

統計データを記述する例文収集の試み ーデータを説明する能力の育成を目的としてー

Collecting Example Sentences that Describe the Statistical Data -for the purpose of improving the ability to explain data in text-

吉根 勝美*¹

Katsumi YOSHINE*¹

*¹ 南山大学経済学部

*¹ Faculty of Economics, Nanzan University

Email: kyoshine@nanzan-u.ac.jp

あらまし：すべての大学生に統計教育を実施する必要性が高まる中、特に社会科学系の学生には、実社会に結びついたデータを用いることが望ましい。統計教育では、単にデータ分析ができるようになるだけではなく、実際のデータを文章で説明する能力の育成も重要だ。その能力を育成する教育をするには、例文となる文章が有用となるだろう。本稿では、行政機関が公表する白書から、実データについて説明している文章の抽出を試み、ある1編の白書から例題となりうる文章をある程度は収集できることを確認した。
キーワード：統計教育、実データ、例文、白書

1. はじめに

近年の統計教育では、実際の社会を反映する実データの必要性が高まっている。文理を問わず全ての大学・高専生が修得すべき「数理・データサイエンス・AI（リテラシーレベル）」の教育プログラム認定制度では、「実データ・実課題（学術データ等を含む）を用いた演習など、社会での実例を題材として、『データを読む、説明する、扱う』といった数理・データサイエンス・AIの基本的な活用法に関すること」を審査観点の一つとして挙げている⁽¹⁾。

特に社会科学系の学生には、実際の統計データによる指導が望まれる。指導の内容には、データ集計・分析の手法、ソフトウェアの操作方法と並んで、データから読み取れることを文章で説明するレポートの作成も含まれる。

各行政機関が公表する白書は、実データに基づいたレポートの典型である。例えば、令和元年度年次報告（公務員白書）⁽²⁾では、集計したデータを図1のようにグラフ化し、これを引用しつつ「海外勤務等の国際経験を有している本府省の幹部職員の割合を見ると、各府省で異なるものの、ほとんどの府省で過半数となっている（図1-1）。」と述べている。

ただし、同じ実データだから同じ文章になるとは限らない。前述の文章では、国際化が進んでいるという肯定的な結論が予想されるが、同文章の後半を「7割以上の府省はまだ半分程度に過ぎない」と変えれば、否定的な結論に導くこともできる。

すなわち、データを説明する文章のテンプレートによる指導は不十分であり、学生には数多くの文章に接する機会が必要である。そのためには、様々な分野から実際の文章を集めた例文集が必要である。

総務省が運営する e-Gov ポータル⁽³⁾によると、現在、28 機関から 52 編の白書が公表されている。そこで本稿では、様々な分野にわたる白書からデータ

を説明する文章を収集することで例文集の作成可能性を検討するため、令和元年度の公務員白書⁽²⁾からデータを説明する文章の抽出を試みる。

図1-1 海外勤務等の国際経験を有している本府省の幹部職員の割合

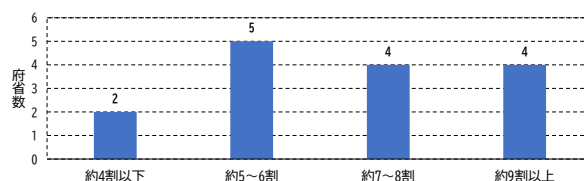


図1(2)のp.21に掲載されている棒グラフの再現

2. 例文を抽出する手順

HTML 形式の令和元年度公務員白書の本編から、p 要素をすべて取り出し、p 要素の内容に句点「。」を含む場合に限り、p 要素の内容を「。」で分割する。ただし、(….) や “…” のように、括弧や引用符の内側の末尾の「。」では分割されないようにする。

こうして得られた 1,213 文から、「図*-*」、「表*-*」、「図*」、「表*」（*は数字）という文字列を含む 83 文を抽出し、『図表を引用している文章』とした。

次に『図表を引用している文章』に多く現れる単語を特徴語として抽出し、もとの 1,213 文について特徴語を含む文章を『データを説明している文章』としてあらためて検索する。この手順を踏むのは以下の 2 つの理由による。1 つは、図表を引用しているが、それが組織図や年表のようにデータではない可能性があるから。もう 1 つは、図表を引用していないが、データを説明する文章である可能性があるからである。

3. 例文の抽出結果

3.1 特徴語リストの作成

公務員白書から抽出した『図表を引用している文

章』83文の特徴語上位40語の出現回数を表1に示す。データの説明に使われそうな「比べる」「前年度」「件数」「減少」「多い」「増加」「割合」「内訳」という単語も見られるが、それ以外の単語の方が多い。

そこで、まったく別の白書について同様の処理を行い、2つの白書で共通して現れる特徴語を使って、データを説明する文章の抽出を試みる。今回は、内閣府「令和2年版交通安全白書」⁽⁴⁾のHTMLファイルから1,933文を抽出し、そのうち『図表を引用している文章』123文から特徴語を抽出した。

2つの特徴語リストを名寄せし、両者の出現回数のうち小さい方で降順に並べ替えたものを表2に示す。前述の単語の多くが上位に上がり、上位40語にはなかった「占める」「傾向」「推移」「最も」もこのリストに現れ、表1よりも『データを説明している文章』の特徴語リストらしくなっている。

表1 特徴語上位40語の出現回数

(1)表	53	(11)試験	14	(20)行政	10	(25)民間	8
(2)図	38	(11)法人	14	(20)女性	10	(32)給与	7
(2)年度	38	(13)機関	13	(20)調査	10	(32)公務員	7
(4)職員	36	(13)比べる	13	(24)資料	9	(32)増加	7
(5)平成	27	(15)前年度	12	(25)見る	8	(35)割合	6
(5)令和	27	(16)月	11	(25)公務	8	(35)研修	6
(7)採用	25	(16)件数	11	(25)災害	8	(35)国際	6
(7)派遣	25	(16)行う	11	(25)多い	8	(35)執行	6
(9)状況	23	(16)年	11	(25)男性	8	(35)対象	6
(10)実施	22	(20)減少	10	(25)府	8	(35)内訳	6

表2 名寄せした特徴語リストと出現回数

a 欄：公務員白書, b 欄：交通安全白書
c 欄：上2つのうち小さい方（降順）

	a	b	c	(10)実施	22	7	7
(5)令和	27	51	27	(35)割合	6	30	6
(5)平成	27	24	24	(16)月	11	6	6
(13)比べる	13	14	13	(41)事業	5	6	5
(16)年	11	35	11	(53)占める	4	36	4
(16)件数	11	26	11	(53)傾向	4	21	4
(20)減少	10	38	10	(53)推移	4	13	4
(2)年度	38	10	10	(53)最も	4	11	4
(16)行う	11	9	9	(53)基本	4	7	4
(25)多い	8	28	8	(53)期間	4	7	4
(25)見る	8	16	8	(20)女性	10	4	4
(32)増加	7	19	7	(35)対象	6	4	4
(9)状況	23	7	7	(41)現在	5	4	4

3.2 特徴語による文章の点数化

次に、1,213文それぞれについて、表2の特徴語を含む文章を検索する（「令和」「平成」は除く）。実際には、特徴語が含まれている文章に、特徴語ごとに表2のc欄を点数として加算する。したがって、点数が高い文章ほど『データを説明している文章』である可能性が高い。図表引用の有無別に、1,213文の点数分布を図2に示す。『図表を引用している文章』の点数の方が全体的に高くなるのは当然である。

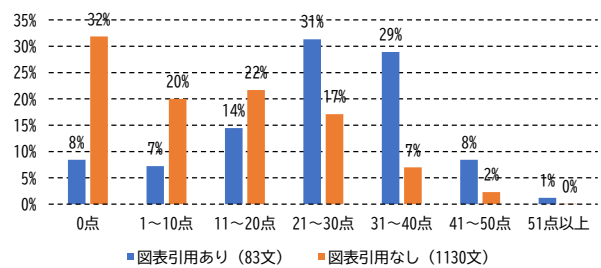


図2 文章ごとの点数分布

3.3 評価

『図表を引用している文章』にもかかわらず特徴語を全く含まない（0点）の文章7文について、『データ説明する文章』に該当するかどうか個別に確認したところ、引用している表が数表ではない文章1文および「…は表*のとおりである」というパターンの文章4文を確認した。したがって、7文中5文は、図表を引用しているが『データを説明する文章』ではないと判定できている。

一方、『図表を引用している文章』ではないが点数が高い文章を『データを説明する例文』とみなして、41点以上の文章27文について個別に確認したところ、前後の文章で図表を引用している文章14文および対応する図表はないものの確かにデータについて言及している文章8文を確認した。したがって、27文中22文は、図表を引用していないものの『データを説明する文章』であると判定できている。

4. まとめ

本稿では、実データを文章で説明する能力を育成するのにふさわしい例文を、行政機関が公表している白書から収集することを試みた。『データを説明する文章』によく現れる特徴語のリストを作成し、ある1つの白書からの文章抽出を試みたところ、単純な方法にもかかわらず、ある程度の文章抽出ができていることは確認できた。

今回の試みで、白書をもとにした例文集の実現可能性は明らかにできた。今後の課題として、文章の抽出手順および評価方法の再検討が挙げられる。さらに、実データを文章で説明する能力を育成する教育をどう実現すべきか検討しなければならない。

参考文献

- (1) 内閣府, “「数理・データサイエンス・AI 教育プログラム認定制度(リテラシーレベル)」の創設について,” https://www.kantei.go.jp/jp/singi/ai_senryaku/suuri_datascience_ai/pdf/ninteisousetu.pdf, (2020)
- (2) 人事院, “令和元年度年次報告書(公務員白書),” <https://www.jinji.go.jp/hakusho/pdf/> (2020)
- (3) 総務省, “白書等 | e-Gov ポータル,” <https://www.e-gov.go.jp/about-government/white-papers.html>
- (4) 内閣府, “令和2年交通安全白書,” https://www8.cao.go.jp/koutu/taisaku/r02kou_haku/zenbun/ (2020)