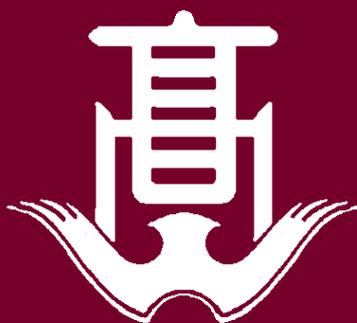


令和 3 年度

# 「データサイエンス教育」の開発

【令和 3 年 4 月～令和 4 年 3 月の取組】



文部科学省スーパーサイエンスハイスクール指定校  
兵庫県立姫路西高等学校

## 目 次

- ・「データサイエンス教育」の概要
- ・資料編
- ・生徒研究成果事例

○ 「データサイエンス教育」の概要

1. カリキュラム開発・・・・・・・・・・・・・・・・	1
2. 統計的探究プロセス PPDAC サイクル・・・・・・・・	2
3. 学校設定科目の授業内容と構成例・・・・・・・・	3
4. 評価の方法・・・・・・・・・・・・・・・・	5



## 【 「データサイエンス教育」の概要 】

「データサイエンス教育」(以下、DS教育)は、本校のSSH事業において、国際理学科・普通科の全生徒が学校設定科目として受講する、新時代を見据えた独自の教育活動である。学習指導要領の改訂に伴い、各教科を横断的に学ぶことにより、効果を高める教育の実現を目指しカリキュラム開発を行っている。

SSH I 期目では、教科「数学」におけるデータの分析、統計的な推測、教科「情報」におけるデータの活用、情報とデータサイエンス(以下、DS)の分野を組み合わせた内容を学習するとともに、「総合的な探究の時間」における探究のプロセスを実践していくことで、DSを基盤とした探究活動を行う。

### 1. カリキュラム開発

本校における3年間のDS教育は、1年次を「探究準備期間」、2年次を「探究実践期間」、3年次を「探究展開期間」と位置づけ、表1のように単位数を設定した。なお、本校では、理数に関する学科である「国際理学科」と、「普通科」が設置されている。

表1 DS教育に関わる科目名と単位数

国際理学科	科目名	単位数
準備(1年)	データサイエンス研究	4
実践(2年)	データリサーチ研究	3
展開(3年)	グローバル研究	2
普通科	科目名	単位数
準備(1年)	データサイエンス探究	2
実践(2年)	データリサーチ探究	2
展開(3年)	グローバル探究	1

各学科において下記のように科目を代替し、3年間の教育課程を編成した。

#### (ア) 国際理学科1年【探究準備期間】

DS教育の基礎・基本を学ぶ学校設定科目「データサイエンス研究(4単位)」を実施し、それによって「社会と情報(1単位)」「課題研究(1単位)」「総合的な探究の時間(2単位)」の代替とする。

#### (イ) 国際理学科2年【探究実践期間】

DSを基盤とした研究実践を行う学校設定科目「データリサーチ研究(3単位)」を実施し、それによって「社会と情報(1単位)」「課題研究(1単位)」「総合的な探究の時間(1単位)」の代替とする。

#### (ウ) 国際理学科3年【探究展開期間】

DSの学びを振り返るための学校設定科目「グローバル研究(2単位)」を実施し、それによって「コミュニケーション英語Ⅲ(1単位)」「課題研究(1単位)」の代替とする。

#### (エ) 普通科1年【探究準備期間】

DS教育の基礎・基本を学ぶ学校設定科目「データサイエンス探究(2単位)」を実施し、それによって「社会と情報(1単位)」「総合的な探究の時間(1単位)」の代替とする。

(オ) 普通科2年【探究実践期間】

DSを基盤とした研究実践を行う学校設定科目「データリサーチ探究（2単位）」を実施し、それによって「社会と情報（1単位）」「総合的な探究の時間（1単位）」の代替とする。

(カ) 普通科3年【探究展開期間】

DSの学びを振り返るための学校設定科目「グローバル探究（1単位）」を実施し、それによって「総合的な探究の時間（1単位）」の代替とする。

## 2. 統計的探究プロセス PPDAC サイクル

平成30年度告示高等学校学習指導要領解説理数編における第1節数学Iの「3(4)データの分析」において、統計的探究プロセスが記載されている。統計的探究プロセスの5つの段階からなる「問題(Problem)－計画(Plan)－データ(Data)－分析(Analysis)－結論(Conclusion)」(以下、PPDACサイクル)に基づき、本校のDS教育の根幹を組み立てた。

### 問題 (Problem)

生徒が自分の興味関心のある研究テーマを決め、課題発見に向けて抽象的な内容から具体的な内容へと細分化・具体化していく。

### 計画 (Plan)

仮説を立てて、必要なデータを考え、仮説を検証するための分析計画を立てる。

### データ (Data)

オープンデータや観察・実験などにより実際にデータを収集する。収集したデータを表計算ソフトやPythonなどのプログラミング等の知識・技能を活かして、データを整理・整形する。

### 分析 (Analysis)

整理・整形したデータをもとに解析を行う。データ解析に関する知識や技能を活かしてデータを解析し、結論を導く。

### 結論 (Conclusion)

データ解析によって得られた結果を、先行研究やフィールドワークなどを通して実態と照らし合わせて考察する。

また、具体的な授業を実践していく上では、5つの段階を細分化し、各教科の内容を割り振りながら授業展開を作成した。PPDACサイクルを細分化した内容が次の図1である。

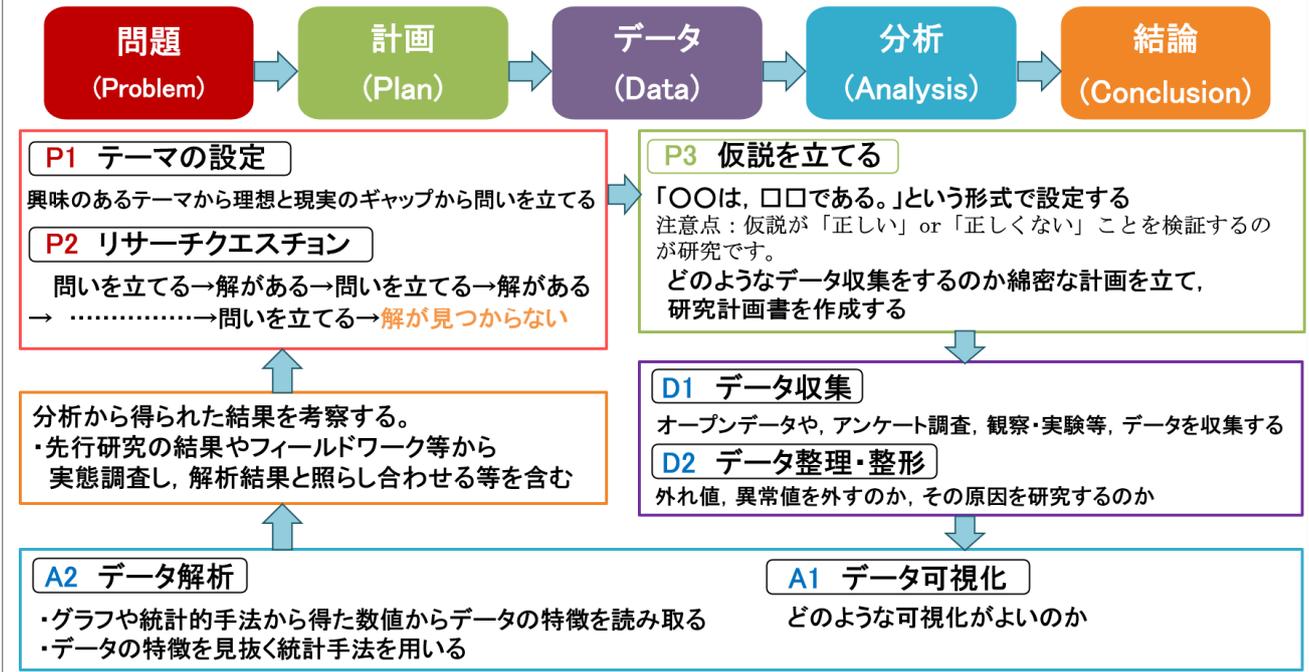


図1 PPDAC サイクルの詳細

### 3. 学校設定科目の授業内容と構成例

実施形態は、通常講座と集中講座を適宜実施している。通常授業は、週に1コマ実施し、集中講座は毎月特定の時間に集中的に実施している。

#### 3.1 「データサイエンス探究」の年間計画

- (1) 対象 普通科1年生（6クラス）
- (2) 教員の配置

通常講座は、3名の教員で担当し、それぞれ場面に応じた授業を実施している。

集中講座は、生徒240名に対して、12名の教員で担当する。1人の教員が4グループ担当し、研究指導を行っている。

表2 データサイエンス探究の年間計画

	通常講座	回数	集中講座	回数
4月 ～5月考査	データの分析 情報モラル・研究モラル MS-Excelの知識・技能	3	DSの概要	1
5月考査終了後 ～7月考査	【D（データ）】に関する講義・演習 MS-Excelの知識・技能 単回帰分析法の理解	7	上級生の研究発表による学び	1
7月考査終了後 ～夏季休業	統計ポスター作製演習	2	研究中間発表会(ポスター) 研究評価の振り返り	2 2

9月 ～10月考査	【A(分析)】に関する講義・演習	3	【P(問題)】に関する講義・演習	1
	西松屋ケースメソッド連携企業データ解析演習	1	【P(計画)】に関する講義・演習 データ分析小論文講座	1 1
10月考査後 ～12月考査	連携企業によるデータ解析PBL	6	データ解析I発表会	2
	【C(結論)】に関する講義・演習 情報デザインに関する講座	1 1	(スライド)	
12月考査後 ～3月	DS教育の基礎・基本の復習	2	2年生研究発表会への参加	3
	新価値創造講座	2	英語研究発表会	1
	次年度のオリエンテーション	1	【A(分析)】に関する講義・演習	1
	研究計画書作成講座	3	先行研究調査	1

### 3.2 「データリサーチ探究」の年間計画

(1) 対象 普通科2年生(6クラス)

(2) 教員の配置

通常講座は、4名の教員で担当し、それぞれ場面に応じた授業を実施している。

集中講座は、生徒240名に対して、12名の教員で担当する。1人の教員が7～8グループ担当し、研究指導を行っている。

表3 データリサーチ探究の年間計画

	通常講座	回数	集中講座	回数
4月 ～5月考査	【A(分析)】に関する講義・演習	1		
	回帰分析法の理解	3		
5月考査終了後 ～7月考査	標準化の理解	2	【P(問題)】に関する講義・演習	2
	統計的仮説検定の理解	2		
	クラスタリングの理解	2		
7月考査終了後 ～夏季休業	統計ポスター作製演習	1	研究中間発表会(ポスター)	2
			研究評価の振り返り	2
9月 ～10月考査	Pythonを活用したデータ処理	1	事例研究演習(計画)	
	推測的な統計の理解	3		
	考察・創造力について	1		
10月考査後 ～12月考査	ループリックに関する講義	1	事例研究演習(評価)	1
	アブストラクトに関する講義	1	事例研究演習(分析・考察)	1
	研究実践	4	研究実践	2
12月考査後 ～3月	メタ認知育成講座	4	データ解析発表会(ポスター)	3
	論文作成講座	4		

## 4. 評価の方法

### 【知識・技能】定期考査

DSに係わるペーパーテストを実施し、DSの知識・技能を測定する。テスト内容に関する従来の科目との関係は下表の通りである。

表4 ペーパーテストに含まれる従来の科目内容

従来の科目	1年前半	1年後半	2年前半	2年後半
数学I	データの分析①	データの分析②	推測統計① 情報とデータサイエンス	データの分析③
数学B				推測統計②
情報I		データの活用①		データの活用②
情報II				

### 【思考力・判断力・表現力】発表会におけるルーブリック評価

(74回生は3観点×5段階評価, 75回生は7観点×5段階評価)

発表会ごとに教員研修を実施し、ルーブリックにおける基準の相互理解を行い、評価をつける。

### 【主体的に学習に取り組む態度】研究記録簿 (Teamsの討議を含む) による研究討議の得点化

研究グループの生徒同士や、担当教員とのやり取りを踏まえて、主体的に研究に取り組む態度を評価する。



○ 資料編

1. 授業教材例 . . . . .	6
2. 具体的指導例 . . . . .	18
3. PPDAC サイクルに対応した研究記録簿 . . . . .	22
4. PPDAC サイクルに対応した研究ルーブリック	23
5. データサイエンスに関するペーパーテスト (一例)	24



## 【 資料編 】

### 1. 授業教材例

(1) 「DS教育」において価値創造を目指すことを意識させるためスライド

膨大な情報(ビッグデータ)から、必要なデータを選択し、データが意味を持つように表現(可視化)し、解析結果を融合することで、新たな価値を生み出す過程の獲得、結論の創造を目指す。

Hyogo Prefecture Himeji Nishi SHS ↑

### データサイエンスで目指すこと

現在は「情報化社会」と言われるが、データサイエンスの世界では、**創造性**  
「扱えるデータが容易に得られる社会」と捉える。

データサイエンスは、ビッグデータから**有益な情報をあぶり出す**ことである。  
多変量のビッグデータからあぶり出す解析を**多変量解析**という。

**データ解析で重要なことは？**

ヒストグラム・箱ひげ図  
散布図をかきまくる  
↓  
相関をみる  
↓  
原因となる相関を見抜く

**格言1**  
データサイエンスは、「宝探し」  
データとデータを繋ぎ合わせ、筋道をつくる！  
異分野のデータや知恵を組み合わせる！

**格言2**  
相関関係と因果関係は異なる  
相関関係から因果関係を見抜く！

(2) 「DS」に必要な学問概念をイメージしたスライド

DSは、文系・理系という概念を超えた分野であることを示す。

Hyogo Prefecture Himeji Nishi SHS ↑

### データサイエンスとは

データの集積のみでは価値は生まれない  
客観的な存在としてのビッグデータを対象  
→ **新しい知見の抽出、価値の発見・創造**

統計学と情報学のドッキング  
+ 人文知の役割

**理系的**  
収集・加工・処理  
データ  
エンジニアリング  
(コンピューター科学)

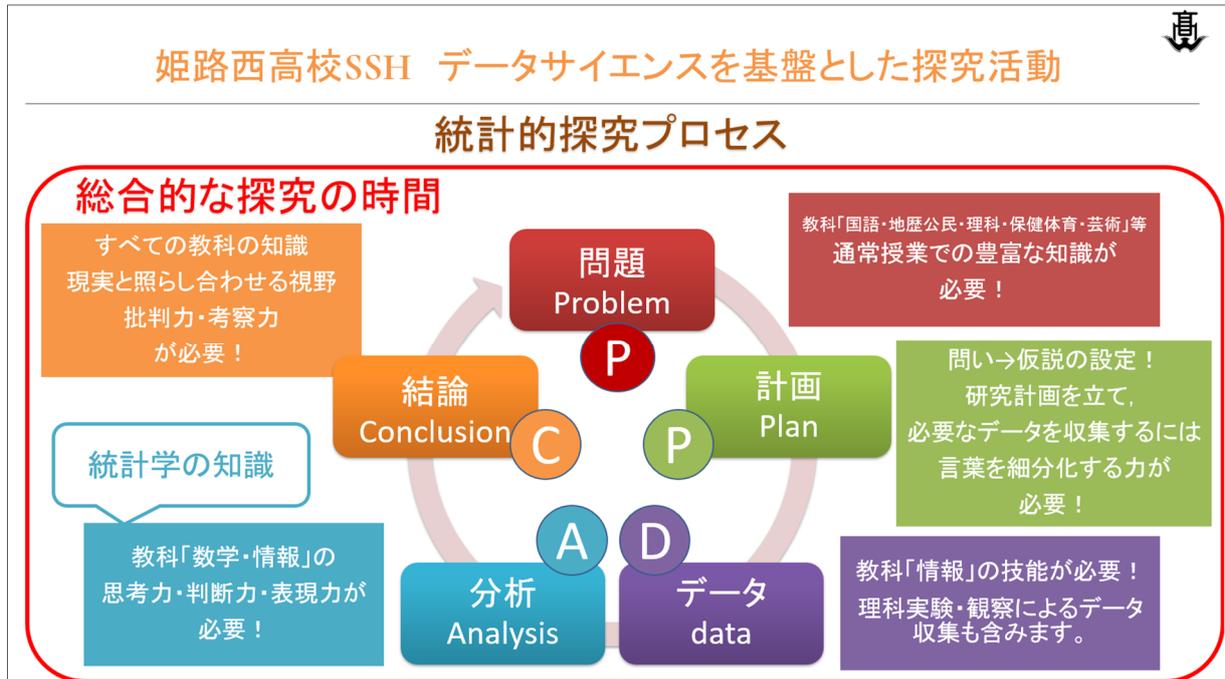
**理系的**  
データ  
分析・解析  
データ  
アナリシス  
(統計学)

**文系的**  
諸分野の  
知識と理解  
人文・社会知  
価値創造

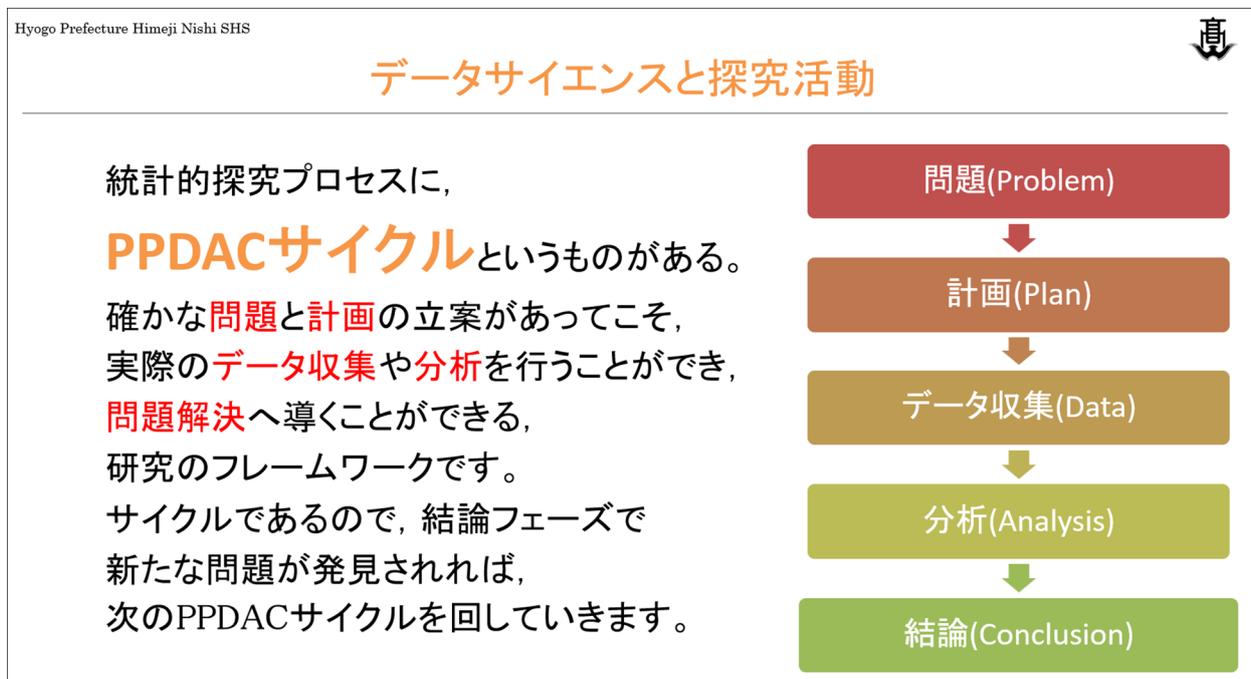
データサイエンティスト

滋賀大学位田隆一学長による講演会資料より抜粋 (2020年7月)

- (3) PPDAC サイクルと各教科の関連を示したスライド  
学校の教育活動における各教科の役割を整理している。



- (4) PPDAC サイクルを説明したスライド  
統計的探究プロセス，PPDAC サイクルという研究のフレームワークを説明する。



(5) 【問題】に関する授業スライド1

各生徒の興味関心のある内容でテーマを設定させていく。その際に、テーマに対して「理想」を述べさせ、「現実」を調査させると、ギャップが出てくる。そのギャップから問題→課題を見つけるきっかけが生まれることを説明する。

Hyogo Prefecture Himeji Nishi SHS

問題(P1 テーマの設定)

### 研究テーマ→問いを立てる→課題を発見する

- ① 研究テーマを決めよう。【抽象的】
- ② ①に対して、どんな理想をもっているか考えよう。
- ③ ①の「現実」について考えよう。(調査しよう。)【具体的】
- ④ ②と③のギャップから問題を見つけよう。【抽象的】
- ⑤ ④の問題から自分が解決したい課題を見つけよう。【具体的】

興味・関心があることからテーマを決めよう

(6) 【問題】に関する授業スライド2

研究テーマが抽象的な場合、スライドのように言葉の意味や定義が明確でない場合が多い。そのため、スライドのように言葉の定義を意識させる。

問題(P1 テーマの設定)

### 研究テーマ→問いを立てる→課題を発見する

参考：課題研究メソッド p.31

#### 1. 言葉の意味や定義を問う「問い」

- 「ニート」って何？
- きちんと調べる(辞書や国・地方自治体のホームページをみる)

・意味や定義が不明確な場合は、自分自身で仮に定義する

- 研究発表の際の表現を確認しよう。

・似ている言葉の違いを比較する。

- 幼稚園と保育園の違いは何？

資料引用元: 啓林館 岡本尚也(2021)「課題研究メソッドよりよい探究活動のために 2nd Edition」

(7) 【問題】に関する授業スライド3

指導の参考事例として、スライドのように抽象的な内容を具体的な内容へ換言していくと、研究内容が深まる。そうしたマジックワードを意識化させることで、生徒と教員が共有した概念を持つことができる。

Hyogo Prefecture Himeji Nishi SHS

問題(P1 テーマの設定)

## 研究テーマ→問いを立てる→課題を発見する

参考：課題研究メソッド p.32

### マジックワード(抽象的な言葉)を使用していますか？

安全な社会

言い換えると、  
・犯罪のない社会  
・戦争のない社会  
・災害による被害が少ない社会

災害による被害が少ない社会

言い換えると、  
・台風による被害  
・大雨による被害  
・竜巻による被害

大雨による被害が少ない社会

言い換えると、  
・洪水や冠水による  
・雨漏りによる浸水被害  
・土砂崩れによる被害

大雨による土砂崩れによる被害が少ない社会

マジックワードは、言葉の言い換えを行い、具体化しよう。

資料引用元：啓林館 岡本尚也(2021)「課題研究メソッドよりよい探究活動のために 2nd Edition」

(8) 【問題】に関する授業スライド4

仮説の立て方として、スライドのように「解が見つからない問い」が仮説候補となる。スライドのように問いを深めていくことで、仮説へとつながることを説明する。

Hyogo Prefecture Himeji Nishi SHS

問題(P2 リサーチクエスション)

## 研究テーマ→問いを立てる→課題を発見する

参考：課題研究メソッド p.44

### ○ 研究テーマに対して「問い」を立てよう

問いを立てる→解がある→問いを立てる→解がある  
→ ……………→問いを立てる→解が見つからない

1. 言葉の意味や定義を問う「問い」 → ○○の意味は？ ○○の定義は？
2. 原因(なぜ)を問う「問い」 → なぜ○○は生じているのか？
3. 信憑性を問う「問い」 → ○○は本当に生じているのか？
4. 比較を問う「問い」 → 他の国・地域ではどのように変化している？
5. 先行研究・先行事例を問う「問い」 → ○○に対して、どのような研究が行われた？
6. 影響を問う「問い」 → ○○によって、どのようなことが起こるのか？
7. 方法や関連性を問う「問い」 → ○○と△△にはどのような関連があるのか？

資料引用元：啓林館 岡本尚也(2021)「課題研究メソッドよりよい探究活動のために 2nd Edition」

- (9) 【計画】に関する授業スライド  
先行研究を調査し、仮説を立てることを示す。

Hyogo Prefecture Himeji Nishi SHS

計画(P3 仮説を立てる)

## 仮説を立てる

○ 仮説

仮説とは、リサーチクエスションに対する**予想される仮の答え**である。  
「○○は□□である」という形で書こう！

**先行研究は必ず調査しよう！**  
先行研究に基づいて研究していることを発表で示すこと

1. 先行研究・事例から根拠をもって仮説を示す
2. 複数の仮説を立てる
3. 仮説を検証する方法を考える

参考：課題研究メソッド p.66～70

資料引用元：啓林館 岡本尚也(2021)「課題研究メソッド よりよい探究活動のために 2<sup>nd</sup> Edition」

- (10) 【データ】に関する授業スライド1  
データ収集の方法を示す。

Hyogo Prefecture Himeji Nishi SHS

データ(D1収集)

## 検証のためのデータ収集方法

○ データ収集のポイント

- ・インターネットにあるデータをそのまま使う → ×
- ・インターネットにあるデータの出典をみる → ○  
⇒ その出典元からデータを取り、自分で**データ整理整形**しよう！
- ・**先行研究**を参考に、データ収集元を探す → ◎
- ・自分でデータを採取する → ◎

- (11) 【データ】に関する授業スライド2  
 オープンデータを収集するための集積サイトを活用させる。

Hyogo Prefecture Himeji Nishi SHS

データ(D1収集)

## オープンデータのデータ収集方法

JDSSP高等学校データサイエンス教育研究会

高校生のための「DS教育コンソーシアム」  
 高校生にデータサイエンス教育を実践する教員のための研究会です。

DS教材の記事一覧

「データ活用」に係わる授業モデル・教材  
2021年2月16日

データサイエンスの評価について  
2020年10月23日

データ収集一覧  
2020年10月23日

<https://ds-education.com/>

オープンに利用できる統計データのサイトが集まっているページ

DS教材  
データ収集一覧  
2020年10月23日

- DATA GO.JP (日本政府全体のデータ一覧サイト)
- e-Stat (政府統計の総合窓口)
- 統計GIS (地図で見える統計)
- RESAS (リーサス) 地域経済分析システム (データがすぐに可視化できるサイト)
- SSDSE (教育用標準データセット) 統計データ分析
- miripo (マイクロデータ利用ポータルサイト)
- 統計ダッシュボード (総務省統計局)
- なるほど統計学園
- 国土交通省 国土地理院
- 国、都道府県、政令指定都市「統計年鑑(統計書、県勢要覧、統計年報)まとめサイト
- 各府省及び独立行政法人等のページ (国の省庁ごとのデータリンクサイト)
- 国立社会保障・人口問題研究所「将来推計人口データベース」

- (12) 【データ】に関する授業スライド3  
 容易にデータが可視化される RESAS (リーサス) というサイトの活用方法を体得させる。

Hyogo Prefecture Himeji Nishi SHS

データ(D1収集)

## オープンデータのデータ収集方法

### RESAS地域経済分析システムの利用

高校生のためのデータサイエンス入門

**第1週第4回  
地域経済分析システム  
(RESAS)の利用**

DS教材  
データ収集一覧  
2020年10月23日

- DATA GO.JP (日本政府全体のデータ一覧サイト)
- e-Stat (政府統計の総合窓口)
- 統計GIS (地図で見える統計)
- RESAS (リーサス) 地域経済分析システム (データがすぐに可視化できるサイト)
- SSDSE (教育用標準データセット) 統計データ分析
- miripo (マイクロデータ利用ポータルサイト)
- 統計ダッシュボード (総務省統計局)
- なるほど統計学園
- 国土交通省 国土地理院
- 国、都道府県、政令指定都市「統計年鑑(統計書、県勢要覧、統計年報)まとめサイト
- 各府省及び独立行政法人等のページ (国の省庁ごとのデータリンクサイト)
- 国立社会保障・人口問題研究所「将来推計人口データベース」

簡単にデータが可視化できます。  
この動画をみると使い方がわかります。

本校は滋賀大学と連携協定を締結しているため、動画を閲覧できます。

資料引用元: 滋賀大学 高校生のためのデータサイエンス入門

(13) 【データ】に関する授業スライド4

エクセルファイル・csv ファイルでのオープンデータの取得が可能になる、政府統計の総合窓口（e-stat）というサイトの活用方法を体得させる。

Hyogo Prefecture Himeji Nishi SHS

データ(D1収集)

## オープンデータのデータ収集方法

### 政府統計の総合窓口(e-stat)の利用

滋賀大学

第1週第5回  
政府統計の総合窓口  
(e-Stat)の利用

オープンデータがまとまっています。  
この動画を見ると使い方がわかります。

本校は滋賀大学と連携協定を締結しているため、動画を閲覧できます。

資料引用元: 滋賀大学 高校生のためのデータサイエンス入門

- DATA GO.JP (日本政府全体のデータ一覧サイト)
- e-Stat (政府統計の総合窓口)
- 統計GIS (地図で見る統計)
- RESAS (リーサス) 地域経済分析システム (データがすぐに可視化できるサイト)
- SSDSE (教育用標準データセット) 統計データ分析
- miripo (マイクロデータ利用ポータルサイト)
- 統計ダッシュボード (総務省統計局)
- なるほど統計学園
- 国土交通省 国土地理院
- 国、都道府県、政令指定都市「統計年鑑(統計書、県勢要覧、統計年報)」まとめサイト
- 各府省及び独立行政法人等のページ (国の省庁ごとのデータリンクサイト)
- 国立社会保障・人口問題研究所「将来推計人口データベース」

(14) 【データ】に関する授業スライド5

地図で見る統計（統計 GIS）というサイトによって、地図上に統計データを可視化するソフトの活用方法を体得させる。

Hyogo Prefecture Himeji Nishi SHS

データ(D1収集)

## オープンデータのデータ収集方法

### 地図で見る統計(統計GIS) 【jSTAT MAPの利用】

滋賀大学

地図で見る統計(統計GIS)

各種統計データを地図上に表示し、視覚的に統計を把握できる地理情報システム(GIS)を提供しています。

「お知らせ」

2021年4月21日 地図で見る統計(jSTAT MAP)のログイン画面を変更いたしました。

2021年3月19日 2018年漁業センサス 都道府県及び市町村の提供を開始いたしました。

> 地図で見る統計(jSTAT MAP)

地図で見る統計(jSTAT MAP)は、誰でも使える地理情報システムです。統計地図を作成する他に、利用者のニーズに沿った地域分析が可能となるようなさまざまな機能を提供しています。防災、施設整備、市場分析等、各種の詳細な計画立案に資する基本的な分析が簡単にできます。 ※システムの動作が著しく遅い場合は、システムが混み合っている可能性があります。時間をおいて再度アクセスをお願いいたします。 また、地図で見る統計(jSTAT MAP)起動時にエラーとなる場合は、ブラウザの閲覧履歴の削除を行い再度お試しください。

地図に  
様々な統計データを  
反映できる

- DATA GO.JP (日本政府全体のデータ一覧サイト)
- e-Stat (政府統計の総合窓口)
- 統計GIS (地図で見る統計)
- RESAS (リーサス) 地域経済分析システム (データがすぐに可視化できるサイト)
- SSDSE (教育用標準データセット) 統計データ分析
- miripo (マイクロデータ利用ポータルサイト)
- 統計ダッシュボード (総務省統計局)
- なるほど統計学園
- 国土交通省 国土地理院
- 国、都道府県、政令指定都市「統計年鑑(統計書、県勢要覧、統計年報)」まとめサイト
- 各府省及び独立行政法人等のページ (国の省庁ごとのデータリンクサイト)
- 国立社会保障・人口問題研究所「将来推計人口データベース」

(15) 【データ】に関する授業スライド6

データの特徴を読み取るために、エクセルを活用してデータを整理整形する。

Hyogo Prefecture Himeji Nishi SHS

データ(D2整理整形)

### 整理整形に活用できるエクセルの関数

統計量とは ... 代表値(最小値・第1四分位数・第2四分位数(中央値)・第3四分位数・最大値)  
最頻値, 平均値, 分散, 標準偏差など

求めたい値	エクセルの関数
合計	=SUM( ○ : ○ )
データ数	=COUNT( ○ : ○ )
平均	=AVERAGE( ○ : ○ )
分散	=VARP( ○ : ○ )
標準偏差	=STDEV( ○ : ○ )
最小値	=QUARTILE(○ : ○ , 0)
第1四分位数	=QUARTILE(○ : ○ , 1)
第2四分位数(中央値)	=QUARTILE(○ : ○ , 2)
第3四分位数	=QUARTILE(○ : ○ , 3)
最大値	=QUARTILE(○ : ○ , 4)

データ	10
	20
	30
	40
	50
	60
	70
	80
合計	360
データ数	8
平均	45
分散	525
標準偏差	22.91
最小値	10
第1四分位数	27.5
第2四分位数(中央値)	45
第3四分位数	62.5
最大値	80

(16) 【データ】に関する授業スライド7

外れ値や異常値に関するデータの取り扱い方を理解させる。

Hyogo Prefecture Himeji Nishi SHS

データ(D2整理整形)

### データの外れ値・異常値

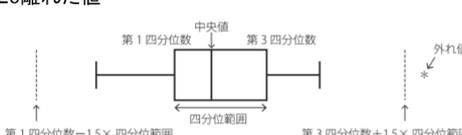
**外れ値** ... 量的データにおいて、他の値から極端にかけ離れたデータのこと

**異常値** ... 測定ミスや入力ミスなどによる値

**外れ値の判断**

まず、明確な外れ値の定義はない。ただ、1つの目安として、下記を示しておく。

- ・通常四分位範囲の1.5倍以上離れた値(「数学」学習指導要領より)
- ・平均値より±2σ離れた値

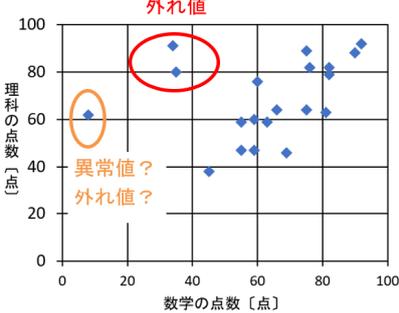


「外れ値」のポイント

① 外れ値を除いて、研究を進めるのも1つの流れです。  
しかし、研究の流れとしては、一度外れ値を含めた散布図を示し、外れ値として判断した基準を示す。  
そして、外れ値を除いた散布図を示すことで、きちんと外れ値を判断したことを伝えることが重要です。

② 外れ値の原因を研究するのも、1つの研究です。

**外れ値**



(17) 【分析】に関する授業スライド1

基本的なデータの可視化について、伝えたい内容に適した方法を理解させる。

Hyogo Prefecture Himeji Nishi SHS

分析(A1 データ可視化)

## 収集したデータの可視化

### 棒グラフ

横軸に対する縦軸の値の大きさの変化や違いを明確に示すことができる

### 円グラフ

全体に対する各項目の割合を明確に示すことができる

- 野球部
- サッカー部
- テニス部
- ソフトテニス部
- バドミントン部
- 水泳部
- バレー部
- バスケットボール部
- その他

### 折れ線グラフ

比較対象が多く、時系列など横軸に対する変化をみる際に適している

### 帯グラフ

割合を示し、複数の項目やその変化を比較できる

学年	10冊以上	6~9冊	3~5冊	1~2冊	ほとんど読まない	不明
高1	3.5	15.1	23.4	24.6	22.2	11.2
高2	23.1	12.7	23.1	57.4		1.6
高3	3.0	11.1	27.4	12.1	41.4	

資料引用元: 啓林館 岡本尚也(2021)「課題研究メソッド よりよい探究活動のために 2nd Edition」

(18) 【分析】に関する授業スライド2

解析する際に、単体のデータで判断せずに複数のデータを比較する手法を理解させる。

Hyogo Prefecture Himeji Nishi SHS

分析(A2 データ解析)

## グラフの特徴的な部分を読み取る

### 事例

図1のグラフでは、特徴的な部分はあるが、本当に特徴的な部分であるのか根拠が少ない

2014年夏物売れ数

図1 2014年夏物服の販売数

毎年同じような傾向がみられる複数のデータから判断する

図2 2014年~2018年の夏物服の販売数

データ解析ポイント

**ポイント①** 図2のように複数年を重ねると、特徴的であることが明確になる。

**ポイント②** さらに、2014年は、他の年度より、非常に売り上げが多いのはなぜ？と研究が深まっていく

図2では、他に特徴的な部分がある。気づきますか？

(19) 【分析】に関する授業スライド3

本校で主に指導する統計手法である。ただし、研究でこれらの統計手法をすべて活用しなければならぬわけではない。研究内容に即した手法を選択する力の必要性も理解させる。

Hyogo Prefecture Himeji Nishi SHS

分析(A2 データ解析)

姫路西高校で学ぶ「統計手法」

回帰分析法(単回帰分析・重回帰分析)

- ・ 未知な事柄を予測するための手法  
キーワード: 回帰式・回帰係数・決定係数・寄与率・p値・予測モデル

標準化

- ・ データ群を, 平均0, 標準偏差1にする手法  
2つ以上を比較するとき利用する。

仮説検定(平均の差の検定)

- ・ データ群とデータ群を比較し, 平均の差が有意であるかないか判断する手法

クラスタリング(k-means法など)

- ・ 統計的にデータをグループ分けする手法

主成分分析法

- ・ 変数間の関係性から共通する因子(潜在変数)を導き出し,  
多くのデータを少ない変数に縮約する分析手法

(20) 【分析】に関する授業スライド4

プログラミングを学ぶ際に、プログラミングを手段として活用することの有効性を伝える。

分析(A2 データ解析)

プログラミング言語 Python を学ぶ

データの整理・整形, 可視化をエクセルで行うことは非常に手間である  
⇒ **プログラミング**を活用すれば, **楽ができる!**

Pythonでプログラミングができる! → **楽ができる!**

大切なことは、Pythonを使って何をしたいのか → **Pythonは単なる手段!**

データサイエンスでは、

- ・ データ収集のために, Pythonを使う
- ・ データ研磨のために, Pythonを使う
- ・ Pythonの出力結果から, データ解析する

(21) 【結論】に関する授業スライド1

PPDACサイクルを何周もまわすことで研究が進化することを理解させる。

Hyogo Prefecture Himeji Nishi SHS

結論(C) 真

### データ解析結果→考察→結論を導こう

PPDACサイクルを繰り返すことで、結論が磨きあげられる！

資料引用元: 啓林館 岡本尚也(2021)「課題研究メソッド よりよい探究活動のために 2nd Edition」

(22) 【結論】に関する授業スライド2

新しい価値を生む創造性に関するメカニズムを理解させる。

Hyogo Prefecture Himeji Nishi SHS

結論(C) 真

### 【創造力】新たな価値を創造しよう

研究

- ・みんなが当たり前と思っていることをデータで検証する
- ・他の人が思いつかない結論 を導く!
- ・予想外の結論 を導く!

新しい価値を生む創造性には

- ・ある分野の内容を、異なる分野に活用して、  
新しいものを創り出すこと ⇒ **リフレーム**
- ・異質なものを結び付けて新たな軸をつくること ⇒ **統合**

解析結果に基づいた解釈していく上で、  
**リフレームと統合**を意識し、新たな価値を創造しましょう!

(23) 【1年生連携企業によるデータ解析 PBL】説明スライド1

1年生前半の学びを実社会のデータの解析に生かす演習。本校と連携しているグローリー（株）の技術を活用した（株）西松屋の顧客データを提携してもらい、そのデータを分析し、（株）西松屋チェーンへ提案する演習である。

## 子ども服販売の社員になろう！

### 演習の主旨

背景： あなたは、子ども服を販売するお店の店舗で子ども服を販売している社員です。

ミッション： 社員として働いている店舗の売り上げ高を向上させるよう社長から指令がありました。

ゴール： **与えられたデータ**に基づいて、**店舗の売上を向上させる施策**を提案してください。



### 提案方法

与えられたデータに基づき、様々な視点から1つ提案してください。

来店者の地域情報・性別・来客人数・来客頻度などのデータを元に、ターゲットを絞った客層に応じた提案でもよいです。

また、客層に応じたSNSの利用による宣伝方法の提案など、幅広い視点での提案をお願いします。

ただし、どのジャンルにおいても、客観的なデータ分析に基づいたプランにしてください。

(スライド提供：AdInte)

(24) 【1年生連携企業によるデータ解析 PBL】説明スライド2

PPDACサイクルを活かし、具体的な施策を（株）西松屋チェーンに提案する。

## 問題解決に向けたデータ解析実践 → スライド発表

### データ分析の視点をレクチャー

### 子ども服販売の社員になろう！

#### データ分析のプロセス

データ分析のプロセスには、大きく分けて以下の二つがあります。

1. データ分析をして課題を探し施策を検討する（帰納的データ活用）  
例えば、「休日に比べ平日の来店数が少なかった。子ども服売上向上のためには、休日来店者のクチコミを広げて平日来店につなげていけばどうだろう？」など
2. 仮説を立て「仮説の正誤」をデータ分析で明らかにし施策の実施判断をする（演繹的データ活用）  
例えば、「子ども服の売上向上にはファミリーの来店を促す必要があるのでは？データ分析の結果、休日に比べ平日の来店数が少なく、「休日来店ファミリーのクチコミを広げて平日来店に繋げたい」など

実際のデータサイエンス・統計の仕事では、1,2.を行き来して提案の説得性と予測性・確度を上げていきます。はじめは、2.から始めるのが発想豊かになりやすいです。

PPDACサイクルの繰り返し

問題(Problem)  
... 課題の細分化

計画(Plan)  
... 仮説の設定

データ(Data)  
... 与えられたデータ + オープンデータの利用

分析(Analysis)  
... 可視化, 解析

結論(Conclusion)  
... 解決策の提示

(スライド提供：AdInte)

## 2. 具体的指導例

### ≪重回帰分析の指導（2時間）≫

#### (1) 重回帰分析法における指導用スライド1

統計手法「重回帰分析法」に関する内容である。単回帰分析と重回帰分析の違いを理解させる。

Hyogo Prefecture Himeji Nishi SHS DR探究・研究

### 目的変数に対して、説明変数1つだけで十分なの？

**質問：** ハンドボール投げの結果を予測したい。  
ハンドボール投げを予測するための説明変数を見つけよう。

握力 → ハンドボール投げ

「目的変数」に対して、  
「1つ」だけの「説明変数」との関係  
⇒「**単回帰分析**」という

50m走  
握力  
上体起こし → ハンドボール投げ

「目的変数」に対して、  
「複数」の「説明変数」との関係  
⇒「**重回帰分析**」という

#### (2) 重回帰分析法における指導用スライド2

エクセルの「データの分析」（アドインで設定）を活用して重回帰分析を実行する。その出力結果を読み取り、予測モデルを構築する。回帰係数の有意差の検定を行い、変数減少法を用いて最適なモデルを探索する。

Hyogo Prefecture Himeji Nishi SHS DR探究・研究

### MS-Excelを利用して、重回帰式を作成する。

「分析ツール」→「回帰分析」からデータを選択する。

回帰統計	
重相関 R	0.85
重決定 R <sup>2</sup>	0.72
補正 R <sup>2</sup>	0.72
標準誤差	3.73
観測数	275

分散分析表					
	自由度	変動	分散	観測された分散比	有意 F
回帰	6	9771.33	1628.56	117.28	0.000
残差	268	3721.40	13.89		
合計	274	13492.73			

「重相関R」→「重相関係数 R」  
回帰モデルによる期待値と実測値との相関係数である。

「重決定R<sup>2</sup>」→「寄与率（決定係数） R<sup>2</sup>」  
0 から 1 の間の値を取り、1 に近いほど、当てはまりがよいと判断する

「補正R<sup>2</sup>」→「自由度調整済み寄与率R<sup>2</sup>（決定係数）」  
複数の回帰式を比較するとき使用する

**p 値**  
変数の係数が 0 のとき、つまり、この係数がないときの影響はどれほどあるのか判断できる  
(0.05が基準となることが多い)

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 99.0%	上限 99.0%
切片	1.44	7.60	0.19	0.850	-13.51	16.40	-18.27	21.15
握力	0.32	0.05	6.78	0.000	0.23	0.41	0.20	0.44
上体起こし	0.21	0.06	3.54	0.000	0.09	0.32	0.06	0.36
反復横とび	0.09	0.03	2.72	0.007	0.03	0.16	0.00	0.18
持久走	-0.97	0.48	-2.00	0.047	-1.92	-0.01	-2.22	0.29
50M走	-1.59	0.54	-2.97	0.003	-2.65	-0.54	-2.99	-0.20
立ち幅跳び	0.04	0.01	2.94	0.004	0.01	0.07	0.01	0.08

係数 正か負であるか符号を確認すること

(3) 演習

授業の目的：過去のデータから、予測モデルを構築する。  
 「重回帰分析」を用いて、まだ測定していないハンドボール投げの予測値を求めよう。

(手順①) 保健体育で測定している過去の生徒の体力測定の結果データをもとにして考える。

部活動	男女	握力	上体起こし	反復横とび	持久走	50M走	立ち幅跳び	ハンドボール投げ
テニス	2	28	26	60	287	8.5	193	17
サッカー	1	32	34	62	343	7.7	223	25
バドミントン	1	34	28	55	378	7.8	230	18
陸上競技	2	28	27	49	262	7.9	191	9
卓球	1	41	27	63	383	7.7	226	21
囲碁将棋	1	47	28	59	391	7.4	224	19
サッカー	1	42	39	64	353	6.7	270	29
E・S・S	2	26	22	42	327	10.4	157	8
音楽 新聞	1	41	36	63	333	7.4	240	28
囲碁将棋	1	34	20	50	412	7.6	233	26
バドミントン	1	25	34	58	332	7.8	180	13

(手順②) 「データ」 → 「データの分析」を選択する。

「データの分析」が表示されない場合は、オプションのアドインで設定する。



(手順③) 回帰分析を選択する。

(手順④) 「入力Y範囲」に目的変数, 「入力X範囲」に説明変数の行データ (複数行が可能) を選択する。ここで、エラーが出る場合はデータが欠損しているなど、データを確認する。

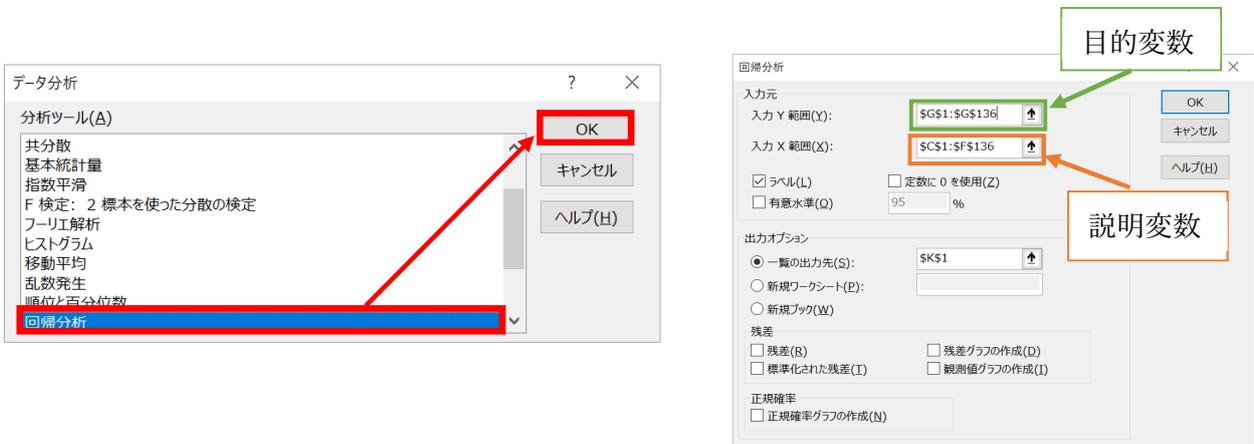


表5 活用するデータ事例

握力	上体起こし	反復横とび	持久走	50M走	立ち幅跳び	ハンドボール投げ
28	26	60	287	8.5	193	17
32	34	62	343	7.7	223	25
34	28	55	378	7.8	230	18
28	27	49	262	7.9	191	9
41	27	63	383	7.7	226	21
47	28	59	391	7.4	224	19
42	39	64	353	6.7	270	29
26	22	42	327	10.4	157	8

説明変数

目的変数

(手順⑤) 分析ツールで実行した数値を読み取り、予測モデルを作る。

表6 重回帰分析の出力結果

	係数	標準誤差	t	P-値
切片	-1.534	9.807	-0.156	0.876
握力	0.345	0.058	5.925	0.000
上体起こし	0.247	0.079	3.128	0.002
反復横とび	0.146	0.048	3.047	0.003
持久走	-1.032	0.520	-1.987	0.049
50M走	-1.594	0.704	-2.264	0.025
立ち幅跳び	0.034	0.019	1.763	0.080

表6から読み取った予測モデル例

$$\begin{aligned} \text{ハンドボール投げ(m)の予測値} &= 0.345(\text{m/kg}) \times \text{握力(kg)} + 0.247(\text{m/回}) \times \text{上体起こし(回)} \\ &+ 0.146(\text{m/回}) \times \text{反復横跳び(回)} - 1.032(\text{m/秒}) \times \text{持久走(秒)} - 1.594(\text{m/秒}) \times \text{50m走(秒)} \\ &+ 0.034(\text{m/cm}) \times \text{立ち幅跳び(cm)} - 1.534 \end{aligned}$$

握力・上体起こし・反復横跳び・持久走・50M走・立ち幅跳び  
の数値を代入し、ハンドボール投げの予測値を求める。

$$+0.034(\text{m/cm}) \times \text{立ち幅跳び(cm)} - 1.534$$

(手順⑥) 予測モデルの寄与率を読み取る。

表計算ソフトの一例としてExcelで示したが、Excelの出力結果は必要に応じて、レポート等では書き直す必要がある。

「重相関R」→「重相関係数R」

「重決定R<sup>2</sup>」→「寄与率(決定係数)R<sup>2</sup>」

「補正R<sup>2</sup>」→「自由度調整済みR<sup>2</sup>」

回帰統計	
重相関R	0.822
重決定R <sup>2</sup>	0.676
補正R <sup>2</sup>	0.665
標準誤差	3.987
観測数	180

この出力から、モデルの適合度を示す寄与率が67.6%であることがわかる。

(手順⑦) 最適な予測モデルを構築するために数値を読み取る。

複数の説明変数の中には、目的変数の予測に役立たないものが含まれている可能性がある。そこで、モデル式の構築にあたっては、説明変数の取捨選択が重要な課題となる。これを変数選択もしくはモデル選択（モデリング）という。その際、寄与率ができるだけ大きくなることが望ましいが、説明変数を増やせば増やすほど単純に大きくなる。しかし、欠点（過剰適合、過学習）も生じる。

ここでは、変数減少法を説明する。

- 回帰係数の有意差の検定により、回帰係数の有意確率（P-値）の列から、1%もしくは5%以下であるか判断する。表6における青色枠の例では、「握力」「上体起こし」「反復横跳び」は1%有意、「持久走」「50M走」は5%有意であるが、「立ち幅跳び」に有意差はないと判断する。

(手順⑧) 最適な予測モデルを構築する。

有意差が出ない変数は「ハンドボール投げ」の値の変化に影響を与えていない可能性が高いと判断される。

「持久走」を含む結果		「持久走」を含まない結果		「持久走」を含まない回帰係数・p値				
回帰統計		回帰統計		係数	標準誤差	t	P-値	
重相関 R	0.822	重相関 R	0.819	切片	9.121	7.769	1.174	0.242
重決定 R <sup>2</sup>	0.676	重決定 R <sup>2</sup>	0.671	握力	0.371	0.057	6.530	0.000
補正 R <sup>2</sup>	0.665	補正 R <sup>2</sup>	0.661	上体起こし	0.255	0.079	3.217	0.002
標準誤差	3.987	標準誤差	4.011	反復横とび	0.166	0.047	3.555	0.000
観測数	180	観測数	180	持久走	-1.061	0.522	-2.030	0.044
				50M走	-2.314	0.577	-4.009	0.000

【手順⑤に比べ、変数が少ない予測モデル】

$$\begin{aligned} \text{ハンドボール投げの予測値}(m) = & 0.371(m/kg) \times \text{握力}(kg) + 0.255(m/回) \times \text{上体起こし}(回) \\ & + 0.166(m/回) \times \text{反復横跳び}(回) - 1.061(m/秒) \times \text{持久走}(秒) - 2.314(m/秒) \times \text{50M走}(秒) + 9.121 \end{aligned}$$

この予測モデルに、説明変数（握力・上体起こし・反復横跳び・持久走・50m 走）の値を代入した値が、ハンドボール投げの予測値である。そして、予測値と実測値（実際にハンドボール投げをして測定した値）を比較する。

参考資料：文部科学省高等学校情報科「情報Ⅱ」教員研修用教材（本編）第3章情報とデータサイエンス

### 3. PPDAC サイクルに対応した研究記録簿

#### 統計的探究のプロセス 研究進捗確認表 (日付 / 記入者名 )

<p>問題(Problem)</p> <ul style="list-style-type: none"><li>・抽象的な課題になっていないか</li><li>・言葉の定義は定まっているか</li><li>・課題が細分化されているか</li><li>・問いから仮説を導き出そうとしたか</li></ul>	<p>計画(Plan)</p> <ul style="list-style-type: none"><li>・どのようなデータを収集するのか</li><li>・仮説を立てているか</li><li>・先行研究を調査したか</li></ul>
<p>データ(Data)</p> <ul style="list-style-type: none"><li>・データ収集元は記録しているか</li><li>・外れ値、異常値はどうするのか</li><li>・データの整理・整形をどうするのか</li></ul>	<p>分析(Analysis)</p> <ul style="list-style-type: none"><li>・データをどのように可視化するか</li><li>・統計量を活用しているか</li><li>・どのようにグラフを読み取るのか</li><li>・統計手法は活用できないか</li></ul>
<p>結論(Conclusion)</p> <ul style="list-style-type: none"><li>・解析結果から考察したか</li><li>・複数の結論を統合したのか</li><li>・新たな仮説は立てられるのか</li></ul>	

## 4. PPDAC サイクルに対応した研究ルーブリック

令和3年度に作成した PPDAC サイクルに基づいた評価用ルーブリックを示す。

【令和3年度研究評価用ルーブリック】

観点/評価点	1	2	3	4	5
問題 Problem	理想だけ、もしくは、現実だけに着目しており、問いも立てておらず、抽象的な問題設定である。	理想と現実のギャップを見出し、立てた問いはすぐに解が見つかり、やや抽象的な問題設定である。	研究の目的が明確であり、適切な問い立てができ、具体的な問題設定である。	複数の問い立てから、具体的な問題設定である。	先行研究を根拠して独自性のある具体的な問題設定である。
計画 Plan	仮説が立てられていない。 (まだ問いであり、仮説になっていない。)	仮説を立てているが、研究による見通しを欠いている。(期限内で終わる見込みがない。)	適切な仮説を立てており、期限までに完成の見込みがある計画を立てている。	問題解決につながるデータ収集の方法までの計画を立てることができている。	問題解決につながるデータ収集・可視化・分析手法の計画を立てることができている。
データ Data	データ収集ができていない。	データ収集ができていないが、整理・整形ができていない。	データ収集ができ、整理・整形ができていない。	問題解決につながるデータ収集ができ、整理・整形ができていない。	創造的な問題解決につながる複数の分野のデータ収集を行い、整理整形ができていない。
分析 Analysis	インターネットや先行研究等、他者による可視化のみである。	データ可視化をしているが、作法として不十分な点が見受けられる。	データの可視化がなされ、適切な数値を扱い、データ解析できている。	問題解決につながるデータの可視化、数値の扱い、データ解析ができている。	データの可視化、適切な統計手法を用いた客観的なデータ解析ができている。
結論 Conclusion	分析結果と結論が繋がっていない。もしくは、分析結果をそのまま示しただけで考察していない。	おおむね結論をまとめることができているが、不十分な点がある。	考察を行い、適切に結論をまとめることができている。	データを適切に分析し、問題解決にむけた説得力のある結論である。	問題の意味を広く認識し、分析結果をもとにさらに広い視野で結論を導いている。

## 5. データサイエンスに関するペーパーテスト（一例）

○ 「情報とデータサイエンス」に関わるテスト内容

問題に該当するデータを与え、エクセルを活用して問題を解く形式である。SSDSE（教育用標準データセット）を用いて、エクセルを用いたペーパーテストを実施している。

兵庫県内の市町村において、2018年度のデータに基づいて、次の問いに答えよ。

(1) 小学校児童数を、小学校教員数から予測する。

このとき、小学校児童数を（ A ）変数、小学校教員数を（ B ）変数という。

次の①～⑥における適した用語を選びなさい。

① 説明    ② 因子    ③ 目的    ④ 決定    ⑤ 潜在    ⑥ 相関

(2) 小学校児童数を（ A ）変数とし、小学校教員数を（ B ）変数とし、回帰分析をしたとき、このモデルの寄与率を次の適切な選択肢から選びなさい。

① 0.119    ② 0.995    ③ 17.119    ④ 678.23

(3) 小学校教員数が98人の市町村では、小学校児童数は何人であると予想できますか。

次の適切な選択肢から選びなさい。

① 約1000人    ② 約1100人    ③ 約1200人    ④ 約1300人

(4) ある市町村 A と比べて小学校教員が100人多い市町村 B の小学校児童数は、市町村 A に比べて何人多いと推測できますか。次の適切な選択肢から選びなさい。

① 約680人    ② 約800人    ③ 約1030人    ④ 約1710人

(5) 次のデータ群はそれぞれ正規分布に従うことを仮定し、平均値の2標本の両側検定、分散は2標本で共通で未知とし、母平均の差を仮説検定で調べた。帰無仮説を「母平均に有意な差はない」、対立仮説を「母平均に有意な差がある」としたとき、有意水準5%とし、帰無仮説を棄却できるデータの組を選びなさい。

① 15歳未満人口の「男」と「女」    ② 75歳以上人口の「男」と「女」  
③ 出生数と死亡数    ④ 婚姻件数と離婚件数

(6) 兵庫県内の市町村における姫路市の婚姻件数において、標準化した値を小数第2位で求めなさい。

(7) 兵庫県内の市町村における姫路市の婚姻件数と離婚件数は、全体分布の位置から判断し、どちらの方が多いと判断できますか。

① 婚姻件数    ② 離婚件数



## 生徒研究成果事例



## 2020年度統計データ分析コンペティション高校生の部 優秀賞

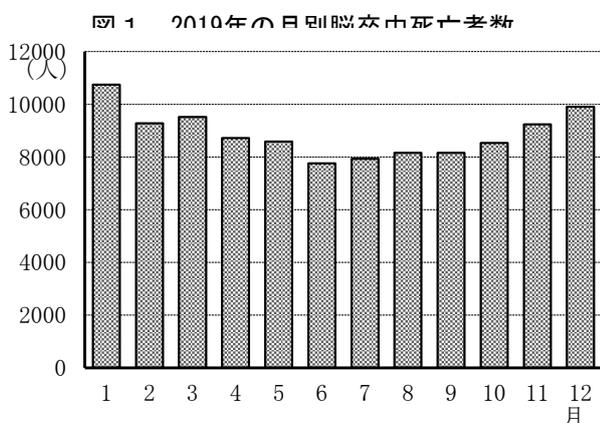
### 気温と脳卒中の発症リスクについて

兵庫県立姫路西高等学校 山野 瑞起・岩見 拓海・黒子 風大・柏木 創太

#### はじめに：研究の背景・目的・方法

2018年人口動態調査によれば、がん、心疾患、脳卒中の三大疾病は、日本人年間死亡者の51%を占める。そのうち、脳卒中死亡割合は8%であるが、発症後に障害が残ることが多い。脳卒中を予防できれば、人々の健康維持に大きく貢献できる。脳卒中は冬に多いというイメージから、環境的要因に影響を受けるのではないかと考え、調べたところ、ヒートショック（短時間の急激な温度変化による血圧の急激な上昇や下降）や脱水症状といった環境的要因が脳卒中を引き起こすことがあり<sup>(1)</sup>、脳卒中の発症リスクは気温や気候と関係するとの先行研究もあった<sup>(2)</sup>。

実際に2019年の脳卒中死亡者数を月別でグラフにすると（図1）、脳卒中による死亡者数は冬に多くなることがわかる。これらから本研究は、気温と脳卒中の発症リスクの間に因果関係があるかを調べることを目的とする。



本研究は、まず気温と脳卒中の発症リスクには相関があるのかを確かめる。次にどのような気温の変化や生活条件で脳卒中発症リスクが高まるのかについて、一連の仮説を立て、相関分析で検討する。

#### 分析に用いたデータ・指標

本研究では、気象庁による日本の12気候区分から9都市を気温変動の異なるモデル都市として選定し、分析の対象とした（表1）。この際、都市による医療格差等に起因する死亡率への影響を抑えるため政令指定都市から選定した。四国、奄美、沖縄の3気候区分から都市を選定しなかったのは政令指定都市に近い大都市が選ばなかったからである。ただし、九州南部には隣接する熊本市を選定している。

表1 選定したモデル都市一覧

気候区分	都市名	気候区分	都市名
北海道	札幌市	近畿	大阪市
東北	仙台市	中国	広島市
関東甲信	さいたま市	九州北部	福岡市
北陸	新潟市	九州南部	熊本市
東海	名古屋市		

分析に使用した変数とその出典の一覧を表2、作成した指標とその計算方法の一覧を表3に示す。ただし表3の「脳卒中死亡率」の計算方法は次のとおりである。

表2 使用したデータおよび出典一覧

データ名	年度	出典
総人口（人）、 15歳未満人口（人）、 15歳以上65歳未満人口 （人）、 65歳以上人口（人）	2015	SSDSE
脳卒中による総死亡数 （人）	2012 ～2019	人口動態調査 （厚生労働省）
日間平均気温（℃）、 日間最高・最低気温（℃）	2011 ～2019	過去の気象データ （気象庁）
2人以上の世帯における 床暖房普及率（%）	2014	全国消費実態調査 （総務省）

表3 作成した指標および計算方法一覧

指標名	計算方法
脳卒中死亡率（10万人対）	後述
月間平均気温（℃）	日間平均気温を月ごとに平均
月間平均最高・最低気温差 （℃）	日最高気温・日最低気温を 月ごとに平均した値

研究対象とするモデル都市ごとに年齢構成が異なり、死亡率に影響が出るので、人口を15歳未満、15歳以上64歳未満、65歳以上の3階級に分け、厚生労働省が示す下記の計算式で死亡率の年齢調整を行った<sup>(4)</sup>。

**年齢調整死亡率＝（階級別粗死亡率×当該階級の基準人口）の各階級の総和／基準人口の総数**

粗死亡率は脳卒中死亡数を総人口で除したものであり、基準人口は厚生労働省にならって昭和60年モデル人口を採用した。このモデル人口は昭和60年国勢調査をもとにベビーブームなどの極端な人口増減を補正している<sup>(4)</sup>。本研究では、年齢調整死亡率にさらに10万をかけた10万人あたりの脳卒中による年齢調整死亡率を脳卒中死亡率と呼ぶ。

## 分析結果

### 気温と脳卒中の発症リスク

脳卒中発症の環境的要因としてヒートショックに着目した。外気温が下がると外気温と室温との差が大きくなるのでヒートショックが起きやすくなり、脳卒中の発症リスクが高まるのではないかと考え、月間平均気温と脳卒中死亡率には負の相関があると予想し、仮説1-1を立てて検討した。

仮説1-1：月間平均気温が低いほど、脳卒中死亡率は高くなる。

各モデル都市で、2019年1月～12月の月間平均気温と脳卒中死亡率の相関を調べると、いずれの都市でも負の相関を示した（表4、図2）。

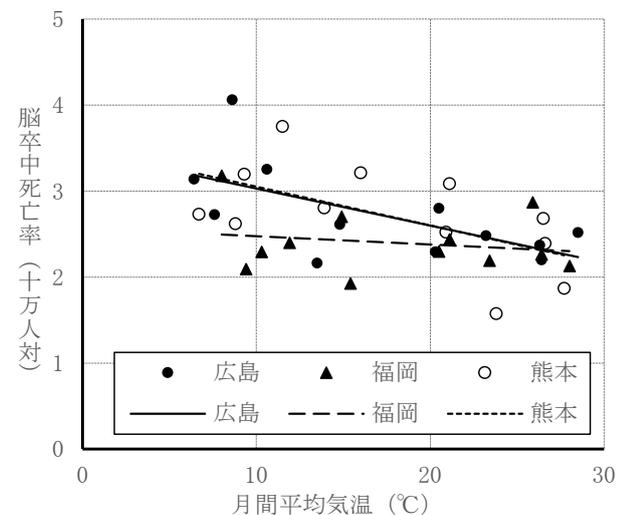
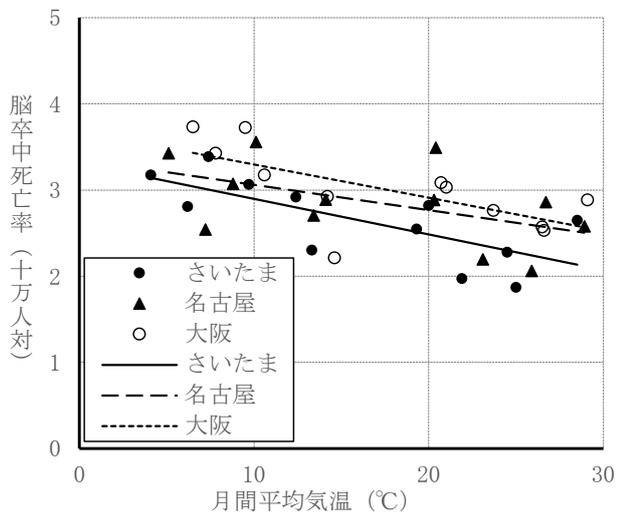
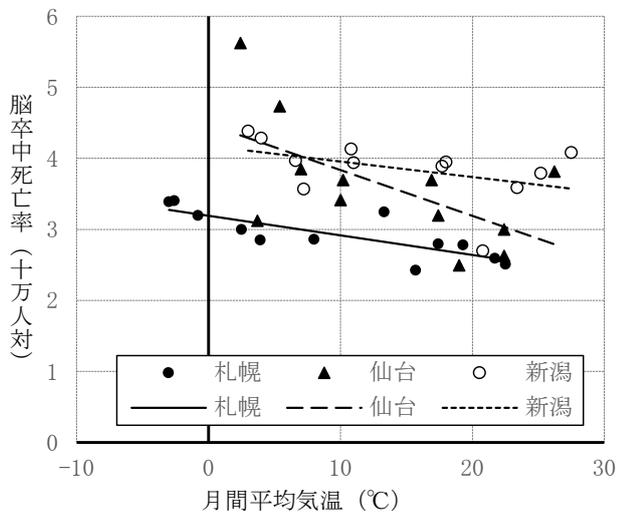
表4 各モデル都市の相関係数

モデル都市	相関係数	モデル都市	相関係数
札幌市	-0.80	大阪市	-0.66
仙台市	-0.60	広島市	-0.63
新潟市	-0.43	福岡市	-0.19
さいたま市	-0.72	熊本市	-0.58
名古屋市	-0.50		

ただし、図2によると、平均気温が低い北日本地域の都市では札幌市を除き脳卒中死亡率が高くなっており、モデル都市間の比較から、札幌市以外は月間平均気温が低いほど脳卒中死亡率は高くなるという仮説を支持する結果が得られた。

札幌市がその他の北日本地域の都市に比べて脳卒中死亡率が低かったことについて、札幌市のように常に低い気温が続き昼夜の気温の差が小さい都市では体が気温に慣れるため脳卒中死亡率が低くなるのではないかと考察した。

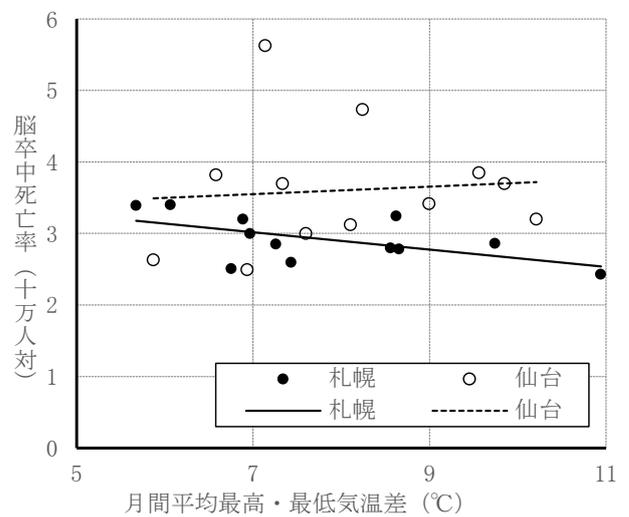
図2 月間平均気温と脳卒中死亡率



仮説 1-1 の検証結果を踏まえ、札幌市と仙台市の傾向の違いの原因として、1日の最高気温と最低気温の差が大きいことがあるのではないかと考えた。そこで、月間平均最高・最低気温差と脳卒中死亡率には正の相関があると予想し、仮説 1-2 を想定した。

仮説 1-2 : 月間平均の最高・最低気温差が大きいほど脳卒中死亡率は高くなる。

図3 月間平均最高・最低気温差と脳卒中死亡率



結果として、札幌市は負の相関を示し、仙台市はほとんど相関がなかった (図3)。したがって、1日の気温差は脳卒中の発症リスクに影響があるとは言えず、仮説 1-2 は検証されなかった。

### 気温と脳卒中の発症リスクの年単位比較

仮説 1-2 は検証できなかったので、平年よりも気温が低い年は脳卒中死亡率が高くなるのではないかと考えた。本研究では脳卒中死亡者数が多い1月に絞って2012年から2019年までの各都市の月間平均気温のデータを抽出し、1月の各都市の月間平均気温と脳卒中死亡率には負

の相関があると予想し、仮説 2-1 を立てた。

卒中死亡率には負の相関があると予想した。

仮説 2-1：月間平均気温が低い年は脳卒中死亡率が高くなる。

図 5 前年 1 月との月間平均気温の差と脳卒中死亡率

仮説 2-1 の検討のため、モデル都市のうち仮説 1-1 の検討で似た傾向を示したものから 1 都市ずつ、札幌市、仙台市、さいたま市、福岡市を選んで分析した。どの都市でも回帰直線の傾きは 0 に近く、相関は見られず、仮説 2-1 は検証されなかった (図 4)。

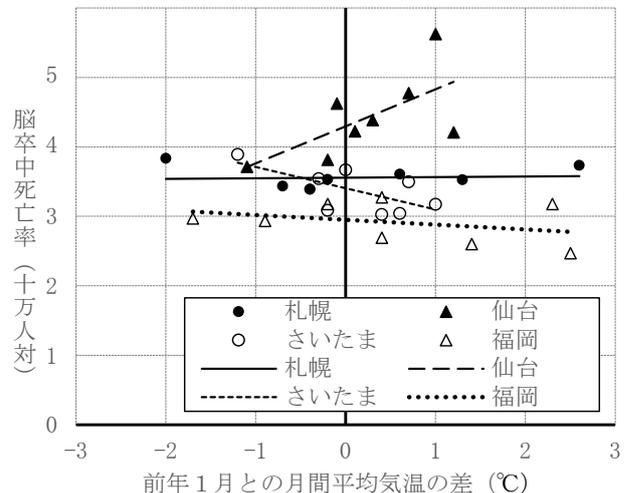
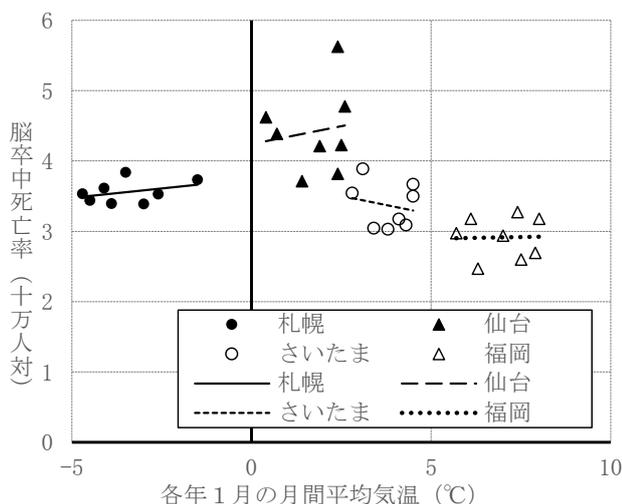


図 4 各年 1 月の月間平均気温と脳卒中死亡率



分析結果の散布図を図 5 に示す。さいたま市や福岡市は前年より気温が低いほど、脳卒中死亡率が上がっていた。対して仙台市は下がっており、札幌市はほとんど変化が無かった。都市によって結果にばらつきがあるため、一概に前年より気温が低ければ脳卒中の発症リスクが高まるとは言い難く、仮説 2-2 も成立するとは言い難いと考えた。

仮説 2-1 も検証されなかったもので、更に前年と比べて気温が低いときに脳卒中の発症リスクが高くなるのではないかと考え、仮説 2-2 を検討することとした。

### 暖房の使用と脳卒中の発症リスク

仮説 2-2：前年に比べて気温が低い年は脳卒中死亡率が高くなる。

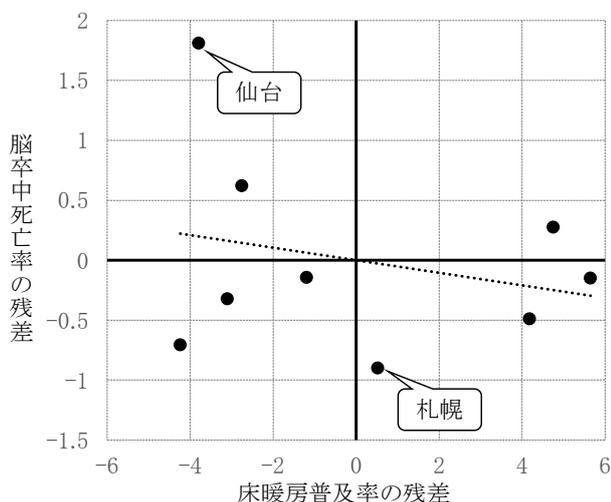
ここまで、外気温の変化によって脳卒中の発症リスクに影響が出るという推測に基づいて一連の仮説検証を行ってきた。しかし、ヒートショックの発生には室温も影響することから暖房の普及率を考慮して分析を行うこととした。暖房が完備された住宅では部屋ごとの室温差が小さいのでヒートショックが起こりにくいのではないかと考え、仮説 3-1 を立て、床暖房普及率と脳卒中死亡率には負の相関があると予想した。

仮説 2-1 と同様に 1 月のデータを用い、前年 1 月との月間平均気温の差 (前年 1 月の月間平均気温 - その年の 1 月の月間平均気温) と脳

仮説 3-1 : 床暖房普及率が高い都市ほど脳卒中死亡率は低い。

仮説 1-1 の分析から外気温は脳卒中死亡率に影響がすることがわかっていたので、2019年1月の各都市のデータを用いて外気温の影響を除いた床暖房普及率と脳卒中死亡率の偏相関分析を行った。その結果、偏相関係数は $-0.25$ となり、弱い負の相関を示した(図6)。

図6 外気温の影響を除いた床暖房普及率と1月の脳卒中死亡率



よって床暖房普及率が高い都市ほど脳卒中死亡率が低いことになり、仮説 3-1 は成立すると考えた。図6に札幌市と仙台市の偏残差プロットを示した。外気温の影響を除いたときに、札幌市の残差は正、仙台市の残差は負である。これから、札幌市は、床暖房普及率が、外気温の影響を除いてもそれ以上に高いことが分かる。このことは、札幌市の脳卒中死亡率が同じ北日本地域の都市である仙台市より大幅に低いことについての1つの説明要因となる可能性がある。

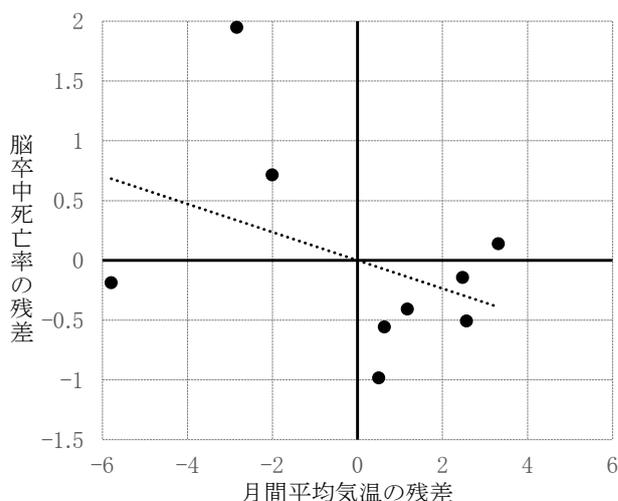
仮説 3-1 の検証より、外気温の影響を除いたとき、床暖房普及率と脳卒中死亡率には負の

偏相関があることがわかった。また、月間平均気温と床暖房普及率との相関関係を調べると、相関係数は $-0.43$ となり、負の相関であった。そこで、床暖房普及率の影響を除いたときに月間平均気温と脳卒中死亡率との相関が弱まるのではないかと考え、仮説 3-2 を立てた。

仮説 3-2 : 床暖房普及率の影響を除くと気温と脳卒中死亡率の相関は弱まる。

仮説 3-2 の検討のため床暖房普及率を第3変数として、仮説 3-1 と同様の偏相関分析を行った(図7)。

図7 床暖房普及率の影響を除いた1月の月間平均気温と脳卒中死亡率



結果として、床暖房普及率の影響を除く前の月間平均気温と脳卒中死亡率の相関係数が $-0.33$ であったのに対して、床暖房普及率の影響を除いた偏相関係数は $-0.40$ とより強い相関を示し、仮説 3-2 に反する結果となった。よって暖房という室内の要因の影響を除いても、外気温と脳卒中死亡率には高い関連性があると考えられる。

## 分析結果の要約と今後の課題

本研究で、気温および室温が脳卒中の発症リスクに大きく関わっていることがわかった。

気温の低い地域ほど脳卒中死亡率は高くなる傾向があることもわかった。これらは概ね予想通りだった。しかし、札幌市では仮説に反し、さほど死亡率が高くならなかった。この差には室温や防寒対策の程度など外気温以外の要素が影響していた。暖房器具の普及率（床暖房）が高いほど脳卒中死亡率は低かった。これは予想通りヒートショックが起こりにくいためと考えられ、札幌市の死亡率が抑えられていた理由も、他の北日本地域の都市と比べ暖房の設置数が多く、ヒートショックなどのリスクが低くなっていたためと説明がつく。しかし、暖房器具の普及率の影響を考慮しても、気温が低くなると脳卒中の発症リスクが高まることも示せた。

本研究の課題としては、地域生活様式による影響、例えば塩分の多い食生活や暖房以外の住環境などが十分考慮されていなかったことや、相関分析に留まり因果関係の解明ができなかったことなどが挙げられる。さらに、脳卒中の起きるメカニズムなど医学的アプローチができれば研究が進むと考える。

最後に、この研究は直接脳卒中の予防に応用できる訳ではないが、今回の分析で得られた傾向は、予防の全容を理解する第一の足掛かりとなり得ると考えている。

## <参考文献>

- (1) 栗田智久：“冬は脳卒中リスクが増大！気づきにくい2つの原因を医師が解説”、マイナビニュース（2016年）、  
<https://news.mynavi.jp/article/20160122-brain/>（2020年8月4日閲覧）

- (2) 大橋唯太：“急性循環器疾患の発症リスクと気象・気候変化との関係性について”、  
[https://www.jstage.jst.go.jp/article/ceispapers/ceis33/0/ceis33\\_301/\\_pdf/-char/ja](https://www.jstage.jst.go.jp/article/ceispapers/ceis33/0/ceis33_301/_pdf/-char/ja)（2020年8月6日閲覧）
- (3) 気象庁：“平年の日本の気候”、  
[https://www.jma.go.jp/jma/kishou/known/kisetsu\\_riyou/tenkou/Average\\_Climate\\_Japan.html](https://www.jma.go.jp/jma/kishou/known/kisetsu_riyou/tenkou/Average_Climate_Japan.html)（2020年8月5日閲覧）
- (4) 平成27年都道府県別年齢調整死亡率の概況（厚生労働省）、  
<https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/other/15sibou/dl/16.pdf>（2020年8月4日閲覧）

# 世界一の桃田選手をデータ解析！高校バドミントンに活かせるか 要旨

バドミントンのシングルスで勝つために必要な要素を見つけることを課題としている。本研究ではシングルスにおける強みの発見を目的とし、その達成のために桃田賢斗選手のショットの速度率をデータ化し、プロのプレーから特徴を見つけ、高校生のデータと検定を用いて比較した。データ解析の結果、テンポ率には有意差がなかったため、プロの速度率を活用したプレーを高校生が実践することでシングルスの技術向上を目指した。この成果はバドミントンをしている高校生にとって、シングルスで勝利するために必要な要素を発見・立証したと考えている。

## 研究動機

本校のバドミントン部は団体戦で勝てないことが課題である。そこで団体戦の要であるシングルスに勝つことを目的とし、シングルスで勝つための要素を発見しようと考えた。

## 研究手法

- 使用する桃田選手のデータ
- ・YouTube上の桃田選手の試合データと高校生の試合をビデオ撮影したものから「羽の滞空時間」、「ショットの場所」についてデータを採取
  - ・「羽の滞空時間」はストップウォッチを使用し手動で計測
  - ・「ショットの場所」はコートをも9分割することでデータ化



図1:ビデオのデータ化の様子

	桃田	相手	合計
ショットによる得点	34	31	65
ミスによる得点	25	25	50
制したラリー数	59	56	115
ショット数	674	683	1357

図2:採取したデータ数のまとめ(桃田選手+高校生)

高校生データ  
合計ショット数1251

総データ数 計2608

## データ採取 → データ解析

### 研究内容①「桃田選手と相手選手のテンポ比較」

- ・桃田選手の滞空時間のデータを解析した
- ・各選手の最後のショットの滞空時間を比較
- ・得点された選手(ラリーの敗者)に対する得点した選手(ラリーの勝者)のショットの滞空時間の割合を計算
- ・得点するために必要となるショットの滞空時間の割合を算出

ラリー番号	ラリーの勝者	ラリーの敗者	ラリーの勝者/敗者
1	0.88	0.79	1.11
2	0.71	0.29	2.45
3	0.45	0.89	0.51
4	1.03	0.98	1.05
5	0.49	1.03	0.48

図3:ラリーの勝者と敗者の滞空時間の割合

### 研究内容②「桃田選手の有効打と有効打以外の比較」

- ・滞空時間のデータを使用
- ・各ラリーで得点した選手のショットの滞空時間のデータのみ使用
- ・得点する直前の2ショット(有効打)とそれまでのショット(有効打以外)の滞空時間の変化に着目
- ・得点につながる際のショットの滞空時間の変化を算出



図4:有効打と有効打以外の説明(1ラリー抜粋)

### 研究内容③「場所ごとのテンポの比較」

- ・桃田選手の滞空時間と場所のデータを使用
- ・場所ごとに有効打と有効打以外の滞空時間を比較

## 参考文献

データ採取に用いた動画  
<https://www.youtube.com/watch?v=sZAAmtQ4VBY>  
<https://www.youtube.com/watch?v=WICT4vXAerM>

謝辞

『情報・システム研究機構 統計数理研究所 医療健康データ科学研究センター』

## データ解析結果・考察

### 解析結果

### 考察

- 研究① 相手よりも速いテンポは66%の確率で点数につながっている → 「得点になるラリーは相手選手より滞空時間が短いショットである」という研究は正しい
- 研究② 滞空時間を0.83倍に変化させると得点につながる → 「得点する前にはショットのテンポが速くなっている」という研究は正しい
- 研究③ 場所ごとに比較しても有効打のほうが速い → 「場所についてデータを分けても有効打の方が速い」という研究は正しい

ラリーの勝者	早いとき	遅いとき	合計
ラリー数	35	18	53
ラリーの割合	66.0%	34.0%	100.0%

図5:研究①相手選手に対するショットのテンポの内訳

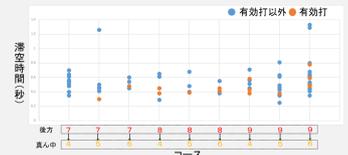


図6:研究③コースごとの有効打と有効打以外の比較

### 研究① 桃田選手と相手選手の「ショットのテンポ比較」

より速いテンポで打つ必要がある

### 研究② 桃田選手のショット「有効打と有効打以外の比較」

0.83倍のテンポが有効打になる

### 研究③ 桃田選手のショット「場所ごとの比較」

場所ごとの比較でも有効打のほうが有効打以外より速い

## 結果からの高校生の実践による検証

### 実践内容

### データ採取

高校生と桃田選手の滞空時間のデータを使用

### データ解析

高校生と桃田選手の様々な場面のデータを可視化し、t検定を用いて比較

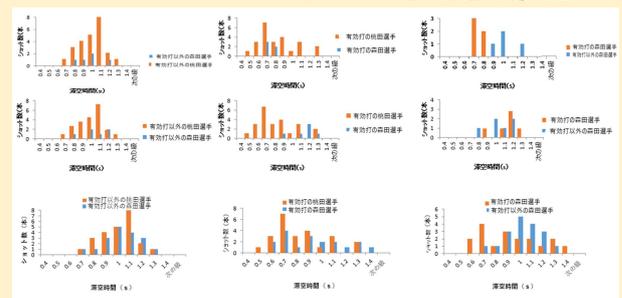


図7:各試合のショットの分布の比較

0.92	0.96	0.71	0.75	0.92	0.71
0.15	0.14	0.05	0.21	0.15	0.05
なし	なし	なし	あり	あり	なし
1.00	0.96	1.10	0.75	1.00	1.10
0.15	0.14	0.13	0.21	0.15	0.13
なし	あり	あり	なし	なし	なし
0.99	0.96	0.89	0.75	0.99	0.89
0.14	0.14	0.24	0.21	0.14	0.24
なし	なし	なし	あり	あり	なし

図8:図7に対応した平均、標準偏差、有意差

平均A	平均B
標準偏差A	標準偏差B
有意差	

## 検証結果

### 結果に基づく練習メニュー作成

- ・予想通り、有効打における桃田選手と高校生のテンポに有意差があった
- ・高校生のデータでも研究内容②のテンポの比率に有意差があった。
- ・有効打と有効打でない打つテンポが0.8倍の比率となるように意識すれば、シングルスが強くなると考えた。

## 展望 研究を利用した新たな練習方法の考案(実施中)

従来、1セット(1分間)プッシュ→スマッシュのノックは、スマッシュを15回打つテンポでノックを実施していた。そのノックの5セット目(1分間)でスマッシュを18回打つテンポでノックを実施している。(テンポ差を身体に覚えさせることを目的としたノック)

# 機械学習を用いたヒット曲予測AIの構築

## 素朴な疑問

日本国内では毎年3000曲以上の楽曲がリリースされるけど、「ヒットする」楽曲には特徴がないのかな？

楽曲にはいろいろな特徴量がある  
メロディ・ハーモニー  
リズム・歌詞など

コード進行だけでヒットするかどうか予測できないか？！

過去のヒット曲の「コード進行」と「ランキング結果」のデータがある。

コード進行って「複雑」そうにみえて規則性がある？  
コード進行にはある一定の「パターン」がありそう？  
ヒットする楽曲のコードには「特徴的な要素」がありそう？

統計手法を用いて数値化

## 予測モデル構築

ニューラルネットワーク  
重回帰分析

## 機械学習の仕組み

ある曲のコード進行  
F G Em7 Am7 F E7  
Am7 Gm7 C7 F G

コード進行とランキングのデータ入手の自動化

予測モデル

予測モデルの自動更新

ランキングにつながる予測値を算出

ヒット曲予測AI

## 1. 研究動機・目的

ある楽曲がヒットするかの予測ができれば、音楽を創る上で参考にできて有用である。また音楽産業にとって、どのような楽曲を売り出すかの戦略を決める上で有益である。このような背景のもと、本研究は日本国内で毎年3000曲以上リリースされるシングルCDの楽曲中で、楽曲ランキング上位となる楽曲の法則を「コード進行」だけで発見し、そのランクインを予測することを目的とする。

## 2. 楽曲データベース

【ヒット曲の収集】 Japan's Billboard Year-End Hot 100 2010~2020年 各年上位20曲  
【コード進行の取得】 U-FRET サビ部分のみ 長調はハ長調、短調はイ短調に移調

## 3. 分析手法

2010年から2019年のヒット曲のコード進行を3つの観点で分析

### 【第一分析】

コード進行の複雑性を数値化  
クラスタリングを行い、得られた複数のクラスタ中心からの距離によって特徴の数値化

### 【第二分析】

コード進行のパターンを数値化

### 【第三分析】

特徴的なコード進行を数値化  
トピック分析によって求めたトピック分布の類似度によって数値化

使用した分析手法

主成分分析・クラスタ分析

N-gram解析・クラスタ分析

トピックモデル分析(LDA)

1 3つの特徴量を説明変数、順位を目的変数  
ヒット曲(順位)を予測する回帰式の作成

重回帰分析法

2 ニューラルネットワークの設計  
入力:3つの特徴量 出力:順位の逆数

ニューラルネットワーク

## 4. 分析

### 第一分析 コード進行の複雑性の分析

手順①

コード種類数  
コード進行中に出現するコードの種類

非ダイアトニックコード率  
コード進行に含まれるダイアトニックコードではないコードの数のコード進行の系列長に対する比率

非繰り返し率  
長さ5以上のコードの繰り返し単位に含まれないコードの数のコード進行の系列長に対する比率

主成分分析

2次元化

寄与率

第一主成分 0.5447  
第二主成分 0.2860

手順② 主成分得点をk-meansクラスタリングを行う。

- ・エルボー法によりクラスタ数を4に決定
- ・クラスタごとに分類(図1) ⇒  $f_A(X)$  の値を算出

結果

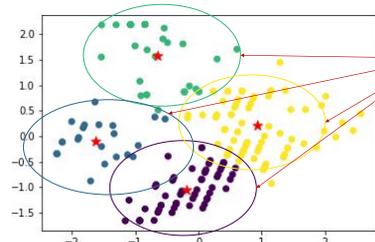


図1 クラスタリング結果

### 第二分析 コード進行のパターンの分析

手順①

ヒット曲に含まれるコード進行の典型的なパターン分析

2010~2019年の各楽曲ごとに連続する3つのコード(トライグラム)の出現回数を調べる。

トライグラムとは(例)

A,B,C,D,B,A,B

(A,B,C), (B,C,D), (C,D,B), (D,B,A), (B,A,B)とわかる

手順②

それぞれの楽曲に特有なトライグラムを見つける

トライグラムを単語、各楽曲を文章とみなしてtf-idf値を算出。

手順③

各楽曲のtf-idfの平均と標準偏差をk-meansクラスタリングを行う。

- ・エルボー法によりクラスタ数を3に決定
- ・クラスタごとに分類(図2) ⇒  $f_B(X)$  の値を算出

結果

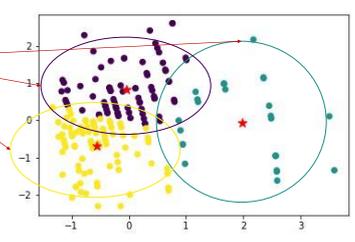


図2 クラスタリング結果

それぞれ異なった特徴を持つヒットしやすい楽曲のクラスタを表す

楽曲数が多いクラスタ → よりヒットする傾向が高いコード進行を持った楽曲のクラスタ

ある楽曲について「ヒットしやすさ度合い」

$$f_A(X), f_B(X) = (\text{所属したクラスタの楽曲数}) \div (\text{中心との距離})$$

### 第三分析 特徴的なコード進行の分析

- 手順① 楽曲のコード進行から  
7割以上出現・10個以下しか  
出現しないコードを削除した。
- 手順② トピック数は評価指標である  
PerplexityとCoherenceによって  
7トピック分けることを決定した。
- 手順③ 楽曲を文章、コードを単語とみなして、  
LDAを実行した。(図4)



図4 トピックの可視化

### 結果

過去のヒット曲のトピック分布と新しい楽曲のトピック分布の類似度を計算することで、ヒットしやすさ度合いを数値化

分析データ中の全ての楽曲(200曲)のトピック分布  $p_n(\theta)$ ,  $n = 1, \dots, 200$  の総和

$$\alpha_k = \sum_{n=1}^{200} p_n(\theta_k)$$

過去のヒット曲のトピックの相対的な偏り  
求めたトピックの相対的な偏りと、  
あるコード進行  $X$  のトピック分布  $p(\theta)$  の  
重み付け和

$n$ : 楽曲のインデックス  
 $\theta_k$ :  $k$ 番目のトピック

$$f_C(X) = \sum_{k=1}^7 \alpha_k p(\theta_k)$$

「ヒットしやすさ度合い」 $f_C(X)$

過去のトピック分布の  
相加平均を  $q(\theta)$  とする

$$f_C(X) = 200 \sum_{k=1}^7 q(\theta_k) p(\theta_k)$$

トピックはヒット曲の  
特徴を反映している  
・同様なトピックの偏りがある  
・特徴が似ている  
⇒ ヒットしやすい

過去のヒット曲のトピック分布とトピック分布  $p(\theta)$   
の類似度を内積によって求めている

## 5. 結果

コード進行の特徴分析によってヒット曲予測するシステムの構築

曲の特徴量  $f_A(X)$ ,  $f_B(X)$ ,  $f_C(X)$

楽曲が音楽チャートにランクインする  
順位の推定につながる予測値を出力

算出した予測値の相対順位と  
実際の順位と比較

### 重回帰分析

説明変数:  $A \cdot B \cdot C$  の値 目的変数: 各曲の順位  $r$   
として, 回帰式  $r = \beta_A \times f_A(X) + \beta_B \times f_B(X) + \beta_C \times f_C(X) + \epsilon$

を用いて, 係数及び切片を最小二乗法により求めた。

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
$\beta_A$	-0.631133775	0.404479353	-1.560360919	0.120287802	-1.42882418	0.16655663	-1.42882418	0.16655663
$\beta_B$	-0.05637123	0.404101941	-0.139497548	0.889200229	-0.853317326	0.740574866	-0.853317326	0.740574866
$\beta_C$	-0.988339587	0.404564946	-2.442968915	0.015452785	-1.786198794	-0.19048038	-1.786198794	-0.19048038
$\epsilon$	10.5	0.402858086	26.06376876	1.31346E-65	9.705506962	11.29449304	9.705506962	11.29449304

求めた回帰式を使い, 2020年のランキング上位20曲の  
サビの部分のコード進行を用いて, 順位を予測する。

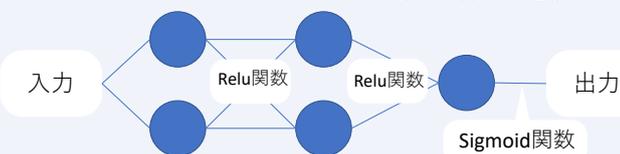
### 2020年ランキングの予測結果

順位の推定に  
つながる予測値

楽曲	アーティスト名	正解	重回帰	NN	平均
楽曲A	アーティスト I	1	8.0	2.6	5.3
楽曲B	アーティスト II	2	6.9	5.5	6.2
楽曲C	アーティスト III	3	11.2	6.8	9.0
楽曲D	アーティスト IV	4	9.3	7.5	8.4
楽曲E	アーティスト V	5	10.8	6.9	8.9
楽曲F	アーティスト VI	6	11.0	5.9	8.5
楽曲G	アーティスト VII	7	9.3	3.7	6.5
楽曲H	アーティスト VIII	8	12.2	7.0	9.6
楽曲I	アーティスト IX	9	11.6	6.7	9.2
楽曲J	アーティスト X	10	11.5	6.3	8.9

### ニューラルネットワーク

入力: 曲の特徴量  $f_A(X)$ ,  $f_B(X)$ ,  $f_C(X)$  出力: 順位の数値

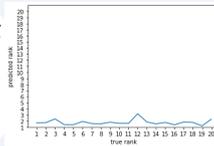


### 補足

サポートベクトル回帰

各楽曲の特徴量の分布が  
非線形である

SVRIによる順位予測



楽曲の予想順位に  
明確な差が  
現れなかった。

## 6. 考察

コード進行の特徴にもとづいた3種類の分析結果を用いてヒット曲の予測を行い, ランキングの1位と2位の  
ヒット曲の予測を行うことができた。

楽曲A	アーティスト I	1	8.0	2.6
楽曲B	アーティスト II	2	6.9	5.5

重回帰分析で最も上位と予測された楽曲  
楽曲Bの予測値6.9 ⇒ 他楽曲と相対的に比較すると1位  
ニューラルネットワークで最も上位と予測された楽曲  
楽曲Aの予測値2.6 ⇒ 他楽曲と相対的に比較すると1位

⇒ 楽曲A~Jで2つの予測値の平均値の  
相対的な順位は, 実際の1位, 2位と一致した!

一方で

その他のランクの予測の精度は不十分

### 改善案

回帰を二段階で行い, 各分析でランキングとの単回帰  
分析を行った結果を使って重回帰分析を行った

t検定のp値0.93であり  
有意差は見られなかった

### 分析精度の向上

分析に用いた楽曲を増やす

サビ以外のコード進行と  
メロディを用いる

ヒットしなかった楽曲を  
含めて分析する

### 今後

コード進行の移調に専門知識が必要 ← 自動化

分析用楽曲データの差し替え → ある条件に沿った  
予測が可能

## 7. おわりに

本研究ではコード進行によってどの程度ヒット曲が予測できるかを  
明らかにするために, ヒット曲のコード進行の特徴量を計算し,  
特徴量とランキングの順位の間での予測を行い, コード進行から未  
来のヒット曲を予測するシステムを構築した。今後, コード進行以外  
の要素も考慮したヒット曲予測システムを構築できるように研究を  
発展させたい。

### 【参考文献】

- [1] Pachet, F.: Hit Song Science Music Data Mining (Chap. 10), Taylor & Francis (2011).
- [2] Dhanaraj, R. and Logan, B.: Automatic Prediction of Hit Songs. Proceedings of the International Conference on Music Information Retrieval, pp. 488-491 (2005).
- [3] Ni, Y., Santos-Rodríguez, R., McVicar, M. and Bie, T. D.: Hit Song Science Once Again a Science?. Proceedings of the 4th International Workshop on Machine Learning and Music, pp. 1-2 (2011).
- [4] Herremans, D., Martens, D. and Soeren, K.: Dance Hit Song Prediction. Journal of New Music Research, Vol. 43, No. 3, pp. 291-302 (2014).
- [5] 川井豊大: ギターコードから見る J-POP の特徴の統計解析, 南山大学卒業論文要旨 (2005).
- [6] 鶴田 兼: ギターコードから見るヒット曲の違いに関する統計的分析, 南山大学卒業論文要旨 (2009).

令和4年3月発行

兵庫県立姫路西高等学校

〒670-0877

兵庫県姫路市北八代2丁目1番33号

Tel 079-281-6621

Fax 079-281-6623

