集計前データの統計教育教材化の可能性

吉根勝美^{*1} *1 南山大学経済学部

Possibility of Using Data before Aggregation as Teaching Material for Statistical Education

Katsumi Yoshine*1
*1 Faculty of Economics, Nanzan University

大学生に対する統計教育では、実際のデータを用いて学習者の興味・関心を喚起したい.特に経済学部でしばしば用いられる経済統計の多くは、調査あるいは業務で得られた個々のデータを集計した統計であるが、むしろ、集計前の個々のデータの方が興味・関心を得やすいだろう.集計前の個々のデータがウェブで公開されている場合があるので、これを自動収集して教材化する事例を示し、統計教育教材としての可能性を考えたい.

キーワード: 大学初年次教育, 統計教育, ウェブスクレイピング

1. はじめに

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 "Use real data"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(1). データの "reality"を求めて、例えば、経済学部における統計教育では、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多い。南山大学経済学部の 1 年次必修科目「データ処理入門」では、愛知県統計年鑑「百貨店・スーパーの事業所数、従業者数及び販売額等」、賃金構造基本統計調査「企業規模別新規学卒者の初任給の推移」等を用いて、表計算ソフトウェア演習のための例題データを作成している。

Δ	Α	В	С	D	E	F	G
1			百貨店販売	額(愛知県	人平成29	年)	
		紳士服・洋	婦人・子供	その他の衣	身の回り品	飲食料品	その他
2		믦	服・洋品	料品			
3	1月	3325	11560	880	6122	8906	9582
4	2月	1885	7701	526	4356	9388	9081
5	3月	2601	11321	826	5616	9208	12059
6	4月	2587	9220	638	4867	7369	9167
7	5月	2562	9000	1570	5064	7420	9327
8	6月	2322	8158	615	5083	9191	10101
9	7月	2641	9667	840	6011	11632	10939
10	8月	1705	6961	548	4535	7432	9520
11	9月	1991	8600	1250	4728	7643	10234
12	10月	2918	9388	668	5194	7701	9512
13	11月	3186	9587	961	5116	10138	10296
14	12月	3691	10729	664	7317	16904	13763

図 1 表計算演習用データの例(百貨店販売額)

経済学部の統計教育における"real data"である公的な統計データは、調査あるいは業務で得られた個々のデータを集計した統計であることが多く、学習者にとっては実感しにくいデータである。図1は、実際の表計算演習で使用している例題で、愛知県統計年鑑の最新刊から百貨店の販売額を抽出したものである。学生には縦横集計をさせて、愛知県の百貨店販売額の年間合計の数値を求めさせたのち、その数値の単位を推測させているが、正答率は必ずしも高くない。学生にとっては、1店舗あたりの年間販売額すら想像がつかないのに、愛知県にある百貨店13店舗の合計年間販売額となると、ますます想像できないだろう。

調査統計や業務統計の場合,学生は,集計されたデータよりも,集計される前の個々のデータの方に,より"reality"を感じるだろう.百貨店販売額は,経済産業省が商業動態統計で毎月調査し,経済産業局別,東京特別区・政令指定都市別,都道府県別に集計した結果が公表されている.こうした集計された数字に比べれば,百貨店が回答した調査票に書かれている月ごとの商品販売額の方に,より"reality"が感じられるだろう.しかし,調査票の記載事項が個別に公表されることは,もちろんあり得ない.

ウェブ公開されている公的な統計データの中には, 情報提供を目的とする検索機能を用いて, 集計前の 個々のデータに1件1件アクセスすることが可能な場 合がある. ただし、実際にデータをまとめて収集する には大変手間がかかるので、 ウェブスクレイピングの 技術が有用である. そこで本稿では, 集計前の個々の データをウェブスクレイピングの技術で自動収集して 教材化する2つの事例を示し、統計教育教材としての 可能性を考えたい. なお, 各事例の開発環境は, Windows 10 Home バージョン 1903 (64 ビット) で 動作する Spyder 4.0.0 であり、使用言語は Python 3.7.5 である. 使用した Python ライブラリは, Requests 2.22.0, BeautifulSoup 4.8.1, Selenium 3.141.0 である. また, ブラウザには Google Chrome 79.0.3945.130 (64 ビット) を使用し、対応する Selenium WebDriver 1th ChromeDriver 79.0.3945.36 である.

2. 事例 1 (地価公示)

2.1 データの説明

国土交通省は、地価公示についてのウェブページ(2)で「地価公示にみる地価の状況」を発表している。名古屋圏については、沿線別駅周辺住宅地・商業地の公示価格例、市の住宅地の平均価格等、地域別・市の対前年平均変動率を公表している。図2は「名古屋圏の市の住宅地の平均価格等」であり、例えば、名古屋市千種区について、標準地数26、平均価格243,900(円/㎡)、上位の価格369,000(円/㎡)、下位の価格158,000(円/㎡)と記載されていて、基本統計量の計算のための題材にしたい。

⊿	A	В	C	D	E	F	G	H	I	J	K	L
				筆	12表 名	古屋圏の	市の住室	地の平均価	格等			
2											(単位:円/㎡
	都 道府県名	市区名	標準地數	平均価格	上位の価格	下位の価格	都 道府 県 名	市区名	標準 地数	平均価格	上位の価格	下位の価格
		名古屋市	347	188,100	1,280,000	72,800	愛知県	安城市	22	126,600	168,000	58,200
5		千種区	26	248,900	868,000	158,000		西尾市	40	62,200	112,000	30,300
3		東区	13	\$92,200	860,000	230,000		犬山市	18	63,500	105,000	35,200
7		北区	19	184,400	222,000	108,000		常滑市	15	45,000	65,300	27,500
		西区	17	178,600	278,000	119,000		江南市	8	76,700	98,000	47,500
П		中村区	16	178,800	220,000	127,000		小牧市	25	86,100	112,000	41,900
)		中区	5	724,400	1,280,000	\$00,000		稲沢市	20	78,000	137,000	35,500
1		昭和区	12	285,100	880,000	232,000		東海市	29	89,900	128,000	60,800
2		瑞穂区	17	241,400	200,000	162,000		大府市	16	121,000	162,000	52,400
3		熱田区	13	194,100	\$00,000	131,000		知多市	16	70,700	98,400	46,000
4		中川区	40	125,700	188,000	82,500		知道市	14	122,100	158,000	98,300
5		港区	20	106,500	138,000	85,800		尾張旭市	20	100,100	134,000	78,500
3		南区	21	131,200	188,000	94,000		高浜市	12	82,800	112,000	66,800
7		守山区	36	118,500	164,000	72,800		岩倉市	10	105,300	142,000	84,200
3		緑区	40	146,800	177,000	99,000		豊明市	11	108,000	130,000	68,500
9		名東区	24	200,100	826,000	117,000		日進市	14	132,500	170,000	58,500
Ó		天白区	28	169,300	239,000	130,000		愛西市	11	54,800	88,400	29,000
1		国時市	72	101,500	158,000	38,700		清須市	27	98,900	107,000	74,500
2		一宮市	46	76,900	260,000	38,800		北名古屋市	18	101,400	140,000	75,200
3		瀬戸市	38	57,000	87,600	20,000		弥富市	11	64,200	82,500	31,300
4		半田市	28	76,300	110,000	44,100		みよし市	9	114,200	147,000	85,800
5		春日弁市	49	102,900	180,000	45,500		あま市	18	78,900	91,500	64,500
6 7		津島市	11	58,800	79,500	46,500		長久手市	13	149,000	185,000	71,500
7		碧南市	19	71,200	104,000	40,500	三重県	四日市市	62	47,200	72,000	20,000
3		刈谷市	29	137,200	189,000	68,500		桑名市	89	50,900	85,900	12,800
9		弗田市	58	105,400	166,000	20.000		www.	7	25,600	40,000	14,100

図 2 地価公示で公表されているデータの例

国土交通省は、国土に関する基礎的な情報を GIS データとして整備した「国土数値情報 ダウンロードサービス」を提供しており、地価公示価格も含まれているが、大変複雑なデータ構造のため使いにくい.

一方で、同省は「標準地・基準地検索システム~国土交通省地価公示・都道府県地価調査~」(3)を提供しており、地価公示を検索することが可能である.ここで、名古屋市千種区の住宅地の地価公示価格をすべて抽出すれば、前述した基本統計量の計算のための題材における例題データにすることができる.そこで以下では、ウェブスクレイピングにより、実際の地価公示価格を抽出する方法を説明する.

2.2 スクレイピングの方法

「標準地・基準地検索システム」で地価を検索する手順は以下の通りである。まず、「検索地域指定(地名入力)」の画面で、愛知県、名古屋市、千種区と入力し「決定」をクリックする。次に、地価の検索条件の設定画面で、対象「地価公示のみ」、用途区分「住宅地」をチェックして「検索」をクリックすると(図 3)、検索結果が 1 ページに 20 件ずつ表示される。次ページが存在する限り「次へ」をクリックすることを繰り返して地価公示価格を閲覧することができる(図 4)。



図 3 地価公示価格の検索条件指定画面





図 4 地価公示価格の検索結果表示画面

以上の検索手順をそのままなぞるように実行する Python プログラムを作成した. ソースプログラムはおよそ 50 行である. 検索条件の指定と検索結果のページ送りには selenium を、検索結果表示画面からデータを抽出するには BeautifulSoup を利用した. 「所在及び地番」欄に記載されている住所と「価格(円/m²)」に記載されている地価を抽出した結果を CSV ファイルに出力して、表計算ソフトウェアで直ちに利用できるようにした(図 5). これは、実際の地価データによる平均、最小値、最大値を計算するための例題データであるとともに、文字列の一括置換機能((円/m²)の削除) やソート機能(価格の昇順、降順)を学ばせる例題にもなっている.

1	A	В
1	address	price
2	愛知県名古屋市千種区新西1丁目613番	158,000(円/m²)
3	愛知県名古屋市千種区清明山1丁目907番	190,000(円/m²)
4	愛知県名古屋市千種区南ケ丘1丁目43番3	188,000(円/m²)
5	愛知県名古屋市千種区池上町2丁目6番15	232,000(円/m²)
6	愛知県名古屋市千種区桐林町2丁目37番	348,000(円/m²)
7	愛知県名古屋市千種区唐山町3丁目16番	312,000(円/m²)
8	愛知県名古屋市千種区向陽町2丁目11番8外	312,000(円/m²)
9	愛知県名古屋市千種区城木町2丁目25番	264,000(円/m²)
10	愛知県名古屋市千種区茶屋が坂1丁目1104番	241,000(円/m²)
11	愛知県名古屋市千種区光が丘2丁目1215番	180,000(円/m²)
12	愛知県名古屋市千種区川崎町1丁目16番1	250,000(円/m²)
13	愛知県名古屋市千種区東山元町5丁目10番4	171,000(円/m²)
14	愛知県名古屋市千種区新池町4丁目7番	286,000(円/m²)
15	愛知県名古屋市千種区徳川山町1丁目25番	223,000(円/m²)
16	愛知県名古屋市千種区清住町2丁目38番1	369,000(円/m²)
17	愛知県名古屋市千種区千代田橋1丁目103番	211,000(円/m²)
18	愛知県名古屋市千種区豊年町1602番	235,000(円/m²)
19	愛知県名古屋市千種区千種3丁目2010番外	272,000(円/m²)
20	愛知県名古屋市千種区東山元町1丁目58番1	200,000(円/m²)
21	愛知県名古屋市千種区今池3丁目907番1	284,000(円/m²)
22	愛知県名古屋市千種区京命1丁目810番	171,000(円/m²)
23	愛知県名古屋市千種区鹿子町5丁目33番	271,000(円/m²)
24	愛知県名古屋市千種区汁谷町68番3外	189,000(円/m²)
25	愛知県名古屋市千種区桜が丘280番	246,000(円/m²)
26	愛知県名古屋市千種区神田町2904番	255,000(円/m²)
27	愛知県名古屋市千種区丸山町2丁目63番3外	283,000(円/m²)

図 5 抽出された地価公示価格の一覧

3. 事例 2 (求人情報)

3.1 データの説明

厚生労働省は、公共職業安定所(ハローワーク)に おける求人、求職、就職の状況をとりまとめ、求人倍 率などの指標を作成し、一般職業紹介状況⁽⁴⁾として毎 月公表している。ウェブページではさまざまな数表が 公表されているが、図 6 にその一例を示す。

	有効求人		有効求職		倍率	新規求人		新規求職	P	倍率	8tm		对有効	対新規	对有効	对新規
	前年	前年比		前年比	1000		前年比	1133.340. 3	前年比	200	200	前年比	就職率	就職率	充足率	充足率
28年度平均	2, 569, 72	6.4	1, 844, 89	-5.6	1, 39	922, 565	5.3	442, 531	-5.9	2.08	149, 686	-5.0	-	33. 8	-	16.2
29年度平均	2, 726, 32	6.1	1, 770, 695	-4.0	1.54	967, 308	4.8	421, 739	-4.7	2.25	142, 882	-4.5		33. 9	-	14.1
3 0年度平均	2, 782, 42	2.1	1, 716, 55	7 -3,1	1, 62	975, 836	0.9	403, 089	-4.4	2.42	133, 974	-6.2	-	33. 2	-	13.
30年3月	2, 893, 47	3 3.1	1, 781, 453	-5.2	1. 62	1, 010, 782	3.1	459, 065	-6.5	2.20	181, 109	-8.0	10.2	39. 5	6.3	17.1
4.8	2, 785, 85	4.1	1, 867, 21	-3.7	1.49	966, 323	4.6	544, 826	-1.9	1.77	157, 731	-4.0	8.4	29. 0	5.7	16.
5月	2, 753, 34	5.3	1, 871, 391	3 -2.6	1.47	962, 465	5.5	446, 742	-1.3	2.15	153, 632	-2.0	8.2	34. 4	5.6	16.
6月	2, 729, 36	3.7	1, 799, 383	2 -3.7	1.52	966, 055	0.2	381, 829	-9.3	2.53	141, 341	-9.4	7.9	37.0	5.2	14.1
7月	2, 722, 69	7 4.1	1, 722, 960	-3.5	1, 58	956, 726	3.7	371, 586	-2.4	2.57	129, 656	-4.5	7.5	34, 9	4.8	13.
8月	2,749,25	7 3.3	1, 696, 983	-4.0	1. 62	975, 576	3.4	383, 599	-4.9	2.54	121, 418	-6.3	7.2	31.7	4.4	12.
9月	2, 730, 63	8 0.4	1, 662, 893	-5.8	1, 64	931, 362	-6.6	352, 638	-14.9	2.64	123, 144	-14.8	7.4	34. 9	4.5	13.
10月	2, 838, 70	1.4	1, 712, 073	3 -3.2	1.66	1, 070, 360	4.6	422, 089	3.0	2.54	140, 718	-3.5	8.2	33. 3	5.0	13.
11,8	2, 824, 815	5 1.4	1, 666, 623	-2.4	1. 69	966, 635	2.6	357, 488	-2.4	2.70	127, 971	-3.9	7.7	35. 8	4.5	13.
12月	2, 745, 55	3 0.8	1, 582, 695	5 -1.8	1,73	869, 219	-5.7	302, 016	-7.0	2.88	111, 196	-7.3	7.0	36, 8	4.1	12.
31年1月	2, 768, 17	0.7	1, 616, 14	-1.5	1.71	1, 058, 693	2.8	435, 168	-1.4	2.43	105, 178	-5.1	6.5	24. 2	3.8	9.
2月	2, 857, 04	0.6	1, 664, 145	-1.5	1. 72	1, 036, 945	2.1	415, 866	-3.6	2.49	130, 040	-5.0	7.8	31.3	4.6	12.
3月	2, 883, 60	-0.3	1, 736, 185	-2.5	1,66	949, 673	-6.0	423, 219	-7.8	2.24	165, 668	-8.5	9.5	39. 1	5.7	17.
4月	2, 768, 21	4 -0.6	1, 821, 621	-2.4	1.52	963, 317	-0.3	522, 046	-4.2	1.85	145, 420	-7.8	8.0	27. 9	5.3	15.
元年5月	2, 696, 25	5 -2.1	1,817,353	3 -2.9	1.48	938, 680	-2.5	417, 227	-6.6	2.25	135, 994	-11.5	7.5	32. 6	5.0	14.
6月	2, 683, 37	8 -1.7	1, 772, 82	-1.5	1.51	925, 624	-4.2	372, 963	-2.3	2.48	130, 152	-7.9	7.3	34. 9	4.9	14.
7月	2, 708, 98	5 -0.5	1,743,90	1.2	1.55	980, 223	2.5	395, 727	6.5	2.48	129, 524	-0.1	7.4	32. 7	4.8	13.
8.8	2, 679, 82	5 -2.5	1, 698, 204	0.1	1.58	917, 772	-5.9	349, 542	-8.9	2.63	108, 712	-10.5	6.4	31. 1	4.1	11.
9月	2, 695, 30	5 -1.3	1, 698, 521	3 2.1	1, 59	917, 174	-1.5	378, 174	7.2	2.43	122, 367	-0.6	7.2	32. 4	4.5	13.
10月	2, 730, 02	7 -3.8	1, 705, 665	5 -0.4	1.60	1, 027, 758	-4.0	389, 973	-7.6	2.64	129, 485	-8.0	7.6	33. 2	4.7	12.
11月	2, 702, 08	4 -4.3	1, 658, 50	-0.5	1.63	901, 638	-6.7	338, 040	-5.4	2.67	115, 142	-10.0	6.9	34. 1	4.3	12.
12月	2, 666, 121	3 -2.9	1, 586, 93	2 0.3	1, 68	887, 713	2.1	314.030	4.0	2.83	104, 964	-5.6	6.6	33. 4	3.9	11.

図 6 一般職業紹介状況で公表されているデータの例

このようなデータは、学習者にマクロ的な視点で観察させるのに必要ではあるが、全国的に集計されたデータに"reality"を感じさせるのは難しい、そこで、本章では、厚生労働省職業安定局が提供する「ハローワークインターネットサービス」(5)を用いて、地域の実際の求人情報を抽出する方法を説明する.

3.2 スクレイピングの方法

「ハローワークインターネットサービス」のうち「求人情報検索」で、愛知県名古屋市瑞穂区での介護、福祉の求人情報を検索する手順は以下の通りである。求人区分では「一般求人」を指定する(フルタイム、パートのチェックは任意)。「就業場所」で愛知県を選び「選択」をクリックし、別画面で名古屋市瑞穂区を選択して「決定」をクリックする。希望する職種で「職種を選択」をクリックする。の関すると、条件に当てはまる求人情報が1ページに30件ずつ表示される。次ページが存在する限り「次へ>」をクリックすることを繰り返して求人を閲覧することができる(図8)。求人1件ごとの表示内容の項目数が異なる場合があるので、スクレイピングするときには注意を要する。



図 7 求人情報の検索条件指定画面



図 8 求人情報の検索結果1件分の表示例

以上の検索手順をなぞるように実行して、求人区分、 雇用形態、賃金(求人種別、賃金形態、賃金額)のデータを抽出して、CSVファイルを出力する Python プログラムを作成した.ソースプログラムはおよそ 80 行である. このプログラムを 2020 年 2 月 6 日に実行したところ、検索結果は 125 件あった. 図 9 には、そのうち最初の 10 件分のデータを示す. この CSVファイルは、"real data"によるデータ分析の例題であると同時に表計算ソフトの操作演習にも活用できる.例えば、文字列操作で賃金額の下限、上限を取り出したり、フィルター機能を用いて、ある条件に該当するデータを抽出したりするような演習に利用可能である.

Δ	A	В	С	D	E
1	求人区分	雇用形態	求人種別	賃金形態	賃金額
2	フルタイム	正社員	一般	月給	231,250円~289,063円
3	パート	パート労働者	一般	時給	1,200円~2,000円
4	パート	パート労働者	一般	時給	1,400円~1,700円
5	パート	パート労働者	一般	時給	1,250円~1,800円
6	パート	パート労働者	一般	時給	1,000円~1,000円
7	パート	有期雇用派遣パート	一般	時給	1,200円~1,400円
8	フルタイム	正社員	一般	月給	237,100円~250,600円
9	フルタイム	正社員	一般	月給	237,100円~250,600円
10	フルタイム	正社員	一般	月給	246,000円~311,000円
11	フルタイム	正社員	一般	月給	256,000円~385,000円

図 9 抽出された求人情報の一覧(一部)

4. おわりに

本稿では、学習者が"reality"をより感じられるデータを教材化することを目的として、ウェブスクレイピングの技術を用いて、地価公示データや求人情報データを収集して、データ分析の授業で利用可能な教材データの作成事例を紹介した。いずれも Python プログラムにより、データの自動収集が可能となった。

ウェブスクレイピングによりデータを自動的に収集 するときには、Web サービスの利用規約でスクレイピ ングが禁止されていないかどうかを確認する必要があ る. また、Web サービスに過度な負担をかけないよう にする必要がある. 今回作成したプログラムでは、待 機時間を随時設けている。

政府が全大学生の受講を目指している初級水準の人工知能教育の一環として、表計算ソフトによるデータ分析用のために、実際の商品の購買データなどを実社会から集め、各大学向けに公開する計画があるという新聞報道もある(6)。このように、大学生の統計教育における"Use real data"の重要性は変わることないので、同様の事例を積み上げてアーカイブ化し、教材データの共同利用の促進を目指したい。

参考文献

- (1) GAISE College Report ASA Revision Committee:

 "Guidelines for Assessment and Instruction in
 Statistics Education College Report 2016",
 http://www.amstat.org/education/gaise (2016)
- (3) 国土交通省: "標準地・基準地検索システム~国土交通省地価公示・都道府県地価調査~", https://www.land.mlit.go.jp/landPrice/AriaServlet?M OD=0&TYP=0
- (4) 厚生労働省: "一般職業紹介状況(職業安定業務統計)", https://www.mhlw.go.jp/toukei/list/114-1.html
- (5) 厚生労働省職業安定局: "ハローワークインターネット サービス", https://www.hellowork.mhlw.go.jp/
- (6) 日本経済新聞: "AI「普段使い」へ教育―政府・大学など モデル案、学生に基礎力 担い手拡大", 2020 年 2 月 4 日 付夕刊 (2020)