Stats for schools (advanced)

大学での学びにつながる

統計で身近な現象や社会の課題を探究するスタディガイド

高校からの

統計・データサイエンス活用

～上級編～

統計的思考力を身につけよう！

総務省政策統括官
（統計基準担当）

はじめに

今、私たちの身の回りにはテレビや新聞・雑誌・インターネット等を通して、統計資料や調査データから作成されたグラフや表に基づくたくさんの情報が手軽に入手できます。とくに、政府や地方自治体による公的統計を始め、種々のデータが公開されるようになってきた現在、情報をどのような状況で活用するのか、情報の内容と質を見極めながら、私たち一人ひとりが正しい判断や価値選択を行う能力を身につけることが期待されています。

また、21世紀の社会は、「人口問題」や「環境問題」など、社会経済・自然現象の両面に渡って、未来がどうなるのか予測がつかない、不確実性が増大している社会です。このような社会では、確かな統計データに基づいて不確実性を科学的に分析し、その結果をエビデンスとして社会課題に論理的に取り組む、思考力や判断力、いわゆるデータサイエンス力の高い人が求められています。

そのため、学校では、自分で見出した具体的な課題を探究する「総合的な学習の時間」が設けられています。

しかし、「総合的な学習の時間」をどのように充実させればよいのか、その方法が分からず困っていませんか？

探究的な学習では、

身の回りでいま何が起きているのか？

それが私たちの日常の暮らしや未来への判断にどのように影響するのか？

よりよい暮らしや社会を実現するために、いま何をすべきなのか？

をできるだけ客観的な統計資料を提示しながら、効果的に表現し伝えることが大切になります。もちろん、テーマは、社会的なことに限らず、スポーツや自然環境、将来の進路や趣味にいたる幅広い対象を自由に選ぶことができます。ただし、何を対象としても、主張や判断が統計的な分析結果によって導かれた科学的なエビデンスに基づいているということが重要です。

本書は、既に、中学生以上の生徒のみなさんを対象に刊行している、探究的な学習の取組み方を学ぶ学習ワークブック（基礎編）に続く上級編として編集されたものです。基礎編にはない実践的な事例を通して、統計的探究プロセスの流れや統計分析を行う上での理論的な背景も理解できる構成になっています。

本書を通して、みなさんの統計・データサイエンス活用能力が更に向上し、探究的な学習や課題研究がスムーズにできるようになれば幸いです。

平成29年３月

総務省政策統括官（統計基準担当）
新　井　　豊

目　次

第６部　地域の課題解決と統計活用

第７部　統計的思考によって大学入試・統計検定を乗り切ろう！

第８部　公的統計を通してセカイを見る！

第９部　公的統計の有用性を知り、統計調査の重要性を学ぶ

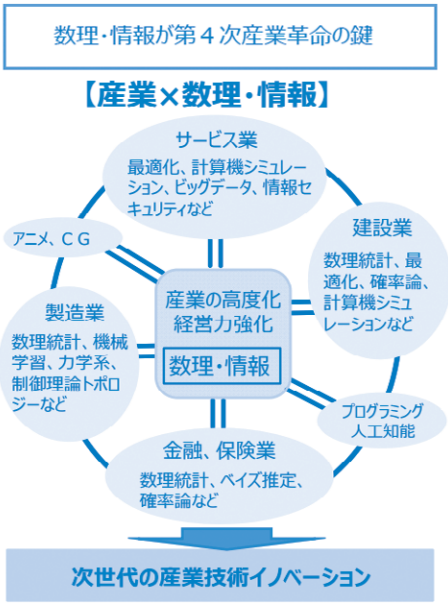
第1部 統計的探究のプロセス

1　いま、求められる“統計・データサイエンス力”とは？

（1）　第４次産業革命の鍵となる統計

21世紀に入り、人工知能を搭載したコンピュータソフトが囲碁や将棋の名人に勝利したり、ロボットが会話したり、自動車が自動運転したり、モノとモノとがインターネットを介して直接データをやりとりするなど、これまでSF の世界とされていたことが、私たちの身の回りで急速に現実化している。この背景には、状況を示す膨大なデータからコンピュータが次々とルール（法則）を学習して、最適な予測や判断を行うことを可能にした統計的なデータ分析技術の進歩がある。

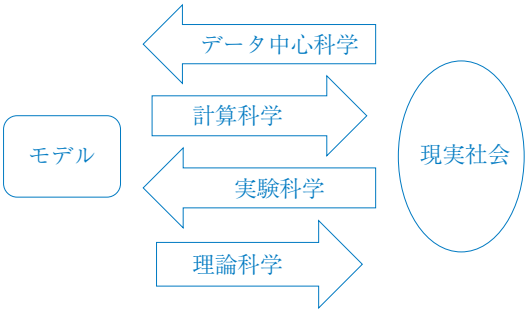
この技術は、既に迷惑メールの検知、クレジットカードの不正使用の検知、数字や顔画像の認識、商品購入のレコメンデーション、医療診断、信用リスクの予測、自然言語処理など、社会で広く応用されているため、現在はデータを中心とする科学技術で第４次産業革命が到来したとまで言われている。



資料：文部科学省「第４次産業革命に向けた人材育成総合イニシアチブ」関連資料（2016年４月）

これまでの産業革命を振り返ってみよう！
　まずは、18世紀後半の工業化の黎明期を語る第１次産業革命。これは、蒸気機関による自動化の時代。
　次に、19世紀後半の大量生産と文明化を語る第２次産業革命。これは、電気による自動化の時代。
　続いて、20世紀後半の電子化による製品・生産設備システムの進化を語る第３次産業革命。
　これは、コンピュータによる自動化の時代。
　そして現代は、第４次産業革命。データ駆動型サービスによる自動化の時代に突入したんだ。

（２）　第４の科学のパラダイム：データ中心主義



データを中心とした変革が進んでいるのは、産業界だけのことではない。大学や大学院に入ってからの研究の方法についても、これまでは、知識の蓄積を背景とした「理論科学」や「実験科学」の方法、コンピュータの発達による「計算科学」の方法が中心であったが、新しく第４の科学のパラダイム（The Forth Paradigm）として、膨大なデータから直接、社会・自然・経済・人間行動等のルール（法則）を発見する「データ中心科学」が急速に広まってきている。

この「データ中心科学」によって、医学、健康科学、生物学、物理学、地学、経営学、経済学、社会学、教育学、スポーツ科学などの多くの領域で、データを活用した創造的な研究成果が生まれてきている（Tony Hey、2009）。

調べてみよう！

データが研究にどのように役立てられているか具体的に調べてみよう

（例）　遺伝子データの医学での活用

遺伝子データが医学研究に活用されるようになって、いろいろな病気の発生リスクがどのようなタイプの遺伝子型と関連しているのかについて研究が進められ、病気の予測や治療方法の選択に役立てられている

大学に「データサイエンス学部」が創設

文部科学省は、平成28年に「大学の数理・データサイエンス教育強化方策について」を公表し、国立大学法人の拠点大学として下記の６大学を選定しています。この背景には、データが豊富に入手できる時代となっているなかで、データとアナリティクスを用いた意思決定を行う企業の割合が世界平均61％であるのに対し、日本は40％と低い状況であること、今後、世界ではますますデータを利活用した新産業創出や企業の経営力・競争力強化がなされるという予想があります。 このため、数理的思考力とデータ分析・活用能力を持つ人材の育成と社会に価値やサービスを生み出すという目的に合致した大学教育システムの構築を目指しています。 拠点大学は、最初は６大学ですが、今後、日本の多くの大学で、統計・データサイエンスの教育拡充が進められる方向です。

数理およびデータサイエンスに係る教育強化拠点大学選定校一覧（国立大学法人が対象）

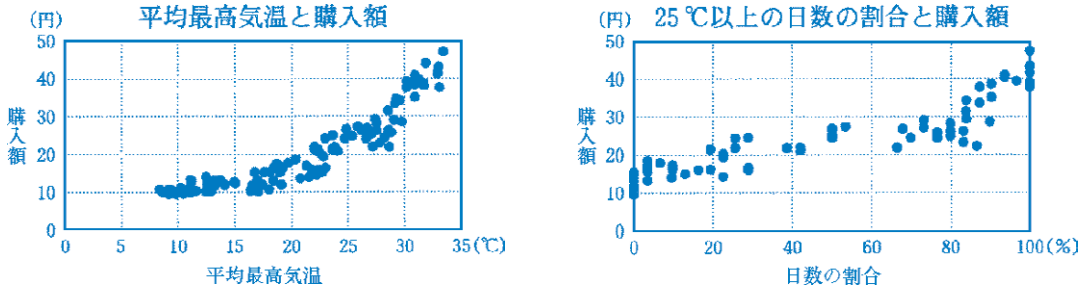
NO	大学名	事業名
1	北海道大学	数理的データ活用能力育成特別教育プログラム ～データサイエンスセンター（仮称）の設置～
2	東京大学	数理・情報教育研究センターの設立
3	滋賀大学	データサイエンス教育の全学・全国への展開 －データリテラシーを備えた人材の育成に向けたカリキュラム・教材の開発－
4	京都大学	データ科学イノベーション教育研究センター構想 －21世紀のイノベーションを支える人材育成－
5	大阪大学	数理・データ科学の教育拠点形成
6	九州大学	九州大学「数理・データサイエンス教育研究センター（仮称）」構想

（３） 統計的探究とデータサイエンスの考え方（センター試験の問題を例に）

データが社会の中心となるなかで、いま、私たちには、身近なデータを活用して新しい知識を創造する探究力が求められている。そのためには、探究のための科学的な思考の方法と統計的にデータを分析する方法を理解し、身に付けなければならない。

社会課題を科学的な思考で取り組むとは、身の回りの課題や地域・社会の課題をいろいろな側面から検討し、広い視野（マクロな視点）で捉えた上で、課題を複数の具体的な現象の関わり合いとして絞り込み、各現象にデータを対応させ、現象間の関連性のルール（法則）を統計的な分析の方法で検証していくという方法をいう。

2016年度の大学入試センター試験の数学Ⅰの出題問題から見てみよう。アイスクリームの購入金額という現象と気温や湿度といった現象の関係を探究して、何がアイスクリームの消費を促すかの法則を見出す設定になっている。そのため、2003年から2012年までの10年間の東京都の月別データとして、１世帯当たりのアイスクリーム消費額（家計調査）と気象庁が公開する気温等のデータを集め、次のような散布図を作成している。



散布図上で、世帯のアイスクリームの消費額という日別や月別で値が変動する変数、気温というやはり値が一定しない変数との関連を具体的なデータで示すことが、特定の品目（アイスクリーム）の消費を気温で評価し、予測するためのエビデンスを得ることに繋がる。また、消費額と気温の関係に、曲線や直線のモデル式を当てはめれば、更に具体的に、消費に与える気温の効果を数量で見積もることもできる。これは、アイスクリームの製造会社が販売量や利益を考える上で、非常に有用な情報（新しい知識）にもなる。

考えてみよう！
① アイスクリームの消費に関係する気温以外の現象は？
② 気温と関係する他の消費品目は？
③ 予測に役立ちそうな２つの関連する現象の他の例は？
④ これが予測できると、こんな課題解決に役立つのでは、と思われる例は？

このように、企業などで何かの判断や決定をするときには、客観的で信頼のおける情報が必要になる。直感や口コミの情報ではなく、できるだけ信頼性のある公的な統計データを使うこと、また、データの数についてもより多くのデータを使うことで、対象とした現象に関して何が起きているのか、その傾向を俯瞰することができる。

現象を表すデータをいろいろな方向から考え、その変動を説明する要因を探し出すことで、予測ができたり、問題解決に繋がる効果的な介入策を考えたりすることができる。これが提案や判断の根拠に繋がる。統計・データサイエンス力とは、統計グラフの種類や個別の分析手法の知識を覚えるだけではなく、現実の社会の課題を捉え、問題を明確にし、その問題を解決する方策を検討し、適切な意思決定を行う力のことである。

次の節で、そのための方法として、PPDAC メソッドを学習しよう。

2　課題発見と問題解決のフレーム：PPDAC メソッドの活用

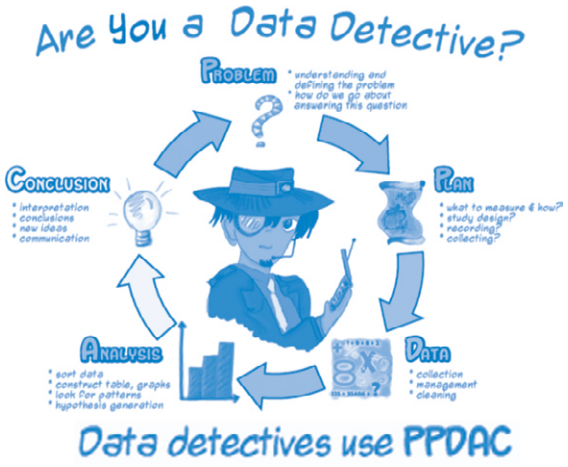
PPDAC メソッドとは、カナダ・米国・ニュージーランド等の学校教育で使用されている科学的探究の手順を示したもので、漠然とした課題をデータで解決可能な問題に落とし込んだ上で統計分析し、元の課題の内容に照らして、状況を判断したり、解決策を提案したりする次の一連の探究活動のフレームをいう。

P	Problem 問題の設定 とらえる	①関心のあるテーマを決め、そこでの**課題**を考える ≪概念図や俯瞰図を作成する≫ ②課題から**問題の構造（原因系と結果系の現象）**を明確にする ≪ロジックツリーや特性要因図・要因連関図などを作成する≫ ③具体的な**研究仮説（リサーチクエッション）**を設定する
P	Plan 計画 みとおす	・問題の重要度を測る指標、その変動に影響を与える要因系の指標など、計測すべき変数（データ）データや統計資料を決め、その収集計画を立てる ・研究仮説を明らかにするための分析の計画を立てる ・分析結果の見通しを立てる
D	Data データ あつめる	・情報（データや統計資料等）を実際に取得し、整理する。データの取得方法（実験か質問紙調査か観察・記録なのかの区別）を意識する。
A	Analysis 分析 まとめる よみとる	表やグラフを作成したり、代表値を計算したりして、データや統計資料を分析する。下記は主な分析の視点である。 ・**全体の傾向（分布）**を見る ・条件の違いやグループに分けて、**比較**する ・指標間の**関連性**を見る ・指標間の**因果関係**を見る ・**時間経過による変化**を見る ・対象を**分類**する
C	Conclusion 結論 いかす	最初に立てた研究仮説に対して判断や結論を示す。同時に、元の課題の内容に戻り、分析に基づいた考察や提言をし、新たな研究課題の提起から次の探究サイクル PPDAC へと繋いでいく。

考えてみよう！

右の図は、ニュージーランドの学校で使用されている PPDAC サイクルのポスターある。それぞれのステップで、どのような内容が書かれているか、英語を訳して考えてみよう。

（例）　Problem
・Define the problem
・Investigative Question



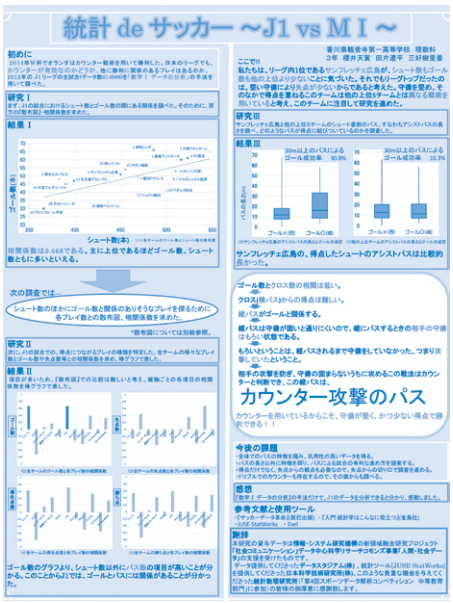
資料：CensusAtSchool NZ

3　高校生の事例で学ぶ PPDAC メソッドの活用

PPDAC メソッドで具体的にどんな分析ができるのか、高校生たちが行った実際の課題研究を例に見てみよう。

◎ Ｊリーグ　チームの強さとプレイの相関分析

【香川県立観音寺第一高等学校】



右のポスターは、2015年度スーパーサイエンスハイスクール生徒研究発表会において、生徒投票賞を受賞した香川県の高校生３名のグループによる研究作品である。2013年のＪ１の全試合データを分析して、リーグ１位のサンフレッチェ広島の強さの要因を解き明かしている。

ここで使用している分析手法は、高校１年生で全員が学習する数学Ⅰの「データの分析」に出てくる、散布図、相関係数、箱ひげ図である。そのため、タイトルに数学Ⅰを表すＭⅠが入っている。最後の「感想」に、「数学Ⅰ　データの分析」の手法だけで、Ｊ１のデータを分析できると分かり、感動しました。」とあるように、基本的なグラフと分析手法だけでも、目的に沿ってそれらを組み合わせることで、オリジナルな研究成果をあげ、新しい知見を見出すことができるという優れた研究事例となっている。では、どういう統計的探究のプロセスをたどったのか、PPDAC のプロセスに沿って考えてみよう。

Problem

◇ テーマと対象、課題の設定

「2014年Ｗ杯でオランダはカウンター戦術を用いて勝利した。日本のリーグでも、カウンターが有効なのかどうか、他に勝利に関係のあるプレイはあるのか、2013年のＪ１リーグの全試合‥を用いて調べた。」とあるように、サッカーＪ１リーグのチームの試合を対象として、何がチームの勝敗を分けるのかがテーマである。

統計的探究における「対象」とは、具体的な観察対象のことで、ある前提条件の下で、複数の対象が観察（測定）可能なものを指す。

「課題」とは何か？

課題とは、「対象」に対しての理想の状態を想定し、現実とのギャップを意識することから見いだされる。サッカーのチームを対象にした今回の場合、理想を“勝利”や“優勝”で捉え、その上で、現実を対比させると、負けもあれば、ランキングで下位になるチームもある。優勝や勝利とのギャップが解くべき課題

＊何がチームを優勝に導くのか

＊何が試合を勝利に導くのか

に繋がる。

■ 課題から問題の構造を見出し、データや統計で解ける研究仮説に落とし込む

ア）評価指標の決定

目的とした「勝利するチーム」という定性的な性質（言葉や感覚で決めた概念）を定量的に測る指標として、まずは決める必要がある。指標が決まらなければ、具体的なデータがとれない。

ここでは、チームの強さを“試合でのゴール数”で計測することとしている。他にも、試合での“得失点差”で測るなどいろいろ考えられる。このように、目的となる指標で、かつ、具体的にデータとして入手できる指標をター

ゲット指標または、最重要評価指標（Key Performance Indicator：KPI）、最重要目的指標（Key Goal Indicator：KGI）といっている。

イ）問題解決は原因分析

データや統計で現実の問題を解くとは、このターゲット指標の値の変動の要因を明らかにし、その値を理想の方向に変える条件や方策を考察することである。

ウ）原因と結果の法則から研究仮説（Research Question）を立てる

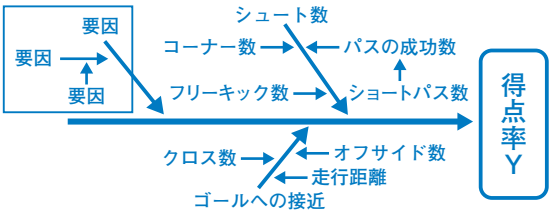
どの要因が最も効果的にターゲットとする指標を変化させるのか？　やみくもにいろいろなデータや資料を集めるのではなく、できるだけ原因と結果の関係に見通し（仮説）を立てた上で、データや統計資料の収集をする必要がある。

仮説を立てる上で、対象に関連するいろいろな現象間の関連性を俯瞰する論理図（**特性要因図**、連関図、ロジックツリーなど）を予め作成することが、統計的探究活動では重要な作業となる。また、このような俯瞰図は、分析の結果を発表する際にも、自分がどのように分析の背景全体を捉えたかを示すためにも、欠かせない重要な資料となる。

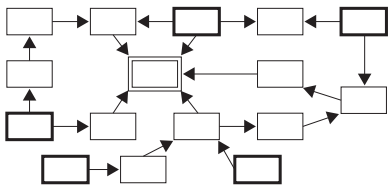
考えてみよう！

ブレーンストーミングのツールを使って、ターゲット指標と要因指標の関係を構造化してみよう

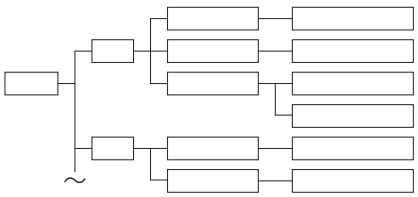
（例）　特性要因図を使うと…



＊連関図を使うと…



＊ロジックツリーを使うと…



ここでは、「試合中のさまざまなプレイがゴール数に関係する」という研究仮説を立てている。

Plan

◇ 仮説の検証に必要なデータや分析の方法を考える

Ｊ１リーグ全チームの2013年の試合における各プレー数、ゴール数等のデータを収集して、散布図や相関係数で関連性を分析する。

Data

◇ データや統計資料を集める・データシートの作成

ターゲット指標に加え、要因系の指標も含めて各チームのデータを１つの行でまとめた、リスト形式と言われる次のようなデータシートを作成する。

順位	チーム名	シュート数	クリア	コーナーキック	直接フリーキック	クロス	パス	キャッチ	ブロック	ドリブル	ファウルする数	ファウルされる数	クリアされる	スルーパス	ゴール数	失点数	得失点差	勝ち点
1	サンフィレッチェ広島	450	748	180	946	648	8397	250	555	557	330	442	746	443	51	29	22	63
2	横浜Fマリノス	469	753	179	543	488	7282	260	618	495	397	534	913	465	49	31	18	62
3	川崎フロンターレ	547	824	177	437	390	8025	311	607	565	419	430	767	600	65	51	14	60
4	セレッソ大阪	510	795	155	371	489	7427	445	554	539	465	360	784	427	53	32	21	59
5	鹿島アントラーズ	516	781	138	450	466	7182	325	550	514	467	437	735	468	60	52	8	59
6	浦和レッズ	486	637	169	448	458	8628	259	489	541	454	438	697	462	66	56	10	58
7	アルビレックス新潟	508	828	220	439	637	6557	322	695	429	448	427	776	579	48	42	6	55
8	FC東京	523	734	148	448	547	7674	313	531	382	396	448	880	577	61	47	14	54
9	清水エスパルス	383	801	152	461	529	6159	315	602	451	435	448	756	406	48	57	-9	50
10	柏レイソル	441	831	176	457	634	7360	240	645	347	494	448	947	533	56	59	-3	48
11	名古屋グランパス	399	839	126	395	565	7860	273	621	370	469	388	798	426	47	48	-1	47
12	サガン鳥栖	455	869	140	448	443	6178	317	699	334	482	440	906	296	54	63	-9	46
13	ベガルタ仙台	500	768	172	421	637	6767	333	608	393	365	413	887	409	41	38	3	45
14	大宮アルディージャ	390	830	157	358	534	6871	291	640	390	430	350	781	529	45	48	-3	45
15	ヴァンフォーレ甲府	366	918	133	429	446	6477	352	542	411	456	427	772	401	30	41	-11	37
16	湘南ベルマーレ	436	1003	144	419	491	6070	331	659	396	455	404	756	457	34	62	-28	25
17	ジュビロ磐田	467	826	222	348	666	7427	314	656	482	477	338	973	469	40	56	-16	23
18	大分トリニータ	385	956	157	406	510	5961	361	638	417	483	405	767	322	31	67	-36	14

Analysis

◇ データを研究仮説に沿って分析する

観音寺第一高校の分析では、一つひとつの分析結果から仮説を次々と進化させ、分析を深めている。

最初の仮説

ゴールに最も結びつくプレーはシュートである。仮説として「シュート数がゴール数に影響する」とする。

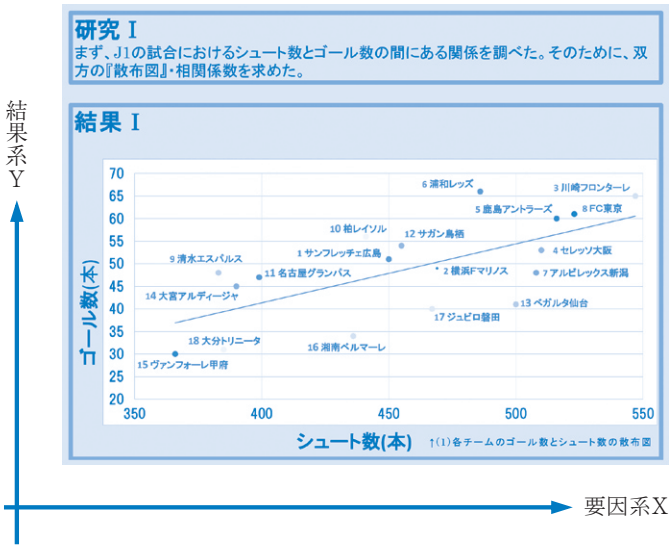
分析の方法

散布図と相関係数

分析結果

シュート数を横軸、ゴール数を縦軸とした散布図では、シュート数が多いチームほどゴール数も多く、逆に、シュート数が少ないチームはゴール数も少ないという正の相関関係（相関係数r＝0.68）が見られた。

その中で、１位のサンフレッチェ広島は、他の上位集団に比べて、シュート数が少ない箇所に位置している。また、ゴール数も特に多いというわけではない。この理由はどこにあるのか？そこで、次の分析の仮説を考えた。



↑(1)各チームのゴール数とシュート数の散布図

次の仮説

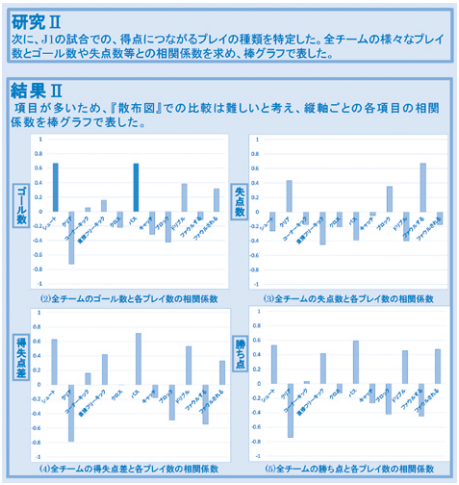
ゴール数以外の勝利に結びつくターゲット指標が存在する。また、シュート以外のプレイも、それらと関係する。

分析の方法

＊勝利に結びつくゴール数以外の指標の洗い出し
＊それらに繋がるプレイの洗い出し
＊各プレイ数とターゲット指標との相関係数の算出
＊プレイごとの相関係数の比較を棒グラフで表現

分析の結果

ゴール数、得失点差、勝点などの指標に対して、シュート数以外に、パス数の相関係数が高い。



覚えておこう！　散布図と相関係数

a　ヒストグラム、箱ひげ図と並ぶ統計３大グラフの１つ
b　２つの数量変数Ｘ と Ｙ の関係のパターンを分析するグラフ
　＊直線傾向の場合（相関関係）
　＊曲線的な傾向もある
c　散布図上で、各対象のポジショニング（位置）が分析できる
　＊傾向に沿う対象
　＊傾向から外れる対象
d　相関傾向の強弱は相関係数ｒで計量化できる
　＊ｒは、－１から＋１の間の値
　＊０に近いほど、相関関係は弱くなる
　＊負の値‥負の相関
　　正の値‥正の相関
　＊絶対値 |r| が１に近いほど、相関関係（直線傾向）が強い、Ｙ の予測モデル（直線）の誤差が小さくなる、Ｙ の変動を Ｘ の変動で説明する説明力が高くなる

Conclusion

◇ 結論とそこから生じる新たな課題を考えるステップ

サッカーの試合に対して、勝利に貢献するプレイとしてシュートがゴール数と関係があることを示しただけではなく、他のプレイも関係すること、そのなかでもパスが重要であることを相関係数によって示している。また、シュート数、ゴール数で特にに大きな特徴がないサンフレッチェ広島の優勝要因が何であるのか、新たな探究課題を見出している。

覚えておこう！　分析を成功させるイロハ

イ）局所管理する

対象の種類の違いが分析結果に影響を与えることを避けるため、対象としているデータはできるだけ同質なものの集合とする

ロ）比較対照をおく

一般的な傾向か固有の傾向かの区別をするため、対照集団（ベンチマーク）をおいて比較する

ハ）繰り返し測定（データの数）する

分析結果の差が単純な標本変動でないことを示すためには、データの数がある程度大きいことが必要

Next Problem

リーグ優勝したサンフレッチェ広島と他の上位チームが、どこで明暗を分けたのか、そこに、サンフレッチェ広島独自の戦術があるのではないか、というテーマを設定し、パスを観測対象とした統計的探究を行う。

Plan

ここで、パスの中でも具体的に、ゴールに結びつくシュート直前のアシストパスに限定して（局所管理）、そのパスによって「シュートが成功したか失敗したか」をターゲット指標に、パスの長さをその関連要因として比較分析することを計画する。

Data

J1の試合で上位チームのアシストパス1本を1行に、下のようなデータシートを作成する。

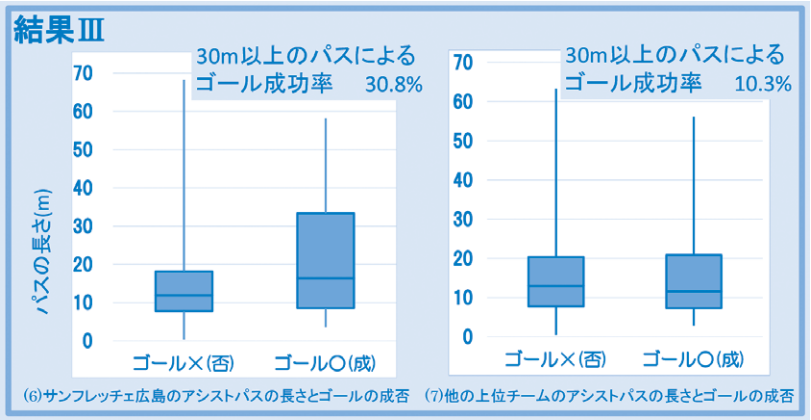
アシストパスID	チーム名	ゴール	パスの距離
1	サンフレッチェ広島	○	38.57
2	サンフレッチェ広島	×	18.34
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・
109	川崎フロンターレ	×	21.08
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・

Analysis

ゴールが成功した場合としなかった場合のそれぞれの「アシストパスの長さ」（量的変数）の分布の比較を五数要約と箱ひげ図で示している。また、その傾向をサンフレッチェ広島とその他の上位チームと比較している。

サンフレッチェ広島		
	ゴール×	ゴール○
最大値	68.27	58.16
第三四分位数	18.14	33.34
中央値	11.89	16.37
第一四分位数	7.79	8.58
最小値	0.37	3.58
アシストパス数	96	12

その他の上位5チーム		
	ゴール×	ゴール○
最大値	63.29	56.14
第三四分位数	20.36	20.86
中央値	13.00	11.60
第一四分位数	7.78	7.32
最小値	0.53	2.87
アシストパス数	595	45

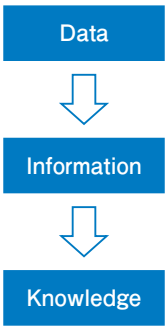


Conclusion

これまでの分析により、サンフレッチェ広島は、ゴール成功時のアシストパスの長さが失敗時に比べて長い傾向にあり、このパスの長さとゴールの成否の関係の傾向は、他の上位チームには、見られないことが分かる。この分析結果から、優勝の要因は、ワールドカップでのオランダチームの優勝要因といわれる、「カウンター攻撃にある」と段階を追った推察で結論付けている。

4　周囲を説得！　できる分析レポートの構成

（1）「データ」→「情報」→「知識」創造を支える分析力



散布図や箱ひげ図といった基本的統計グラフの作成、相関係数や四分位数などの統計量の計算だけでも、「データ」から「情報」を得ることができ、それが背景の文脈の枠組みのなかで考察されることで、その領域固有の「知識」となる。客観的なデータに基づいて得られた「情報」や「知識」は、個人や組織において、提案や提言に説得性をもたせる基礎資料になったり、判断や意思決定の信頼できる基準となる。ここで取り上げた事例は、スポーツデータの分析だったが、PPDAC メソッドの手順は、スポーツ以外のいろいろな課題の発見と問題解決に応用できる。

（2）　説得力をあげる分析レポート：競争優位なレポートとは？

ここでは、どのような視点で分析を加えていくと、レポートの優位性が出てくるのかについて一般的な段階を追って見てみよう：

レベル０：何が起きたのか？　起きたことだけを報告した基礎レポート
「試合に負けた。テストで80点をとった。海岸には空き缶ゴミが捨てられている。給食の残飯が多い。…」など、起きたことだけを報告するレポートは、レベル０。
「サイコロを振ったら２が出た」としているだけで、「サイコロを振ったときに出る目」という現象全体を確率的現象として理解していないことに相当する。

レベル１：どこで、いつ、どうしたら、など、５Ｗ１Ｈに回数や確度を報告した調査レポート
全体の起きうる事象を洗い出し、その起きる回数や確度（相対度数・統計的確率）を調査した上で、現在、起きていることが滅多に起きないことか、よく起きることかを考察したレポート。「サイコロの目は１から６までの数字があり、それぞれ１/６の確率で起きる。その中で２が出た。」というように、試合の結果やテストの得点など、身の回りの現象に分布を対応させて考えることが大切である。

レベル２：そのようなことが起きた問題はどこにあるのか？　考察を加えたレポート

レベル３：取り急ぎ解決に必要なアクションを示し、対策まで加えたレポート

レベル４：統計的に関連性を分析し、なぜそれが起きたかの仮説を加えた基礎統計分析レポート

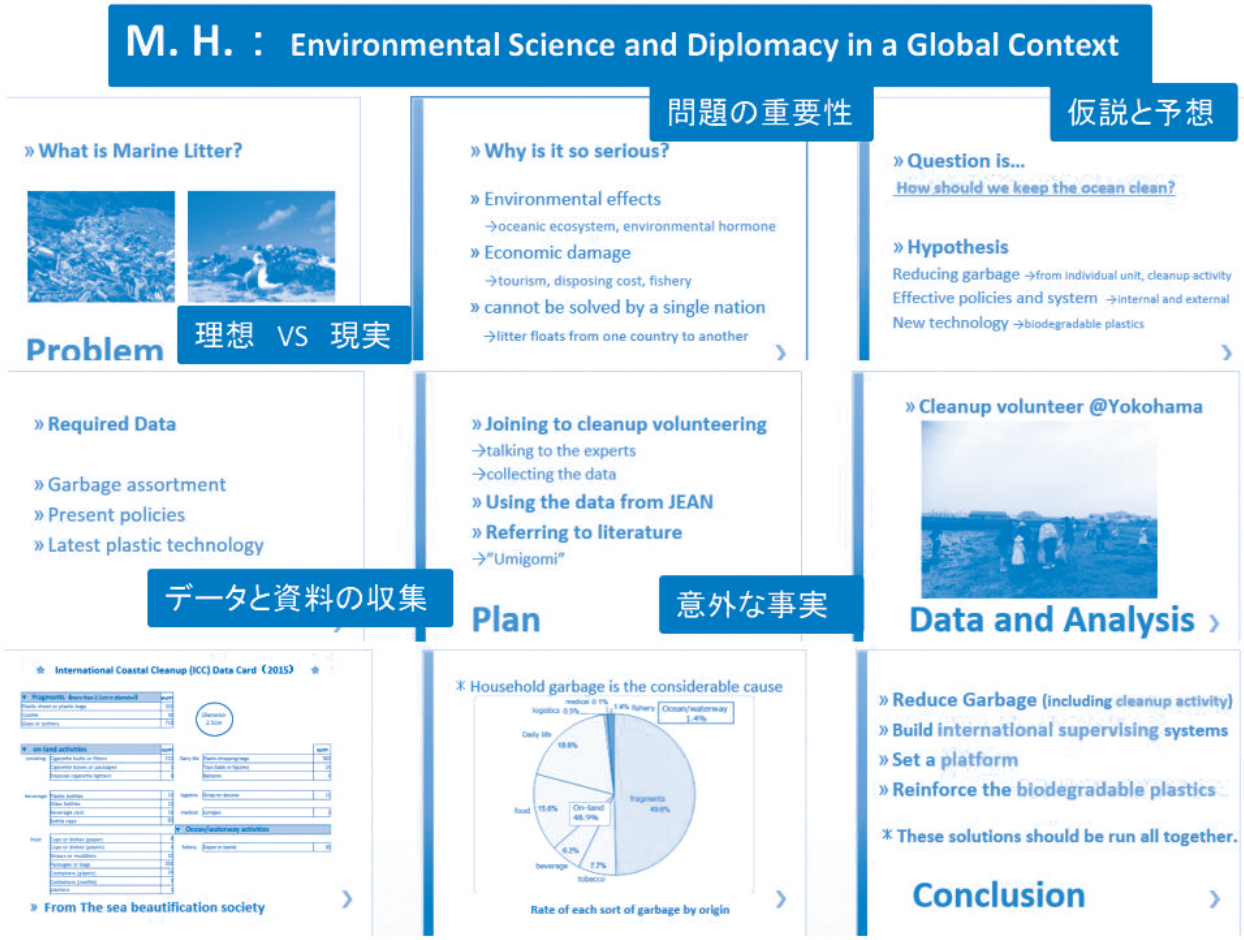
レベル５：この傾向が続けばどうなるのか、予測を示した統計分析レポート

レベル６：複数の要因に対して、それぞれの要因を動かしたらどうなるのかの分析を加たレポート（予測モデルなど高度な統計分析）

レベル７：要因分析を踏まえた上で、取るべき最適な戦略を示した高度な統計分析レポート

下記は、高校生が、PPDAC メソッドに沿って、海岸線のごみの問題を探究した研究成果である。

理想と現実の対比から問題を捉え、問題の重要性を示し、問題解決に有用な方策を想定し、実際のデータを収集・分析によって得られた事実に基づき、最終的な提言を行っている。



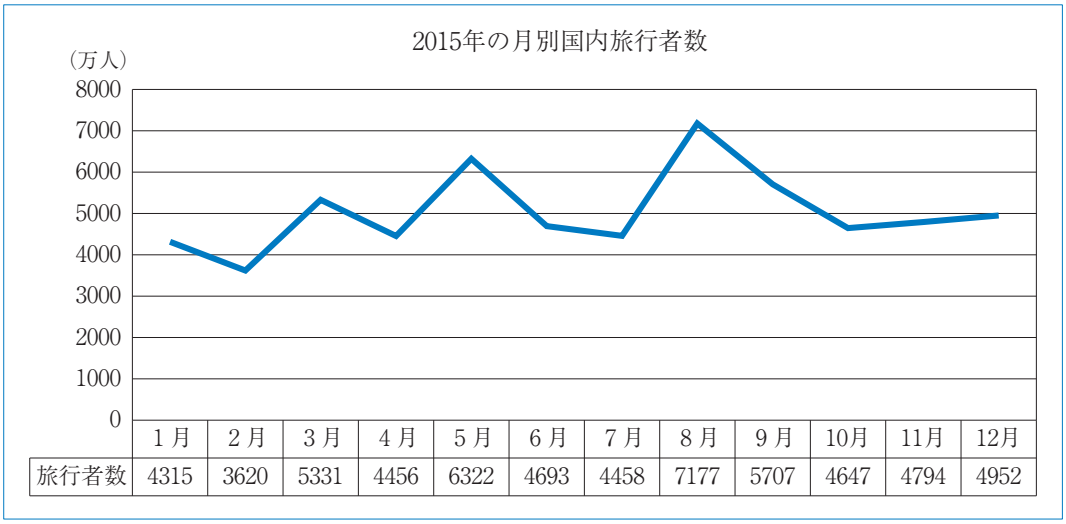
引用：スーパーグローバルハイスクール
平成27年度第２回 SGH 連絡会分科会資料
「課題解決に向けた PPDAC メソッド活用」
資料１より抜粋
http://www.sghc.jp/wp/wp-content/uploads/2015/12/2.pdf

本書は、このような PPDAC メソッドを軸にして、具体的な課題発見と問題解決のスキルが身につくように構成されているので、実際に、探究する気持ちで学習してみよう。

第2部

統計的探究の実践 Ⅰ
～データから有用な情報を引き出す～

1　夏の避暑地の気候の特徴～夏の避暑地が快適な理由は？［データの整理］



	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
旅行者数	4315	3620	5331	4456	6322	4693	4458	7177	5707	4647	4794	4952

観光庁「2015 年旅行・観光消費動向調査」

日本への外国人旅行者はこの２、３年に急増しているが、日本人の国内旅行者の動向を月別に見ると、上の図から、月毎に変動していることが分かる。

Q１：上のグラフは2015年の国内旅行者数の推移を表している。８月が突出して多いのは何故だろう？

STEP 1：Problem 問題　課題の設定

◇ 東京都に比べて、夏の避暑地は本当に過ごしやすいか？

日本では、夏に避暑地を訪れることを好む人が多いが、避暑地にはどのような特徴があるのだろうか？都内の高校の休み時間に、航平、理恵、公介、馨、健三の５人は、それぞれの夏休みについて、次のように話していた。

航平：軽井沢は東京に比べて、過ごしやすかったよ
理恵：東京も今年は涼しい日もあったけど、すごく暑い日が多かったわ
公介：熊谷の祖母の家にいたけど、東京以上に暑かった
馨　：沖縄は暑かったけど、慣れてしまえば逆に過ごしやすかったわ
健三：札幌は過ごしやすかったけど、大阪に行ったときは東京と同じように暑かったな

それぞれの場所で、本当に暑さに違いはあったのだろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 夏の暑さを調べる指標を探る

気象庁の HP には、夏の暑さを調べる指標として、１日の（A）平均気温、（B）最高気温、（C）最低気温の３つが掲載されている。

国土交通省 気象庁 Japan Meteorological Agency

本文へ | ENGLISH | ご意見・ご感想 | サイトマップ

キーワードを入力し検索ボタンを押下ください。 検索

ホーム | 防災情報 | 各種データ・資料 | 知識・解説 | 気象庁について | 案内・申請

ホーム > 各種データ・資料 > 過去の気象データ検索 > 日ごとの値

日ごとの値

一覧表 | グラフ | 見出しの固定 | メニューに戻る

主な要素 | 詳細（気圧・降水量） | 詳細（気温・蒸気圧・湿度） | 詳細（風） | 詳細（日照・雪・その他）

前年 | 前月 | 前日 | 翌日 | 翌月 | 翌年 | 月ごとの値 | 旬ごとの値 | 半旬ごとの値 | 日ごとの値

東京 2015年8月（日ごとの値） 主な要素

日	気圧(hPa)		降水量(mm)			気温(℃)			湿度(%)		風向・風速(m/s)					日照時間(h)	雪(cm)		天気概況	
	現地	海面									平均	最大風速		最大瞬間風速			降雪	最深積雪	昼	夜
	平均	平均	合計	最大 1時間	最大 10分間	平均	最高	最低	平均	最小	風速	風速	風向	風速	風向		合計	値	(06:00-18:00)	(18:00-翌日06:00)
1	1007.6	1010.3	--	--	--	30.5	35.3	26.6	76	50	2.6	5.1	南	8.5	南	8.9	--	--	晴	晴一時曇
2	1008.2	1011.9	0.0	0.0	0.0	30.2	35.1	26.3	71	54	2.6	5.6	南南東	9.3	南東	9.6	--	--	晴	曇時々晴
3	1008.7	1011.4	--	--	--	29.8	35.0	26.1	69	52	3.1	6.7	南	10.8	南	12.0	--	--	晴	晴
4	1007.4	1010.1	--	--	--	30.0	35.1	26.5	69	53	3.4	6.6	南南東	10.5	南南東	11.4	--	--	晴	晴
5	1007.4	1010.1	--	--	--	30.2	35.2	25.7	69	47	4.1	8.2	南南東	13.1	南南東	12.8	--	--	快晴	快晴

資料　気象庁 HP「東京　2015年８月（日ごとの値）主な要素」からの抜粋

Q２：気温について調べる際、（A）、（B）、（C）のどのデータを選ぶのが適当か？

航平君たちは、夏の暑さが関係しているのではと考え、１日の最高気温のデータを収集することに決めた。
それでは、どのようにデータを集めれば良いだろうか？

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 夏の暑さを調べるため、気象観測データを活用しよう

気象庁の HP から、2015年 8 月の東京（東京都）、軽井沢（長野県）、熊谷（埼玉県）、石垣島（沖縄県）、札幌（北海道）、大阪（大阪府）の 6 地点の 1 日の最高気温のデータをダウンロードし、表に低→高の順に整理した。

表 1　2015 年 8 月の 1 日の最高気温のデータ（31日間の昇順；℃）

順位	東京	軽井沢	熊谷	石垣島	札幌	大阪
①		15.9	20.7	28.7	21.7	26.6
②		16.9	21.3	28.8	22.1	27.4
③		17.1	22.6	29.2	22.5	28.1
④		17.8	23.4	30.5	23.0	29.2
⑤		20.0	24.3	30.6	23.3	30.1
⑥		20.3	24.6	30.6	23.5	30.9
⑦		20.4	26.0	30.8	23.6	31.5
⑧		21.6	27.0	30.9	24.2	31.6
⑨		21.8	27.5	30.9	24.9	31.7
⑩		23.3	28.1	31.1	25.0	32.1
⑪		23.7	28.8	31.2	25.4	32.2
⑫		23.8	29.4	31.3	25.4	32.2
⑬		24.3	31.0	31.4	26.0	32.3
⑭		24.5	32.6	31.4	26.3	32.3
⑮		26.6	33.2	31.4	26.4	32.6
⑯		27.0	33.8	31.5	26.5	33.1
⑰		27.1	33.9	31.6	26.7	33.1
⑱		27.3	34.0	31.7	26.9	33.1
⑲		27.4	34.1	31.7	27.3	33.5
⑳		27.4	34.2	32.0	27.5	34.3
㉑		27.8	34.4	32.1	27.7	34.7
㉒		28.3	34.4	32.2	27.8	35.1
㉓		28.5	34.6	32.7	27.8	36.2
㉔		29.8	35.7	32.8	27.8	36.3
㉕		30.0	36.6	32.8	28.2	36.3
㉖		30.1	37.5	33.0	28.2	36.4
㉗		30.2	37.5	33.1	28.7	36.4
㉘		30.2	38.0	33.3	29.0	36.5
㉙		30.3	38.2	33.3	29.5	36.7
㉚		30.4	38.3	33.4	31.9	37.5
㉛		31.1	38.6	33.6	34.5	38.0

表 2　東京の各日の最高気温

2015年 8 月	東京
1 日	35.3
2 日	35.1
3 日	35.0
4 日	35.1
5 日	35.2
6 日	35.9
7 日	37.7
8 日	32.6
9 日	33.4
10日	31.9
11日	35.5
12日	33.7
13日	30.5
14日	31.8
15日	33.1
16日	31.9
17日	28.0
18日	31.9
19日	31.4
20日	27.0
21日	29.4
22日	32.7
23日	31.4
24日	29.3
25日	22.9
26日	21.3
27日	27.3
28日	22.9
29日	21.0
30日	22.5
31日	24.1

資料：気象庁ホームページ「過去の気象データ・ダウンロード」
http://www.data.jma.go.jp/gmd/risk/obsdl/index.php

Q３：表 2 に示す、東京の最高気温のデータを昇順に並べ直し、表 1 を完成しなさい。

データを収集する際には、「データを小さい順に並べ替える」、「必要に応じて36000人→3.6万人とみなす」等の分析しやすい工夫をしておくと負担が減って楽だよ。データを順に並べたものを順序データといい、そこから**中央値**や**四分位数**が容易に求められるよ。

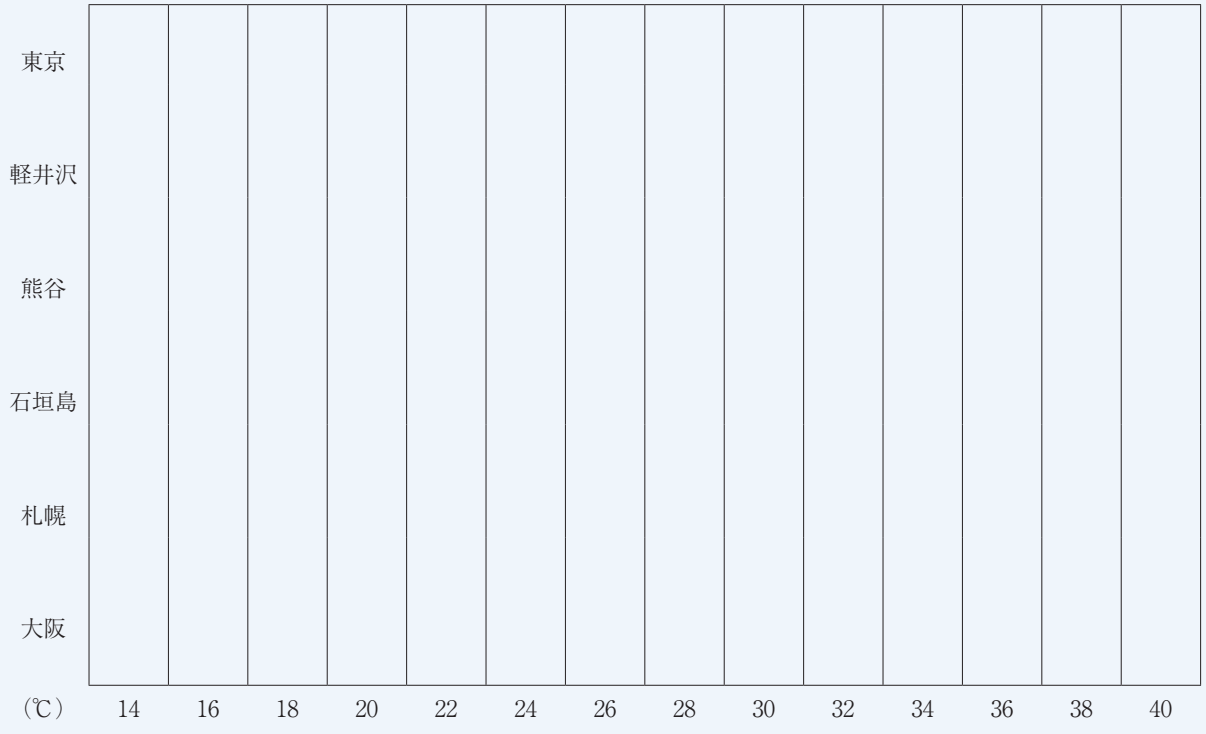
STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ 複数のデータの傾向を捉えるためには、並行箱ひげ図を活用

Q４：表１のデータをもとに、**五数要約（最大値、最小値、四分位数）**と**四分位範囲**をそれぞれ求め、**並行箱ひげ図**を作成しなさい。

表３　６地点の１日の最高気温の五数要約および四分位範囲

	東京	軽井沢	熊谷	石垣島	札幌	大阪
最大値						
第３四分位数						
第２四分位数（中央値）						
第１四分位数						
最小値						
四分位範囲						



２つのデータのバラツキは、**ヒストグラム**や度数折れ線を用いても比較できるが、複数（３つ以上）のデータのバラツキは、並行箱ひげ図を用いて比較するのが良いね。

Q５：作成した並行箱ひげ図と表１のデータに基づいて、各地点の特徴について分かったことを、次の観点からまとめなさい。

① 東京や大阪のような大都市は、避暑地と比べて暑い日が多いかどうか。

② 避暑地として人気の高い軽井沢は高原にあるが、北海道とどう違うのか。

③ 熊谷や沖縄は暑い地域として有名だが、それぞれで違いはあるか。

避暑地の特徴：

STEP 5：Conclusion 結論　結論を導く

◇ 夏の避暑地は、なぜ過ごしやすいのか？

Q６：軽井沢や札幌は夏の避暑地として人気が高いことで知られている。その理由をまとめなさい。

理由：

夏の蒸し暑さを表す指標：不快指数

平均気温と平均湿度から夏の蒸し暑さを数量的に表す指標として、不快指数が知られている。気温を t、湿度を H とすると

不快指数＝0.81t＋0.01H(0.99t－14.3)＋46.3　で求められる。不快指数が75になると人口の約１割が不快を感じ、85になると全員が不快になる。（三省堂編集所（1988)、『大辞林』p.2096、三省堂）

たとえば、沖縄県那覇市の2015年８月は平均気温が29.5℃、平均湿度が74% です。したがって、不快指数は81.2と算出され、次の表を参考に判断すると、不快指数は高いため、多くの人が不快（蒸し暑い）と感じる気候であると判断できます。

表１　不快指数と不快と感じる人数の目安

不快指数	70～75	75～80	80～
程度	１割の人が不快	半数の人が不快	全員が不快

資料：新村出（2008）『広辞苑第６版』p.2430、岩波書店

STEP 3 に示した６地点の不快指数を求めると、表２のようになる。

沖縄は、亜熱帯気候で知られているだけあって不快指数は高く、人気の高い避暑地である軽井沢や札幌は、不快指数は安定して低く、快適な気候であったと判断できる。

2017年において、高校の数学Ⅰでは、データのバラツキ（散らばり）を表す指標として、分散や標準偏差といった要約統計量を学習する。そして、これらを算出することで、データのバラツキ（散らばり）度合を判断できる。

表２　2015年８月１日から31日までの不快指数

８月	東京	軽井沢	熊谷	石垣島	札幌	大阪
１日	83.1	73.0	82.5	81.2	72.0	83.0
２日	81.8	69.6	79.7	81.5	71.6	82.8
３日	80.9	69.6	79.6	81.0	72.7	82.0
４日	81.2	70.3	81.8	80.7	76.4	82.2
５日	81.5	70.9	81.4	81.2	77.6	82.6
６日	82.4	71.1	80.4	82.3	75.1	82.0
７日	82.1	71.8	81.7	80.9	72.7	82.7
８日	77.9	70.9	78.6	80.5	71.0	81.7
９日	78.0	70.1	78.4	82.1	72.4	81.5
10日	79.4	70.2	79.4	81.8	75.4	81.2
11日	80.5	70.2	80.2	82.3	73.7	79.9
12日	80.2	70.5	79.7	82.7	72.7	79.2
13日	79.2	68.4	78.1	82.9	72.5	78.5
14日	78.4	69.2	76.0	82.9	72.2	78.7
15日	78.7	69.5	78.8	83.0	71.3	78.9
16日	79.0	67.9	77.6	83.4	71.6	77.5
17日	77.8	66.6	76.0	83.1	72.0	77.3
18日	79.5	70.6	79.4	82.0	64.6	77.8
19日	78.1	68.7	77.7	82.5	66.0	76.4
20日	77.2	67.7	75.7	82.8	67.5	77.5
21日	77.4	66.5	76.3	82.8	69.1	78.5
22日	80.2	71.3	79.9	81.6	71.3	78.8
23日	77.0	66.6	75.9	80.2	67.9	76.9
24日	72.5	64.2	72.7	81.2	66.7	77.0
25日	68.5	56.5	68.1	79.9	65.0	76.7
26日	67.0	58.5	66.3	80.4	65.2	76.9
27日	72.8	64.6	73.1	80.4	66.3	76.6
28日	69.9	62.5	70.7	81.2	65.5	77.0
29日	67.9	60.6	68.9	81.2	65.8	77.0
30日	69.9	60.8	69.1	81.2	66.1	75.8
31日	71.3	62.8	71.4	81.0	69.3	74.5

〔本節の解答〕

Q１：ゴールデンウィークや夏休み等、長期休みに国内旅行に出かける人が多いから。

Q２：（B）最高気温

理由：一番暑かったときの気温で、その日の暑さの印象が決まるから。

Q３：2015年８月の最高気温（℃）；昇順

順位	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳	㉑	㉒	㉓	㉔	㉕	㉖	㉗	㉙	㉚	㉛	㉛
東京	21.0	21.3	22.5	22.9	22.9	24.1	27.0	27.3	28.0	29.3	29.4	30.5	31.4	31.4	31.8	31.9	31.9	31.9	32.6	32.7	33.1	33.4	33.7	35.0	35.1	35.1	35.2	35.3	35.5	35.9	37.7

Q４：箱ひげ図は略

表３　６地点の１日の最高気温の五数要約および四分位範囲

	東京	軽井沢	熊谷	石垣島	札幌	大阪
最大値	37.7	31.1	38.6	33.6	34.5	38.0
第３四分位数	35.0	29.8	35.7	32.8	27.8	36.3
第２四分位数（中央値）	31.9	27.0	33.8	31.5	26.5	33.1
第１四分位数	27.3	21.6	27.0	30.9	24.2	31.6
最小値	21.0	15.9	20.7	28.7	21.7	26.6
四分位範囲	7.7	8.2	8.7	1.9	3.6	4.7

Q５：

①東京・大阪 vs 軽井沢・札幌

（例）　人気の高い避暑地として知られる軽井沢や札幌は、東京や大阪と比べて、各日の最高気温が低いため、多くの観光客が訪れている。

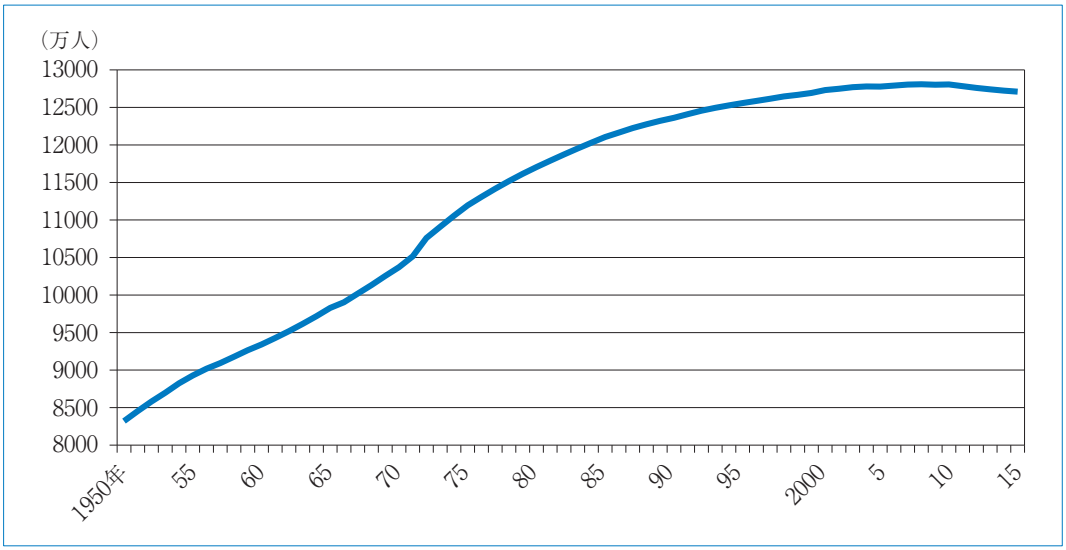
②軽井沢 vs 札幌

（例）　軽井沢と札幌のいずれも涼しいが、札幌の方が四分位範囲が狭いため、安定して涼しい。軽井沢は最高気温が20℃を下回る日もあるため、少し涼しすぎると感じた人がいる可能性もある。

Q６：略

2　地域の豊かさの格差は拡大しているか？［特性値の活用］

図１　日本の人口の推移



資料：総務省「国勢調査」、「人口推計　各年10月１日の推計人口」

日本の人口が1920年の国勢調査開始以来、初めて減少したことが、2015年国勢調査（総務省統計局）で明らかになった。５年前に比べて人口が減少したのは39の道府県にのぼる。東京を中心とした大都市に人口が集中する一方、地方の人口減少は大都市部から離れた県ではかなり以前から始まっている。第６部で取り上げる島根県では、1955年の92万９千人をピークとして2015年には69万４千人まで人口が１／４減少した。

STEP 1：Problem 問題　課題の設定

◇ 経済的な豊かさは地域間で拡大している？

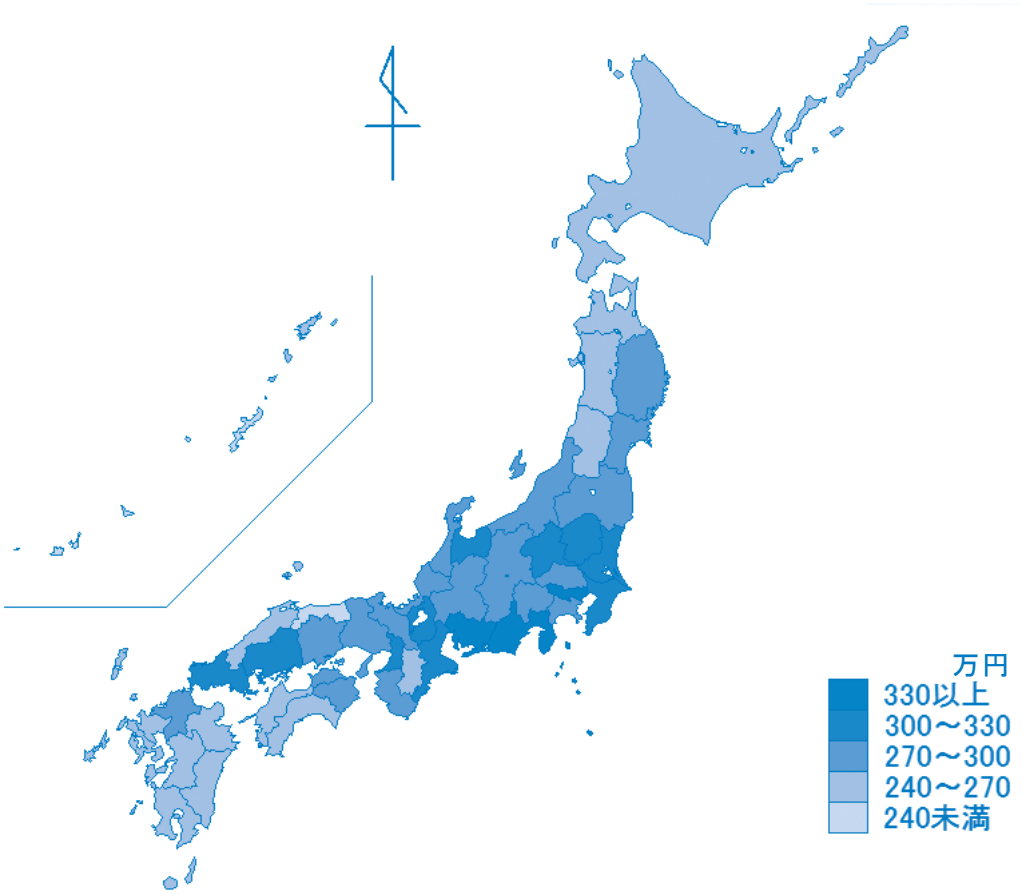
最近、世界各国で貧富の差が拡大していると言われている。我が国では、人口減少時代に入って、地方では人通りが少なく、かつての繁華街がシャッター通りとなっているところも増えている。人口は経済・社会の基盤を成すものであり、人口の増減は経済的な豊かさと密接に関わっていると言われる。近年の人口変動によって、地域間で経済的な豊かさの格差は拡大したのであろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ どのような指標で豊かさを捉えたら良いか？

地域ごとの経済的な豊かさを捉える指標を都道府県別の１人当たり県民所得とする。県民所得は企業を含めて県民全体の経済水準を表すもので、内閣府が推計する国民所得に準拠して作成されているので、都道府県間で比較可能な統計データである。各都道府県は人口規模が大きく異なるので、県民１人当たりの所得とすれば、地域の経済的な豊かさを捉える指標として適当であろう。

１人当たり県民所得（2013年度）



資料：内閣府「平成25年度　県民経済計算」

格差の大きさを１人当たり県民所得の都道府県間のバラツキで評価することとし、バラツキがこの40年間で拡大しているか否かで、地域ごとの経済的な豊かさに不均等が生じているかを明らかにする。

国民所得：国民全体が得る所得の総額をいう。個人や企業の所得は経済活動によって産み出された付加価値総額の配分であるので、経済活動の規模を表す指標である。
県民所得：国民所得の県民版

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ １人当たり県民所得の時系列データを47都道府県について収集しよう

１人当たり県民所得は内閣府「県民経済計算」から利用することができる。なお、人口については、我が国のすべての人を対象にして、５年に１回、「国勢調査」（総務省）が実施されているが、毎年の１人当たり県民所得のベースとなる人口は、「10 月１日現在推計人口」（総務省）に拠っている。表１に2013年度の１人当たり県民所得について、47都道府県ごとにその平均を示す。

表１　１人当たり県民所得（2013年度；万円）

01	北海道	255	17	石川県	297	33	岡山県	280
02	青森県	243	18	福井県	285	34	広島県	306
03	岩手県	270	19	山梨県	292	35	山口県	312
04	宮城県	286	20	長野県	271	36	徳島県	288
05	秋田県	246	21	岐阜県	273	37	香川県	280
06	山形県	263	22	静岡県	333	38	愛媛県	254
07	福島県	279	23	愛知県	358	39	高知県	245
08	茨城県	314	24	三重県	317	40	福岡県	283
09	栃木県	325	25	滋賀県	327	41	佐賀県	251
10	群馬県	305	26	京都府	297	42	長崎県	242
11	埼玉県	286	27	大阪府	300	43	熊本県	242
12	千葉県	302	28	兵庫県	282	44	大分県	256
13	東京都	451	29	奈良県	253	45	宮崎県	241
14	神奈川県	297	30	和歌山県	282	46	鹿児島県	240
15	新潟県	277	31	鳥取県	234	47	沖縄県	210
16	富山県	316	32	島根県	242		全県平均	307

資料：内閣府「県民経済計算」

表１の01～47の符号は1970年に統計に用いる標準地域コードとして定められたもので、各都道府県に概ね北東から南西に順に割り振られているよ！

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ 格差の変動を捉えるために、バラツキ（散らばり）の指標を活用する

1 人当たりの県民所得を Y_i（$i=1, 2, \cdots, 47$）として、1 人当たりの県民所得のバラツキの指標として、**標準偏差** $s=\sqrt{\sum_{i=1}^{47}(Y_i-\overline{Y})^2/47}$（$\overline{Y}$ は平均で $\overline{Y}=\sum_{i=1}^{47}Y_i/47$）を利用する。標準偏差は**分散** s^2 の平方根である。

図 2 は、1975～2013年度の 1 人当たり県民所得の標準偏差を示している。表 1 に示されている2013年度のデータと同様の各年度データから計算されている。

図 2　1 人当たり県民所得の標準偏差の推移

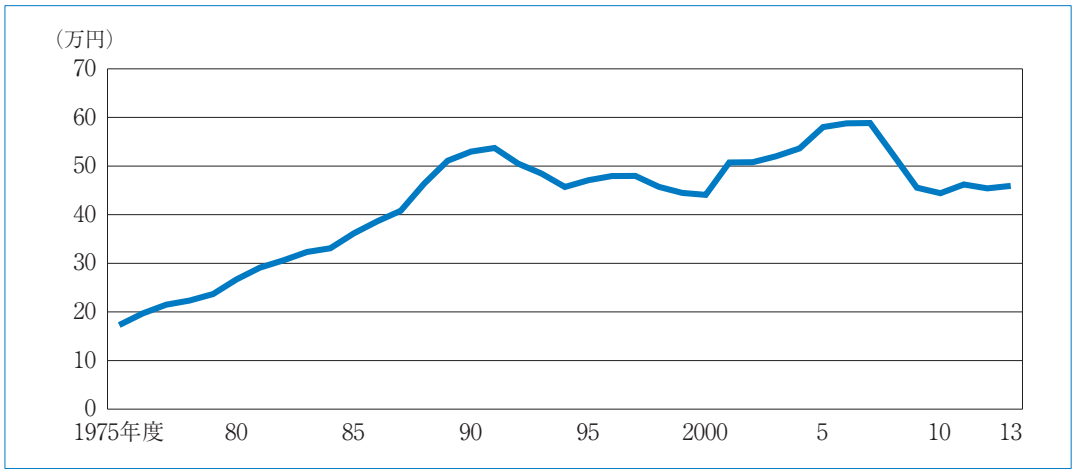
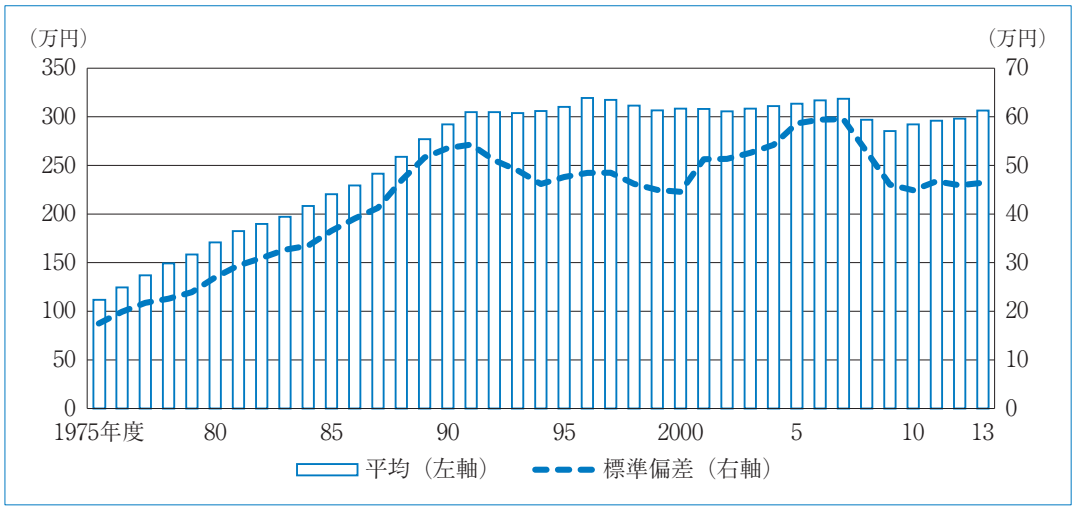


図 2 から、1975年～91年度のバラツキ（標準偏差）は一貫して増加しており、格差は拡大しているように見える。図 2 に各年度の 1 人当たり県民所得の平均を追加したのが図 3 である。

図 3　1 人当たり県民所得の平均と標準偏差の推移



1975年～91年度の期間は標準偏差が大きくなっているが、平均も同様に増加している。経済成長に伴って、所得が増加した結果、標準偏差も大きくなったことに注意しなければならない。所得水準が異なる時点で相違すれば、各時点の所得のバラツキを評価する際、各時点の所得水準を調整した指標に基づかないと適切な比較ができないことを理解しよう。標準偏差は平均からのバラツキの度合いを求めるもので、その大きさは対象とするデータの平均で代表される水準によって異なった値となる。仮に、Y_i を円単位と万円単位の２種類の数値で標準偏差 s を計算すると、前者の s の値は後者の s の値の１万倍になることを容易に確認できるであろう。

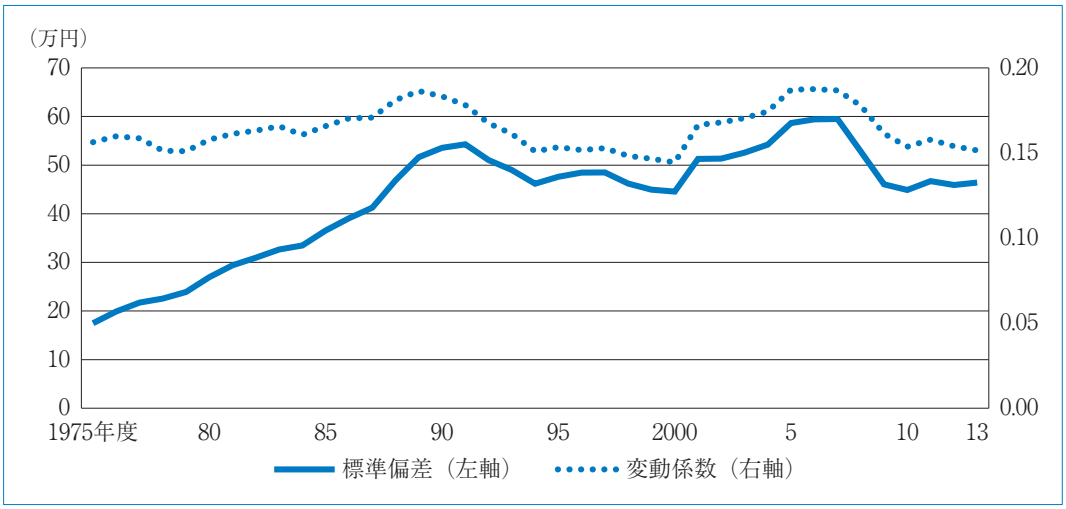
Q１：2013年度の１人当たり県民所得が万円単位で与えられたとき、その標準偏差は45.9である。１人当たり県民所得が円単位で示されたとき、その標準偏差はどうなるか？

１人当たり県民所得の標準偏差＝

バラツキを評価するのが同じデータ項目であっても、異なる時点や異なる属性のグループのバラツキを相互に比較する際、標準偏差を平均で割って求める変動係数が広く用いられている。**変動係数**は分母となる平均で水準を調整しているので、これを用いれば、より的確に１人当たり県民所得の各時点のバラツキを相互に比較することができる。

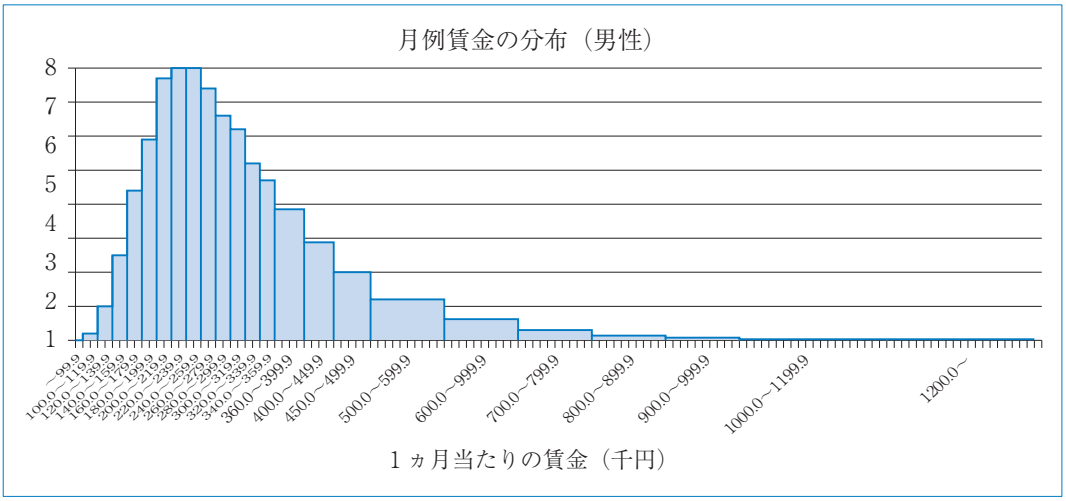
図４は、１人当たり県民所得の標準偏差と併せて変動係数を図示している。標準偏差が大きくなった1975～91年度の期間の変動係数はそれほど大きくなっていないことが分かる。変動係数の値は時期によって相違するが、2013年度は0.15であり、1975年度の0.16と比べてほぼ同様な水準にある。

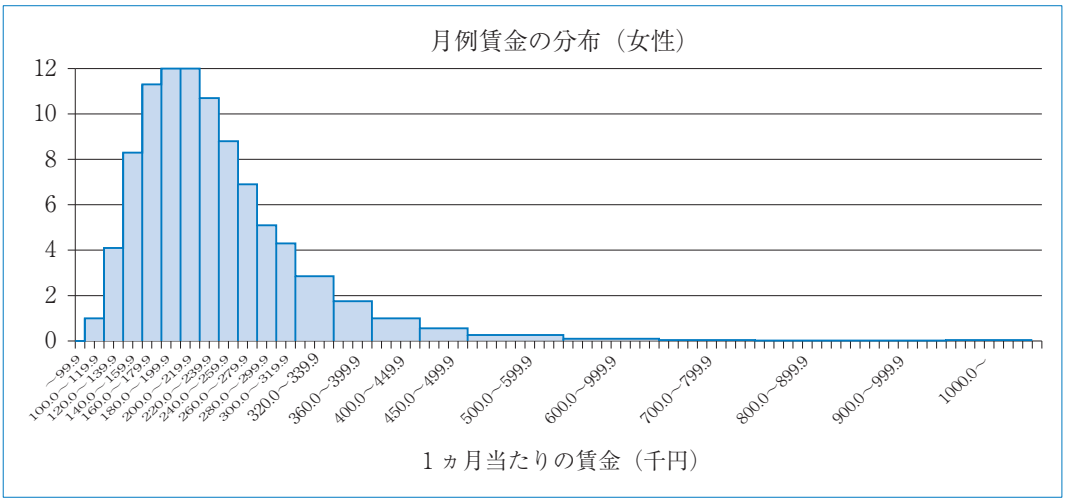
図４　１人当たり県民所得の標準偏差と変動係数



歪度（skewness）と尖度（kurtosis）

図５　賃金分布





資料：厚生労働省「平成27年賃金構造基本調査」

図５は2015年の男女別の賃金分布をヒストグラムで図示している。男性と女性のいずれの賃金分布も右裾が長くなっていて左右対称ではなく、平均が中央値より大きく乖離している。このような分布の歪みの程度を示す統計量が**歪度**（わいど）であり、$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$ で与えられる（$\bar{x}$ は平均、s は標準偏差）。分布が右に歪んでいるとき、歪度は算式の３乗項によって右裾にある少数のデータの影響を大きく受けて正の値をとる。他方、左に歪んでいる分布のとき、同様の理由で歪度は負の値をとる。分布が対称のとき、歪度は０となるので、歪度は分布の対称性を知る特性値となる。

その他、分布の尖り具合（裾の厚さ）の程度を示す統計量が**尖度**（せんど）であり、$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4$ で与えられる。第４部で紹介する正規分布の場合、尖度は３となるので、尖度が３より大きい分布のとき、正規分布より分布の裾が厚い、あるいは平均から大きく離れた値が多いと判断できる。また、外れ値の存在を知る指標ともなりうる。

このように、平均、分散（標準偏差）、歪度、尖度等の特性値を求めれば、分布の形状をある程度知ることができることが理解されるであろう。

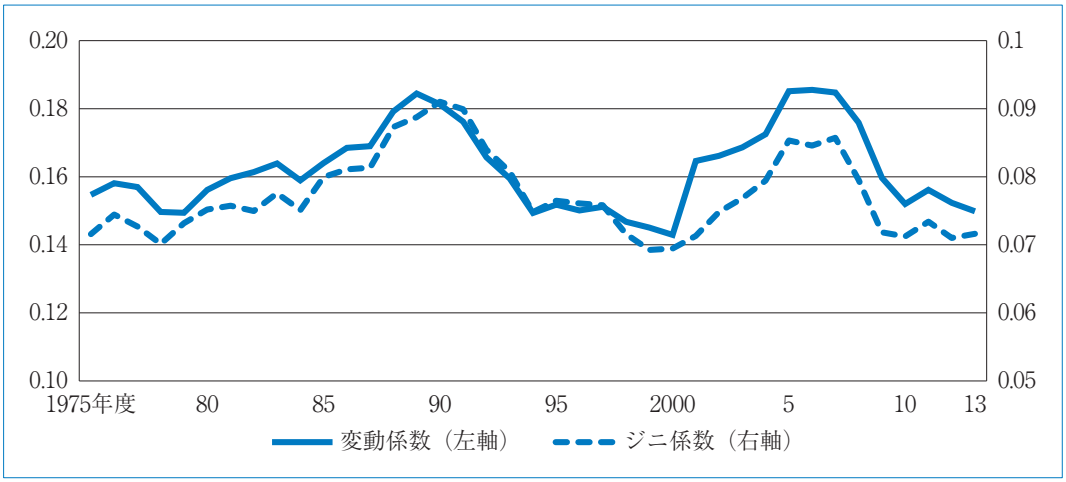
近年、所得の不平等を示す指標として、**ジニ係数**が広く一般に知られるようになってきており、昨今の国会論戦やニュースの中でもたびたび登場する。

ジニ係数（G）は、1936年にイタリアの数理統計学者のコッラド・ジニ（Corrado Gini）が社会における所得分配の不平等さを測る指標として考案したもので、

$G=\Sigma\Sigma|Y_i-Y_j|/2n^2\overline{Y}$として求められる。$0 \leqq G \leqq 1$であり、Gが０のとき完全平等を示し、１に近づくに従って不平等の程度が大きくなる。

図６にジニ係数と変動係数を併せて示す。

図６　１人当たり県民所得の変動係数とジニ係数



２つの指標はほぼ同じ変動を示しており、いずれに基づいても2013年度に至る約40年で都道府県間の所得格差が拡大したことは示していない。

ジニ係数とローレンツ曲線

所得分布の状況を図示したのが**ローレンツ曲線**で、米国の統計学者のマックス・ローレンツ（Max Lorenz）が1905年に考案しました。横軸に所得額の大きさの順に所得人数の累積百分比を、縦軸にそれに対応した所得金額の累積百分比をとって得られる曲線をいいます。所得分布が完全に平等であれば、曲線は対角線に一致し（均等分布線）、そこから下方に位置すればするほど不平等度が大きいという性質をもつので、所得分布の不平等度を測定することができます。

２つのグループの所得分布の不平等度を比較するとき、ローレンツ曲線が交わってしまうと、どちらがより不平等であるかが判断しづらくなります。ローレンツ曲線に基づいて考案されたのがジニ係数です。ジニ係数はローレンツ曲線と均等分布線によって囲まれた面積の正方形の面積に対する割合を２倍した値として求められます。均等分布線からの乖離が大きいほどジニ係数の値は大きくなるので、複数のグループの不平等度を数量的に比較できます。

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ 地域間格差は拡大しているか？

変動係数やジニ係数の約40年間の推移を見る限り、地域間で経済的な豊かさの格差が拡大したとはいえない。ただし、1980年代後半のバブル期に至る時期にかけてやや格差が拡大した。その後の景気低迷の中で、景気対策としての地方への公共事業の重点配分などの結果、地域間格差は低下した。景気対策の後遺症として、国と地方の財政赤字は膨れ上がり、小泉政権下での公共事業抑制や行政改革への転換に伴って、地域間格差は拡大した。その後の政策転換と2008年９月のリーマンショック後の世界不況のもとで地域間格差は縮小し、現在に至っている。バブル崩壊後やリーマンショック後の景気の大きな後退局面では、地域間格差が縮小している。

Q２：なぜ、景気後退期に地域間格差は縮小しているのだろうか？

〔本節の解答〕

Q１：１人当たり県民所得の標準偏差＝459,000

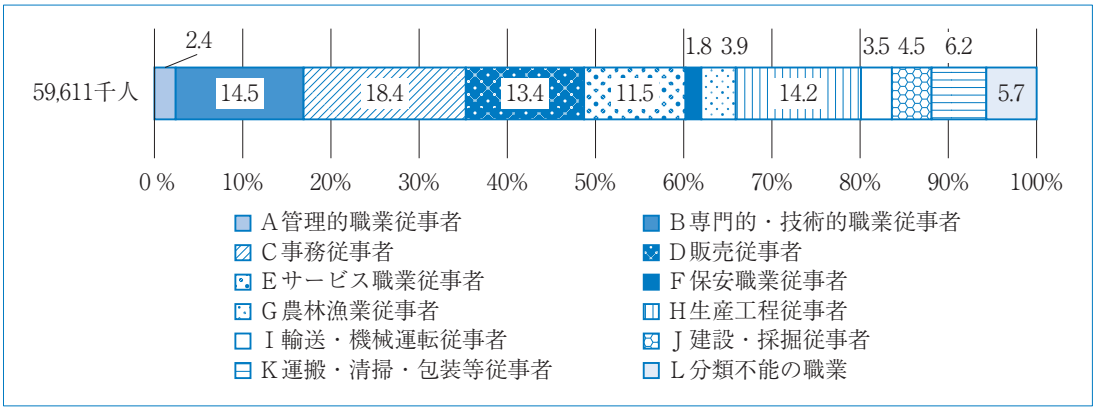
Q２：大都市圏から離れた地域では、地域経済のなかで農林水産業、建設業、公務の占める役割が大きいところが多い。

3　サービス経済化の状況とその背景を探る［関係の度合］

高校生の航平君は将来の進路を考えている。

「僕の父はIT企業、母は病院に勤務している。公介君の父は学校の先生、母は流通企業、健三君の父は金融業、……のように、周りの友達の親の勤務先はサービス業が多い。ところが、祖父は日本の高度経済成長が始まるころ就職したそうで、製造業が花形産業で友達の多くもそこで働いていたという。」

図１　就業者の職業構成



資料：総務省「平成22年国勢調査」

職業分類と産業分類

同じような仕事を一まとめにして統計データを有効に利用するために、職業分類が設けられています。図１の「就業者の職業構成」で表示されている職業は、総務省が定めている「**日本標準職業分類**」の大分類項目です。報酬を伴う仕事に就いている人の職業を大括りにしたもので、たとえば、「管理的職業従事者」には経営者や管理者等が分類されます。「専門的・技術的職業従事者」には学者、技術者、教員、医者、弁護士等が分類されます。それぞれの大分類の下に中分類、小分類が設けられていて、その分類項目からより詳細な職業に関するデータが利用できます。

同じような事業活動を行っている工場や事務所、営業所、店舗等を一まとめにしたのが産業です。職業分類と同様に、総務省が統計データを産業ごとに表すために「**日本標準産業分類**」を定めています。「農業、林業」、「建設業」、「製造業」、「情報通信業」、「卸売業、小売業」、「金融業、保険業」など、20の大分類項目が設けられていて、その下に、中分類、小分類、細分類があって、細分類は1460の項目数です。

STEP 1：Problem 問題　課題の設定

◇ なぜ近年はサービス業（第 3 次産業）で働いている人が多いのだろう？

サービス経済化といわれて久しいが、いま、どのような状況であるのか、また、その主な要因は何だろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ サービス経済化の状況とその背景を探る

サービス業で仕事をしている人が実際に増加して、全体の中で比率が上昇しているかを確認する。併せて、サービス業の生産活動についても高まっているかを確認する。次いで、サービス経済化を解明するための仮説として広く受け入れられている、**ペティ＝クラークの法則**を統計データに基づいて実証する。

ペティ＝クラークの法則

経済の発展につれて、国民経済に占める第 1 次産業の比重は次第に低下し、第 2 次産業、次いで第 3 次産業の比重が高まるという、産業構造の高度化を説明したものです。自然界から採取する農林業、漁業等の産業を第 1 次産業、その産出物を加工する製造業、建設業等の産業を第 2 次産業、それ以外の産業を第 3 次産業と大別したのは、ケンブリッジ大学の経済学者コーリン・クラーク（Colin Clark）で、統計学の始祖とも称されるウィリアム・ペティ（William Petty）が『政治算術』の中に記した内容を整理して1941年に提示しました。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ サービス経済化を就業者と GDP の産業別構成比から確認しよう！ サービス経済化の要因は？

産業別の就業者数は総務省「労働力調査」から1953年以降のデータが利用可能である。また、産業別の経済活動については、内閣府「国民経済計算」から利用できるが、SNA（国民経済計算）の基準改定によって1990年以降とそれ以前で経済活動の範囲が若干変更していることは留意しておこう。表 1 に1955年以降の産業別就業者数の推移を示す。

表１　産業別就業者数

年次	総数（万人）	第１次産業	第２次産業		第３次産業
			建設業	製造業	
1955	4090	1581	195	757	1557
56	4171	1539	197	805	1630
57	4281	1517	217	853	1694
58	4298	1453	223	898	1724
59	4335	1396	243	896	1800
60	4436	1383	253	946	1854
61	4498	1341	274	1011	1871
62	4556	1308	290	1066	1892
63	4595	1227	290	1108	1968
64	4655	1179	308	1129	2038
65	4730	1142	328	1150	2109
66	4827	1098	350	1178	2201
67	4920	1062	359	1252	2247
68	5002	1015	370	1305	2305
69	5040	970	371	1345	2348
70	5094	906	394	1377	2408
71	5121	834	414	1383	2484
72	5126	771	433	1383	2530
73	5259	718	467	1443	2619
74	5237	689	464	1427	2646
75	5223	677	479	1346	2712
76	5271	661	492	1345	2764
77	5342	653	499	1340	2839
78	5408	648	520	1326	2904
79	5479	625	536	1333	2976
80	5536	588	548	1367	3019
81	5581	567	544	1385	3073
82	5638	558	541	1380	3145
83	5733	541	541	1406	3229
84	5766	520	527	1438	3260
87	5911	497	533	1425	3432
88	6011	481	560	1454	3487
89	6128	470	578	1484	3566
90	6249	457	588	1505	3668
91	6369	433	604	1550	3752
92	6436	417	619	1569	3801
93	6450	389	640	1530	3862
94	6453	379	655	1496	3893
95	6457	373	663	1456	3939
96	6486	362	670	1445	3979
97	6557	357	685	1442	4039
98	6514	349	662	1382	4084
99	6462	341	657	1345	4078
2000	6446	331	653	1321	4102
01	6412	318	632	1284	4133
02	6330	301	618	1202	4158
03	6316	298	604	1178	4176
04	6329	290	584	1150	4236
05	6356	285	568	1142	4284
06	6389	275	560	1163	4320
07	6427	277	554	1170	4352
08	6409	273	541	1151	4370
09	6314	267	522	1082	4380
10	6298	258	504	1060	4411
11	6289	252	502	1049	4431
12	6270	243	503	1032	4430
13	6311	236	499	1039	4445
14	6351	233	505	1040	4474
15	6376	231	500	1035	4509

資料：総務省「労働力調査」

注：総数には就業先不詳を含む。1972年以前の数値には沖縄県は含まれていない。2002年以降の数値は日本標準産業分類改定を踏まえているが、それ以前は製造業とサービス業の間で若干の業種移動がある。

Q１：表１に基づいて、第３次産業就業者の比率を求めよう！

総数には就業先産業が不詳の就業者を含むので、第３次産業の就業者比率は第１次産業、第２次産業、第３次産業の就業者の合計から求めることが必要だよ！

図２に1955年以降の第３次産業就業者の比率の推移を示す。また、図３に第３次産業の総生産（GDP）の構成比の推移を示す。

図２　第３次産業就業率

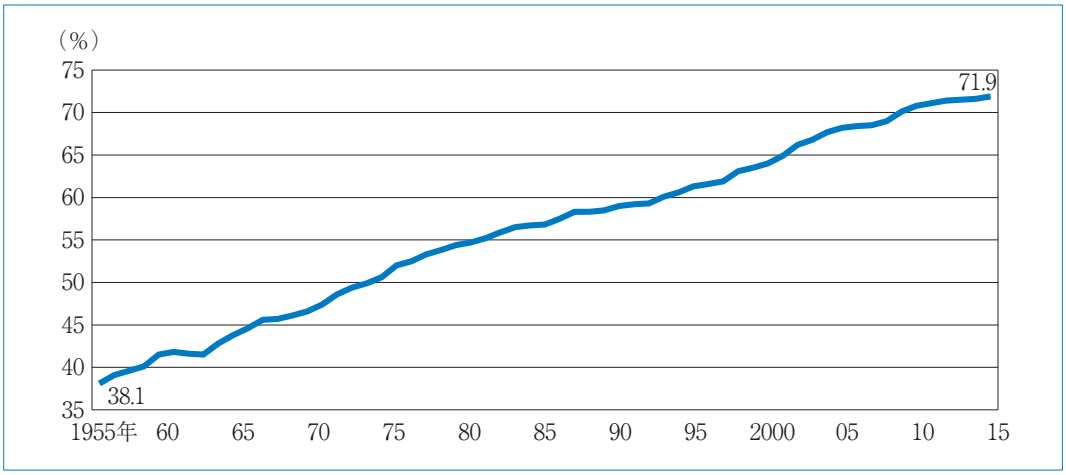
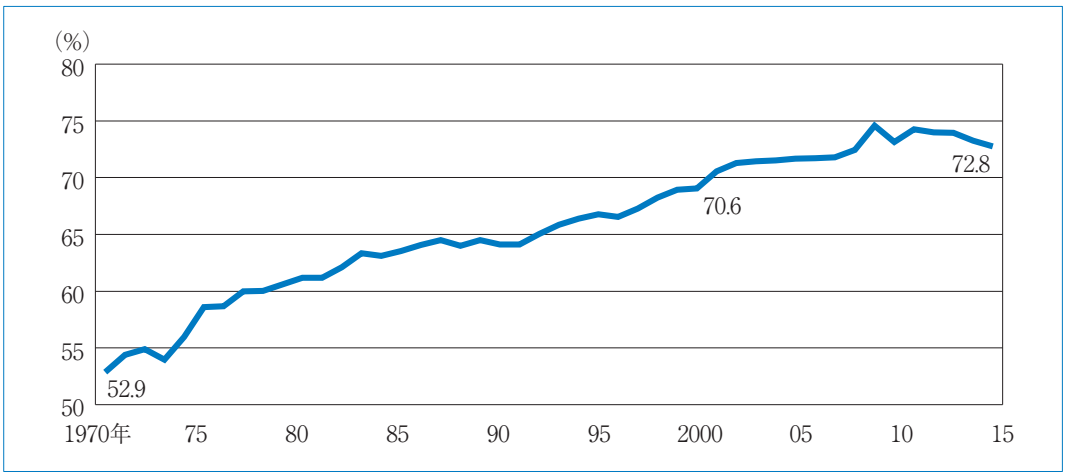


図３　第３次産業の GDP 構成比



資料：内閣府「国民経済計算」、総務省「労働力調査」

Q２：経済の発展を表す指標として、どのようなデータが適切であろうか？

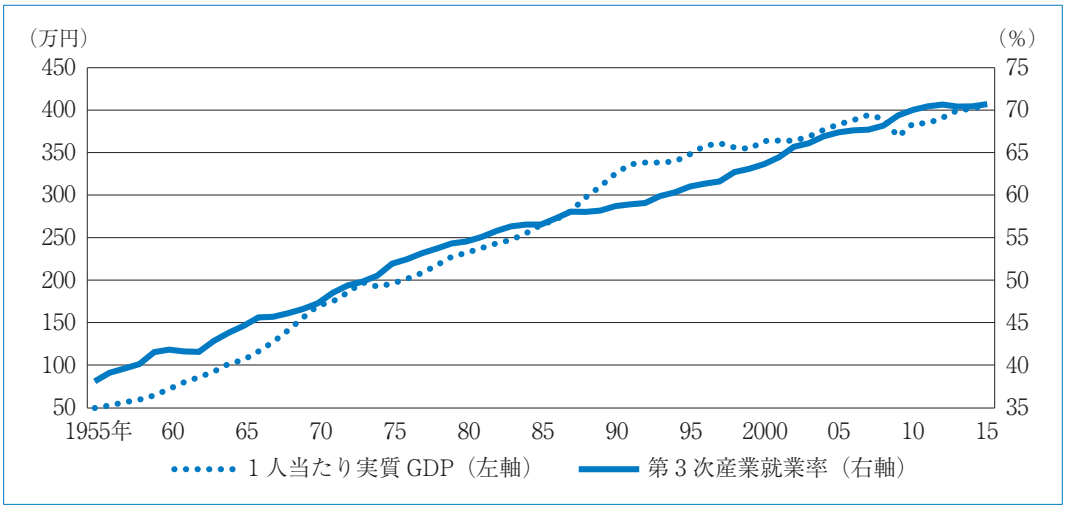
STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ サービス経済化の背景を探るため散布図を活用

第３次産業の就業者数は1955年の1557万人から2015年には4509万人へと約60年間に2.9倍に急増し、全産業に占める比率も38.1％から71.9％へ上昇の一途をたどっている。同様に、GDP についても、第３次産業の構成比は拡大しており、2001年には70.6％と70％を上回り、2015年は72.8％となった。

図４に１人当たり実質 GDP と第３次産業の就業者比率を併わせて図示すると、両者の推移が軌を一にしていることが分かる。

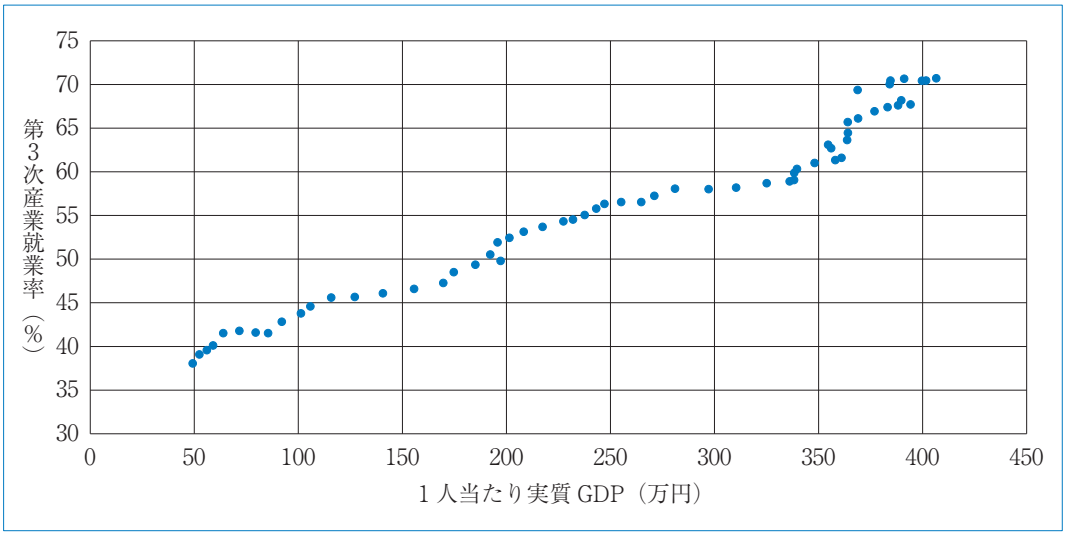
図４　１人当たり実質 GDP と第３次産業就業率の推移



資料：内閣府「国民経済計算」、総務省「労働力調査」

ペティ＝クラークの法則に従って、横軸に国の豊かさを表す指標として物価変動を調整した１人当たり**実質** GDP、縦軸にサービス経済化の指標として第３次産業就業率をとって**散布図**（図５）を描くと、右肩上がりの関係が明確に読み取れる。

図５　１人当たり実質 GDP と第３次産業就業率



1人当たり実質GDPをX、第3次産業就業率をYとしたとき、XとYの関係の度合いを**相関係数**rの値から量的に捉えることができる。XとYのn個のデータが与えられたとき、

rは $\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})/\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}$　$(i=1,2,\cdots,n)$ の式で求められる。

$-1 \leqq r \leqq 1$であり、$r=1$のとき完全に正の相関、$r=-1$のとき完全に負の相関といい、いずれもXとYのすべてのデータが直線上に位置する。また、rが0近傍のとき、XとY の間にはほとんど関係がない。rが± 1に近いほど相関の度合いが高い。

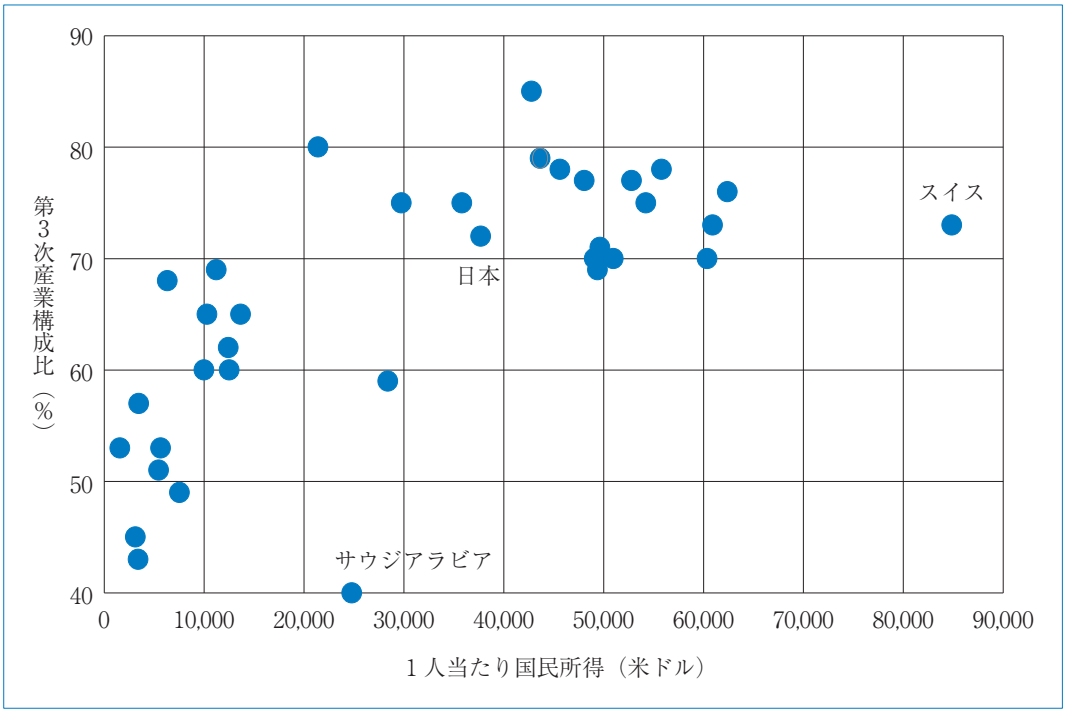
相関係数（correlation coefficient）

相関の概念を提唱し、相関係数rを最初に用いたのはゴールトン（Francis Galton）です。ゴールトンはダーウィンの従兄弟で、ダーウィンの「種の起源」に刺激を受け、親から子への遺伝の法則を探すために、親と子の身長や体重、骨格などを測定し、そこから相関係数の考えに至りました。ただ、ゴールトンは数学が得意ではなかったため、相関係数を定式化したのは弟子のカール・ピアソン（Karl Pearson）で、通常、相関係数といえば、ピアソンの名を冠して、ピアソンの積率相関係数と称されます。他に、分布に関する仮定を置かないノンパラメトリックな相関係数として、スピアマンの順位相関係数、ケンドールの順位相関係数などがあります。

日本において、経済の発展とサービス経済化には強い関係があることが分かったが、ペティ＝クラークの法則を世界各国のデータからも検証できるであろうか？

図6は2014年の各国の1人当たり国民所得（米ドル換算）と第3次産業のGDP構成比を図示している。サウジアラビアを除けば、全体として2つの指標の関連度はかなり高いことを確認できる。

図6　1人当たり国民所得と 第3次産業構成比



資料：日本統計協会「世界の統計」

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ サービス経済化は経済的豊かさの象徴

日本について、長期的な経済発展とともにサービス経済化が進展し、今日の状況にあることをデータに基づいて実証することができた。また、このような関係を示すペティ＝クラークの法則について、特定時点の世界各国の統計データからも、同様に実証することができた。

Q３：サウジアラビアは１人当たり国民所得（米ドル換算）と第３次産業の GDP 構成比についての世界各国で観察される関係から、何故、外れて位置するのだろうか？サウジアラビアを除くと関係の度合いはどのように変わるのだろうか？

〔本節の解答〕

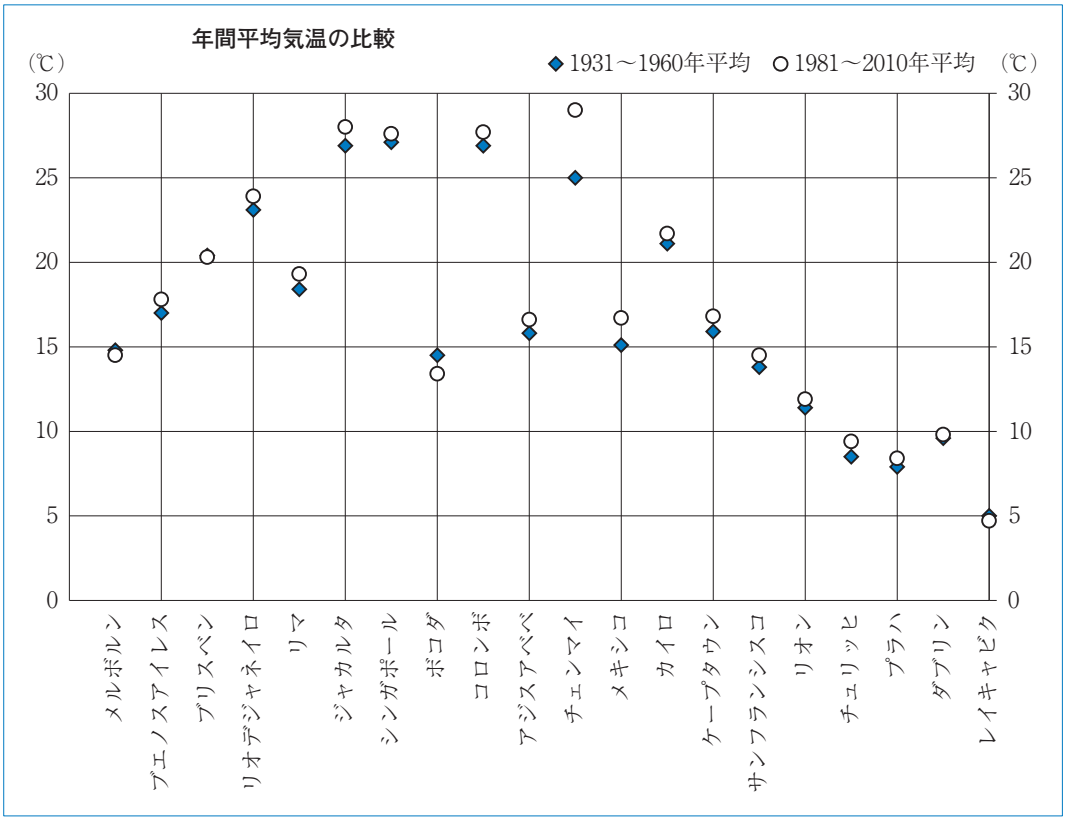
Q１：

1955年	38.1	65	44.6	75	51.9	85	56.5	95	61.0	05	67.4
	39.1		45.6		52.4		57.3		61.3		67.6
	39.6		45.7		53.1		58.1		61.6		67.7
	40.1		46.1		53.7		58.0		62.7		68.2
	41.5		46.6		54.3		58.2		63.1		69.4
60	41.8	70	47.3	80	54.5	90	58.7	2000	63.6	10	70.8
	41.6		48.5		55.1		58.9		64.5		71.1
	41.5		49.4		55.8		59.1		65.7		71.4
	42.8		49.8		56.3		59.9		66.1		71.5
	43.8		50.5		56.5		60.3		66.9		71.6
										15	71.9

Q２：１人当たり国民所得、１人当たり GDP、１人当たり所有資産等、人口１人当たりで求めることが適当である。

Q３：（前半の問いの解答は略）サウジアラビアを含めた相関係数は0.687であるが、サウジアラビアを除くと0.728となり、関係の度合いが高まることが見て取れる。

4　都市の平均気温と緯度はどんな関係？［散布図・相関分析による問題解決］



資料：国立天文台「理科年表」
注：数ヶ国で平均気温の対象年次が表記と異なっている

地球の温暖化の影響なのか、50年前と比べると主要な都市の平均気温は上昇している。ただし、各地域の平均気温の上昇よりも、地域別の平均気温の相違が著しい。地球上のさまざまな地域の年間平均気温は何によって決まっているのだろうか？

STEP 1：Problem 問題　課題の設定

◇ 自然現象に関わっている自然の法則を統計的に考察しよう！

先生：地球温暖化が進行していると言われています。既に避暑地の問題のときに議論したように、信州が、東京より夏に涼しいこともよく分かっています。そもそも私たちが住む地球は、赤道付近は暑いし、極地に近づくほど寒くなることは誰でも知っていますね。ここでは、世界のさまざまな地域の標準的な年平均気温はどのように決まっているのかといった問題を統計的に解決してください。

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 地域の年間平均気温に影響を与える要因を考えてみよう

先生：世界の各地域で年間平均気温は異なっています。各地域での年間平均気温に影響を与える原因にはどのようなものが考えられるか、皆で議論して、**特性要因図**を描いてください。その原因候補の中で、皆さんが比較的簡単にデータをとれるものを考えてみましょう

航平：地域の人間の活動が影響を与えているのではないかな？

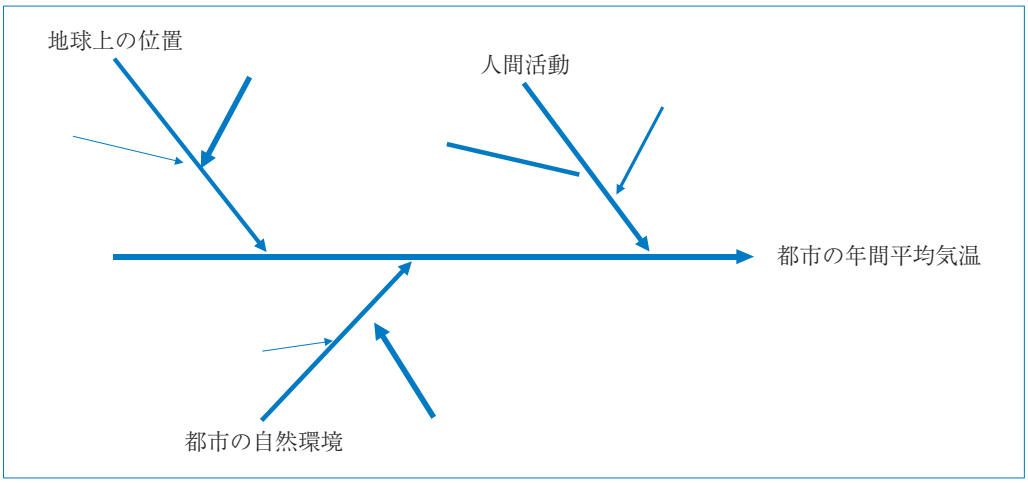
理恵：都市の自然環境も重要でしょう。軽井沢は涼しいって言ってたじゃない

航平：自然環境といっても、いろいろ考えられそうだね

公介：北海道が涼しいのは北にあるからだよね

先生：地域の地球上の位置、都市の自然環境、人間活動が地域の年間平均気温に影響を与える要因（原因候補）と考えられるようだね。それを特性要因図に描くと図 1 のようになります。これは大きな原因候補（大骨）を網羅しただけです。

図 1　特性要因図：都市の年間平均気温に影響を与える要因の網羅



Q 1：図 1 の特性要因図の大きな原因候補に関係し、データをとれる可能性のある具体的な原因候補を特性要因図の大骨上の対応する位置（中骨）に書き込んでください。

航平：大骨とか中骨って何か変な名前ですね

先生：特性要因図は、アメリカでは「魚の骨図（Fish Bone Diagram）」と呼ぶこともあるのです。原因の候補を魚の骨の大きな骨、それを具体化した原因を小さな骨に例えているイメージなのです。

理恵：特性要因図はアメリカで考案されたのですか？

先生：いいえ、日本の石川馨先生という方が考案したものです。海外でも石川ダイアグラムと呼ぶ人もいます。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 世界の都市の平均気温などを集めてみよう

世界各都市の平均気温は、理科年表や気象庁のホームページで調べることができる。航平君たちは、表１のようなデータを理科年表からとってきた。ただし、別途１か所だけ都市ではなく、南極の昭和基地のデータを調べた。

表１　世界の25都市の年間平均気温（℃）、緯度（北緯、南緯はーで表示）、標高（m）

地名	平均気温	緯度	標高	地名	平均気温	緯度	標高
昭和基地	−10.5	−69.00	18	ドーハ	27.0	25.15	11
メルボルン	14.5	−37.39	132	カイロ	21.7	30.06	116
ブエノスアイレス	17.8	−34.35	25	ケープタウン	16.8	33.58	46
ブリスベン	20.3	−27.23	4	東京	15.4	35.42	25
リオデジャネイロ	23.9	−22.55	5	サンフランシスコ	14.5	37.37	6
リマ	19.3	−12.01	12	北京	12.9	39.56	55
ジャカルタ	28.0	−6.11	8	サラエボ	10.4	43.52	630
シンガポール	27.6	1.22	5	リオン	11.9	45.43	197
ボゴダ	13.4	4.42	2547	チュリッヒ	9.4	47.22	555
コロンボ	27.7	6.54	7	プラハ	8.4	50.06	380
アジスアベベ	16.6	9.02	2354	ダブリン	9.8	53.26	68
チェンマイ	29.0	13.00	13	レイキャビク	4.7	64.08	54
メキシコ	16.7	19.24	2309				

資料：国立天文台「理科年表」

STEP 4：Analysis 分析　グラフから緯度と平均気温との関係を捉える

◇ 平均気温と関係があるデータを探ってみよう！

公介：データの間の関係性が強いかどうかは**相関係数**を計算すれば良いと習いましたので計算してみます
先生：それは本当かな？

Q２：緯度と平均気温の相関係数、標高と平均気温の相関係数を求めなさい。

緯度と平均気温の相関係数＝
標高と平均気温の相関係数＝

航平：どうも相関係数はあまり高くないようだな？

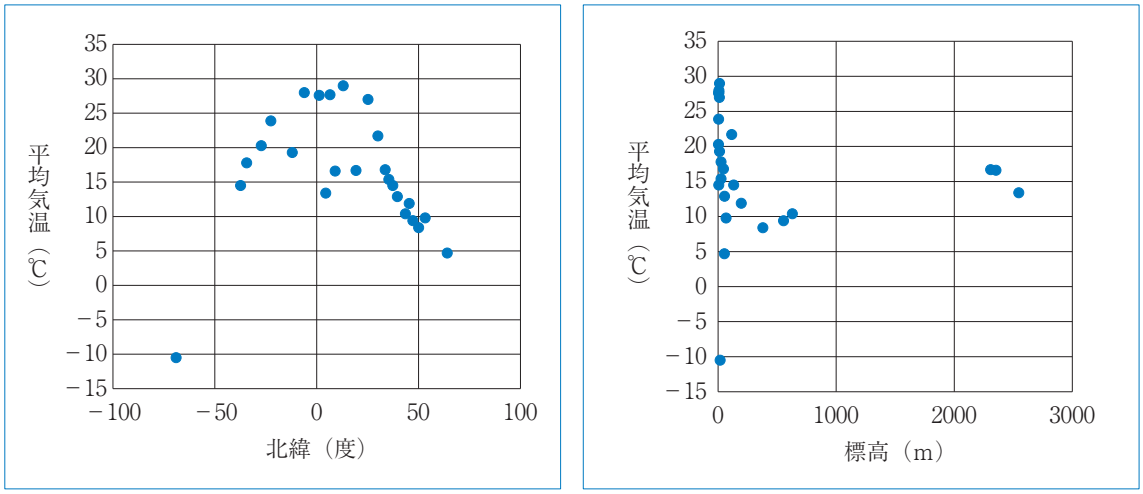
相関係数は、データの直線関係の強さを示します。散布図でデータが直線の上に乗っていれば、±１になるのですが、２次関数の上に乗っていても、必ずしも±１になるわけではありません。

◇ 散布図を注意深く眺めて関係性を想像しよう

先生：まずは、関係性が直線的かどうか、散布図というグラフで確認することが分析の第一歩です

表 1 のデータで緯度を横軸に、平均気温を縦軸にとった散布図と標高を横軸に平均気温を縦軸にとった**散布図**を図 2 に示す。

図 2　緯度・標高と平均気温の関係（散布図）



先生：図 2 の散布図から、どのような関数関係が想像できますか？

理恵：関数というのは、平均気温のデータをきちんと緯度で表現できる式のことですか？

先生：近似的に表現できる式ならば良いのです

公介：近似という意味が良く分かりません？

数学では変数 x と y の関係性は、関数 $y = f(x)$ を用いて表現する。一方、世界のさまざまな都市の平均気温 y と緯度 x との関係は、散布図を観察すれば、近似的にしか関係性は成立していない。実際、各都市のデータは、図 2 において想像された関数の上に正確に乗っているわけではない。散布図上の n 個の観測データの座標を (x_i, y_i), $i=1, \cdots, n$ とすれば、想像した関数に対して、その座標は

$$y_i = f(x_i) + e_i$$

と表現できる。e_i は、関数 $f(x_i)$ をデータに当てはめたときのハズレ（乖離）と考えれば良い。この e_i を統計では**残差（residual）**と呼ぶ。

このように、データ y は、関数で説明できる項と説明できない項の和に分けて表現される。

航平：緯度と平均気温の関係を示す図 2 の左側の散布図を見ると、緯度 0 度の赤道近辺を頂点として、緯度と平均気温は、上に凸の 2 次関数のような形状を示しているように思えます

理恵：南緯と北緯では、符号が違うけれど赤道について対称な関係のようです

公介：標高と平均気温の関係は、表 1 のデータが標高の低いところに固まっていて、その関係は思ったより無いね

◇ 関係性を定量的に表現してみよう

先生：それでは緯度と平均気温の関係をデータから示してみてください

一同：どうやれば良いのでしょうか？

データが近似的に示す関数関係を推察する簡単な方法は、条件付き平均値を求めることです。ある緯度帯を設定して、その緯度帯に入る都市を調べて、それらの都市の年間平均気温を平均や標準偏差を求めると良い。

Q３：緯度と平均気温との関係性を定量的に導きなさい

＊＊ヒント＊＊　表１の25都市データを緯度の絶対値の昇順に５都市ずつに並べ替え、その順番に５つの群に区分し、各群に属する都市を低緯度順に以下に示す。

群１　シンガポール、ボコタ、ジャカルタ、コロンボ、アジスアベバ
群２　リマ、チェンマイ、メキシコ、リオデジャネイロ、ドーハ
群３　ブリスベン、カイロ、ケープタウン、ブエノスアイレス、東京
群４　サンフランシスコ、メルボルン、北京、サラエボ、リオン
群５　チューリッヒ、プラハ、ダブリン、レイキャビク、昭和基地

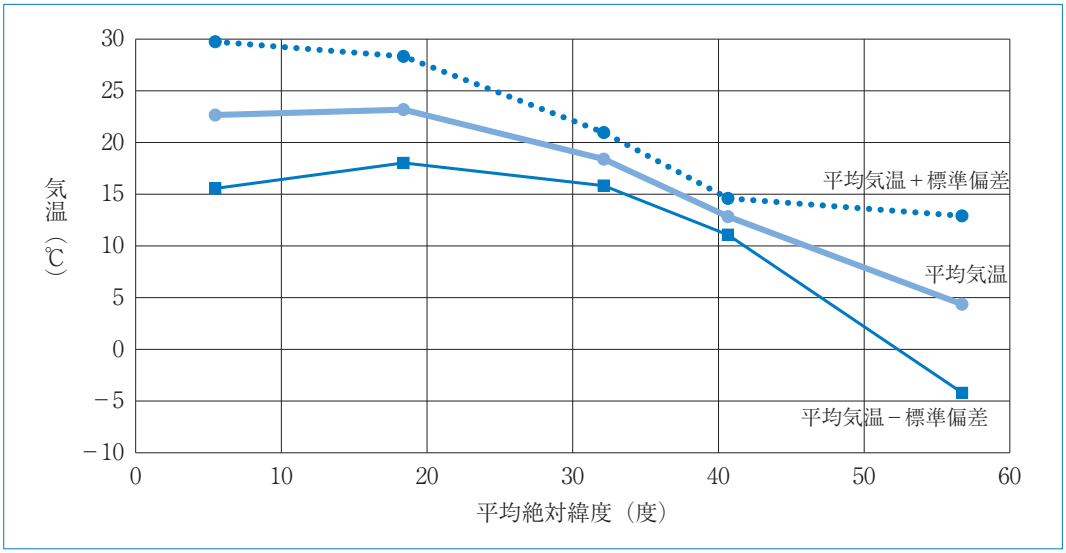
これから、各郡の緯度（の絶対値）の平均と平均気温の５都市平均・標準偏差を求め、表２に書き込みなさい。

表２　緯度の５層区分別の平均絶対緯度および平均気温の平均と標準偏差

群	平均 絶対緯度	平均気温の 平均	平均気温の 標準偏差
1			
2			
3			
4			
5			

５つの群について、表２で求めた平均絶対緯度を横軸として、縦軸に平均気温の平均と平均気温の平均±標準偏差を折れ線グラフに記すと図３となる。

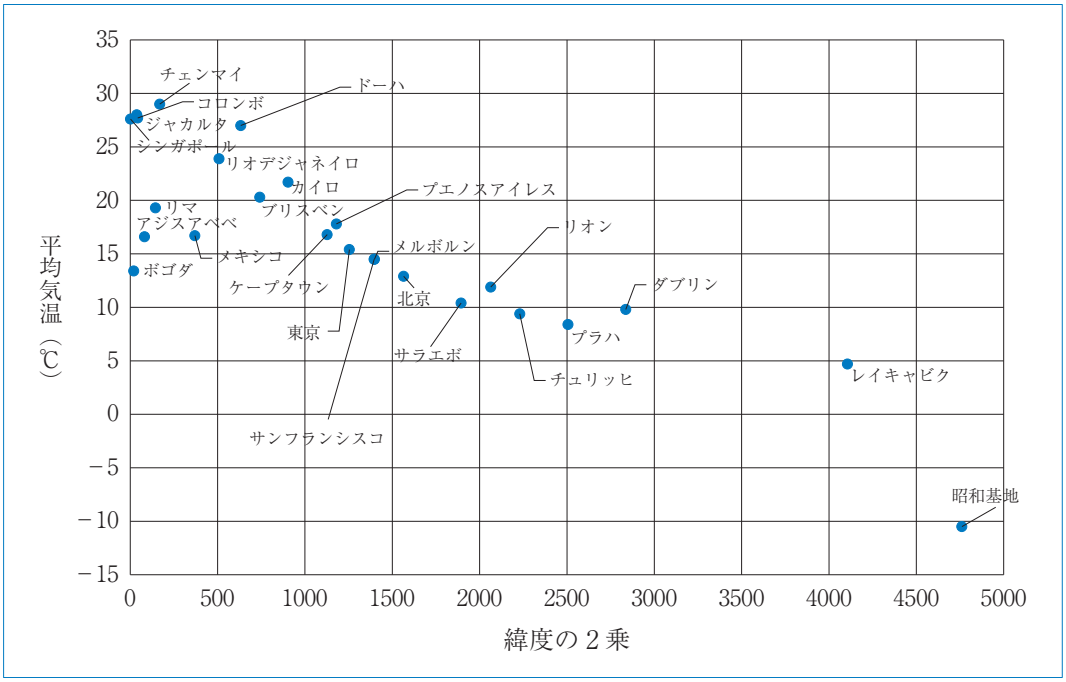
図３　平均緯度（平均）と平均気温、平均気温の平均±標準偏差



航平：図３をみると緯度と平均気温の関係は２次関数に見えますね
先生：２次関数だけが候補だとは思えないけれどね
理恵：三角関数みたいなものかもしれないわね
公介：ためしに、横軸に緯度の２乗、縦軸に平均気温をとった散布図を描いてみよう

公介君が描いた、散布図は図４のようになった。

図４　緯度の２乗と平均気温との散布図



先生：この散布図の関係性はどう見えますか
公介：直線関係に近いように見えます。この場合には相関係数を計算しても意味がありそうです。

Q４：緯度の２乗と平均気温の相関係数を求めなさい。

相関係数＝

航平：なぜ、緯度の２乗と平均気温との相関が高いのかな？
理恵：緯度は90度までしかないのに２次関数というのは奇妙よね。緯度100度があったら、もっと気温が下がるということですものね。もう少し、物理的裏付けのある緯度の変換はないかしら？

STEP 5：Conclusion 結論 1

Q５：次の文章の（　）の中に適当な語句を埋めなさい。

都市の平均気温は、緯度の（　）と直線的な関係にあることが分かる。
散布図から、おおよそどのような関係式になるだろうか？
地域の平均気温≒

理恵：直線関係に見える散布図から、一応、緯度の２乗と平均気温との関係式を導きましたが、もう少し数学的な方法はないのでしょうか？
先生：最小２乗法という方法で関係式を導くことができます。**回帰分析**とも呼ばれています。ただ、高校の数学の範囲を超えてしまいます

新たに見つかった課題解決のための、第２巡目の PPDAC サイクルに入る。

STEP 6：Problem 問題発見

◇ 関係性から外れているデータの統計的処理

航平：平均気温と関係のある緯度の２乗変換の散布図を眺めると、他の都市と異なって、直線関係上から少しずれている都市がいくつか見出せます
先生：なぜ、それらの都市が直線関係からずれるのかを考えてみましょう

多くのデータが示す散布図上の関係性とは少しだけ異なるように見えるデータは、**外れ値**と呼ばれる。外れ値が存在するということは、まだ結果の変化を説明する際、考慮されていない原因がデータに影響を与えているということが考えられる。

Q６：図４で外れの程度が大きい都市を外れの程度の大きい順に示しなさい。

外れの程度が大きい都市：

STEP 7：Plan 2

◇ 外れ値が生じている原因の考察

先生：外れている都市には、よく似たところがありませんか？もう一度、図１の特性要因図を眺めなおして、外れ値が生じる原因は何なのか考えてください。

公介：せっかくデータがあるのだから、外れている都市の標高を調べてみよう

Q７：先生と公介君の対話を参考に、外れ値が生じる要因を絞り込みなさい。

STEP 8：Data & Analysis 2

◇ 大きな外れ値を無くすデータの加工と解析

先生：都市の標高による平均気温への影響を調整すると、平均気温と緯度の関係はどうなるだろうか？

航平：表１の都市の平均気温の代わりに、都市の標高が０m だったとしたらどういう平均気温になるかを考えれば良いのでしょう

理恵：そうすると、（平均気温＋0.006×標高）を縦軸にした散布図を描いて、相関係数を求めれば良いのね

Q８：すべての都市の標高が０m だとしたら、その平均気温がどのようになるかを考えて、散布図を描き、相関係数を求めなさい。

相関係数＝

STEP 9：Conclusion 結論 2

Q９：都市の平均気温と緯度、標高の関係はどのようなものか、散布図から読み取りなさい。

都市の平均気温≒

重回帰分析

平均気温と緯度、標高との関係を計算で導く方法には、重回帰分析があります。

単回帰分析は**被説明変数**（目的変数）に対して**説明変数**が１つでしたが、説明変数を２つ以上に増やして回帰分析を行うことができます。これを重回帰分析といいます。

被説明変数を y、説明変数を x_1、x_2とした場合、重回帰分析から、切片 a、回帰係数 b_1、b_2を求めると、重回帰式は次のように表すことができます。

$y = a + b_1x_1 + b_2x_2$

重回帰式から y と x_1、x_2の２つの変数との関係を量的に求めることができるので、x_1と x_2の値から y を予測することができます。

第３巡目の PPDAC サイクルに向けて

上の結論で求めた関係から外れている都市はないだろうか？それは何故だろうか？

航平：依然として、リマやレイキャビクは少し、外れているように見える。リマやレイキャビクの気候について、Wikipedia で調べてみよう

航平君の調べた結果は、次のとおりである。

リマ：南緯12度と低緯度であるが沿岸を北流するペルー海流の影響によって気温は低く、最暖月の２月で22.5℃、最寒月の８月で15℃となる。 https://ja.wikipedia.org/wiki/ リマ より引用

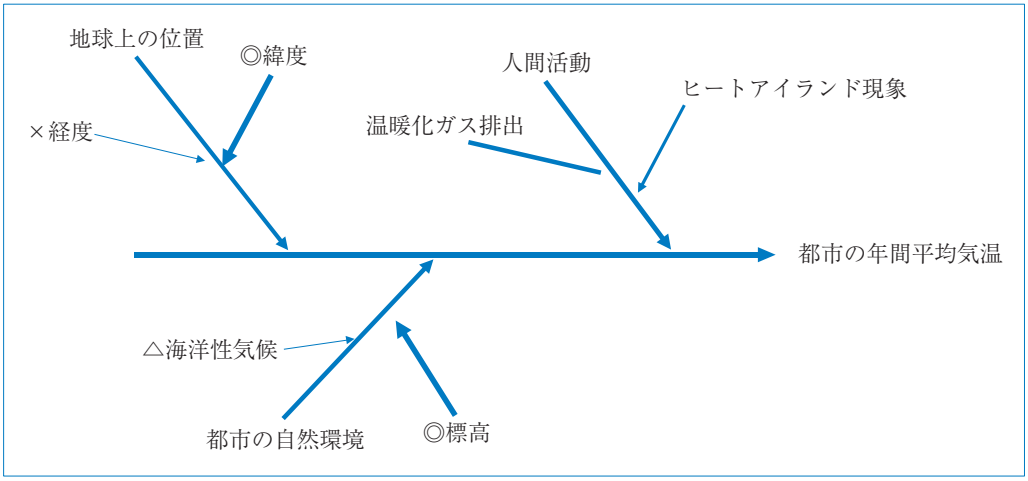
レイキャビク：アラスカのフェアバンクス、東シベリアのヤクーツクといった同緯度の地域と比べ、非常に温暖であることが最大の特徴である。ヤクーツクなど東シベリアでは内陸部中心に真冬ともなると－50℃前後の値を観測する事があるが、レイキャヴィークの最低気温は１年で最も寒い日でも－10℃程度にしかならない。この緯度のわりに温和な気候は、沖合を流れる暖流（北大西洋海流）、南から吹く偏西風に起因している。

https://ja.wikipedia.org/wiki/ レイキャヴィーク より引用

実際、リマは、図４の関係性から、リマは下側に、レイキャビクは上側にずれているように見え、ウィキペディアの記述と整合的である。したがって、図４に生じた外れ値は、海流などの影響と考えることができる。この海流や偏西風の影響の調整は、理科年表や気象庁から収集しているデータではできないので、今後の課題となる。

〔本節の解答〕

Ｑ１：特性要因図の一例



Ｑ２：緯度と平均気温の相関係数＝－0.045

標高と平均気温の相関係数＝－0.107

Q３：

群	平均 絶対緯度	平均気温の 平均	平均気温の 標準偏差
1	5.462	22.66	7.09
2	18.390	23.18	5.15
3	32.128	18.40	2.57
4	40.654	12.84	1.76
5	56.724	4.36	8.55

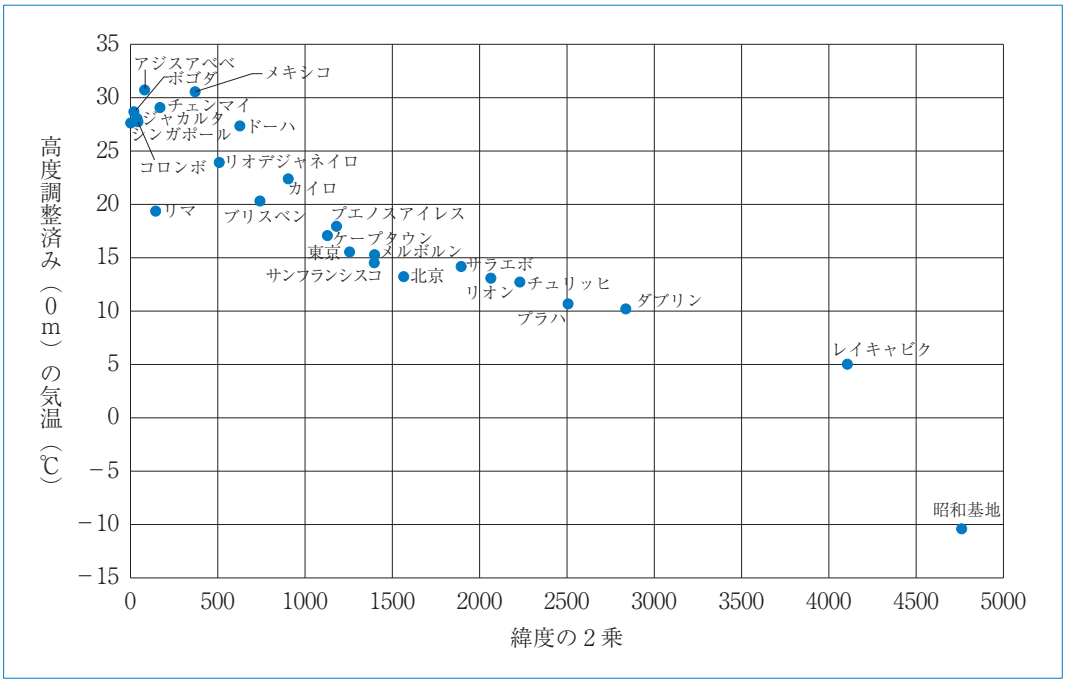
Q４：−0.877（４）２乗（コサインも正解）

Q５：おおよそ　$24-0.006\times$北緯^2、あるいは$-20+44\times\cos$（北緯）

Q６：図４で、他の都市が示す関係性から最も外れているのは、コロンビアの首都のボコダである。次に外れが大きいのは、メキシコの首都のメキシコ、エチオピアの首都のアジスアベバである。リマやレイキャビクもやや外れているように見える。

Q７：外れ具合が大きい都市は、すべて標高が2000m 以上である。このことは、図１で特性要因図に標高を入れることと整合的である。

Q８：



相関係数＝−0.943

Q９：都市の平均気温≒$28-0.007\times$緯度２$-0.006\times$都市の標高
あるいは、都市の平均気温≒$-24+\cos$(緯度)$-0.006\times$都市の標高

閑話休題　母集団と標本

第２部では、分析の対象とする個人、企業、地域など、集団の構成単位のすべてについて、データを利用できる状況を前提としていました。この場合、集団の特徴を度数分布、ヒストグラム、箱ひげ図等のグラフや平均、中央値、分散等の統計量から知ることができました。このような分析手法は、**記述統計学**（descriptive statistics）と呼ばれます。

一方、ある集団の構成単位のすべてに関する情報を得ることができず、集団の一部だけに関する情報を利用して、集団の特徴を推測したいことがあります。このような場合、知りたい（推測したい）集団全体を**母集団**（population）と呼び、取り出した一部を**標本**（sample）と呼びます。標本が母集団を正しく代表すると考えられるなら、その情報を利用することによって母集団に関して何らかの結論を導くことができるでしょう。母集団の特性を知ることを目的として、母集団から標本を選んで（**標本抽出、サンプリング** sampling）、それを統計的に分析し、母集団について推測する手法を、**統計的推測**（statistical inference）と呼びます。

統計的推測に基づいて有効な結論を導くためには、標本が母集団を適切に代表することが必要です。次の事例はそのことを十分に理解させてくれることでしょう。

～ 米国大統領選挙の予測の失敗～

無作為抽出の意義が広く認められるようになった事件があります。1936年の米国大統領選挙の選挙予測の**世論調査**で、出版社のリテラリー・ダイジェスト誌は回収数200万人超の大規模調査結果に基づいて共和党のアルフレッド・ランドンの勝利を予測しました。一方、世論調査会社のギャラップ社はわずか3000人を対象とした調査結果から民主党のフランクリン・ルーズベルトの勝利を予測しました。結果は後者の的中という大番狂わせでした。その理由は、両者の調査方法、特に標本抽出の方法にあります。リテラリー・ダイジェスト誌による調査は、自誌の購読者名簿や自動車・電話保有者名簿から対象を選んだため、結果として比較的高所得者に偏っていた可能性が指摘されています。これに対して、ギャラップ社は、有権者を性、年齢、社会階層、人種等の属性でグループに分け、それぞれのグループから規模に比例した割合で対象を抽出する**割当法**によることで、抽出された標本が母集団により近いものになったといわれています。

ギャラップ社が導入した割当法は、その後の世論調査で広く利用されるようになりましたが、12年後の1948年の大統領選挙では暗転しました。共和党のトマス・デューイが民主党のハリー・トルーマンに勝利するとの世論調査の予想結果はギャラップ社を含めてことごとく外れました。割当法において属性ごとの調査対象者数の割り当てを受けた調査員は、それに合致する人を必要数まで現地で選出する際、どうしても調査しやすい人を選びがちで、それによって調査結果に偏りを生じさせる可能性を残します。

世論調査ではこのような経験を踏まえて、偏りのない調査を行うため恣意的な要素を含まない無作為抽出法の重要性が認識されるようになりました。

記述統計から推測統計へ

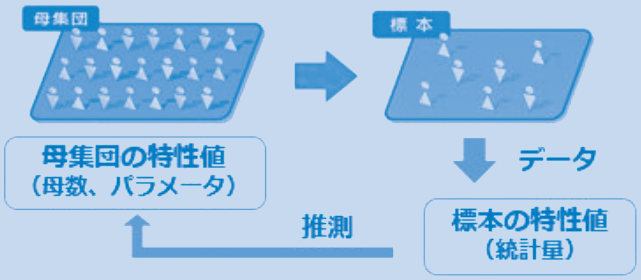
標本が与えられたとき、平均やバラツキを計算したり、度数分布を作成する「記述統計」の手法を用いることによって、標本がもっている情報を整理することができます。それでは、母集団の分布や平均などを知るための統計的推測においては、標本の分布や平均をそのまま使うことができるでしょうか。標本が母集団を適切に代表しているのであれば、標本から得られた平均は母集団の平均に近いことが予想されますが、実際に、どの程度近いのでしょうか。別な問題として、ある政策に関する意識に関する世論調査の例では、標本における支持率は分かりますが、母集団における支持率は、標本の支持率とどの程度近いのでしょうか。いずれの例でも客観的な評価の方法が必要となります。

このような問題に答えるための方法の１つは、標本を客観的な基準で選ぶことです。たとえば、テレビなどで政策に関する意見として街頭のインタビューの結果を放送することがありますが、これでは適切な標本とは言えません。インタビューの時間帯にその場所にいる人には事務系の会社員は少ない、社会人より学生・生徒が多い、女性より男性が多い、などの偏りがあります。午前10時から午後４時までの電話調査についても、電話番号をどのように選ぶかの他、日中、自宅にいる人は会社員ではない、有権者に限ると主婦が多いなどの偏りがあります。ある学校で生徒の意識を聞く標本調査の場合に、男子と女子に分けてそれぞれ「代表的」と考える生徒を選んだとしても、それが適切な標本であるかどうかは、選んだ教員以外には判断が困難ですし、他の教員が選ぶ生徒は違う可能性が高いでしょう。このような場合に、どちらの結果の信頼性が高いかについては、客観的な判断の根拠が乏しいことになります。

客観性を保証する最も簡単な方法が無作為抽出と呼ばれるもので、選挙人名簿から同じ確率で有権者の標本を選ぶという方法です。さらに丁寧に調査するなら、性別、年齢階級別に選挙人名簿を分けて、それぞれのグループから無作為に有権者を選ぶ方法もあります。ただし、この場合には、母集団における性別、年齢別の有権者数の情報を利用する必要があるので、手間がかかります。したがって、母集団をその特徴に応じて分割するにしても、ある程度の限界があります。

無作為抽出という方法で標本抽出が適切に行われた場合、標本は、偶然によって選ばれますから、標本の平均や度数分布も確率的に変動します。そこで、標本から母集団に関する推論を行うためには、このような確率的な変動がどの程度の大きさとなるかを明らかにする必要があります。この意味で、確率に関する理解が必要となります。

統計的推測に先立って、第３部で確率の基本的知識、第４部で代表的な確率分布である二項分布、正規分布等を紹介します。



第3部 統計的探究の実践 Ⅱ
～不確実な事象を理解する～

1　途中で中断したゲームの勝敗の帰趨は？［確率の概念］

可能性の大きさを測る確率の考え方に関する試行錯誤の例を紹介する。本節の素材は、パスカル（Blaise Pascal）とフェルマー（Pierre de Fermat）という２人の偉大な数学者によって、最初に不確実性が扱われた歴史的な問題である。

◇ 歴史的な問題

２人の人物 A、B があるゲームをして、先に３勝した者が賞金を貰えるものとする。A が２勝１敗の状態でゲームを中断したとき、賞金をどのように分配したら公平だろうか。問題を明確にするために、コイン投げ（表なら A の勝ち）のように規則を決めることにする。A が２勝１敗の状態から、どちらかが勝つまでゲームを続けたらどうなるかを考えれば良いだろう。

パスカルが考えた方法は次のとおりである。４回目のコイン投げで、A が勝てば賞金の全額を受け取るが、B が勝てば２勝２敗と五分になり分配金を半分ずつとするのが公平である。したがって、ゲームを中断したときの A の権利は（1＋1/2）/2＝3/4となる。

フェルマーは、次のように考えた。ゲームは最大であと２回で終了するから、すべての場合を列挙して、結果（*aa*、*ab*、*ba*、*bb*）はすべて1/4の可能性を持つと考える。ただし、*a*、*b* はそれぞれ、A、B の勝ちを表す。このうち、A は最初の３つの場合に勝ち、最後の場合は、B が勝つから、A への公平な分配金額は、3/4といえる。

フェルマーの方法では、４つの場合があるとしているが、次に A が勝ったらゲームは１回で終了するから、*aa* と *ab* はあり得ない。そうすると、起きうる場合は *a*，*ba*，*bb* の３通りしかないが、それらは同程度に確からしくないようにみえる。これは、この議論が行われた当時に、実際に提示された疑問とされる。確からしさについて、どのように考えたら良いのだろうか。

パスカルとフェルマーは次の問題も考えた。先に４勝した者が勝ちというゲームで、A が３勝１敗のときに中断したときの A の権利は、どのように求めたら良いだろうか。パスカルの方法を使えば、次のようになる。４回目のコイン投げで A が勝てば全額（分配金は１）を受け取るが、負けた場合は３勝すれば勝ちのゲームで２勝１敗のときと同じ状態となり、このときの分配金はさきほど計算した3/4となる。A が勝つ可能性と B が勝つ可能性は等しいから、A の権利は（1＋3/4）/2＝7/8となる。

Q１：この問題をフェルマーの方法で解いてみよう。

Q２：先に４勝した者が勝ちというゲームで、A が３勝２敗のときに中断すると、A の分配金はいくらとなるか、パスカルの方法とフェルマーの方法で、それぞれ計算してみよう。

◇ 確率の求め方

賞金を１とすると、勝つ可能性、すなわち、**確率**を合理的な分配金額とみなすことができるから、確率を求める問題として考えよう。

公正なサイコロやゆがみのないコインなど、すべての結果（n 通り）が同等に確からしい場合であれば、ある結果（r 通り）が起きる確率を r/n と定義することはもっともらしい。

これは確率の先験的な定義と呼ばれるもので、実際、多くの教科書ではこれを確率の定義と記しているが、この定義が適用できない問題があるだろうか。
（後に、相対度数による定義と比較して検討する）

ところで、A が２勝１敗の状態でゲームを中断した例では、ゲームが終わるまで続けたときの結果は、*a*、*ba*、*bb* の３通りであって、それらは同等に確からしいとはいえない。このような場合は、同等に確からしい結果を並べる必要があり、フェルマーが示したように *aa*、*ab*、*ba*、*bb* の４通りとすると、確率の先験的な定義が使えて、このうち A が勝つのは *aa*、*ab*、*ba* の３通りで、その確率は3/4と求められる。しかし、この方法では、実際に実現しない結果 *aa*、*ab* を含めているため、不自然という指摘も納得できる。そこで、もう少し納得できる方法を考えるために、準備として、確率に関する基本的な事項を確認しよう。

確率の考え方

簡単のために、毎回のコイン投げで表が出る確率が1/2であり、また、各回の結果は他の回の結果に影響しないという状況に限定しよう。表、裏をそれぞれ、H、T（英語の表裏は Head と Tail である）として、P(　)という記号で確率を表すと、P(H)＝P(T)＝1/2である。各回のコイン投げ結果はお互いに無関係だから、「１回目が表、かつ２回目が裏」（これを HT と表す）となる確率は掛け算で求めることができて、P(HT)＝P(H)P(T)＝(1/2)(1/2)＝1/4である。以下の確率も同様に求められる。

P(TH)＝P(T)P(H)＝1/4、P(HH)＝P(H)P(H)＝1/4、P(TT)＝P(T)P(T)＝1/4

このように掛け算で確率的に求められる場合、確率では独立と呼ぶが、詳しくは、後にあらためて考える。

決定木

途中でゲームが終わる場合も考えて、決定木（decision tree）という道具を使う。

図１　決定木の例

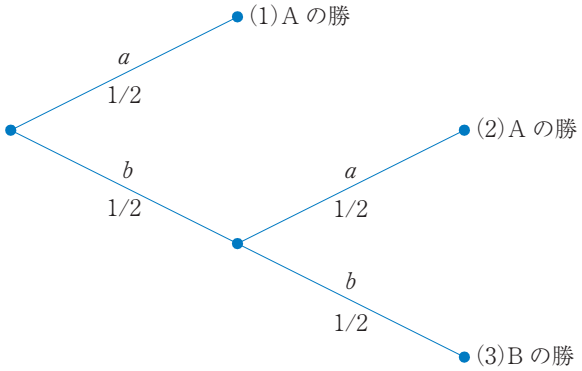


図１では、一番左の黒丸がゲームを中断した状態を表し、その後、ゲームを継続するときに起きうる結果を枝で表している。各節（黒丸）から分かれる枝に、*a*、*b* の結果とその確率が示されていて、右端の点でゲームが終了する。*a* が(1)、*ba* が(2)、*bb* が(3)という結果に対応する。(1)でゲームが終了する確率は P(H) = 1/2である。同様に、(2)で終了する確率は P(TH) = P(T)P(H) = 1/4、(3)で終了する確率は P(TT) = P(T)P(T) = 1/4である。A が勝つのは(1)と(2)の場合だから、確率はそれらの合計として、1/2 + 1/4 = 3/4と求められる。このようにすれば、ゲームにおいて実際には起きない HH および HT という場合を考える必要はなくなる。

実験による検証

実際にコイン投げなどの実験を何回か行って、以上の方法の妥当性を確かめよう。A の２勝１敗という状態から、決着がつくまでゲームを継続して、最終的に A が勝った回数と B が勝った回数を比べてみる。

A の２勝１敗という状態から始めて、ゲームを100回行った結果を示す。

H　TT　H　H　H　TT　TH　H　TT　H　TH　H　H　H　TH　H　TT　TH　TH　TT　(A) 15
TH　TH　H　TT　H　H　H　TH　H　TH　H　H　H　TT　H　H　TT　TH　TT　TT　(A) 15
H　TT　TH　TH　TH　H　H　H　TT　TH　H　TT　H　H　H　TT　H　TT　TT　H　(A) 14
H　TH　TT　H　H　H　H　H　TH　TT　H　H　H　TH　H　TH　H　H　TT　TH　(A) 17
H　H　TH　H　TH　H　H　TH　H　H　TT　TT　H　H　H　TT　TT　TT　H　H　(A) 15

実験は20回ずつ区切って、各行ごとに A が勝った回数を記している。100回の実験のうち、A が勝った回数の合計は76回、比率にすると0.76となる。これはこれまでに求めた確率の値3/4 = 0.75にかなり近い。

読者が実験する場合は、同じコインを投げ続ける代わりに、コンピュータが発生する疑似乱数を使っても良い。各回の実験で表が出る確率を1/2となるようにプログラムを作る。

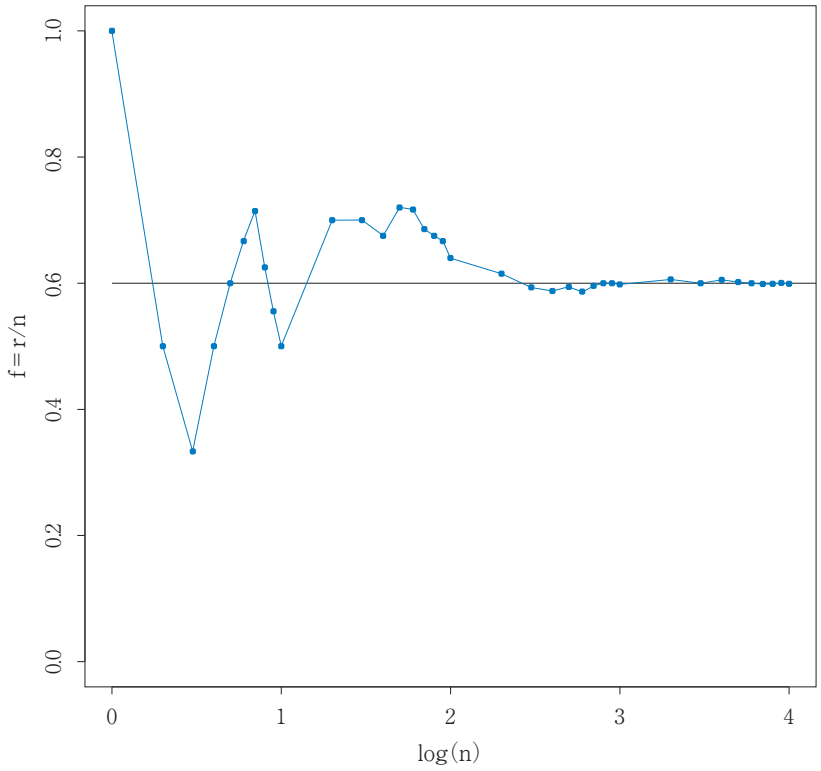
◇ 一般的な確率の考え方

最初の問題で確率を求めるためには、決定木を用いる解答が最も自然であろう。ここでは、もう少し進んだ問題を考える。実際には、サイコロやコインはゆがんでいることもあり、同等に確からしいという前提が満たされない場合にも確率を考える必要がある。20世紀まで標準的であった解釈では、相対度数によって確率を定義する。ゆがんでいるかもしれないコインを投げたときに表が出る確率を例にして、説明しよう。

コインを投げ続けて、n 回中表が r 回出たときに r を度数（または頻度、frequency）、$f=r/n$ を**相対度数**（relative frequency）と呼ぶ。私たちの経験では、実験を続けていると、最初のうちは相対度数 $f=r/n$ は変動しているが、実験の回数を増やしていくと相対度数は次第に安定して、ある値に近づくと考えられる。この値を表が出る確率と呼ぶことができる。

図 2 は、表が出る確率が0.6のコイン投げを、コンピュータで仮想的に実験した結果である。

図 2　コイン投げの実験（横軸は実験回数 n の常用対数）



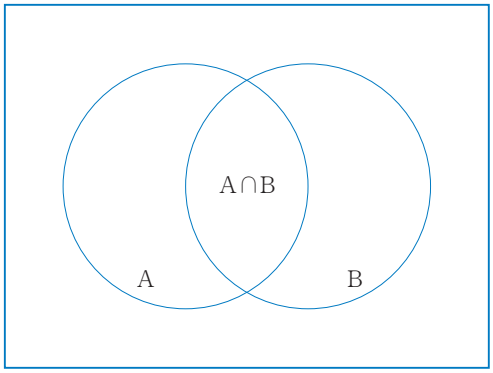
1, 2, …, 10, 20, …, 100, 200, …, 1000, 2000, …, 10000と回数を区切って、その回数までの相対度数を示している。横軸は常用対数で表示しているから、1, 2, 3, 4が、それぞれ10, 100, 1000, 10000に対応する。この例では、1000回を超えたあたりから0.6に近づいている。

一般の問題でも、ある事象 A に関する仮想的な実験において、A が起きる相対度数 r/n の極限を考えて、これを事象 A の確率 P(A) と呼ぶ。これが確率の相対度数による定義である。

確率論では確率を定義する対象を**事象**と呼び、記号 A，B，…，E，F，…などで表す。とくに、確実に起きる事象である**全事象** S の確率を 1 と定義する。A または B という事象（**和事象**）を A∪B，A かつ B という事象（**積事象**）を A∩B（AB または A・B という記法もある）と表す。A ではないという事象（**余事象**）を $\overline{A}$ と表す。とくに、全事象 S の余事象 $\overline{S}$ は**空事象**と呼ばれる「あり得ない事象」であり、これを記号 ϕ で表す。

図3は、英国の数学者ジョン・ベン（John Venn）による事象を表す図の例で、中央の部分がA∩Bである。同じコインを2回投げる例で、Aを HH、Bを HT、Cを TH、Dを TT、Eを1回目がH（2回目は何でも良い）という事象とすると、A∪B＝E、A∩E＝A、A∩B＝ϕ、E∪C∪D＝Sとなる。この例のA、Bのように、A∩B＝ϕとなる（同時に起きない）事象を排反と呼ぶ。

図3　ベン図の例



相対度数による定義を採用すれば、**排反な事象** A、Bに対して $P(A\cup B)=P(A)+P(B)$ が成り立つ。とくに、$\overline{A}$ とAは排反だから、$P(\overline{A})=1-P(A)$ が成り立つ。事象の数が多くなっても A_1，A_2，…，A_n が互いに排反 $A_i\cap A_j=\phi$ $(i, j, =1, \cdots, n, i\neq j)$ のときは、次の式が成り立つ。

$$P(A_1\cup A_2\cup\cdots\cup A_n)=P(A_1)+P(A_2)+\cdots+P(A_n)$$

この性質は加法性と呼ばれ、どのような定義を用いても、確率が満たすべき基本的な性質である。なお、排反でない事象も含めるときは、$P(A\cup B)=P(A)+P(B)-P(A\cap B)$ となることが**ベン図**から分かる。

伝統的な確率論では、事象Aが与えられたときの事象Bの**条件付確率**は、形式的に次の式で定義される。

$$P(B|A)=\frac{P(B\cap A)}{P(A)}\quad \text{ただし、} P(A)\neq 0$$

とくに、事象Aと事象Bについて、$P(B|A)=P(B)$ が成り立つとき、事象Aと事象Bは独立であるという。

Q3：$P(B|A)=P(B)$が成り立つとき、逆方向の関係 $P(A|B)=P(A)$ も成り立つことを確かめよう。ただし、$P(B)\neq 0$とする。

条件付確率の定義を書き換えた次の式は、**乗法法則**と呼ばれることがある。

$$P(A\cap B)=P(A)P(B|A)$$

乗法法則は、まず事象Aが起きたことを観測して、次に事象Bが起きる確率を求めるときに、自然な求め方となっている。ここで、事象Aと事象Bが独立であれば、$P(A\cap B)=P(A)P(B)$だから、$P(B)=P(B|A)$ となる。

初等的な問題では、2つの事象が発生する過程が物理的に無関係であれば、それらを独立と想定すれば良い。

練習のため、以下ではゆがんだコインを投げ続ける問題を考える。ただし、このコインで表が出る事象Hについては $P(H)=p$ $(0<p<1)$ であり、$p=1/2$とは限らない。毎回のコイン投げが物理的に無関係なら、独立と想定して良い。たとえば、順番に（H, H, T, H）と出る確率は、

$$P(HHTH)=P(H)P(H)P(T)P(H)=p\cdot p\cdot(1-p)\cdot p=p^3(1-p)$$ となる。

ゆがんだコイン投げなどの場合についても、相対度数による定義から、条件付確率の意味を明らかにできるだろうか。

Q4：このコインを投げ続けて、5回目に初めて表が出る確率はいくらか。

Q5：このコインを5回投げて、表が3回出る確率はいくらか。

◇ さらに理解を深めるために、練習問題を問いてみよう

Q６：青いカードｂ枚、白いカードｗ枚、合計ｎ＝ｂ＋ｗ枚を入れた箱から、よくかき混ぜてカードを抜きだす実験を考えよう。結果について、１枚目が青ならばB_1、２枚目が白ならばW_2などと記すことにする。

(a)　元に戻さずに抜き出すとき、条件付き確率$P(W_2|B_1)$はいくらか。
(b)　元に戻さずに抜き出すとき、確率$P(B_1 \cap W_2)$はいくらか。　ヒント：乗法公式
(c)　元に戻さずに抜き出すとき、確率$P(B_1 \cap W_2 \cap W_3)$はいくらか。　ヒント：乗法公式
(d)　元に戻さずに抜き出すとき、確率$P(W_2)$はいくらか。　ヒント：$P(W_2)=P(B_1 \cap W_2)+P(W_1 \cap W_2)$
(e)　毎回、元に戻しながら抜き出すとき、確率$P(B_1 \cap W_2 \cap W_3)$はいくらか。

Q７：箱の中に数字を書いた６枚のカードがあり、その数字は、３枚が「１」、２枚が「２」、１枚が「４」である。よくかき混ぜて、毎回元に戻しながらカードを抜き出す独立な実験を考える。以下の確率はいくらか。

(a)　３回の実験で、いずれも「１」が出る。
(b)　３回の実験で、「１」、「１」、「２」の順番に出る。
(c)　３回の実験で、「１」、「２」、「１」の順番に出る。
(d)　３回の実験で、「１」が２回、「２」が１回出る。
(e)　５回目に初めて「４」が出る。
(f)　「１または２」が出る前に「４」が出る確率はいくらか。

Q８：前問の箱で、元に戻さずにカードを取り出す（独立でない実験）場合、以下の確率はいくらか。

(a)　１枚目に「１」が出たという条件の下で、２枚目に「４」が出る。
(b)　３回の実験で、「１」、「１」、「２」の順番に出る。
(c)　３回の実験で、順番にかかわらず「１」が２枚、「２」が１枚出る。
(d)　１枚目の結果にかかわらず、２枚目に「２」が出る。
(e)　判断確率（主観確率）：１枚目を伏せた状態で２枚目に「２」が出たとき、１枚目を見たら「２」となる。

Q９：A、B、Cがゲームを続けて、Bが勝つ前にAが勝つ確率はいくらか。ただし、１回のゲームでそれぞれが勝つ確率は、$P(A)=a$、$P(B)=b$、$P(C)=c$とする（$a+b+c=1$）。

〔本節の解答〕

Q１：B の権利は1/8、A の権利は7/8

Q２：B の権利は1/4、A の権利は3/4

Q３：（省略）

Q４：$p(1-p)^4$

Q５：$10p^3(1-p)^2$

Q６：(a)　w/(n－1)
(b)　bw/{n(n－1)}
(c)　bw(w－1)/{n(n－1)(n－2)}
(d)　w/n
(e)　bw^2/n^2

Q７：(a)　1/8
(b)　1/12
(c)　1/12
(d)　1/4
(e)　$(5/6)^4$（1/6）
(f)　1/3

Q８：(a)　1/5
(b)　1/10
(c)　3/10
(d)　1/3
(e)　1/6

Q９：$a/(a+b)$

確率の話題

ここで扱った問題は、実際に、1654年８月24日付のパスカルからフェルマーに宛てた手紙で論じられている問題です。これは確率の計算というより公平な賞金の分配の問題であり、現在の用語なら「A が受け取る金額の期待値はいくらか」と言い換えても良い。期待値の話題は次節で取り上げます。

以下、パスカルとフェルマーが扱った、もう少し複雑な問題を紹介します。A、B、C の３人によるゲームで最初に２点取った者が勝つとして、１回目に A が１点取った時点で中断する場合の公平な分配方法はどうなるでしょうか。

フェルマーの方法では、この場合は決着がつくまでには最大で３回のサイコロ投げが必要です。すべての場合を列挙すると次の27通りとなります。

aaa　aab　aac　aba　abb　abc　aca　acb　acc
baa　bab　bac　bba　bbb　bbc　bca　bcb　bcc
caa　cab　cac　cba　cbb　cbc　cca　ccb　ccc

これらが同等に確からしいと考えて、A が勝つ場合の数を調べると17通りですから、答えは17/27となります。丁寧に数えるのは多少面倒で間違えやすく、決定木を使う方が確実です。

不思議なことに、パスカルは、勝負がついた後にサイコロ投げを続けると「勝敗が影響される」として、A が勝つのは *a* が１つ以上現れ、*b*、*c* が２つ未満の13通りに、*abb*、*bba* のように *b* または *c* が２回現れる６通りの半分の３通りを加えた16通りとしました。後者の場合は、B または C と折半するからです。そうすると、A、B、C の取り分はそれぞれ、16/27、5.5/27、5.5/27になります。しかし、これはゲームの規則が違う問題の答えであり、パスカルのような天才が誤りを犯したことは謎とされています。なお、パスカルはこの直後に神秘的な体験をして、数学を捨てて信仰を中心とする生活に入りました。

先験的な確率の定義を適用するための「同等に確からしい」という判断は必ずしも容易ではありません。18世紀フランスの書物『百科全書』の著者の１人である数学者のダランベール（Jean Le Rond d'Alembert）は、２枚のコインを投げて２枚とも表になる確率を1/3とする有名な間違いを記してしまいました。ダランベールについて、「区別できないコインに対する可能な結果 HH、HT、TT が同等に確からしいとする」という前提であれば、数学的には誤りとは言えません。しかし、通常は「２枚のコインは区別できる」から、同等に確からしい可能な結果は HH、HT、TH、TT の４通りと想定できます。実際にコインを投げてみれば、多少はゆがみがあっても、この想定に近い結果が得られます。

もう少し複雑な歴史的な例に、ガリレオ（Galileo Galilei）が解いた問題があります。それは３つのサイコロを同時に投げ、その目の和が９になる確率と10になる確率を求める問題で、やはり３つのサイコロを区別することが要点です。単純に数えれば、目の和が９になるのは（1, 2, 6）（1, 3, 5）（1, 4, 4）（2, 2, 5）（2, 3, 4）（3, 3, 3）の６通り、目の和が10になるのも（1, 3, 6）（1, 4, 5）（2, 2, 6）（2, 3, 5）（2, 4, 4）（3, 3, 4）の６通りですが、先験的な確率が適用できる事象を列挙するためには、３つのサイコロを青・赤・白のように区別して考える必要があります。すると、異なる目の出方は$6^3=216$通りで、この１つ１つが同等に確からしい。３つの目の組合せを数えると、（1, 3, 6）となるのは６通り、（2, 2, 5）となるのは３通りだから、それらの確率は異なります。この手順を用いると、目の和が９になる確率は（6＋6＋3＋3＋6＋1）/216＝25/216です。同様にして、目の和が10になる確率は（6＋6＋3＋6＋3＋3）/216 ＝ 27/216とわずかに大きいことが分かります。

ところで、量子力学の世界では、区別できないコインに基づく確率が現れる場合があります。２個の粒子がそれぞれ２つの状態を取り得る場合、それらが区別できるならコインと同様に４通りの状態があり、いずれも1/4の確率で生じると考えたいところですが、量子力学においては粒子を区別できず、出現確率は HH、HT、TT が1/3ずつとなる状況があります。

このような場合も含めると、無条件で先験的な確率の定義を適用することには注意が必要です。ただし、日常生活で現れる事例では、近似的には同等に確からしいと想定できるような状況も少なくないので「先験的な確率の定義」による確率の計算は最も基本的な考え方といえます。

最後に、ニュートン（Isaac Newton）が答えた、いくつかのサイコロを投げて６の目が出る確率を求める問題を紹介しましょう。

（A）　６個のサイコロを投げて、少なくとも１つ「６の目」が出る確率
（B）　12個のサイコロを投げて、少なくとも２つ「６の目」が出る確率
（C）　18個のサイコロを投げて、少なくとも３つ「６の目」が出る確率

答えは、次のようになります。式は難しくないけれど、この計算には、ニュートンも疲れたのではないでしょうか。

$$P(A)=1-\left(\frac{5}{6}\right)^6=\frac{31031}{46656}\fallingdotseq 0.6651$$

$$P(A)=1-\left(\frac{5}{6}\right)^{12}-12\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{11}=\frac{1346704211}{2176782336}\fallingdotseq 0.6187$$

$$P(A)=1-\left(\frac{5}{6}\right)^{18}-18\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{17}-\frac{18\times 17}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{16}=\frac{6066641080916}{10155995668460}\fallingdotseq 0.5973$$

2　保険料をどのように決める？［確率の応用］

確率の考え方が実際に応用されている例として、保険の話題を取り上げる。

◇ 期待値

住宅と家財を含めて2000万円の資産価値があり、１年間に火災などの原因でその価値が失われる確率が１％である場合を考えよう。個人が火災などのリスクを負担すると、１％の確率とはいえ、その損害は重大である。確率を考慮するとき、重大な損失を防ぐためには、どの程度の金額を用意しておけば良いのだろうか。保険の制度は、このような問題に対する１つの解答である。その仕組みを理解するためには、**期待値、標準偏差、大数の法則**などが役に立つ。

ある「くじ」から得られる賞金をどれだけ期待できるかを表す数値が期待金額であり、**確率変数の期待値**の本来の意味である。次のくじについて考えよう。

１から100までの番号がついた100個の玉が入っている箱の中から（毎回、元に戻しながら）玉を１個取り出す。賞金 u の額は、番号が１から60のときは $u_1=0$、60から90のときは $u_2=1$、91から100のときは $u_3=10$とする（単位は千円）。

（１）　このくじを１回引いて、u_1、u_2、u_3という結果が起きる確率は、それぞれ、$p_1=0.6$、$p_2=0.3$、$p_3=0.1$と考えてよい。

（２）　このくじを $n=10$回引いて、u_1、u_2、u_3となる回数をそれぞれ、n_1、n_2、n_3（$n_1+n_2+n_3=n$）とする。n 回のくじから得られる金額の合計は、$u_1n_1+u_2n_2+u_3n_3$で与えられる。n_1、n_2、n_3は確率的に変動するから、合計金額も確率的に変動する。

（３）　ここで、n を大きくして100万回と想定してみよう。このとき $u=u_1$、u_2、u_3が得られるそれぞれの相対度数 $f_1=n_1/n, f_2=n_2/n, f_3=n_3/n$ は、それぞれ、確率 $\mathrm{P}(u=u_1)=p_1=0.6$、$\mathrm{P}(u=u_2)=p_2=0.3$、$\mathrm{P}(u=u_3)=p_3=0.1$ に近いと考えて良い。

（４）　多数回のくじから得られる１回当たりの金額は、$(u_1n_1+u_2n_2+u_3n_3)/n=u_1f_1+u_2f_2+u_3f_3$と求められ、これは近似的に $u_1p_1+u_2p_2+u_3p_3$に等しい。

（５）　確率変数 u の期待値は $\mathrm{E}(u)=u_1p_1+u_2p_2+u_3p_3$と定義され、$n$ が大きい場合の１回当たりの金額と解釈できる。

以下、本節では、確率変数を x, u などの小文字で表す。特に区別する場合には、$\tilde{x}$を確率変数、x をその実現値と表現する。これはベイズ統計学では標準的な表記法である。

確率変数の期待値（定義と解釈）

確率変数 x が、確率 $p_1, p_2, \cdots$（$p_1+p_2+\cdots=1$）で $x_1, x_2, \cdots$という値をとるとき、x の期待値 $\mathrm{E}(x)$ は次の式で定義される。

$$\mathrm{E}(x)=x_1p_1+x_2p_2+\cdots=\sum_i x_ip_i$$

ここで、E は expectation の頭文字である。また、平均（mean）の頭文字 m に対応するギリシア文字μが、期待値の標準的な記号である。$p_i=p(x_i)$ と書いて、$\mu=\sum_x xp(x)$という表現もよく用いられる。

確率変数 x について、実験を n 回行った観測結果で、$x_1, x_2, \cdots$となる度数をそれぞれ、$n_1, n_2, \cdots$として、相対度数をそれぞれ、$f_1=n_1/n, f_2=n_2/n, \cdots$とすると、算術平均は $\bar{x}=(n_1x_1+n_2x_2+\cdots)/n=f_1x_1+f_2x_2+\cdots$として与えられる。ここで、$n$ が大きくなるとき、$f_1, f_2, \cdots$はそれぞれ、確率 $p_1, p_2, \cdots$に近づき、$\bar{x}$は期待値μに近づく。

宝くじの期待値

１枚300円のある宝くじの賞金と当選確率は表１のように与えられている。

表１　宝くじの賞金・当選確率と期待値および分散の計算

賞	賞金	当選確率
１等	3億円	1/1000万
１等前後	1億円	1/500万
２等	1千万円	1/250万
３等	10万円	1/1万
４等	1万円	1/500
５等	2000円	1/100
６等	300円	1/10
外れ	0円	（＊）

注：（＊）は差で求める

u	$p(u)$	$up(u)$	$(u-\mu)^2p(u)$
3×10^8	$1/10^7$	30	9.0000×10^9
1×10^8	$2/10^7$	20	2.0000×10^9
1×10^7	$4/10^7$	4	3.9999×10^7
1×10^5	$1/10^4$	10	9.9732×10^5
1×10^4	$2/10^3$	20	1.9468×10^5
2000	1/100	20	3.4820×10^4
300	1/10	30	2.7556×10^3
0	0.8878993	0	1.5943×10^4
合計	1.0000	$\mu=134$	1.1041×10^{10}

このくじの賞金を u として、その期待値を計算すると、表の右側のように $\mu=\mathrm{E}(u)=134$円となる。現金300円を持っているより損をしているが、宝くじは夢を買うものとされる。

◇ 確率変数の分散と標準偏差

期待値が度数分布から計算される平均値の極限とするのと同様、度数分布から計算される分散の極限を確率変数の分散とすることも自然に理解できるだろう。

x の観測値からは、分散は次の式で計算される。

$$s^2=\frac{1}{n}\left\{n_1(x_1-\bar{x})^2+n_2(x_2-\bar{x})^2+\cdots\right\}=\frac{1}{n}\sum_i n_i(x_i-\bar{x})^2=\sum_i f_i(x_i-\bar{x})^2$$

ここで、n が大きくなると $f_i \fallingdotseq p_i$ となることから、確率変数 x の分散は、次の式で定義する。

確率変数の分散（および標準偏差）

確率変数 x が、確率 $p_1, p_2, \cdots (p_1+p_2+\cdots=1)$ でそれぞれ、$x_1, x_2, \cdots$ という値を取るとき、確率変数 x の分散は、$\mu=\mathrm{E}(x)$ として、次の式で定義される。

$$\sigma^2=\mathrm{V}(x)=p_1(x_1-\mu)^2+p_2(x_2-\mu)^2\cdots=\sum_i p_i(x_i-\mu)^2=\sum_x p(x)(x-\mu)^2$$

分散の平方根 σ を確率変数 x の標準偏差と呼ぶ。ここで V は分散の英語（variance）の頭文字、σ は標準偏差の英語（standard deviation）の頭文字 s に対応するギリシア文字である。

宝くじの標準偏差

表１の宝くじの賞金の標準偏差を求めてみよう。表の右側で計算されているように、分散は $\sigma^2=1.1041\times10^{10}$ である。バラツキの尺度としては標準偏差が理解しやすい単位であり、この例では $\sigma=105077.3$円である。期待値は134円と小さいが、最高の賞金が大きいため、標準偏差が大きくなっている。$\sigma \fallingdotseq 0$ なら確実に損をするが、結果が大きく変化する可能性が夢を表していると考えられる。

◇ 確率変数の和

2つの確率変数 x と y が独立のとき、その和 $T=x+y$ について、x と y の期待値を $\mathrm{E}(x)=\mu_x$、$\mathrm{E}(y)=\mu_y$、分散を $\mathrm{V}(x)=\sigma_x^2$、$\mathrm{V}(y)=\sigma_y^2$とするとき、T の期待値と分散は $\mathrm{E}(T)=\mathrm{E}(x)+\mathrm{E}(y)=\mu_x+\mu_y$、$\mathrm{V}(T)=\mathrm{V}(x)+\mathrm{V}(y)=\sigma_x^2+\sigma_y^2$と求められる。

さらに、一般に次の結果が知られている。

独立な変数 x, y の一次式 $L=ax+by$ の期待値と分散

$\mathrm{E}(L)=a\mathrm{E}(x)+b\mathrm{E}(y)=a\mu_x+b\mu_y$、$\mathrm{V}(L)=a^2\mathrm{V}(x)+b^2\mathrm{V}(y)=a^2\sigma_x^2+b^2\sigma_y^2$

類似の結果は3つ以上の確率変数の場合にも成り立つ。

たとえば、互いに関係のない（独立な）2つのくじ x, y について、期待値と標準偏差が $\mu_x=70$、$\mu_y=65$、$\sigma_x^2=10^2$、$\sigma_y^2=15^2$とするとき、合計 T については、期待値が $\mathrm{E}(T)=70+65=135$、分散が $\mathrm{V}(T)=100+225=325$となる。

したがって、標準偏差は$\sigma_T=\sqrt{325}\fallingdotseq 18.03$である。

表1の宝くじを100枚買うときの期待値を計算してみよう。発売されるくじの枚数が1000万枚と大きければ、100枚のくじの結果を独立な確率変数とみなすことができる。ただし、1等前後賞があるので、連続番号のくじは買わないものとする。それぞれの宝くじの賞金の期待値はいずれも134円となり、合計で13400円が期待金額である。1枚当たりにすると、何枚買っても期待金額には変化がない。ところで分散は100倍になり、標準偏差はその平方根だから10倍、1枚当たりにすると10分の1になる。つまり、夢が小さくなる。100枚ではなく、10000枚買えば標準偏差は100分の1の1050.7円とさらに小さくなる。あまりたくさん買うと独立な結果とはみなせなくなるが、買い占めた場合を考えれば、標準偏差はゼロとなるから、正確に1枚当たりの期待値134円を受け取り、確実に損をする。

◇ 保険の役割

火災保険の例も、くじの1種と考えることができる。個人については、期待損失額は表2のように計算される。

表2　期待損失額の計算（単位：万円）

事象	損失額（u）	確率（p）	積（up）
無事	0	0.99	0
災難	2000	0.01	20
合計		1.00	20

災難が起きる確率が小さいので、損失額の期待値（**期待損失額**）は20万円と、それほど大きくはならないが、多くの人は、災難が起きる場合、2000万円という大きなリスクを個人で負うことは賢明ではないと考える。リスクを少なくする方法の1つが火災保険の制度で、1年間20万円程度の保険料を支払うことによって、万一、損害が発生した場合には2000万円の保険金を受け取ることができる。

保険の仕組みを理解するために、同じような状況の人が n 人いて、少しずつお金を出し合って、そのお金を災

難が起きた人に提供して経済的に助け合う制度を考えよう。

n 人の中で 1 年間に災難が発生する件数は、裏が出る確率が $p=0.01$のゆがんだコインを n 回投げる実験で裏が出る回数とみなすことにしよう。火災の発生確率が等しく、それぞれの火災発生が独立であれば、この想定は合理的である。 1 年間に災難が発生する件数を x とすると、損失額の合計は $u=2000x$ となる。ここで、二項分布に従う x の期待値は $\mathrm{E}(x)=np$ だから、損失額の期待値は $\mathrm{E}(u)=2000np$ となる。$n=100$人のときは、期待値は $\mathrm{E}(x)=np=1.0$（件）、$\mathrm{E}(u)=2000\mathrm{E}(x)=2000$万円となる。これを $n=100$人で割ると 1 人当たり $\mathrm{E}(u/n)=20$（万円）となって、表 2 の期待値と変わらない。

表 3　$n=100$人のとき、$x=0, 1, \cdots, 8$までの確率（%）小数点以下 3 桁まで

	$x=0$	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$	$x=7$	$x=8$
確率	36.603	36.973	18.486	6.100	1.494	0.290	0.046	0.006	0.001
累積	36.603	73.576	92.063	98.163	99.657	99.947	99.993	99.999	100.000

また、xの確率は表 3 で与えられる。この表によると、損害が 9 件以上となる確率は10万分の 1 より小さい。仮に、10件の損害が発生したとしても、 1 人当たりの負担額は2000万円 ×（10/100）＝200万円で収まる。個人で2000万円の損害が発生する確率 1 %と同じ水準の危険なら 4 件以下となり、その場合の 1 人当たりの負担額は2000万円 × 4/100＝80万円となる。このように、$n=100$人でもある程度のリスクの軽減ができているが、数十万人、数百万人と大勢になれば、 1 人当たりの損害負担額は、期待金額とほとんど変わらないことが予想できるだろう。これは、確率論の重要な定理である**大数（たいすう）の法則**（law of large numbers）によって保証される。

現実には、このような契約は保険会社が提供しているもので、期待損失額に管理経費を加えた程度の保険料を受け取ることで、保険会社の健全な経営が可能となる一方、契約する個人にとっては、大きな損失の可能性というリスクを軽減することが可能となっている。

◇ 生命保険の仕組み

生死や病気、ケガなど人に関わる生命保険も同様な原理による。保険に加入している人が死亡した場合に、残された家族に必要な保険金が支払われるのが死亡保険である。

近代的な生命保険が生まれるまでの歴史

中世ヨーロッパでは、商人たちは職業ごとに同業者組合「ギルド」を作り、冠婚葬祭など組合員の経済的マイナスを組合全体で分担しあっていたことから、このギルドを生命保険の起源とする説があります。17世紀のイギリスにおいて、教会の牧師たちが組合を作り、自分たちに万が一のことがあった場合に遺族へ生活資金を出すために保険料を出し合う制度を始めました。しかし、この制度では全員が同じ金額の保険料を支払っていました。人の死亡率は年齢とともに上がっていくので、若い人よりも年をとった人の方が有利となったため、組合はほどなく解散することになりました。その後、イギリスのジェームス・ドドソンという数学者によって、公平な保険料分担の方法が発見され、1762年に世界で初めて近代的な保険制度に基づく生命保険会社が設立されました。日本においては、福沢諭吉による「西洋旅案内」で初めて生命保険が紹介された後、1881年に欧米の近代的保険制度を手本として生命保険会社が設立されました。

（生命保険協会ホームページの解説による）

生命保険においては、保険料（収入）と保険金（支出）が等しくなることが基本で、仮に保険金と保険料を一律に定めると、以下の関係が成り立つ必要があります。

保険金×死亡者数＝保険料×契約者数

Q１：さまざまな年齢の男女100万人が構成している集団を考える。この集団の構成員が１人死亡するたびに、生命保険会社がその家族に100万円を支払う場合、構成員１人が負担すべき保険料はいくらになるか。ただし、この集団の死亡率は２％とし、保険料は全員同じとする。

実際の死亡率

厚生労働省によって、毎年の出生者数、死亡者数などが報告されています。

2014（平成26）年の死亡数は127万3004人で前年より4568人増加し、死亡率（人口千対）は、10.1で前年と同率です。死亡数と死亡率の年次推移をみると、明治から大正にかけて、死亡数は90万～120万人、死亡率は20台で推移してきました。昭和に入って初めて死亡率は20を割り、1941（昭和16）年に死亡数は115万人、死亡率は16.0まで低下しました。第２次世界大戦後の1947年に死亡数は114万人、死亡率は14.6でしたが、医学や医療の進歩および公衆衛生の向上などにより死亡の状況は急激に改善され、1966年には死亡数が最も少ない67万人、1954年には死亡率が最も低い6.0となりました。その後、人口の高齢化を反映して緩やかな増加傾向に転じ、2003年に死亡数は100万人を超え、死亡率も上昇傾向にあります。また、年齢階層でみると、14歳以下の死亡数は、明治から昭和初期にかけて多かったが、戦後、急激に減少しています。近年では人口の高齢化を反映して65歳以上の死亡数が増加し、特に80歳以上の死亡数の増加は顕著で、全死亡数に占める割合は増加しており、2014年では60.0％となっています。なお、2014年の性別死亡率（人口千対）は男10.8、女9.5です。これを都道府県別にみると、死亡率が最も低いのは、男では沖縄県が8.7、次いで、神奈川県9.0、東京都9.1、女では沖縄県が7.4、次いで、神奈川県7.6、埼玉県7.8です。また、最も高いのは男では秋田県15.5、次いで、青森県14.3、島根県14.1、女では秋田県で13.8、次いで、高知県と山形県で13.2となっています。都道府県別にみた死亡率と65歳以上人口割合は、ほぼ同様の傾向です。

「2014年我が国の人口動態（平成26年までの動向）」
http://www.mhlw.go.jp/toukei/list/dl/81-1a2.pdf

これによると、医療の進歩などによって死亡率は減少してきたものの、年齢別・性別に違いがあることが分かります。

都道府県別にみた死亡率と65歳以上人口割合を、次の方法で比較してみよう
１．横軸に都道府県、縦軸に死亡率と65歳以上人口割合をとる棒グラフと折れ線グラフの組み合わせ
２．横軸に65歳以上人口割合、縦軸に死亡率をとる散布図
３．高齢者が多い地域の死亡率が高いことが確認できただろうか

図１　性別・年齢別死亡率の変化

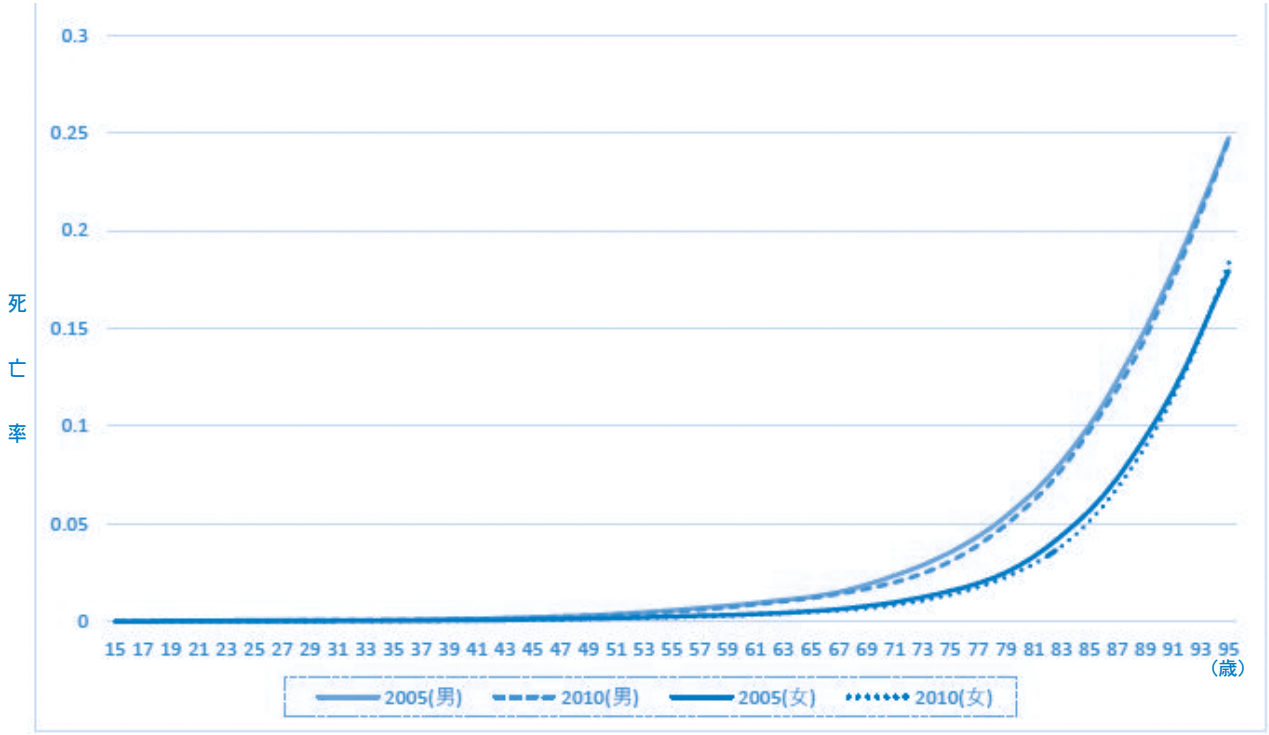


図１は15歳から95歳まで年齢ごとの死亡率で、ある年齢に達した人が１年以内に死亡する割合として求められている。これは無作為に選ばれた人の死亡確率と考えて良い。死亡率には男女間で差があるが、2005年と2010年の比較では、性別・年齢別に比較的安定している様子が見える。なお、若い年齢層の違いはグラフでは読み取りにくいので、５歳ごとの死亡率を表４に記している。これを見ると、死亡率は15歳以上では年齢とともに増加し、また、男より女の方が一貫して低いことが分かる。たとえば、2005年の15歳・男の死亡率は0.00023で、これは2010年の20歳・女の死亡率0.00024に近い。

性別・年齢別の調整　Ｑ１の例では集団全体の死亡率を２％と想定したが、実際は、統計から死亡率は年齢とともに高くなることや、同じ年齢でも男性と女性では死亡率が異なることが知られている。そのため、すべての人が同じ保険料を負担すると、不公平となってしまう。「生命保険の誕生」の例でも、この点が原因で継続できなかった。実際には、生命保険会社では契約者の集団における死亡率を年齢別・男女別に計算して、契約者ごとの保険料が公平になるよう算出している。

◇ 大数の法則

保険で公平な保険料を算出するときには、大数の法則と呼ばれる確率の性質が利用されている。ここでは比率に関する大数の法則について説明しよう。

表 4　死亡率の比較（2005年、2010年生命表）

年齢	男		女	
	2005	2010	2005	2010
15	0.00023	0.00019	0.00012	0.00012
20	0.00056	0.00051	0.00026	0.00024
25	0.00067	0.00064	0.00032	0.00026
30	0.00074	0.00069	0.00037	0.00036
35	0.00098	0.00085	0.00053	0.00048
40	0.00143	0.00128	0.00075	0.00071
45	0.00227	0.00198	0.00113	0.00108
50	0.00357	0.00317	0.00176	0.00167
55	0.00579	0.00507	0.00265	0.00236
60	0.00883	0.00810	0.00364	0.00340
65	0.01277	0.01214	0.00536	0.00498
70	0.02123	0.01842	0.00890	0.00767
75	0.03555	0.03087	0.01574	0.01381
80	0.05998	0.05568	0.02898	0.02600
85	0.10068	0.09785	0.05696	0.05155
90	0.16453	0.16041	0.10563	0.10160
95	0.24758	0.24695	0.17947	0.18367

比率に関する大数の法則

1回の実験で事象Aが起きる確率をpとするとき、この実験を独立にn回繰り返したときに事象Aがx回起きたとすると、nが大きくなるに従って、相対度数$\hat{p}=x/n$は確率pに近づく。

この問題では、xは二項分布に従う確率変数であり、期待値は$\mathrm{E}(x)=np$、分散は$\mathrm{V}(x)=np(1-p)$となる。なお、相対度数の記号として$\hat{p}=x/n$を用いて、観測された$\hat{p}$によって未知のpを推定するという意図を表している。標準偏差は分散の平方根で$\mathrm{sd}=\sqrt{np(1-p)}$だから、相対度数$\hat{p}=x/n$はその期待値である$p$から$\sqrt{p(1-p)/n}$の数倍程度しか離れないという解釈ができる。

ここで必要なのは相対度数$\hat{p}=x/n$の期待値が$\mathrm{E}(\hat{p})=\mathrm{E}(x)/n=p$となることと、標準偏差が$\mathrm{sd}=\sqrt{p(1-p)/n}$となることである。これから、$\hat{p}$と、その期待値である$p$との距離は$\sqrt{p(1-p)/n}$の数倍程度と考えることができて、この距離は$n$が大きくなればゼロに近づく。以上が大数の法則が成立する理由である。前節で相対度数による確率の定義を紹介したが、大数の法則から、この定義の正当性が保証される。

保険に関して、一人ひとりの死亡を予測することはほとんど不可能だが、社会全体として死亡率は非常に安定している。このことが保険会社が死亡率を利用して保険料を定めることを可能にしている。現実には、性別・年齢別の各集団で保険会社が支払い義務のある保険契約者の数は、数万人程度である。そこで、集団の大きさと死亡率を想定して、いくつかの場合に死亡者数xの期待値npとその標準偏差(sd)$\sqrt{np(1-p)}$を計算すると、表5のようになる。

表 5　集団の大きさ（人）・死亡率と死亡者数の期待値・標準偏差・変動係数

n	p	np	sd	cv	n	p	np	sd	cv
20,000	0.0005	10	3.16	0.316	50,000	0.0005	25	5.00	0.200
20,000	0.001	20	4.47	0.223	50,000	0.001	50	7.07	0.141
20,000	0.002	40	6.32	0.158	50,000	0.002	100	9.99	0.100
20,000	0.005	100	9.97	0.100	50,000	0.005	250	15.77	0.063
20,000	0.01	200	14.07	0.070	50,000	0.01	500	22.25	0.044
100,000	0.0005	50	7.07	0.141					
100,000	0.001	100	9.99	0.100					
100,000	0.002	200	14.13	0.071					
100,000	0.005	500	22.30	0.045					
100,000	0.01	1000	31.46	0.031					

最後の列は標準偏差を期待値で割った変動係数 cv＝sd／np＝$\sqrt{(1-p)/np}$ であり、cv は相対的な変動を表すものである。cv をみると、集団の大きさが10万人であれば、どの死亡率に対しても期待値の10％程度の変動であり、十分安定していることが分かる。一方で、２万人程度の場合は相対的な変動が大きいため、支払う保険金額が大きくなって経営が不安定となる可能性がある。契約者数が少ない場合は、このように死亡者数の変動幅が比較的大きいため、現実の保険料は経営の安全性を見込んで、標準偏差の２倍程度と大きめに設定することが多い。

〔本節の解答〕

Q 1 ：１人当たりは100万円×（100万人×0.02）÷100万人＝２万円となる。

3　確率の意味するもの［確からしさの実践］

◇ 生まれてくるこどもの性別

ある家族に次に生まれてくる子どもが男の子である確率は1/2だろうか。実際に、男女の比率は等しいかどうか調べてみよう。「男女の生まれる確率が等しい」ときは、ゆがみのないコイン投げと同様に、二項分布を用いて確率を計算することができる。

◇ 歴史的な例

1629年～1710年にロンドンで生まれた子どもの数は5000人から15000人の間だったが、比率（男／女）は表１のとおりであり、82年間のすべての年について男が多い。男女の生まれる確率が等しければ、男が女より多い確率と、女が男より多い確率は等しい。そのとき、82年続けて男が多くなる確率は近似的に$1/2^{82}=2.067952\times10^{-25}$であり、これは極端に小さい。このような小さな確率で発生する事象が実際に起きたときに、単に珍しいとするのではなく、確率を計算する前提とした「男女の生まれる確率が等しい」という想定（これを**仮説**と呼ぶ）を疑うのが、**統計的仮説検定**の考え方である。

表１　ロンドンで1629年～1710年に生まれた子どもの性別比率（男／女）

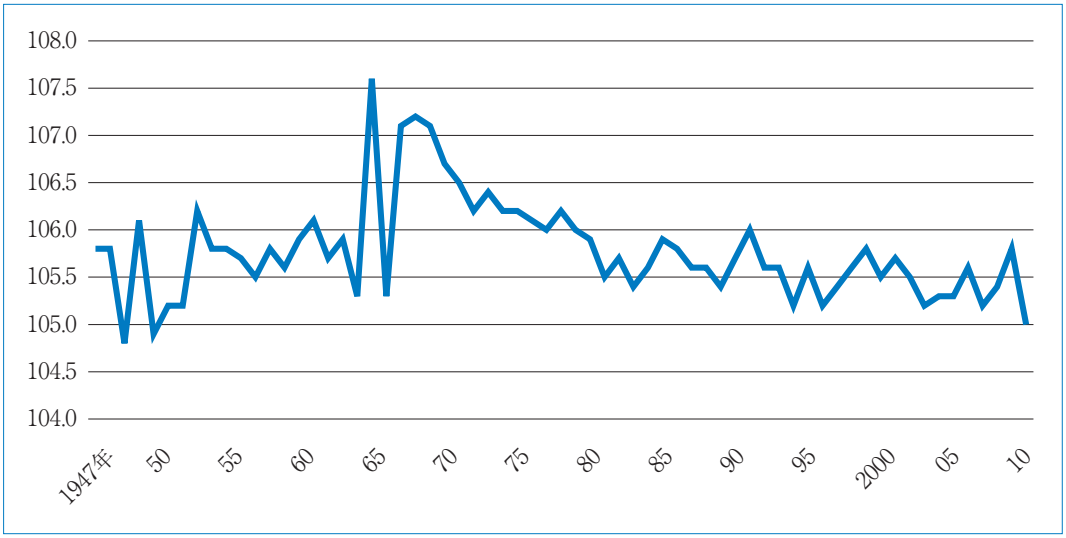
1.114	1.090	1.078	1.088	1.066	1.045	1.036	1.068	1.055	1.082	1.122	1.035	1.052
1.112	1.038	1.028	1.033	1.110	1.074	1.057	1.121	1.062	1.138	1.107	1.080	1.082
1.091	1.085	1.033	1.048	1.154	1.147	1.156	1.086	1.109	1.063	1.053	1.083	1.055
1.092	1.116	1.098	1.064	1.053	1.043	1.065	1.060	1.121	1.035	1.089	1.034	1.040
1.044	1.024	1.059	1.063	1.033	1.064	1.072	1.054	1.061	1.083	1.037	1.039	1.026
1.051	1.082	1.056	1.038	1.105	1.062	1.073	1.078	1.049	1.011	1.065	1.075	1.072
1.090	1.081	1.062	1.048									

◇ 日本の出生性比

人口学では、生まれてくる子どもについて、100×（男／女）を**出生（しゅっしょう）性比**と呼ぶ。図１を見ると、第２次世界大戦後の日本については、出生性比は105から107の近くで比較的安定している。65年間連続で男が多い確率は近似的に$1/2^{65}=2.710505\times10^{-20}$であり、これも非常に小さいため、日本についても「男女の生まれる確率が等しい」とは言えない。

なお、特定の家族では、あらかじめ子どもの男女比を知ることはできないが、日本全体でみると男女の比率は非常に安定している。ここでも、**大数の法則**の傾向が確かめられる。

図１　日本の出生性比（1947年～2011年）



◇ 仮説検定の考え方

以上の計算では、非常に確率の小さな事象が発生したため「男女の生まれる確率が等しい」という仮説を疑った。しかし、単に確率の小さな事象を観測しただけでは、仮説を疑うことはできない。子どもの数を２、４、10、100、1000と増やしていくと、ちょうど半数が男となる確率は、0.5、0.375、0.24609375、0.07958924、0.02522502と次第に小さくなる。もし、100万人の新生児のうち、ちょうど50万人が男だったら、この仮説をおかしいとは考えないであろう。しかし、その確率は0.0007978844とかなり小さい。したがって、男の生まれる確率が1/2であるという仮説を疑う根拠については、さらに考える必要がある。男が半数より多い年を数える場合でも、「65年中、63年は男が多い」、「65年中、62年は男が多い」などの結果が得られたときに、仮説を疑うべきだろうか。伝統的な統計学では、ある結果が起きたとき「それと同等以上に珍しい結果すべてを合わせた確率が小さいとき」に仮説を疑う。これを**すその確率**（tail area probability）と表現する。

二項分布 $B(n,\ p)$ における仮説検定

実際の観測値 x_{obs} に基づいて、確率に関する仮説 $H: p=p_0$ の妥当性を判断する手順は、以下のようになる。

$B(n, p_0)$ において、確率変数 x が $|x-np_0| \geq |x_{obs}-np_0|$ を満たす確率を計算して、これが非常に小さい場合に仮説を疑う。

ここで np_0 は、仮説が正しいときの確率変数 x の期待値であり、np_0 に近い x の確率が大きくなる。統計学ではすその確率を **P 値**（または p 値、P-value）と呼ぶ。

例：ある実験の成功確率が $p_0=0.8$ という仮説を考える。$n=1000$ 回のうち $x_{obs}=780$ 回が成功だったとすると、$np_0=800$ だから、二項分布 $B(n, p_0)$ において $|x-np_0| \geq |x_{obs}-np_0|=20$、すなわち、$x \leq 780$ または $x \geq 820$ となる確率である P 値を求める。その結果は0.126と、かなり大きいため、仮説を疑う根拠は小さい。なお、この場合でもちょうど $x=780$ となる確率は、0.009とかなり小さいことに注意してほしい。

この例で、もし $x_{obs}=760$ 回が成功だったとすると、P 値は0.0022となり、仮説は強く疑われる結果となる。

ひのえうまの出生比率

図１を見ると、1965年から1967年にかけて、105.3、107.6、105.3という例外的な変動を示しています。この時期の出生数は180万人から200万人程度だったのが、この３年間の出生数は、1,823,697人、1,360,974人、1,935,647人と激しく変動していて、1966年が激減しています。1966年に生まれた男は705,463人、女は655,511人でした。1961年から1970年の10年間の平均では出生性比は106.3、男が生まれた比率は106.3/(100＋106.3)＝0.515269です。これを1966年に男が生まれる確率として、仮説 $p=p_0$ (＝0.515269) に対するＰ値を計算してみましょう。期待値は $np_0=1360974\times0.515269=701268$ だから、$|x-np_0|\geq|x_{obs}-np_0|=4195$ となる確率を求めると、6.15951×10^{-13} と、Ｐ値は非常に小さい。つまり1966年の出生比率が他の年と違っているのは偶然とは考えられません。

その原因が「丙午（ひのえうま）」の迷信によって女の子が困らないように配慮した人たちの行動によるものであることはよく知られています。丙午の年には出産を控えただけでなく、年初や年末に生まれた女子の出生届を前後の年にずらした可能性が高いと考えられます。

◇ ベイズの公式

次の問題を考えよう。

Ｑ１：ある試験の受験者に占める女性の割合は0.6で、合格率は女性が0.4、男性が0.3であった。合格者のうち女性の割合はいくらか。

これは確率ではなく、比率の問題であり、容易に解くことができる。次の問題はどうだろうか。

Ｑ２：ある試験の受験者に占める女性（Ｆ）の割合は0.6で、合格率は女性が0.4、男性（Ｍ）が0.3であった。合格者（Ｅ）の受験票を１枚取り出したとき、それが女性である確率 $P(F|E)$ はいくらか。

これらの２つの問題がそっくりなのは、**Ｑ２**では受験票を繰り返し抜き出す実験を想定することができて、相対度数による確率の解釈が自然にできるからである。

もう少し一般的に拡張した次の問題を考える。

Ｑ３：$F_1, F_2, \cdots, F_m$ を排反全事象（$F_1, F_2, \cdots, F_m$ は互いに排反で、そのうちの１つが必ず起きる）とする。確率 $P(F_1), P(F_2), \cdots, P(F_m)$ および、各 F_j が与えられたときの事象Ｅの条件付き確率 $P(E|F_j)$（$j=1, \cdots, m$）が与えられている。事象Ｅが起きたときの事象 F_i の条件付き確率 $P(F_i|E)$ はいくらか。

Q４：（病気の診断）友人が住んでいる地域で、ある病気に感染している人の割合は0.1％である。最新のデータベースによれば、血液検査で陽性（A）になる確率は、感染している（M）とき $P(A|M)=0.8$、感染していない（$\overline{M}$）とき $P(A|\overline{M})=0.1$である。昨日、友人が検査を受けたところ陽性と告げられた。友人が感染している確率 $P(A|\overline{M})$はいくらと判断したらよいか。

Q５：（病気の診断、つづき）友人がさらに別な検査を受けたところ、その結果は陽性であった。この検査で陽性（B）になる確率は、感染している（M）とき $P(B|M)=0.9$、感染していない（$\overline{M}$）とき $P(B|\overline{M})=0.2$である。友人が感染している確率 $P(M|A\cap B)$はいくらと判断したらよいか。なお、検査結果の事象 A と事象 B は M，$\overline{M}$ のいずれのときも、確率的に独立と想定してよい。

◇ 判断確率の考え方

Q４と Q５の例では、感染しているかどうかは、精密な検査を受ければ確定している。したがって、同じ友人について実験を繰り返しても、M か $\overline{M}$ かは変らない。つまり、友人が病気である確率を相対度数によって理解することは難しい。このように、特定の人に対して病気の可能性を確率で表現することは妥当なのだろうか。

ベイズ統計学では、相対度数ではなく、**判断確率**（**主観確率**ともいう）による定義を用いる。そこでは、合理的な行動に関する基準から、判断の根拠となる指標が導かれ、それが確率の法則を満たすことが導かれている。簡単な例で判断確率の考え方を紹介しよう。

◇ 判断確率の構成

今日の夕方の天気に関する問題を考える。可能な結果は F_1（雨）、F_2（曇りまたは晴れ）の２通りとする。ここで、ある人「あなた」に２つのくじを提供する。賞金はいくらでもよいが、小さな額とする。くじ ℓ_1では、天気が雨のときに賞金を受け取り、曇りまたは晴れのときは受け取れない。くじ ℓ_2は次のように作る。箱の中に100個の玉があり、そのうち a 個が赤、それ以外は白とする。この箱をよくかき混ぜて１個を取り出し、それが赤のときだけ賞金を受け取る。$a=100$の場合、ℓ_2を選べば確実に賞金がもらえる。$a=0$の場合は、天気に関する判断に関わらず ℓ_1を選ぶのが合理的である。a の値を変えていって、ある値で、どちらのくじでも同じ程度に好ましいと判断する場合、その a が事象 F_1に関する「あなた」の判断確率である。

判断確率と呼ぶ理由は、「あなた」が天気予報の専門家であるか、最近の天気図や気象衛星の画像を見たなど、持っている情報によって判断が異なる可能性を認めているからである。

Q６：（前節で扱った問題）箱の中に数字を書いた６枚のカードがあり、その数字は、３枚が「１」、２枚が「２」、１枚が「４」である。元に戻さずにカードを取り出して、１枚目を伏せた状態で２枚目に「２」が出た（この事象を E とする）とき、１枚目が「２」である確率 $P(F_2|E)$はいくらか。

この問題は、判断確率を理解するために有用である。１枚目のカードを取り出す手順から $P(F_2)=2/6$ は明らかであり、伏せてあるカードの数字は、２枚目を見ても変化しない。それにも関わらず判断確率は変化する。

１枚目のカードが伏せてある状態で、その数字が「２」である確率を問う問題で、相対度数の定義を使う場合、繰り返し実験とは、そのカードを開いてみることに対応する。この実験は何回繰り返しても、数字が「２」の相対度数は０または１である。一方、判断確率であれば、ベイズの公式に従って、自然に確率の評価が修正される。極端な場合、残りのカードをすべて取り出してみれば、伏せてあるカードの数字は確率１で当てることができる。

このように、判断確率を認める立場では、追加的な情報が与えられれば確率が変化することがある。その合理的な更新の手順を与えるものがベイズの公式であり、これは、正確には、個人の合理的行動に関する公理体系から導かれる定理である。

◇ 相対度数による確率と判断確率

厳密に考えると、繰り返し可能な実験はほとんど無意味になる。サイコロ投げについては、よく回転させるなどの同じ条件で実験を繰り返す必要があることは、ある目を上にしてそっと置けば、結果は確実にその目になることから分かる。しかし、正確に同じ条件で回転を与えて、同じ位置から転がせば、結果は同一になるから、相対度数による定義は意味がなくなる。病気の例で「ある人の検査の結果が陽性（A）であったとして、実際に病気（M）である確率はいくらか」という問題でも、同じ人に検査を繰り返しても病気であるか健康であるかは変わらず、相対度数は０または１だから、判断としての確率以外は意味を持たせることは難しい。

◇ 判断確率の表記

確率の評価は前提となる知識 H によって異なることを明示的に表すため、ある命題 F の判断確率を、$P(F|H)$ と書くことがある。追加的な情報 E が与えられたときの判断確率は $P(F|H\cap E)$ と書くことになるが、この事後確率を求める手順がベイズの定理である。ただし、事前情報 H を固定して議論を進める場合には H を省略して、$P(F)$ や $P(F|E)$ と略記しても混乱はない。

◇ 独立性と情報

新たな情報である E を観測しても、判断確率 $P(F)$ が変化せず、$P(F|E)=P(F)$ となる場合があるだろうか。ベイズの公式 $P(F|E)=P(F\cap E)/P(E)$ の分母を払うと $P(E)P(F|E)=P(F\cap E)$ である。これを、通常の独立性の定義である $P(E)P(F)=P(F\cap E)$ と比較すると、$P(F|E)=P(F)$ が成り立つのは、E と F が独立の場合であることが分かる。事象 E と事象 F が無関係であれば、追加的な情報はないため、判断確率も変化しないのは当然である。

◇ 事象と命題

判断確率の場合は、繰り返しを想定する**事象**（event）だけではなく、原理的に繰り返せないような**命題**（proposition）に関しても、確率を考えることができる。たとえば、「紫式部が宇治十帖を書いた」という命題の確からしさについては、相対度数の考え方は適用できないが、判断確率であれば適用することができる。実際、文献の著者に関する判断確率は次第に利用される機会が増えてきている。

Q７：（３つの箱のパラドックス）
箱 B_1，B_2，B_3があり、それぞれの中身は B_1：(G, G)、B_2：(S, S)、B_3：(G, S)とする（G, S はそれぞれ、金貨、銀貨を表す）。いま、１つの箱を選んだとき、１枚目が金貨だったとすると、２枚目も金貨である確率はいくらか。

Q８：（３人の囚人のパラドックス）仮釈放を申請している３人の囚人 A，B，C は同程度に良好な服役記録がある。判事が３人のうち２人を釈放する決定を下したことは囚人に知れたが、どの２人かは知らされていない。囚人 A は、看守に他の２人の囚人のうち釈放される１人の名前を聞いてみようと考えた。この質問をする前には A は次のように考えた。
　自分が釈放される確率は2/3である。A は、仮に看守が「B は釈放される」と答えたならば、釈放されるのは「A と B」か「B と C」かのいずれかだから、A が釈放される確率は1/2に減ってしまう。
　結局 A は、質問をすることは中止した。しかし、A の計算は誤りである。このことを説明せよ。

〔本節の解答〕

Q１：$2/3 \fallingdotseq 0.67$

Q２：$2/3 \fallingdotseq 0.67$

Q３：$P(F_i|E) = \dfrac{P(F_i)P(E|F_i)}{\sum_j P(F_j)P(E|F_j)}$

Q４：$(0.001 \times 0.8)/(0.001 \times 0.8 + 0.999 \times 0.1) \fallingdotseq 0.0079$

Q５：$(0.0079 \times 0.9)/(0.0079 \times 0.9 + 0.9921 \times 0.2) \fallingdotseq 0.131$

Q６：$P(F_2|E) = 1/5$

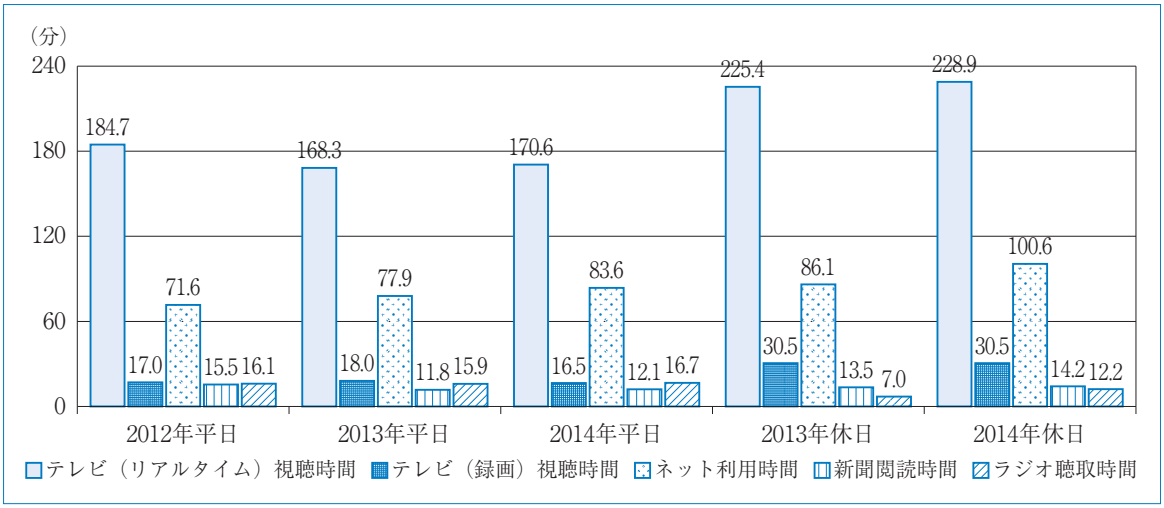
Q７：$P(B_1|G) = 2/3$

Q８：B と C が釈放される場合に、看守が「B が釈放される」と答える（この命題を b とする）確率 π に依存して A が釈放される（命題 A とする）事後確率が次のように求められる。$P(A|b) = 1/(1+\pi)$、特に $\pi = 1/2$ なら $P(A|b) = 2/3$

第4部 統計的探究の実践 Ⅲ
～モデルに基づいて現象を理解する～

1　視聴率調査の仕組みは？[視聴率データの分布は正規分布で近似できる]

図１　主なメディアの平均利用時間の推移（全年齢）



資料：総務省「2014 年情報通信メディアの利用時間と情報行動に関する調査」

メディアの利用時間のなかで、テレビの視聴時間（平日）は、過去３年を見れば最も長く、次のデータから、年齢が高齢になるほど長くなることが分かる。テレビ業界にとって、どのようなテレビ番組が多く見られるかは一大関心事で、視聴率の調査結果に一喜一憂することとなる。

	全年齢	10歳代	20歳代	30歳代	40歳代	50歳代	60歳代
テレビの視聴時間（分）	170.6	91.8	118.9	151.6	169.5	180.2	256.4

Q１：あなたの回りの人は、平日の１日に何分間テレビを視聴しているだろうか。次の表に調べた結果をまとめなさい。

（例）A さん							
100分							

STEP 1：Problem 問題　課題の設定

◇ テレビの視聴率調査はどのように実施されている？

関東地区（関東１都６県（東京都島部を除く））の世帯数は約1800万世帯（2016年10月３日現在）である。これらの世帯の中で、ある番組を見ている世帯の割合を表したものが番組視聴率である。全世帯の視聴率を完全に把握

するためには全世帯を調査するしかないが、現実的には不可能である。そのため、実際に視聴率を知るためには、標本となる世帯を抽出し、その結果から全世帯の視聴率を推定する手順をとる。代表的な調査機関であるビデオリサーチ社の**視聴率調査**では、関東地区で調査対象となる世帯数は900世帯である。（ビデオリサーチ『視聴率調査ハンドブック』2016より引用）

Q２：約1800万の世帯に対して、視聴率調査の**調査対象**が900世帯ということは、調査対象となる**標本**は全体の何％になるのか？

全体の　　％

なぜ、関東地区において、たった900世帯で視聴率を推定することができるのだろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 視聴率調査の模擬実験

航平君たちは、視聴率調査の正確性と調査対象世帯数の関係を調べるために、青玉を「番組を見た世帯」、白玉を「番組を見ていない世帯」として、視聴率調査の模擬実験を考えた。

模擬実験用の道具例

＜道具＞

・白玉：60個、青玉：20個、玉を入れる袋

・玉を取り出す器（紙コップの底：12個、磁石のふた：15個、ステンレス皿：30個等）

＜方法＞

① 「**サンプルサイズ** 4（取り出す玉の数）で推定」、「サンプルサイズ12で推定」、「サンプルサイズ15で推定」、「サンプルサイズ30で推定」の４つのグループに分けて200回の実験を行う。

② 袋の中身をよくかき混ぜ、決められた数の玉を“器”または“素手（サンプルサイズ４）”で取り出す。決められた個数に足りなかった場合は、袋の中身を見ずに手で取り出して調整する。

③ 青玉が何個入っていたかを記録する。

④ １回の調査ではサンプルサイズの違いによる傾向は分からないので、取り出した玉を袋に戻して、よくかき混ぜ、②を繰り返す。

⑤ 結果をヒストグラムにまとめて傾向を調べ、青玉の割合（視聴率）を推定する。

よくかき混ぜて玉を取り出さないと、取り出した標本が偏ってしまい、結果に大きく影響するので、注意しなければいけないね。

Q３：実際の視聴率調査では、どのように調査世帯を選び、標本を抽出しているのだろうか。予想しなさい。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 模擬実験の結果を表にまとめよう！

模擬実験をそれぞれ200回行った結果、次の結果が得られた。

表 1 視聴率調査の模擬実験の結果

① サンプルサイズ：4

青玉の個数	0 個	1 個	2 個	3 個	4 個
青玉の比率	0 %	25%	50%	75%	100%
実験結果（回）	70	80	40	10	0

② サンプルサイズ：12

青玉の個数	0 個	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個	9 個	10個
青玉の比率	0 %	8 %	17%	25%	33%	42%	50%	58%	67%	75%	83%
実験結果（回）	5	30	50	45	50	15	5	0	0	0	0

11個	12個
92%	100%
0	0

③ サンプルサイズ：15

青玉の個数	0 個	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個	9 個	10個
青玉の比率	0 %	7 %	13%	20%	27%	33%	40%	47%	53%	60%	67%
実験結果（回）	3	6	30	42	54	36	18	6	5	0	0

11個	12個	13個	14個	15個
73%	80%	87%	93%	100%
0	0	0	0	0

④ サンプルサイズ：30

青玉の個数	0 個	1 個	2 個	3 個	4 個	5 個	6 個	7 個	8 個	9 個	10個
青玉の比率	0 %	3 %	7 %	10%	13%	17%	20%	23%	27%	30%	33%
実験結果（回）	0	2	1	3	5	7	28	46	56	18	14

11個	12個	13個	14個	15個	16個	17個	18個	19個	20個	21個	22個	23個
37%	40%	43%	47%	50%	53%	57%	60%	63%	67%	70%	73%	77%
10	8	2	0	0	0	0	0	0	0	0	0	0

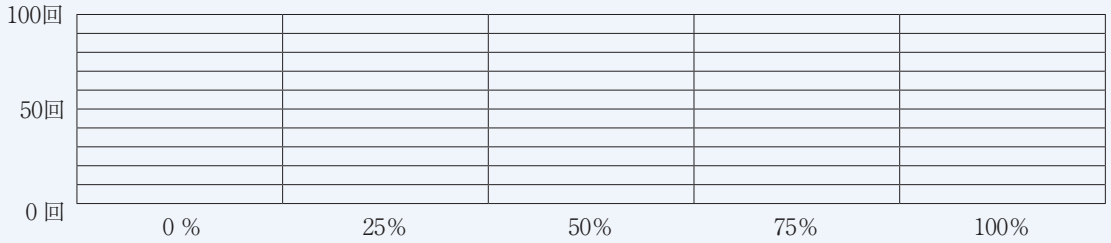
24個	25個	26個	27個	28個	29個	30個
80%	83%	87%	90%	93%	97%	100%
0	0	0	0	0	0	0

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

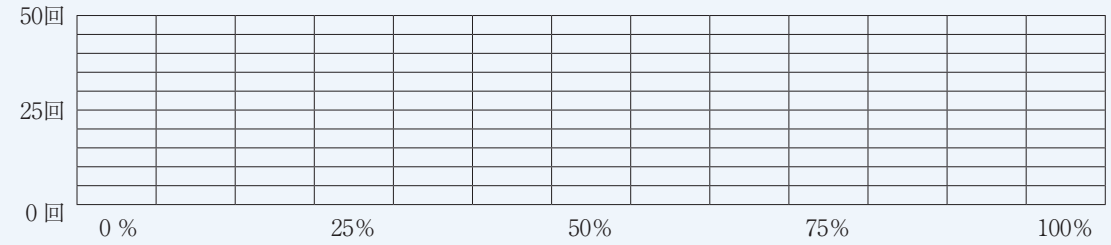
◇ 模擬実験の結果をグラフにまとめ、傾向を比較

Q 4 ：表 1 のデータをもとにグラフを作成しなさい。

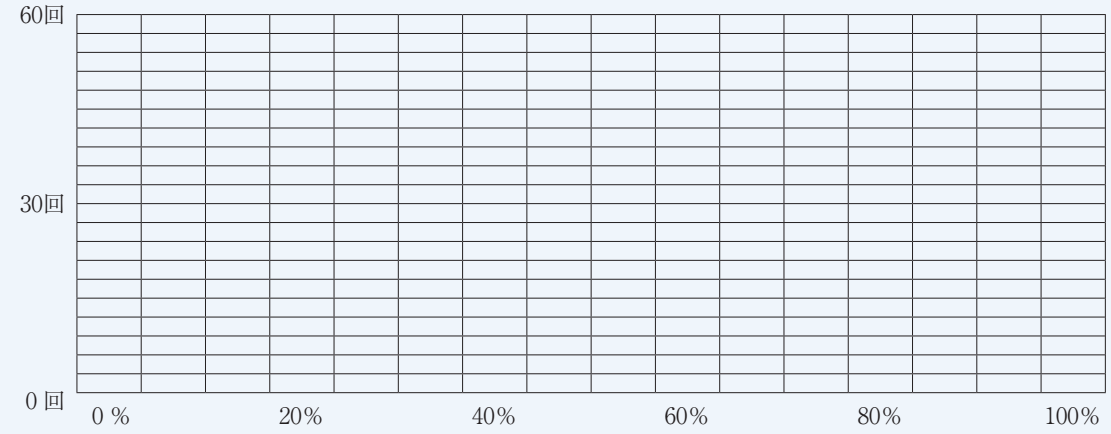
① サンプルサイズ： 4



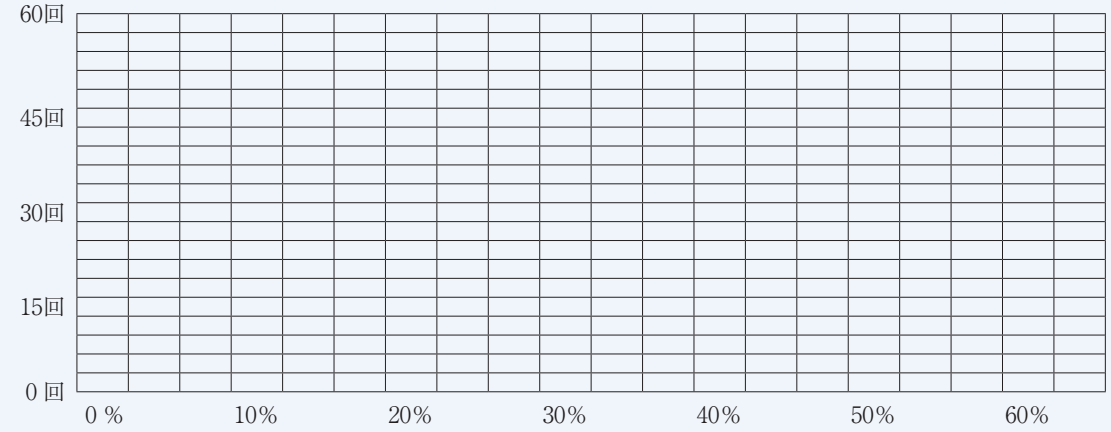
② サンプルサイズ：12



③ サンプルサイズ：15



④ サンプルサイズ：30

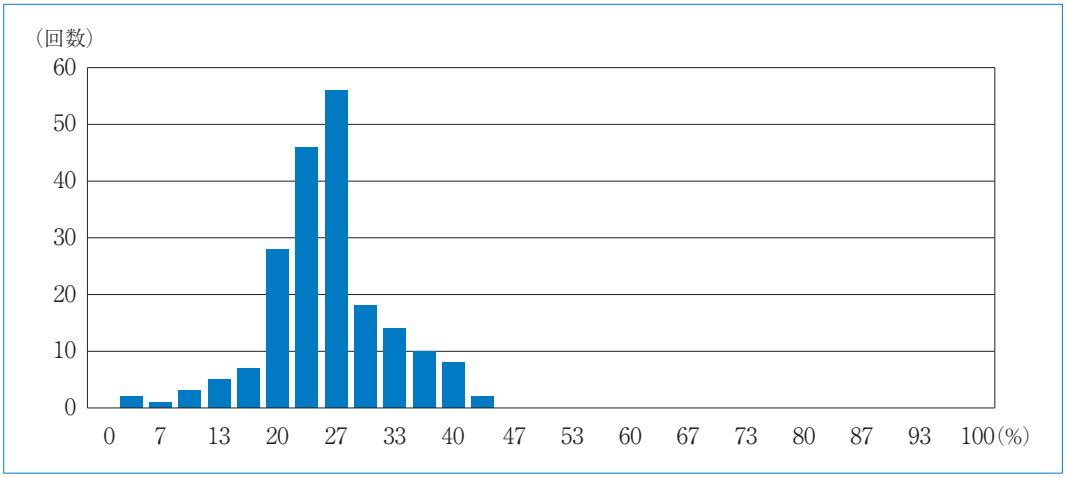


グラフからどんなことが読み取れるか、考えよう。

STEP 5：Conclusion 結論　結論を導く

◇ 模擬実験の結果からみる視聴率調査の仕組み

図 2　視聴率調査の模擬実験の結果：サンプルサイズ30の場合



Q５：実験結果のグラフをサンプルサイズの大きさの順に眺めると、すべての玉に占める青玉の割合25％の付近に、次第にグラフが収まってくることが分かる。それは何故か？

Q６：模擬実験の結果を踏まえ、関東地区では、たった900世帯だけ調べるので良い理由について、まとめなさい。
〔説明〕

実際の視聴率調査では、後で学ぶ区間推定の公式から、900世帯中180世帯が番組を見ていた場合は約20％±2.7％、90000世帯中18000世帯が見ていた場合には約20％±0.27％と視聴率を推定できるんだ。

さらなる発展を目指してみよう！

◇ 視聴率調査の模擬実験から得られた山形の分布は？

航平：視聴率調査の実験結果の分布は山形の分布になったね

理恵：サンプルサイズが大きい分布ほど、きれいな山形に見えるわ

公介：一般に、標本平均（標本比率）の分布は、サンプルサイズ（実験回数）が大きいほど**正規分布**と呼ばれる**単峰**性の釣鐘形の分布に近づくことが知られているんだ

正規分布

全国の高校生の身長の分布は、右の図のように、平均の近くの人数が一番多く、そこから遠く外れるほど人数が少なくなり左右対称の釣鐘型になることが知られています。このような分布の型を正規分布と言います。正規分布のグラフは中央が一番高く、両側に向かってだんだん低くなっていき、左右対称の釣鐘型をしていますが、正規分布の場合、この中央の一番高い位置が平均となります。

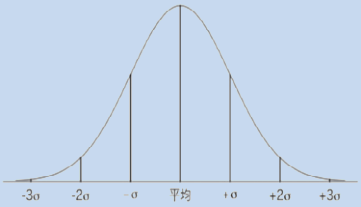
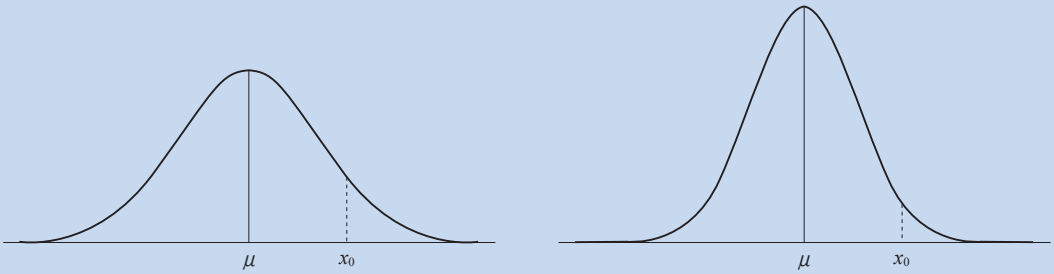


図３は２つの正規分布のグラフを表わしています。２つのグラフはいずれも平均が μ の正規分布ですが、右の正規分布の標準偏差 σ（分散の平方根）は、左の図の標準偏差に比べて小さい値になっています。標準偏差の値が大きくなるほど釣鐘型の曲線が横に伸びて裾野が広がる形になりますが、これは形が横に伸びただけで、正規分布の曲線の本質的な形状は決まった形をしています。

図３　いろいろな正規分布の曲線



平均値や標準偏差（分散）がどのような値でも、正規分布は次の性質をもっています。

正規分布の性質

平均を μ、標準偏差 σ としたとき、

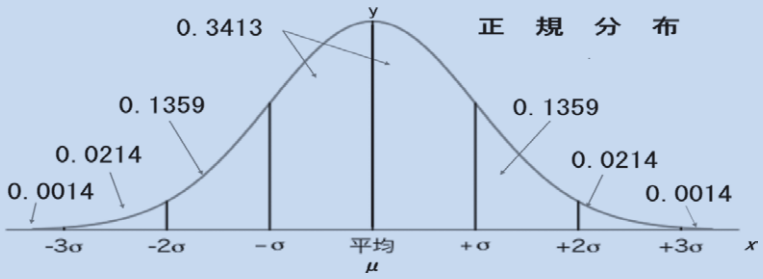
区間（$\mu-\sigma$、$\mu+\sigma$）　に入る確率は、0.683である。

区間（$\mu-2\sigma$、$\mu+2\sigma$）に入る確率は、0.954である。

区間（$\mu-3\sigma$、$\mu+3\sigma$）に入る確率は、0.997である。

この μ と σ の値で示される区間の確率が決まっている性質をグラフ上に表したのが、図４です。

図４　正規分布の性質



社会現象あるいは自然現象に現れるバラツキ（散らばり）は正規分布（normal distribution）に従うと見なせるものが多く、正規分布は統計学の理論上も応用上も非常に重要な分布です。確率的に変動する変数（**確率変数**）X のとる値の範囲に対応して確率が与えられる関数（**確率密度関数**）は、正規分布の場合、次の式で与えられます。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

平均 μ は分布の中心となる位置を表し、分散 σ^2 の標準偏差 σ は平均からのバラツキの大きさを表します。正規分布は、英語表記の “normal distribution” の頭文字の N を使って、$N(\mu, \sigma^2)$ とも表されます。

正規分布は平均と分散（標準偏差）の違いによってグラフの形状が違ってくるので、平均と標準偏差の値で変動しないような形に変換することができれば、異なる平均と標準偏差の正規分布を互いに比較することができます。そこで、位置の尺度である平均を 0 、バラツキの尺度である標準偏差を 1 に変換することを考えます。確率変数 X を次式で変換すると、変換された確率変数 Z は、平均が 0 、標準偏差が 1 の分布になります。

$Z=(X-\mu)/\sigma$

この変換を**標準化**といい、標準化された Z は、**標準化得点**、または Z 値（z score）などと呼ばれ、平均が 0 、標準偏差が 1 の正規分布 $N(0,1)$ に従うことになります。この $N(0,1)$ を特に**標準正規分布**（standard normal distribution）といい、その確率密度関数は次の式で与えられ、グラフは一意的に定まります。

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

偏差値（standard score）

試験結果に記される**偏差値**は広く知られているにもかかわらず、その意味が十分に理解されていない、統計学に基づいた指標です。あたかも試験得点のように 1 点上がったとか、下がったとか、受験生の間で会話されることがありますが、意味も分からず話されていることも多いと思われます。

偏差値は、対象となる集団の得点分布において、分布の中心からどの程度乖離しているかを表す数値であり、標準化された得点(Z)を用いて次式で定義されます。

偏差値＝50＋10×Z＝50＋10×（得点－平均）÷標準偏差

このように、偏差値とは、得点について、中心を50として、平均からの乖離を 1 標準偏差分が10となるように変換された値を示します。

Z の式において、分子が点数で分母も点数だから、当然、偏差値は単位のない数になります。1970年頃から受験業界で学力成績の指標として使用されるようになりました。得点の分布がほぼ左右対称の釣鐘型（正規分布）となるなら、全体の中でその人がどのくらいに位置するかが分かります。たとえば、偏差値70の成績順位は全体の上位2.3％に位置します。受験者が1000人ならば、23位前後の順位だとおおよその順位の見当もつけられます。

それでは、「標準正規分布」は、一体どのように活用できるのだろうか？

［例題 1］

毎年行われる学校保健統計調査によると、高校 3 年生（17歳）の男子の身長の平均は約170cm、標準偏差は 5 cm である。また、身長のデータは正規分布に従うことが知られている。このとき、身長180cm 以上の高校 3 年生は、全体で何％いるか。

［解説］

男子の身長を Xcm とすると、X の分布は N（170, 5^2）である。

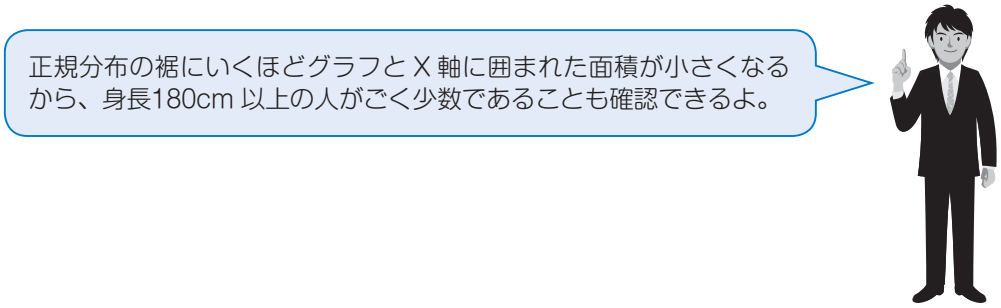
したがって、Z＝(X－170)/5 は N（0, 1）の標準正規分布に従う。

X＝180のとき、Z＝(180－170)/5＝2 であるから、

$$P(X \geqq 180) = P\left(\frac{X-170}{5} \geqq \frac{180-170}{5}\right)$$

$$= P(Z \geqq 2) = 0.5 - P(0 \leqq Z \leqq 2) = 0.5 - 0.4772 = 0.0228$$

したがって、身長180cm 以上の高校 3 年生は、全体で約2.3％いる。



Q 7：公介君の身長は176cm である。公介君は全国の高校 3 年生で高い方から数えて何％位の位置にいるか。

〔本節の解答〕

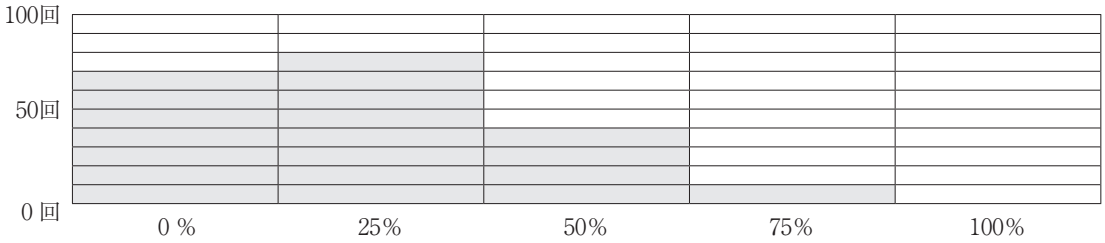
Q 1：略（授業内のグループメンバーの視聴時間等）

Q 2：900 ÷ 18000000 ＝ 0.00005　したがって、全体の0.005％が調査対象

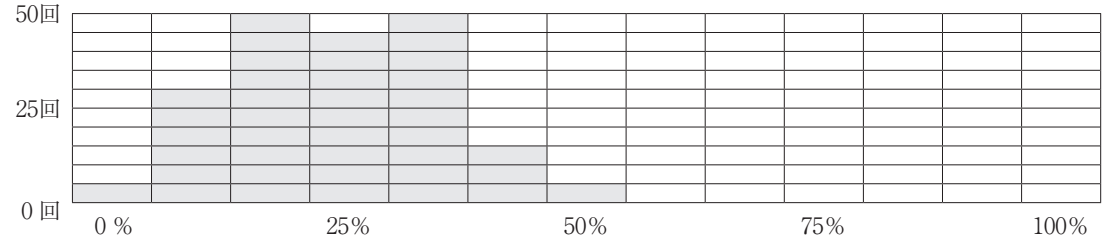
Q 3：標本抽出の方法：系統抽出法

国勢調査の調査区を無作為に抽出し、抽出された調査区の世帯を住民基本台帳等から調べ、調査区内の世帯数を求め、各世帯に番号を振る。次に、最初に抽出する世帯をランダムに選び、そこを起点として、（調査区内の世帯数 / 抽出世帯数）の値を間隔として世帯を抽出し、選ばれた世帯に調査協力をお願いする。

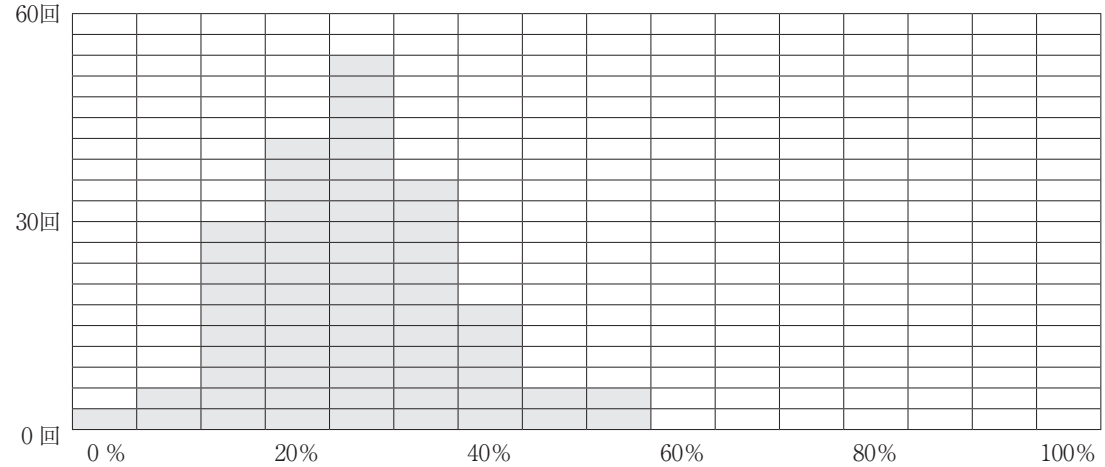
Q４：①　サンプルサイズ：４



②　サンプルサイズ：12



③　サンプルサイズ：15



④　サンプルサイズ：30　略

サンプルサイズが大きくなるにつれて、実験結果として得られた青玉の比率は20～35％の付近に密集し、グラフは尖った山の形に近づく。

Q５：略

Q６：略

Q７：例題１と同様にして、X＝176のとき、

$P(X \geqq 176) = P\left(\dfrac{X-170}{5} \geqq \dfrac{176-170}{5}\right) = P(Z \geqq 1.2)$ を利用すればよい

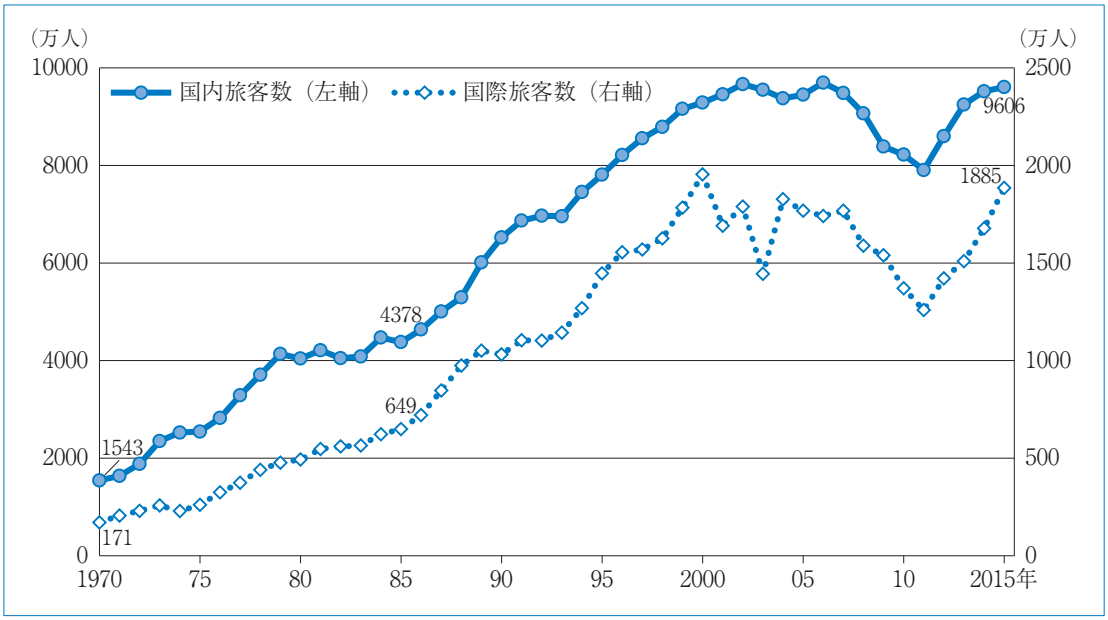
答：高い方から数えて、11.5％位の位置にいる。

正規分布表 P(Z≧z)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

2　日本では航空交通が一番安全！？
［めったに起きない事象の分布はポアソン分布で近似できる］

図１　日本の航空旅客数の推移



資料：国土交通省「航空輸送統計年報」

日本の国内交通の旅客数は、2006年度をピークに右肩下がりとなっていたが、東日本大震災からの復興需要、LCC 参入による需要増により、2012年から増加に転じ、2015年には国内旅客数は9606万人、国際旅客数も1885万人となった。

Q１：グラフからみると、30年前の1985年の国内旅客数は4378万人、国際旅客数は649万人である。2015年の旅客数は、それぞれ1985年の旅客数の何倍になったといえるか？

国内旅客数：約　　倍、　国際旅客数：約　　倍

STEP 1：Problem 問題　課題の設定

◇ 日本の航空交通は一番安全！？

日本の交通手段の中で、最も安全なのは航空機と言われている。その根拠として用いられるのが、航空アナリスト杉浦一機氏が提示した「輸送実績１億人キロ当たりの死亡乗客数＝0.04人」、「10万飛行時間当たりの死亡事故件数＝0.07件」という指標である。（『知らないと損するエアライン＜超＞利用術』杉浦一機（2001）平凡社より引用）

Q２：「10万飛行時間当たりの死亡事故件数＝0.07件」を「死亡事故件数１件当たりの飛行時間」として算出すると、「死亡事故件数１件当たり、約143万飛行時間」であることが分かる。この結果に基づけば、東京－ニューヨーク間を２日かけて往復飛行を続けたとき、事故に遭うのはおよそ何年か？　なお、東京－ニューヨーク間の片道飛行時間は14時間、１年を365日間として計算しなさい。

東京－ニューヨーク間の便に乗り続けるとおよそ　　年。

さすが、「航空機は世界で最も安全な移動手段」と呼ばれることだけあるね。実際、下記の航空事故の発生件数には、落雷等の災害による機体の損傷等も含まれるから、重大な事故は限りなく少ないことが予想できるね。

それでは、最近の航空事故の発生件数はどうなっているのだろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 航空事故の発生確率を求めよう

2006年から2015年までの10年間の航空事故（大型機）の発生件数は、次のとおりである。

年	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
航空事故発生件数	3	5	3	6	0	1	8	1	4	3

資料：国土交通省運輸安全委員会 HP「航空事故の統計」

Q３：上記のデータから、この10年間の航空事故の平均発生件数を求めなさい。

平均発生件数：　　件

1年間に成田空港だけで、航空機の離着陸数は20万回以上ある。そのため、大型機だけでも離着陸数が最低3万回あり、Q3の結果を踏まえると、1年間の航空事故（大型機）の発生確率は、およそ0.0001以下であると見積もることができる。

1年間の航空事故の発生確率を0.0001、離着陸数を3万回として、1年間に航空事故の発生件数が4件以下である確率は、どれくらいなのだろう？

事故が「起きる」または「起きない」という2つの場合があるから、二項分布を使えば確率が分かるね。二項分布については、次ページのコラムに詳しく説明しているよ！

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 航空事故の確率分布（二項分布）表を完成しよう！

航空機が3万回離着陸するとき、航空事故がx回起きる確率をp(x)とすると、
x＝0, 1, … 7　までの確率は下記のとおりである。

$$p(0)=\left(\frac{9999}{10000}\right)^{30000}=0.0497\cdots$$

$$p(1)={}_{30000}C_1\left(\frac{1}{10000}\right)^{1}\left(\frac{9999}{10000}\right)^{29999}=0.1493\cdots$$

$$p(2)={}_{30000}C_2\left(\frac{1}{10000}\right)^{2}\left(\frac{9999}{10000}\right)^{29998}=0.2240\cdots$$

$$p(3)={}_{30000}C_3\left(\frac{1}{10000}\right)^{3}\left(\frac{9999}{10000}\right)^{29997}=0.2240\cdots$$

$$p(4)={}_{30000}C_4\left(\frac{1}{10000}\right)^{4}\left(\frac{9999}{10000}\right)^{29996}=0.1680\cdots$$

$$p(5)={}_{30000}C_5\left(\frac{1}{10000}\right)^{5}\left(\frac{9999}{10000}\right)^{29995}=0.1008\cdots$$

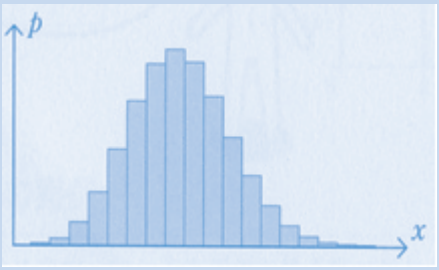
$$p(6)={}_{30000}C_6\left(\frac{1}{10000}\right)^{6}\left(\frac{9999}{10000}\right)^{29994}=0.0504\cdots$$

$$p(7)={}_{30000}C_7\left(\frac{1}{10000}\right)^{7}\left(\frac{9999}{10000}\right)^{29993}=0.0216\cdots$$

Q4：小数第3位を四捨五入し、小数第2位までで確率を表し、次の確率分布（二項分布）表を完成しなさい。

x	0	1	2	3	4	5	6	7	…
事故の発生確率 p(x)									…

二項分布（Binominal distribution）



サイコロを10回投げて 1 の目がちょうど 3 回出るとき、「1 の目が出る場合（確率 1／6）」と「1 の目が出ない場合（確率 5／6）」の 2 つがあり、その確率は、${}_{10}C_3\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^7$ で求めることができる。いま、サイコロを n 回投げて 1 の目が出る回数を x とおくと、$p=\frac{1}{6}$、$q=\frac{5}{6}$ として、$\mathrm{P}(\mathrm{X}=x)={}_nC_x\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{n-x}$ と表される。ここで、確率変数 X について、X の確率関数が $\mathrm{p}(x)=\mathrm{P}(\mathrm{X}=x)={}_nC_xp^xq^{n-x}$ $(p+q=1)$ であるとき、X の確率分布を二項分布といい、$\mathrm{B}(n, p)$ で表す。

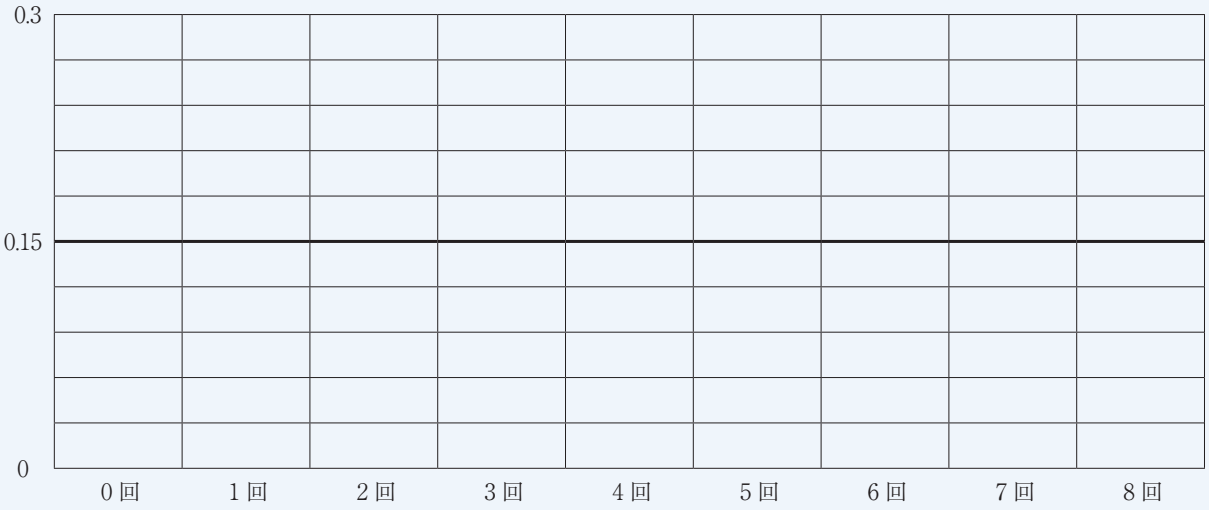
また、平均 μ、分散 σ^2 は次の式で表される。　$\mu=np$、$\sigma^2=npq$

それでは、「標準正規分布」は、一体どのように活用できるのだろうか？

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ 航空事故の発生件数が 4 件以下である確率を求めよう

Q 5 ： STEP 3 の確率分布表から次のグラフを完成させ、航空事故の発生件数が 4 件以下である確率を求めなさい。



航空事故の発生件数 X が 4 件以下である確率を $\mathrm{P}(\mathrm{X}\leqq 4)$ とおくと

$\mathrm{P}(\mathrm{X}\leqq 4)=\mathrm{P}(\mathrm{X}=0)+\mathrm{P}(\mathrm{X}=1)+\mathrm{P}(\mathrm{X}=2)+\mathrm{P}(\mathrm{X}=3)+\mathrm{P}(\mathrm{X}=4)$

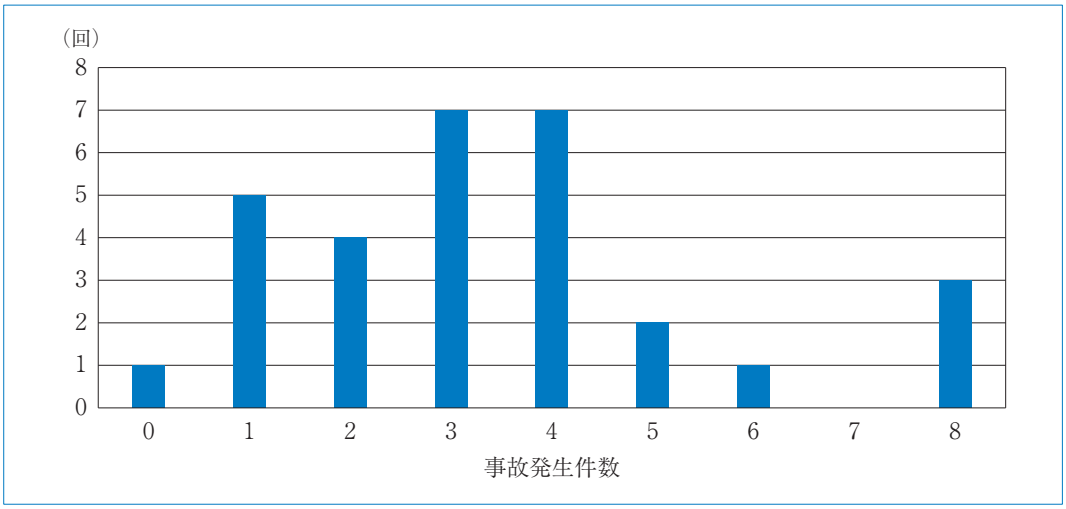
$=$

航空事故の発生件数が 4 件以下である確率：

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ 推定した航空事故の発生確率と過去30年の結果の違いは？

図 2　1986～2015年の30年間の航空事故の発生件数



資料：国土交通省 運輸安全委員会 HP「航空事故の統計」

図 2 によれば、過去30年間で事故の発生件数が 4 件以下の年度は24回あった。相対度数は24/30＝0.8であるので、過去30年間のデータから、大まかに航空事故の発生確率は0.8と見積もることができる。

Q 6 ：二項分布を用いて求めた確率と過去30年のデータから大まかに見積もった確率を比較した際、航空事故の発生確率にどのような違いがあるか。

Q 7 ：「日本の交通手段の中で航空交通は最も安全である」と言われている。自分の考えをまとめなさい。

〔発展〕２周目のサイクルへ：新たな課題を立て、解決する

◇ めったに起きない事象の確率分布：ポアソン分布

航平：正規分布では、「標本平均（標本比率）の分布は、サンプルサイズが大きいほど正規分布に近づく」と学んだよね？

理恵：正規分布を使えば、二項分布の大変な計算をしなくて助かるわ

健三：実際に正規分布を使って、航空事故の発生件数が 4 件以下である確率も求められるはずだね
正規分布を用いると、航空事故の発生確率はどのように求められるだろうか

正規分布を用いた場合

航空事故の発生件数を X とすると、発生確率 $p=0.0001$、離着陸数 $n=30000$から、

X の平均は、$E(X)=np=30000\times0.0001=3$

X の標準偏差は、$\sqrt{V(x)}=\sqrt{npq}=\sqrt{30000\times0.001\times0.9999}=1.731\cdots\fallingdotseq1.73$

したがって、$Z=\dfrac{X-np}{\sqrt{npq}}=\dfrac{X-3}{1.73}$

$$P(X\leqq4)=P\left(Z\leqq\frac{4-3}{1.73}\right)=P(Z\leqq0.578\cdots)$$

$0.578\cdots\fallingdotseq0.58$として、標準正規分布表から、求める確率は

$P(Z\leqq0.58)=0.5+P(0\leqq Z\leqq0.58)=0.5+0.219=0.719$

航空事故の発生件数が 4 件以下である確率は約0.72

航平：あれ、二項分布を用いたときの結果とずいぶん違うよ

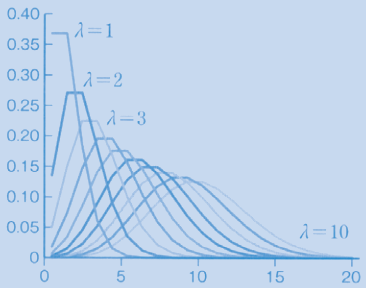
理恵：そうね、なぜかしら

Q８：航空事故に関する（Q５）のグラフを利用して、結果が違った理由がなぜだか考えよう。

一般に、航空事故の例のように、「試行回数 n が限りなく大きく」かつ「その事象の起きる確率が限りなく小さくなり 0 に近い」とき、その平均 $E(X)=np$ は一定の値 λ とみなすことができる。そのため、確率変数 X について、二項分布 $B(n,p)$の代わりに、**ポアソン分布** $P_o(\lambda)$を用いることが多い。

ポアソン分布

ポアソン分布とは、稀にしか起きない事象を大量に観測した際に用いられ、ポアソン（Siméon Denis Poisson）が二項分布から導きました。二項分布に比べて、実際の数値計算が簡単であるという特徴をもっています。プロシア陸軍で馬に蹴られて死亡した兵士の実際の数が、ほぼポアソン分布にあてはまるというのは、ロシアの統計学者のボルトキィヴィッチによる有名な実例です。ポアソン分布は、$\mu = np$（一定）で、$n \to \infty$、$p \to 0$が成り立つとき、



確率変数 X について、X の確率関数が

$p(x) = P(X = x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$（$\lambda$ は正の定数、$x = 0, 1, 2, \cdots$）

の確率分布を、$P_o(\lambda)$で表します

なお、e は定数で、$e = 2.71828\cdots$です。

また、平均 μ、分散 σ^2 は次の式で表されます。

$\mu = \lambda$、$\sigma^2 = \lambda$（平均と分散の値が一致）

ポアソン分布を用いた場合

ポアソン分布を用いて、航空事故の発生件数が 4 件以下である確率を求めてみよう。

ただし、$e = 2.7$とする。

航空事故の発生件数を X として、発生確率 $p = 0.0001$、運行回数 $n = 30000$より、

$\lambda = np = 3$ となる。

いま、$P(X = x) = \dfrac{3^x}{x!} e^{-3}$ であるから

$$P(X \leqq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$= \frac{3^0}{0!} e^{-3} + \frac{3^1}{1!} e^{-3} + \frac{3^2}{2!} e^{-3} + \frac{3^3}{3!} e^{-3} + \frac{3^4}{4!} e^{-3}$$

$$= \left(1 + 3 + \frac{9}{2} + \frac{9}{2} + \frac{27}{8}\right) e^{-3} = 16.375 \times \frac{1}{e^3} = 0.8319\cdots \fallingdotseq 0.83$$

航空事故の発生確率は約0.83

上に図示したポアソン分布のグラフから、$\lambda = 10$になるとほぼ正規分布の形になるから、ポアソン分布は、λ が小さい整数値のとき、使うと便利な分布なんだね。

試行回数 n に関係なく確率を計算できるところも、ポアソン分布の魅力だね。

したがって、航空事故はめったに起きない事象であるため、二項分布の代わりにポアソン分布を用いて、航空事故の確率を容易に見積もることができる。

ポアソン分布を用いると、めったに起きない事象の確率を簡単に求めることができた。

Q 9：$\lambda = 3$ として、航空事故（大型機）が年間 6 件以上起きる確率を求めなさい。

年間に成田空港だけでも20万回以上の離着陸数があるなかで、$\lambda = 3$ と仮定した場合で航空事故（大型機）が 6 件以上起きる確率は0.07程度である。そのため、この中には災害による機体の損傷等も含まれ、実際の航空事故の発生確率は非常に低いことを考慮すると、日本で「航空交通が最も安全である」と言われる理由も納得がいくのではないだろうか。

〔本節の解答〕

Q 1：国内旅客数：約2.19倍、国際旅客数：約2.90倍

Q 2：約280年

Q 3：$(3+4+1+8+1+0+6+3+5+3) \div 10 = 34 \div 10 = 3.4$
したがって、航空事故の平均発生件数は3.4件

Q 4：

x	0	1	2	3	4	5	6	7	…
事故の発生確率 p(x)	0.05	0.15	0.22	0.22	0.17	0.1	0.05	0.02	…

Q 5：グラフ略
航空事故の発生件数が 4 件以下である確率を $P(X \leqq 4)$ とおくと
$P(X \leqq 4) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) = 0.81$
航空事故の発生件数が 4 件以下である確率：0.81

Q 6：二項分布から求めた確率は0.81、過去30年間のデータから見積もった確率は0.8であり、航空事故の発生件数が 4 件以下である確率は、ほとんど変わらない。

Q 7：略

Q 8：二項分布を用いた（Q 5）の分布は、やや左よりの分布であり、正規分布のような対称性をもつ分布にならないから。

Q 9：$P(X \geqq 6) = 1 - P(X \leqq 5)$

$$= 1 - \{P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)\}$$

$$= 1 - \left(\frac{3^0}{0!} e^{-3} + \frac{3^1}{1!} e^{-3} + \frac{3^2}{2!} e^{-3} + \frac{3^3}{3!} e^{-3} + \frac{3^4}{4!} e^{-3} + \frac{3^5}{4!} e^{-3} \right)$$

$$= 0.0651\cdots \fallingdotseq 0.07$$

航空事故の発生件数が 6 件以上起きる確率は約0.07

3　二項分布を利用した問題解決 [データに基づく仮説の検討方法－背理法の拡張]

◇ コイントスが、公平であるか否かを検討してみよう

先生：コインを10回トスして、表が１回も出なかったとしましょう。皆さんはこのコインは公平と思いますか？
航平：思いません
先生：それはどうしてですか？
一同：？？？

このコイントスが公平であるか否かについて、データに基づいて検討する方法が、**仮説検定**（hypotheses testing）である。仮説検定は、数学の背理法の考え方に類似している。

背理法では、ある命題を否定したとき、数学的矛盾が生じることを示して、この命題が真であることを証明する。これに対して、仮説検定では、前提とした仮説の下で求められた結果が、現実のデータと整合しないことを確率的に示す手順を採る。この棄却したい仮説のことを**帰無仮説**（null hypothesis）と呼ぶ。仮説検定の手順を示せば次のようになる。

[手順１：帰無仮説の設定]
仮説検定では、まず背理法と同じように、「このコインは公平である」と仮定する。

[手順２：観測事象が帰無仮説の仮定の下で生じる確率の評価]
帰無仮説の下で、現実に観測した事象（event）が、どれくらいの確率で起きるかを評価する。実際、上のような事象が、公平なコインを用いて生じる確率は、二項分布を用いて計算すれば、${}_{10}C_0/2^{10}=1/1024$となる。つまり、1024回に平均１回程度しか起きない事象であると評価される。

[手順３：確率が小さい場合に帰無仮説を棄却]
仮説検定の論理では、帰無仮説の下で稀な事象が起きたと考えるのではなく、データは帰無仮説を否定する**証拠（evidence）**を示したと考え、前提とした帰無仮説が誤っていたと考える。そして、このような証拠がデータから示されたとき、帰無仮説を**棄却**（reject）して、「コインは、表が出る確率が１/２より小さくなるような偏りをもつ」と結論づける。

航平：でも本当に稀なことが起きているかもしれません
先生：確かにそうです

> Q１：仮説検定と背理法を比較して、仮説検定の論理の持つ危険性を議論しなさい。

PPDAC サイクルに沿って、背理法と仮説検定の手順を整理すると、表１のようになる。

表 1　問題解決から見た背理法と仮説検定

手順	背理法	仮説検定
Problem 立証すべき仮説の起案	証明したい命題の明確化	検証したい命題の明確化
Plan 否定すべき作業仮説	作業仮説（もとの命題を否定した命題）の設定	帰無仮説の設定
Data＋Analysis 推論	数学的知識を用いて論理矛盾を導出	データに基づいて確率的矛盾を導出
Conclusion 結論	作業仮説の否定による証明したい命題の採択（結論に誤りはない）	帰無仮説の否定による検証したい命題の採択（結論は一定の確率で誤りが生じる）

表が出やすいのか裏が出やすいのかが事前に分からない場合

コイントスの例で、10回のうち表が10回出たとすれば、これも公平なコインを仮定すれば、そのようなことが起きる確率は 1 /1024となる。その場合には、やはり帰無仮説を棄却して、表が出る確率は 1 / 2 より大きかったと考えることになる。しかし、私たちは、表が出る確率が 1 / 2 より大きくなるか小さくなるか、普通は実験前に予想できていないことが多い。したがって、表の出る確率が大きい方にも小さい方にも偏ることを問題にして、表が 1 回も出ない場合にも、あえて「表が 1 回も出ない」または「裏が 1 回も出ない」確率を帰無仮説の下で評価するのが、より安全な態度といえる。その確率は、(1 / 1024) × 2 = 1 / 512ということになる。これも十分稀な事象と考えられるであろう。

◇ 帰無仮説の下で評価すべき事象とは？…より極端な事象が生じる確率の評価

航平：表が 1 回も出ないときにその確率を考えることは分かりましたが、10回中 1 回出た場合はどう考えるのでしょうか？

先生：そのときは、表が出るのが 1 回である事象ではなく、表が出るのが 1 回以下である事象の確率を評価します。帰無仮説の下で、観測された事象よりも偏っているすべての事象の確率の総和を評価するのです。

二項分布を用いれば、この確率は $({}_{10}C_0 + {}_{10}C_1)/2^{10} = 11/1024$ と評価される。これも常識的にはかなり小さな確率なので、表が出る確率が 1 / 2 より小さくなっているという証拠をデータが示していると考えることができる。上で述べたことと同様、帰無仮説からのずれに関して、確率が 1 / 2 より大きくなる場合と小さくなる場合とが、事前には予想がつかない場合には、両方の偏りの可能性が事前には考えられていたのだとして、この確率を 2 倍にして、11/512としておくのがより公平な態度と考えられる。

このように、帰無仮説の下で、どの程度稀なことが観測されたかを確率で表した数値を**有意確率**（significance Probability）、ないしは**P 値**（P-value）と呼ぶ。

航平：どのくらい小さな確率ならば、めったに起きない事象と考えたら良いのですか？

先生：常識的に考えて十分小さな値となったとき、帰無仮説を否定すれば良いのです

航平：その常識がないから聞いているのです

理恵：５％とか１％といった確率は小さいのではないかと思うけれども

公平なコイントスに関する仮説検定を応用すると、PPDAC サイクルの Conclusion の確実性を上げることができる。つまり、単なる観察や記述に基づく発見ではなく、観察した事実を基に「仮説が検証された」という Conclusion（結論）を導くことができる。もちろん、このためには、PPDAC の Problem、あるいは Plan の段階、つまり Data を収集する前までに、自身の問題に対する仮説を明確に提示する必要がある。

ここでは、**第２部４の「散布図・相関分析による問題解決」**の事例を用いて、仮説の検証手順を示す。すでに、都市の緯度と年間平均気温に関係性があることを発見している。PPDAC サイクルの A、C のステップにおいて、この関連性が偶然とは言えないことを仮説検定で検証する。

STEP 1：Problem 問題　課題の設定

「都市の平均気温は緯度の高さと関係している」ことを検証したい。

STEP 2：Plan 計画　どのような方法で分析するか

「高緯度地域は年間平均気温が低い」との仮説を検証するために、帰無仮説として「都市の緯度が高いことと平均気温には全く関係がない」を設定し、仮説検定を行う。

STEP 3：Data 収集　仮説検定のためのデータの収集と整理

第２部の「４　散布図・相関分析による問題解決」の表１（38ページ）に掲げた25都市の平均気温、緯度、標高のデータを活用する。

Q２：38ページの表１のデータについて、緯度（北緯ベース）は絶対値に変換して、３つのデータのそれぞれについて、25都市の中央値を求めなさい。そして、データの値が中央値より大きければ＋、中央値ならば０、中央値未満ならば―、という符号に変換したデータを構成し、次の表１を完成しなさい。

表１　アジスアベバの平均気温、ケープタウンの緯度の絶対値、ケープタウンの標高の中央値を閾値として±に符号化したデータ

地名	平均気温	緯度	標高	地名	平均気温	緯度	標高
昭和基地				ドーハ			
メルボルン				カイロ			
プエノスアイレス				ケープタウン		0	0
ブリスベン				東京			
リオデジャネイロ				サンフランシスコ			
リマ				北京			
ジャカルタ				サラエボ			
シンガポール				リオン			
ボゴダ				チュリッヒ			
コロンボ				プラハ			
アジスアベベ	0			ダブリン			
チェンマイ				レイキャビク			
メキシコ							

STEP 4：Analysis 分析　帰無仮説の下での確率の評価

Q 3：平均気温と緯度について、それぞれの中央値より大きいか、小さいかに関する符号条件が一致するものと一致しないものは、何都市ずつになっているか？　ただし、どちらかの符号が 0 となっているものは数えない。

帰無仮説の下では、平均気温と緯度には関係性がないので、緯度の絶対値が中央値より高ければ平均気温が中央値より高くなる事象（符号が共に＋となること）、あるいは、いずれも中央値より低くなる事象（符号が共に－となること）は、平均気温と緯度の符号が一致しない事象と同じ確率で起きるはずである。

もし、緯度が高ければ平均気温も高くなるのならば、符号が一致することが多くなるはずであるし、緯度が高ければ平均気温が低くなるのならば、符号は一致しにくくなるはずである。そして、私たちはこの後者の関係性を実証したいと考えている。

Q 4：緯度が高ければ平均気温が低くなる事象の P 値を求めなさい。

STEP 5：Conclusion 結論　統計的に検証された結論

仮説検定の論理に基づいて高緯度ほど年間平均気温が低いと主張できる。

このように、中央値より大きいか小さいかといった検定の考え方を用いれば、事象間の関連性について簡便に検証できる。

もう 1 つオマケで仮説検定の方法に習熟しよう！

STEP 1：Problem 問題　課題の設定

◇ ボルトが出場した組には強い選手がそろっているのではないか？

2016年リオデジャネイロ・オリンピックの100m 準決勝第 2 組には、ウサイン・ボルトが出場した。ボルトの出た第 2 組には、第 1 組や第 3 組に比べて強い選手が多いのではないかという問題意識を持った。

STEP 2：Plan 計画　どのような方法で分析するか

帰無仮説として「第 2 組とそれ以外の組の記録の分布は同じである」を設定し、準決勝の実際の記録を調べて、分布を比較する。

STEP 3：Data 収集　仮説検定のためのデータの収集と整理

◇ 2016年オリンピック100m 走の準決勝の走破記録を集めよう！

準決勝で記録を残した22名の中央値は10.05秒（２名同タイム）である。検定を行うために、それより速い記録の選手10名を太字（符号＋）に、それより遅い記録の選手10名を斜体（符号－）で示し、選手名の頭に第２組は＋、第１組と第３組は－の符号を付けて加工したデータが表２である。

表２　2016年リオデジャネイロ五輪男子100m 走準決勝第２組と第１組、第３組の走破記録

第２組選手	記録（秒）	第１組選手	記録（秒）	第３組選手	記録（秒）
＋ボルト	**9.86＋**	**－ピコ**	**9.95＋**	**－ガトリン**	**9.94＋**
＋デクラッセ	**9.92＋**	**－メイテ**	**9.97＋**	**－ブレーク**	**10.01＋**
＋ブロメル	**10.01＋**	**－シンビネ**	**9.98＋**	*－リメートル*	*10.07－*
＋ Ujah	**10.01＋**	**－ハーベイ**	**10.03＋**	*－蘇*	*10.08－*
＋山県	10.05	－アシュミード	10.05	－ブラウン	*10.13－*
＋コリンズ	*10.12－*	*－ブレーシー*	*10.08－*	*－ダサオル*	*10.16－*
＋グリーン	*10.13－*	*－謝*	*10.11－*	*－飛鳥*	*10.17－*
フィッシャー	失格	－タフィティアン	*10.23－*	ベイリー	棄権

表２で中央値より小さい数値を持つ人数（太字の数）は、第２組が４名、第１組と第３組を合わせて６名である。また、中央値より大きい、あるいは小さい数値を持つ人数は、第２組が６名、第１組と第３組を合わせて14名である。

STEP 4：Analysis 分析　帰無仮説の下での確率の評価

帰無仮説の下では、組の符号と走破記録の符号の両者が存在する20名（第２組６名、第１組と第３組を合わせて14名）の中で、符号が一致する確率は0.5となる。なぜならば、第２組とそれ以外の組での記録の分布が同じならば、記録の符号は、それぞれの組の中で五分五分の確率で生じるからである。

一方、第２組が他の組よりも成績が上位になる傾向が高ければ、組の符号と記録の符号が一致する確率は、0.5より大きくなるはずである。実際に、符号が一致しているのは12名である。したがって、P 値は、次のように計算される。

$$P\text{ 値} = ({}_{20}C_{12} + {}_{20}C_{13} + {}_{20}C_{14} + {}_{20}C_{15} + {}_{20}C_{16} + {}_{20}C_{17} + {}_{20}C_{18} + {}_{20}C_{19} + {}_{20}C_{20}) / 2^{20} = 0.251$$

STEP 5：Conclusion 結論　統計的に検証された結論

仮説は棄却できない。

P 値に基づけば、この程度の記録の偏りは、帰無仮説の下でも４回に１回程度起きうることで、稀な事象が生じたとは言い難い。したがって、「第２組とそれ以外の組の記録の分布は同じである」との帰無仮説を棄却して、「ボルトが出場した組とそれ以外の組とで記録の分布が異なっている」は、二項分布を用いた検定では主張することはできない。

航平：帰無仮説が棄却できないということは、帰無仮説が正しいということでしょうか？
先生：それは違います。背理法で作業仮説の矛盾を示す証明に失敗したときは、どう考えますか？
理恵：証明ができないというだけです。作業仮説が正しいわけではありません。
先生：そうです。帰無仮説が棄却できないような状況は、帰無仮説を否定する十分な根拠が得られなかったと考えるのが正しい態度です。

Q５：二項検定を応用できそうな状況をいろいろと議論してみよう。

〔本節の解答〕

Q１：帰無仮説が正しいのが真実であった場合、仮説検定の論理に従って決断すれば、小さな確率で誤りを起こす。

Q２：

地名	平均気温	緯度	標高	地名	平均気温	緯度	標高
昭和基地	−	+	−	ドーハ	+	−	−
メルボルン	−	+	+	カイロ	+	−	+
ブエノスアイレス	+	+	−	ケープタウン	+	0	0
ブリスベン	+	−	−	東京	−	+	−
リオデジャネイロ	+	−	−	サンフランシスコ	−	+	−
リマ	+	−	−	北京	−	+	+
ジャカルタ	+	−	−	サラエボ	−	+	+
シンガポール	+	−	−	リオン	−	+	+
ボゴダ	−	−	+	チュリッヒ	−	+	+
コロンボ	+	−	−	プラハ	−	+	+
アジスアベベ	0	−	+	ダブリン	−	+	+
チェンマイ	+	−	−	レイキャビク	−	+	+
メキシコ	+	−	+				

Q３：符号一致：２都市（ブエノスアイレス、ボコダ）、符号不一致：21（それ以外の都市）となる。

Q４：$({}_{23}C_0 + {}_{23}C_1 + {}_{23}C_2)/2^{23} = 0.0000330$

Q５：２つの時系列データの間に関連性があるか否かの検討

第5部 統計的探究の実践 Ⅳ
～標本データから全体を推測する～

1　どの味のラーメンが好まれるだろうか？［標本誤差の評価］

ラーメンは、中国の麺料理が日本に伝わったものであるが、日本各地で独特の進化を遂げ、今や日本の国民食の地位を得ているといっても過言ではない。それにとどまらず、中国への逆輸出を始め、海外にも盛んに進出していて、世界のあらゆる都市で日本のラーメン屋が店を構えるほどになっている。

ラーメンの味には大きく分けて、みそ、しょうゆ、塩、とんこつの４種類があるが、はたしてどの味のラーメンが最も好まれているのであろうか。ラーメンの味に関する調査は数多く行われていて、オリコンスタイル・トレンドリサーチの調査によると、日本全国では４種類の味の中で最も好きなものの割合は表１の結果であった。

表１　ラーメンの味の好みの調査結果

しょうゆ	みそ	塩	とんこつ	合計
28.6%	25.3%	22.0%	24.1%	100.0%

表１は全国の結果であるが、ラーメンの味の好みには地域差があり、北海道・東北地方ではみそ味が好まれ、九州・沖縄地方ではとんこつ味の人気が高いとの調査結果となっている。さらに細かく見れば、ご当地ラーメンがそれぞれの地方にあり、味の好みも千差万別のようである。そこで、あなたのいる地方では、どの味のラーメンの味が好まれるのかに興味がわくのではないか。これは単なる興味ではなく、地域のラーメン屋にとっては死活問題であるし、スーパーマーケットなどでの即席ラーメンの仕入れ量にもかかわってくる問題である。

STEP 1：Problem 問題　課題の設定

◇ **自分の住む地域において、ラーメンの４種類の味（しょうゆ、みそ、塩、とんこつ）のうち、どれが最も好まれるかを知りたい**

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ **４種類の味のうちどれが最も好まれるかについて、その結果が妥当かを検討する**

データの集め方には種々のものが考えられる。たとえば、ラーメン屋に行ってどの味のラーメンの売り上げが最も多いかを調べる、スーパーマーケットに行ってどの味の即席麺が最も売れているかを調べる、などである。しかし、街のラーメン屋がみそラーメン専門店であったり、スーパーの売り上げは企業秘密であったりと、望ましいデータが得られないことも考えられる。そこで、**アンケート調査**を実施することとする。

アンケート調査は比較的手軽に実施できることから、あまり計画性もなく行われることも多いが、正しい結果を得るためには相応の計画と準備が必要である。まず、母集団をある程度特定する必要がある。母集団をあらかじめ特定し、そこからデータを得るのが原則であるが、逆に、実際に得ることができるデータは何かという考察から母集団が決まることもあるだろう。母集団からの無作為（ランダム）な標本抽出を原則とするならば、調査可能でない部分集団は母集団に含め得ないことになる。

データは母集団から無作為に抽出するのが理想的であるが、実際問題ではなかなか無作為抽出は難しい。したがって、性別や年齢層など、結果に影響を及ぼすと想定される要因については、なるべく偏らないようにすること！
また、回答拒否などによるデータの欠損（欠測）への対処法も考えるのが大事だね。実際のデータ収集では、計画通りにデータが得られることは稀であり、それへの対応もできる範囲で事前に想定しておくことが必要だ。

Q 1 ：母集団からの偏りのない調査法にはどのようなものがあるだろうか。また、はたしてそれは実施可能か。もし、実施が難しかった場合、代わりにどのような調査法があるだろうか。

Q 2 ：インターネット調査では偏りが起きる可能性が否定できない。起きうる偏りにはどのようなものがあるだろうか。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ アンケート調査でデータを収集しよう！

ここでは、高校の生徒たちが自分自身を含め、彼らの親兄弟や身近な人からデータを得ることを想定する。高校のあるクラスの生徒のラーメンの味の好みが地域の他の集団と著しく異なるのであれば、この調査結果を地域の住民全体に一般化することはできないが、そのようなことはないと仮定する。

何人分のデータを集めたら良いかは重要な問題である。データ数が少ないと確かな結論は得られないが、多くのデータを集めるには時間も費用もかかる。そこで、参考となるのが標準偏差の値である。サンプルサイズを n として母集団比率 p を標本比率 $\hat{p}$ で推定する場合、$\hat{p}$ の標準偏差は $\sigma(\hat{p})=\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ である。この値は p が分からないと計算できないが、その最大値は $p=0.5$のときであり、$\sigma(\hat{p})=\frac{1}{2\sqrt{n}}$ である。仮に、標準偏差の目標値が５％であったとすると、$\frac{1}{2\sqrt{n}}\leqq 0.05$ より、$n\geqq\frac{1}{4\times 0.05^2}=100$ が得られる。そこで、サンプルサイズを100以上にすれば、すべての p に対して標準偏差を５％以下にできることになる。

Q３：母集団比率が0.3程度とみなされるとき、標準偏差を２％とするためには、サンプルサイズ n はいくつ以上必要か。

世論調査の有効性

2016年には、当初の調査結果を覆す「事件」が２つ起きました。１つは英国の国民投票によるEU離脱の決定であり、もう１つは米国大統領選挙でのトランプ候補の勝利です。とくに、トランプ候補の勝利は、それまでの世論調査の結果が概ねクリントン候補の勝利と出ていただけに、大いなる驚きを与えたと同時に、世論調査への疑問も投げかけられる結果となりました。

大統領選挙当日まで多くの調査が行われましたが、選挙直近ではどの調査においても、両候補の支持率は極めて接近したものでした。したがって、世論調査で導くべき結論は「支持率は接近していて、どちらが勝ってもおかしくない」であったはずです。その証拠に、全米での支持率はクリントン候補が上回っていましたが、トランプ候補が大統領に選出されたのは米国の大統領選挙の仕組みによるところが大きかったといえます。

日本でのマスコミ論調では、トランプ候補を揶揄したような報道が多く、それに接した日本人の導いた結論では、圧倒的にクリントン候補に支持が集まっていました。日本でのデータに基づかない判断よりも、米国でのデータに基づく調査結果のほうが正しく結果を言い当てていたのです。

いろいろな調査結果を比べてみると、好みの味の割合が調査ごとに異なっていることが分かる。これは、どの調査結果が正しくて他の調査結果が間違いということではない。日本のすべての人のラーメンの味の好みなどというのは、もちろん調べることはできない。調査は、日本全体の人々の中から選んだごく一部の人に対して行われるもので、調査ごとに対象者が違うため調査結果がそれぞれ異なるのは当然であるといえよう。

統計用語では、調査対象となる全体の集団を母集団といい、実際に得られたデータを標本（サンプル）という。そして母集団から標本を得ることを**標本抽出**（**サンプリング**）という。調査では、何人の人から得たデータであるか（これを**標本の大きさ**、あるいは、**サンプルサイズ**という）が重要な鍵である。また、得られた標本が母集団の特徴を偏りなく表しているかどうかも重要な視点となる。

標本抽出で、標本に含まれる人が調査ごとに異なることによる不可避的な調査結果のバラツキを**標本誤差**という。調査の計画が妥当なものであれば、標本誤差はサンプルサイズが大きいほど小さくなり、調査結果は信頼に足るものとなる。しかし、標本抽出が妥当なものでなく、標本が母集団の特徴をうまく表していないとすると（これを**非標本誤差**という）、サンプルサイズをいくら大きくしても、調査結果は信用のおけないものとなってしまう。

STEP 4：Analysis 分析　統計量で傾向を捉える

◇ アンケート結果を母集団における値の推定値とみなし、その標準偏差を計算する

データが得られた場合には、まず、データのクリーニング、すなわち年齢が180歳などというあり得ないデータを拾い出し、訂正すべきものは、根拠をもって訂正しなくてはならない（根拠のない訂正はデータの改ざんにつながる恐れがある）。その後、集計表の作成やグラフ化により結果を分かりやすく表示する。そしてその後、必要に応じて標準偏差の計算など推測統計的な手法を適用する。

たとえば、100人の生徒への調査で、表２のようなデータが得られたとする。みそ味の人気が高いように見えるが、そのように結論付けて良いだろうか。

表２ ラーメンの味の好みの割合と標準偏差

好みの味	しょうゆ	みそ	塩	とんこつ	合計
割合	22%	42%	19%	17%	100%
標準偏差	4.52%	4.35%	4.14%	4.28%	

各味の好みの割合の標準偏差を計算すると表２の下段のようになる。これによると、各好みの割合の平均は４％から５％程度の標本誤差を含んでいることが分かる。しかし、みそと２番人気のしょうゆとの差は20％もあることから、この調査から、みその首位は動かないものとみても差し支えないと考えられる。

コンピュータ・シミュレーション

標本誤差の大きさを実感するため、コンピュータを使って実験してみましょう。表１の割合を母集団における値として、ここからランダムに50人および200人を抽出して、それぞれの味を好む人がどれくらいの割合となるかについて、コンピュータの乱数を用いてシミュレーションします。ここでは、標本抽出のシミュレーションを100回繰り返します。

表３は、シミュレーションの最初の５回分の結果です。各値は、それぞれの味を好んだ人の割合と、表１の設定で最も人気の高かったしょうゆを好んだ人の割合の４つの味の中での順位です。サンプルサイズｎ＝50の５回のシミュレーションでは、しょうゆが最も好まれたのは１回だけですが、ｎ＝200では３回となっています。ｎ＝50の第５回目のシミュレーション結果では、しょうゆの割合が最も小さくなっています。それに対して、ｎ＝200では、しょうゆの順位は１位もしくは２位のみです。ここからも、標本誤差はそれなりにあること、およびサンプルサイズが大きいほど結果は安定することが分かります。

表３　シミュレーションの最初の５回分

ｎ＝50	しょうゆ	みそ	塩	とんこつ	順位
1	0.26	0.28	0.22	0.24	2
2	0.22	0.28	0.20	0.30	3
3	0.28	0.36	0.22	0.14	2
4	0.28	0.20	0.24	0.28	1
5	0.22	0.30	0.24	0.24	4

ｎ＝200	しょうゆ	みそ	塩	とんこつ	順位
1	0.320	0.275	0.185	0.220	1
2	0.310	0.240	0.210	0.240	1
3	0.260	0.285	0.215	0.240	2
4	0.295	0.280	0.220	0.205	1
5	0.255	0.235	0.230	0.280	2

表４に100回のシミュレーションにおける標本ごとの割合の平均と標準偏差を示します。表４からは、シミュレーション100回の平均は、ｎ＝50とｎ＝200のいずれについても、表１の値に近いことが分かります。それに対して、値のバラツキを表す標準偏差は、ｎ＝200のときの値がｎ＝50での値の約半分になっていることが見て取れます。また、図１は、100回のシミュレーションにおけるしょうゆの割合のヒストグラム（ｎ＝50, 200）としょうゆの順位の棒グラフ（ｎ＝50, 200）です。図１の（a）と（c）より、ｎ＝200ではｎ＝50のときに比べて、しょうゆの割合の値のバラツキが小さいことが見て取れ、図１の（b）と（d）からは、しょうゆの順位が正しく１位となった回数は、ｎ＝200ではｎ＝50のときに比べて大きくなり、４位となってしまうことはなかったことが分かります。

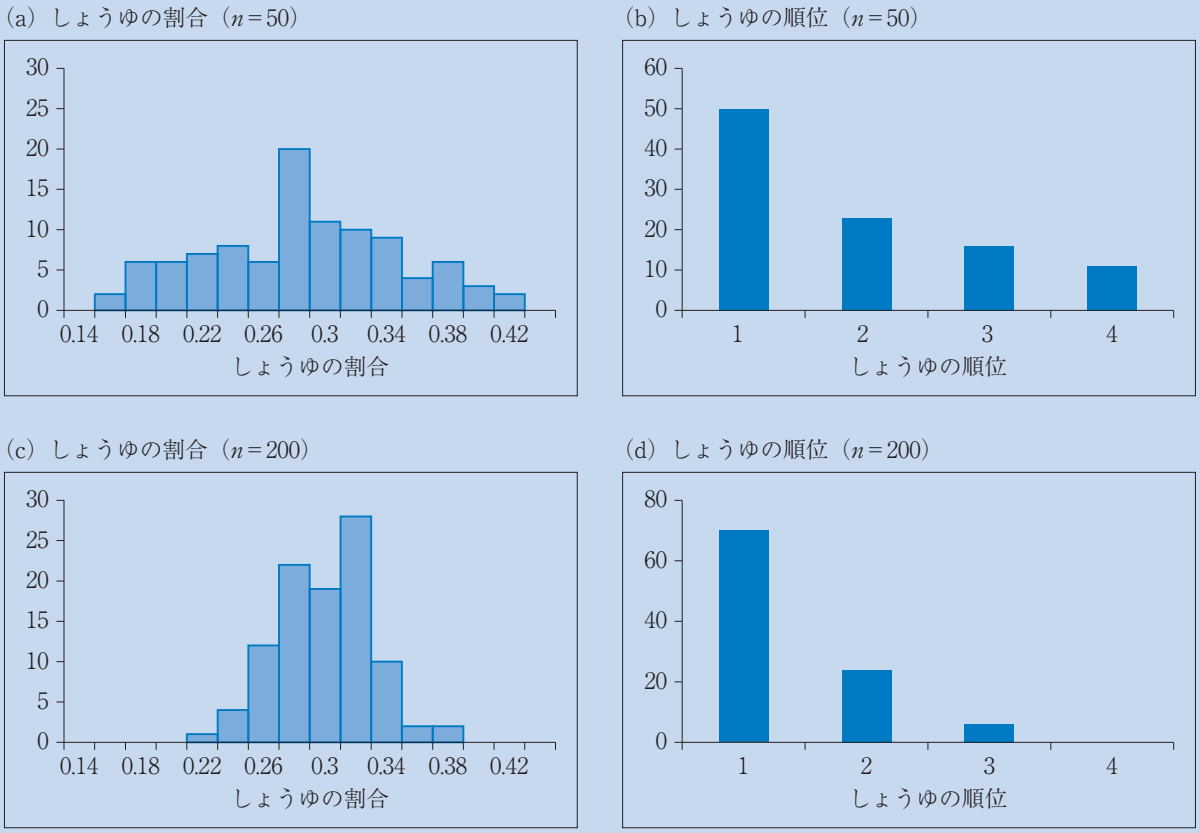
このように、標本抽出が妥当であれば（このシミュレーションでの標本抽出は完全無作為抽出といって、最も妥当な標本抽出法であるとされる）、サンプルサイズが大きければ大きいほど標本誤差は小さく、母集団での特徴を正しく表すことができることになります。

表４　100回のシミュレーションにおける標本別割合の平均と標準偏差

ｎ＝50	しょうゆ	みそ	塩	とんこつ	順位
平均	0.285	0.253	0.219	0.242	1.880
標準偏差	0.062	0.054	0.054	0.064	1.047

ｎ＝200	しょうゆ	みそ	塩	とんこつ	順位
平均	0.292	0.250	0.218	0.241	1.360
標準偏差	0.030	0.026	0.029	0.029	0.595

図 1　100回のシミュレーションの結果のグラフ表示



コンピュータを用いて、母集団からの標本抽出をシミュレーションしてみよう。乱数発生は Excel の分析ツールの「乱数発生」で容易にできるよ。

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ 調査結果が妥当なものかどうかを吟味しよう！

データの収集と統計分析の結果、表 2 のデータからは、表 1 の全国における値に比べて、みそ味の人気が高いことが分かった。プラスマイナス 4 %程度の誤差はあるが、この地域でみそ味が好まれる割合はおおよそ40%である。

この結果を次の段階につなげることが重要だ。より一層みそ味のラーメンの人気を高めるべきなのか、あるいは、他の味のラーメンをさらに拡充していくべきなのか。いずれにせよ、意思決定には客観的なデータと統計分析の利活用が必要となる。

2　フライドポテトの重量は公表値と同じ？［区間推定］

ハンバーガー・チェーンは、M 社が1971年に東京・銀座に 1 号店を出店して以来日本全国に展開され、今やファーストフードの王様的存在になっている。ハンバーガー店での人気メニューの 1 つがフライドポテト（フレンチフライ）であり、通常、L、M、S の 3 サイズが用意されている。それぞれのサイズの重量は概ね決められていて、あるチェーン店では M サイズのフライドポテトは135g と公表されている。実際にハンバーガー店に行って注文すれば分かるが、フライドポテトの重量を逐一測ってお客さんに提供していたのではサービスに時間がかかるため、店舗スタッフが目分量で判断して提供していることが多い。そこで、本当にフライドポテトの重量が公式発表の135g となっているかどうかの疑問がわく。

STEP 1：Problem 問題　課題の設定

◇ あるハンバーガーショップのフライドポテト M サイズの重量は、公表値の135g なのだろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ このハンバーガーショップから、M サイズのフライドポテトを実際に10個購入した上で重さを測定し、得られたデータに基づいて、公表値が適正かどうかについて区間推定を使って検証する

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 購入日と時間帯を変えて、M サイズのフライドポテト10個の重量データを得よう

10個のフライドポテトの重量の測定結果は表 1 のとおりであった。

表 1　10個のフライドポテトの重量

重量（g）	118	122	125	126	128	130	132	135	136	138

STEP 4：Analysis 分析　統計量で傾向を捉える

◇ データのバラツキについて評価しよう

購入したフライドポテトの重量の平均129.0g は公式発表の値よりも小さかったとはいえ、たった10個のデータであり、実際、表 1 からは公表値の135g 以上のフライドポテトも10個中 3 個あるのも事実である。データのバラツキが考慮されていないので、標本誤差も加味した上での分析が必要であろう。統計学の手法を用いて精緻に検証しよう！

前節では、調査には標本誤差が不可避的に生じることを学んだ。前節では比率のデータを用いたが、ここでは、フライドポテトの重量のような連続型のデータのバラツキとその評価法を考えよう。

データが得られたら、その特徴を示す基本統計量を計算するのが統計分析の第一歩である。表 1 のデータに Excel の「分析ツール」の「基本統計量」を適用すると表 2 が出力される。本節ではそれらの基本統計量のうち標準誤差、標準偏差、**信頼度**（95.0%）に着目する。

表 2　Excel の「基本統計量」

重量	
平均	129
標準誤差	2.03
中央値（メジアン）	129
最頻値（モード）	#N/A
標準偏差	6.43
分散	41.34
尖度	−0.80
歪度	−0.24
範囲	20
最小	118
最大	138
合計	1290
データの個数	10
信頼度（95.0%）	4.59

「統計情報」と「平均の信頼度の出力」にチェックを入れると求められる。数値は小数第 3 位を四捨五入している。
Excel において量的データの場合、最頻値は意味がないため、表 2 の最頻値の欄には #N/A の利用不可のマークが表示されている。

n 個の測定値を $x_1, \cdots, x_n$ として、それらの標本平均を $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ とする。そして、**標本分散**および**標本標準偏差**（standard deviation）をそれぞれ

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2,\ s=\sqrt{s^2}$$

とする。また、標準誤差（standard error）は $\frac{s}{\sqrt{n}}$ で求められる。

高等学校の教科書では標本分散の定義式において除数は n であるが、ここでは、大学以降での定義式および Excel での計算式に合わせて n−1としている。n が大きい場合には数値的にあまり差はない。

Q１：表１のデータを用いてExcelで表２を出力しなさい。また、標準誤差の値2.03は、標準偏差の6.43を$\sqrt{10}$で除して得られることを確かめなさい。

Q２：ハンバーガーショップはフライドポテトの公表値135gをどのように捉えるべきであろうか。平均値が135gであれば良いのか。最小値が135g、すなわち135gを下回る商品があってはならないのか。135gを下回る商品が５％程度あるのは許容するのか。それぞれの基準を満たすためには、どのように商品の提供をするのが良いか。

標本標準偏差は測定データのバラツキを表す指標であり、母集団全体での値のバラツキを反映した値である。すなわち、標本標準偏差は母集団全体でのバラツキを示す標準偏差（母標準偏差σ）の推定値であるともみなされる。したがって、サンプルサイズを大きくすると標本標準偏差はσに近づく。それに対して、標準誤差は、その定義式の分母にサンプルサイズnを含むことから、nを大きくすると０に近づく。

標準誤差の意味の理解には、前節でも議論した母集団からの無作為抽出が重要な役割を果たす。母集団からの無作為な標本抽出によってn個の測定値を得て、それらから標本平均$\bar{x}$を計算する、という作業を考えよう。この「n個の測定値から標本平均$\bar{x}$を計算する」という作業は、実際は１度きりしか行われていなくても、概念的には何回でも繰り返すことができる。そして、その度ごとに$\bar{x}$が得られるが、標本抽出は無作為であるので、標本誤差により$\bar{x}$はそれぞれ異なる値を取り、何らかの分布に従って分布することになる。これを$\bar{x}$の標本分布という。母集団分布の**母平均**がμで、**母分散**がσ^2のとき、$\bar{x}$の期待値は同じくμで、分散は$\frac{\sigma^2}{n}$となることが示される。この分散の平方根$\frac{\sigma}{\sqrt{n}}$を$\bar{x}$の標準誤差と呼ぶのである。母標準偏差σは通常未知であるので、標本標準偏差sをその代わりに用いた$\frac{s}{\sqrt{n}}$を標準誤差と呼ぶことが多い。

前節では、推定値に、その精度の情報を加味して（推定値±推定値の標準偏差）と表示した。ここでは、推定値の標準偏差のことを標準誤差と呼んでいるので、これは推定値±標準誤差と表現することができる。表２の結果については、129.0±2.03となり、区間（126.97, 131.03）を意味している。その区間に確率的な意味を持たせたものが**信頼区間**である。

信頼区間

母集団を特徴付ける定数は**パラメータ**と呼ばれるが、ここではパラメータの推定について扱う。推定は 1 つの値で行われることが多い。たとえば、母平均 μ の推定値として標本平均 $\bar{x}$ といった 1 つの値が用いられる。このように、1 つの値でパラメータを推定することを、1 点で推定するという意味で点推定という。しかし、推定値の提示だけでは、それが精度の情報を持たないため適切であるとはいえない。そこで、標準誤差の値を加味した（126.97, 131.03）のような表示が用いられるが、そこで示される区間は確率的な意味が不明確である。そこで、確率的な意味を持つ区間を導出し、その区間を用いて母集団パラメータを推定しようというのが区間推定である。そして、そのときの区間を信頼区間という。n 個の測定値が正規分布に従う場合、α を**有意水準**として、$Z(\alpha/2)$ を標準正規分布の上側 $100(\alpha/2)$ %点としたとき、$P(Z \leqq -Z(\alpha/2)) + P(Z \geqq Z(\alpha/2)) = \alpha$ である。ここで有意水準とは、ある事象が起きる確率が偶然であると判断する確率をいい、0.05または0.01が用いられることが多い。母集団パラメータのうち、母平均 μ の信頼度（**信頼係数**）$100(1-\alpha)$ %の信頼区間は、

$\left(\bar{x} - Z(\alpha/2)\frac{\sigma}{\sqrt{n}},\ \bar{x} + Z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right)$ で与えられる。

このハンバーガーショップで販売されているフライドポテトの母平均の信頼度95%の信頼区間を求める。母標準偏差σの値は与えられていないが、データから得られた値6.43が母標準偏差であるとしよう。すなわち、ここでは計算手順を理解する目的で$\sigma = 6.43$としておく。信頼度は95%であるので、α は0.05であり、標準正規分布表から $z(0.025) = 1.96$であるので、**誤差限界**は $1.96 \times \frac{6.43}{\sqrt{10}} \approx 3.99$ となる。したがって、信頼度95%の信頼区間は、129.00 ± 3.99 = (125.01, 132.99) と計算される。

Q 3 ：信頼度を90%あるいは99%として信頼区間を求めなさい。

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ 公表値135g は怪しい！

母標準偏差を既知とした場合、信頼区間は公表値135g をその中に含んでいない。信頼度95%の信頼区間が公表値135g を含まないということは、信頼度95%でこのハンバーガーショップの販売するフライドポテトの母平均が135g ではないと言い切れる。ただし、10個の観測データには大きな標本誤差が存在することを踏まえたうえで、結論を導かなくてはならない。母平均についてさらに精度の高い推定が望まれるのであれば、サンプルサイズをもっと増やした調査が必要である。

Q 4 ：信頼区間幅を狭めるための手立てとして、どのようなものがあるのかについて議論しなさい。

3　フライドポテトの重量は公表値通りか？［統計的検定］

◇ 勘と経験よりも客観的なデータ分析！

勘や経験だけでなく、客観的なデータの分析や数学的な計算に基づく行動や決定は、多くの場合、良い結果を生むことになろう。良い結果を生まないまでも、データ分析や数学計算の限界を知ることができる。そのためには、基本的な事項を学習し、実際の現場に応用する力を養う必要がある。努力は必ず報われる。報われない努力はないと思いたい。

STEP 1：Problem 問題　課題の設定

◇「ハンバーガー店のフライドポテトの重量が公表値通りかどうか」を検証する

前節で、あるハンバーガー店のフライドポテトのMサイズの重量が公表値135gかどうかについて、信頼区間を用いて検証した。ここでは、それについて統計的検定という方法論を用いて評価する方法を実践する。

STEP 2：Plan 計画　どのような方法で分析するか

◇ 統計的仮説検定を行う

「駅前のハンバーガー店のMサイズのフライドポテトの重量が公表されている通りかどうか疑わしい」と考え、これを検証するために、帰無仮説として「フライドポテトの重量が公表値135gとおりである」と設定し、購入データに基づいて仮説検定を行う。

帰無仮説の棄却とは

統計的検定では、P値という値が有意水準（有意確率）よりも小さいとき、帰無仮説が正しくてありそうもないことが起きたとは考えないで、帰無仮説を棄却する。すなわち、帰無仮説は誤りであると判断する。しかし、P値が有意水準よりも大きいときでも、帰無仮説が正しいことが証明されたわけではない。そのことを表現するため、帰無仮説を棄却しないといい、帰無仮説を採択するとは言わない（「採択」に正しいと認めるという意味が含まれているため）。

たとえば、ある高校には男子生徒と女子生徒とが半分ずついるという帰無仮説を考えよう。すなわち、男子生徒の割合をpとしたとき、帰無仮説は$H_0 : p = 0.5$である。いま、この高校からランダムに4人の生徒を抽出して性別を調べたところ、全員が女子であったとしよう。そうなる確率は$(0.5)^4 = 0.0625$であり、有意水準を$\alpha = 0.05$とすると、確率はそれよりも大きいため、帰無仮説は棄却されない。このとき、帰無仮説が正しく、この高校には男子生徒が半数いるといって良いだろうか。実はこの高校は女子高であって、性別を何人調べても全員が女子である可能性はあるのである。なお、5人調べて全員が女子であれば、帰無仮説の下でそうなる確率は$(0.5)^5 = 0.03125$であるので、H_0は棄却される。

すなわち、帰無仮説が棄却されないときは、単なる情報不足という可能性を含んでいるのである。

STEP 3：Data 収集　仮説検定のためのデータの収集と整理

◇ 実際にフライドポテトを購入してその重さを測る

高校生のグループが手分けして、駅前のハンバーガー店の M サイズのフライドポテトを10個購入し、その重量を計測した。その結果を下記の表 1 に示す。

表 1　購入した10個のフライドポテトの重量

重量（g）	120	124	126	130	130	131	132	133	134	140

このデータについて、Excel の分析ツールの「基本統計量」で計算した出力結果が表 2 である。このデータを用いて、このハンバーガー店のフライドポテトの重量が公表値135g といえるかどうかを統計的検定の枠組みで検討する。まず、統計的検定について簡単に述べておく。

表 2　Excel の「基本統計量」

重量	
平均	130
標準誤差	1.77
中央値（メジアン）	130.5
最頻値（モード）	130
標準偏差	5.60
分散	31.33
尖度	0.53
歪度	−0.14
範囲	20
最小	120
最大	140
合計	1300
データの個数	10
信頼度（95.0%）	4.00

STEP 4：Analysis 分析　帰無仮説の下での確率の評価

◇ 帰無仮説の下で検定のための統計量を求め、その実現する確率を導く

母集団の分布は、母数（パラメータ）によって規定される。いま、駅前店で販売するフライドポテトの重量の分布が正規分布に従っているとすると、分布の型は、平均 μ と分散 σ^2の 2 つのパラメータで定まる。ここでは、店で販売するフライドポテトの重量が135g であるかどうかが問われているので、$\mu = 135$と判断することが適当か否かを統計的に検定することになる。

このとき、帰無仮説 H_0は $H_0 : \mu = 135$と表される。それに対して、帰無仮説が否定されたときに採用する仮説を対立仮説（alternative hypothesis）といい、それを H_1で表す。対立仮説には、**両側仮説**と**片側仮説**の 2 種類がある。両側対立仮説は $H_1 : \mu \neq 135$と表される。検証すべき値135g の両側を考えている。

一方、片側仮説は135g より多いか、少ないかの片側を考えており、$H_1 : \mu > 135$または $H_1 : \mu < 135$のいずれかとなる。両側と片側のいずれの仮説を立てるかは問題による。

ここでの設定については、対立仮説として、消費者側として μ が135g より小さいかどうかに関心があるのであれば、$H_1 : \mu < 135$の片側仮説を選択するが、ハンバーガー店としては、逆向きの $H_0 : \mu > 135$に興味があるかもしれない。そこで、ここでは中立的に両側仮説の $H_1 : \mu \neq 135$を立てることにする。

いま、店で販売するフライドポテトの重量が正規分布 $N(\mu, \sigma^2)$ に従っているとして、高校生のグループが無作為（ランダム）に購入した10個のフライドポテトの重量を（$X_1, X_2, \cdots X_{10}$）としよう。そのとき、それらの標本平均 $\overline{X} = (X_1 + \cdots + X_{10})/10$ は正規分布 $N(\mu, \sigma^2/10)$ に従う。母分散 σ^2 が既知であるとすると、X を標準化した統計量 $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{10}}$ は標準正規分布 $N(0, 1)$ に従う。

分散が既知で $\sigma = 6$ とすると、帰無仮説 $\mu = 135$ の下で、データから求めた Z^* の値は、$Z^* = \dfrac{130 - 135}{6/\sqrt{10}} \approx -2.64$ となる。したがって、両側対立仮説に対する（両側）P 値は、正規分布表から $P = P(|Z| \geq |-2.64|) \approx 0.0084$ と求められる。

検定の有意水準を $\alpha = 0.05$ とすると（有意水準は検定を実行する前にあらかじめ決めておくべきである）、$P < \alpha$ であるので、この場合の検定は有意水準５％で有意であり（有意水準１％でも有意である）、$\mu = 135$ という帰無仮説は棄却される。すなわち、有意水準５％でこのハンバーガー店のフライドポテトの平均値は135g ではないことになる。

片側対立仮説の場合は、検定統計量の値は同じく $z^* = -2.64$ であるが、P 値は $p = P(Z \leq -2.64) \approx 0.0042$ となり、この場合の結論は、有意水準５％でフライドポテトの平均値は135g よりも小さいといえる、となる。

統計的検定では、帰無仮説の真偽とデータに基づく判断では、下の表のように４通りの結果となる。すなわち、２通りの誤り方を区別するのである。

		人間の判断	
		H_0 を棄却する	H_0 を棄却しない
神のみぞ知る	H_0 が真	第１種の過誤	正しい判断
	H_0 が偽	正しい判断	第２種の過誤

誤りを２通り考える考え方は重要で、たとえば、病気の診断のとき、病気であるのにそれを病気でないとして見逃す誤りと、病気でないのに病気であるという誤りとを考える必要がある。どちらかの確率を小さくしようとすると片方が大きくなるという二律背反的な性格をもつ。統計的検定では、第１種の過誤を犯す確率をコントロールするという考え方を取る。実際、第１種の過誤を犯す確率が有意水準となる。

実際の問題では、２種類の過誤のコストを考えての判断が必要となる。病気の診断で、病気であるのにそれを見逃す誤りの確率をなるべく少なくしたいのであれば、ちょっとした兆候を持つ人をことごとく病気であると診断すればよい。そうすると病気を見逃す危険性は減るものの、病気でないのに病気と言われて再検査を受けるまで心配するのも精神衛生上よいものではない。

STEP 5：Conclusion 結論　統計的に検証された結論と新たな課題

◇ 仮説検定の結果、駅前店のフライドポテト平均重量は公表値よりも少ないと判断される

高校生グループからフライドポテトの重量が公表値以下であるとの指摘を受けたハンバーガー店の店長は、それでは平均値をどのくらいに設定したら良いかを、逆に高校生グループに問いかけた。もし、平均値を135g に設定して販売すると、およそ半数の客から自分のポテトの重量が135g より少ないとクレームがきそうである。かといって、平均の重量を135g より必要以上に重めに設定すると店の利益に響いてしまう。売り場に秤を置いて重さを逐一測るのが良いかもしれないが、それでは混雑時に時間がかかってしまいサービスが低下する。重量の多少のバラツキはやむを得ないとして、どうしたら良いか、ということである。

高校生グループは、授業で学習した確率計算の実践的な応用例として次のように考えた。まず、ポテトの重量のバラツキには特に理由があるわけではなく偶然的なものであることから誤差とみなすことにした。誤差であれば、そのバラツキへの正規分布の仮定は妥当性があることから、正規分布 $N(\mu, \sigma^2)$ に基づく確率計算を行うことにした。

X をフライドポテトの重量を表す確率変数とすると、それを標準化した $Z=(X-\mu)/\sigma$ は標準正規分布 $N(0, 1)$ に従う。X が135g 未満であるとクレームがくる可能性があるので、クレーム率として $q=P(X<135)=P(Z<(135-\mu)/\sigma)$ を求めることにした。

これまでの経験から、バラツキの大きさを表す標準偏差は $\sigma=6$ 程度であったとのことであるので、$P(Z<(135-\mu)/6))$ とすると、q はμを定めれば定まる値、すなわち μ の関数であるので、横軸に μ を取り、縦軸に q を取ってグラフに表すことにした。また、$\sigma=3$ としたグラフも同時に描くことにした。図１がそのグラフである。

図１　母平均（μ）とクレーム率

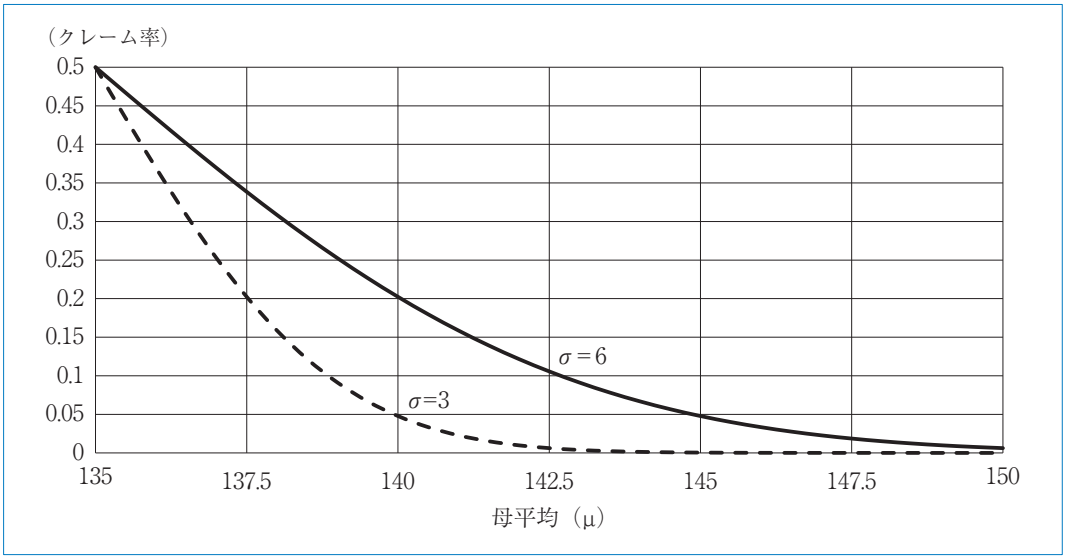


図１から、$\sigma=6$ とした場合、クレーム率を0.05程度にするためには母平均を145g 程度と、公表値よりも10g 多くしなければならないことが分かった。５g プラスの140g を平均値に設定すると、約20％のクレーム率となる。スタッフのサービスのスキルを向上させてバラツキを半分の $\sigma=3$とすると、140g を母平均に設定した場合のクレーム率は５％程度に下がることも、図１から分かる。たかが５g の差に見えるが、１日のフライドポテトの販売量はかなり多いので、かなりの節約になる。バラツキを抑える努力は報われるという一例であろう。

第6部　地域の課題解決と統計活用

1　米産地新潟ブランドをいかに維持するか！

コメ王国 新潟 崩壊か!?　１等米“20%ショック”が県内を襲う

2010年（平成22年）秋、県農林水産部、農協や米生産流通関係者に衝撃が走った…

食味の良さで他県を圧倒するコシヒカリを武器に、「ブランド米産地」の名をほしいままにしてきた新潟県で、平成22年産米（水稲うるち米）の１等米比率が20%となり、前年比70ポイントも急落し、過去最低の結果となった。

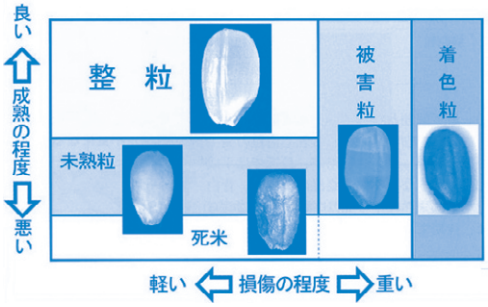
日本一のコメ産地のはずが全国で最低の結果となった…。原因究明と分析を行うための対策会議が急きょ開かれ、有識者による「コメ品質に関する研究会」も設置された。

品質

玄米の外観品質であり、成熟度合と損傷の程度から１等～３等、規格外に分類される。未熟粒や死米、被害粒等を除いた整粒が70%以上で１等級となる。

また、着色粒の割合が0.1%を超えると２等に格付けされる。格付けは農林水産大臣の登録を受けた民間の検査機関が行う品位等検査によって行われる。

玄米品質の構成



資料：新潟県農林水産部「おいしい米づくりのポイント」

STEP 1：Problem 問題　課題の設定

◇「品質」の良い米を作るには

研究会の分析によると「2010年の夏、新潟県は記録的な猛暑に襲われたため、米に栄養分が行き届かず、白濁した粒が増えて品質が低下した」との報告。気象と品質にはどのような関係があるのだろうか。

また、品質が低下する可能性がある気温は何℃以上からなのだろうか。

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

稲の茎から穂の先端が出る出穂～成熟の期間

◇ 農林水産省公表の１等米比率と、特に品質に影響を与えるとされる登熟期間である８月の気温（平均、最高、最低）、降水量、日照時間の気象データとの関係を分析する

【１】　１等米比率と各気象データの関係を**時系列**で調べる。

【２】　１等米比率と各気象データの**相関関係**を調べる。

【３】「記録的な猛暑に襲われたため、…品質が低下した」と研究会の報告にある。品質が低下する可能性がある気温を**回帰分析**により予測する。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 必要なデータをダウンロードし、データクリーニングをしよう

表 1 を参考に Excel を使って分析できる状態にデータをまとめる。

データを分析できる状態に整えること

データ・ダウンロード先

・気象データ…気象庁 HP　http://www.jma.go.jp/jma/index.html
「過去の気象データ・ダウンロード」 地点・項目・期間を選択して（CSV 形式）でダウンロードできる。
・ 1 等米比率…農林水産省 HP　http://www.maff.go.jp/j/tokei/index.html
「政策情報＞研究会等＞水稲の作柄に関する委員会＞第 3 回配付資料＞資料 No. 2-4「作況指数、10a 当たり収量、平年収量および一等米比率の推移」（PDF 形式）」

表 1　1 等米比率と新潟市 8 月の気象データ（1979年～2015年）

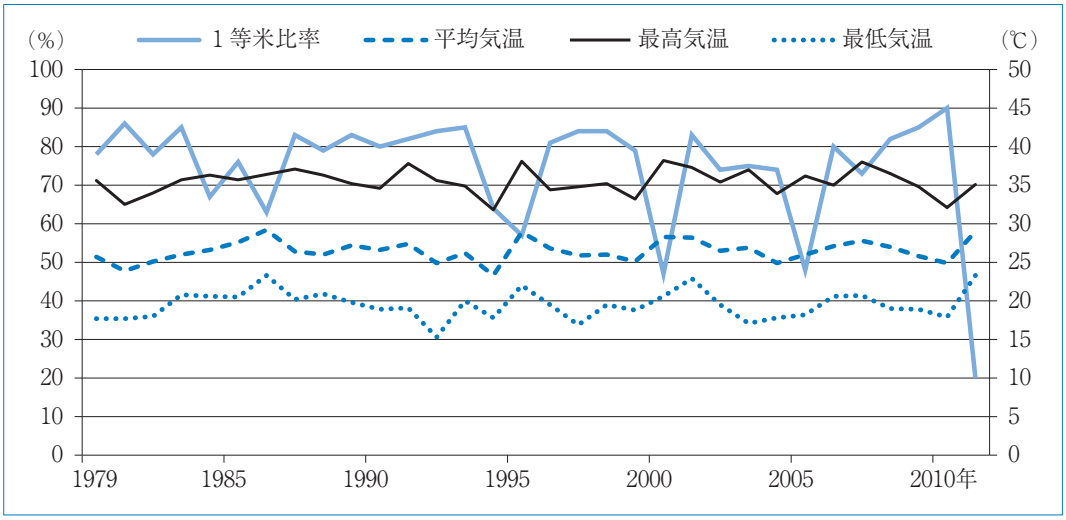
年	平均気温（℃）	最高気温（℃）	最低気温（℃）	降水量（mm）	日照時間（時間）	1 等米比率（%）
1979	25.7	35.6	17.7	130.0	217.3	78
80	23.9	32.5	17.7	250.0	175.8	86
81	25.1	34.0	18.0	213.0	214.7	78
82	26.0	35.7	20.8	38.5	202.5	85
83	26.6	36.3	20.6	89.5	216.2	67
84	27.6	35.7	20.5	17.5	301.7	76
85	29.2	36.4	23.3	3.5	342.6	63
86	26.4	37.1	20.2	67.5	236.4	83
87	26.0	36.3	20.9	166.0	168.2	79
88	27.2	35.2	19.8	40.0	214.7	83
89	26.6	34.6	18.9	116.5	209.4	80
90	27.4	37.8	19.1	85.0	268.0	82
91	24.9	35.6	15.3	156.5	187.1	84
92	26.2	34.9	20.0	88.0	216.4	85
93	23.3	31.8	17.8	181.5	99.0	64
94	28.9	38.1	22.0	65.0	292.6	57
95	26.8	34.4	19.5	303.5	185.2	81
96	25.9	34.8	16.9	84.0	243.5	84
97	26.0	35.2	19.5	129.5	216.7	84
98	25.1	33.2	18.8	616.0	111.2	79
99	28.3	38.2	20.6	167.0	230.4	47
2000	28.2	37.3	22.9	11.5	288.9	83
01	26.5	35.4	19.5	135.5	223.7	74
02	26.9	37.0	17.1	128.0	205.1	75
03	24.9	33.9	17.8	182.5	109.2	74
04	26.0	36.2	18.2	178.0	241.4	48
05	27.1	35.0	20.6	193.0	201.2	80
06	27.8	38.0	20.7	78.0	274.3	73
07	27.0	36.5	19.0	286.5	220.6	82
08	25.8	34.8	18.9	201.0	189.7	85
09	24.9	32.1	17.9	142.0	155.1	90
10	29.0	35.1	23.3	54.5	257.7	20
11	27.0	34.8	20.1	51.0	196.5	79
12	27.9	35.6	20.3	28.0	283.9	65
13	26.9	35.1	20.4	264.5	203.2	77
14	26.1	35.2	19.5	163.5	145.2	75
15	25.8	34.5	18.0	114.5	183.9	79

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇【1】2010年を含む過去の1等米比率と気象データ（平均気温、最高気温、最低気温、降水量、日照時間）を時系列で分析してみよう

表1のデータから、2010年以前の1等米比率と気温（平均、最高、最低）の推移を示す図1をExcelで作成した。

図1　1等米比率と気温の推移



Q1：図1から、2010年と同じような傾向を示す年をすべてあげなさい。

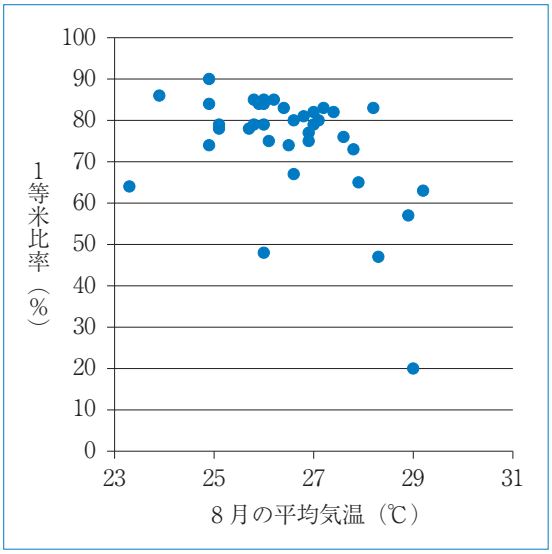
Q2：図1から、1等米比率と気温（平均、最高、最低）には、それぞれどのような関係があると読み取れるか。

Q3：1等米比率と（降水量、日照時間）にはどのような関係があるか。図1を参考にグラフを作成し、傾向をまとめなさい。

◇【２】１等米比率と気象データ（平均気温、最高気温、最低気温、降水量、日照時間）の相関関係を調べてみよう

Excelで１等米比率と平均気温の散布図を作成した（図２）。

図２　平均気温と１等米比率の関係



２つの変数の相関係数を求める関数（CORREL は相関係数 correlation coefficient の略）

また、CORREL 関数を使って相関係数を求めたところ

r＝－0.460となった。

相関係数と関連の強さの大まかな目安
（判断方法に統一のルールはない）
相関係数の絶対値
0.8以上…強い相関がある
0.4以上…弱い相関がある
0.2未満…ほとんど相関がない

Q４：図２と同様に、〔最高気温、最低気温、降水量、日照時間〕と１等米比率の散布図をそれぞれ作成し、相関係数を求め、下の表を完成させなさい。

	平均気温	最高気温	最低気温	降水量	日照時間
相関係数	－0.460				

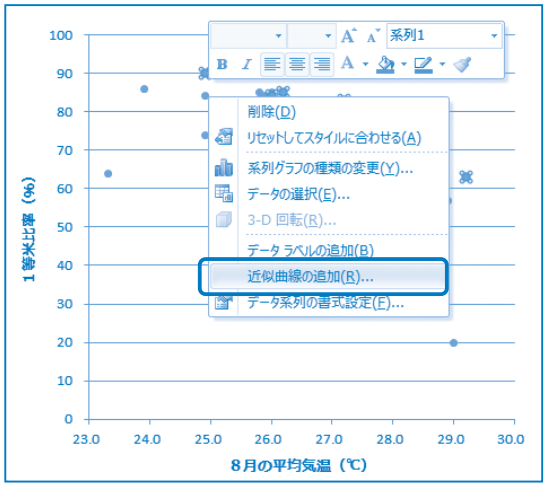
Q５：平均気温、最高気温、最低気温、降水量、日照時間のうち、１等米比率と相関関係があると考えられるものをあげなさい（相関係数0.4以上で相関があるとした場合）。

◇【３】回帰分析を用いて、品質が低下する可能性がある気温を求めてみよう

回帰分析とは、ある変数に他の変数がどのように影響するかを量的に求める分析手法である。

Excelで回帰式を求めるには「分析ツール」機能を使う方法もあるが、ここでは近似曲線の追加による方法を紹介する。

◆ Excel の散布図に近似曲線を追加する機能を使って回帰式を求める方法



①散布図上の**マーカー（点）**を右クリックして、**「近似曲線の追加（R)」**を選択する。

②**「線形近似（L)」**を選択し、**「□グラフに数式を表示する（E)」**と**「□グラフに R-2 乗値を表示する（R)」**にチェックを入れる。

図３　平均気温と１等米比率の関係

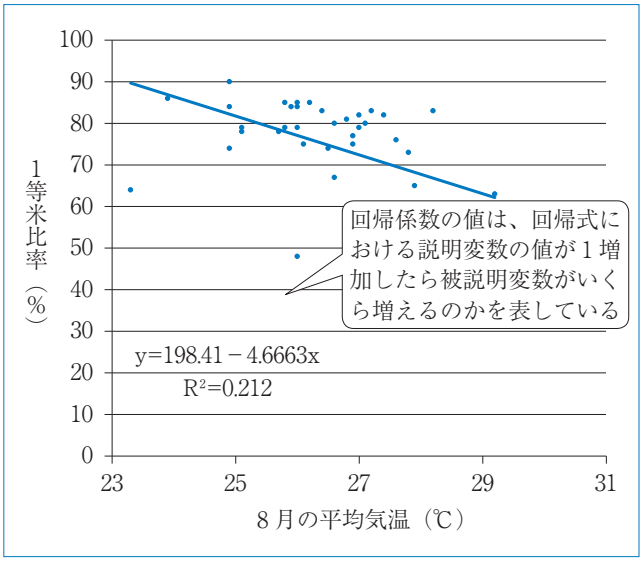


図２の１等米比率と平均気温の散布図に近似曲線を追加した。

グラフ上にプロットされたデータの傾向を視覚的に表したい時に引く。回帰直線もその１つである。

その結果を図３に示す。

表示された直線を回帰直線と呼び、その直線を表す式 y＝198.41－4.6663x が回帰式である（切片 198.41　回帰係数－4.6663）。

求めた回帰式に被説明変数と説明変数を書き入れると

１等米比率＝198.41－4.6663×平均気温と表すことができる。

Q６：回帰式を利用して、１等米比率が1979年～2015年の平均値74.7％以下になる平均気温は、何℃以上からであると考えたら良いか。

決定係数 R^2

R^2は決定係数と呼ばれ、被説明変数のバラツキが、説明変数のバラツキによってどの程度説明できているかを示します。また、被説明変数について、観測された値と回帰式で求められる値の間の相関係数を２乗した値と一致し、常に０～１の値になります。１に近いほど、見かけ上当てはまりの良い回帰式といえます。

今回の分析では、決定係数 $R^2=0.212$と表示されており、「被説明変数を説明変数で21.2％説明できている」と解釈します。

◆ 決定係数 R^2値ができるだけ高くなるような分析を試みる

決定係数は、説明変数と目的変数の直線関係の強さの尺度であり、値が大きければ回帰式による説明力が高い。決定係数を高くするためには、散布図を眺めて直線性があるか否かを確認する必要がある。たとえば、ある温度以上のデータに着目する、あるいは、被説明変数や説明変数を変数変換するなどの試みを行う。

平均気温27.0℃以上のデータに着目してみる！

図４　平均気温と１等米比率の関係

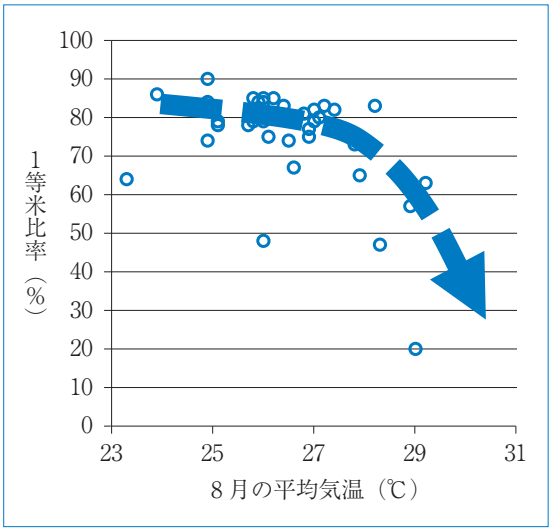


表２　平均気温27.0℃以上の年の１等米比率

年	平均気温(℃)	１等米比率(%)
1985	29.2	63
2010	29.0	20
1994	28.9	57
1999	28.3	47
2000	28.2	83
2012	27.9	65
2006	27.8	73
1984	27.6	76
1990	27.4	82
1988	27.2	83
2005	27.1	80
2007	27.0	82
2011	27.0	79
平均	27.9	68.5

図４の散布図を見ると、平均気温が上がると１等米比率が直線的に下がる傾向にあり、27.0℃付近を境に急激に１等米比率が低下する傾向が読み取れる。そこで、表２のように、平均気温が27.0℃以上の年のデータだけを抽出し、相関係数を求めたところ $r=-0.735$となった。したがって、このデータで単回帰分析を行うと決定係数 R^2の値は高くなり、回帰式の当てはまりも良くなると考えられる。

Ｑ７：表２の平均気温が27.0℃以上の年のデータを使い、回帰式と決定係数を求めなさい。

Ｑ８：Ｑ７で求めた回帰式を使って、１等米比率が平均気温が27.0℃以上のときの平均値68.5％を下回る平均気温は何℃以上からであると考えたら良いか。

Q９：Q７で求めた回帰式を使って、平均気温29.0℃〔2010年８月の平均気温〕のときの１等米比率を求めなさい。

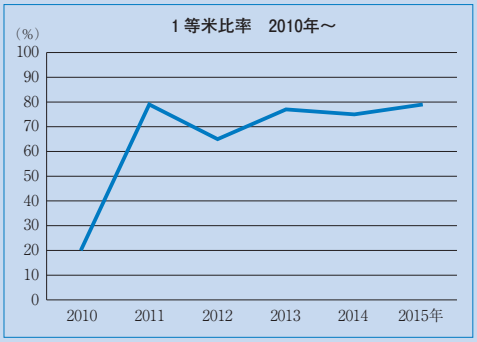
Q10：Q９の予測結果と2010年の１等米比率20％には差があるだろうか。差がある場合、その差についてどのようなことが考えられるか。

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

Q11：気象と品質の関係をまとめ、品質低下の対応策を考えなさい。

2010年（平成22年）のその後

新潟県では平成22年産米の１等米比率20％（過去最低）の結果を受けて、気候変動に対応できるさまざまな取組を進めてきました。そこでも統計は活用されています。たとえば、気象観測データから出穂期を予測することや移植期（田植えをする時期）と出穂期の相関関係を分析し、近年の夏の異常高温期に登熟期（米が成熟する時期）が重ならないように移植期を調整することなどの生産技術開発へのデータの活用です。そのような成果に加えて、県農林水産部や生産者の絶え間ない努力の結果、近年では安定的に高品質米が生産されています。



新品種「新之助」開発の経緯

新潟県では、新潟米ブランド力のさらなる強化を図るため、地球温暖化の進行に備え、高温条件下でも登熟性に優れ、品質が良く、食味に秀れた晩生新品種の開発を2008年から進めてきました。約20万株の個体から選抜と育成を繰り返し、最終的にコシヒカリとは異なるおいしさや特長をもつ品種を選定し、2015年９月に品種名を「新之助」とすることを発表しました。2017年から一般販売の予定であり、コシヒカリと並ぶトップブランド米となることが期待されています。

資料：新潟県農林水産部「平成27年度新潟県の農林水産業」

品質低下年の異常気象

1994年：生育期全般にわたる高温・多照の気象であり、一部地域では干ばつの被害がありました。出穂期は幼穂伸長期間の高温により著しく早まりました。登熟期間は過高温に推移し、８月は記録的な少雨でした。

1999年：７月下旬から８月中旬まで過高温、高夜温、無降雨により、登熟初期からデンプンの転流阻害が発生しました。

2004年：７月下旬の出穂期前から８月中旬の登熟初期まで高温で経過し、その後８月末まで低温で経過しました。新潟県中越地域で７月中に記録的豪雨による水害が発生し、また、沿岸平野部では８月下旬から９月上旬にかけて度重なる台風による潮風害が発生しました。

2012年：５月中下旬の低温、強風の影響で初期生育はやや不良となり、７月の梅雨明け以降の猛暑で過去最低の著しい低品質でした。

資料：新潟県農林水産部「水稲栽培指針」

〔本節の解答〕

Ｑ１：1985年、1994年、1999年、2004年

Ｑ２：１等米比率が例年より低い年は、平均気温、最高気温、最低気温が高くなる傾向がある。

Ｑ３：１等米比率と降水量の推移…目立った傾向は読み取れないが、品質低下を助長する要因になっている可能性もある。
１等米比率と日照時間の推移…日照時間が多いと１等米比率は低下する傾向が読み取れる。

Ｑ４：最高気温－0.25／最低気温－0.42／降水量0.15／日照時間－0.28

Ｑ５：平均気温、最低気温

Ｑ６：26.5℃

Ｑ７：$y = 555.78 - 17.471x$、$R^2 = 0.541$

Ｑ８：27.9℃

Ｑ９：49.1％

Q10：予測値と実際の値は29.1％と大きく乖離している。2010年の１等米比率の低下は平均気温だけが原因ではなく、他の要因も考えられるのではないか。

Q11：（例）１等米比率は特に気温との相関が高い。気象予報で平均気温が27.9℃を超えるような猛暑が予想出来る場合は、品質低下の可能性があるため、田植えの時期を遅らせ、登熟期間をずらす工夫をしたり、高温耐性の新品種を開発したりする必要がある。

新潟米“おいしい米”には秘密がある!?

米は新潟の代表的な特産物であり、特にコシヒカリは、1956年（昭和31年）に新潟県の奨励品種となってから60年を経た現在も全国一の生産・流通量を誇っており、消費者から「おいしい米」としての評価が定着している。高評価を得続けるためには品質とともに食味の高安定化を図る必要がある。

食味

食べたときのおいしさであり、食味検査で評価する。食味検査は、官能検査と理学的検査（分析検査）の２つに分けられるが、現状ではいずれの検査方法にも、まだ問題があるので、食味を正しく評価するためには両者の検査結果から総合的に判定する手法が主流となってきている。

食味官能検査

外観、香り、味、粘り、硬さおよび総合の６項目を基準米（基準米の評価値は０）と比較して評価する。各項目は－３（基準よりかなり悪い）～＋３（基準よりかなり良い）の７段階の数値で複数の評価者によって評価され、評価スコアは評価者の平均値で算出している。

外観：炊飯米の光沢の強さや白さ、煮崩れの有無等を評価
香り：ご飯特有の新米の香りの有無を評価
味　：ご飯のうまみで、喉ごしの感じの良い滑らかさ、噛むと感じる甘みを判断
粘り：ご飯の粘りの強弱を判断
硬さ：ご飯の硬軟の程度を判断
総合：各項目の評価スコアの合計ではなく、あくまでも総合的な判断

STEP 1：Problem 問題　課題の設定

◇「食味」の良い米を作るには

おいしい米の評価基準である食味を良くするにはどうしたら良いだろうか。

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 食味の評価を決める食味官能検査の「総合」評価を上げるために、「総合」と「外観」、「香り」、「味」、「粘り」、「硬さ」との関係を調べる

食味は米に含まれる玄米タンパク質と関係があることが分かっている。「総合」、「外観」、「香り」、「味」、「粘り」、「硬さ」と玄米タンパク質含有率の関係を明らかにするとともに、良食味米（食味が良いと評価された米＝「総合」評価が良い米）にはどれくらい玄米タンパク質が含まれているのか調べる。

【１】 Excel の「分析ツール」を使用し、以下の相関関係を調べる。
　① 「総合」と「外観」、「香り」、「味」、「粘り」、「硬さ」
　② 「玄米タンパク質含有率」と「総合」、「外観」、「香り」、「味」、「粘り」、「硬さ」

【２】「総合」評価を基準米より悪い（０未満）、基準米より良い（０以上）のデータに区分し、箱ひげ図を作成し、玄米タンパク質含有率の分布をみる。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 食味官能検査結果を使用する

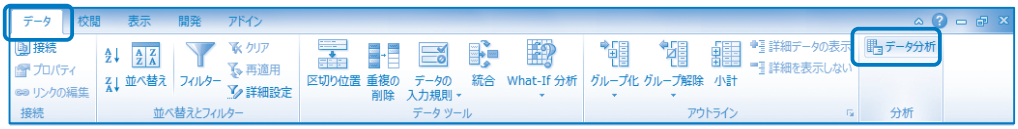
詳しい情報は、URL 参照 → http://www.jstat.or.jp/content/statsforschoolsadvanced/

（新潟県農業総合研究所作物研究センター「水田作栽培試験成績書」より引用）

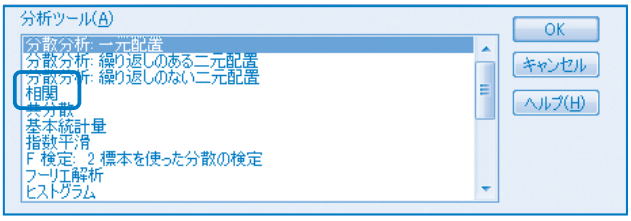
STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

【1】「総合評価」と他の評価項目（「外観」、「香り」、「味」、「粘り」、「硬さ」）の相関関係を Excel の「分析ツール」を使って調べよう。
また、玄米タンパク質含有率と各評価の相関関係も調べよう。

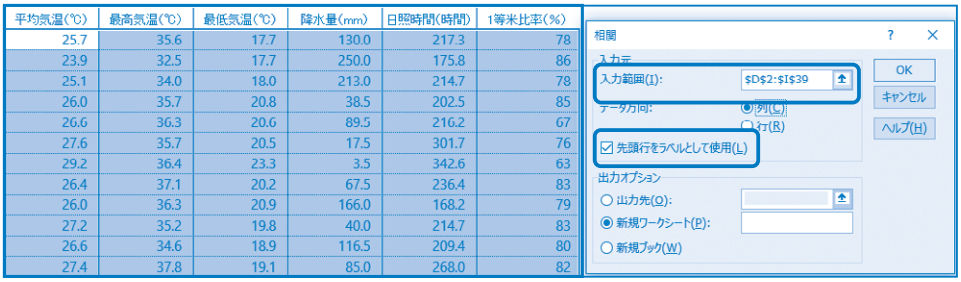
◆ Excel の「分析ツール」で 1 度に全ての組み合わせの相関係数を求める◆



① 「データ」タブの「分析」グループから、**「データ分析」**を選択する。



② 「データ分析」ウィンドウから「相関」を選択して「OK」を押す。



③ 「入力範囲」に相関係数を求めたいデータの範囲を指定する。この際にデータラベルを含めて範囲指定をし、「先頭行をラベルとして使用（L)」にチェックを入れ「OK」を押す。

表 3　相関係数行列

	平均気温（℃）	最高気温（℃）	最低気温（℃）	降水量（mm）	日照時間（時間）	1 等米比率（%）
平均気温（℃）	1					
最高気温（℃）	0.730	1				
最低気温（℃）	0.776	0.445	1			
降水量（mm）	−0.466	−0.418	−0.370	1		
日照時間（時間）	0.802	0.683	0.564	−0.624	1	
1 等米比率（%）	−0.460	−0.247	−0.422	0.150	−0.283	1

求める相関係数は小数第 3 位までの表示にしておこう。

相関係数が表 3 のように求められる。この表を相関係数行列という。
日照時間と平均気温の相関係数は0.802、1 等米比率と最高気温は −0.247である。

Q１：「総合評価」と相関関係がある評価項目を調べ、どのような傾向があるかまとめなさい。
（Excel の「分析ツール」で相関係数を求めるとともに、散布図を作成する。）

Q２：Q１と同様に、玄米タンパク質含有率と相関関係がある評価項目を調べ、どのような傾向があるかまとめなさい。

【２】 良食味米にはどれくらい玄米タンパク質が含まれているのかみてみよう。

Q３：全データを「総合評価」が、基準米より悪い〔基準米未満（０未満）〕、基準米より良い〔基準米以上（０以上）〕の２群に区分し、それぞれの箱ひげ図を作成してみよう。

Q４：作成した箱ひげ図から、どのようなことが読み取れるか。

STEP 5：Conclusion 結論　結論を導く

Q５：良食味米を作るにはどのようにしたら良いか。総合評価を上げる観点と米に含まれる玄米タンパク質の観点からまとめなさい。

SPAD 値を利用した玄米タンパク質含有率の調整

葉緑素計（SPAD）を用いて測定する稲葉身の葉色（緑色）の濃さは葉の葉緑素濃度と比例し、葉緑素濃度は窒素濃度と相関関係があります。このことから、葉色を測定することにより、稲体の栄養状態を簡易に知ることができます。出穂後の葉色（SPAD 値）が大きいと玄米タンパク質含有率が高いことが分かって、玄米タンパク質含有率が高いと食味が劣ることから、出穂後の葉色値を高めないことが重要です。とくに、出穂期15日後の葉色との相関が高く、この時期に水田の稲葉色を測定すると、収穫前でもお米の食味が推定できます。
（新潟県農業総合研究所作物研究センター）

資料：新潟県農林水産部「おいしい米づくりのポイント」

〔本節の解答〕

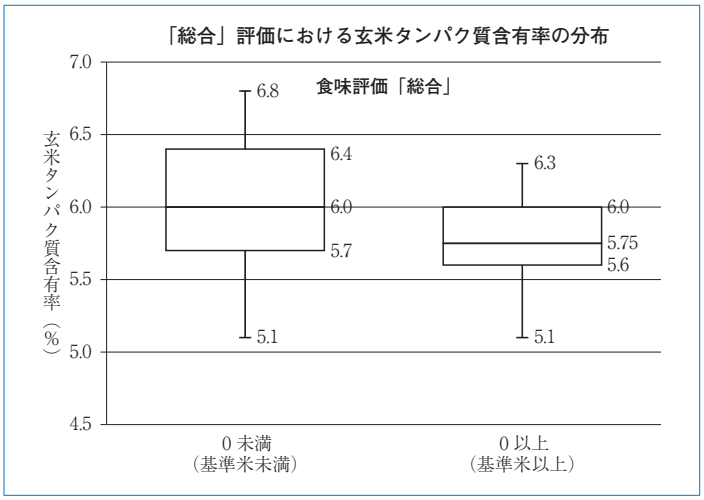
Ｑ１：散布図は省略

	総合	外観	香り	味	粘り	硬さ	玄米タンパク質含有率（%）
総合	1						
外観	0.334	1					
香り	0.340	0.275	1				
味	0.785	0.269	0.212	1			
粘り	0.662	0.252	0.189	0.478	1		
硬さ	－0.403	－0.149	－0.122	－0.227	－0.498	1	
玄米タンパク質含有率（%）	－0.439	－0.156	－0.097	－0.346	－0.543	0.532	1

「総合」評価と「味」、「粘り」、「硬さ」評価と相関関係がある。

Ｑ２：玄米タンパク質含有率は「総合」、「粘り」、「硬さ」評価と相関関係がある。散布図は省略。

Ｑ３：



Ｑ４：基準米以上米の玄米タンパク質含有率は5.6～6.0％にデータが集中していて、基準米未満米のデータの分布と比較すると、玄米タンパク質含有率のデータは総じて低く分布している。

Ｑ５：食味官能検査の「総合」評価は「味」、「粘り」、「硬さ」の評価と相関があるため、この３点を意識して米作りをしていく必要がある。そして、玄米タンパク質含有率が高いほど粘りが弱く、硬く、「総合」評価が低くなる傾向があるため、玄米タンパク質含有率が高くならないように稲を栽培する必要がある。

2　AED で救える命を増やそう

AED の必要性や問題点について考え、自分たちが住む地域では AED が必要に応じて設置されているかどうかを調べる活動は重要である。ここでは、**地理情報システム**（**GIS**：Geographic Information System）を活用して、300m ごとに設置するという条件に合っているかどうか、距離だけではなく必要なところに設置されているかどうか、どの場所に増設すべきかという観点で探究しよう。

＜ボロノイ図＞

AED マップ（どの AED 設置場所へ行くべきかを示す地図）の作成には**ボロノイ図**が有効です。実際に、AED 設置場所をボロノイ図にしたものをインターネットで公開している地域もあります（例えば、群馬県の高崎市・前橋市）。ボロノイ図とは、平面上にいくつかの点が配置されているとき、その平面内の点を、どの点に最も近いかによって分割してできる図のことです。ボロノイ図は、さまざまな分野で利用されており、AED 設置場所に類似するものでは、救急車の担当地区の決定、PHS の基地局の探索や新しい基地局の決定などが挙げられます。

STEP 1：Problem 問題　課題の設定

◇ AED の適正配置はどうあるべきか？

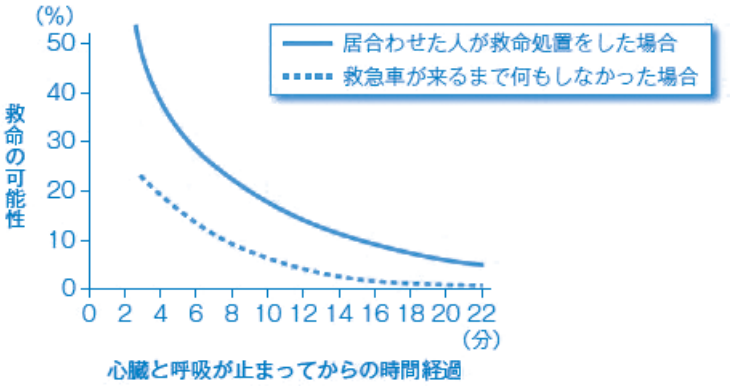
AED は、心停止した時に心臓の状態を判断して、電気ショックが必要な場合には、電気ショックを与えることで、正しい心臓のリズムに戻す機器である。私たちが住む地域には、AED が必要なところに十分に設置されているのだろうか。

心肺停止時間と救命率

心停止時間と救命率の関係を示す図 1 のグラフから、救命処置の開始時間と救命の可能性について分かること。

- 心停止からの経過時間が 2 分経つごとに救命の可能性が10％低くなる
- 救命処置をした場合、何もしなかった場合に比べて、救命の可能性が 2 倍になる
- 救命の可能性と心停止や呼吸停止からの経過時間の関係は反比例のようだ
  →心停止からなるべく短い時間で使用できるように設置するのが理想

図 1　救命の可能性と時間経過

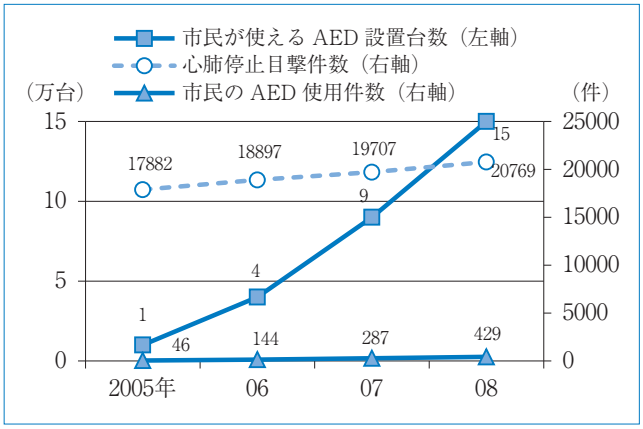


資料：厚生労働省「救急蘇生法の指標2015」

市民が使える AED 設置台数と市民の使用件数、心肺停止目撃件数の2005～08年の推移を示す図 2 のグラフから分かる設置台数と使用台数についての問題点を示すと

・市民が使える AED 設置台数は増加しているが、AED 使用件数はほとんど変化していない
・使用件数が増えない理由があるのでないか
・AED の設置場所が周知されていないのではないか

図 2　AED の設置台数と使用件数および心肺停止目撃件数



資料：消防庁「救急蘇生統計」、厚生労働省「平成21年度循環器疾患等の救命率向上に資する効果的な救急蘇生法の普及啓発に関する研究」

AED を何 m ごとに設置すれば良いかの基準

たとえば、次のような考えが予想される。

・片道 2 分程度で取りに行くとする。10秒で50m を走る速さなら 2 分間（120秒）走り続けることは可能。50m×12＝600m より600m ごとに設置する必要がある。
・片道 1 分程度で取りに行くとする。10秒で50m を走る速さなら 1 分間（60秒）走り続けることは可能。50m×60＝300m より300m ごとに設置する必要がある。

なお、一般財団法人日本救急医療財団が示している AED の適正配置に関するガイドラインには「現場から片道 1 分以内の密度で配置」とある。また、日本心臓財団の HP（http://www.jhf.or.jp/aed/arrangement.html）には「300m ごとに AED が設置されていると、150m/分で早足に取りに行けば、その間のどこからでも 1 分以内で AED が届き、5 分以内に除細動が可能となる。」とある。

300m ごとに AED を設置することを基準としているかを確認するのが第一歩である。

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ AED は十分に設置されているのか、また、AED は距離に基づいて設置すれば良いのか？

ここでは、**jSTAT MAP**（https://jstatmap.e-stat.go.jp/）を利用して、東京都練馬区を対象として分析する。jSTAT MAP では、地図上に年齢別や男女別の情報などを表示させることができる。AED の適正な設置場所の検討において、適切なデータを選択し、それに基づいて AED の新たな設置場所を提示しよう。

Q 1：AED 設置場所を検討するに当たって、あなたなら、どのデータを選ぶだろうか？
選びたいデータ：
〔理由〕

STEP 3：Data 収集　必要なデータ・統計資料を集めよう

◇ jSTAT MAP で実際にデータを地図上に表示しよう

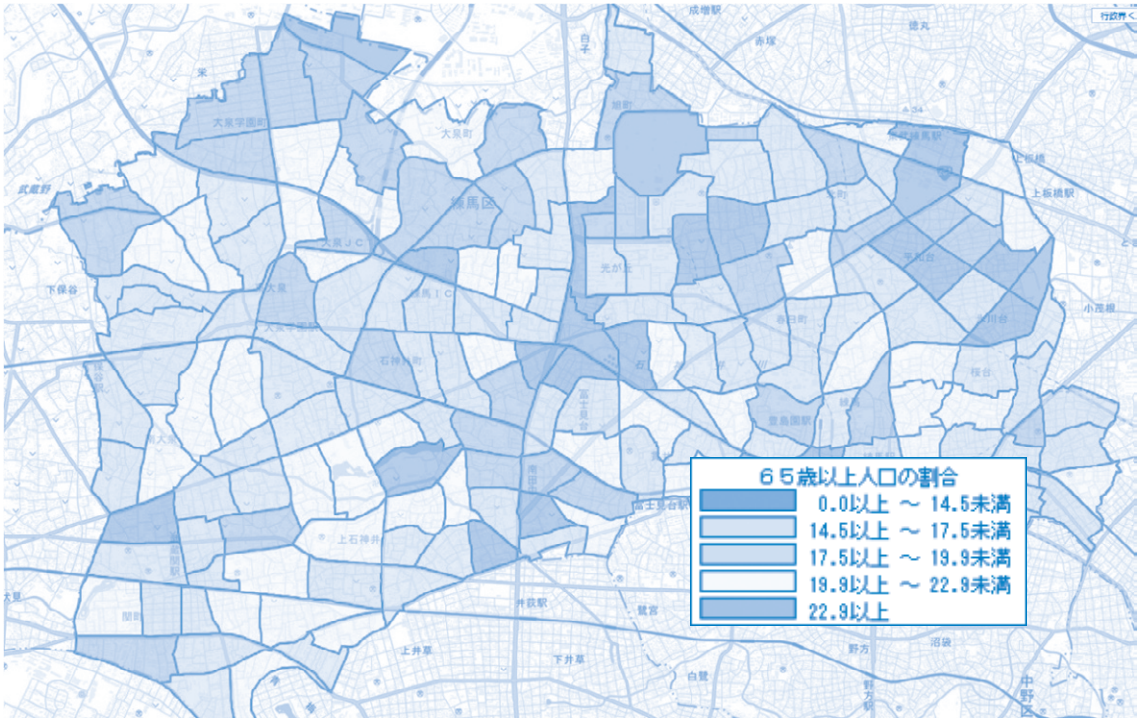
図 3 は、東京都練馬区における AED の設置場所を示す。

図 3　東京都練馬区における AED の設置場所

◇ AED の設置の必要性と関連する指標を jSTAT MAP で図示しよう

ここでは、AED の使用頻度が高いと考えられる高齢者（65歳以上人口）の割合を取り上げる。

図 4　東京都練馬区の65歳以上人口の割合



Q２：AED の設置場所を検討する際に、65歳以上人口の割合の他に、どのようなデータを選んだら良いか考えよう！

STEP 4：Analysis 分析　グラフや統計量で傾向を捉えよう

◇【１】AED が300m ごとに設置されているか調べる

図 5 について

① ２つの AED の距離が300m 以内になっているかどうかを定規やコンパスで調べる

② AED を中心にして半径150m の円を描いてみる

③ jSTAT MAP で、AED の設置場所を中心にして半径150m の円を地図上に描く

図 5　AED の設置場所と半径150m の円

Q３：図 5 から、AED の設置場所について問題点をあげてみよう。

円の外部にある場所は AED までの距離が遠いことが分かり、このような場所には増設すべきと考えられる。

AED は距離だけを考えて設置すれば良いのだろうか？

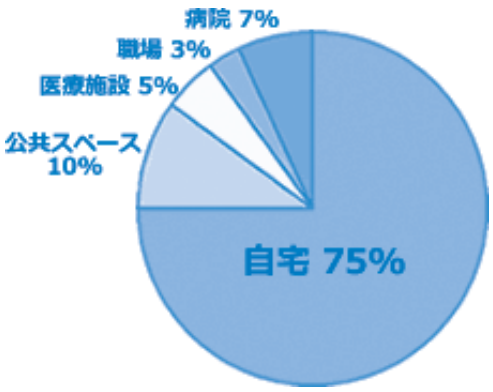
300m ごとに設置するというのは１つの基準ではあるが、使用する市民のことも考えてみよう。

図６の円グラフは、心原性心肺停止の発生場所を表したものである。心臓が突然動かなくなって倒れるのは自宅がほとんどであることが分かる。

また、心肺停止になる人たちはどのような人たちに多いのだろうか。

AED を使用する確率がより高い場所に設置されているのだろうか。

図６　心肺停止の発生場所

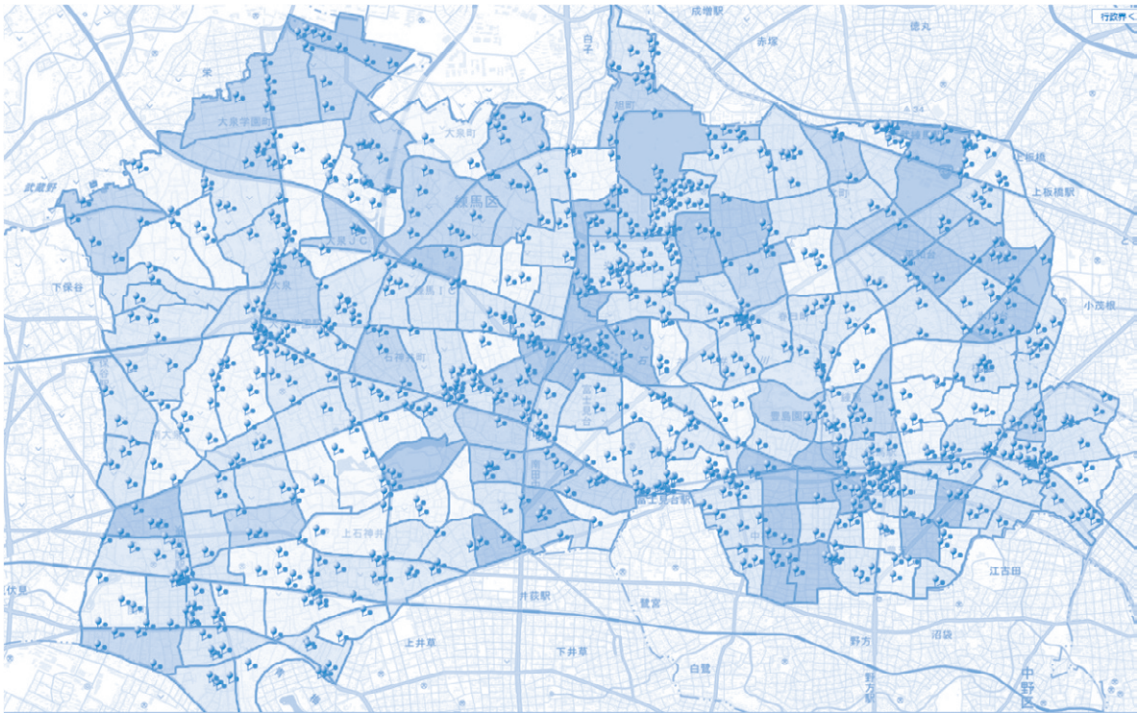


資料：ウツタイン大阪プロジェクトより引用

◇【２】AED 設置場所は、必要度に対応しているか？

AED 設置場所と65歳以上人口の割合のデータを重ねて表示した図７をもとに、問題点を考察しながら、新たな設置場所を検討する。

図７　AED 設置場所と65歳以上人口の割合



Q４：図７から分かった問題点は何か？

STEP 5：Conclusion 結論　分かったことをまとめ・読み取ろう

◇ データをもとにして、AED 設置場所を提案しよう！

Q５：何故そのデータが必要なのかを明らかにして、提案書を作成しよう！

AED マップを作ろう

どの AED 設置場所へ行くべきかを示す AED マップを作りたい。

（ア）　AED の設置場所を Ａ１、Ａ２とし、Ａ１と Ａ２の間にある場所で AED が必要になったとき、どちらの AED 設置場所へ行くべきかを一目で分かるようにするにはどうするか。

→Ａ１と Ａ２の垂直二等分線を作図する。垂直二等分線上の地点はどちらからも距離が等しい地点であるので、垂直二等分線は境界線となる。

（イ）　右の地域の AED マップを作ってみよう。

→地図上にある AED 設置場所４箇所のうち、近接する２箇所ずつをとり、それぞれ垂直二等分線を引いていく。

〔本節の解答〕

Q１：高齢者の人口

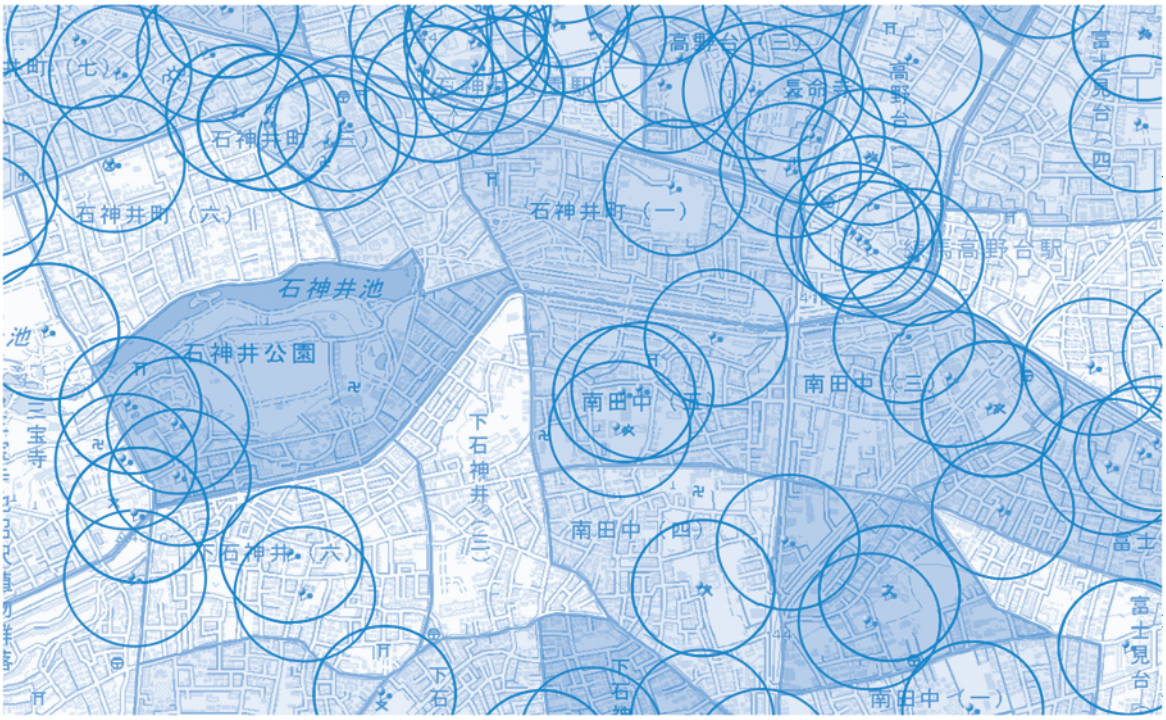
Q２：70歳以上の人口の割合
　　65歳以上の単身世帯の割合
　　診療所・病院の所在

Q３：駅周辺に住んでいる人は比較的若い人が多い。
　　埼玉県に近いところは65歳以上人口の割合が高い。

Q４：駅周辺は比較的若い人が多いにもかかわらず、たくさんの AED が設置されている。
　　65歳以上人口の割合が高い地域でも AED が設置されていない地域がある。

Q５：図８の石神井地区では、65歳以上人口の割合が高いのに、AEDが設置されていない場所が随所にあるので、
　　これらの地域に AED の設置を提案する。

図８　65歳以上人口の割合が高いのに、AED が設置されていない場所に焦点を当てた地図



地図による小地域分析「jSTAT MAP」で実際に統計地図を作ってみよう！

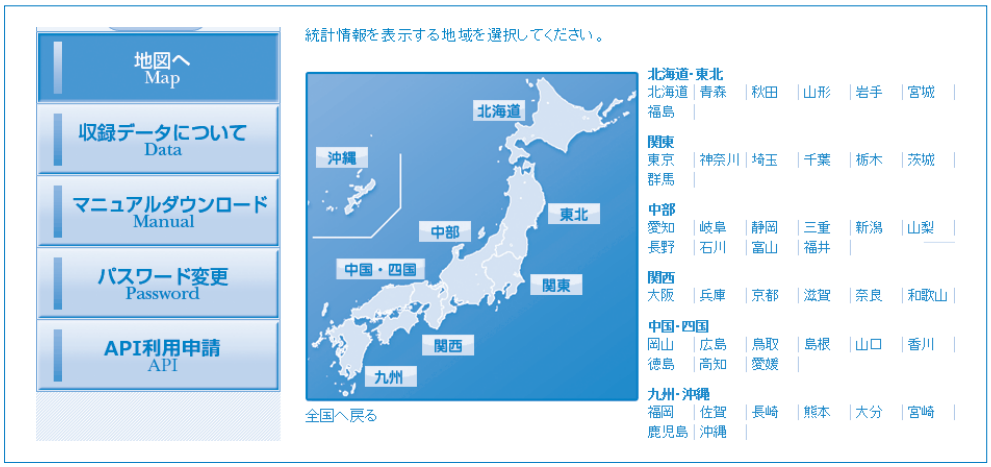
「jSTAT MAP」を利用すれば、自分たちの住む地域を概括することができる。統計データとして、国勢調査の人口などの小地域集計結果や経済センサスの企業・事業所数、従業者数がすでに収録されており、簡単に地域分析が可能である。また、ジオコーディング（住所などから地図上にポイントデータとして表現（登録）すること）や小地域の特徴など主題別に地図を作成することもできる。「jSTAT MAP」の初期画面には、「e-Stat」（政府統計共同利用ポータルサイト）から入ると分かりやすい。

1　「e-Stat」の「地図や図表で見る」の中から「地図による小地域分析（jSTAT MAP）」を選択し、初期画面を表示する

図　jSTAT MAP の初期画面

ログインをクリック。ID、パスワードを入力。※ID、パスワードは利用申込みから事前に入手する必要がある。

ログインすると日本地図が出てくる。利用する都道府県を選択すると県庁所在地の県庁を中心にした地図が表示される。



なお、マニュアルには、利用ガイドもあり、「jSTAT MAP」の機能についてダイジェスト的に分かりやすく説明されている（マニュアルダウンロードをクリックし参照する。）。

＜練馬区の AED の分布を地図に表してみよう＞

2　プロット（plot）機能を使い、ある場所（住所）を地図上に表示（ジオコーディング）する

［準備するもの］

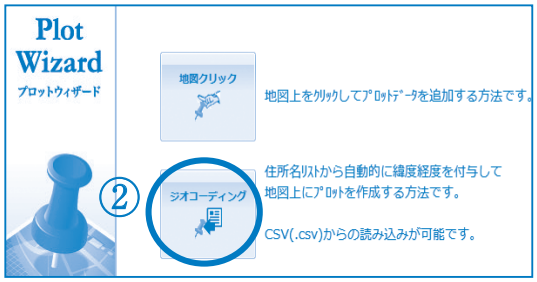
住所名リストイメージ

施設名	住所	登録番号	一連番号
株式会社イノベーションクリエイト	東京都練馬区向山4-17-10	495803	447
グループホーム石神井の森	東京都練馬区上石神井2－20－18	577259	698
日本通信（株）板橋営業所	東京都練馬区旭丘2-22-15	464460	400
北練馬総合病院	東京都練馬区旭丘1-24-5	343832	111
東京北信用金庫江田支店	東京都練馬区旭丘2-27-3	504459	479
そとうみ歯科クリニック	東京都練馬区旭丘2-54-10	589224	710
やすやま 江古田店	東京都練馬区旭丘1-74-17	381386	207
北武鉄道株式会社　江田駅	東京都練馬区旭丘1-76-13	499896	456
ロピー産業株式会社	東京都練馬区旭丘1-9-8	420649	283

住所名リストを作成する場合は２列目に必ず住所が来ること。１度に地図に表示できるのは2000地点と限度があるのでデータが2000以上の場合はリストを分割すること、データ形式はCSV で保存すること。この３点に注意して作成する。

①　プロット（plot）をクリック
②　ジオコーディングを選択
③　参照をクリックし、作成した住所名リストを呼び出す
④　アップロードし、ジオコーディング結果を確認
⑤　「登録する」をクリックすると地図上アイコンが表示される。

※④の「ジオコーディング結果を確認」については、129ページを参照。



下表はジオコーディング結果を確認（評価）するためのリスト例であるが、登録する前にジオコーディング結果ファイルのダウンロードを行う。ファイル例ではＨ欄のマッチングレベルを確認する。７以下の数値であれば、リストを確認するか地図上で確認し、位置を移動（修正）する必要がある。

	A	B	C	D	E	F	G	H	I	J
1	名称	設置施設住所	設置場所補	登録番号	番号	マッチング緯	マッチング経	マッチングレ	説明	
2	S	S	S	S	S					
3	株式会社千曲テクノクリエイト	東京都練馬区向山4-17-9	1F事務室	495803	447	35.74364	139.6362	9	建物・ランドマーク	
4	小規模多機能　グループホームし	東京都練馬区上石神井2-20-13	1Fエントラ	577259	698	35.72853	139.5944	9	建物・ランドマーク	
5	日本通運(株) 板橋営業課	練馬区旭丘1-22-13		464460	400	35.73382	139.6778	9	建物・ランドマーク	
6	練馬総合病院（1Fエントランス）	練馬区旭丘1-24-1		343832	111	35.73351	139.6771	9	建物・ランドマーク	
7	東京東信用金庫江古田支店	練馬区旭丘1?27?9	営業室(カ	504459	479	35.7331	139.6761	9	建物・ランドマーク	
8	うちうみ歯科クリニック	練馬区旭丘1?54?9アイウッド1F	受付	589224	710	35.73625	139.6733	9	建物・ランドマーク	
9	やすだ 江古田店	練馬区旭丘1-76-7		381386	207	35.7371	139.6726	9	建物・ランドマーク	
10	西武鉄道株式会社　江古田駅	練馬区旭丘1?78?7		499896	456	35.7375	139.6729	9	建物・ランドマーク	
11	トピー工業株式会社 社宅	練馬区旭丘1-9-6		420649	283	35.73299	139.6783	9	建物・ランドマーク	
12	公益財団法人東京都医療保健協	練馬区旭丘1丁目24?1		508222	481	35.73351	139.6771	9	建物・ランドマーク	
13	練馬区立旭丘小学校	練馬区旭丘2-21-1		336783	89	35.7367	139.6769	9	建物・ランドマーク	
14	株式会社　安田屋　やすだ江古田	練馬区旭丘一丁目76番7号		510195	488	35.7371	139.6726	9	建物・ランドマーク	
15	江古田駅前交番	練馬区旭丘一丁目78番2号	交番内	511394	508	35.73792	139.6713	9	建物・ランドマーク	
16	株式会社　ヤナセ東京営業本部	練馬区旭町1?22?8		482805	430	35.76444	139.62	9	建物・ランドマーク	
17	ドラゴンマンション光が丘公園壱番	練馬区旭町1-40-12		398562	245	35.76651	139.6207	9	建物・ランドマーク	

＜練馬区の AED の分布と65歳以上人口割合の統計を重ねてみよう＞

3　jSTAT MAP で地図（レイヤー）を重ねて表示するには、レイヤー名をクリックし、表示更新をクリックする

4　作成されたエリア円内の人口・世帯数を調べることもできる

統計データ（graph）、種類で平成22年国勢調査（小地域）を選択する

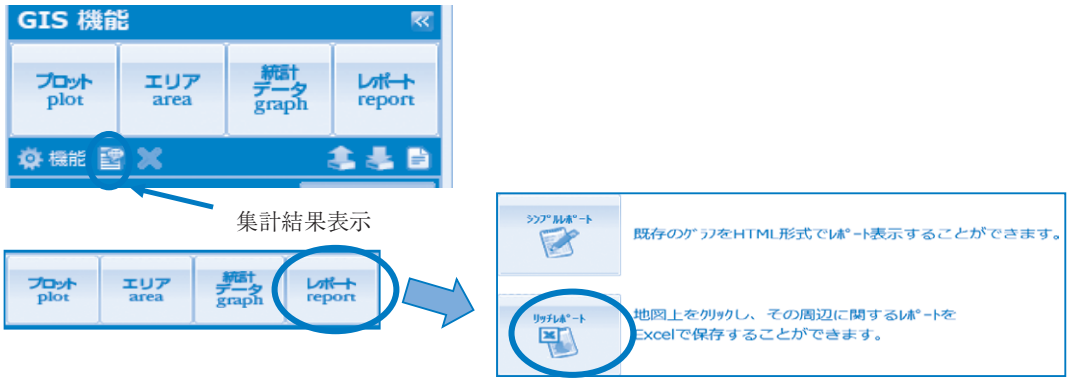
分類で男女別人口総数および世帯総数を、指標で人口総数および世帯総数を選択し、指標選択をクリックし、次の画面に進み、エリアを選択、既存エリアグループから集計を行うエリアをクリックし、「按分する」を選択、集計開始をクリックすると、円内150m の内の人口および世帯数が集計され、それぞれの階級別の人口および世帯数を色別に表示することができる。

また、集計結果は機能の横のファイルを開くと画面の下に表示される。

なお、機能をクリックし、集計結果を地図上に表現することもできる。

5　レポート機能を使って、リッチレポートを作成することもできる

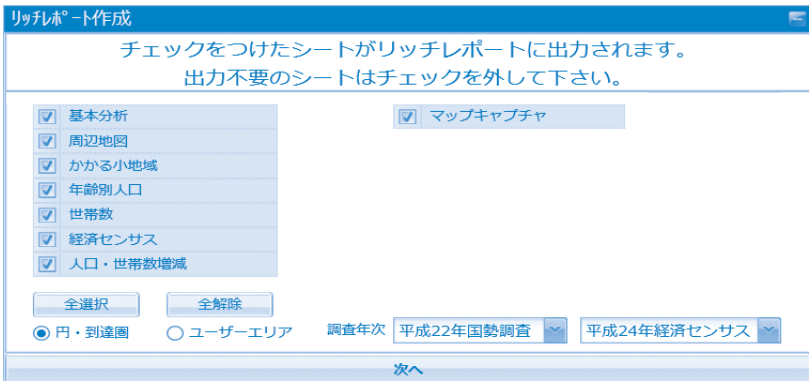
リッチレポートは、わずかなクリック数で指定した地点の周辺について、統計データを集計し、人口ピラミッド、年少人口・生産人口・老年人口などとしてエクセル形式でレポート出力する機能である。

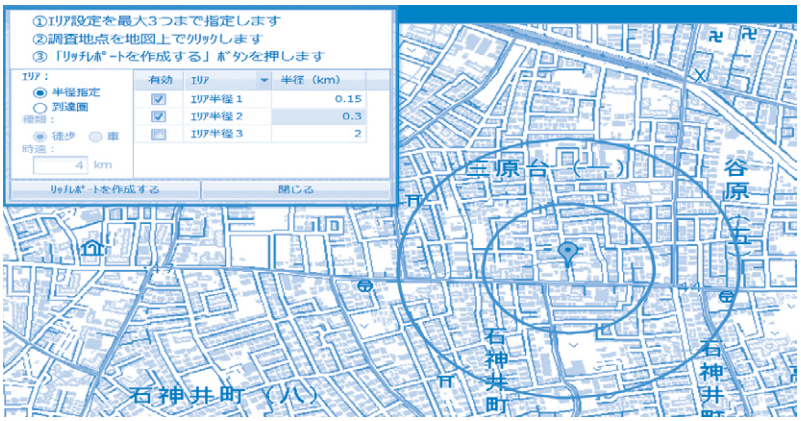


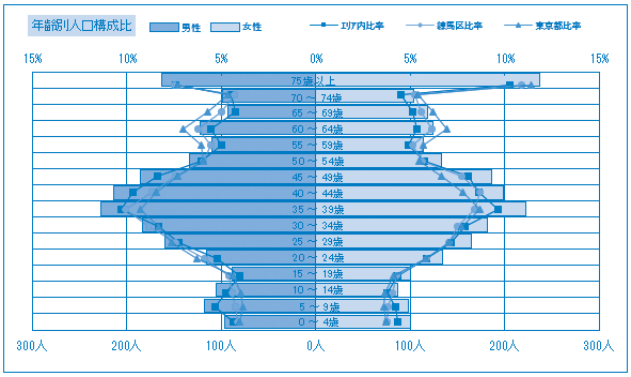
レポート（report）→リッチレポート→次へ→エリア半径設定→中心ポイントを指定→「リッチレポートを作成する」の流れで作成する。

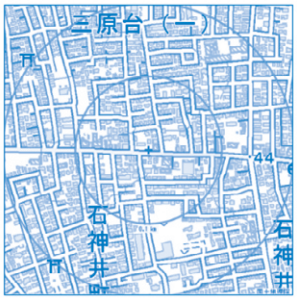
下例では、ある地点から半径150m、300m の地域について、どのような地域なのか、統計データを使って自動的に集計している。レポート結果はエクセル形式により基本的に 9 シートで構成されている。また、周辺地図についてもキャプチャしてくれる特徴をもっている。

レポート結果の一部を紹介すると、人口ピラミッド、男女別人口、５歳階級別人口など、次のとおりである。









データ名	人口				
	1次エリア	2次エリア	3次エリア	練馬区	東京都
人口総数	1,171	4,535		716,124	13,159,388
男人口	573	2,219		350,647	6,512,110
女人口	597	2,316		365,477	6,647,278

75歳以上	104	400		65,691	1,215,904
70-74	53	202		34,136	654,931
65-69	55	211		37,798	771,396
60-64	63	246		44,113	905,914
55-59	58	224		38,204	760,764
50-54	70	266		41,338	740,091
45-49	97	371		54,037	905,561
40-44	107	411		61,777	1,053,232
35-39	115	450		65,583	1,164,057
30-34	92	365		58,098	1,038,768
25-29	83	325		51,541	949,354
20-24	65	250		41,864	785,911
15-19	49	189		31,543	546,573
10-14	50	192		30,246	492,799
5-9	56	215		28,770	484,303
0-4	50	197		28,241	500,269

年少人口（0歳～14歳）	157	604		87,257	1,477,371
生産年齢人口（15歳～64歳）	798	3,097		488,098	8,850,225
老年人口（65歳以上）	211	814		137,625	2,642,231
15歳以上就業者数	525	2,039		321,148	6,012,536
後期高齢者数（75歳以上）	104	400		65,691	1,215,904

データ名	世帯数				
	1次エリア	2次エリア	3次エリア	練馬区	東京都
一般世帯総数	544	2,114		335,952	6,382,049
単身世帯	225	879		142,811	2,922,488
2人以上世帯	319	1,235		193,141	3,459,561
核家族世帯	295	1,140		176,059	3,078,860
夫婦のみの世帯	99	383		61,195	1,081,892
夫婦と子供から成る世帯	154	595		88,960	1,516,499
6歳未満親族のいる世帯	52	200		26,930	473,941
65歳以上親族のいる世帯	146	563		94,201	1,837,074
持ち家世帯	259	1,003		154,373	2,927,775
民営借家世帯	237	932		140,274	2,533,628

平成22年国勢調査

3　人口減少社会に向かう地域の課題と取り組み

地域の課題を考える上で最も基礎的なデータの１つが地域の人口である。人口の推移は単に人の動きを示しているだけではない。都道府県や市町村の補助金・交付金の算定や選挙の定数の算定に使われるほか、将来人口の推計を通して、公共施設の計画や経済・福祉の施策策定の基礎資料などにも使われる。

私たちが地域の課題に取り組む際も、人口の状況を確認することは非常に大切である。

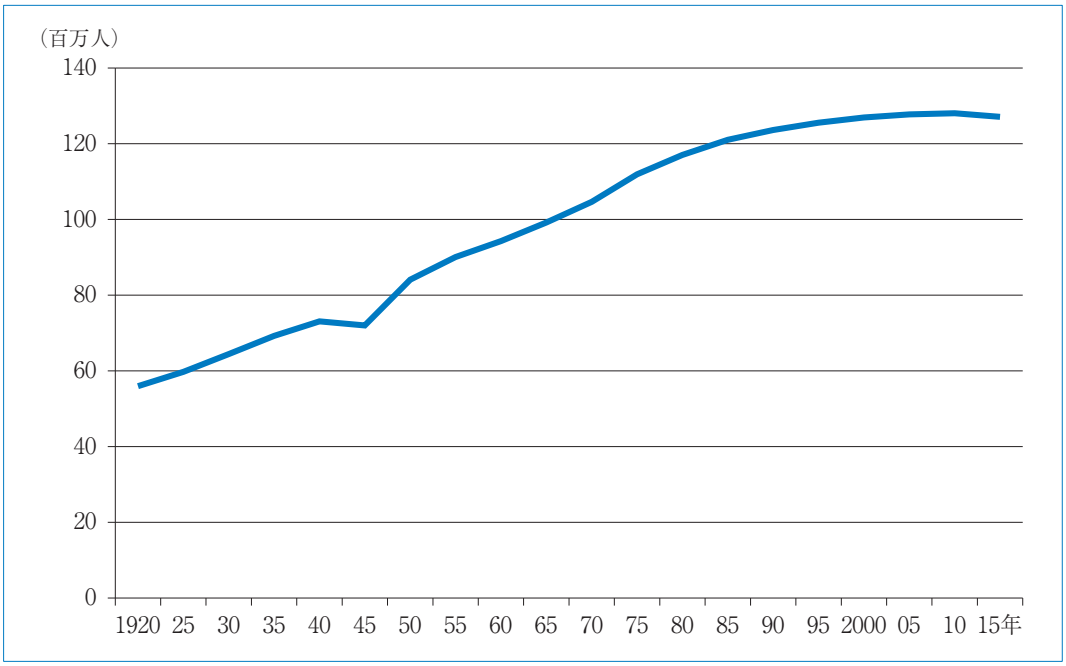
STEP 1：Problem 問題　課題の設定

◇ 減少が進む地域人口を探る！

日本の人口に関する最も基礎的で大規模な統計調査が**国勢調査**である。1920年の第１回調査以来、５年ごとに10月１日現在の全国の人口・世帯数を詳しく調べており、調査結果もすべて公開されている。

図１は、1920年～2015年の全国の人口を図示したものである。

図１　全国の人口推移



データは、e-Stat に掲載されている国勢調査の時系列データである。ただし、1945年は国勢調査が実施されなかったため、1945年人口調査の数字となっており、沖縄県は調査されなかったため０になっている。

Q 1 ：人口の推移を分析するため、図 1 に人口推移の傾向を表す直線を書き入れ、増加の特徴を詳しく見てみよう。

図 1 を見ると、人口は1965年頃までは直線的に増加しているが、1985年以降は出生率の低下によって、人口増加が鈍化している。さらに、2015年国勢調査では人口は減少に転じており、今後も従来とは異なって推移すると考えられる。実際、国の研究機関である国立社会保障・人口問題研究所は、これからの日本の人口は減少が続くと予測している。

国立社会保障・人口問題研究所は、厚生労働省に所属する国立の研究機関で、人口や世帯の動向を捉えるとともに、内外の社会保障政策や制度についての研究を行っている。

日本全体の人口が減少するなかで、地域の人口減少はどのように進むのだろうか。

STEP 2：Plan 計画　どのようなデータ・統計資料を集めるか

◇ 人口減少が先行している地域は？

地域の人口減少がどのように進むのかについて検討するためには、いち早く減少が始まっている地域の動きが参考になるだろう。

都道府県の人口推移を見てみよう。都道府県の人口データの入手方法や分析方法は、「生徒のための統計活用～基礎編～」（http://www.soumu.go.jp/toukei_toukatsu/index/seido/stkankyo.htm）で紹介されているので、この方法を参考に、過去からの人口データを e-Stat から入手し、分析する。

ただし、そのままの数字で比較しても、桁が違いすぎて、推移の違いが分かりにくいので、1920年（大正 9 年）の人口を100としたときの数値（**指数**）にする。人口推移を指数化して折れ線グラフで示したのが図 2 である。

指数は変化の様子を表すため、ある基準を100としたときの比率。異なる種類の数字の動きを比較するとき便利だよ。

Ｑ２：自分が住む都道府県はどれだろうか。また、特徴的な変化をしている都道府県はあるだろうか。

図２　第１回国勢調査（1920年）の人口を100としたときの都道府県別人口の推移

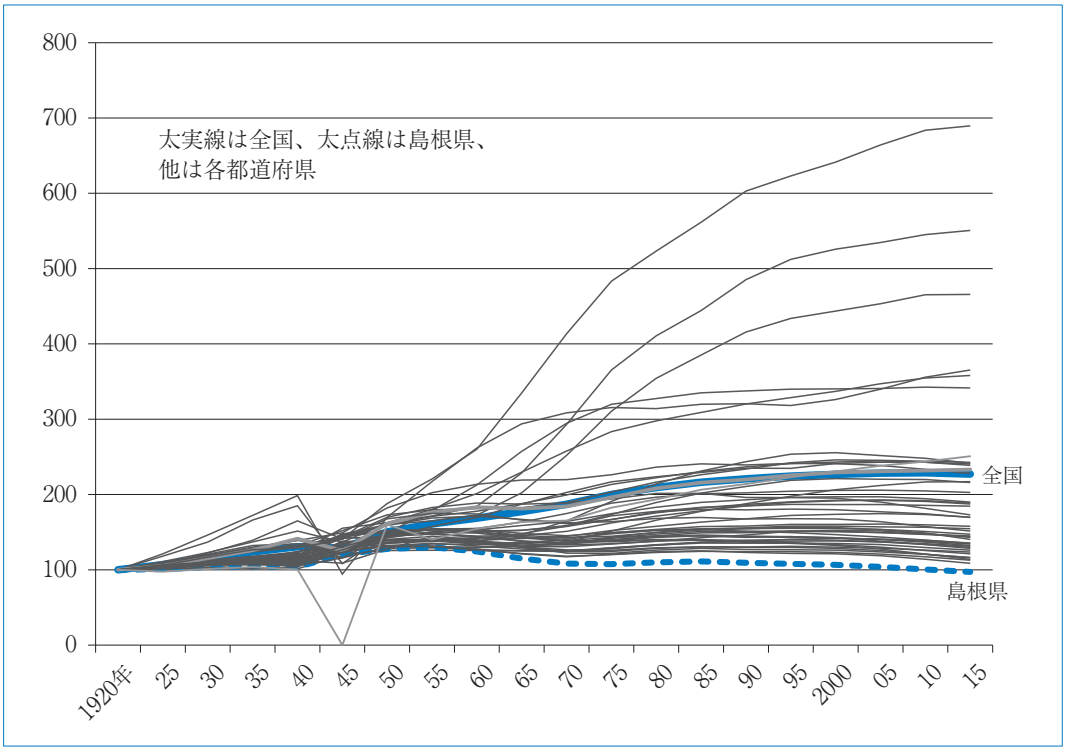


図２によれば、2015年の国勢調査結果が1920年の人口を下回ったのは、島根県のみであることが分かる。そこで、地域人口の推移を探るために、全国で最も早く人口減少が進んだ島根県の人口動向に着目することにする。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 島根県の人口推移は？

国勢調査の結果によれば、島根県の人口は表１の１行目のように推移した。これを、1920年を100とした指数で表すと、表１の２行目のようになる。なお、島根県で最も人口が多かったのは、1955年の929,066人であった。

表１　島根県の人口の推移

島根県	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010	2015年
人口総数（千人）	715	740	741	913	889	774	785	781	762	717	694
指数（1920年＝100）	100	103	104	128	124	108	110	109	107	100	97

島根県の人口減少はどのような形で進行したのかを、詳しく見ることにする。

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ 人口減少を世代別に分析してみよう

国勢調査の結果から、島根県の人口減少にどの年齢層が大きく影響したのかを見てみよう。表２－１は、年齢階級を５歳刻みにして表示した1955～75年の人口である。1955年は島根県の人口が最も多かった年であり、それから20年間、人口減少が続いた。とくに、1960～65年の減少率は、全国で最も大きなものであった。

表２　1955～75年の年齢階級別にみた人口推移

表２－１　人口の総数

年齢階級	1955年	1960年	1965年	1970年	1975年
総　数	929,066	888,886	821,620	773,575	768,886
０～４歳	95,321	72,500	59,331	52,296	56,482
５～９歳	119,435	93,089	69,219	58,285	53,313
10～14歳	101,415	117,007	89,853	67,876	58,277
15～19歳	77,344	67,581	78,566	62,425	53,285
20～24歳	75,171	56,643	45,483	51,575	42,692
25～29歳	71,152	68,036	50,893	45,640	56,957
30～34歳	61,115	67,490	62,934	49,761	47,266
35～39歳	51,992	58,492	63,692	61,111	49,946
40～44歳	49,687	49,741	55,702	61,693	60,487
45～49歳	44,725	47,261	47,133	53,451	60,291
50～54歳	42,018	42,272	44,142	44,959	51,718
55～59歳	38,323	39,304	39,213	41,425	43,043
60～64歳	31,203	34,753	35,528	36,133	39,256
65～69歳	26,254	27,132	30,124	31,598	33,120
70～74歳	20,551	21,147	21,975	25,041	27,176
75～79歳	13,837	14,698	15,177	16,108	19,337
80歳以上	9,514	11,740	12,655	14,198	16,198
年齢不詳	9	0	0	0	42

表２－２　人口の変化率

55－60	60－65	65－70	70－75
▲ 4.3	▲ 7.6	▲ 5.8	▲ 0.6
▲ 23.9	▲ 18.2	▲ 11.9	8.0
▲ 22.1	▲ 25.6	▲ 15.8	▲ 8.5
15.4	▲ 23.2	▲ 24.5	▲ 14.1
▲ 12.6	16.3	▲ 20.5	▲ 14.6
▲ 24.6	▲ 19.7	13.4	▲ 17.2
▲ 4.4	▲ 25.2	▲ 10.3	24.8
10.4	▲ 6.8	▲ 20.9	▲ 5.0
12.5	8.9	▲ 4.1	▲ 18.3
0.1	12.0	10.8	▲ 2.0
5.7	▲ 0.3	13.4	12.8
0.6	4.4	1.9	15.0
2.6	▲ 0.2	5.6	3.9
11.4	2.2	1.7	8.6
3.3	11.0	4.9	4.8
2.9	3.9	14.0	8.5
6.2	3.3	6.1	20.0
23.4	7.8	12.2	14.1
▲ 100.0	－	－	－

表２－３　年齢階級別の人口変化率への寄与度

55－60	60－65	65－70	70－75
▲ 4.3	▲ 7.6	▲ 5.8	▲ 0.6
▲ 2.5	▲ 1.5	▲ 0.9	0.5
▲ 2.8	▲ 2.7	▲ 1.3	▲ 0.6
1.7	▲ 3.1	▲ 2.7	▲ 1.2
▲ 1.1	1.2	▲ 2.0	▲ 1.2
▲ 2.0	▲ 1.3	0.7	▲ 1.1
▲ 0.3	▲ 1.9	▲ 0.6	1.5
0.7	▲ 0.5	▲ 1.6	▲ 0.3
0.7	0.6	▲ 0.3	▲ 1.4
0.0	0.7	0.7	▲ 0.2
0.3	▲ 0.0	0.8	0.9
0.0	0.2	0.1	0.9
0.1	▲ 0.0	0.3	0.2
0.4	0.1	0.1	0.4
0.1	0.3	0.2	0.2
0.1	0.1	0.4	0.3
0.1	0.1	0.1	0.4
0.2	0.1	0.2	0.3
▲ 0.0	－	－	－

表２－２と表２－３は、年齢階級ごとの人口の変化率とその**寄与度**を示しており、そこから、人口減少率がどの年齢階級で大きくて、全人口の変化への影響度合が大きいかが分かる。

寄与度とは、全体の変化に対して、内訳部分の変化がどの程度貢献したかを示す指標。ここでは、０～４歳から年齢不詳までの寄与度を足すと、全体の変化率となるように示していることを知っておくこと。

Q３：表２からどのようなことが分かるだろうか。

変化率や寄与度を見ると、島根県の人口減少には、25歳以下の人口減少が大きく影響している。この年齢層の死亡者が多くなるとは考えにくいから、減少の主な要因は出生数の減少と県外への流出である。

国勢調査は５年ごとに調査が実施されるので、生まれ年が同じ人の集まりを５年おきに追跡して推移をたどることができる。たとえば、1955年調査で５～９歳だった人は、次の1960年調査では10～14歳の年齢階級に移る。このように、出生の時期を同じくする人口の集まりを**コーホート**といい、コーホートに着目して時系列的に追跡して分析する方法を**コーホート分析**という。

コーホートとは、出生の時期を同じくする人の集まりのこと。コーホートに着目すると、年齢や時代の影響以外の影響を捉えることができるんだ。

コーホートの中でも、1946～50年生まれのコーホートは、1947年のベビーブーム世代を含んでいるため人数が多い。こうした点は、特に長期間にわたって年齢別の人口を比較するときには注意が必要である。表３は生まれ年のコーホートに着目して、各コーホートの人口の５年ごとの推移を各行で示したものである。

表３　1955～75年の島根県の人口コーホートによる表示

表３－１　人口の総数

	1955	1960	1965	1970	1975年	
生年	929,066	888,886	821,620	773,575	768,886	1975年年齢
1971～75					56,482	０～４歳
1966～70				52,296	53,313	５～９歳
1961～65			59,331	58,285	58,277	10～14歳
1956～60		72,500	69,219	67,876	53,285	15～19歳
1951～55	95,321	93,089	89,853	62,425	42,692	20～24歳
1946～50	119,435	117,007	78,566	51,575	56,957	25～29歳
1941～45	101,415	67,581	45,483	45,640	47,266	30～34歳
1936～40	77,344	56,643	50,893	49,761	49,946	35～39歳
1931～35	75,171	68,036	62,934	61,111	60,487	40～44歳
1926～30	71,152	67,490	63,692	61,693	60,291	45～49歳
1921～25	61,115	58,492	55,702	53,451	51,718	50～54歳
1916～20	51,992	49,741	47,133	44,959	43,043	55～59歳
1911～15	49,687	47,261	44,142	41,425	39,256	60～64歳
1906～10	44,725	42,272	39,213	36,133	33,120	65～69歳
1901～05	42,018	39,304	35,528	31,598	27,176	70～74歳
1896～1900	38,323	34,753	30,124	25,041	19,337	75～79歳
1891～95	31,203	27,132	21,975	16,108	16,198	80歳以上
1886～90	26,254	21,147	15,177	14,198		
1881～85	20,551	14,698	12,655			
1876～80	13,837	11,740				
1871～75	9,514					
	9	0	0	0	42	年齢不詳

表３－２　人口の変化率

55－60	60－65	65－70	70－75
▲ 4.3	▲ 7.6	▲ 5.8	▲ 0.6
			1.9
		▲ 1.8	▲ 0.0
	▲ 4.5	▲ 1.9	▲ 21.5
▲ 2.3	▲ 3.5	▲ 30.5	▲ 31.6
▲ 2.0	▲ 32.9	▲ 34.4	10.4
▲ 33.4	▲ 32.7	0.3	3.6
▲ 26.8	▲ 10.2	▲ 2.2	0.4
▲ 9.5	▲ 7.5	▲ 2.9	▲ 1.0
▲ 5.1	▲ 5.6	▲ 3.1	▲ 2.3
▲ 4.3	▲ 4.8	▲ 4.0	▲ 3.2
▲ 4.3	▲ 5.2	▲ 4.6	▲ 4.3
▲ 4.9	▲ 6.6	▲ 6.2	▲ 5.2
▲ 5.5	▲ 7.2	▲ 7.9	▲ 8.3
▲ 6.5	▲ 9.6	▲ 11.1	▲ 14.0
▲ 9.3	▲ 13.3	▲ 16.9	▲ 22.8
▲ 13.0	▲ 19.0	▲ 26.7	0.6
▲ 19.5	▲ 28.2	▲ 6.5	▲ 100.0
▲ 28.5	▲ 13.9	▲ 100.0	
▲ 15.2	▲ 100.0		
▲ 100.0			
▲ 100.0	－	－	－

表３－３　人口変化率への寄与度

55－60	60－65	65－70	70－75
▲ 4.3	▲ 7.6	▲ 5.8	▲ 0.6
			7.3
		6.4	0.1
	6.7	▲ 0.1	▲ 0.0
7.8	▲ 0.4	▲ 0.2	▲ 1.9
▲ 0.2	▲ 0.4	▲ 3.3	▲ 2.6
▲ 0.3	▲ 4.3	▲ 3.3	0.7
▲ 3.6	▲ 2.5	0.0	0.2
▲ 2.2	▲ 0.6	▲ 0.1	0.0
▲ 0.8	▲ 0.6	▲ 0.2	▲ 0.1
▲ 0.4	▲ 0.4	▲ 0.2	▲ 0.2
▲ 0.3	▲ 0.3	▲ 0.3	▲ 0.2
▲ 0.2	▲ 0.3	▲ 0.3	▲ 0.2
▲ 0.3	▲ 0.4	▲ 0.3	▲ 0.3
▲ 0.3	▲ 0.3	▲ 0.4	▲ 0.4
▲ 0.3	▲ 0.4	▲ 0.5	▲ 0.6
▲ 0.4	▲ 0.5	▲ 0.6	▲ 0.7
▲ 0.4	▲ 0.6	▲ 0.7	0.0
▲ 0.5	▲ 0.7	▲ 0.1	▲ 1.8
▲ 0.6	▲ 0.2	▲ 1.5	
▲ 0.2	▲ 1.3		
▲ 1.0			
▲ 0.0	－	－	－

先の表２から、1955～75年にかけて、15～24歳の人口減少が大きいことをみた。進学・就職が原因と考えられる。他方、表３によれば1941～45年生まれのコーホートは、1965～75年に人口が増加しており、1946～50年生まれのコーホートは1970～75年に人口が増加している。いずれのコーホートも、1970～75年の増加は少なくない増加率であり、その背景を探ることが、人口減少への対策のヒントになるかもしれない。

ちなみに、島根県の人口は、その後一時回復するが、再び減少に転じ、2015年の国勢調査では1920年の第１回国勢調査よりも人口が少なくなった。

Q４：自分が住む都道府県・市町村の人口を、コーホートごとに見たときに、人口が急に増えたり減ったりする世代はあるだろうか？

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ 人口減少の主たる要因は？

日本全体の人口が増加しているなかで、島根県の人口が1960～70年に大きく減少しているのは、15～24歳代の若年層が大幅に減少したことによる。これらの年齢階級の人口減少が島根県の人口減少に大きく影響したのみならず、出生数の大幅減少にもつながった。当時は20歳代前半が婚姻年齢であり、1955～70年に15～24歳代の年齢層が20～30％減少したことによって、1960～70年の出生数が激減した。これが島根県の1960～70年の人口大減少の最大の要因である。若年層の人口減少は進学や就職に伴う県外への流出による。学ぶ場所・働く場所が島根県に少なかったことがその根幹にあると推察される。

ここでの人口減少の確認と原因究明のための分析は、行政の政策の企画・立案にとっても重要な作業である。人口に限らず、統計の分析は、その結果が実際に使われることで、はじめて活きてくる。とくに、地域の課題を解決するためには、データを見るだけでも、現場を見るだけでも不十分である。適切な分析を行い、その分析結果が地域住民にも理解されるように、分析結果を分かりやすく広範囲に周知することが必要である。

実際の課題解決の事例

島根県隠岐郡海士町は、島根半島の沖合60km の日本海に浮かぶ離島の町です。国勢調査によると、人口は最盛期だった1950年の6,986人から、2010年には2,374人にまで減少し、このままでは町から人がいなくなってしまうと危惧されました。

そこで、地域データを活用して、現状の分析や課題の整理を行い、住民とともに、地域を守る施策の原案を作成しました。行政もこれをもとに施策を行うようになりました（注１）。たとえば、「ひと」を課題として、若者のＵターン・Ｉターン者を役場などに登用したりするなど、町内の学校教育の充実や、地域の内外との交流、人材育成などを推し進めました。また、「産業」を課題として、海産物加工の施設を整備したり、観光に力を入れたりするなど、地域資源を有効に活用するための基盤整備や人材育成などに力を入れました。いずれも、学校に通う若い世代が少なく高齢者が多い現状や、第１次産業が著しく減少してきた過去からの推移を示す統計データを踏まえての取り組みです。

こうした取り組みの結果、海士町では流入人口が増え、2015年の国勢調査は2,353人と、国立社会保障・人口問題研究所の人口推計（注２）を上回り、2010年からの減少率も緩やかになりました。

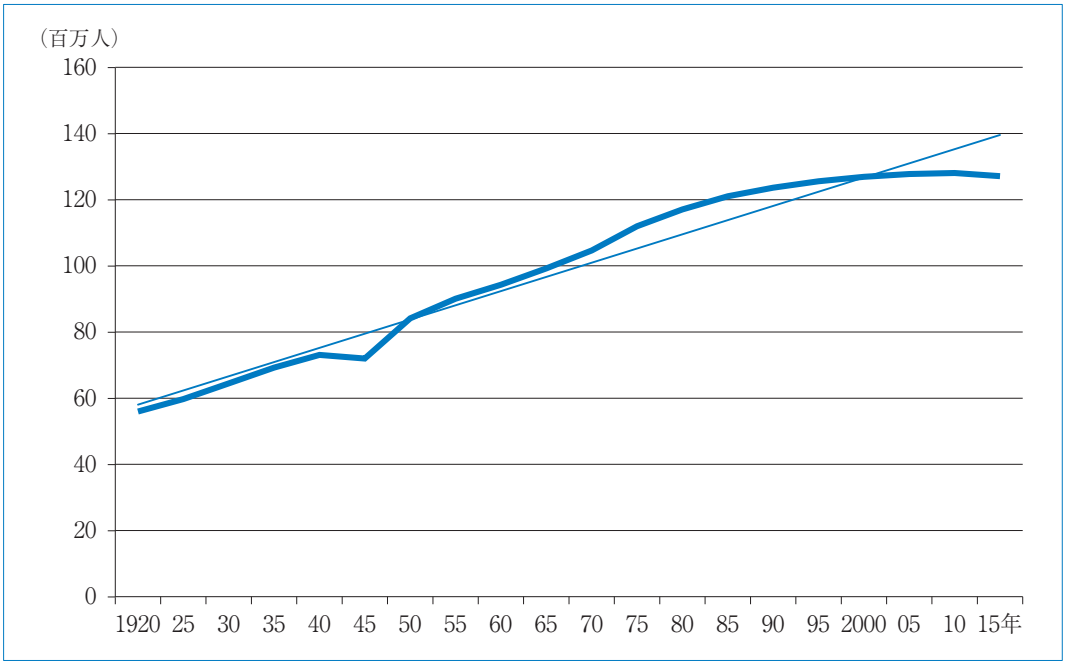
とはいえ、人口の減少が止まったわけではありません。今後も、統計データを確認しながら活動を継続することが必要です。実際、海士町でも、新たな国勢調査の結果を踏まえて、次の取り組みに向かっています。

注１：海士町ホームページ「第四次総合振興計画」
（http://www.town.ama.shimane.jp/gyosei）

注２：国立社会保障・人口問題研究所「日本の将来推計人口（2012年１月推計）」
（http://www.ipss.go.jp/syoushika/tohkei/newest04/hh2401.asp）

〔本節の解答〕

Ｑ１：x を第１回国勢調査が実施された1920年から５年おきの西暦、y をその年の人口（百万人）として、回帰直線 $y=a+bx$ を求めると、$y=-1590+0.858x$ となる。この回帰式をグラフに書き入れると次のとおりとなる。



Ｑ２：略

Ｑ３：・25歳以下の人口減少が顕著である。
・1955～60年の10～14歳、1960～65年の15歳～19歳、1965～70年の20歳～24歳は、全体が減少しているにもかかわらず、大きな増加をしている。いずれも生まれ年は1945～50年で、その前の世代が戦時中で出生が少なく、かつ死亡も多かった世代であるのに対して、戦後生まれの団塊の世代は出生者が激増したことによる。異なる時点で同じ年齢階級の人口を比較することには十分な注意が必要であることを示す結果である。

Ｑ４：略

第7部

統計的思考によって大学入試・統計検定を乗り切ろう！

～大学入試・統計検定問題からみる統計的思考力～

2020年度（平成32年度）から、現在の「大学入試センター試験」に代わって「大学入学希望者学力評価テスト（仮称）」が実施される。そこでは、一体どのような問題が出題されるのだろうか？

2016年３月に公表された文部科学省の報告書では、例えば、国語のイメージ問題例として下記の問題が取り上げられている。

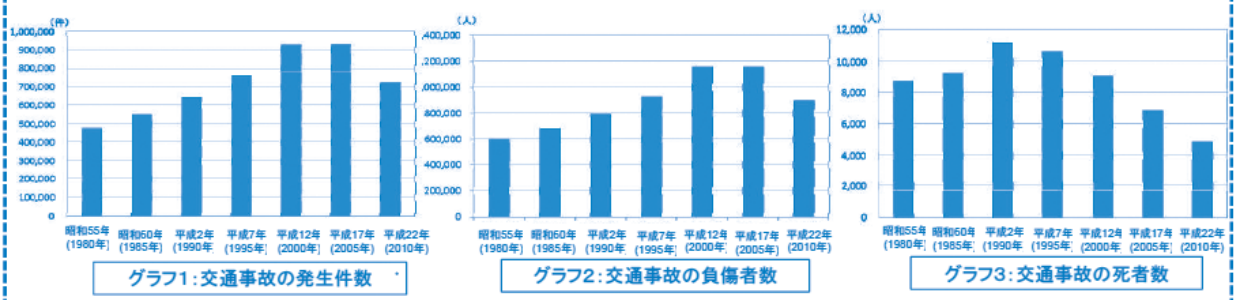
図１　高大接続システム改革会議「最終報告」参考資料イメージ問題例＜国語＞の抜粋

問題イメージ＜例1＞

国立教育政策研究所「特定の課題に関する調査（論理的な思考）」（平成24年2月実施）より一部改題

次の文章とグラフを読み，後の問いに答えよ。

次に示すのは，警察庁事故統計資料に基づいて作成された交通事故の発生件数，負傷者数，死者数のグラフと，この3つのグラフを見て，交通事故の死者数が他よりも早く，平成2年（1990年）以降減少傾向になっていることについて，4人の高校生が行った話し合いの一部である。



グラフ1：交通事故の発生件数　グラフ2：交通事故の負傷者数　グラフ3：交通事故の死者数

Aさん：交通事故の死者数が他よりも早く，平成2年（1990年）以降減少傾向になっているのは，交通安全に関する国民の意識の変化が関係しているのではないかと思います。
その裏付けとなる 資料として，「交通違反で検挙された人数の推移が分かる資料」があると思います。その資料を見れば，飲酒運転やスピード違反など，死亡事故につながるような重大な違反の割合が少なくなっていることが分かるはずです。

Bさん：私は，この30年間で販売されてきた自動車の台数と安全性に関係があると思います。
(a)つまり，自動車の台数は年々増加し続けているので事故件数と負傷者数はなかなか減らなかったけれども，［ア］ということです。
例えば，最近30年間における，「車の総販売台数の推移が分かる資料」と，「車の安全に関する装置の装備率の推移が分かる資料」があれば，このことを裏付けることができると思います。

136

資料：高大接続システム改革会議「最終報告」

国語の問題なのに、統計のグラフが出ているのが意外だね！

この問題では、統計資料から自動車事故の現状を把握し、アの文章を構成することも求められているんだ！

この問題では、棒グラフと会話文から情報を解釈し、問題場面に適した考えを構成し、表現することが求められている。新テストでは、教科の枠を超えて統計を活用する能力が求められており、これから統計的思考力を養うことはますます重要になってくるといえる。

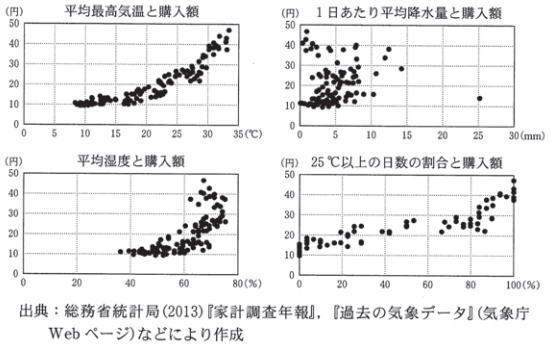
以降では、大学入試や統計検定の過去問題を参考に、現在、どのような統計的思考力が求められているかについて考えていこう。

1　「数学Ⅰ」：データの分析の問題から見る統計的思考力

現行の「数学Ⅰ」データの分析の単元では、箱ひげ図や散布図、相関係数といった記述統計に関わる事柄を学ぶ。2016年度の大学入試センター試験では、アイスクリームの購入金額と気象データとの散布図から、「アイスクリームが売れる条件」を分析する問題が出題された。

〔例題１〕2016年度　大学入試センター試験「数学Ⅰ・A」第２問

〔２〕 次の４つの散布図は，2003 年から 2012 年までの 120 か月の東京の月別データをまとめたものである。それぞれ，１日の最高気温の月平均(以下，平均最高気温)，１日あたり平均降水量，平均湿度，最高気温 25 ℃ 以上の日数の割合を横軸にとり，各世帯の１日あたりアイスクリーム平均購入額(以下，購入額)を縦軸としてある。



出典：総務省統計局(2013)『家計調査年報』，『過去の気象データ』(気象庁 Web ページ)などにより作成

次の［ス］，［セ］に当てはまるものを，下の⓪～④のうちから一つずつ選べ。ただし，解答の順序は問わない。

これらの散布図から読み取れることとして正しいものは，［ス］と［セ］である。

⓪ 平均最高気温が高くなるほど購入額は増加する傾向がある。
① １日あたり平均降水量が多くなるほど購入額は増加する傾向がある。
② 平均湿度が高くなるほど購入額の散らばりは小さくなる傾向がある。
③ 25 ℃ 以上の日数の割合が 80 % 未満の月は，購入額が 30 円を超えていない。
④ この中で正の相関があるのは，平均湿度と購入額の間のみである。

〔解説〕

⓪…「平均最高気温と購入額」の散布図では、正の相関を持つので ○

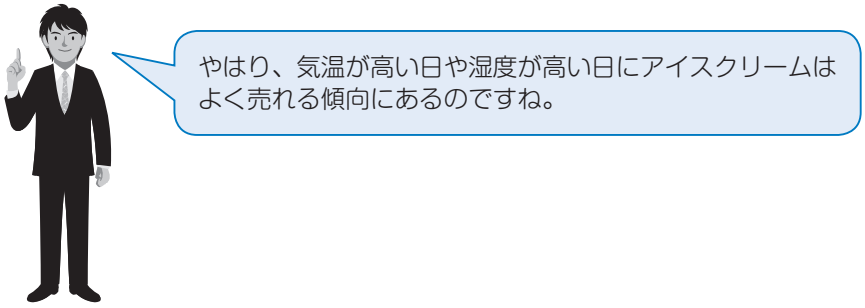
①…「１日あたり平均降水量と購入額」の散布図では、負の相関を持つので ×

②…「平均湿度と購入額」の散布図では、平均湿度が高くなるほど購入額の散らばりは大きくなるので ×

③…「25℃以上の日数の割合と購入額の散布図」では、横軸が80％未満の範囲で、縦軸の購入金額が30円を超えていないので ○

④…「平均湿度と購入額」のほか「平均最高気温と購入額」にも正の相関があるので ×

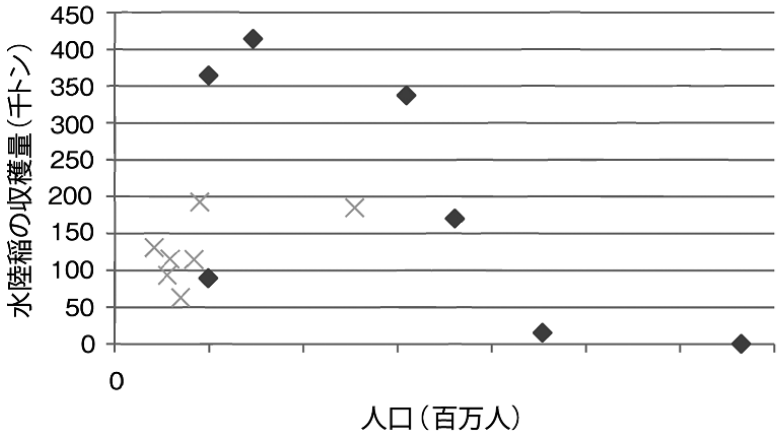
答：⓪と③



〔練習 1〕社会データの分析

（1） 2014年11月実施　統計検定 3 級問題「水陸稲の収穫量と人口の相関」

〔2〕 各都道府県の人口と水陸稲の収穫量の相関係数を関東地区の 7 都県（群馬，栃木，茨城，埼玉，千葉，東京，神奈川）と九州地区の 7 県（福岡，佐賀，長崎，熊本，大分，宮崎，鹿児島）について調べたところ，散布図は次のようになり，それぞれの相関係数は関東地区で −0.67，九州地区で 0.60 であった。



資料：総務省「人口推計」および農林水産省「作物統計」

これについて，次の I〜III の記述を考えた。

I. 人口と水陸稲の収穫量について，関東地区には負の相関がある。

II. 人口と水陸稲の収穫量について，九州地区には正の相関がある。

III. 関東地区と九州地区を合わせた 14 都県での相関係数は $(-0.67 + 0.60) \div 2 = -0.035$ と計算できる。

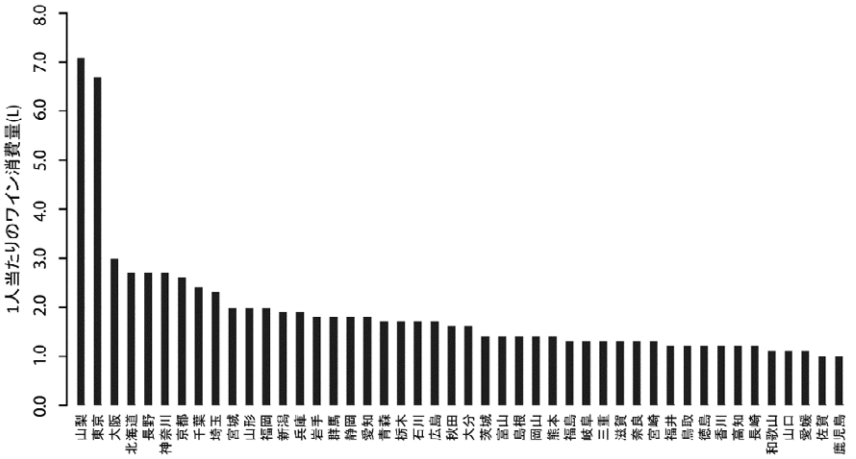
この記述 I〜III に関して，次の ① 〜 ⑤ のうちから最も適切なものを一つ選べ。

16

① I のみ正しい。
② II のみ正しい。
③ I と II のみ正しい。
④ II と III のみ正しい。
⑤ I と II と III はすべて正しい。

（２） 2014年６月実施　統計検定３級問題「ワインの消費量を表す箱ひげ図の選択」

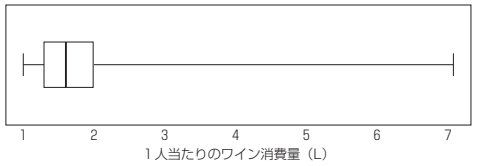
問17　次の棒グラフは、平成22年度の沖縄県を除く46都道府県の成人１人当たりのワイン消費量（L）を表している。



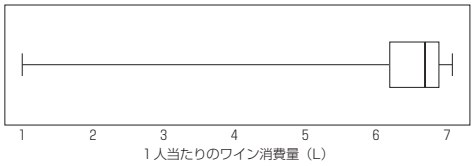
資料：国税庁課税部酒税課「平成22年度酒類販売（消費）数量表」

この結果の中で，7.1Lの山梨県が最も多く，次いで東京都の6.7Lであり，最も少ないのは鹿児島県，佐賀県で消費量は1.0Lであった。この棒グラフから作成した46の都道府県のワイン消費量の箱ひげ図として①～⑤のうちから最も適切な者を一つ選べ。 **24**

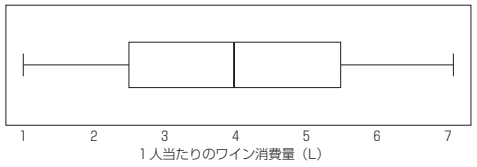
①



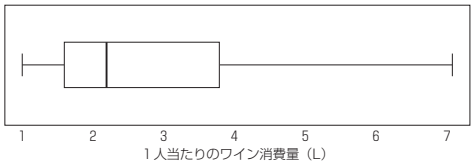
②



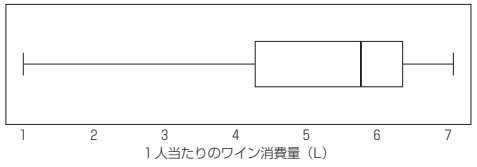
③



④



⑤



2 「数学 B」：確率分布と統計的な推測の問題から見る統計的思考力

現行の「数学 B」確率分布と統計的な推測の単元では、二項分布や正規分布、区間推定といった推測統計に関わる事柄を学ぶ。2015年度の大学入試センター試験では、標準正規分布の性質を利用した区間推定の公式そのものの意味理解を問う問題が出題された。

〔例題 2 〕2015年度　大学入試センター試験「数学Ⅱ・B」第 5 問

第 5 問　（選択問題）（配点　20）

以下の問題を解答するにあたっては，必要に応じて 29 ページの正規分布表を用いてもよい。

また，小数の形で解答する場合，指定された桁（けた）数の一つ下の桁を四捨五入し，解答せよ。途中で割り切れた場合，指定された桁まで⓪にマークすること。

(2)　確率変数 Z が標準正規分布に従うとき

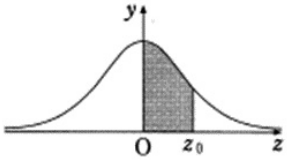
$$P\left(-\boxed{\text{タ}} \leqq Z \leqq \boxed{\text{タ}}\right) = 0.99$$

が成り立つ。 タ に当てはまる最も適切なものを，次の⓪～③のうちから一つ選べ。

⓪ 1.64　　① 1.96　　② 2.33　　③ 2.58

正　規　分　布　表

次の表は，標準正規分布の分布曲線における右図の灰色部分の面積の値をまとめたものである。



z_0	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

(3)　母標準偏差 σ の母集団から，大きさ n の無作為標本を抽出する。ただし，n は十分に大きいとする。この標本から得られる母平均 m の信頼度(信頼係数) 95 % の信頼区間を $A \leqq m \leqq B$ とし，この信頼区間の幅 L_1 を $L_1 = B - A$ で定める。

この標本から得られる信頼度 99 % の信頼区間を $C \leqq m \leqq D$ とし，この信頼区間の幅 L_2 を $L_2 = D - C$ で定めると

$$\frac{L_2}{L_1} = \boxed{\text{チ}}.\boxed{\text{ツ}}$$

が成り立つ。また，同じ母集団から，大きさ $4n$ の無作為標本を抽出して得られる母平均 m の信頼度 95 % の信頼区間を $E \leqq m \leqq F$ とし，この信頼区間の幅 L_3 を $L_3 = F - E$ で定める。このとき

$$\frac{L_3}{L_1} = \boxed{\text{テ}}.\boxed{\text{ト}}$$

が成り立つ。

〔解説〕

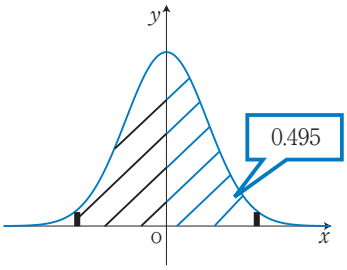
(2) $P(0 \leqq Z \leqq z_0) = 0.99 \div 2 = 0.495$

したがって、正規分布表より右の図の状況を満たす

z_0は2.57と2.58の間であるから

$$z_0 = \frac{(2.57 + 2.58)}{2} = 2.575 \fallingdotseq 2.58$$

答：③



(3) (2)と同様の考え方を用いて、$P(-1.96 \leqq Z \leqq 1.96) = 0.95$

母平均 m、母標準偏差 σ、大きさ n の標本の標本平均を $\overline{X}$ とすると、$\overline{X}$ は平均 m、分散 σ^2/n の標準正規分布に従うので $P\left(-1.96 \leqq \dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leqq 1.96\right) = 0.95$ と表せる。

母平均 m の信頼度95％信頼区間は、上のかっこ内の式を変形して、

$$\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

したがって、$L_1 = \overline{X} + 1.96\dfrac{\sigma}{\sqrt{n}} - \left(\overline{X} - 1.96\dfrac{\sigma}{\sqrt{n}}\right) = 2 \times 1.96\dfrac{\sigma}{\sqrt{n}}$ …①

同様に、$P(-2.58 \leqq Z \leqq 2.58) = 0.99$であるから、母平均 m の信頼度99％信頼区間は

$$\overline{X} - 2.58\frac{\sigma}{\sqrt{n}} \leqq m \leqq \overline{X} + 2.58\frac{\sigma}{\sqrt{n}}$$

したがって、$L_2 = \overline{X} + 2.58\dfrac{\sigma}{\sqrt{n}} - \left(\overline{X} - 2.58\dfrac{\sigma}{\sqrt{n}}\right) = 2 \times 2.58\dfrac{\sigma}{\sqrt{n}}$…②

①、②より

$$\frac{L_2}{L_1} = \frac{2.58}{1.96} = 1.316... \fallingdotseq 1.3$$

答：$\dfrac{L_2}{L_1} \fallingdotseq 1.3$

また、大きさ 4 n の標本の標本平均を $\overline{X}'$ とすると、$\overline{X}'$ は平均 m、分散 $\sigma^2/4n$ の標準正規分布に従うので、母平均 m の信頼度95％信頼区間は

$$\overline{X}' - 1.96\frac{\sigma}{\sqrt{4n}} \leqq m \leqq \overline{X}' + 1.96\frac{\sigma}{\sqrt{4n}}$$

したがって、$L_3 = \overline{X}' + 1.96\dfrac{\sigma}{\sqrt{4n}} - \left(\overline{X}' - 1.96\dfrac{\sigma}{\sqrt{4n}}\right) = 2 \times 1.96\dfrac{\sigma}{\sqrt{4n}} = 1.96 \times \dfrac{\sigma}{\sqrt{n}}$…③

①、③より

$$\frac{L_3}{L_1} = \frac{1}{2} = 0.5$$

答：$\dfrac{L_3}{L_1} = 0.5$

正規分布の曲線によって囲まれた面積は１で、「平均付近は起きやすいが、裾の方は起きにくい」といった単純なモデルから区間推定の公式が導き出せるんだね！

〔練習 2 〕正規分布の利用

2014年度　鹿児島大学 2 次試験の問題「合格最低点の推定」

ある企業の入社試験は採用枠300名のところ500名の応募があった。試験の結果は500点満点の試験に対し、平均点245点、標準偏差50点であった。得点の分布が正規分布であるとみなされるとき、合格最低点はおよそ何点であるか。小数点以下を切り上げて答えよ。ただし、確率変数 Z が標準正規分布に従うとき、$P(Z>0.25)=0.4$、$P(Z>0.5)=0.3$、$P(Z>0.54)=0.2$とする。

〔鹿児島大学　理・工・医（医）・歯〕

3　「大学」：統計学の問題から見る統計的思考力

大学の統計学の授業では、正規分布以外の確率分布や検定といった、さらに専門的な推測統計学に関わる内容を学ぶ。高等学校では、「正規分布の裾の部分が起きにくい」といった程度しか言及しないが、実は、正規分布の裾の見方は、検定で重要となる “棄却域” の考え方につながる。たとえば、2014年 6 月に実施された統計検定 2 級では、私たちも馴染み深い「二項分布」を題材に検定の考え方を問う問題が出題された。

〔例題 3 〕2014年 6 月実施　統計検定 2 級問題「公正なサイコロか？」

問 13　サイコロを 7 回投げたときに 3 の目が 4 回出たことから，公正なサイコロ（どの目も出る確率は 6 分の 1）であるかどうか疑問となった。

〔1〕サイコロが公正であるとして，サイコロを 7 回投げたときに 3 の目が 4 回出る確率を求める式として正しいものはどれか。次の ① ～ ⑤ のうちから適切なものを一つ選べ。 20

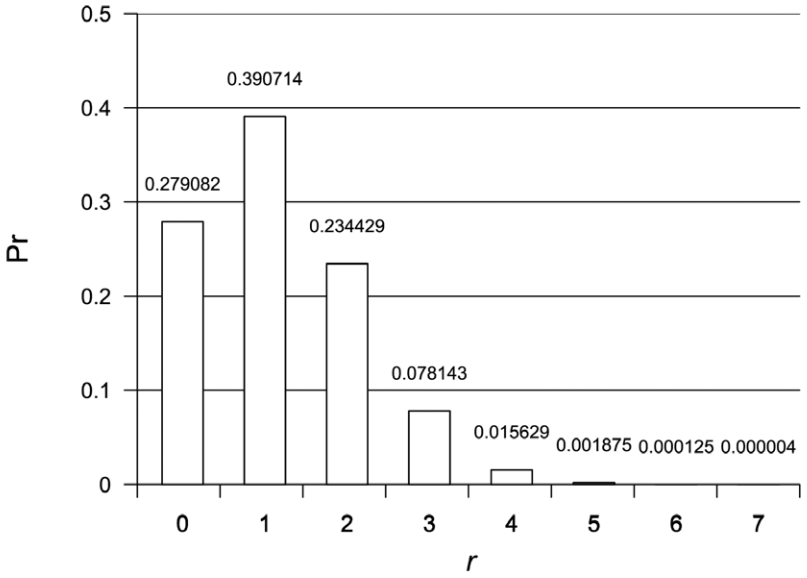
① $840\times\left(\frac{1}{6}\right)^4\times\left(\frac{5}{6}\right)^3$　　② $35\times\left(\frac{1}{6}\right)^4$

③ $\frac{1}{35}\times\left(\frac{1}{6}\right)^4$　　④ $35\times\left(\frac{1}{6}\right)^4\times\left(\frac{5}{6}\right)^3$

⑤ $\frac{1}{35}\times\left(\frac{1}{6}\right)^4\times\left(\frac{5}{6}\right)^3$

〔2〕 このサイコロが公正なサイコロであるかどうかを調べることにする。そのために，このサイコロが公正なサイコロであった場合に，サイコロを 7 回投げたときに 3 の目が r 回出る確率を求め，判断することにした。

その確率 Pr を r = 0, 1, … , 7 の場合についてグラフにすると，次の図のようになる。



r = 0, 1, … , 7 のときの確率

帰無仮説として「このサイコロの 3 の目の出る確率は 6 分の 1 である」を用い，有意水準 5% で仮説検定を行ったとき，結論として正しい判断をしているものはどれか。次の ① ～ ⑤ のうちから適切なものを一つ選べ。 21

① 3 の目が 4 回以上出る確率は 5% より小さい。したがって，帰無仮説を棄却して，公正なサイコロでないと結論する。

② 3 の目がちょうど 4 回出る確率は 0.015629 となり，5% より小さい。したがって，帰無仮説を棄却して，公正なサイコロでないと結論する。

③ 3 の目が何回出るかは事前にわからず，今回はたまたま 4 回出ただけである。したがって，公正なサイコロと結論する。

④ 3 の目が出る回数が 4 回未満である場合で最も確率が小さいのは 3 回の 0.078143 で，5% より大きい。したがって，公正なサイコロと結論する。

⑤ 3 の目が 4 回出る場合の前後を考慮し，3 回，4 回，5 回出る確率を足すと 0.095647 となり 5% より大きい。したがって，帰無仮説は棄却できず，公正なサイコロと結論する。

〔解説〕

(1)　サイコロを７回投げたときに３の目が出た回数を確率変数Ｘとおく。

３の目が４回出る確率は、

$$P(X=4)={}_7C_4\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^3=35\times\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^3$$

答：④

(2)　①…３の目が出る確率は１／６であると仮定し、右片側検定で考えると、

３の目が４回以上出る確率は、

$$P(X\geqq 4)=P(X=4)+P(X=5)+P(X=6)+P(X=7)$$

$$=0.015629+0.001875+0.000125+0.000004=0.017633$$

有意水準５％で、$P(X\geqq 4)<0.05$であるから、サイコロを７回投げて３の目が４回以上出ることはめったにないといえる。したがって、帰無仮説は棄却できるので ○

②…４回ちょうどの場合を考えても、片側検定・両側検定の考え方に適さないので ×

③…仮説検定の枠組みになっていないので ×

④…３の目が４回も多く出たことが疑問であるので、４回未満の左片側検定をそもそも考える必要がないので ×

⑤…３の目が４回出る前後を考えても、②と同様に仮説検定の考え方に適さないので ×

答：①

二項分布で、右裾の３の目が４回以上出る確率はすごく小さい。だから、それが「有意水準５％未満の確率ならば、めったに起きないことが起きたと判断する」といった、仮説検定ならではの考え方が問われているんだね。

〔練習 3〕仮説検定の考え方

2014年 6 月実施　統計検定 2 級問題「仮説検定における用語の確認」

問 14　次の文章中の（ア）～（オ）にあてはまる用語の正しい組合せとして，下の ① ～ ⑤ のうちから最も適切なものを一つ選べ。 **23**

統計的推測では，母集団からの標本として得られたデータを用いて，測定値を得た対象だけに限定することなく，母集団について何らかの判断を下す方法論を扱っている。統計的推測は大きく 2 つに分けられる。（ア）と（イ）である。（イ）は，はじめに母集団に対して（ウ）と呼ばれる特定の仮説を設定し，観測したデータがこの仮説を否定するかどうかを調べる手法である。このとき，仮説を支持するか否かは，確率を伴う判断が必要となる。（ウ）が正しいにも関わらず，（ウ）を否定してしまう確率をある値以下にする必要がある。この値を（エ）と呼ぶ。（エ）を決めることで，（オ）と呼ばれる判定のための領域を決めることができる。

① （ア）推定，（イ）検定，（ウ）帰無仮説，（エ）第 1 種過誤の確率，（オ）検出力
② （ア）検定，（イ）推定，（ウ）対立仮説，（エ）P-値，（オ）有意水準
③ （ア）推定，（イ）検定，（ウ）対立仮説，（エ）有意水準，（オ）棄却域
④ （ア）検定，（イ）推定，（ウ）帰無仮説，（エ）P-値，（オ）有意水準
⑤ （ア）推定，（イ）検定，（ウ）帰無仮説，（エ）有意水準，（オ）棄却域

大学入試や統計検定では、実際に分析をするために必要な個別の知識・技能だけでなく、問題を解決するために必要な思考力や判断力も問われている。だからこそ、算数・数学科に限らず、教科の枠を超えた日頃の授業や探究活動の中で、統計的思考力を育む機会を大切にしたいものだ！

〔本節の解答〕

〔練習 1〕 (1)　③　(2)　①

〔練習 2〕 試験の得点を X 点とし、500名中300名が合格することから、

$$P\left(\frac{X-245}{50} \geqq z_0\right) = \frac{300}{500} (=0.6)$$

$P\left(\frac{X-245}{50} \geqq z_0\right) = 1 - P\left(\frac{X-245}{50} \leqq z_0\right)$であるので、$P\left(\frac{X-245}{50} \leqq z_0\right) = P(Z \leqq z_0) = 0.4$ を満たす X は不合格となる最高点を与える。

正規分布は左右対称であるから、$P(Z \leqq z_0) = P(Z \geqq -z_0) = 0.4$

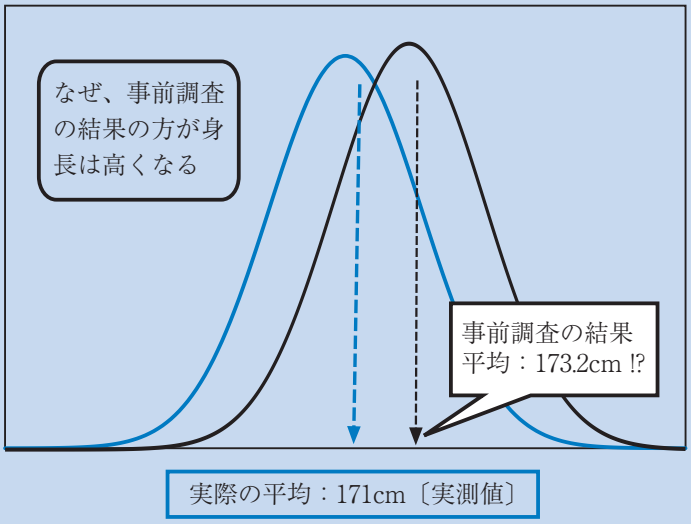
$-z_0 = 0.25$より、$X \leqq 245 - 50 \times 0.25$であるので、232.5点以下の得点は不合格となる。したがって、合格最低点は約233点。

〔練習 3〕 ⑤

正規分布の歪みから見る若者たちの身長の真実

文部科学省の「平成27年度学校保健調査」によると、日本の17歳の男子生徒の平均身長は、170.7cm です。ある地域では、どの年も、17歳の男子生徒の身長の測定結果は、全国結果とほぼ変わらない傾向であることが知られていました。しかし、事前にアンケート調査を実施すると、その結果（平均身長173.2cm）は、実際の測定結果（平均身長171cm）より高い数値を示しました。このことから、正規分布を用いることで、「生徒たちは、事前調査では、実際の身長より高めの身長を回答した」と推測できます。

ある地域の17歳の男子生徒の身長の分布（cm）



一方、この話題とは逆に、統計学者のアドルフ・ケトレー（Adolphe Quételet）は、1844年に男性の身長が正規分布に従うことを利用して、フランス軍の徴兵検査の際に測定された若者たちの身長について、次のウソを見抜きました。

当時の記録では、157cm よりやや背が高い者が少なく、逆に157cm よりやや背が低い者の数が極端に多かったようです。そのため、身長の分布が一部凹んだ正規分布になりました。この結果を受け、ケトレーは、当時のフランス軍は身長157cm 以上の若者を徴兵していたことを踏まえ、「157cm よりわずかに身長が高い若者たちの何人かが、身長を低くごまかし、徴兵から逃れた」と推測しました。

したがって、この正規分布の歪みは、「徴兵を逃れたい、157cm よりやや高い若者たちのごまかしによって、現れた歪みである」と結論付けることができます。

フランスの徴兵検査の際には、モデルとして正規分布を利用したが、グラフの一部が歪んだことで、ケトレーにより記録の偽りが見抜かれました。このように、現実の事象を分析する際には、状況に合わせた確率分布をモデルとして用いることで、さまざまな統計分析を行うことができます。

第8部 公的統計を通してセカイを見る！

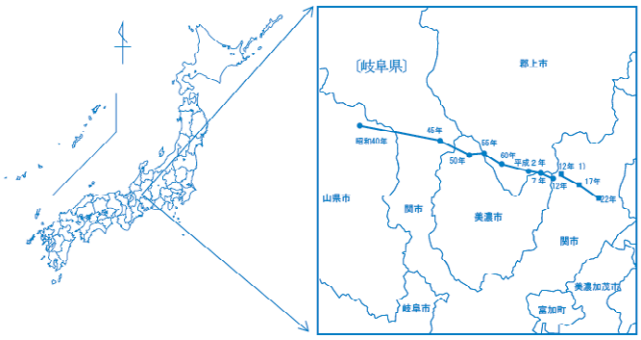
1 日本の国土と気象

日本の“へそ”はどこ

◇ 日本の国土の重心は富山県沖

国土の重心の位置
東経137度42分44秒
北緯 37度30分52秒

◇ 日本の人口の重心は東へ移動中（2010年10月１日現在で岐阜県関市）



人口の重心の位置
東経137度01分45秒
北緯 35度35分35秒

◇ 人口分布の比重は東

（%、km）

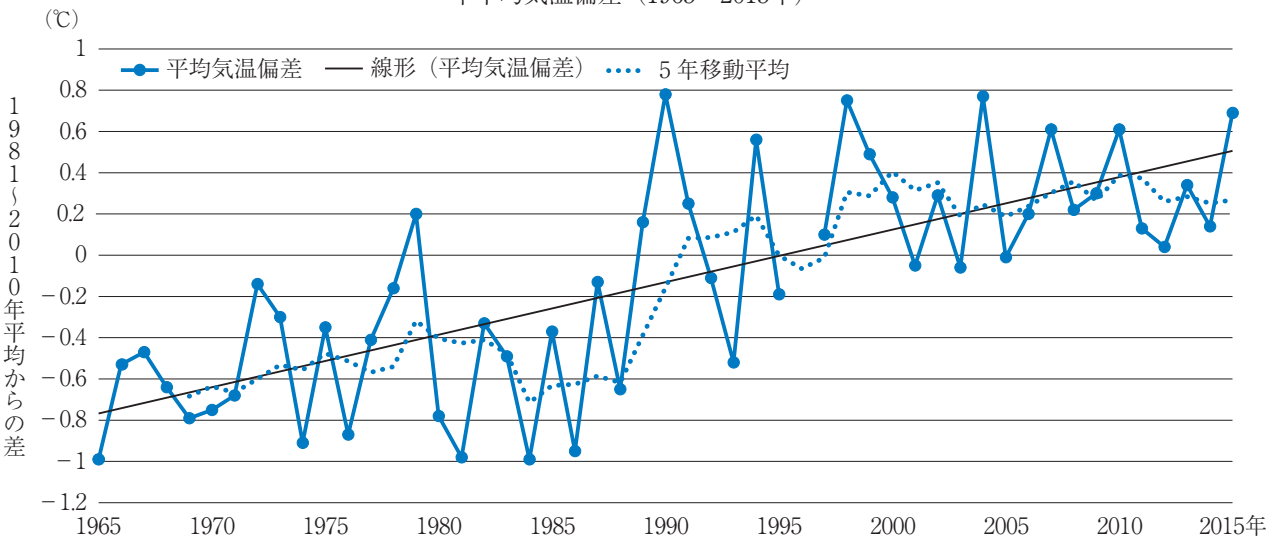
年		1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
人口割合	東北日本	53.39	53.90	54.22	54.80	55.22	55.41	55.56	55.79	55.81	56.03
	西南日本	46.61	46.10	45.78	45.20	44.78	44.59	44.44	44.21	44.19	43.97
人口の重心の移動距離			8.3	3.3	1.5	1.8	2.1	1.3	1.4	2.1	2.4

資料：国土地理院ホームページ、総務省ホームページ「国勢調査」
注：東北日本：北海道、青森県、岩手県、宮城県、秋田県、山形県、福島県、茨城県、栃木県、群馬県、埼玉県、千葉県、東京都、神奈川県、新潟県、富山県、石川県、福井県、山梨県、長野県、岐阜県、静岡県
西南日本：愛知県、三重県、滋賀県、京都府、大阪府、兵庫県、奈良県、和歌山県、鳥取県、島根県、岡山県、広島県、山口県、徳島県、香川県、愛媛県、高知県、福岡県、佐賀県、長崎県、熊本県 、大分県、宮崎県、鹿児島県、沖縄県

日本列島を取り巻く温暖化

◇ 平均気温は長期的に上昇傾向、この50年の気温の上昇は１℃を超える

年平均気温偏差（1965～2015年）



◇ 日本近海の海面の50年前の水温は、100年前に比べて日本海中部と南西部を除き、低くなっている

1915～1965年の50年間の海面水温変化

◇ 最近50年間は、いずれの海面でも水温が上昇している 特に太平洋側の上昇が著しい

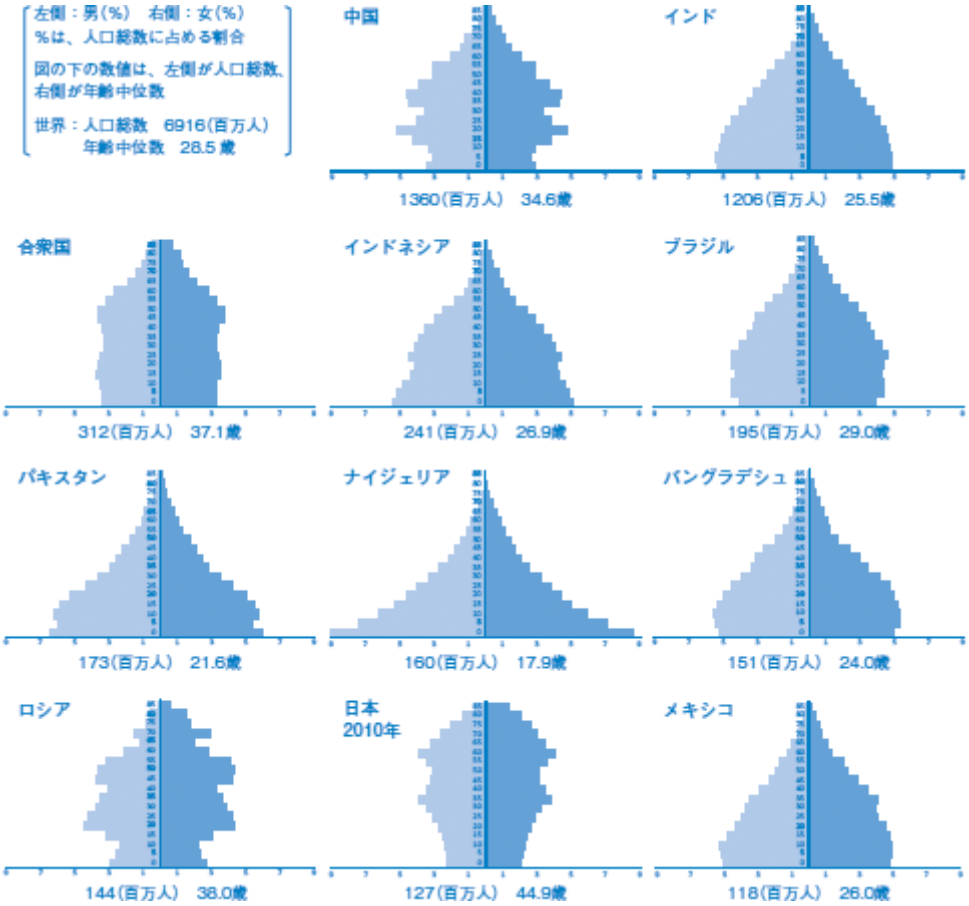
1965～2015年の50年間の海面水温変化

資料：気象庁ホームページ　http://www.jma.go.jp/jma/menu/menureport.html
注：平均気温偏差は、各年の平均気温と1981～2010年の30年間の平均気温の差

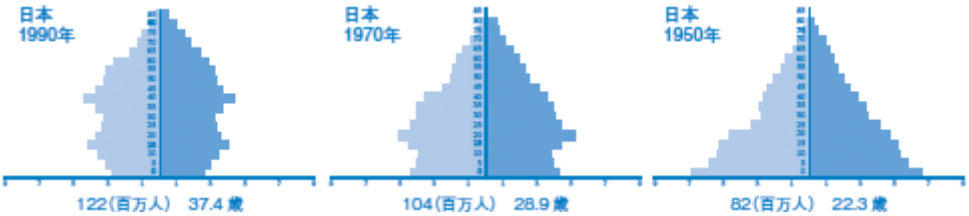
2　世界の人口とこれから

人口１億人以上の国の人口ピラミッド：2010年

◇ 国により著しく異なる男女年齢・構成比



◇ どの国の人口ピラミッドが日本のどの年の人口ピラミッドに似ているか？

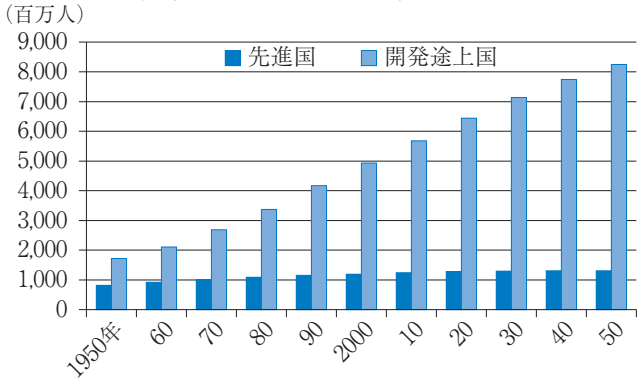


資料：United Nations, World Population Prospects, The 2012 Revision.、総務省「国勢調査」

これからの世界人口

◇ ほぼ横ばいの先進国に対して開発途上国は急増

世界（先進国・開発途上国）人口の推移と展望



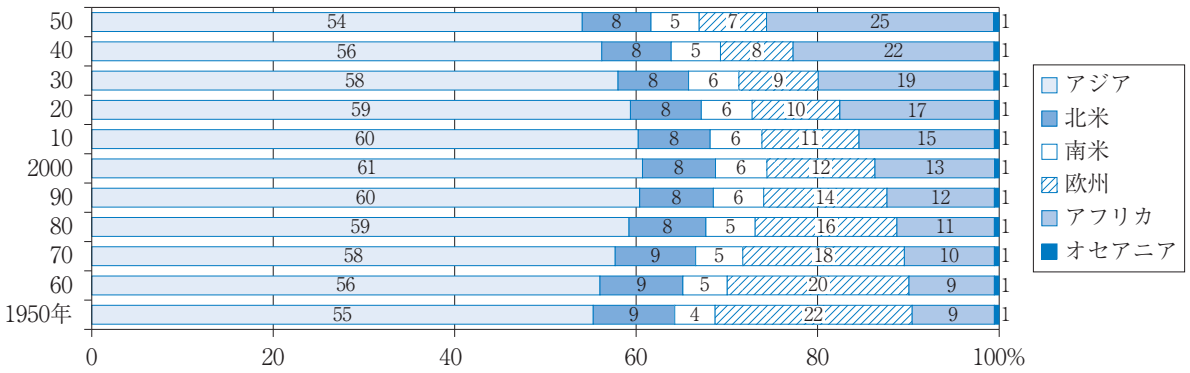
資料：UN, World Population Prospects：The 2012 Revision
注：先進国（日本、米国、カナダ、オーストラリア、ニュージーランド、欧州）、開発途上国（先進国を除く全ての国・地域）以下の図も同じ

（参考）日本の人口の推移

	人口（万人）	構成比（対世界％）	出生率（対人口千人当たり）	死亡率（対人口千人当たり）
50	9708	1.0		
40	10728	1.2		
30	11662	1.4		
20	12410	1.6		
10	12806	1.9	8.5	9.5
2000	12693	2.1	9.5	7.7
90	12361	2.3	10	6.7
80	11706	2.6	13.6	6.2
70	10467	2.8	18.8	6.9
60	9430	3.1	17.2	7.6
1950年	8412	3.3	28.1	10.9

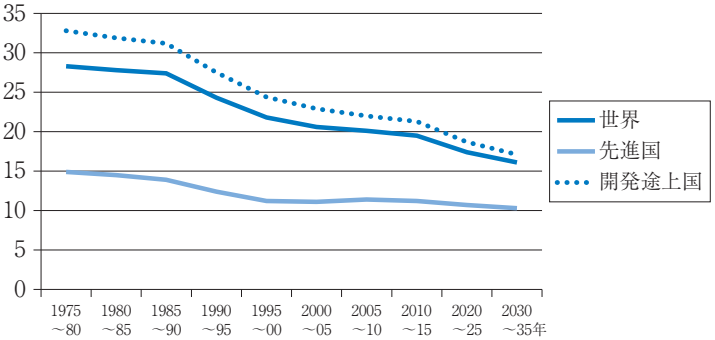
◇ アジアの人口構成比がピークアウトする一方、アフリカは急拡大

地域別人口構成比の推移



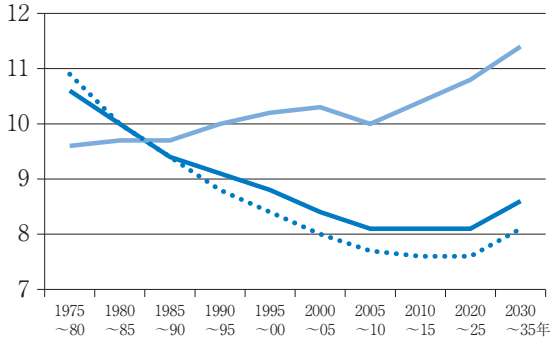
◇ 出生率は開発途上国も低下の一途

出生率（人口1000人当たり）の推移



◇ 死亡率は開発途上国も低下から上昇へ

死亡率（人口1000人当たり）の推移

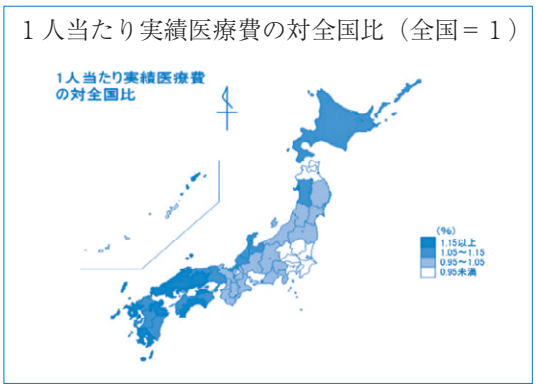


資料：総務省「国勢調査」、国立社会保障・人口問題研究所「将来推計人口」
注：2020年以降の人口は出生中位、死亡中位の仮定による推計値

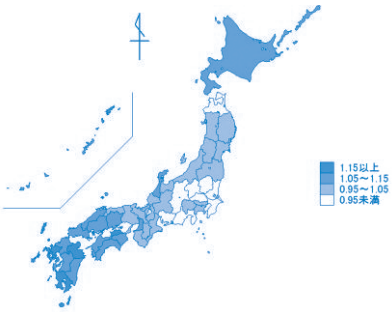
3　統計地図で見える日本の地域特性

日本の医療費問題を考える

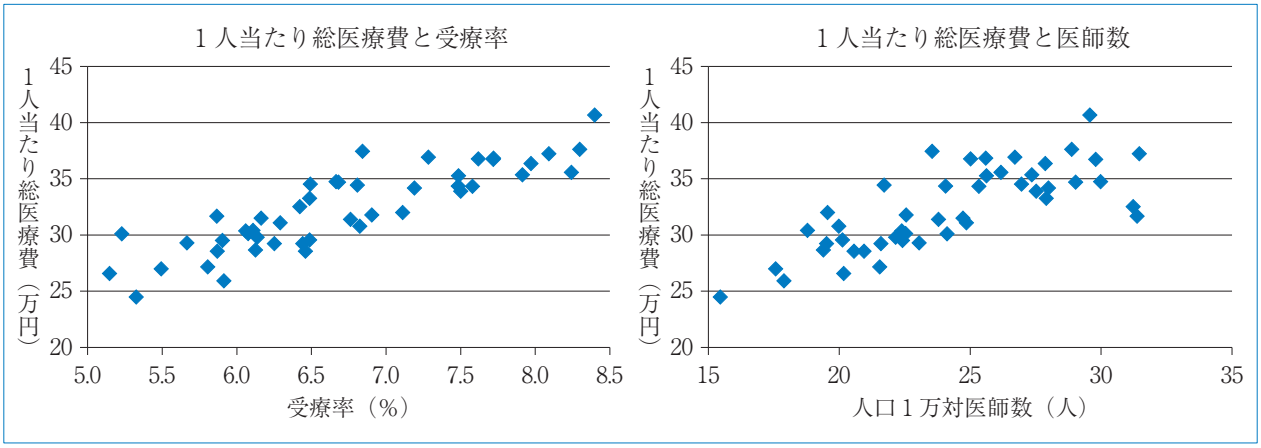
◇ 医療費の地域差は西高東低

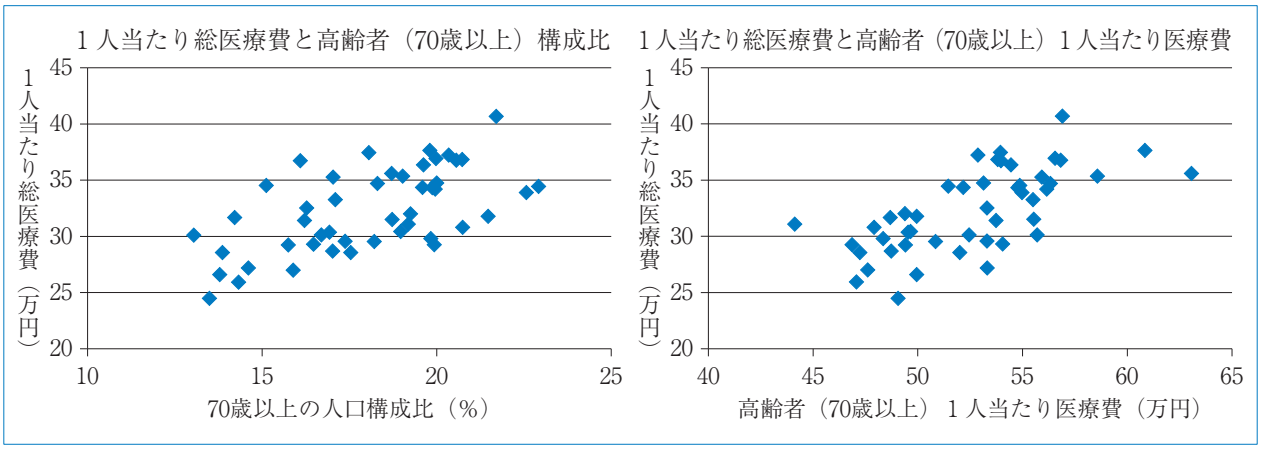


地域差指数（年齢補正後）



◇ 医療費の差異の要因を探る

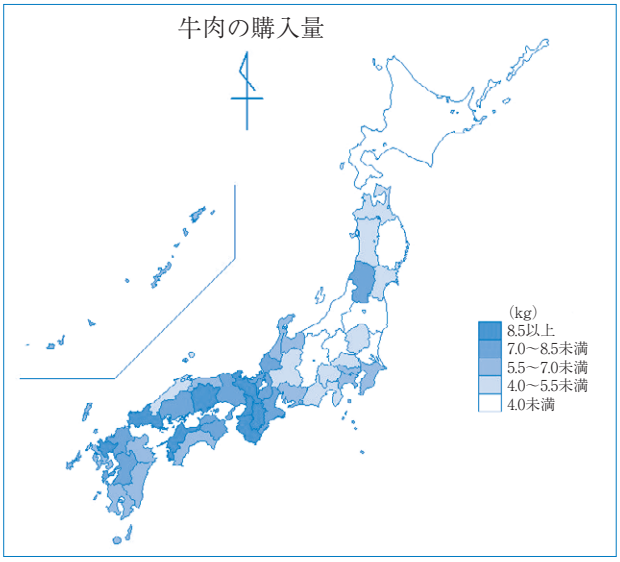


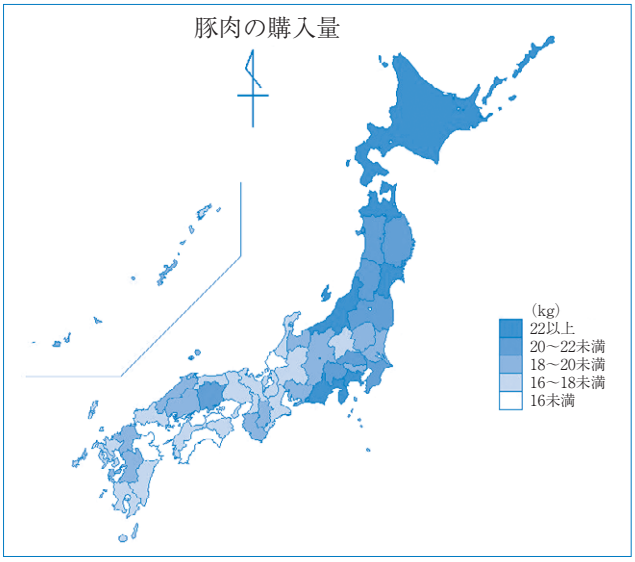


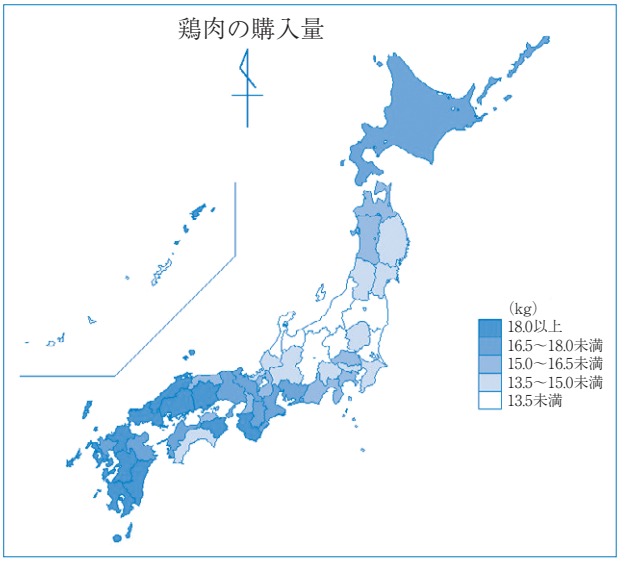
資料：「1人当たり総医療費」は厚生労働省「平成25年度国民医療費」と総務省「人口推計」から算出、「受療率」は厚生労働省「平成20年患者調査」の入院と外来の合計、「人口1万対医師数」は厚生労働省「平成24年医師・歯科医師・薬剤師調査」、「70歳以上高齢者構成比」は総務省「人口推計」、「高齢者1人当たり医療費」は厚生労働省「平成23年度老人医療事業年報」

地域色が表れる食生活の違い

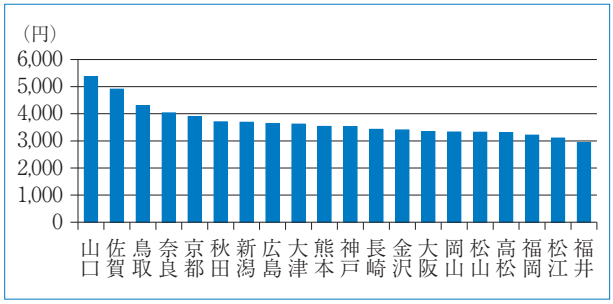
◇ 西日本は牛肉、東日本は豚肉、九州は鶏肉が好き



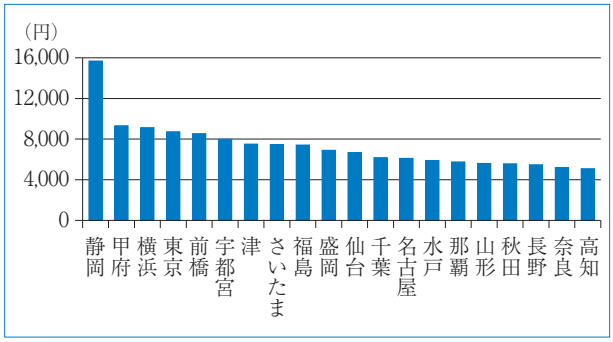




◇ 白身魚（たい・かれい）の支出は西日本が多い！
（たい＋かれい）の１世帯当たり年間支出額上位20都市



◇ まぐろの支出は東日本が多い！静岡は突出
まぐろの１世帯当たり年間支出額

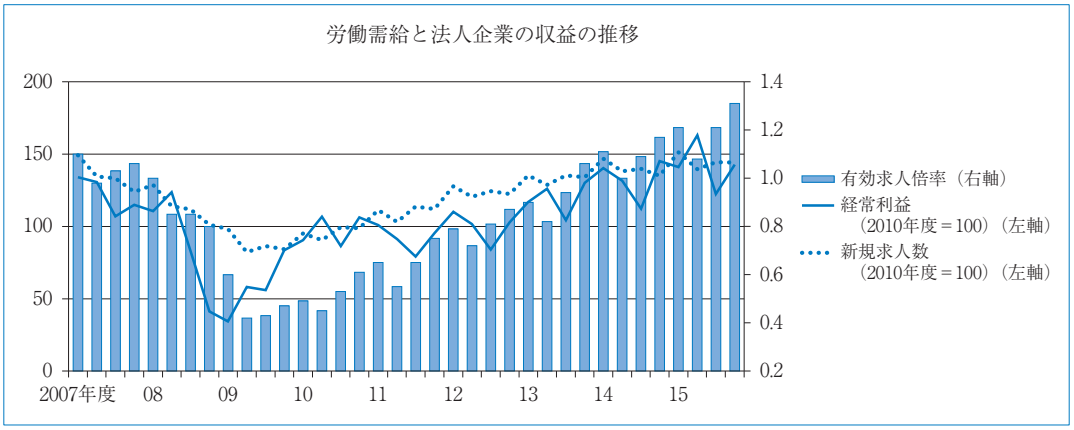


資料：総務省「家計調査」
注：都道府県庁所在市・地方別における２人以上世帯の１世帯当たり年間購入数量および支出額（2015年）

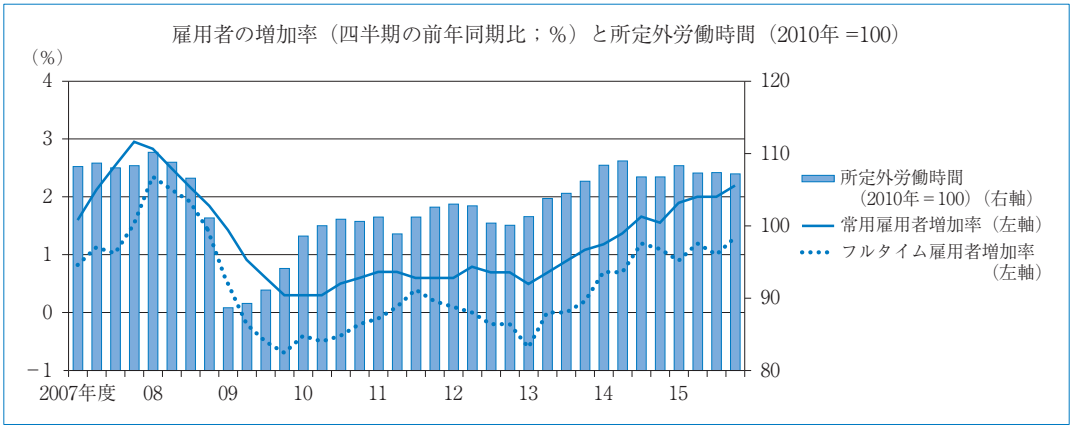
4　関心を高める雇用・賃金・物価

改善の兆しが見えてきた雇用・賃金状況

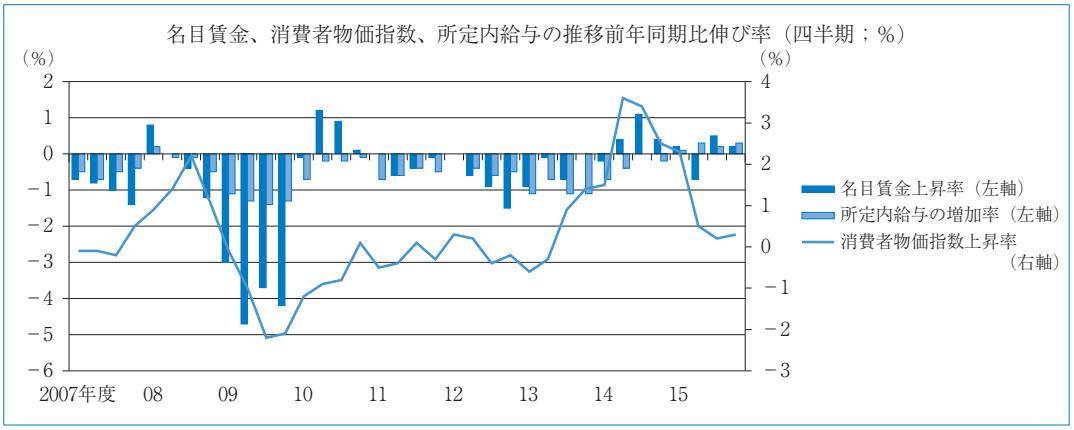
◇ 企業収益の好転とともに新規求人数は増加し、労働需給は改善



◇ 所定外労働時間の増加が一服し、フルタイム雇用者が増加に転じる



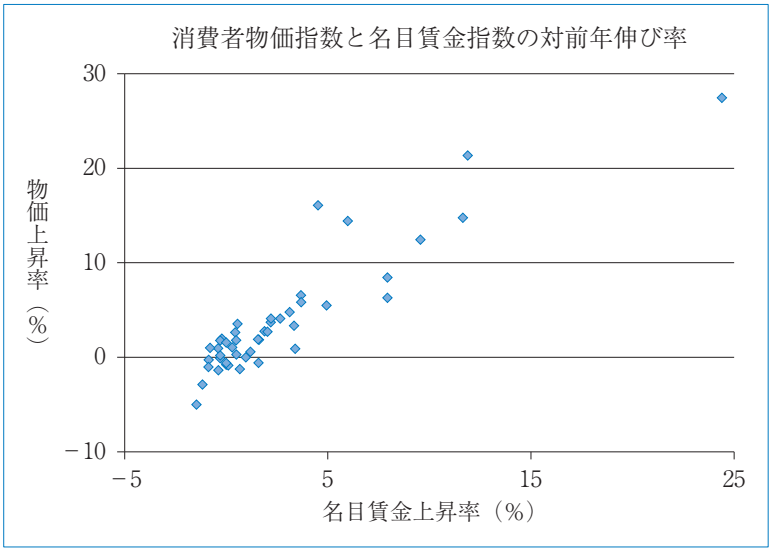
◇ 名目賃金が増加する中で、ようやく増加に転じた所定内給与



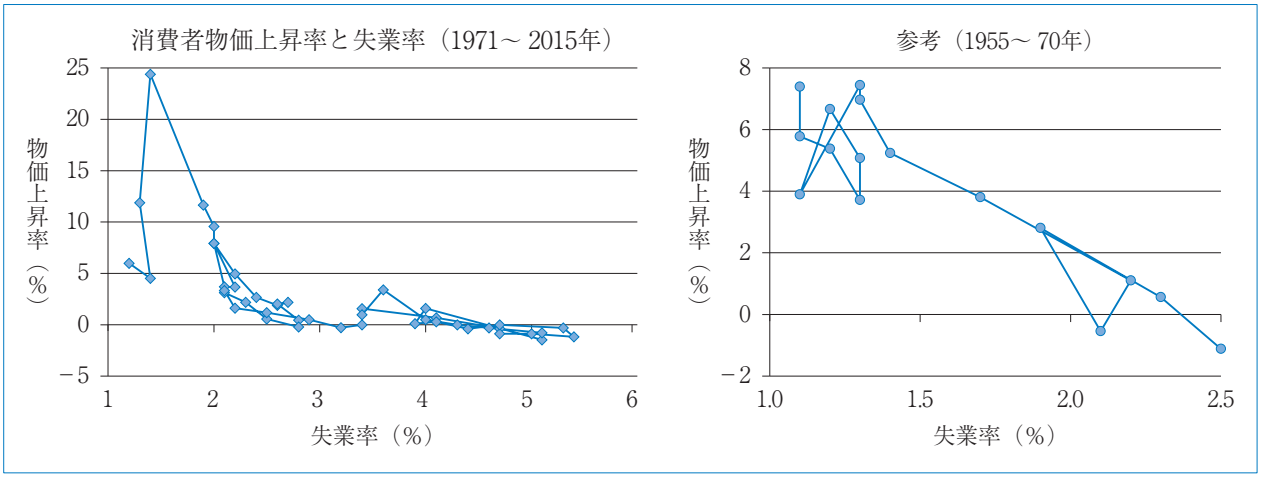
資料：厚生労働省「職業安定業務統計」・「毎月勤労統計」、財務省「法人企業統計季報」、総務省「消費者物価指数」
注：求人数は新規学卒者を除きパートタイムを含む、経常利益は金融・保険を除く全産業。雇用者・賃金・給与の調査対象は雇用者５人以上の事業所。

失業率と物価（賃金）の関係は？フィリップス曲線から知る！

◇ 名目賃金上昇率と物価上昇率は極めて高い相関



◇ フィリップス曲線（失業率と物価上昇率の関係）の形状は理論通り

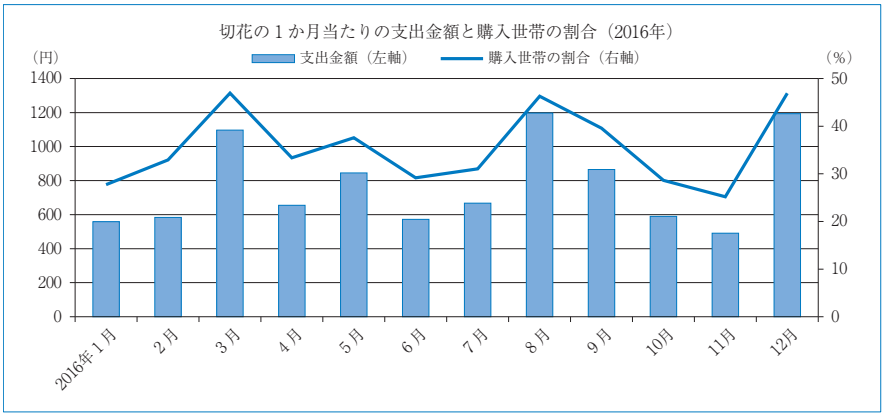


資料：総務省「消費者物価指数年報」・「労働力調査」、厚生労働省「毎月勤労統計調査」

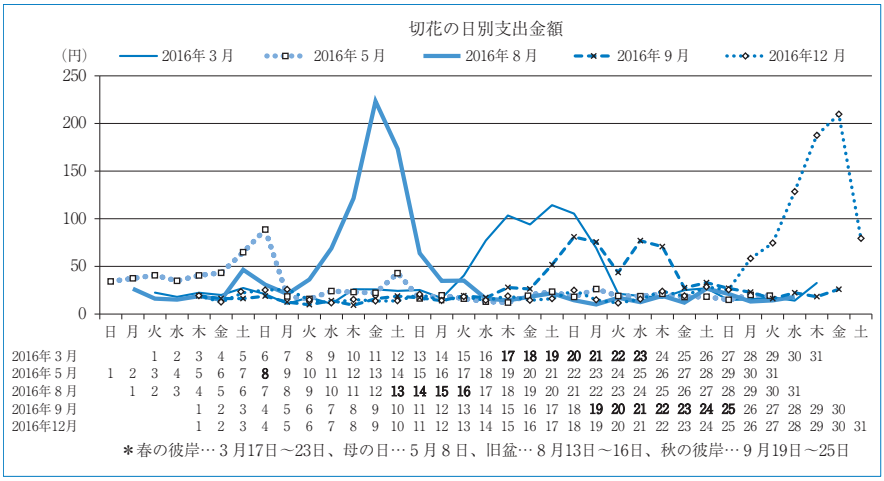
5　統計データに表れる日本の歳時記

切花は季節行事に密接に関わる

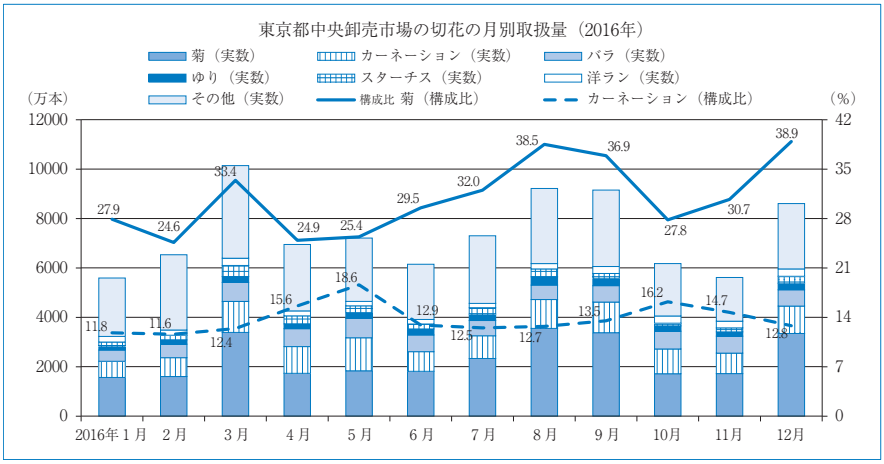
◇ 切花の１か月当たり支出金額と購入世帯は３月、５月、８月、９月、12月に多い



◇ ３月は春の彼岸、５月は母の日、８月は旧盆、９月は秋の彼岸、12月は大晦日前日（30日）にピーク



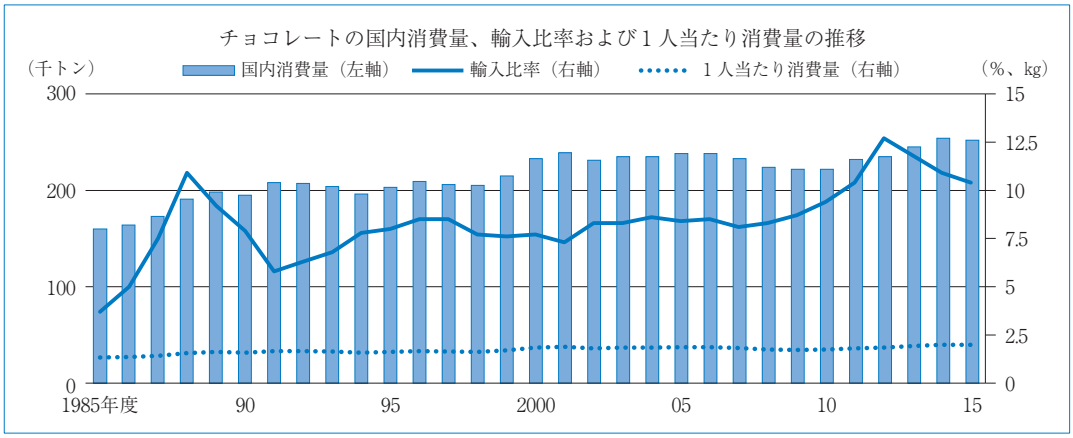
◇ ３月、８月、９月、12月は菊の花、５月はカーネーションの構成比が高い



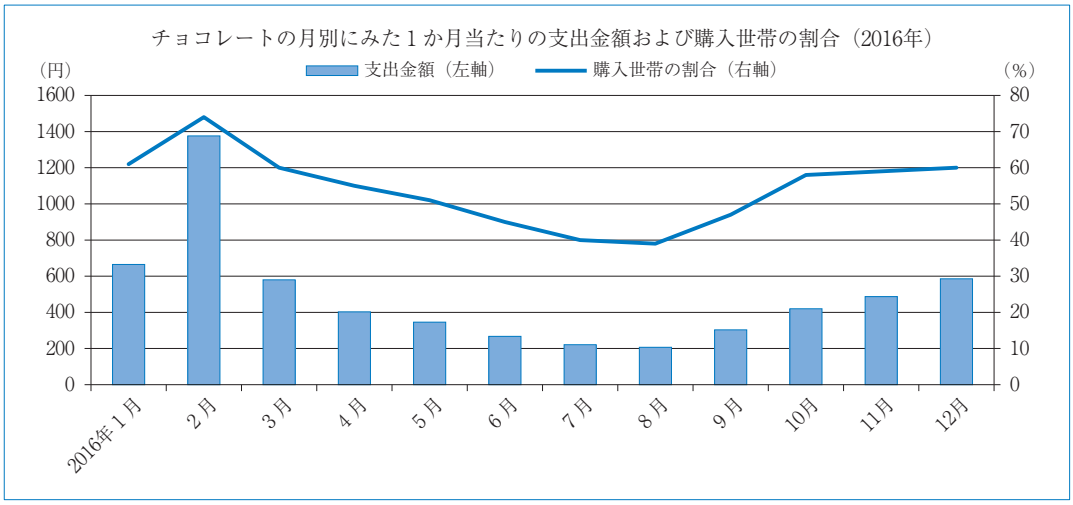
資料：総務省「家計調査」、東京都卸売市場ホームページ
　注：支出金額・購入世帯は２人以上の世帯

バレンタイン・チョコのいま？

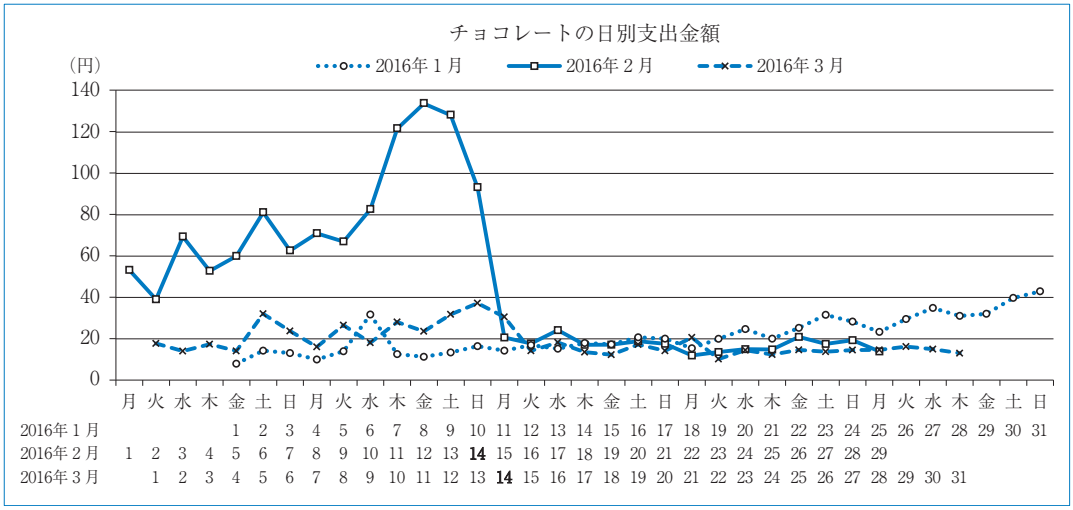
◇ チョコレートの人気は堅調、近年輸入が増加



◇ ２月に突出するチョコレート支出と購入世帯比率



◇ バレレンタイン・チョコはバレンタインデー（２月14日）に向けて１月下旬から始まり、前日までの数日間に盛り上がる



資料：日本チョコレート・ココア協会ホームページ、総務省「家計調査」
　注：数量はいずれも推計値である。支出金額・購入世帯は２人以上の世帯

第9部 公的統計の有用性を知り、統計調査の重要性を学ぶ

Ⅰ　大量で多様なデータを駆使するビッグデータ時代を生き抜く！

1　「統計」の重要性を知り、データサイエンスを身に付ける

統計は国家の基本ともいうべきものです。

日本の歴史において、律令制による中央集権国家の形成に当たり、統計は徴税、兵役などの目的で作成されたと考えられます。戦国時代においても、北条早雲（1432～1519）が初めて検地を自領地で実施し、全国的には、豊臣秀吉（1537～1598）が大名の申告制によって、検地を実施しています。諸外国においては、古代エジプトで帝政ローマの初代皇帝アウグストゥス（紀元前63～14）が、ピラミッド建設のために人口や土地を調べる調査（Census）を行ったと言われています。また、フランスでは、統計の重要性に着目したナポレオン（1769～1821）によって、1801年に統計局が設置され、政府が統計を整備するようになりました。このように、統計はいつの時代でも国家の運営に不可欠なものと位置づけられています。

また、私たちの暮らしのなかにおいても、ICT（情報通信技術）の急速な進展とともに、携帯電話、スマートフォン、タブレット、パソコンなど、ICT 機器は、今や持っていない人はいないほど普及しています。すぐに知りたいことを、日本のみならず世界中から集めることが容易になり、いつでも、どこでも、タイムリーに情報を入手できて、日常の生活のなかで、調べる、知る、使う、そして行動する時代になっています。こうした時代において、公的統計のオープンデータや民間のビッグデータを活用することは、的確な判断を下す上で欠かせぬものであり、情報をどのような状況で活用するかに際して、あふれる情報の内容と質を見極める能力が必要です。

データサイエンス（統計的探究力・データ エンジニアリング）の学習を通して、データを加工し、そこから有効な情報を引き出す能力が身に付きますが、その前提として情報の価値を正確に認識し、適切に利用することが肝要です。

2　いつでも、どこでも、すぐに入手できる統計データの正しい活用を知る

統計データは、「政府統計の総合窓口（e-Stat）」から入手できます。e-Stat は各府省が公表する統計データを 1 つにまとめ、統計データの検索を始めとした、さまざまな機能を備えた政府統計のポータルサイトです。各府省が公表している統計表を Excel・CSV 形式でダウンロードしたり、データベースを使って人口ピラミッドなどのグラフを作成する機能や統計データを地図上に表示する機能など、利用者のニーズに応えうる多くの機能を備えており、誰でも、手軽に、分析することができます。

Ⅱ　住みよい街づくりは統計調査への協力から ～労働力調査を具体例として～

重要な公的統計の多くは、社会の実態を把握するため、各府省が世帯や事業所・企業に対して統計調査を実施することによって作成されています。その結果は、政府が政策を立案したり、景気を判断する等において欠かせぬ情報基盤を成しています。しかし、統計調査において、世帯や事業所・企業から調査への協力が得られない、あるいは、適切に回答されなかった場合、社会の実態を正確に把握できず、政策運営や景気判断で間違った決定を行うことにもなりかねません。そうならないためには、誰もが統計調査に協力することが必要で、的確な統計情報は、より良い行政、社会、経済の基盤となって、豊かな街、住みよい街づくりに直結します。

具体例として、国民に最もなじみのある統計の１つである「労働力調査」を取り上げ、労働に関する公的統計がどのような経緯で始まり、現在に至るまで、どのようにして作成されてきたのかを説明します。さらに、調査がどのように実施されているかを知り、また、外国と比較して調査結果を見れば、統計データへのより一層の理解が深まることでしょう。

１　就業者と失業者を把握することの重要性の高まり

（１）　経済の発展と景気循環（好況と不況の循環）

産業革命以降、著しく発達した世界経済は、人々の暮らしを豊かにし、住宅や金融資産への投資が拡大するなど、それ以前に比べて、お金の流通量が加速度的に増加し、好不況の波が大きくなりました。

第１次世界大戦が終わった後、ヨーロッパの復興需要の高まりなどを背景として、米国を中心とした世界経済は、好循環を生みだし、1910年代末から1920年代の後半まで長い間の繁栄を維持しました。しかし、その後は反転し、株式市場で1929年10月に勃発した「暗黒の木曜日」以降、行き過ぎた投資の反動による世界恐慌と呼ばれる大不況時代に突入しました。これにより、米国の失業率は25％近くの高水準にまで上昇し、仕事を探す人たちが都市にあふれ、政府にとっても失業率の改善が最大の関心事になりました。それまでの経済学では、労働市場における需要と供給のギャップは、賃金の変動により調整され、失業は解消するという考え方が一般的でした。しかし、世界恐慌以降は、ケインズ（1883～1946）が提唱した新しい経済理論により、労働需要が不足している状態では、賃金の調整によって失業は解消されないという考え方が一般的になり、景気循環との関連で失業率が注目されるようになりました。

（２）　失業率の計測が重要な課題

世界恐慌を経験した後、米国は、ニューディール政策といわれる積極的な公共事業により需要を喚起する経済政策を推進し、不況からの転換を図ることを目標としました。そこでは、経済状況を判断する上での経済指標として、正確な就業者数の把握と失業率の計測が重要な課題となりました。このような背景の下、米国では、“Current Population Survey”（米国における現行の労働力調査）の原型となる調査が実施され、現在まで継続しています。この調査に基づいて、現在も、就業者数、失業者数、失業率などの数値が毎月公表されています。

我が国では、第２次世界大戦後の壊滅的な経済状況からの復興のため、GHQ により基礎的な統計調査が整備されることとなり、その１つとして労働力調査が始まりました。労働力調査は、戦後の混乱した社会経済の実態を把握するため、1946年９月から試験的に開始され、約１年間の試験期間を経て、1947年７月から本格的に調査が行われるようになりました。

（３）　国際基準としての ILO 決議（国際労働統計家会議）の発布

各国が労働統計を整備していくなかで、就業者、失業者等の諸概念は、1940年代に登場し、その後、次第に明確なものへと整備されてきました。これらの諸概念を整理し、国際基準の策定を担ってきた組織が国際労働機関（ILO）です。また、ILO の１機関として、労働統計の作成業務を担当する代表者を招集する国際会議「国際労働統計家会議」（1923年第１回開催）が発足し、おおむね５年毎に開催され、2013年までに19回開催されています。

とくに、1947年に開催された第６回において、現在の労働力調査の概念と同一の「調査時点における活動状態」（我が国では、月末１週間の就業状態）を調査することを決議しています。これは、定義をより厳密にすることによって、就業者や失業者を正確に把握し、かつ、時系列で各月を比較することを容易にするなどの利点があることから、国際的に広がっていくことなりました。

現行の労働力、就業、失業に関する指針は、1982年に開催された第13回の会議において採択されました。とりわけ、経済活動との関連性を重視し、労働統計と生産統計が整合するように概念の統一がなされました。つまり、就業者は国民経済計算体系（SNA）に計上されている商品およびサービスの生産に向けた労働力を供給する人のことを指し、また、失業者は労働力のうち、利用可能でありながら、それが生産に活用されていない人のことを指す

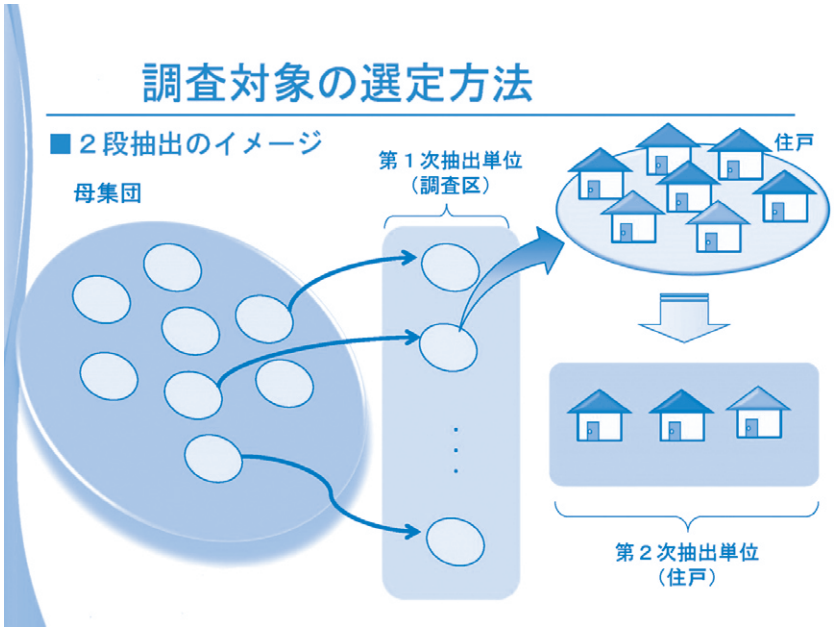
よう明確に定義されました。

2013年に開催された第19回会議では、雇用形態の多様化の進展や潜在的労働力の有効利用のための指標が検討されました。具体的には、就業者のうち、仕事はしているが、もっと追加的に仕事をしたい者や、現在仕事をしていないけれども、条件が合えば、仕事をすることを望んでいるといった潜在的な労働力を把握することの重要性が議論されました。そして、これらの未活用労働を的確に計測することにより、各国の失業率を含む未活用労働指標が整備されるよう各国に協調を呼びかけました。

2　一部を調査することにより日本全体の現状を推計する

労働力調査が明らかにしようとするのは、就業者数、失業者数など労働に関する15歳以上の人口の総数であり、我が国の15歳以上人口は、2014年時点で１億１千万人です。労働力調査においては、その約１/1100（これを抽出率といい、抽出率の逆数を乗率という）の10万人を毎月調査して全体を推計しています。このように一部を調べて全体を推計するための調査を標本調査といい、対象全体の１億１千万人が母集団（労働力調査の場合は、15歳以上人口）であり、調査対象として抽出された10万人が標本となります。

標本抽出の方法は、まず、国勢調査から得られる調査区を第１段で抽出し、抽出された調査区の中から無作為に調査対象を抽出する方法を採っています。これは、層化２段抽出といわれる方法で、より具体的には、国勢調査から得られる情報を元に、農家が多い地域、製造業に勤める者が多い地域などをグループ化し、グループごとに定められた調査区を抽出（第１次抽出）し、次に抽出された調査区から住戸に１番からの番号を付与し、等確率で無作為に抽出（第２次抽出）するという手法で、この方法を採用することによって、層化せずに抽出するよりも良い標本になるようにしています。



3　調査の方法と ICT を活用した調査方法へ

公的統計の多くは、統計調査員が対象となる世帯や企業、店舗を回って調査して作成されます。労働力調査の対象は、我が国に居住する世帯となるため、調査の対象となる調査区について、準備調査の段階で、調査区内の地図を作成し、住戸が正確に分かるように地図に書き込みます。その住戸の中から等確率で調査対象を選び、選ばれた世帯を対象世帯と定めます。当該世帯に対して、調査票を配布し、15歳以上の世帯員全員について、月末１週間の就業状態について調査票に記入してもらい、それを後日、調査員が回収します。

外国では、統計調査員が対象者に対して、対面形式で質問し、答えてもらう方法で調査をしている国もありますが、我が国の労働力調査では、対象となる回答者に調査票を記入してもらう方式で実施しています。また、郵送で調査票を送付する方法も考えられますが、この方法では、回収を催促することが困難であることから、労働力調査では、統計調査員が訪問して、調査票を配布・回収する方法を採用することで高い回収率を確保しています。

近年、特に都市部においては、居住者のプライバシーに配慮したオートロックマンション等の形態の住戸に住む住民が多いことや、共働き世帯の割合が増加し、調査員が日中に訪問しても対象となる世帯に会うことが難しい状況にあります。そのような厳しい状況下であっても、調査員が何度も世帯に足を運ぶなどの努力の結果として、毎月、正確な統計調査結果が得られています。

また、国勢調査では、プライバシー意識の向上などから、希望者はインターネットを通じたオンライン調査で回答することが可能となっており、公的統計においても、一般的な調査方法の１つとなりつつあります。労働力調査でも、将来的にオンライン調査の導入を検討しており、調査の回答者も ICT を利用したさまざまな方法を選択することが可能となることでしょう。

4　回収された調査票から結果の集計、公表まで

回収された調査票は、都道府県に集められ、記入内容の審査を経て、労働力調査を所管する総務省へ配送されます。総務省では、回収された調査票をセキュリティー対策が施された倉庫に収め、厳重な管理をした上で、次のステップである集計作業に進みます。

全国から集められた調査票は、全都道府県分がそろった後、OCR と呼ばれる調査票内容を読み取る機械にかけて、10万人分の記入内容を読み取ります。ここからアナログの紙情報からデジタル化された集計用データへと生まれ変わります。

いくつかのデータ処理工程において、記入内容をチェックし、データを補定するなどの処理を施した後、１枚の調査票情報を約1000人の代表として乗率を掛けて計算することで、我が国で働く就業者の数、失業者の数といった調査項目ごとに統計データを作成していきます。多くの結果表は、男性・女性の別、年齢階級別、地域別などさまざまな項目ごとに、我が国の労働状況がどのような状況にあるのかを分析することが可能となるように作成されます。

労働力調査結果として作成された統計表は、毎月、報告書として冊子にまとめられるほか、統計局のホームページでも公開されます。主要な経済指標は、その時点での景気動向を表わし、統計数値の公表結果が金融マーケットに影響を及ぼす可能性があることから、公表時間は厳格に守られ、公表数値は国家公務員法で守秘義務が課された者だけで管理し、事前に公表数値が漏洩しないように厳重に管理しています。

また、統計の利用者は行政施策を企画する国の官庁だけでなく、民間企業でもさまざまな用途で利用されているため、多様な利用者に利用しやすい媒体で提供されています。e-Stat や過去からのデータを収録したデータベースから、利用者がデータを検索し選択することによって、必要なデータをダウンロードできるシステムも提供されており、ICT の利用により利用者の利便性を高めています。

さらに、国際機関である OECD（経済協力開発機構）では、日本の労働力調査のデータを含め、世界中の就業者、失業者などの数値を収集し、各国のデータをデータベースで管理し、必要なデータをダウンロードできるサービスを提供するなど、さまざまな利用者の要求に応えるようなシステムを構築しています。アクセス方法は比較的容易なことから利用者が多いと思われます。

5　統計データから得られる情報

（１）　失業率の推移

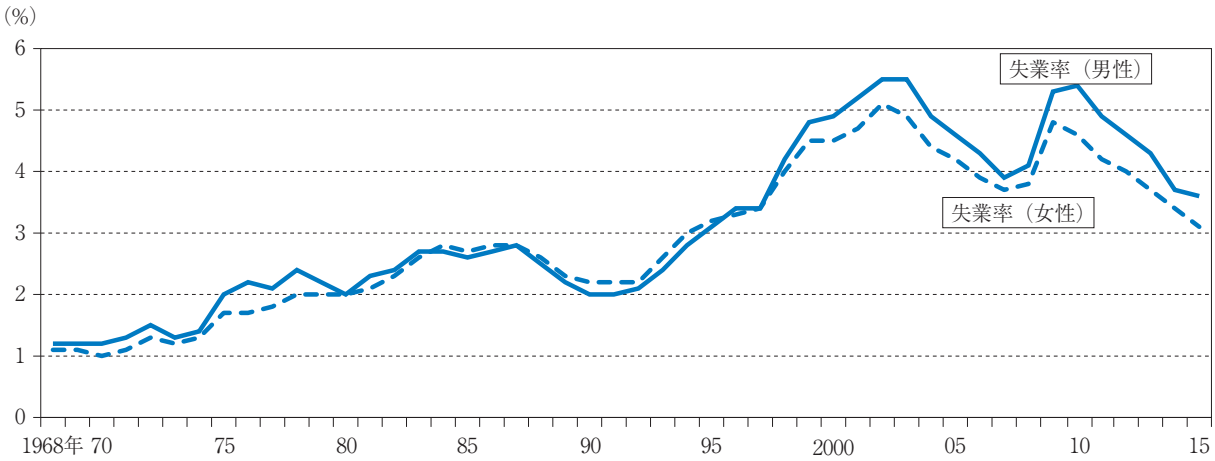
労働力調査から最も利用者に関心のある失業率について、現在の調査と比較可能な1968年（昭和43年）からの推移を図１から見ると、第１次オイルショック以前の1960年代～70年代前半にかけては、男女ともに１％台で推移し

ていたことが分かります。しかしながら、1973年に勃発した中東戦争をきっかけに起きた第１次オイルショック後に男性で初めて２％台に達した後、なだらかに上昇していることが分かります。

ただし、1960～70年代の我が国では、日本型雇用システムといわれる安定した雇用形態が定着しており、諸外国に比較して低い失業率が継続していました。日本型雇用システムとは、(a) 終身雇用、(b) 年功賃金、(c) 企業内組合制度という特徴を持ち、新規に学校を卒業した者が正社員として雇用された後、同一企業に生涯を通じて継続して働き続け、企業内でさまざまな職種を経験するものの、企業内での継続就業を前提とした働き方となっていました。このような仕組みのことをメンバーシップ型雇用システムといい、ヨーロッパの雇用システムをジョブ型という人もいます。メンバーシップ型の長所としては、景気変動によって仕事が減ったとしても企業内の他の仕事につくことで正社員としての身分が保障されるということが挙げられます。このことが日本の失業率を低い水準で継続していた要因といわれています。

しかし、1990年代のバブル景気の崩壊後、我が国の雇用慣行は大きく変化し、日本型雇用システムが変容し、非正規雇用者が増加、雇用が流動化することで失業率も上昇することとなりました。とくに、最近の動向をみると、2008年に発生したリーマンショックといわれる金融機関の破綻を引き金にした世界的な景気後退期には、失業率が高くなって５％台の半ばまで達していることが分かります。

図１　失業率の推移



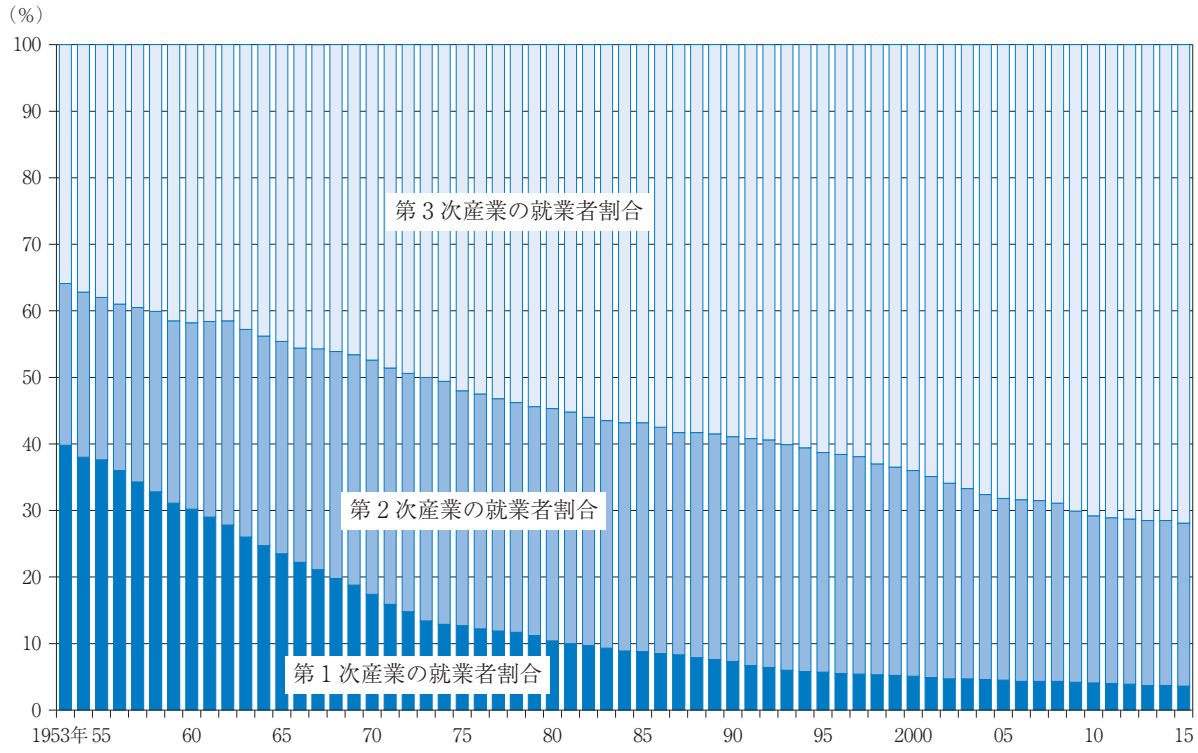
資料：総務省「労働力調査」

（２）　産業別の就業者割合の推移

労働力調査は、就業者の勤務する産業を調査しています。図２から、産業別の就業者の割合を第１次産業、第２次産業、第３次産業別に見てみると、1953年当時は、農業や漁業を含む第１産業が39.8％と最も高いことが分かります。その後、1960年代～70年代には、製造業や建設業を含む第２次産業の就業者割合が上昇しました。当時の我が国では、1964年に東京オリンピックが開催され、1970年には大阪万博が開催されるなど、高度成長を後押しする特需があったほか、新幹線や高速道路の建設等により、製造業、建設業に多くの労働需要があったことにより、これらの就業者数が増加していました。当時の日本の GDP は1968年に西ドイツを抜き、アメリカに次ぐ世界第２位となるなど、第２次世界大戦後から急速な復興を果たしました。このような世界に例を見ない我が国の急速な経済発展は、「東洋の奇跡」と言われるほどの著しい成長でした。

一方、1970年代の後半頃から第３次産業の就業者の割合が上昇していることが分かります。これは、ペティ＝クラークの法則（「**第２部３　サービス経済化の状況とその背景を探る**」のコラムを参照）といわれ、経済が成熟する過程では、就業者は第１次産業から第２次産業へ移行し、そして先進国のような経済が成熟した社会ではサービス業等の第３次産業の就業者が上昇することを示しています。このような傾向はどの国でも観察され、日本では、現在約７割の就業者が第３次産業で働いています。

図２　産業別の就業者割合の推移



資料：総務省「労働力調査」

（３）　M 字カーブ（女性の年齢階級別労働力人口比率）の推移

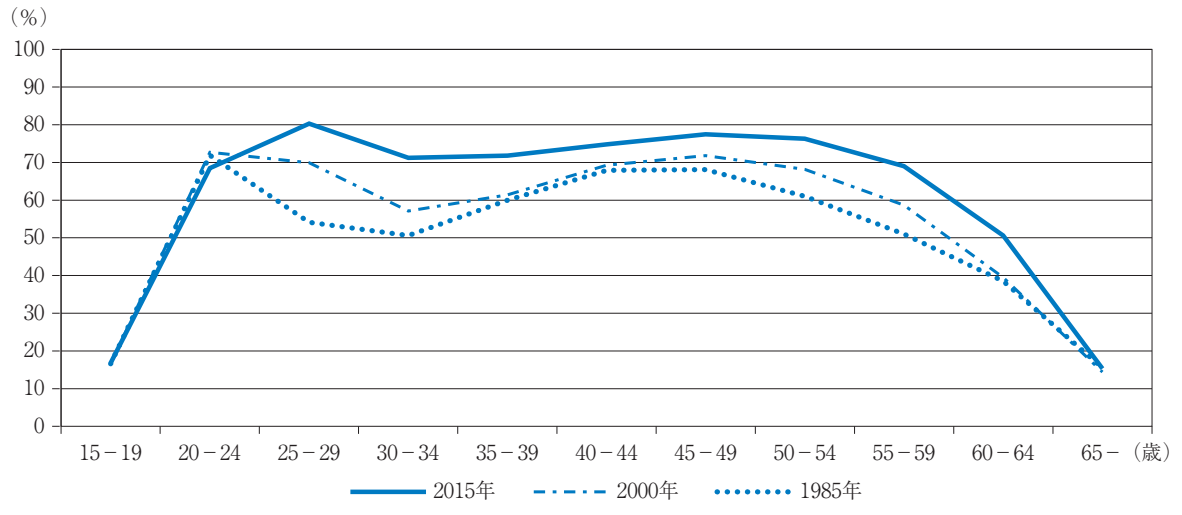
1985年の女性の年齢階級別労働力人口比率を示す図３を見ると、20歳代前半と40～50歳代の労働力人口比率が高く、出産から子育てを担う25～29歳および30～34歳で比率が低い M 字型のカーブを描いていたことが分かります。

一方、30年後の2015年について見ると、1985年にM字カーブの底であった30～34歳の比率が顕著に上昇しており、M 字カーブの谷はかなり浅くなっていることが分かります。また、図４の配偶者がいる女性の年齢階級別労働力人口比率を見ても、2015年は1885年、2000年に比べて、20～24歳、25～29歳、30～34歳の各階級で大幅に高くなっていることが分かります。

これは、男女雇用機会均等法（1986年施行）が施行されて以降、多くの女性が労働市場に参加してきた結果を反映しているものです。かつての日本では、男性が労働力として企業などで生産活動に携わる一方で、女性は家事・育児を担うという伝統的な家族スタイルが一般的でしたが、ライフスタイルの変化や価値観の多様化などにより、男女の役割分担が変化し、女性が労働力として社会進出することが一般的になったことが主な要因であると考えられます。

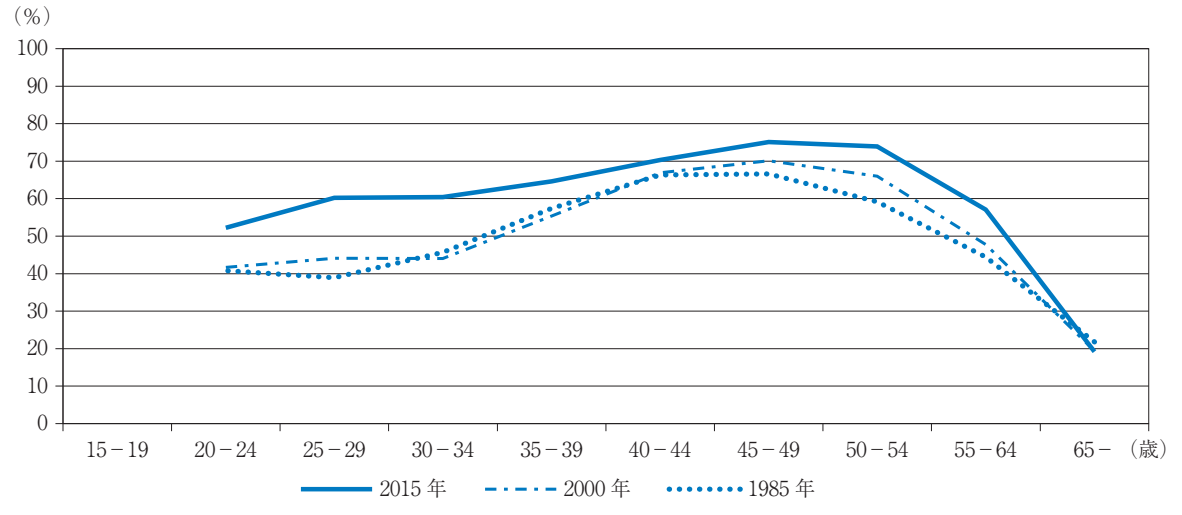
また、最近の我が国の課題として、現在、急速に進行している人口の高齢化に伴う労働力人口の不足に陥る状況を回避するために、女性が活躍できる社会の構築が政府の最重要課題となっており、この M 字カーブの解消のためにさまざまな施策を推進しているところです。

図３　女性の年齢階級別労働力人口比率の推移



資料：総務省「労働力調査」

図４　女性の年齢階級別労働力人口比率の推移（有配偶）



資料：総務省「労働力調査」

このように、労働力調査から現在の我が国の雇用情勢が分かるさまざまなデータが提供されています。公的統計は、社会を写す鏡といわれることがありますが、我が国の状況を的確に、かつ、客観的な数値で表すことができるとても重要なデータです。統計データが無ければ適切な行政施策が実行できなくなります。統計調査員等の多くの人が苦労して作成している公的統計は、とても貴重であり、経済の安定や、安心・安全な社会を作るために、非常に重要な役割を担っています。

データサイエンス　トピックス No. 1

統計データが欲しい人、データサイエンスを身に付けたい人は、ここを見てください！

①　統計データが欲しいときは・・・こちらから

「政府統計の総合窓口（e-Stat）」

http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

各府省が公表する統計データを１つにまとめ、統計データの検索の他、さまざまな機能を備えた政府統計のポータルサイトです。

各府省が公表している統計表を Excel・CSV・PDF 形式でダウンロードでき、また、データベース化されたデータを使って人口ピラミッドなどのグラフを作成する機能、統計データを地図上に表示する機能など、利用者のニーズの高い機能を数多く備えています。

②　データサイエンスを身に付けたいときは・・・こちらから

生徒のための統計活用～基礎編～（平成28年５月 刊行）

http://www.soumu.go.jp/toukei_toukatsu/index/seido/stkankyo.htm

生徒が身近な現象や社会の課題を研究することを通して、課題学習や自由研究の取り組み方を学ぶ、中学生以上向け教材である学習ワークブックです。

データサイエンス・オンライン講座「社会人のためのデータサイエンス入門」

https://www.youtube.com/c/stat-japan

統計学のプロフェッショナルが分かりやすく解説するオンライン講座です。

初めて学ぶ統計

https://www.youtube.com/c/stat-japan

統計の見方や使い方について、体系的に学習できる講座です。

なるほど統計学園高等部

http://www.stat.go.jp/koukou/index.htm

統計を分かりやすく学べる高校生用のサイトです。情報化社会を生き抜くために必要な、統計を「読み解く力」と「活用する力」を養うための材料が満載されています。

なるほど統計学園

http://www.stat.go.jp/naruhodo/index.htm

統計を学ぼう、知ろう、楽しもう、小学校高学年から中学生向け学習サイトです。

日本統計学会「統計学Ⅰ：データ分析の基礎」および「統計学Ⅱ：推測統計の方法」講座

https://lms.gacco.org/courses/course-v1:gacco+ga014+2017_04/about

ビッグデータ時代に必要とされる統計的な考え方やデータの要約と分析の基礎的な手法、統計学Ⅰで学んだデータ分析の基礎に続いて、推定・検定・回帰分析などの推測統計の方法について学習できる講座です。

データサイエンス　トピックス No. 2

現在の国勢を詳明せざれば 政府すなわち施政の便を失う
過去施政の結果を鑑照せざれば 政府その政策の利弊を知るに由なし

【意味】
現在の国の情勢を詳細に明らかにしなければ、政府は政治を執り行うことができない。また、過去の施政の結果と比較してみなければ、政府はその政策のよしあしを知ることができない。

この言葉は、２度にわたり内閣総理大臣を務めた大隈重信（1838－1922）が、統計院設立（明治14年５月30日）の建議の中に書いたもので、明治維新後における“明治という日本の新たな国づくり”のためには、社会経済の実態を詳しく捉えた統計データに基づく施策が必要であることを、強く訴えていたメッセージとして現在でも引用されています。

索　引

な 行

は 行

ま 行

や 行

ら 行

わ 行

監修　長　尾　篤　志　　文部科学省初等中等教育局視学官

統計教育のための学習用教材（上級編）開発研究会名簿

座長　渡　辺　美智子　　慶応義塾大学大学院　健康マネジメント研究科教授
　　椿　　広　計　　独立行政法人統計センター　理事長
　　岩　崎　　学　　成蹊大学　理工学部情報科学科教授
　　西　村　圭　一　　東京学芸大学　教育学部教授
　　美　添　泰　人　　青山学院大学　経営学部招聘教授
　　塩　澤　友　樹　　東京都立白鷗高等学校附属中学校　数学科教諭
　　南　雲　裕　介　　新潟県総務管理部統計課　統計情報班主任
　　森　　永　壽　　島根県政策企画局統計調査課　調査分析グループ企画員
　　舟　岡　史　雄　　一般財団法人日本統計協会　専務理事

　　執筆協力者
　　本　田　千　春　　東京学芸大学附属国際中等教育学校　数学科教諭
　　和　田　　弘　　一般財団法人日本統計協会　研究員・事業部長（兼編集担当）
　　和　田　澄　子　　一般財団法人日本統計協会　研究員（兼編集担当）

協力　全国統計教育研究協議会
　　一般社団法人日本統計学会　統計教育委員会
　　一般社団法人日本品質管理学会　TQE 特別委員会
　　統計関連学会連合　統計教育推進委員会
　　統計教育連携ネットワーク（JINSE）

平成29年３月　発行

大学での学びにつながる

高校からの

統計・データサイエンス活用

～上級編～

編集・発行　総務省政策統括官（統計基準担当）付統計企画管理官室
〒162-8668
東京都新宿区若松町19-１


ZEHN DEUTSCHE
10
DEUTSCHE BUNDESBANK
Banknote
10
Deutsche Mark
© DEUTSCHE BUNDESBANK