JSiSE Research Report
vol.33,no.6(2019-3)

政府統計からデータ処理向け教材を生成する手法の検討

吉根 勝美[*1]
*1 南山大学

Generating Data Processing Teaching Materials from Government Statistics

Katsumi Yoshine[*1]
*1 Nanzan University

社会科学系の大学新入生には，問題発見・解決能力の向上のため，政府統計の収集と分析のスキルを獲得させたいので，新入生対象の情報教育の一環としての表計算ソフトウェア指導では，実際のデータを教材として使用したい．政府統計は，調査方法やデータの読み方のコツなどがそれぞれに異なるが，新入生には多数の政府統計に接触させたい．本稿では，データ処理向けの教材作成に目的を特化して，教材生成の手法を検討する．

キーワード: 政府統計，教材開発，データ処理

1. はじめに

大学生に対する問題発見・解決能力の育成の重要性がうたわれる中(1)，そのための教材開発では，学生の学年ばかりではなく，学問分野も考慮する必要がある．すなわち，学生の成長過程に応じて教材の難易度を上げていくことは当然として，学生が所属する学部の専門性に合わせた教材開発が求められる．

社会科学系の大学新入生には，問題発見・解決能力の基礎的なスキルとして，政府統計の収集と分析のスキルを獲得させたい．例えば，私情協の分野別「学士力考察」によると，経済学教育における到達目標のひとつとして，「経済データを活用して経済の状態を正しく理解するために、経済指標の背景を理解し、自ら適切なデータの収集、加工ができるようにさせなければならない。そのため、経済指標が生成される背景を理解し、経済データを種々のデータベースから取得し、それを統計理論に基づいて実証分析できることを目指す」と示している(2)．

また，大学生に対する統計教育において実際のデータを使用することの重要性は，アメリカ統計学会による「統計教育における評価と指導のガイドライン」の College Report 2016 で示されている(3)．同ガイドラインでは 6 項目の推奨事項が提示されており，その中のひとつ Recommendation 3: Integrate real data with a context and a purpose. の前提となるのが“Use real data”である．データの“reality”についても，まったく文脈を持たないデータ(Naked data)から始まり，適当な意味付けをして現実のように見えるデータ(Realistic data)，小テストと試験の成績のように実際のデータ(Real data)ではあるが，特に応用が見込めないデータを経て，実際の研究から得られたデータ(Real Data, from a Real Study)に至るまで，“reality”のスペクトルを，同ガイドラインの中で説明している．

新入生対象の情報教育の一環として，表計算ソフトウェアの指導が行われることが多く，その中で使用するデータに“reality”を求めるには，社会科学系の学生が対象ならば，行政機関や地方自治体が発表している公的な統計データを使用することが効果的である．実際，南山大学経済学部の 1 年次必修科目「データ処理入門」における表計算ソフトウェアの実習においては，愛知県統計年鑑のうち「百貨店・スーパーの事業所数，従業者数及び販売額等」や，賃金構造基本統計調査のうち「企業規模別新規学卒者の初任給の推移」などを利用して教材を作成している．

統計教育の及ぶ範囲は大変広い上に，個別の統計に依存する部分が多い．例えば，表 1 に示した大学基礎統計学の知識と問題解決力を測る統計検定 2 級の出題範囲は，記述統計から推計統計までの統計学の内容に加え，統計ソフトウェアの活用も含む広範囲にわたる[4]．さらに，実際の統計表を読み取る能力（統計リテラシー）も求められ[5]，経済統計というジャンルでは，個別の統計ごとに，その目的，調査方法，作成方法があることを理解し，データが持っているクセやデータの読み方のコツも個別に学ぶ必要がある[6]．

表 1 統計検定 2 級の出題範囲

大項目	小項目
データソース	身近な統計
データの分布	データの分布の記述
1変数データ	中心傾向の指標
	散らばりなどの指標
	中心と散らばりの活用
2変数以上のデータ	散布図と相関
	カテゴリカルデータ
データの活用	単回帰と予測
	時系列データの処理
推測のための データ収集法	観察研究と実験研究
	標本調査と無作為抽出
	実験
確率モデルの導入	確率
	確率変数
	確率分布
推測	標本分布
	推定
	仮説検定
線形モデル	回帰分析
	実験計画の概念の理解
活用	統計ソフトウェアの活用

広範囲に及ぶ統計教育を大学新入生対象に実施するには教育内容を絞る必要があるが，表計算ソフトウェアを使えるようにすることは，適切な目標設定の一つである．なぜなら，高校数学で記述統計を学習はしているものの，表計算ソフトウェアを利用する統計計算を経験しているとは限らないからである．また，大学初年次終了までに表計算ソフトウェアが使える状態になっていれば，その後の大学における学習や研究においてデータ分析を行う必要が生じても直ちに実行できるという利点もある．

実際の公的な統計データから表計算ソフトウェア実習のための教材を作成する際，インターネットを通じてダウンロードした公的な統計データのファイルをそのまま使用することはできないので，何らかの加工を施す必要がある．実際，愛知県統計年鑑のうち「百貨店・スーパーの事業所数，従業者数及び販売額等」の統計表から，百貨店の欄から紳士服・洋品，婦人・子供服・洋品，その他の衣料品，身の回り品，飲食料品，その他の販売額について，1 月～12 月のデータを抽出したものを，表計算ソフトウェアの基本的な操作方法を学ぶための教材とした．また，賃金構造基本統計調査のうち企業規模別に作られている「企業規模別新規学卒者の初任給の推移」の統計表から，昭和 51 年以降の男女別大卒初任給額を，企業規模 1,000 人以上，100～999 人，10～99 人それぞれの統計表から抽出して，ひとつの表にまとめたものを，表計算ソフトウェアにおける時系列データ処理を学ぶための教材とした．

公的な統計データから教材を作成するときは，統計データが更新されるたびに教材も再作成するべきであり，教材作成のために使用する公的な統計データ自体も常に見直すべきであるが，これらは教材作成者には大変な負担である．統計データが更新されるたびに統計データのファイルをダウンロードし，ファイルを加工する必要があるが，フォーマットが同じである保証はないため，手作業で教材を作成しているのが実情である．また，教材作成に使用する公的な統計データをひとつに決めたとしても，その中には複数の統計表が含まれ，ひとつの統計表にも多数の調査項目があることが多いので，実習にふさわしいデータを見つけ出すには多くの時間を割く必要がある．

本稿では，教材作成者の負担軽減もひとつの目的としつつ，大変幅広い統計教育のうち，社会科学系新入生を対象とする表計算ソフトウェアにおけるデータ処理の指導に特化して，政府統計からデータ処理向け教材を簡便に開発する手法を明らかにすることを目的としている．ここでは，データ分布の可視化実習を想定して，実際のデータとして総務省統計局による小売物価統計調査年報 平成 29 年のうち「調査品目の月別価

格及び年平均価格【県庁所在市及び人口 15 万以上の市】」を用いて，ヒストグラム作成と散布図作成のそれぞれにふさわしいデータを見つけるためのインターフェースを試作し，この試作を通じて，教材開発の簡便な手法を検討する．

2. 前処理

本章では，実際の公的な統計データから教材開発に至る過程における最初の段階の具体例として，小売物価統計調査（動向編）から，教材開発のための元ファイルを作成する手順を説明する．この調査は総務省統計局が実施しており，消費者物価指数やその他物価に関する基礎資料を得ることを目的として，国民の消費生活上重要な財 700 以上の品目の小売価格を，1950 年 6 月から毎月調査している(7)．

本稿では，データ分布の可視化実習を仮定して，同調査の年報のうち「調査品目の月別価格及び年平均価格【県庁所在市及び人口 15 万以上の市】」の統計表を使用して，特定の品目の年平均価格について 81 都市ごとの価格の分布を可視化するための教材を作成する状況を想定する．データ分布の可視化実習の具体例として，ヒストグラム作成と散布図作成の２つを想定し，それぞれにふさわしいデータを 700 以上ある品目から見つけやすくするインターフェースを試作する．

この統計表を総務省統計局ウェブサイトからダウンロードには，エクセルファイルをダウンロードする方法と，データベースにアクセスして検索条件を指定して得られた結果をダウンロードする方法があるが，ここでは後者の方法でデータを入手する．前者の場合，県庁所在市及び人口 15 万以上の市にて調査対象となっている品目の月別価格及び年平均価格が 32 個のファイルに分割されて格納されているので，これらをすべてダウンロードしなければならない．後者の場合，データベースにアクセスして，表示項目選択で時間軸（年・月）に「2017 年」だけを指定し，レイアウト設定で，行に「地域」，列に「銘柄（H27 年基準）」をそれぞれ指定すると，調査対象となっている品目の年平均価格の地域別一覧表が得られるので，これをダウンロードする．

ダウンロードした得られたデータに含まれる欠損値はすべて空欄とする．この統計では，“***”（数字が得られないもの），“-”（調査銘柄の出回りがなかったもの），“...”（当該市町村で調査を行わないもの，又は調査期間の定めがあるため調査を行わないもの）の 3 種類の欠損値があり，これらを置換機能で空欄にした．

これまでの処理を施した結果をエクセルファイルとして保存し，その一部を図 1 に示す．1 行目は各列の名称であり，1 列目以外は調査品目を表しており，調査品目ごとに 4 桁のコードがついている．例えば，2 列目の調査品目は「1001 うるち米(単一品種，「コシヒカリ」)」である．このように品目によっては銘柄まで指定されることがある．2 行目～82 行目は，県庁所在市及び人口 15 万以上の市の 81 都市を表している．空欄は欠損値を表している．

	A	B	C	D	E	F	G	H
1	地域	1001 うる	1002 うる	1011 もち	1021 食パ	1022 あん	1023 カレ	1031 ゆで 1
2	札幌市	2408	2092	513	473	78	126	377
3	函館市	2261	2270	497	299	75	97	
4	旭川市	2211	1915	503	443	75	116	
5	青森市	2344	2052	607	424	89	94	380
6	盛岡市	2330	2032	580	378	78	98	323
7	仙台市	2126	1921	623	386	80	92	406
8	石巻市	2291	1897	624	421	78	105	
9	秋田市	2440	1969	515	326	80	98	322
10	山形市	2224	2000	610	442	82	95	384
11	福島市	2328	2147	627	399	86	104	298
12	郡山市	1968	1972	638	240	78	90	
13	水戸市	2073	2069	570	518	81	88	365
14	日立市	1952	1932	524	414	76	86	
15	宇都宮市	2230	1996	537	337	90	111	342
16	足利市	2171	1969	512	430	84	109	
17	前橋市	2233	1993	476	238	76	92	469
18	さいたま市	2504	2184	625	344	94	101	358

図 1 前処理済みのデータファイル

3. インターフェースの試作

本章では，前章で作成した前処理済みのデータファイルから，ヒストグラム作成と散布図作成のそれぞれの実習にふさわしい品目を見つけやすくするためのインターフェースを試作し，検討する．試作環境は Windows 10 Home で動作する RStudio Desktop Open Source Edition (Version 1.1.456)であり，使用言語は R (version 3.5.1)である．

3.1 R 言語によるインターフェース試作

インターフェースの試作は，合わせて 100 行未満の R 言語のプログラムで実現した．エクセルファイルを読み込むパッケージ readxl とインタラクティブな Web アプリケーションを容易に構築することができ

るパッケージ shiny を使用している．

インターフェースのプログラムは，2 つのファイル ui.R と server.R からなる．ui.R では，インターフェースの画面を以下のように定義する．

```
shinyUI(
 navbarPage("小売物価統計調査(2017 年平均)",
  tabPanel("ヒストグラム表示", fluidPage(
   titlePanel("～",
    fluidRow(
     sidebarLayout(
      sidebarPanel(selectInput(～)),
      mainPanel(plotOutput("distPlot1"))
     )
    ),
    fluidRow(
     ～
    )
  )),
  tabPanel("散布図表示", fluidPage(
     ～
  ))
))
```

選択した調査品目に対するヒストグラム，散布図，相関係数の値を表示する機能を記述する server.R の一部を以下に示す．

```
shinyServer(function(input, output) {
 df <- read_excel("retail_price_survey_2017.xlsx")
 output$distPlot1 <- renderPlot({
  item1 = input$select1
  hist(as.matrix(df[c(item1)]),main="",xlab="")
 })
 output$distPlot3 <- renderPlot({
  item3 = input$select3
  item4 = input$select4
  plot(as.matrix(df[c(item3)]),
   as.matrix(df[c(item4)]),main="",xlab="",ylab="")
  })
 output$distText1 <- renderText({
  item3 = input$select3
  item4 = input$select4
  paste("相関係数 ",cor(as.matrix(df[c(item3)]),
   as.matrix(df[c(item4)])),sep="")
 })
})
```

3.2 ヒストグラム

表計算ソフトウェアの実習においてヒストグラムを作成させる場合，操作手順を教えるばかりではなく，ヒストグラムの特徴の読み取り方を指導する必要がある．例えば，ヒストグラムの単峰性，多峰性，対称性（左右対称，右裾が長い，左裾が長い），外れ値の存在を観察させる必要がある．

ヒストグラムの特徴を見つける指導をするには，さまざまなパターンのヒストグラムを実例として見せる必要がある．教科書的な指導法は，特定の特徴を持つように作られた架空のデータによるヒストグラムをいくつか見せることである（例えば(8)の 18～21 ページ）．

本稿では，架空のデータではなく，多数のデータが含まれる実際の統計表から，さまざまな分布の特徴を持つデータを見つけやすくするインターフェースを試作する．そのため，ひとつの画面に複数のヒストグラムを描画し，教材作成者がデータを切り替えると，ヒストグラムも再描画されるようにする．このインターフェースは，複数のヒストグラムを比較しながら，教材作成者がさまざまなデータのヒストグラムを次々と閲覧して，教材にふさわしいものを見つけ出すことを想定したものである．

図 2 ヒストグラムを描画するインターフェース

図 1 の調査品目の年平均価格のデータファイルを読み込んで，2 つの調査品目を選択するとヒストグラムが描画されるようなインターフェース画面を図 2 のように試作した．画面左側の品目欄で調査品目を切り替えると，連動してヒストグラムが再描画される．

3.3 散布図

表計算ソフトウェアの実習において散布図を作成させる場合，操作手順を教えるばかりではなく，散布図の特徴の読み取り方を指導する必要がある．例えば，正の相関関係，負の相関関係，無相関，強い相関関係，弱い相関関係を観察させるとともに，相関係数の値との関連を理解させる必要がある．

散布図の特徴を見つける指導をするには，さまざまなパターンの散布図を実例として見せる必要がある．教科書的な指導法は，特定の相関関係を持つように作られた架空のデータによる散布図をいくつか見せることである（例えば(9)の 16 ページ）。

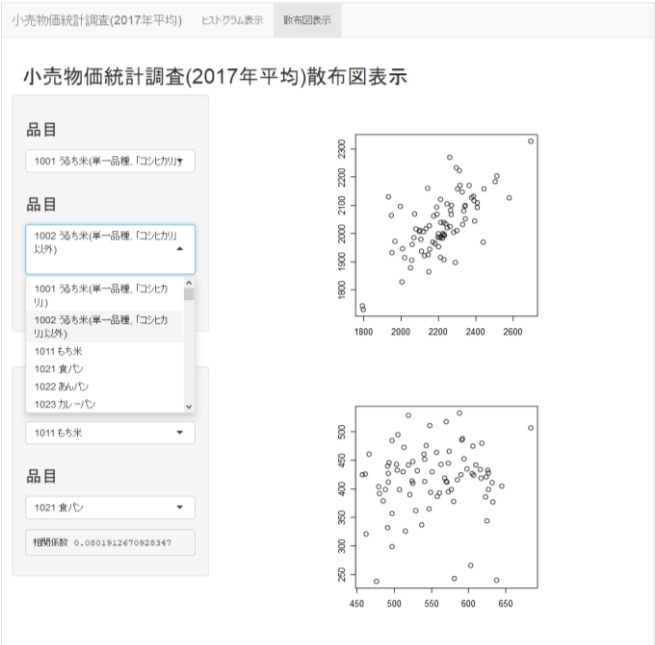

図 3 散布図を描画するインターフェース

本稿では，架空のデータではなく，多数のデータが含まれる実際の統計表から，さまざまな特徴を持つような散布図となるデータを見つけやすくするインターフェースを試作する．そのため，ひとつの画面に複数の散布図を描画し，教材作成者がデータを切り替えると，散布図も再描画されるようにする．合わせて相関係数の値も再計算される．このインターフェースは，複数の散布図を比較しながら，教材作成者がさまざまなデータの散布図を次々と閲覧して，教材にふさわしいものを見つけ出すことを想定したものである．

図 1 の調査品目の年平均価格のデータファイルを読み込んで，調査品目を選択すると散布図が描画されるとともに相関係数の値が表示されるようなインターフェース画面を図 3 のように試作した．画面左側の品目欄で調査品目を切り替えると，連動して散布図と相関係数の値が再表示される．

3.4 試作プログラムの検討

R 言語を用いて簡潔にプログラムが記述されていることは，教材作成者が求めるデータが得られるようにプログラムの改変が容易にできるという利点がある．例えば，図 2 において，上段のヒストグラムは右裾が長いものに限定されるようにプログラムを改変するには，調査品目ごとにデータの歪度を算出し，それがある一定値より大きいものだけに絞ればよい．そのために，歪度を算出するためにパッケージ e1071 を使用して以下のようなプログラムをファイル ui.R に追加すると，変数 sk に調査品目ごとの歪度の計算結果が得られる．ただし，データ不足のため歪度が算出できない場合があるので，計算結果が非数値である場合を除外して sk1 としている．

```
library(e1071)
sk <- apply(df[c(-1)],2,
 function(x){skewness(x,na.rm = TRUE)})
sk1 <- sk[!is.nan(sk)]
```

そして，図 2 の上段の品目欄について，歪度の値が 1 より大きい調査品目だけが選択肢に表れるようにするには，次のようにプログラムを書き換えると実現できる．

```
selectInput("select1", label = h3("品目"),
 choices = names(sk1)[sk1 > 1])
```

このようにプログラムを改変した結果の画面を図 4 に示す．

さらに，教材作成者自身によりプログラムを部分的に修正することが容易であることも，R 言語による簡潔なプログラムの有用性である．例えば，図 4 の右裾が長いヒストグラムを，左裾が長いものに変更するには，sk1 > 1 の部分を sk1 < -1 に変更すればよいことは容易に想像がつくはずである．また，数値の大きさ

を変えることで､裾を引く程度の調整ができることも，教材作成者自ら気付くだろう．

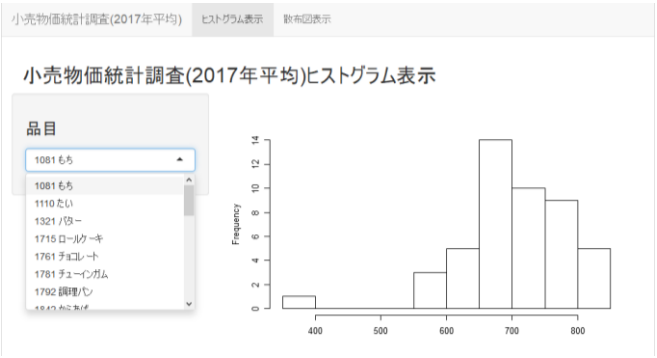

図 4 右裾が長いヒストグラムに限定したもの

4. おわりに

本稿では，総務省統計局による小売物価統計調査年報 平成 29 年のうち「調査品目の月別価格及び年平均価格【県庁所在市及び人口 15 万以上の市】」を政府統計の具体例として，実際のデータから，ヒストグラム作成と散布図作成のそれぞれにふさわしいデータを見つけるためのインターフェースを開発し，その開発過程を通じて検討した．この開発には R 言語を用い，インタラクティブな Web アプリケーションを容易に構築することができるパッケージ shiny を使用することにより，簡潔なプログラムで実現することができた．プログラムの簡潔さは，教材作成者が求めるデータが得られるようにプログラムを改変することが容易にできるという利点をもたらすとともに，プログラミングの経験が少ない教材作成者が，自らプログラムの修正を試みようとすることに対する障壁の高さを下げている．

本稿で取り上げている表計算ソフトウェアの使い方の指導は，大変広範囲に及ぶ統計教育のごく一部に過ぎないが，表計算ソフトウェア実習用の教材に，実際の公的な統計データを用いることは，個別の統計に対する理解の一助になっている．政府統計は，調査方法やデータの読み方のコツなどがそれぞれに異なるが，新入生には多数の政府統計に接触できるという利点もある．

今回試作したインターフェースは，教材作成者がさまざまなデータに対するグラフを次々と閲覧しながら，教材にふさわしいものを見つけるということを想定して作られたインターフェースである．今後の検討課題として，例えば，特徴が類似しているデータ，まったく特徴が異なるデータ，対照的な特徴をもつ 2 群に分けられるデータのように，教材作成者の要求に沿うデータを探索する手法の開発が挙げられる．

参 考 文 献

(1) 中央教育審議会：“第 3 期教育振興基本計画の策定に向けた基本的な考え方”, http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2017/02/06/1381849_01_1.pdf (2017)

(2) 私立大学情報教育協会：“「大学教育への提言」－未知の時代を切り拓く教育と ICT 活用―”, http://www.juce.jp/LINK/teigen.html (2012)

(3) GAISE College Report ASA Revision Committee：“Guidelines for Assessment and Instruction in Statistics Education College Report 2016”, http://www.amstat.org/education/gaise (2016)

(4) 統計質保証推進協会：“統計検定 2 級出題範囲表”, http://www.toukei-kentei.jp/wp-content/uploads/grade2_hani_181214.pdf (2018)

(5) 佐藤朋彦：“数字を追うな 統計を読め”，日本経済新聞出版社 (2013)

(6) 梅田雅信, 宇都宮浄人：“経済統計の活用と論点 第 3 版”, 東洋経済新報社 (2009)

(7) 総務省統計局：“小売物価統計調査（動向編）について（2018 年 4 月現在）”, http://www.stat.go.jp/data/kouri/doukou/1.html (2018)

(8) 総務省統計局：“データサイエンス・スクール あなたの統計力・初級テキスト”，http://www.stat.go.jp/dss/getting/pdf/index.html (2014)

(9) 総務省統計局：“データサイエンス・スクール あなたの統計力・中級テキスト”，http://www.stat.go.jp/dss/getting/pdf/mid.html (2014)