

高3選択科目「データ科学」での授業実践

馬場 国博・慶應義塾湘南藤沢中・高等部

〒252-0816 藤沢市遠藤 5466

TEL (0466)49-3585 FAX (0466)47-5077

E-mail: baba@sfc-js.keio.ac.jp

1. はじめに

本校では、高校3年生を対象にした2単位分の選択必修科目（学校設定科目）を18講座用意しており、文理のコースに拘わらず、1学年6クラス約240人の生徒が原則希望した科目を履修する。数学科では毎年「データ科学」という講座を開設しており、コンスタントに10名前後が履修する。

担当者により取り扱う内容は異なるが、R言語を用いてデータを解析する力を身につけさせるという目的は一貫している。全生徒を対象にしては、関数電卓やそのiPadアプリケーション版を用いて簡易的にデータの分析を学習させているが、学習意欲のある履修者を対象にデータ解析の基礎を身につけさせようとするとき、また大学進学後の統計教育を見据えたとき、データサイエンス分野で広く使用されているR言語を使用することが望ましいと考えている。

この科目は、大学・大学院で統計学を専攻した3名の数学科教員が交代で担当しており、8名の履修者を迎えた2020年度は3年ぶりに発表者が担当した。

前半は、2018年度に新設された「統計データ分析コンペティション」（総務省統計局、統計センター、統計数理研究所、日本統計協会の共催）への応募を前提とした授業展開を行い、後半は多変量解析の幾つかの手法を扱った。

2. オンライン授業による「統計データ分析コンペティション」への挑戦

今年度はコロナ禍により4月～7月が完全オンライン授業となった。その状況下で、自宅で自分のペースで作業を進めることができるであろうことから、年度の前半は「統計データ分析コンペティション」（以下「コンペ」と表記）への応募を前提とした授業展開を行うことを決定した。

この「コンペ」は、都道府県や市町村、県庁所在市別の公的統計を集めた3種類の教育用標準データセット(SSDSE)を用いた統計分析によりアイデアと解析力を競うもので、A4用紙で2～8ページにまとめる。

(1) 授業展開

① 4月～5月中旬

まだオンライン授業の態勢も整わない中で、9月10日締切りの「コンペ」に全員が応募することを宣言し、HPに掲載されている2019年度を受賞論文を読み込み、使用するSSDSEを眺めて、自分の論文テーマ案を考え始めるよう指示した。

② 5/21

各自の自宅にあるPCにRをインストールさせた上で、Google Meetを通じてオンライン(リアルタイム)授業を開始した。

Rの基本文法、基本統計量の算出、グラフ表現(ヒストグラム、箱ひげ図、茎葉図、パレート図など)

③ 5/28 ～ 6/11

SSDSEのRへの取り込みを行った後、2019年度に統計活用奨励賞を受賞した大段利々子さん（広島大学附属高等学校）の論文「日本で暮らす外国人の動向から見た他民族化」をなぞるように、同論文に掲載されたグラフや統計量をRで作成しながら、Rの使用方法やデータの見方を学ぶ授業(1回2時間)を3回にわたって行った。

具体的には、データの加工、行名札や列名札の管理、散布図、回帰直線、決定係数、for文、加工したデータのcsvファイルでの書き出し、「jSTAT MAP」を利用した地図表現など

④ 6/18 ～ 7/2

各自のテーマに沿ったデータ分析の演習時間を3回とり、Meetを通じて個別に質問対応を行った。

⑤ 7/9, 16

1人10分ずつ中間発表を行い、発表後には履修者が互いに疑問点や改善点を指摘しあう時間をとった。

⑥ 7/30

中間発表を受けて、夏休み中の論文作成にあたって履修者と Meet により個別に助言を行った。

⑦ 8月

夏休み中には必要に応じて個別に相談に乗る対応を行った。他の変数の影響を除いた2変数間の関係について助言を求められ、偏相関についての説明やRでの算出方法について指導することもあった。

⑧ 9/3

初めて対面で履修者と顔を合わせ、完成間近の論文について個別に助言を行った。全員が9/10までに論文の応募を行った。

(2) 応募論文

8名の履修者が応募した論文タイトルは以下の通り

- ・学習、SNS いじめ、労働環境から見たいじめの地域格差
- ・教育分野における地域間格差
- ・大都市「東京」がもたらした副作用
- ・定住人口の増加策
- ・外国人の暮らしやすさとグローバル化に向けて
- ・健康3原則からみる糖尿病
- ・投票率の決定要因に関する分析
- ・自治体ごとのふるさと納税の必要性を定義する

SSDSEを用いた分析を行うことから、地域間格差を扱う類似したテーマが多くなる点は致し方ないところと考える。その中でも独自の着想に基づいた論文も散見された。

結果としては、最後に記載したテーマの岡本涼夏の論文が「統計活用奨励賞」を受賞することができた。

(3) 反省点

様々な種類のグラフ表現や統計量を算出してデータ分析を行っていることから、大段さんの論文をなぞる形で、Rによるグラフ作成、統計量算出の手法を学ばせた。論文のイメージをつかませる意味でも有用であったと考えるが、どうしても数人の生徒はこの論文に引けられて、似たようなテーマとなってしまった。

論文作成時のグラフ作成ソフトはRに限定せず、

Excel等でも可としたが、テーマ選び、Rの文法の習得、論文の作成を5ヶ月の間に行わせることは時間的に厳しかった。

また、オンライン授業であったため、個人のPCで解析を行う形となり、生徒間で教えあうような機会をもつことができず、またWindowsとMacが混在するなど個別対応が必要なことも多かった。

生徒諸君がオンライン授業というコミュニケーションが取りづらい環境下でありながら、半年間でデータ分析を行い論文を作成するというハードな取り組みに真摯に取り組んでくれたことを評価していただいたのか、このコンペにおいて学校賞を受賞することができた。またこの授業とは関係しないが、本校高1の生徒である朝倉翔汰君が自主的に「人口増加と『住みやすい街』の関係」と題した論文でこのコンペに応募し、総務大臣賞を受賞することができた。

3. 多変量解析の学び

10月～11月の2ヶ月では、対面授業で以下の多変量解析の手法を扱った。

・重回帰分析

回帰係数に対する検定、偏相関係数、AIC・自由度調整済み決定係数に基づく変数選択

・主成分分析

・正準判別分析

最終授業で実データに対して各手法を適用させる実技試験を行ったところ、各手法の意味や有用性についてよく理解してくれたことが確認できた。中には1学期に教えてもらえたら論文作成にも活かされたのに、といった感想も寄せられた。一方で、各手法の理論的側面については理解を伴わない部分も多かった。

4. 最後に

「コンペ」に応募することを前提とした授業展開は、生徒にかなり大きな負担がかかるものの、学習のモチベーションや、データ分析力、論文作成力を身につけさせる点では得るものも大きいことがわかった。なお、本研究は慶應義塾学事振興資金による共同研究「初等中等教育段階におけるデータサイエンスプログラムの開発」の助成を受けて行った。