イプシロン　2019. Vol.61，141 － 146

Web 上のリソースを扱った統計教材の開発
―プログラミングを用いた大規模データの分析に焦点を当てて―
＜修士論文要旨＞

愛知教育大学大学院　教育学研究科
数学教育専攻　数学科教育学領域
216M054　**早　川　和　希**

論文構成

第１章　本研究の目的

ビッグデータ, IoT などと言われる時代になっていく中で，統計の重要性は高まっている．次期学習指導要領においても，扱う内容なども拡充されるとともに，その指導においてテクノロジーを利用することが必要になることが指摘されている．

統計的なデータを取得する源の一つは,測定データである．別の種類のデータとして，インターネット上のデータもさまざまな形で利用されている．

現在の統計教育の指導の中で扱われているデータは, (1) 実際に調査する (2) 既存のデータセットを使うことが中心になっているが，統計的な探究を深めていく上で，将来的には,データの収集に関して Web 上のリソースを扱い探究することも必要になるのではないかと考える．

本研究では, Web 上のリソースを使うことで統計的探究を深めていく教材開発の可能性について取り組んでいくことが目的である．

第２章　統計教育の現状と課題

現行の学習指導要領解説から，高校数学 I では

「データの分析」が必修となった.この流れには，社会での統計の重要性が増していることがある.渡辺（2014）も，「データ活用能力自身がひろく一般の国民に涵養されるべき基礎コンピテンシーになったと考えるべき」と述べており，社会の中で生きていくために必要な知識とされている.さらには，新学習指導要領より小学校では，「データの活用」という領域が新設され，小学校 4 年時よりデータの活用の目標の中で「問題解決」という文言が加えられている.つまり，知識技能ベースの学びから問題解決のための統計的探究が重要視されていくことは明らかである.

統計的な問題解決をするにあたり，青山(2015)が，「1 変数や 2 変数を収集し，処理するだけで統計的な探究や問題解決が完結するとは考えにくい」と述べているように，多変数の分析やより大規模なデータの分析が小学校段階より行われる可能性を示唆している.増田(2016)は，高校生向けに多変数のデータセットを用いた授業を行い，高校生が多変数の大規模データセットを扱い統計的探究プロセスを通じて問題解決をすることは可能であるということを明らかにした.また，青山(2015)は，小学校 3 年生を対象とした多変数データの分析の授業実践から，多変数の分析が，必要なデータカードを使うなど必要な手立てを行えば，可能であることを示唆した.このことより，大規模なデータセットを用いて授業することは，可能であり，次期学習指導要領改訂に向け重要視されていくものであると考える.つまり，大規模なデータセットを教員側が用意することも求められる可能性はある.ただ，分析を行えるような大規模なデータを作成することは容易ではない.そこで，大規模なデータセットを作成・教材化することを考えた.その中で，本論文ではデータの収集にあたり，Web 上のデータリソースの収集に注目し，次のように研究課題を設定した.

① Web 上に存在するデータに関する整理と教材化可能性の調査
② インターネット上のリソースデータを収集するソフトの開発を行う.
③ 実際に得られたデータの教材化可能性を探る

第３章　Web 上のリソースを用いたコンテンツの開発の可能性

３－１ Web 上に存在するデータの分類

Web 上にあるデータを以下のように分類・整理した.分類の観点としては，

・静的であるか動的であるか
・加工データとして存在しているか

に注目し，以下のように分類をした.データベースは，数量的なデータがあるような Web サイトとしてとらえた.電子書籍のようなものはテキストデータとしてとらえている.

	静的	動的
データベース	(1) MLB の試合・選手データ，国勢調査等	(2)Google Trend，株価データ等
テキストデータ	(3) ニュースやブログ，電子書籍データ等	(4) SNS サービスの現在のトレンド紹介機能
その他のデータ	(5)静止画等 例：顔認証	(6)動画等 例：防犯カメラ映像

表１　Web 上にあるデータの分類

３－２　本研究で扱う Web 上のデータの焦点化

(1)に関しては，データを取り出す作業は比較的容易に行うことができる.酒折他(2010)により「科学の道具箱」という複合デジタル教材が開発されており，csv 形式のファイルのデータセットを統計的処理する環境は整っている.逆に，(3)のようにテキストデータに対するデータ収集は行われていない.そこで，(3)に注目し，データを収集できないかと考え，その足掛かりとして，書籍の中の「文字」に注目し，教材化の可能性を考察した.

３－３　大規模なテキストデータを扱うことによる探究と実現するためのソフト群についての明確化

３－３－１　教科書にある教材例

教材化するにあたって，教科書の中でテキストデータを扱った教材はないか調査した.その結果教育出版の中学数学 3 の教科書に図 2 のような教材があることが分かった.この教材は，明治時代の作品と最近の文学作品との比較を漢字の使用率とひらがなの使

用率を題材として標本調査を通して調べるという課題である．１０ページを無作為抽出し調査している．

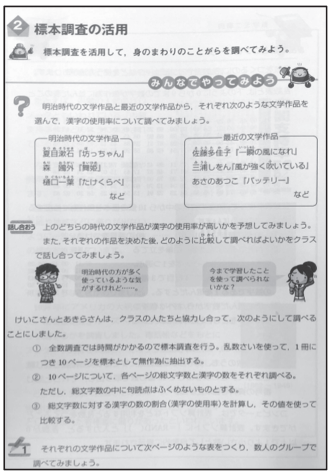

図１　教育出版　中学数学３（p. 216）

この教材は，標本調査の必要性を実感する上では適している．手作業で行ってみると，大学院生の筆者でさえ，1 ページあたり 10 分かかった．比較的短編である「坊っちゃん」すべてについて調べるのでさえ丸 1 日の時間が必要となる．その状況下であるため，この教材は，一定の誤差があるとしても，短時間に結果を知る上で，標本調査を行うことが有効であることを実感することができ，その概念と方法を知ることができる．

しかし，逆に言えば，手作業だけではこの教材の主なねらいは，標本調査の必要性と概念，手順を習得することにとどまらざるを得ない．それを解決する方策として，テクノロジーを利用する方法を示す．

３－３－２　教材開発に向けてのソフトの利用方法の考案

（1）標本を複数回抽出することによる標本平均等の基礎的理解のための教材

1 ページの漢字の使用率を手作業でクラス 40 人で調べたとしても 40 ページを 1 回きりしか調べられない．標本をとる回数を増やせば，誤差がどの程度なのかを体感的に学ぶことができる．発展的には，95% 信頼区間などの理解に結びつけることができる．さらに，標本数を増やせば誤差は減り，より精度が向上していく大数の法則のような学びも行える．

（2）扱う素材に則した探究を目的とした教材

漢字の使用率のように，1 つの文字をデータとしてとらえ，国語科的な探究も行える．文章を題材とする探究の可能性を考察し以下にまとめた．

① 「坊っちゃん」の全文での分析
② 夏目漱石の全作品を分析
③ 著書が，夏目漱石と他の作家では違い
④ 単語の出現頻度の検索

以上の作業を行えるソフト開発をし，文章に対する探究の可能性を考察していく．

第４章　青空文庫を利用したソフト開発

ソフト開発にあたり，Web サイト上で分析できるソフトであれば児童生徒が扱える．そこで HTML＋Java Script 言語を使い考察する．

４－１　HTML＋Java Script でのソフト開発

４－１－１　入力と出力

文章の分析をかけるソフトを作るには，まず基本的な枠組みとして

① 文章を打ち込むためのスペース
② 打ち込んだ文章に対し分析をかけるボタン
③ 分析結果が表示されるスペース

が必要となるため，開発した．

４－１－２　文章の入力と文字のカウント

前節を踏まえ文字の入力とカウントのプログラムを開発した．実行した結果が図 3 である．文章は，無料の電子書籍のデータベースの青空文庫の html データからコピー&ペーストして得ている．実行して結果はうまくいっているように見えるが，問題点が１つある．本文にある漢字には「ルビ」が振られ，「親譲おやゆずり」というように表示されてしまうことだ．つまり，原本のデータからルビを除き，テキストデータ化する必要がある．

そこで，ルビを含めより高度な処理をするために「python」言語を用いる．インターネット環境さえあれば Web サイトを直接開くこともなく html データを取り出すことができるためだ．

図２　実行結果

４－２　python でのソフト開発

「python」では，Web 上のデータの処理のためのライブラリが存在する.それを用いて,プログラムによる文章データの取り出しと，ルビの問題を解決した.「Beautiful Soup」というライブラリを用い，60 行ほどのプログラムでルビを取り除くことができた.次に，４－１同様文字を区切り，分析をする.

４－２－１　特定の作品に対する分析

文字の出現頻度によって整理できるようにした.

この作業にかかるのは３秒ほどである.これによりURL さえわかればすぐに分析ができる.

４－２－２　特定の作家に関する全作品の分析

同作家の他の作品もまとめてより大きなデータを分析していきたい.自動化するためには，Web サイトのリンクの構造を理解しないことが分かった.Web サイトをつなぐ構造を把握し，インデックスから作家を選び，個々の作品を分析できるようなシステムを構築した.200 行ほどのプログラムとなり，夏目漱石であればおおよそ 2 分程度で分析できた.

４－２－３　単語の分析

1 文字ずつに区切り，ソートし分析できる形からさらに発展させていきたい.そこで，指定した単語が，文章中に何回出るかを確認できるソフトを開発した.

４－２－４　サンプリング

出発点とした標本調査では，手作業であるため，時間がかかる.そこで，サンプリングを文字数，回数を決めてサンプリングできるようにした.「坊っちゃん」でひらがなの割合を４００文字ずつの標本で検証した結果約 66%であった.

第５章　Web コンテンツを利用した探究事例

５－１　夏目漱石の作品による分析

先述した教科書の問題では，夏目漱石の「坊っちゃん」を題材としていた.坊っちゃんのデータを４章で作成したソフトを用いて分析を行った.結果は「い」が一番多かった.坊っちゃんの中では，このような結果が得られたが，これより文章は，「い」最もが多くなるということは言えない．坊っちゃんだけでなく，夏目漱石の他の作品にも注目して，データをとってみることを考え,「吾輩は猫である」で調べると「の」が一番多いことが分かった.さらに，夏目漱石の全作品では何が最も使われているのかを調査すると，「の」が圧倒的に多い結果となった.

５－２　作者別の比較

夏目漱石で得られた結果は，一般的なのかを検証すべく，他の作家との検証を行う.

まず，青空文庫にある中で，作品数が多い芥川竜之介を例として調査した.その結果を夏目漱石の結果と合わせて下の表にまとめた.

	夏目（ひらがな）	芥川（ひらがな）	夏目（漢字）	芥川（漢字）
1	の	の	人	一
2	い	に	事	人
3	た	た	見	見
4	に	は	分	云
5	て	い	一	彼

表２　文字の出現頻度の順位

表２より，共通して「の」が多い.そして，上位５つで見ているが，ひらがなは，4 文字が，漢字は 3 文字が共通している．芥川と夏目の文章では，使われる文字に関して共通点も多くあることがわかる.「人」や「事」を使うことは共通して多く，小説を書く際には，登場人物や出来事を描写することが多いため，多く出てくるのではないかと考えられる.

５－３　単語の出現頻度による分析

ここまで一文字ずつ区切り分析することに焦点を当ててきた.しかし，それだけでは限界があるため単語の出現数の分析をした.

（１）吾輩は猫であるの文中に，「吾輩」は何回出てくるのか

図 4 を見ると吾輩は猫であるの文章の中で，「吾」と「輩」がほぼ同数用いられているのがわかる.

('お', 499), ('吾', 484), ('思', 479), ('間', 477), ('生', 460), ('輩', 449), ('ご', 401), ('時', 399), ('方', 395), ('子', 367), ('行', 363), ('中', 357),

図４　「吾輩は猫である」の検索結果の一部抜粋

タイトルにもあるように登場する主人公の猫の一人称が「吾輩」でもある.それゆえに，この「吾」と「輩」は，吾輩という単語のために何回使われているのかという疑問に対し検証した．調べてみると，447 回使われていることが分かった.つまり，全文の中で「吾」は 484 回，「輩」は 449 回あるので，ほとんどが吾輩という単語のために使われているということがわかる.夏目漱石の全作品では，「吾」は 1041 回，「輩」は 688 回，また「吾輩」は 540 回用いられていた.

このことから「吾輩」という言葉を夏目漱石は頻繁に使っていないことがわかる.それに対して，夏目漱石は猫に対して何か特別な思いを抱き「吾輩」という言葉を最適な表現だと捉え用いたと考察できる.

（2）「一」と「二」と「人」からの考察

芥川と夏目の分析をした際に，「一」，「人」が共通して多く存在した.これより，「一人」という単語を連想した.漢数字と「人」に注目してみると，表 3 のようになった.これより，「一」の出現比率に対して，夏目は芥川よりも極端に多く漢数字の「二」を使っていることがよくわかる．そこで，芥川と夏目の漢数字の使う目的に関して「人」という字に注目して考察を行ってみた.その結果が表 4 のようになった.

この結果より，夏目漱石は，「一人」という言葉よりも，「二人」という言葉を多く使い，芥川竜之介は「一人」という言葉を「二人」問言葉よりも使っていることがわかる.さらに，その 2 語の比率を見れば，明らかな差が出ていることがわかる.書籍の中で，芥川は個人に対する描写が多く，夏目は複数人に関する描写が多いのではないかという考察ができる.

	夏目漱石	芥川竜之介
一	10393 語（0.3%）	9030 語（0.5%）
二	4817 語　(0.15%)	2896 語（0.15%）
人	13210 語（0.4%）	8885 語（0.5%）
文字数	3016374 語	1840101 語

(　)内は全文字数に対する比率

表３　夏目漱石と芥川龍之介の比較①

	夏目漱石	芥川竜之介
一人	１１２０（10.78%）	１１８８（13.16%）
二人	１３２７（27.55%）	６０１（20.75%）

表４　夏目漱石と芥川龍之介の比較②

５－４　児童文学と随筆等のジャンルの違いに注目した比較

これまで，小説家である夏目漱石と芥川竜之介に関して注目し探究を行ってきた.次に異なるジャンルの作家も考察していきたい.そこで，随筆作家である寺田寅彦，児童文学作家である新美南吉，和算の研究家である三上義夫を抽出して取り上げた.またそれを csv ファイルとして保存できるような形にした.

まず新美南吉に関して，上位 50 語の中に漢字は「人」が唯一含まれているのみで，残りは句読点やひらがな，かたかなである.児童文学では，漢字の使用率がほかの作者に比べて少ないことが見て取れる.

次に，鍵括弧の量である.夏目漱石，芥川竜之介，新美南吉はともに上位 50 位内にランクインしているが，寺田寅彦と三上義夫はランクインしていない.これは，会話文の多さに差があることが数値から考察することができた.また，句読点の数に注目すると，寺田寅彦や三上義夫は他の 3 人に比べ順位が低いことも見て取れる.そこで，一文の長さを調べると表 5 ようになった.

	寺　田 寅彦	芥川竜 之介	夏　目 漱石	三　上 義夫	新美南 吉
句点	52 語	36 語	36 語	54 語	32 語
読点	57 語	26 語	37 語	33 語	23 語

表５　句点読点に注目した作家別の比較

小説家と随筆家との間には大きな差がみられた.また，新美南吉が一番文の長さが短く，読点も多く使っていることがわかる.これは，小さい子供でも読み

やすくしてあると考察できる.また, 同じ小説家の芥川竜之介と夏目漱石を比べても, 読点の数は大きく異なっている.これより, 作者によっても特徴がみられることがわかる.

５－５　考察

これまで, テキストデータに対する算数・数学科における教材化は, 教科書の中では標本調査のやり方を学ぶことにしか用いられてこなかった.それは, 調査するには膨大な時間がかかってしまっていたことが原因として挙げられる.今回のソフト開発により, 短時間で大量のデータを一気に入手することができ, その結果様々な考察を行うことができた.

６章　まとめと今後の課題

本論文の課題は, 以下の３点であった.

① Web 上に存在するデータに関する整理と教材化可能性の調査
② インターネット上のリソースデータを取り出すソフト開発を行う.
③ 実際に得られたデータを探究し教材化可能を探る

①に関して, Web 上にあるデータのうち, 今回は静的なテキストデータに注目して考察を行った.その結果, 無料で扱える電子書籍のデータベースの「青空文庫」の利用が適していると考え教材化可能性を考えた.その結果, 教科書の事例を参考に教材化の可能性を見出し, プログラミングを利用することによって必要なソフトを開発することができた.

②に関して, ソフト開発を行い, 様々な統計的処理を行えるようにし, データを収集するためのツールを作成することができた.

③に関して, 実際に探究することで文章の構造に対する理解が深まり, 新たな発見をすることができ, そこからまた新たな疑問, そして追究するといったプロセスの過程を体験することができた.

今後の課題として, 今回はテキストデータに注目したが, それ以外の Web のデータに関してもプログラミングを通してデータセットの作成, 探究を行っていきたい.また, 授業としての位置付けを明確化し, 今後の教育に役立てていきたい. また, 5 章で考察したソフトは, python 上で動かしているだけに過ぎない.これを児童・生徒が使えるようにするためには, このソフトをパッケージ化する, もしくは Web 上で動かせるようにすることが必要となる.今回はそこまでに至れなかったため今後の課題とする.今現在考えているのは, 先ほど挙げた後者にあたる Web 上で動かせるデバイスとすることである.そのためには, 4 章の前半で用いた, html 言語と python 言語を組み合わせる必要がある.そのためには CGI という技術が必要ということまではわかっている.その　つなぎ合わせを行い, 世に広く使えるようなソフトとしていきたい.

参考引用文献

青山和裕(2015)　小学校指導における多変数データ利用について　イプシロン vol.57, pp.39-50
青山和裕・柗本新一郎(2015)　ニュージーランドの統計指導　日本数学教育学会誌, 第 97 巻, 第 7 号, pp.13-22
酒折他(2010)　統計的思考力育成のためのデジタル複合教材 : 数学科で活用する『科学の道具箱』　日本数学教育学会誌, 第 92 巻, 第 1 号, pp.20-28
渡辺美智子（2013）知識基盤社会における統計教育の新しい枠組み 日本統計学会誌, 第 42 巻, 第 2 号, pp253-271
渡辺美智子(2014) 不確実性の数理と統計的問題解決の育成 - 次期学習指導要領改訂に向けて－　日本科学教育学会誌, 第96巻第1号, pp.33-37
瀬沼花子(2004)　企業の算数・数学教育への期待　データに基づく予測と論理的思考力の協調と指導法の改善　科学教育研究, 第 28 巻, 第 1 号
増田朋美他(2016)　多変数の教材「ウイニングイレブン」を使ったデータの分析－学統計から使う統計のための教材開発－　研究紀要, 愛知教育大学附属高等学校, 第 43 号, pp.53-67
澤田利夫(2012)　中学数学３　教育出版
青空文庫　https://www. aozora. gr. jp/　（2019. 2. 10 最終確認）
文部科学省(2018)　小学校学習指導要領解説算数編