統計・データサイエンス教育における数学の取り扱い

宮崎大学　藤井　良宜

University of Miyazaki

1　はじめに

21 世紀に入り、インタネットの普及や情報通信機器の急速な発展によって、膨大なデータが蓄積されてきている。これらのデータをどのように生かしていくのかが大きな社会的な課題となっている。この間、データマイニングやビッグデータなどの言葉がブームとなって、データサイエンティストという新しい職業も脚光を浴びている。その後生成 AI の開発は、社会に大きなインパクトを与え、これからの社会では AI を適切に使いこなすことができる人材が求められている。その意味で、AI やデータサイエンスに関する教育が重要になってきており、統計学はその基盤の学問と捉えられている。

日本の学校教育における統計教育については、算数・数学で中心的に行われてきている。これまでの学習指導要領の改訂の歴史を見ると、1998 年、1999 年の改訂の際には、学校５日制が導入され、学習内容の大幅な削減が行われた。この時の改訂では、小学校から高等学校までに導入された「総合的な学習の時間」の中で統計に関する内容が取り扱われることが期待された。この２つの状況が重なって、算数・数学の内容の中から統計的な内容がかなり削られた状況であった。

この状況に危機感を感じて、2005 年に日本統計学会が主体となり、「21 世紀の知識創造社会に向けた統計教育推進への要望書」を文部科学省中央教育審議会と日本学術会議に提

〒889-2192 宮崎市学園木花台西１－１
本研究の一部は、日本学術振興会科学研究費補助金（課題番号 22K02498) の助成を受けて行われたものである。

出するなどの取組が行われた。このあたりの経緯については、渡辺（[9]) に詳しく述べられている。その影響もあって、その後の 2008 年、2009 年の学習指導要領の改訂や 2017 年、2018 年の改訂においては、統計教育の充実が図られ、小学校の算数科と中学校数学科を通して、「データの活用」という領域で統一的に統計教育が進められることになった。

大学においても、滋賀大学にデータサイエンス学部が創設されたのをきっかけに、全国的に統計教育やデータサイエンス教育を行う学部が多く作られ、統計やデータサイエンスの教育に力が入れられている。その後も AI 技術の進展もあり、今後もこの流れは続くことが期待される。その一方で、多くの人を対象として統計やデータサイエンスの教育を行うこともあり、統計教育の内容についても工夫が必要になっている。そこで、本稿ではその際の数学の取り扱いを中心に今後の統計・データサイエンス教育のあり方について議論する。

2　統計教育の目標

2.1　数学の取り扱いによる 3 つの分類

統計教育の目標に関する議論においては、内容の面から統計的リテラシー (statistical literacy)、統計的推理力 (statistical reasoning)、統計的思考力 (statistical thinking) の 3 つに分けて考えられることが多い（Ben-Zvi and Garfield [2])。統計的なリテラシーは、統計的な情報や統計的な研究の結果を理解する際に用いられる、基本的で重要なスキルを表し、統計的推理力は、データのバラツキや偏りなどの統計的なアイデアを用いて、統計情報を十分理解し、その意味を説明できることや統計的なプロセスを説明でき、統計解析の結果を解釈できることを意味し、統計的思考力は、問題の概要をつかみ、それに基づいて自分自身で調査を計画し、データの収集から解析までの全体のプロセスを理解し、解決された問題の結果を批判的な視点から評価できることであると捉えられている。現行の学習指導要領では、統計的な問題解決のプロセスや批判的な思考の部分が強調されており、統計的な思考力が求められていることがわかる。もちろん、統計教育の目標としてとらえるならば、統計的な思考力が重要ではあることは否定しないが、藤井（[5]) でも述べているように、算数や数学の授業の中では、統計的な推理力にも力を入れるべきである。

その一方で、算数・数学の中で統計を取り扱う場合の課題もある。delMas ([3]) は，数学的な推理との違いに着目しながら、統計的推理力の難しさとして，抽象的な概念の理解の難しさ，論理的推論の難しさ，統計的状況がもたらす難しさ，の 3 つを挙げている．藤井・添田（[6]）では、この 3 つの難しさへの対応について議論しているので、そちらも

参照してほしい。また、米国の統計学会がまとめた大学における統計教育の指針であるGAISE College Report([1]) では、統計ソフトウエアを活用することで、統計的な概念の理解や実際のデータ分析に力を入れるよう推奨されている。それでは、大学における統計教育においては、数学的な内容をどのように取り入れていけばよいのだろうか。この点については、その科目においてどのような目標を立てているのかが大きく影響する。そこで、ここでは教育目標を大きく次の３つの場合に分けて、そこでの数学の取り扱いについて考えていくことにする。

ユーザーとしての教育　基本的には、統計ソフトウエアを利用することを前提に、数式等はできるだけ使わずに、統計手法の利用の仕方を強調する。

分析者としての教育　統計的な手法の意味を理解するために必要な最小限の数学を用いて、他の手法との比較や適切な統計手法の用い方を身に付ける。

理論家としての教育　統計量や統計量の分布を導出するのに必要な数学的な内容も含めて理解できるようにし、既存の手法の特徴の分析や新しい手法の開発ができる基盤を作る。

この３つの場合については、以下に挙げる２つの事例を参考に理解を深めてほしい。

2.2　事例１：回帰分析

回帰分析は、目的となる変数 Y を、いくつかの説明変数 $X_1, X_2, \ldots, X_k$ を使って予測する際に用いられる統計手法である。実際には、これらの変数を観測ユニットごとに測定をするため、それを記号であらわすと $y_i, x_{1i}, x_{2i}, \ldots, x_{ki} (i = 1, 2, \ldots, n)$ という形のデータが得られる。この時点で、まず用いられる記号の多さとそれぞれの意味の違いを理解するのに苦労する学生も多い。そして、変数 Y を次のような１次式（回帰式という）で予測することになる。

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

ここでは、データに基づいて１次式の係数である $\beta_0, \beta_1, \ldots, \beta_k$ を求めることになる。具体的には、

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$$

を最小にする $\beta_0, \beta_1, \ldots, \beta_k$ を回帰式の係数の推定値とする。この手法を理解するためには、回帰式の意味を理解することやどのように回帰式の係数を決めるのかが重要である。

さらに、回帰分析においては、求めた回帰式が予測に用いるのに十分当てはまりがよいのか、用いる変数として何を選んだらいいのかなど考えるべきことが非常に多い。

ユーザーとしての教育　具体的な統計ソフトウエアを利用して回帰分析を行うと、かなり多くの情報が提示される。まず、ユーザーとしてはこの中から必要な情報を適切に読み取ることが重要である。例えば、回帰式の係数の推定値を読み取り予測モデルを特定することや、回帰式の当てはまりの良さをチェックすること等が考えられる。もちろん、回帰分析では、複雑な状況をかなり単純化した回帰式で予測するため、回帰式で予測すること自体の限界についても気を付ける必要がある。

分析者としての教育　回帰分析は、予測を行う際の入門的な手法であり、分析者としては手法の基本的な理解だけでなく、他の手法との違いをしっかり理解することが必要である。そのためには、ある程度数式による表現を用いたり、数式の意味を理解することも大切である。特に、モデルを選択する際の基本的な基準として、観測値と予測値の差の 2 乗和を最小とすることを基準とすることを理解することは不可欠である。もちろん、発展形として情報量基準やクロスバリデーション等の手法についても意識しておく必要がある。さらには、複雑なモデルを用いると本来のモデルよりもデータの傾向に引っ張られてしまう可能性があることも理解しておく必要がある。

理論家としての教育　回帰式の係数の推定値を導出する方法や推定値のバラツキや予測誤差などのバラツキを表現する分布の理解が必要である。回帰分析では、多変量の取り扱いが必要であり、行列やベクトルに関する性質なども含めた数理的な内容についてもしっかり理解しておくことが重要である。さらに、近年多くの複雑なモデルの提案が行われており、それらの手法の違いや新しい手法の開発に向けた数理的な理解も必要である。

2.3　事例２： 2 x 2 分割表の解析

2x2 分割表の解析は、カテゴリカルデータの解析においては重要な方法のひとつである。ここでは、独立性の検定でのピアソンのカイ 2 乗統計量を用いた解析を学習する場合の目標について考える。基本的な統計的検定の考え方については既に学習しているものとし、2x2 分割表の分析の内容を学習することが目的である。

一般的には、表 1 のように、観測される各セルの観測度数と期待度数を用いて説明され

表 1　2x2 分割表の観測度数と期待度数

観測度数

	C_1	C_2	合計
R_1	f_{11}	f_{12}	f_{1+}
R_2	f_{21}	f_{22}	f_{2+}
合計	f_{+1}	f_{+2}	f_{++}

期待度数

	C_1	C_2	合計
R_1	m_{11}	m_{12}	m_{1+}
R_2	m_{21}	m_{22}	m_{2+}
合計	m_{+1}	m_{+2}	m_{++}

ることが多い。このときの数学的な部分は、期待度数の計算方法

$$m_{ij} = \frac{f_{i+} f_{+j}}{f_{++}} \qquad i, j = 1, 2$$

や検定統計量の計算方法

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(f_{ij} - m_{ij})^2}{m_{ij}}$$

の説明や、帰無仮説の下での検定統計量 χ^2 の分布が近似的に自由度 1 のカイ 2 乗分布になることの説明をすることになる。統計解析用のソフトウエアが普及する以前は、統計解析を行う人が自分で計算する必要があるため、これらの式を説明することは必須の内容であった。しかし、現在では、ソフトウエアを利用すればこれらの値を自分で計算する必要はない。その点を考慮すると、上の 3 つの目標に対応して、次の事が求められる。

ユーザーとしての教育　ユーザーとしての教育であれば、情報の応じてこの手法を適用するのが適切であるかを判断して、統計ソフトウエアで計算される結果を読み取り、検定結果を判断することが必要である。

分析者としての教育　分析者としては、この検定手法だけでなく他の手法との比較も必要となる。この場合であれば、フィッシャーの直接確率法との比較が問題となるであろう。このように分析法の理解となると、それぞれの手法の意味や違いを理解する必要が生じる。そのため、統計量の形や帰無仮説のもとでの分布の違いなども理解する必要がある。統計量の式の意味を理解しておくことやカイ 2 乗分布の近似の良さの問題なども重要となる。

理論家としての教育　この検定の場合に難しさを感じるのは、統計量の形である。観測度数と期待度数のずれを評価していることは理解できるが、なぜこの形になっているのかという点は意外と難しい。このこととも関連するが、帰無仮説の下での統計量の分布が近似的に自由度 1 のカイ 2 乗分布となることについて理解するためには、いろいろな知識を導入する必要がある。

少し特殊ではあるが、2x2 分割表の解析の場合の難しさとして、データの収集方法の問題がある。それは、表 1 の形にまとめられる状況が 3 つある点である。ひとつは、ある集団に質問紙調査等を実施して、２つの質問を組み合わせて集計した場合である。たとえば、煙草を習慣的に吸っていますか、という質問と週 3 回以上飲酒をしますか、という質問があったとする。この場合には、回答者はこの２つの質問への回答によって、４つに分類されることになる。すなわち、全体の数 f_{++} は回答者の数なので固定されているが、４つのセルの観測度数は調査によってバラツクことになる。そのため、確率モデルとしては多項分布モデルが用いられ、喫煙習慣と飲酒習慣の間に関連性がないという仮説が帰無仮説となる。もう一つは、あらかじめ２つにグループ分けされた集団に質問をして、その答えで２つに分類する場合である。たとえば、20 代、30 代の男性と 40 代、50 代の男性からそれぞれ標本を選び、煙草を習慣的に吸っているかどうかを質問する場合が、これにあたる。この場合は、調査の段階で標本サイズは固定されているので、行または列の合計はあらかじめ決まっていると考える。そのため、確率モデルとしては２つの世代それぞれで２項分布モデルを適用して、２つの世代の喫煙率が等しいという帰無仮説を考えることになる。この２つの場合は、スタートする数学的なモデルは異なっているが、標本サイズが大きければ、どちらの場合にも上で述べたカイ２乗検定を適用することができる。もう一つの状況は、フィッシャーの直接確率法が提案された際のベースになった状況で、紅茶の実験として有名なものである。紅茶の実験では、ミルクティーを作るのに、紅茶を先に入れた方がおいしいか、ミルクを先に入れた方がおいしいかを調べるために、あらかじめ情報としてミルクを先に入れたものの数と紅茶を先に入れた方の数を与えて、与えらえたカップがどちらか判断してもらう実験を想定している。この場合、行の合計も列の合計も固定されていて、その範囲でバラツキを考えることになる。

この３つの場合の確率モデルの違い等については、柳川 [10] で詳しく書かれているので参考にしてほしい。

3　確率に対する理解

統計やデータサイエンス教育においてはバラツキへの対応が不可欠である。例えば、Garfield([7]) で示されている統計的推理力を評価する際の８つの項目の中には、確率の正しい解釈、確率の正しい計算、独立性の理解や標本変動の理解など確率や確率分布に関する項目が多く含まれている。この中でも独立な標本の和の分布が特に重要である。しかし、この独立な標本の和の分布を計算することは簡単ではない。たとえば、１から６の目が同じ確率で出るサイコロを投げた時の出た目の和の分布を考えてみよう。この問題は、

中学校２年生の数学の授業で２個のサイコロの出た目の分布を取り扱っているように、確率の学習では基本的な問題のひとつである。しかし、サイコロの数が３個、４個と増えていくと、計算は急に難しくなっていく。現実的な場面を考えると、統計では独立な標本の和の分布はよく出てくる問題であるが、その場合には３個や４個ではなく、100 個や 1000 個といった数の和の分布の問題になるので、もっと複雑である。

確率に関する有名な問題として、Fischbein ら ([4]）が示した病院の問題がある。この問題では、一日に 15 人程度生まれる小さな病院と一日に 45 人程度生まれる大きな病院において、1 日に生まれた子どもの中で男の子の数の割合が 60% 以上となる日を１年間まとめた場合に、どちらの病院が 60% 以上の日が多いかという問いが用いられている。問題文自体はもっと長いので、問題の理解だけでも苦労する問題ではあるが、意外と正解の割合が低いことが示されている（松浦 [8])。この問題は、割合の推定においては標本サイズが大きな意味を持つことと繋がっている。標本サイズが小さい場合には、割合の推定のバラツキが大きくなるため信頼ができないのである。割合の推定量の分布については、2 項分布を用いて計算することができるので、推定量の分散と標本サイズの関係を求めることで説明は可能である。

統計学では、標本平均の分布が良く用いられるが、一般的にはこの計算は簡単ではない。教育においては、この計算を正確に行うことは難しいので、コンピュータシミュレーションを活用することも考える必要があるだろう。また、近年の統計手法においては、ブートストラップ法やランダムフォレスト解析のようにシミュレーションを活用したものも多く見られる。また、クロスバリデーションのように、統計的なモデルを選択する際に用いられる方法においてもシミュレーションは不可欠になってきている。ただ、これらのシミュレーションの結果を、人々がどのように理解をしているのか、また理解を深めていくためにどのような教育を行っていくのかについても、今後研究を進めていく必要がある。

4　終わりに

本稿では、統計やデータサイエンス教育においてどのように数学を取り扱っていくべきかについて、議論を進めてきた。コンピュータやソフトウエアの開発によって、統計解析を行う際に必要となる計算は統計ソフトウエアを用いることで容易になり、多くの人にとって統計やデータサイエンスが身近なものになってきている。また、統計手法についてもシミュレーションを利用した方法なども提案されどんどん複雑になってきている。それらの統計手法を適切に利用するための教育は、これからより重要になっていくと考えられ、より多くの人に理解してもらう努力をする必要がある。その上で、それぞれの目的に

応じて、数学の利用の仕方も工夫をしていくことが今後重要である。

参考文献

[1] American Statistical Association. “Guidelines for assessment and instruction in statistics education college report 2016”, 2016, https://www.amstat.org/docs/default-source/amstat-documents/gaisecollege_full.pdf.

[2] Ben-Zvi, D. and Garfield, J., “Statistical literacy, reasoning, and thinkin:Goals, definitions, and challenges”, The challenge of developing statistical literacy, resoning and thinking(Ben-Zvi, D. and Garfield, J. ed.), Springer, pp.2-15, 2004.

[3] delMas, R. “A comparison of mathematical and statistical reasoning”, The challenge of developing statistical literacy, resoning and thinking(Ben-Zvi, D. and Garfield, J. ed.), Springer, pp.79-95, 2004.

[4] Fischbein, E. and Schnarch, D. ”The Evolution With Age of Probabilistic, Intuitive Based Misconceptions”, Journal for Research in Mathematics Education, Vol.28, pp.97-105, 1997.

[5] 藤井良宜, “統計を生涯学び続ける生徒を育成する授業を目指して”, 日本科学教育学会年会論文集, Vol.48, pp.189-192, 2024.

[6] 藤井良宜, 添田佳伸, ”統計教育の到達目標の設定と木法達成のためのアプローチ” 日本統計学会誌, Vol.36, No.2, pp.251-262, 2007.

[7] Garfield J., “Assessing statistical reasoning”, Statistics Education Research Journal, Vol.2, pp.22-38, 2003.

[8] 松浦武人,”児童の確率判断の実態に関する縦断的・横断的研究”, 全国数学教育学会誌, 数学教育学研究, Vol.12, pp.141-151, 2006.

[9] 渡辺美智子, “知識創造社会を支える統計思考力の育成 - アクションに繋がる統計教育への転換 - ”, 日本数学教育学会誌, Vol.89, pp.29-38, 2005.

[10] 柳川堯,”離散多変量データの解析”, 共立出版, 1986,