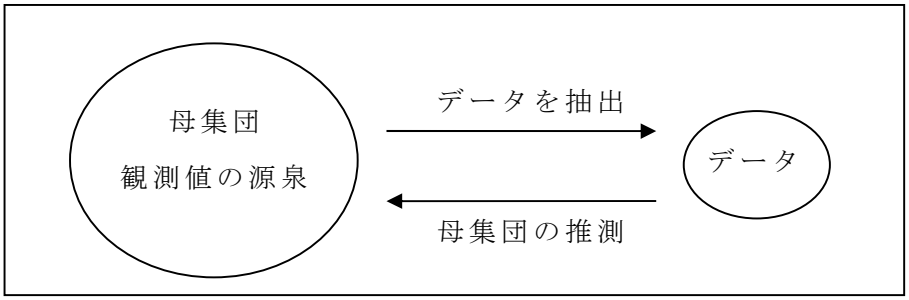
たとえば、みなさんがテストを受けた場合を考えてください。学生は自分の得点しか知りませんが、当然ながら教員は学生全員の得点を把握しています。学生全員の得点は、データにほかなりません。このデータを使って平均や偏差値を計算し、教員は学生の理解度を把握することができます。資産運用にも統計学の知識は欠かせません。企業の決算書というデータを分析すれば、優良な投資対象を決定することができます。また、中央銀行である日本銀行が行う景気判断にも、統計学の知識は不可欠です。日本銀行の調査統計局という部署では、全国からデータを収集、分析し、景気判断および将来予測を行っています。

本章では、さまざまな事例を紹介しながら、統計学の全体像を説明します。これらの事例から統計学の有用性を認識できるはずです。全ての物事は多様な視点から分析することで、真実の姿を明らかにすることができます。統計学はその重要な一面を見せてくれる学問なのです。本章を読んで、これから勉強する統計学の概要を把握していくと同時に、統計学を勉強する必要性を感じとっていってください。

1.1. 統計学の概要

統計学（statistics）の最も基本的な概念である**データ**と**母集団**（population）について説明します(図 1-1 参照)。データとは**観測値**の集まりのことであり、**標本**（sample）ともいわれます。また、データに含まれる観測値の数を**サンプルサイズ**（sample size）といいます[1]。母集団はデータの源泉であり、より具体的には､データの抽出元である集団やデータを生成する構造自体を意味します。

図 1-1：概念図



統計学の関心は母集団そのものの性質を調べることです。しかし、母集団全てを調査することは、コストと時間の両面から困難です。そこで、母集団の一部であるデータを抽出し、母集団の性質を調査しようというのが統計学です。これは、大鍋に入ったスープの味を確認するために、小さじ 1 杯分だけを味見するのと同じことです。

たとえば、日本の内閣支持率を考えてみましょう。内閣支持率とは、有権者のうち内閣を支持している人々の割合です｡母集団は有権者全員となりますが、有権者は日本全国で約 1 億人もおり、当然ながら全数調査は困難です。そこで実際は 1000 人程度に聞取り調査をしています｡統計学の表現を用いると、サンプルサイズは 1000、データは聞取り調査の結果ということになります。

統計学は、記述統計と推測統計と呼ばれる 2 つの分野から構成されます。**記述統計**はデータを分かりやすくまとめることを目的とし、**推測統計**はデータから母集団の特性を推測することを目的としています｡全数調査が可能であれば、データはまさに母集団そのものなので、その特性を直接知ることができます。しかし、多くの場合、全数調査は時間とコストの両面から困難であって、推測

[1] 観測値の数は、標本数ということもありますが、標本数は混乱を招く言葉であるため使わないことをお勧めします。たとえば、標本数が 10 というと、標本（観測値の集まり）が 10 個あるかのようです。観測値の数は、サンプルサイズまたはサンプルの大きさという方が正確です。

統計が必要となります。

母集団とデータの理解を、以下の 2 つの例を通じて確認しましょう。

例 1(テレビの視聴率):テレビ局の収入源の多くは広告収入です。したがって、テレビ局にとっては、視聴率は広告主から広告料をとるうえで大事な交渉材料となります。視聴率が高いほど広告効果も高いため、広告主である企業から高額の広告料を得ることができるからです。テレビ局は、視聴率つまり「全世帯のうち何%の世帯がある番組を見ているか」を調べています。実際には、ビデオリサーチ社が、テレビ局からの依頼で地域ごとの視聴率を調査しています。同社による関東地方の視聴率調査では、全世帯数 1500 万世帯のうち 600 世帯だけが調査対象とされています。つまり、母集団が 1500 万世帯で、サンプルサイズは 600 世帯です。600 世帯のうち 120 世帯がある番組を見ていた場合、その視聴率は 20%と推定されます。

視聴率は限られた情報から計算されており(データは母集団の 0.004%)、その推定には誤差が生じます。こうしたことから、上記のビデオリサーチ社は「本当の視聴率が 10%なら、その誤差は 2.4%の範囲に収まる」としています。このことは、本当の視聴率が 10%であれば、視聴率は 7.6～12.4%の範囲に収まることを意味します（数字の根拠は 7 章で解説します)。

例 2（サイコロ）: 正 6 面体のサイコロを 1000 回振って、出た目を全て記録したとします。この場合のデータは、記録した結果を指します。他方、このときの母集団は、これまでの例のように何らかの集団があって、そこからデータを抽出したわけではないので、1、2、3、4、5、6 の目が 1/6 の確率で生じるというデータの生成構造自体を指します。サイコロ投げは無限に繰返しが可能です。このような母集団をとくに**無限母集団**といいます。

1.2. データ

抽出された個々のデータには観測番号を付けます。そして、それぞれの番号に対応した変数 X の値を記録します。たとえば、無作為に選ばれた 5 人に身長を聞いたところ、その身長が 172.9cm、180.3cm、142.1cm、120.2cm、172.3cm であれば、表 1-1 のようにまとめることができます。ID が観測番号で、変数 X が身長です。

表 1-1：観測表

ID	X
1	172.9
2	180.3
3	142.1
4	120.2
5	172.3

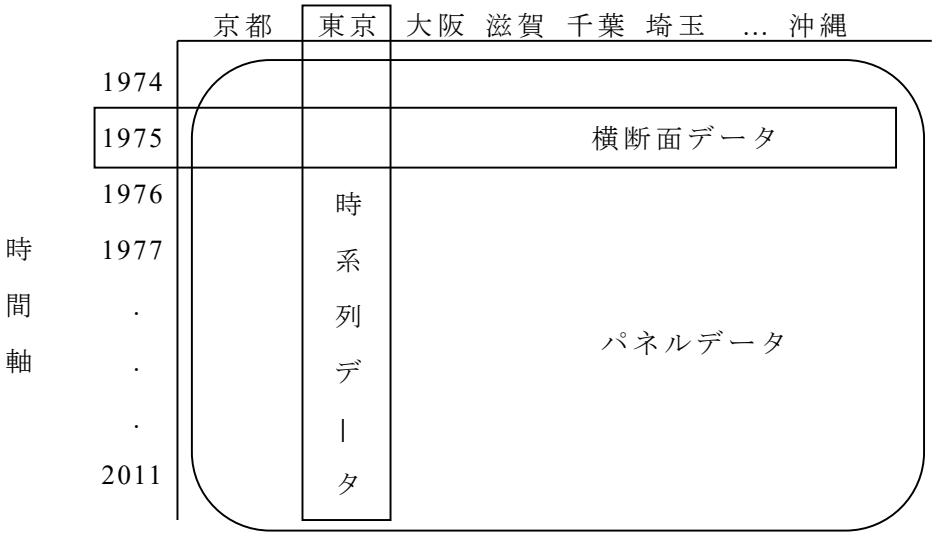
身長のデータは、{172.9, 180.3, 142.1, 120.2, 172.3}です。より一般的な形で$\{x_1,x_2,x_3,x_4,x_5\}$と表すこともあります。x_iは ID が i 番目の人の身長で、たとえば、x_1は ID が 1 番目の人ですから 172.9 となります。なお、統計学では、できるだけ一般的な表記を用いて、さまざまな法則が証明されます。表記が一般的であれば、その適用範囲も広がるからです。次章からは、一般的な表記を用いた説明を行いますが、頑張って慣れていってください。

身長の値のように連続的な値をとりうる変数を**連続変数**、離散的な値だけをとりうる変数を**離散変数**といいます。通常の計測器で計測した身長は連続的な値はとりませんが、理論的には小数点が何桁も存在しますから、身長は連続変数であるといえます。これに対し、ある試合での勝ち負けを記録したデータは、勝ったら 1、負けたら 0 をとる変数と考えられますから、離散変数であるといえます。内閣支持率の聞取り調査も、支持すると答えたら 1、支持しないと答えたら 0 とすると、これも離散変数といえます。

データには、**時系列データ**（time-series data）、**横断面データ**（cross-section data）、**パネルデータ**（panel data）があります。時系列データは時間の経過とともに観

測されるデータであり[2]、横断面データとはある 1 時点において複数の対象を記録したデータです。また、横断面データが複数年にまたがって利用できるとき、このデータはパネルデータといわれます。たとえば、数十年間にわたる日本全体の国内総生産（Gross Domestic Product: GDP）を記録したものは時系列データです。図 1-2 を 1974～2011 年にわたる 47 都道府県の県内総生産の記録であると考えてください。たとえば、1974～2011 年の東京都の県内総生産であれば時系列データ、1975 年だけの 47 都道府県の県内総生産であれば横断面データ、データ全体はパネルデータとなります。

図 1-2：概念図



出典）北村行伸(2007)「パネルデータ分析」『ESP』をもとに作図

1.3. データの収集

1.3.1. 無作為抽出

データを抽出する際には、できるだけ**バイアス**がない形で、母集団を代表するデータを取り出す必要があります。バイアスとは偏りのことで、母集団を代表しないデータを取り出してしまうことを意味します。たとえば、大学生の意識調査をするため、友人にだけ聞取り調査をしたとします。「類は友を呼ぶ」というように、あなたの友人はあなたに似た人が多いかもしれません。たとえば、

[2] 時系列データも観察頻度によって、その言い方が異なったりします。年単位で記録されたデータを年次データ、四半期単位を四半期データ、月単位を月次データといったりします。

男性が多いかもしれませんし､性格も偏りがあるかもしれません｡したがって、あなたの友人という調査対象は母集団である大学生を代表していない可能性があります。バイアスのあるデータの分析からは、バイアスのある結果が出てしまいます。

バイアスの発生を回避する方法に**無作為抽出**（random sampling）があります。無作為抽出とは、母集団を構成するどの個体もデータとして選ばれる確率が同じになる抽出法です。大学生の意識調査の例でいえば、男性も女性も、まじめな人もそうでない人も、同じ確率で選ばれるような方法です。

では、無作為抽出は、どのように行うのでしょうか。1 つの方法は、各面に 0 から 9 までの数字が書かれた 10 面体のサイコロを使うことです(エクセルを用いた無作為抽出の方法は付録 B を参照してください)。たとえば、1000 人の学生から 1 人を無作為抽出するには､まず全員に 000 から 999 までの番号を付けます。番号の付け方はどのような順番でもかまいません。そして、10 面体のサイコロを 3 回振ります。たとえば、サイコロの 1 番目が 3、2 番目が 7、3 番目が 4 であれば 374 番の学生、同様にして、サイコロの目が 0、9、1 の順であれば 091 番の学生､0､0､0 の順であれば 000 番の学生を選ぶというものです。こうすれば、1 人の学生が選ばれる確率は等しくなります。背が高いから選ばれやすいとか、低いから選ばれやすいなどのバイアスはありません。

無作為抽出は理論的には簡単なものですが、現実には誤ったデータ抽出が頻繁に行われており、バイアスのあるデータが多数存在しています。以下で、そうした例を紹介しましょう。

例 1（米国大統領選）: 1936 年の米国大統領選において､『リテラシー・ダイジェスト』誌が勝利者を予想するため、電話や自動車の保有者などから選ばれた約 237 万人に聞取り調査を行いました。その結果、共和党候補ランドン氏が圧倒的な優勢とされましたが、実際の選挙では民主党候補ルーズベルト氏の勝利となりました。なぜ調査結果は誤ったのでしょうか。

共和党は競争政策を重視し、富裕層からの支持が多いのに対し、民主党は弱者保護を重視し、貧困層からの支持が多い政党です。当時、電話や自動車は富裕層が保有するものであったため、電話や自動車の保有者への調査は富裕層に

対する調査にほかならなかったのです。すなわち、共和党支持者に共和党を支持しているかを聞く調査に過ぎず、その結果がランドン氏の優勢と出るのは当たり前のことです[3]。

コラム 1-1：無作為抽出の難しさ

内閣支持率の実際の調査では、どのように無作為抽出が行われているのでしょうか。何ら制約がなければ、調査を行う民間企業が、全ての有権者に番号を割り振り、無作為抽出で選ばれた個人に連絡を取ります。しかし、現実には、個人情報保護法によって、民間企業は有権者の連絡先を自由に知ることはできません。このため、民間企業は厳密な無作為抽出ではなく、できるだけ無作為抽出に近い形で調査を行っています。ここでは日本経済新聞社の **RDD 方式**（random digit dialing）を紹介します。手順は以下のとおりです。

（1）固定電話の番号は、局番（市外局番＋市内局番）+加入者番号（0000～9999）からなる。そこで、局番を小さい順に配列したうえで、1 万件の局番を無作為抽出する。次に、抽出された局番から、それぞれ 1 個の加入者番号を無作為抽出する[4]。

(2)抽出された番号のうち、現在使われていない番号を除去する。この結果、経験的に約 1600 件の世帯番号が得られる。オペレーターは、約 900 件程度の協力を目標として電話をする。

（3）協力してもらえる場合、まず各世帯の有権者の人数を確認する。そして、有権者の人数以下の乱数を発生させ、年齢が上から○番目（乱数）の有権者に答えてもらう。たとえば、人数が 3 人であれば、1、2、3 のどれかをとる乱数を発生させる。回答者が不在であれば、帰宅時間を聞いて、再度、連絡をする。いったん決まった回答者は変更できない。

RDD 方式は、無作為抽出として優れた調査方法に見えますが、問題点もあります。まず、固定電話だけを扱っており、携帯電話だけを所有している世帯が

[3] 同選挙において、ギャラップ社は約 3000 人に対する調査で、ルーズベルトの勝利を正しく予測しました。ギャラップ社も厳密な無作為抽出ではありませんでしたが、『リテラシー・ダイジェスト』誌よりも母集団を代表する抽出法がとられていました。

[4] 1 個の加入者番号を無作為抽出するとは、0000 から 9999 のどれかをとる乱数を発生させて、それを加入者番号とするということです。0 から 9 までの数字が書かれた 10 面体を 4 つ用いれば、乱数を発生させることが可能となります。

除かれています。つまり、母集団は有権者全員ではなく、固定電話を所有している人々だけです。次に、1600 件中 900 件程度の協力を目標としていますが、調査に非協力的な人々（およそ 700 件）すなわち無回答はデータから除外されます[5]。最後に、自分が購読している新聞社であれば、調査に協力する傾向があるかもしれません。たとえば、日本経済新聞社の調査では『日本経済新聞』の購読者が多く含まれ、『日本経済新聞』の考え方に同調している人々が多いと思われます。新聞社によって読者層は異なり、データに偏りが生まれてしまいます。たとえば、民主党の鳩山内閣発足時（調査期間：2009 年 9 月 16～17 日）の支持率は、『毎日新聞』77%、『日本経済新聞』75%、『読売新聞』75%、『朝日新聞』71%、『産経新聞』69%となっており、最大でおよそ 8%もの差がありました。

無作為抽出の難しさを理解していただけたでしょうか。どれほど良い調査機関であっても、完全な無作為抽出はできず、できるだけ完全な無作為抽出に近づけることが求められているといえます。

例 2（ビギナーズラック）: ビギナーズラックとは、「賭事などで、初心者が往々にして好結果を収めること」ですが、それは存在するでしょうか。実際に、ギャンブル好きな人に同じ質問をすると、「初めてのギャンブルで勝った」と答える傾向があるようです。これは、ビギナーズラックの存在を示唆するものでしょうか。

そこで、初めてギャンブルをして勝ったらギャンブルを続ける傾向があり、負けるとギャンブルをやめる傾向があると仮定します。筆者は、大学生のとき初めてのパチンコで、あっというまに 5000 円を失った経験があります。大金を失ったショックから、以降パチンコはやめました。このような経験を持っているのは、筆者だけではないようなので、上記の仮定は現実味を帯びてきます。

この仮定が正しければ、ギャンブルを続けている人々だけを対象に調査を行えば、ビギナーズラックがあったと答えるのは当然の結果となります。ビギナ

[5]有効回答率とは、調査対象人数のうち有効な回答が得られた割合です。有効回答率が高いほど、調査の信頼性は高いといえます。データ収集においては、有効回答率を高めることがとても大事です。この場合、有効回答率は 56%（=900/1600）で、回答率は低いとはいえませんが、データが母集団を代表していない可能性があります。

ーズラックの存在を検証するためには、ギャンブルをやり続けた人々だけでなく、やめた人々も調査しない限り、正しい答えは得られないということです。

1.3.2.質問の仕方

テレビの視聴率や内閣支持率の調査は**社会調査**と呼ばれます。社会調査とは、人々の意識や行動などの実態をとらえるための調査ですが、実施するうえで、無作為抽出のほかにも注意すべきことがあります。以下では、とくに3つの注意点を述べます。

第1に、面接員の誘導的な質問は回答に影響を与えます。郵政民営化に賛成か反対かを調査するとき、面接員が「市場原理主義によって格差が拡大しています。あなたは郵政民営化に賛成ですか」と聞いたとします。原理主義や格差という言葉は悪い意味で使うことが多いものです。質問の中に「民営化は悪い」というメッセージが暗に含まれており、回答者は民営化に賛成と言い難い状況に置かれることなります。

第2に、質問の設定の仕方によって回答が変わる場合があります。コラム 1-2では、雑誌販売において選択肢の設定が重要であることを説明しています。

コラム 1-2：選択肢の設定で答えが変わる？

聞取り調査の際、選択肢の設定の仕方によって回答が変わる場合があり、注意が必要です。ダン・アリエリー『予想通りに不合理』（早川書房、2008年）には、ある雑誌の購読料金の設定が例にあげられています。

この雑誌社では、ある雑誌の購読料について次の料金設定をしています（数値は原書と異なります）。

① ネットでの雑誌閲覧	5000円
② 雑誌の購読	1万円
③ 雑誌の購読＋ネットでの閲覧	1万円

これらの選択肢が与えられると、③を選ぶ人が最も多くなるようです。②と③は同じ値段ですが、③はネットでも雑誌閲覧ができますから②を選ぶ人はいません。では、なぜ雑誌社は意味のない選択肢②を設けているのでしょう

か。試しに意味のない選択肢②を取り除いてみましょう。

① ネットでの雑誌閲覧　　5000 円

③ 雑誌の購読＋ネットでの閲覧　1 万円

こうすると③を選ぶ人が減って、①を選ぶ人が増えてしまいます。①、②、③の選択では③が一番ですが、①と③の選択では①が一番となる人が存在するのです。つまり、雑誌社は③を選ばせるために、②という一見すると無意味な選択肢を加えるのです。

なぜ人々は選択肢が変わると行動を変えるのでしょうか。アリエリーは、「これは人々が相対性を意識して選択を行うからである」としています。人々は比較できるものは意識しますが、比較できないものは無視する傾向があります。たとえば、あなたが友人を合コンに連れていくとき、どのような友人を連れていけば、自分の人気が上がるかを考えてみましょう。自分と違ったタイプを連れていっても、自分との比較はできませんから、あなたを良く見せることもありません。相手に良い印象を与えるためには、自分と比較可能な人で劣っている人、つまり、自分とタイプは似ているが劣っている人が望ましいといえます。

第 3 に、答えにくい質問を聞かれた場合には、回答者は嘘をつく可能性があります。たとえば、先生（もしくは上司）が「未成年のとき飲酒したことがありますか」、「男女差別は許容されますか」とあなたに質問した場合、たとえ「はい」という答えを持っていたとしても、違法性や反道徳性を気にして、率直に「はい」と答えることは難しいと思われます。このような質問に対して正直に答えてもらうための方法に、回答のランダム化があります。

例 1(回答のランダム化)：ここでは「未成年のとき飲酒したことがありますか」という質問に正直に答えてもらうため以下の方法をとります[6]。まず、回答者は質問者に見えないようにサイコロを振り、そして、6 の目が出たらどのような質問に対しても「はい」と答え、1～5 の目が出たら「未成年のとき飲酒したこ

[6] ジェフリー・ローゼンタール『運は数学にまかせなさい』（参考文献[8]）に、この方法が分かりやすく解説されています。

とがありますか」という質問に正直に答えてもらいます。つまり、回答者が「はい」と答えても、質問者にとっては本当の「はい」なのか、6 の目が出たことによる「はい」なのかが分からないため、回答者は正直に回答しやすくなります。

こうした回答のランダム化を実施しても嘘をつく人はいるでしょう。しかし、回答のランダム化によって嘘をつく動機は大きく低下し、通常の調査に比べて本当の答えが得られる可能性が高くなります。たとえば、1200 人の聞取り調査で、700 人が「はい」と答えたとします。1200 人の 1/6 である 200 人は 6 の目が出たから「はい」と答えており、意味がある回答はそもそも 1000 人だけです。また、意味がある「はい」は 700-200=500 人だけです。したがって、意味がある「はい」の割合は 500/1000、つまり 50%の人が飲酒経験ありと推察されます。

社会調査を行う際は、質問の内容や仕方によって結果が大きく左右される点に留意する必要があります。質問内容によっては、回答者が本当のことを答えないので、データの取り方（つまり質問の仕方）を工夫し、できるだけ回答者が本当のことを話してくれるような環境を作る必要があるのです。

社会調査は、我々の意識や行動の実態を捉える重要な調査です。しかし、世の中の調査には、データ収集の基本を踏まえないで行われたものが数多く存在しています。谷岡一郎氏は、社会調査の現状について次のように述べています。

「世の中に蔓延している社会調査の過半数はゴミである。始末の悪いことに、このゴミは参考にされたり引用されることで、新たなゴミを生み出している。では、なぜこのようなゴミが作られるのか。それは、この国では社会調査についてのきちんとした方法論が認識されていないからだ。いい加減なデータが大手を振ってまかり通る日本－データラメ社会―を脱却するために、我々は今こそゴミを見分ける目を養い、ゴミを作らないための方法論を学ぶ必要がある」（谷岡一郎『「社会調査」のウソ―リサーチ・リテラシーのすすめ』文藝春秋、2000 年）。

新聞やテレビで流されている情報をそのまま鵜呑みにするのではなく、それが意味のある情報かを判断する目を養う必要があります。ぜひ本書を通じて、情報の真偽を判断する統計的視点を養ってもらえたらと思います。

1.4. 確率の計算

米国の政治家であり、物理学者でもあったB・フランクリンは、「死と税金のほかには、確実なものはなにもない」と語っています。世の中の多くは不確実であり、その事実を無視するのではなく、不確実性を確率的に把握して上手に対応することは、人生を賢く生きていくうえで欠かせません。

サイコロの目がそれぞれ1/6の確率で生じるというような簡単な確率の計算は正確に行うことができますが、少し複雑な確率の計算となるとそうはいきません。通常は頼りになる直観も、なぜか確率に関してはうまく働かないことが多いのです。以下で、我々の直観がいかに頼りにならないかが明らかになる2つの事例を紹介します。これらの事例を読めば、確率を勉強して正しく確率の計算ができることの重要性が理解できるはずです。

例1（HIV検査の偽陽性問題）[7]：HIV検査は、HIVウイルスへの感染の有無を調べる検査です。反応は、陽性＋か陰性－の2つで、陽性ならHIV感染の「疑いあり」、陰性なら「疑いなし」です。検査は完璧ではなく、感染者なら100%の確率で陽性反応が出ますが、非感染者でも検査に反応する抗体を持っている可能性があり、1%の確率で陽性反応を示すとします。そこで、全人口の0.1%だけが感染者であると仮定した場合に、陽性の検査結果が出たとき、その人がHIVに感染している確率はどのくらいでしょうか。

多くの人は、感染者なら100%の確率で陽性反応が出るため、陽性反応が出たら高い確率でHIVに感染していると考えがちです。しかし、「感染者が検査をして陽性反応が出る確率」と「陽性反応が出たときに、その人が感染している確率」は違います。結論からいうと、陽性反応の出た人がHIVに感染している確率は、わずか9%にしか過ぎません。

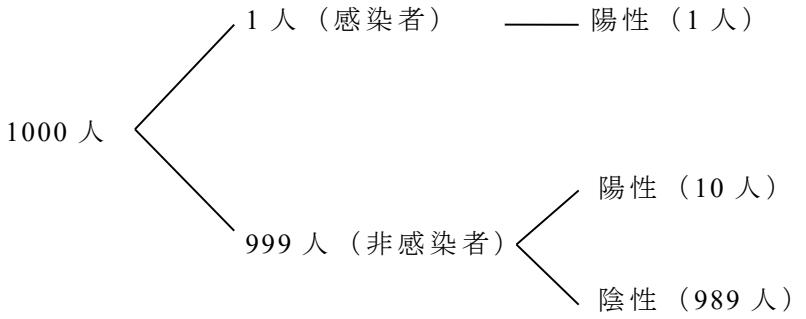
なぜ答えが9%となるかを説明しましょう（図1-3参照）。全人口を1000人とします。このとき、1人（全人口の0.1%）が感染者、残りの999人が非感染者です。感染者は100%の確率で陽性反応が出ますから、この1名の感染者は必ず陽性と診断されます。非感染者であっても1%の人は陽性として診断され

[7] この例は、4章で勉強するベイズの定理を用いて厳密に解くことができます。

ることから、999×0.01 から約 10 人が陽性[8]、残りの約 989 人は陰性と診断されます。以上より、陽性と診断されるのは 10+1=11 人ですが、実際に感染者は 1 人に過ぎません。したがって、陽性結果を受け感染している確率は 9%（=1/11）となります。換言すれば、陽性と診断されても、本当は陽性ではない偽陽性の確率が 91%もあるのです。

健康診断で精密検査の必要ありと診断されたが、精密検査を受けたら何も問題がなかった、という経験がある人も多いでしょう。HIV 検査に限らず、簡易検査で陽性反応であっても精密検査では陰性ということは、以上のような確率の計算からすれば自然なことであるといえます。

図 1-3：HIV の検査結果



コラム 1-3：偽陽性の実体験

『読売新聞』の記事を紹介します。この記事を読めば、物事を確率的に正しく判断する重要性が理解できると思います（以下は記事からの引用）。

「神奈川県内の主婦 B 子さん（21）は昨年 2 月、妊娠 4 か月の時、産婦人科で HIV 検査の結果を陽性と告げられた。考えもしなかった病名を突然告げられ、おなかの赤ちゃんと私、それに夫はどうなるのかと混乱し、涙が出た。会計を待つ待合室でも、車を運転しながら帰る途中も、泣き続けた。通常、HIV 検査はスクリーニング検査と確認検査の 2 段階で行われる。スクリーニング検査は、あくまでもふるいわけのための安価で簡易な検査。そこで陽性と出ても、その後、より精密な確認検査で陰性と判明する偽陽性である場合が少なくない。特に妊婦の場合は、社会全体よりも陽性者の割合が低

[8] 999×0.01=9.99 のところを、約 10 人としています。

いため、スクリーニング検査で陽性とされる人の 9 割以上が偽陽性だという。B 子さんの場合も、偽陽性だったことが後で分かった。スクリーニング検査の結果が陽性だった場合、偽陽性の可能性を十分に説明することが重要だ。ところが、医師が説明なしに陽性と伝えてしまうと、妊婦が『自分は感染している』と思い込み、パニックに陥ってしまうケースがある。エイズ予防財団の矢永由里子さんは、『偽陽性の妊婦が、陰性と判明するまでの間に受ける心の傷については、これまで顧みられてこなかった。しかし、一時的とはいえ陽性と言われることで、夫婦間にひびが入る例もあり、取り返しがつかないことになる』と指摘する」(『読売新聞』(2007 年 5 月 16 日付)。

例 2(1 人っ子政策):中国には、人口を管理する目的から子どもの出産は 1 人までという制度があります。しかし、両親の面倒をみるのは男子であるという考えから、女子が生まれたときは戸籍に残さず、男子が生まれるまで子を産み続けるという事象が生じ、問題となっています。

そこで、中国にいる子どもの男女比を考えてみましょう。単純化のため、全ての夫婦が男子を産むまで子どもを産み続けるとします。たとえば、一番目が男子ならば子作りをやめます。一番目が女子なら男子を産むことを諦めず、もうひとり子どもを産みます。もし二番目が男子なら終りで、二番目も女子なら 3 番目の子を産みます。男子が生まれるまで子どもを産み続けるならば、男子の比率が高くなると考えられます。これは本当でしょうか。

実は、このような行動がとられても男女比は同じままです。何番目の子どもを産むかに関係なく、常に、男子は 50%の確率で生まれます。これこそが、男女比がちょうど 50%となる理由です。もちろん、どのような夫婦も男子が生まれたら出産をやめるため、全ての夫婦に男子は 1 人だけです。しかし、どの夫婦についても女子は 0 人以上います。0 人かもしれませんし 5 人かもしれません。あるいはもっと多い可能性もあります。このため、1 人っ子政策は、ある家族における女子の人数にばらつきを与えますが、全体の男女比には影響を与えません[9]。

[9]男子の出生確率を 0.5 としましたが、実際には男子の生まれる確率の方が高くなっています。世界的には約 0.51 ですが、中国では約 0.54 となっています。その理由としては、何らかの産み分け

1.5. 仮説検定

仮説検定（hypothesis testing）とは、仮説がデータと整合的かどうかを検証する方法です。統計学の用語では、検証したい仮説を**帰無仮説**（null hypothesis）といい、帰無仮説が誤っていたときの受け皿としての仮説を**対立仮説**（alternative hypothesis）といいます。仮説検定では、これらの仮説を設定したうえで、どちらの仮説がデータと整合的であるかを判断します。

ある会社での社内恋愛を考えてみましょう。若手社員の太郎君は、社内に気になる女性がいます。ところが、太郎君は、彼女がある社内の男性と恋愛中との噂を聞きました。その真偽を確かめるため、太郎君は仮説検定を行うことにしました。帰無仮説を「2 人の間に恋愛関係がない」、対立仮説を「2 人の間に恋愛関係がある」とします。これらの仮説を検証するため、太郎君は、2 人が有給休暇を取っているタイミングがどれだけ一致していたか、を調べました。その結果、2 人の有給休暇のタイミングは完全に一致していました。これは偶然と考えるには不自然です。年末年始、お盆休みなどで休みが偶然重なることはありえます。しかし、有給休暇の取得日が完全に一致しているのを、偶然と考えるのは無理がありそうです。この結果から、太郎君は「2 人の間に恋愛関係がある」と判断し、太郎君の片思いは終わりを迎えました。

太郎君の悲しい恋の結末はともかく、これを統計学的にいい表すと、太郎君は帰無仮説（2 人の間に恋愛関係なし）を棄却して、対立仮説（2 人の間に恋愛関係あり）が正しいと判断したということです。さらに厳密にいうと、太郎君は、帰無仮説（2 人は無関係）が正しいとき、休みが一致する確率は低いにもかかわらず、2 人の休みが一致していたので、そもそも帰無仮説が誤っていると考えたのです。

少し難しい表現かもしれませんが、仮説検定は何も統計学に独自の考え方ではなく、生活の中で自然に行っている考え方なのです。

1.6. 回帰分析

経済学や物理学などを理解すれば、さまざまな変数間の関係を推察できます。

をしている可能性も考えられます。

たとえば、金利が上がれば、企業にとっては借入れコストが上昇するため、設備投資が減少します。また家計は預金から得られる収入が増加するため、消費を減らし貯蓄を増やすでしょう。このように、金利の上昇は国内需要（投資や消費など）の減少をもたらし、GDP を低下させます。これは経済理論から分かります。しかし、経済理論から分かることは方向性（増える、減るなど）のみです。これに対して、金利の上げ下げを行う日本銀行が知りたいのは、金利 1%の上昇が GDP を何%減らすかといった具体的な政策効果です。このような具体的な数値を知る 1 つの方法が、変数間の関係を数量的に測る方法である**回帰分析**（regression analysis）です。

以下では、銀行などの融資業務で用いられている回帰分析の例として、クレジットスコアを紹介します。

例 1（クレジットスコア）: クレジットスコア（以下スコア）とは 3 桁の数値で表される個人の支払い能力を測る尺度で、いわば信用度の偏差値のようなものです。たとえば、回帰分析を用いて、借り手が今後 2 年間で債務不履行に陥る確率を推定します。債務不履行確率が低ければスコアは高くなり、逆に高ければスコアは低くなります。スコアの高さは信用度の高さを意味し、数値が高いほど融資が受けやすくなります。スコアを用いた融資手法は米国で 1960 年代に利用され始め、同国における利用件数は 2000 年には年間 100 億件以上までに至っています（日本でも 1998 年より利用が開始）[10]。

クレジットスコアでは、たとえば、勤続年数、職業、借家か持ち家か、信用照会された回数、預金残高、過去の支払い滞納歴などが、債務不履行の確率に影響を与える変数として考慮されます。勤続年数が長かったり、公務員だったり、持ち家だったり、預金残高が多かったりすると、統計的に債務不履行の確率が低くなることが分かっており、その結果、スコアは高くなります。逆に、過去に支払いを滞納した記録があると債務不履行の確率が高くなることが分かっているので、スコアは低くなります。また、金融機関からの信用照会の回数が多いと、多額の資金を必要としていると判断され、スコアは低くなります。

[10]米国の状況は、カイザー・ファング『ヤバい統計学』（参考文献［5］）の中で詳しく解説されています。

スコアは、これら変数の情報をパソコンに入力するだけで容易に求めることができます。融資審査では、事前にスコアが何点以上なら融資をすると決められています。

先ほどの太郎君は独身ですが、マンションを購入することにしました。太郎君は住宅ローンを組むため、ある銀行に融資審査を受けにいきました。この銀行では 700 点を足切りラインと決めていたとします。融資担当者が、太郎君の情報（勤続年数=2 年、持ち家か借家か=借家、居住年数=5 年、信用照会=1 件、過去の支払い滞納歴=なし、など）をパソコンに入力すると、スコアは 720 点と算出されました。太郎君のスコアは 700 点を超えていますから、融資担当者は太郎君への融資を行うことにしました。

個人向けの小口融資審査でも、従来は熟練の融資担当者が必要でした。しかし、クレジットスコアを使えば複数案件を簡単に処理できるようになります。なお、一般的には、融資額の低い案件ではクレジットスコアが活用され、融資額の高い案件は熟練の担当者がより多くの指標を用いて総合的判断をしていることが多いといわれます。

1.7. 統計学以外の視点

社会をより深く理解するうえで、統計学の知識がいかに重要であるかを説明してきました。しかし、統計学から重要な視点を得ることはできますが、それはあくまでも 1 つの視点に過ぎません。バランスの良い正しい判断を行うには、統計学以外のさまざまな分野の知識が不可欠です。

以下には、統計学の知識のみに基づいた判断の危険性を、3 つの事例を通じて確認しましょう。

例 1（黒人への差別）:「タクシーに乗車した黒人による犯罪率は高い」とされることがあります。「犯罪率が高い」という情報を真に受ければ、黒人はタクシー強盗を行う可能性が高いという印象を受けます。しかし、黒人を取り巻く環境を考慮すると、この情報は異なる意味を持っていることが分かります。かつては黒人に対して差別や悪印象が残っていました。黒人が手を挙げてタクシーをつかまえようとしても、車はすぐには止まらないことが多々ありました。こ

れでは、多くの黒人はタクシーを利用しなくなると考えられます。それでもタクシーを利用するのは、乗車により大きな利益を得られる人々、つまり、最初から犯罪を行うことを考えている人々かもしれません。このように考えると、そもそもタクシーに乗車する黒人の数が少ないうえ、乗車する者は犯罪を行うことを企てている可能性が高いことから、タクシーに乗車する黒人の犯罪率は高くなります。黒人への差別が、タクシーに乗車した黒人による犯罪率を高めている可能性があるのです。

例 2（ホットハンド）: スポーツの世界では、ある選手が連続的に成功する（調子がとても良い）時期を、その選手はホットハンドにあるといわれます。この現象の真偽を確かめるために、バスケットボールの試合結果を調査した研究があります[11]。プロ選手の全シュート結果を調べ、シュートが決まったあと、次のシュートが決まる確率を調べたのです。もしホットハンドが本当に存在するならば､いったんシュートが決まった後のシュート成功率は上昇するはずです。しかし、この調査では、その成功率は低下するという予想に反した結果が得られました。

なぜホットハンドという現象が存在すると人々は信じてしまうのでしょうか。その理由として、小数の法則が指摘されています。**小数の法則**(law of small numbers)は､少ない情報から一般法則を人々が誤って見出してしまうことです。たとえば、150 回コインを投げて、表が出たら 1、裏が出たら 0 と記録しましょう。筆者が行ったコイン投げの結果は以下のとおりです。

111100001011000001110010000111
011001000000111101001001111101
100011000010000010101100010101
010110111110011000101011100010
001011110100111001111001110110

コイン投げですから、もちろん表と裏は**ランダム**(random)に生じています。しかし、人々にはランダムがランダムに見えず、表と裏が連続して起こる傾向

11 この研究に興味のある方は、トーマス・ギロビッチ『人間この信じやすきもの―迷信・誤信はどうして生まれるか』（新曜社、1993 年）を参照してください。

を勝手に見出してしまいます。たとえば、最初の 4 回は連続して表で、その後 4 回連続して裏が出ており、ランダムには見えません。実は、真のランダムは繰返しを生み、人々はこの繰返しに勝手に意味を見出してしまうのです。ここで 1 をシュートの成功、0 を失敗と考えてください。こう考えると、かりにシュートの成功や失敗がランダムでも、人々はホットハンドのような現象があると信じてしまうのも理解できます[12]。

これとは逆の見方もあります。つまり、データからホットハンドが支持されなかったとしても、その存在が否定できない可能性があるのです。たとえば、もし調子の良い選手がいたら、相手チームのマークは厳しくなるでしょう。このため、ある選手がホットハンドの状態になっても、相手チームのマークが厳しく、シュート成功率は下がる可能性があります。どちらの見方が正しいのかは分かりませんが、数字の裏で何が起こっているかを、バランスよく判断する必要がありそうです。

例 3（検挙率低下の理由）: 警察は、被害届や事件の通報を受け犯罪発生を認知します。認知件数とは、こうした犯罪の認知数をカウントしたものです。捜査の結果、容疑者を特定し事件を「解決」することを検挙したといいます。検挙率とは、認知件数のうち検挙された割合のことで[13]、検挙率が高いほど、捜査能力が高いと見なされます。図 1-4 は、1990～2010 年の認知件数と検挙率を表したものです[14]。横軸が時間軸（年）であり、左の縦軸が認知件数、右の縦軸が検挙率を表します。1999 年から認知件数が急増し検挙率は急低下しています。つまり、被害届や通報数は増えていますが、事件の解決が追いついていません。この結果は、捜査能力の低下や犯罪の凶悪化を意味しているのでしょうか？

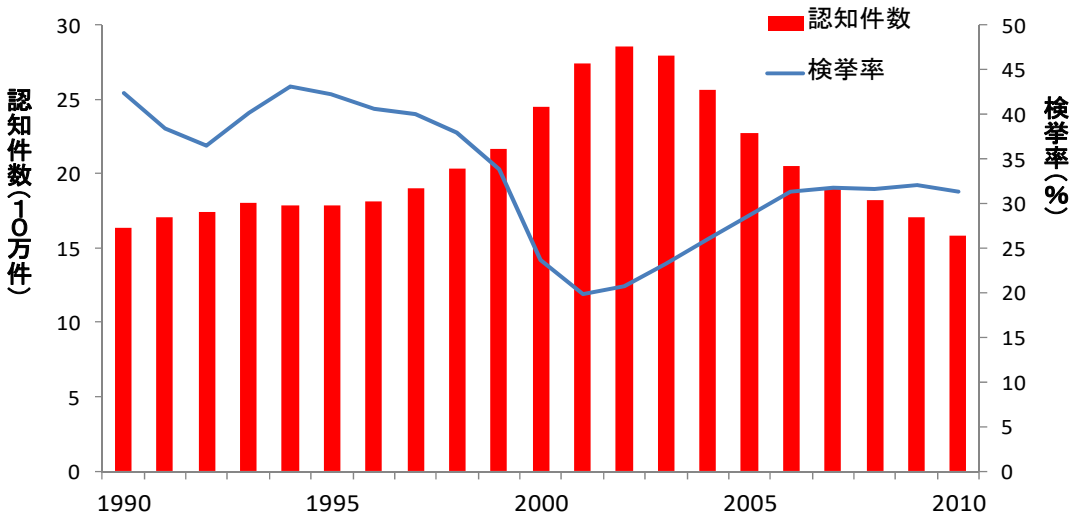
---

12 ムロディナウ『たまたま』（文献[3]）の中で、アップル社が同様の問題に直面していたことが、次のように紹介されています。「アップル社は、音楽プレーヤーiPod で最初に採用したランダム・シャッフリングの方法でその問題にぶつかった。というのは、真のランダムネスはときどき繰り返しを生み出すが、同じ歌が同じアーティストによって繰り返し演奏されるのを聞いた iPod ユーザーが、シャッフルはランダムではないと思ったからだ。そこで『もっとランダムな感じにするために少しランダムではなくした』と、アップル社の創業者スティーブ・ジョブズは言った」（259 頁）。つまり、ランダムであれば偏りを生むものですが、これは iPod ユーザーが必ずしも望んでいたものではなかったのです。

13 正確には、警察で事件を送致送付または微罪処分にしたら検挙といいます。検挙しても、冤罪の可能性もあり、本当の解決になっていない可能性もあります。

14 全ての刑法犯、ただし道路上の交通事故などに関わるものは除いている。

図 1-4：認知件数と検挙率の推移



（出所）警察庁犯罪情勢。

もちろん、その可能性は否定できませんが、それ以外の可能性もあります。別の 1 つの可能性を示すのが、1999 年に起こった「桶川ストーカー事件」です。同事件では、女子大生が元交際相手によるストーカー被害を警察に相談したのに取り合ってもらえず、元交際相手により殺害された事件です。谷岡一郎『データはウソをつく―科学的な社会調査の方法』（筑摩書房、2007 年）の中で、次のように述べられています。「このあとマスコミを始め、世間が警察を批判したのは当然で、新しく就任した警察庁長官は、ある決断をしたのでした。その決断とは、各地方の責任者を集め、今後は直接の被害が現出していない些細なケースでも、人々の相談にのり、捜査をスタートさせるという方針でした……相談件数が増えますと書類上の認知件数も増える。しかも、その増加分は、具体的被害がないがゆえに犯人がわからないケースが多く、当然のように解決（検挙）にいたることはできません」。つまり、谷岡氏の考えによれば、認知件数が増え、検挙率が下がったのは、捜査能力の低下でも凶悪犯罪の増加でもなく、警察が以前よりも被害者の声に耳を傾けるようになったからということになります。

ところで、もう一度、図 1-4 を見てください。2002 年頃から認知件数が下がり、検挙率が上がり始めています。このことは、日本は安全になっていること

を示しているのでしょうか。先と同じ理由から、日本が安全になったのではなく、些細なケースでも捜査をスタートする方針を変えてしまった可能性もあります。データを見て、そのままの数字だけで判断するのは危険なことです。裏にある状況や、データ特有のクセを吟味することも大事なのです。

本節では統計学のみに基づいた判断の危険性を説明しました。しかし、これらの事例は「統計学を用いたデータ分析が意味を持たない」ことを示しているわけではありません。全ての物事は多様な視点から分析することで、真実の姿を明らかにできます。統計学はその重要な一面を見せてくれる学問です。本書を読み進めることで、統計学という新しい視点が読者の身につき、複合的な見方で物事の分析ができるようになり、生きていくうえでの確かな力となるはずです。

練習問題

1.1 ①ある 1 時点の学生 100 人分の身長の記録、②ある 1 時点の 1000 人分の内閣を支持しているかを調査した記録（支持なら 1、不支持なら 0 と記録）、③過去 10 年分の東京都の降水量の記録、④社員 100 人の 5 年分の給与の記録。データ①、②、③、④について、連続変数と離散変数のいずれか、時系列、横断面、パネルデータのいずれかを答えてください。

1.2 日本の大学生全体の統計学の理解状況を調査するために、ある大学の全学生 1000 人に理解度を測るテストを行いました。(1) サンプルサイズを答えてください。(2) データ収集は適切でしょうか。適切でないと考える場合、その理由と解決法をあげてください。

1.3 銀座の街頭で年収に関するアンケート調査を 200 人に対して行った結果、有効な回答者の平均年収は 800 万円でした。一方、国税庁「平成 21 年分 民間給与実態統計調査」によると、平成 21 年の平均年収は 405.9 万円です。この街頭調査の何が問題でしょうか。

1.4 ある国立大学は地方出身者が多く、大学周辺の不動産屋では入学者を事前に取り込む目的から「事前予約」というシステムを導入しています。このシステムを用いれば、受験者は合格発表日まで部屋を押さえることができ、不合格であれば無料でキャンセルできます。地方からの受験者である三郎君は、受験直後に大学周辺を歩いていると、不動産屋の「部屋を事前予約した人は合格率 75%」という広告を見つけました。この大学は難関校であり、通常の合格率は 30%以下でした。三郎君は縁起をかついで、事前予約を用いて部屋を押さえることにしました。事前予約した人の合格率が高くなる理由について答えてください。なお、不動産屋は嘘をついていないとします。

1.5 政府からある会社に対して、アンケートへの協力依頼がありました。アンケートは、詳細な取引内容を尋ねるものでしたが、匿名性が保たれると明記されていました。この会社は、政府に協力的であることが有利と考えて、協力することにしました。調査に問題があれば述べてください。

1.6 ある調査会社が、成人男性の読書量を調べるために対面調査を行うことにしました。ある女性調査員が、調査の一環として会社員の太郎君のもとを訪れ

ました。調査員はとても魅力的であり、太郎君は彼女に良い印象を与えたいと思いました。調査員が「週何冊の本を読んでいますか」と聞くと、太郎君は「週 10 冊読んでいますよ」と答えました。この調査に問題点があれば述べてください。

1.7 面接員が直接会って聞取りをする対面調査について、その利点と問題点を述べてください。

1.8 インターネットによる聞取り調査の利点と問題点を述べてください。

1.9 1982 年のカリフォルニア州知事選挙で、白人候補と黒人候補の対決がありました。世論調査では黒人候補が圧倒的有利でしたが、実際には白人候補が勝利しました。この調査の問題点を述べてください。なお、調査では適切な無作為抽出が行われていたとします。

1.10　無作為に選ばれた大学生 1000 人に、「未成年のとき喫煙したことがありますか」と質問したところ、200 人が「はい」と答えたとします。(1) この調査では、喫煙経験のある人は全体の何%でしょうか。また、この調査の問題は何でしょうか。(2)回答者に質問のリスト「①あなたは大学生ですか、②未成年のとき喫煙したことがありますか」を見せます。回答者は質問者に見えないようコインを投げて、コインが表なら①、裏なら②の質問に答えてもらいました。その結果、1000 人中 700 人が「はい」と答えました。このとき、喫煙経験者は全体の何%でしょうか。(3) ここで、(1) と (2) で推定結果が異なるのはなぜでしょうか。

1.11　マンモグラフィーは代表的な乳ガン検査です。陽性と診断されれば「乳ガンの疑いあり」、陰性であれば「乳ガンの疑いなし」とされます。この検査では、本当に乳ガンであれば 100%の確率で陽性と診断されますが、乳ガンでなくても 9%の確率で誤って陽性と診断されるとします。また、全体の 0.3%が乳ガンにかかっているとします。陽性診断が出たとき、その人が本当に乳ガンである確率は何%でしょうか。