

多変量データを活用した統計的探究の重要性

ほそだ こうき
細田 幸希

1. はじめに

現行の中学校学習指導要領の「データの活用」の指導では、目的に応じてデータを収集・処理しながらその傾向を読み取って判断したり、データに基づく判断や主張を批判的に考察したりすることが大切にされています（文部科学省，2018）。このような統計的探究を促進するためには、実際に授業で中学生が分析を行うデータが非常に重要になってきます。

本稿では、複数のデータから構成される多変量データを活用した統計的探究に着目し、それを可能にする教材の一つである「十種競技データの問題」（Hosoda, 2022）を紹介しながら、統計指導における重要性について考察します。

2. 多変量データを活用した統計的探究のプロセス

前述の通り、「データの活用」の指導では統計的探究を通して統計的に問題解決する力を養うことが強調されています。一方で、統計的探究を促すためのデータについては、「一般的に数値的な情報だけでなく画像や映像などもデータに含まれることもあるが、ここでは特に数値的なものを主に扱う。」（文部科学省，2018，p.55）と示されているのみで、データの変数の数や規模については触れられていません。また、統計指導の多くが1変数，2変数程度しか扱われておらず、ビッグデータの利活用の拡大等の社会的背景を考慮した場合、多変数からなる多変量データの取り扱いを強化することが求められています（青山・小野，2016）。

Kazak et al.(2023)は、Lehrer & English(2018)のデータモデル化アプローチ（図1）を基に、多変量データを活用した統計的探究の重要性を指摘しています。この図の上部では問題の提示や属性（変数）の生成と選択、標本の生成といった統計調査を計画し実行するために相互に必要な概念が含まれています。図の下部ではデータの整理や構造化、測定、表現などのばらつきをモデル化して推論を行うプロセスが示されています。これらのデータモデル化アプローチにおける重要な考え方として、個々の値に着目するのではなくデータを一つの集計値として捉え推論することを挙げています。つまり、グラフ等を活用してデータのばらつきを可視化したり、統計量を用いてばらつきを指標化したりすることが大切になるということです。

さらに、データモデル化アプローチでは手元のデータに基づいて推測を行うことが強調されています。統計学における推測とは一般的に、標本データに基づいて母集団またはそのプロセスに関する結論を、仮説検定や区間推定などの方法を用いて導き出すことを指します。

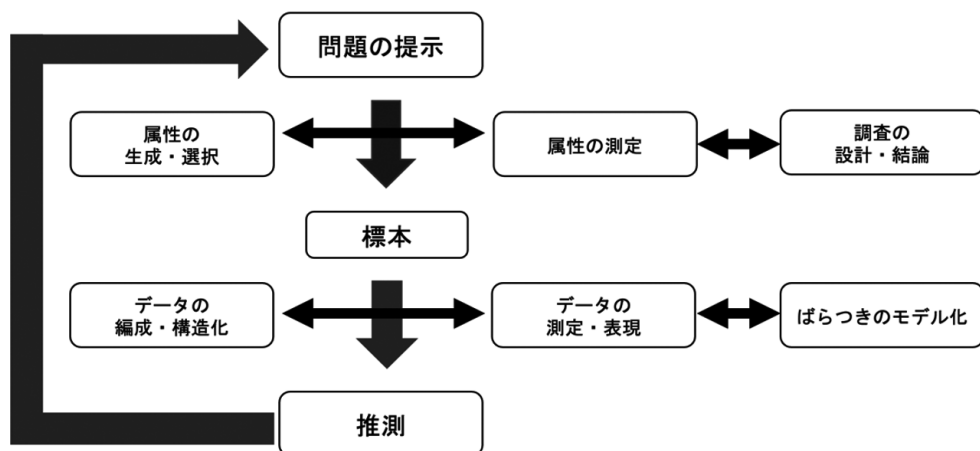


図1. データモデル化アプローチの構成要素 (Lehrer & English, 2018, p.232)

一方で、近年の国際的な統計教育ではインフォーマルな統計的推測 (Informal statistical inference) と呼ばれる、上述のフォーマルな統計的推測の方法を必ずしも用いずに、手元にあるデータの特徴や傾向をもとに推測する方法を取り入れた統計指導が注目されています (Makar & Rubin, 2018)。具体的には、① 推測の根拠としてデータを使用すること、② データを超えて一般化を行うこと、③ 統計的推測を行う際には不確実性を明確化することの3つが、インフォーマルな統計的推測のプロセスに含まれています。これらのプロセスによる統計指導を行うことによって、データからの推測や予測には不確実性が伴うことの理解を促すことが求められています。

よって、データモデル化アプローチは、ばらつきの測定や統計的推測といった統計学の活用に必要な基礎的な概念や考え方を学ぶために必要不可欠なものです。

3. 教材：十種競技データの問題

本稿では、Hosoda (2022) の「十種競技データの問題」の題材をもとに、中学校の統計指導で扱うことができる教材例をご紹介します。十種競技データの問題は以下の通りです。

「キング・オブ・アスリート」を決めると言われる陸上十種競技は、2日間で短距離走、中長距離走、跳躍、投擲（とうてき）に関する10種目の競技を行い、各記録を得点に換算し、その総合得点を競うスポーツです。

そこで、2021年に行われた東京オリンピックのデータをもとに、総合得点が高い選手にはどのような特徴があるのか調べてみましょう。

十種競技データの問題で扱うデータは、当時の出場選手23名の競技ごとの記録（100 m 走、走り幅跳び、砲丸投げ、走り高跳び、400 m 走、110 m ハードル走、円盤投げ、棒高跳び、槍投げ、1500 m 走）と総合得点から構成されています。データは International Association of Athletics Federations のウェブサイト (<https://worldathletics.org/>) で公開されています。このような第三者や機関が提供しているデータを是非活用して統計授業を設計してみてください。また、授業で扱えるソフトウェアについても一つをご紹介します。ここでは、Common Online Data Analysis Platform¹⁾ (以下、CODAP) と呼ばれる統計教育のソフトウェアを用いていきます。CODAP は、Web ブラウザを通して誰でも無料で使用することができ、小学

校で学習するドットプロットやヒストグラム、そして中学校で学習する箱ひげ図などの作成をドラッグとドロップで簡単に作成できることが特徴的です。本教材は、このような ICT を活用しながらデータを探索的に分析し、総合得点が高い選手の特徴を、生徒が主体的に考察することをねらいとしています。

4. 各種目のばらつきの測定

中学生が十種競技データの問題に取り組んだ際に、どのような統計的探究が起こり得るのかを具体的に考えてみましょう。まずは各種目の記録のばらつきについて考察することが考えられます。ここでは、総合得点の上位 5 名の選手に着目して特徴を考えていきます。

まず、ドットプロットや箱ひげ図による可視化を通して、各種目の記録のばらつきについて考察します。すると、100 m 走や 110 m ハードル走の競技において、総合得点が高い 5 名の選手は、それぞれの競技でも上位の記録を残していることが分かります（図 2）。

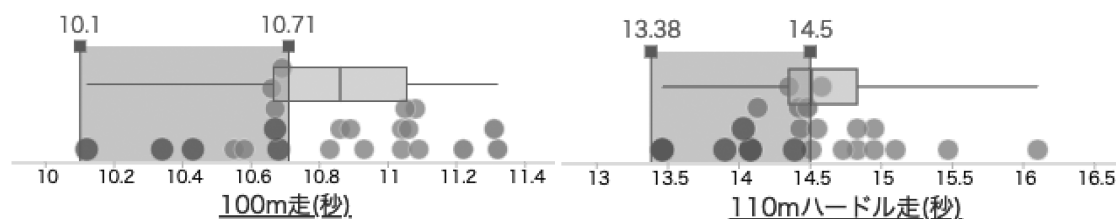


図 2. 100 m 走と 110 m ハードル走のドットプロット及び箱ひげ図

一方、円盤投げや砲丸投げの競技では、総合得点が高い選手であっても上位の記録とは限らないことが分かります（図 3）。加えて、各競技における 5 名の選手の記録の範囲を可視化することで、代表値や四分位範囲とあわせて選手の特徴をより詳細に明らかにすることができます。

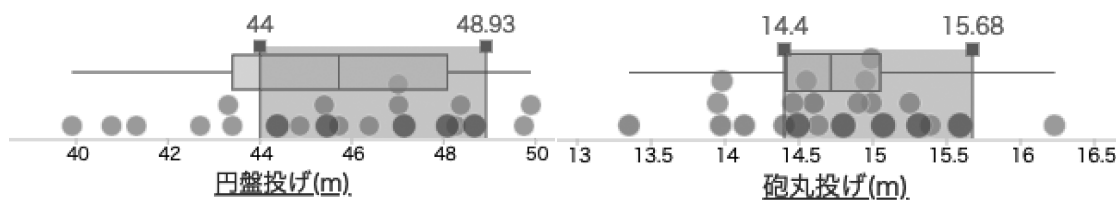


図 3. 円盤投げと砲丸投げのドットプロット及び箱ひげ図

5. 種目と総合得点の関連性

次に、各競技の記録と総合得点との関係に着目して考察していきます。ここでは先ほどの、100 m 走、110 m ハードル走の記録と総合得点の関係をグラフで可視化してみます（図 4）。

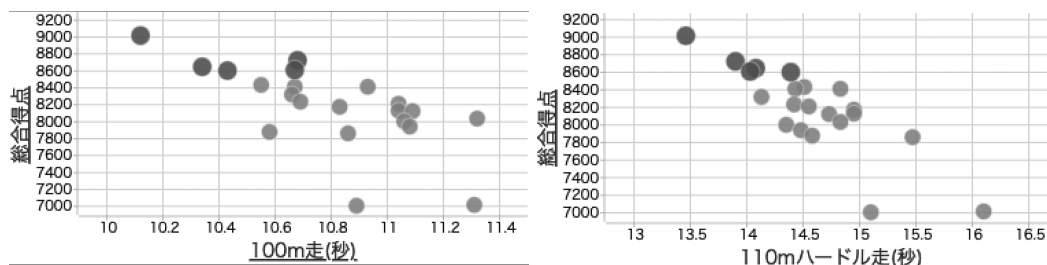


図 4. 100 m 走と総合得点及び 110 m ハードル走と総合得点の散布図

図 4 のグラフは散布図と呼び、本来であれば高等学校の相関の学習指導で扱われますが、実

データを活用した授業であれば中学校でも扱うことは十分可能であると考えます。このグラフから、100 m 走や 110 m ハードル走の記録が良い選手ほど総合得点が高い傾向があることが読み取れます。したがって、総合得点が高い 5 名の選手の特徴として、100 m 走や 110 m ハードル走を得意としていることが分かります。このように、特定の 1 変量や 2 変量の特徴について箱ひげ図や散布図を活用して探索的に分析することは、これからのデータ駆動型社会におけるあらゆる分野や領域で必要となります。さらに、今回の分析で明らかになった特徴はあくまでも東京オリンピックのデータのみに限定されているため、決定論的に結論づけることができないこと（不確実性が伴うこと）を理解することも重要になります。

6. おわりに

十種競技データの問題は中学生が仮説を立てて多変量データから変数を選択し、探索的に分析することができる教材であり、統計的探究に主体的に取り組むことができるでしょう。今年はパリオリンピックも開催される予定ですので、様々なデータを通してスポーツに関する興味や関心を深めてみてはいかがでしょうか。

註

1) CODAP は次のリンク (<https://codap.concord.org/>) からアクセスが可能です。またサイトのヘルプから日本語のマニュアルがダウンロードできます。

文献

青山和裕・小野浩紀 (2016). 多変数を扱う小学校算数での統計授業について：統計的探究プロセスによる授業構想と多変数による授業の広がり. 日本数学教育学会誌, 98(8), 3-10.

https://doi.org/10.32296/jjsme.98.8_3

Hosoda.K. (2022). Principles of task design to promote informal statistical inference in secondary school: Covariational reasoning using multivariate data. *Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022)*.

<https://doi.org/10.52041/iase.icots11.T14A4>

Kazak, S., Fujita, T., & Turmo, M. P. (2023). Students' informal statistical inferences through data modeling with a large multivariate dataset. *Mathematical Thinking and Learning*, 25(1), 23-43. <https://doi.org/10.1080/10986065.2021.1922857>

Lehrer, R., & English, L. (2018). Introducing children to modelling variability. In: Ben-Zvi, D., Makar, K., Garfield, J. (Eds.), *International handbook of research in statistics education* (pp. 229-260). Cham, Switzerland: Springer International Publishing.

https://doi.org/10.1007/978-3-319-66195-7_7

Makar, K., & Rubin, A. (2018). Learning About Statistical Inference. In: Ben-Zvi, D., Makar, K., Garfield, J. (Eds.), *International handbook of research in statistics education* (pp. 261-294). Springer, Cham. https://doi.org/10.1007/978-3-319-66195-7_8

文部科学省 (2018). 中学校学習指導要領 (平成 29 年度告示) 解説 数学編, 日本文教出版.

(目白大学 専任講師)