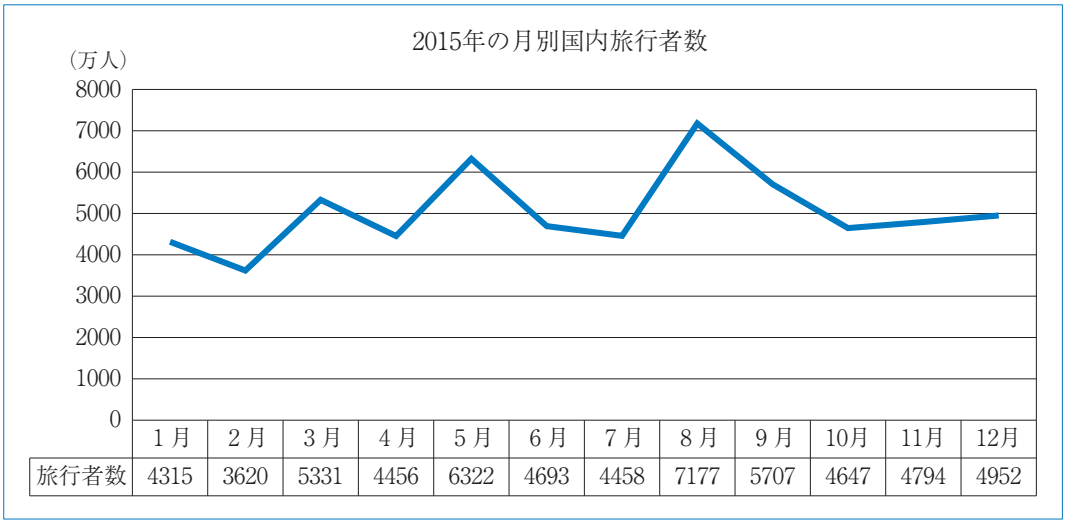
第2部

統計的探究の実践 Ⅰ
～データから有用な情報を引き出す～

1　夏の避暑地の気候の特徴～夏の避暑地が快適な理由は？［データの整理］



	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
旅行者数	4315	3620	5331	4456	6322	4693	4458	7177	5707	4647	4794	4952

観光庁「2015 年旅行・観光消費動向調査」

日本への外国人旅行者はこの２、３年に急増しているが、日本人の国内旅行者の動向を月別に見ると、上の図から、月毎に変動していることが分かる。

Q１：上のグラフは2015年の国内旅行者数の推移を表している。８月が突出して多いのは何故だろう？

STEP 1：Problem 問題　課題の設定

◇ 東京都に比べて、夏の避暑地は本当に過ごしやすいか？

日本では、夏に避暑地を訪れることを好む人が多いが、避暑地にはどのような特徴があるのだろうか？都内の高校の休み時間に、航平、理恵、公介、馨、健三の５人は、それぞれの夏休みについて、次のように話していた。

航平：軽井沢は東京に比べて、過ごしやすかったよ
理恵：東京も今年は涼しい日もあったけど、すごく暑い日が多かったわ
公介：熊谷の祖母の家にいたけど、東京以上に暑かった
馨　：沖縄は暑かったけど、慣れてしまえば逆に過ごしやすかったわ
健三：札幌は過ごしやすかったけど、大阪に行ったときは東京と同じように暑かったな

それぞれの場所で、本当に暑さに違いはあったのだろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 夏の暑さを調べる指標を探る

気象庁の HP には、夏の暑さを調べる指標として、１日の（A）平均気温、（B）最高気温、（C）最低気温の３つが掲載されている。

国土交通省 気象庁 Japan Meteorological Agency

本文へ | ENGLISH | ご意見・ご感想 | サイトマップ

キーワードを入力し検索ボタンを押下ください。 検索

ホーム | 防災情報 | 各種データ・資料 | 知識・解説 | 気象庁について | 案内・申請

ホーム > 各種データ・資料 > 過去の気象データ検索 > 日ごとの値

日ごとの値

一覧表 | グラフ | 見出しの固定 | メニューに戻る

主な要素 | 詳細（気圧・降水量） | 詳細（気温・蒸気圧・湿度） | 詳細（風） | 詳細（日照・雪・その他）

前年 | 前月 | 前日 | 翌日 | 翌月 | 翌年 | 月ごとの値 | 旬ごとの値 | 半旬ごとの値 | 日ごとの値

東京 2015年8月（日ごとの値） 主な要素

日	気圧(hPa)		降水量(mm)			気温(℃)			湿度(%)		風向・風速(m/s)					日照時間(h)	雪(cm)		天気概況	
	現地	海面									平均	最大風速		最大瞬間風速			降雪	最深積雪	昼	夜
	平均	平均	合計	最大 1時間	最大 10分間	平均	最高	最低	平均	最小	風速	風速	風向	風速	風向		合計	値	(06:00-18:00)	(18:00-翌日06:00)
1	1007.6	1010.3	--	--	--	30.5	35.3	26.6	76	50	2.6	5.1	南	8.5	南	8.9	--	--	晴	晴一時曇
2	1008.2	1011.9	0.0	0.0	0.0	30.2	35.1	26.3	71	54	2.6	5.6	南南東	9.3	南東	9.6	--	--	晴	曇時々晴
3	1008.7	1011.4	--	--	--	29.8	35.0	26.1	69	52	3.1	6.7	南	10.8	南	12.0	--	--	晴	晴
4	1007.4	1010.1	--	--	--	30.0	35.1	26.5	69	53	3.4	6.6	南南東	10.5	南南東	11.4	--	--	晴	晴
5	1007.4	1010.1	--	--	--	30.2	35.2	25.7	69	47	4.1	8.2	南南東	13.1	南南東	12.8	--	--	快晴	快晴

資料　気象庁 HP「東京　2015年８月（日ごとの値）主な要素」からの抜粋

Q２：気温について調べる際、（A）、（B）、（C）のどのデータを選ぶのが適当か？

航平君たちは、夏の暑さが関係しているのではと考え、１日の最高気温のデータを収集することに決めた。
それでは、どのようにデータを集めれば良いだろうか？

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 夏の暑さを調べるため、気象観測データを活用しよう

気象庁の HP から、2015年 8 月の東京（東京都）、軽井沢（長野県）、熊谷（埼玉県）、石垣島（沖縄県）、札幌（北海道）、大阪（大阪府）の 6 地点の 1 日の最高気温のデータをダウンロードし、表に低→高の順に整理した。

表 1　2015 年 8 月の 1 日の最高気温のデータ（31日間の昇順；℃）

順位	東京	軽井沢	熊谷	石垣島	札幌	大阪
①		15.9	20.7	28.7	21.7	26.6
②		16.9	21.3	28.8	22.1	27.4
③		17.1	22.6	29.2	22.5	28.1
④		17.8	23.4	30.5	23.0	29.2
⑤		20.0	24.3	30.6	23.3	30.1
⑥		20.3	24.6	30.6	23.5	30.9
⑦		20.4	26.0	30.8	23.6	31.5
⑧		21.6	27.0	30.9	24.2	31.6
⑨		21.8	27.5	30.9	24.9	31.7
⑩		23.3	28.1	31.1	25.0	32.1
⑪		23.7	28.8	31.2	25.4	32.2
⑫		23.8	29.4	31.3	25.4	32.2
⑬		24.3	31.0	31.4	26.0	32.3
⑭		24.5	32.6	31.4	26.3	32.3
⑮		26.6	33.2	31.4	26.4	32.6
⑯		27.0	33.8	31.5	26.5	33.1
⑰		27.1	33.9	31.6	26.7	33.1
⑱		27.3	34.0	31.7	26.9	33.1
⑲		27.4	34.1	31.7	27.3	33.5
⑳		27.4	34.2	32.0	27.5	34.3
㉑		27.8	34.4	32.1	27.7	34.7
㉒		28.3	34.4	32.2	27.8	35.1
㉓		28.5	34.6	32.7	27.8	36.2
㉔		29.8	35.7	32.8	27.8	36.3
㉕		30.0	36.6	32.8	28.2	36.3
㉖		30.1	37.5	33.0	28.2	36.4
㉗		30.2	37.5	33.1	28.7	36.4
㉘		30.2	38.0	33.3	29.0	36.5
㉙		30.3	38.2	33.3	29.5	36.7
㉚		30.4	38.3	33.4	31.9	37.5
㉛		31.1	38.6	33.6	34.5	38.0

表 2　東京の各日の最高気温

2015年 8 月	東京
1 日	35.3
2 日	35.1
3 日	35.0
4 日	35.1
5 日	35.2
6 日	35.9
7 日	37.7
8 日	32.6
9 日	33.4
10日	31.9
11日	35.5
12日	33.7
13日	30.5
14日	31.8
15日	33.1
16日	31.9
17日	28.0
18日	31.9
19日	31.4
20日	27.0
21日	29.4
22日	32.7
23日	31.4
24日	29.3
25日	22.9
26日	21.3
27日	27.3
28日	22.9
29日	21.0
30日	22.5
31日	24.1

資料：気象庁ホームページ「過去の気象データ・ダウンロード」
http://www.data.jma.go.jp/gmd/risk/obsdl/index.php

Q３：表 2 に示す、東京の最高気温のデータを昇順に並べ直し、表 1 を完成しなさい。

データを収集する際には、「データを小さい順に並べ替える」、「必要に応じて36000人→3.6万人とみなす」等の分析しやすい工夫をしておくと負担が減って楽だよ。データを順に並べたものを順序データといい、そこから**中央値**や**四分位数**が容易に求められるよ。

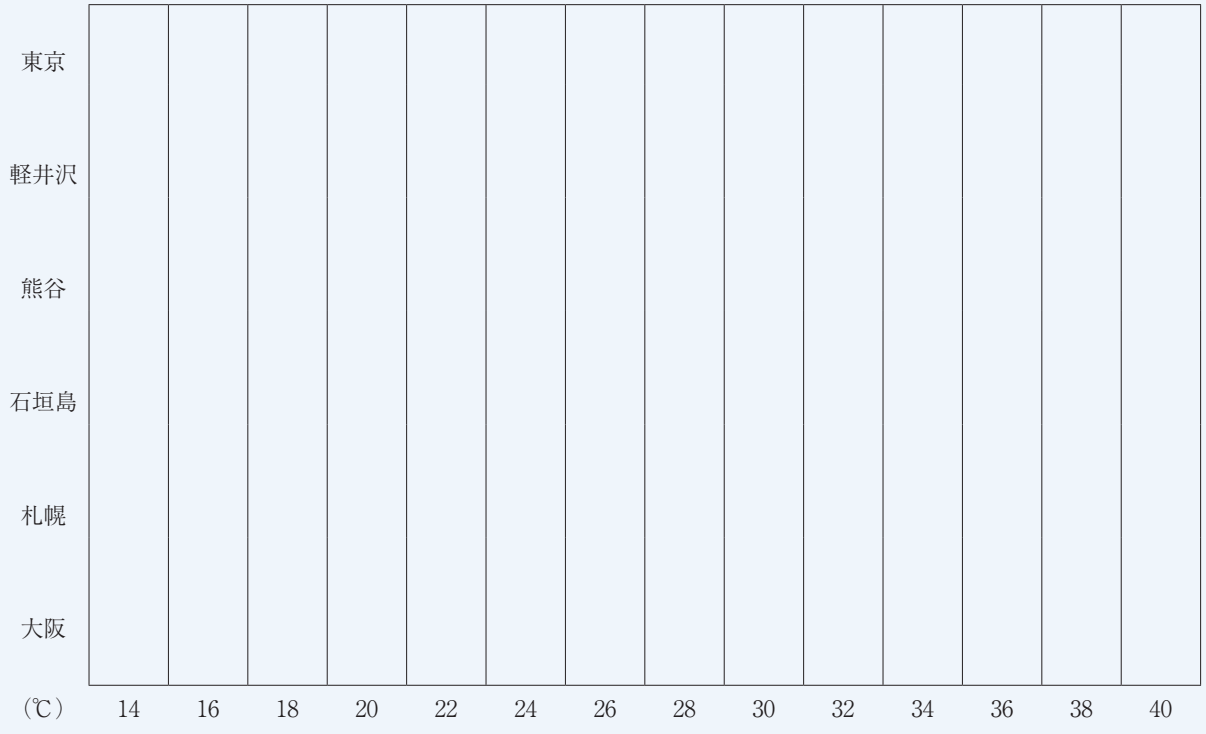
STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ 複数のデータの傾向を捉えるためには、並行箱ひげ図を活用

Q４：表１のデータをもとに、**五数要約（最大値、最小値、四分位数）**と**四分位範囲**をそれぞれ求め、**並行箱ひげ図**を作成しなさい。

表３　６地点の１日の最高気温の五数要約および四分位範囲

	東京	軽井沢	熊谷	石垣島	札幌	大阪
最大値						
第３四分位数						
第２四分位数（中央値）						
第１四分位数						
最小値						
四分位範囲						



２つのデータのバラツキは、**ヒストグラム**や度数折れ線を用いても比較できるが、複数（３つ以上）のデータのバラツキは、並行箱ひげ図を用いて比較するのが良いね。

Q５：作成した並行箱ひげ図と表１のデータに基づいて、各地点の特徴について分かったことを、次の観点からまとめなさい。

① 東京や大阪のような大都市は、避暑地と比べて暑い日が多いかどうか。

② 避暑地として人気の高い軽井沢は高原にあるが、北海道とどう違うのか。

③ 熊谷や沖縄は暑い地域として有名だが、それぞれで違いはあるか。

避暑地の特徴：

STEP 5：Conclusion 結論　結論を導く

◇ 夏の避暑地は、なぜ過ごしやすいのか？

Q６：軽井沢や札幌は夏の避暑地として人気が高いことで知られている。その理由をまとめなさい。

理由：

夏の蒸し暑さを表す指標：不快指数

平均気温と平均湿度から夏の蒸し暑さを数量的に表す指標として、不快指数が知られている。気温を t、湿度を H とすると

不快指数＝0.81t＋0.01H(0.99t－14.3)＋46.3　で求められる。不快指数が75になると人口の約１割が不快を感じ、85になると全員が不快になる。（三省堂編集所（1988)、『大辞林』p.2096、三省堂）

たとえば、沖縄県那覇市の2015年８月は平均気温が29.5℃、平均湿度が74% です。したがって、不快指数は81.2と算出され、次の表を参考に判断すると、不快指数は高いため、多くの人が不快（蒸し暑い）と感じる気候であると判断できます。

表１　不快指数と不快と感じる人数の目安

不快指数	70～75	75～80	80～
程度	１割の人が不快	半数の人が不快	全員が不快

資料：新村出（2008）『広辞苑第６版』p.2430、岩波書店

STEP 3 に示した６地点の不快指数を求めると、表２のようになる。

沖縄は、亜熱帯気候で知られているだけあって不快指数は高く、人気の高い避暑地である軽井沢や札幌は、不快指数は安定して低く、快適な気候であったと判断できる。

2017年において、高校の数学Ⅰでは、データのバラツキ（散らばり）を表す指標として、分散や標準偏差といった要約統計量を学習する。そして、これらを算出することで、データのバラツキ（散らばり）度合を判断できる。

表２　2015年８月１日から31日までの不快指数

８月	東京	軽井沢	熊谷	石垣島	札幌	大阪
１日	83.1	73.0	82.5	81.2	72.0	83.0
２日	81.8	69.6	79.7	81.5	71.6	82.8
３日	80.9	69.6	79.6	81.0	72.7	82.0
４日	81.2	70.3	81.8	80.7	76.4	82.2
５日	81.5	70.9	81.4	81.2	77.6	82.6
６日	82.4	71.1	80.4	82.3	75.1	82.0
７日	82.1	71.8	81.7	80.9	72.7	82.7
８日	77.9	70.9	78.6	80.5	71.0	81.7
９日	78.0	70.1	78.4	82.1	72.4	81.5
10日	79.4	70.2	79.4	81.8	75.4	81.2
11日	80.5	70.2	80.2	82.3	73.7	79.9
12日	80.2	70.5	79.7	82.7	72.7	79.2
13日	79.2	68.4	78.1	82.9	72.5	78.5
14日	78.4	69.2	76.0	82.9	72.2	78.7
15日	78.7	69.5	78.8	83.0	71.3	78.9
16日	79.0	67.9	77.6	83.4	71.6	77.5
17日	77.8	66.6	76.0	83.1	72.0	77.3
18日	79.5	70.6	79.4	82.0	64.6	77.8
19日	78.1	68.7	77.7	82.5	66.0	76.4
20日	77.2	67.7	75.7	82.8	67.5	77.5
21日	77.4	66.5	76.3	82.8	69.1	78.5
22日	80.2	71.3	79.9	81.6	71.3	78.8
23日	77.0	66.6	75.9	80.2	67.9	76.9
24日	72.5	64.2	72.7	81.2	66.7	77.0
25日	68.5	56.5	68.1	79.9	65.0	76.7
26日	67.0	58.5	66.3	80.4	65.2	76.9
27日	72.8	64.6	73.1	80.4	66.3	76.6
28日	69.9	62.5	70.7	81.2	65.5	77.0
29日	67.9	60.6	68.9	81.2	65.8	77.0
30日	69.9	60.8	69.1	81.2	66.1	75.8
31日	71.3	62.8	71.4	81.0	69.3	74.5

〔本節の解答〕

Q１：ゴールデンウィークや夏休み等、長期休みに国内旅行に出かける人が多いから。

Q２：（B）最高気温

理由：一番暑かったときの気温で、その日の暑さの印象が決まるから。

Q３：2015年８月の最高気温（℃）；昇順

順位	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳	㉑	㉒	㉓	㉔	㉕	㉖	㉗	㉙	㉚	㉛	㉛
東京	21.0	21.3	22.5	22.9	22.9	24.1	27.0	27.3	28.0	29.3	29.4	30.5	31.4	31.4	31.8	31.9	31.9	31.9	32.6	32.7	33.1	33.4	33.7	35.0	35.1	35.1	35.2	35.3	35.5	35.9	37.7

Q４：箱ひげ図は略

表３　６地点の１日の最高気温の五数要約および四分位範囲

	東京	軽井沢	熊谷	石垣島	札幌	大阪
最大値	37.7	31.1	38.6	33.6	34.5	38.0
第３四分位数	35.0	29.8	35.7	32.8	27.8	36.3
第２四分位数（中央値）	31.9	27.0	33.8	31.5	26.5	33.1
第１四分位数	27.3	21.6	27.0	30.9	24.2	31.6
最小値	21.0	15.9	20.7	28.7	21.7	26.6
四分位範囲	7.7	8.2	8.7	1.9	3.6	4.7

Q５：

①東京・大阪 vs 軽井沢・札幌

（例）　人気の高い避暑地として知られる軽井沢や札幌は、東京や大阪と比べて、各日の最高気温が低いため、多くの観光客が訪れている。

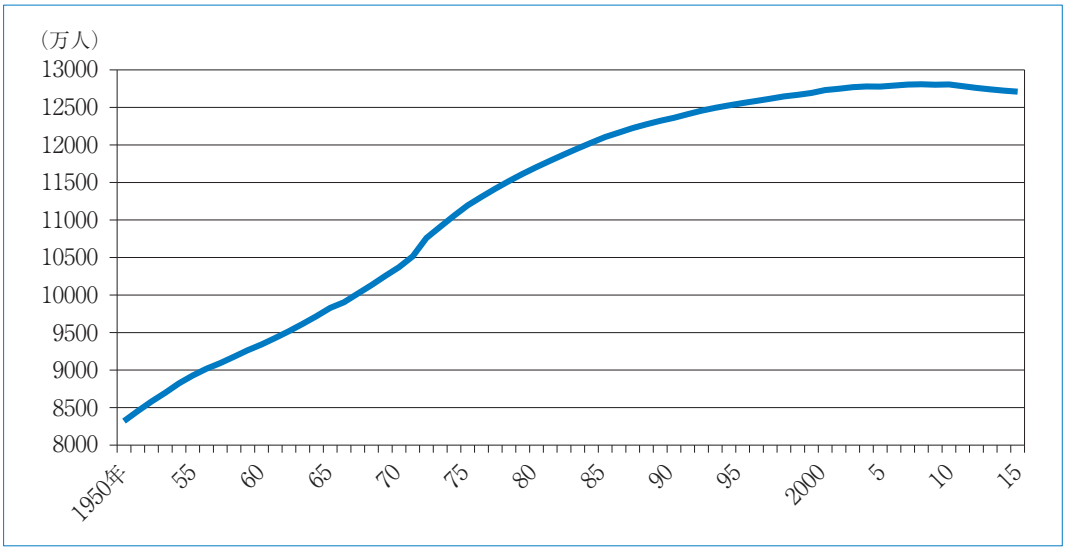
②軽井沢 vs 札幌

（例）　軽井沢と札幌のいずれも涼しいが、札幌の方が四分位範囲が狭いため、安定して涼しい。軽井沢は最高気温が20℃を下回る日もあるため、少し涼しすぎると感じた人がいる可能性もある。

Q６：略

2　地域の豊かさの格差は拡大しているか？［特性値の活用］

図１　日本の人口の推移



資料：総務省「国勢調査」、「人口推計　各年10月１日の推計人口」

日本の人口が1920年の国勢調査開始以来、初めて減少したことが、2015年国勢調査（総務省統計局）で明らかになった。５年前に比べて人口が減少したのは39の道府県にのぼる。東京を中心とした大都市に人口が集中する一方、地方の人口減少は大都市部から離れた県ではかなり以前から始まっている。第６部で取り上げる島根県では、1955年の92万９千人をピークとして2015年には69万４千人まで人口が１／４減少した。

STEP 1：Problem 問題　課題の設定

◇ 経済的な豊かさは地域間で拡大している？

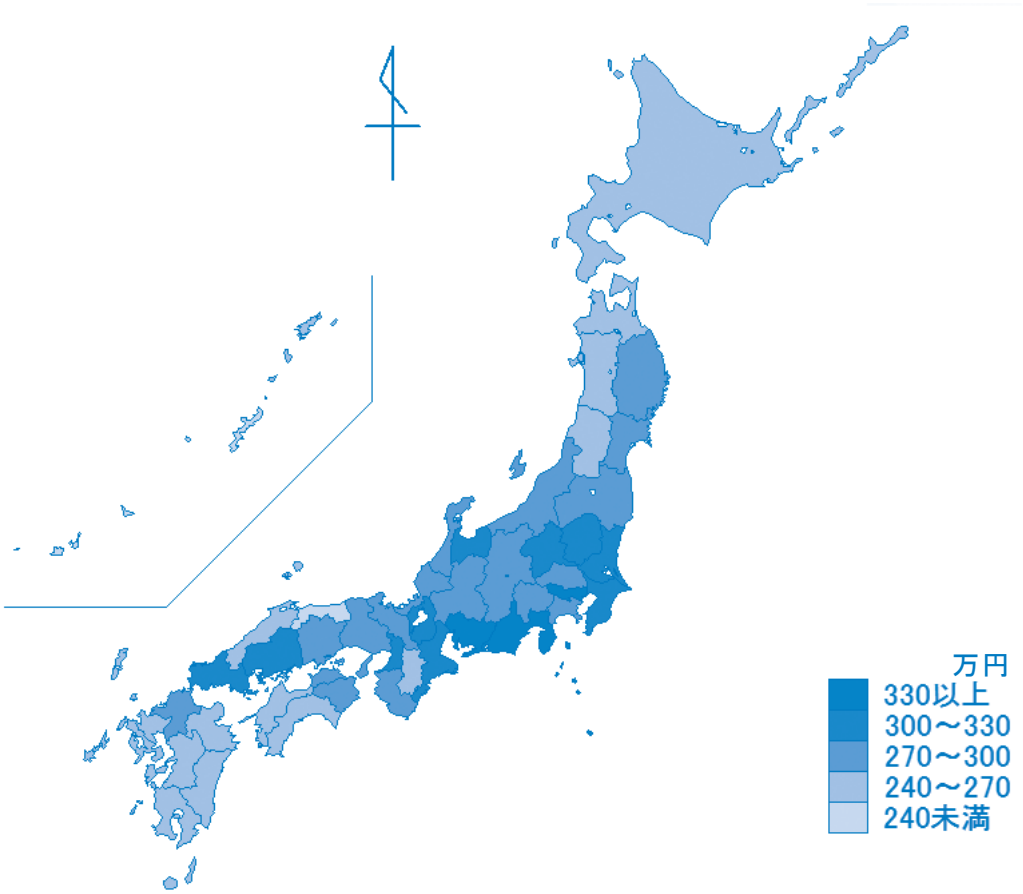
最近、世界各国で貧富の差が拡大していると言われている。我が国では、人口減少時代に入って、地方では人通りが少なく、かつての繁華街がシャッター通りとなっているところも増えている。人口は経済・社会の基盤を成すものであり、人口の増減は経済的な豊かさと密接に関わっていると言われる。近年の人口変動によって、地域間で経済的な豊かさの格差は拡大したのであろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ どのような指標で豊かさを捉えたら良いか？

地域ごとの経済的な豊かさを捉える指標を都道府県別の１人当たり県民所得とする。県民所得は企業を含めて県民全体の経済水準を表すもので、内閣府が推計する国民所得に準拠して作成されているので、都道府県間で比較可能な統計データである。各都道府県は人口規模が大きく異なるので、県民１人当たりの所得とすれば、地域の経済的な豊かさを捉える指標として適当であろう。

１人当たり県民所得（2013年度）



資料：内閣府「平成25年度　県民経済計算」

格差の大きさを１人当たり県民所得の都道府県間のバラツキで評価することとし、バラツキがこの40年間で拡大しているか否かで、地域ごとの経済的な豊かさに不均等が生じているかを明らかにする。

国民所得：国民全体が得る所得の総額をいう。個人や企業の所得は経済活動によって産み出された付加価値総額の配分であるので、経済活動の規模を表す指標である。
県民所得：国民所得の県民版

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ １人当たり県民所得の時系列データを47都道府県について収集しよう

１人当たり県民所得は内閣府「県民経済計算」から利用することができる。なお、人口については、我が国のすべての人を対象にして、５年に１回、「国勢調査」（総務省）が実施されているが、毎年の１人当たり県民所得のベースとなる人口は、「10 月１日現在推計人口」（総務省）に拠っている。表１に2013年度の１人当たり県民所得について、47都道府県ごとにその平均を示す。

表１　１人当たり県民所得（2013年度；万円）

01	北海道	255	17	石川県	297	33	岡山県	280
02	青森県	243	18	福井県	285	34	広島県	306
03	岩手県	270	19	山梨県	292	35	山口県	312
04	宮城県	286	20	長野県	271	36	徳島県	288
05	秋田県	246	21	岐阜県	273	37	香川県	280
06	山形県	263	22	静岡県	333	38	愛媛県	254
07	福島県	279	23	愛知県	358	39	高知県	245
08	茨城県	314	24	三重県	317	40	福岡県	283
09	栃木県	325	25	滋賀県	327	41	佐賀県	251
10	群馬県	305	26	京都府	297	42	長崎県	242
11	埼玉県	286	27	大阪府	300	43	熊本県	242
12	千葉県	302	28	兵庫県	282	44	大分県	256
13	東京都	451	29	奈良県	253	45	宮崎県	241
14	神奈川県	297	30	和歌山県	282	46	鹿児島県	240
15	新潟県	277	31	鳥取県	234	47	沖縄県	210
16	富山県	316	32	島根県	242		全県平均	307

資料：内閣府「県民経済計算」

表１の01～47の符号は1970年に統計に用いる標準地域コードとして定められたもので、各都道府県に概ね北東から南西に順に割り振られているよ！

STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ 格差の変動を捉えるために、バラツキ（散らばり）の指標を活用する

1 人当たりの県民所得を Y_i（$i=1, 2, \cdots, 47$）として、1 人当たりの県民所得のバラツキの指標として、**標準偏差** $s=\sqrt{\sum_{i=1}^{47}(Y_i-\overline{Y})^2/47}$（$\overline{Y}$ は平均で $\overline{Y}=\sum_{i=1}^{47}Y_i/47$）を利用する。標準偏差は**分散** s^2 の平方根である。

図 2 は、1975～2013年度の 1 人当たり県民所得の標準偏差を示している。表 1 に示されている2013年度のデータと同様の各年度データから計算されている。

図 2　1 人当たり県民所得の標準偏差の推移

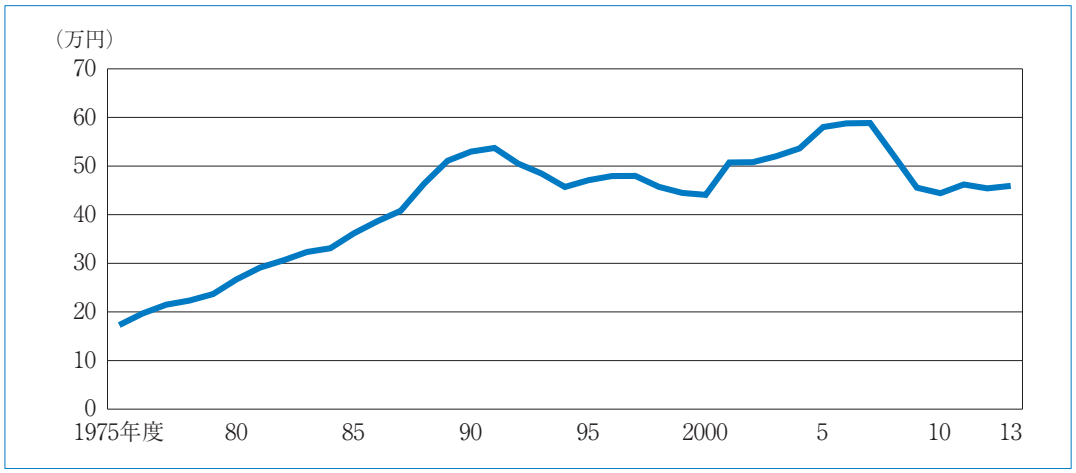
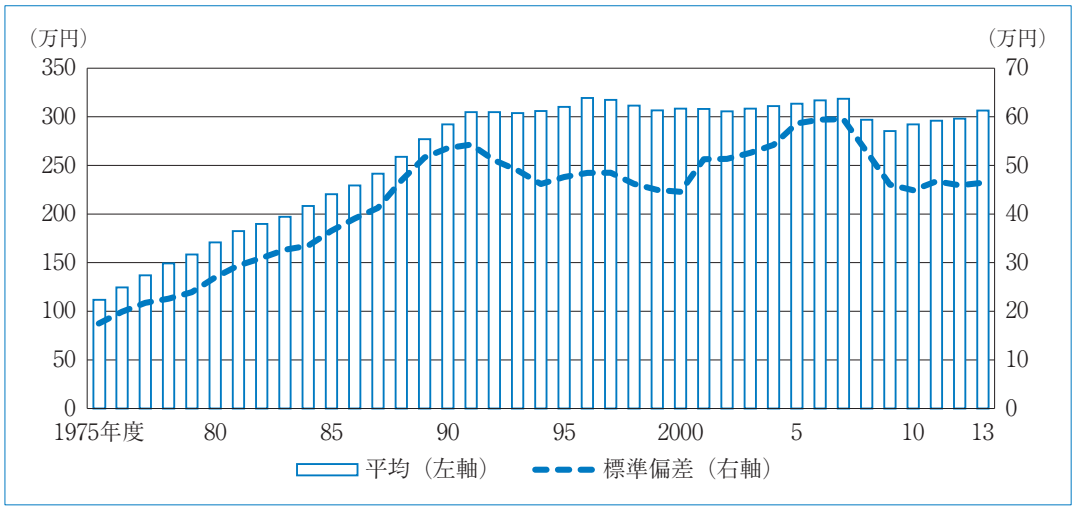


図 2 から、1975年～91年度のバラツキ（標準偏差）は一貫して増加しており、格差は拡大しているように見える。図 2 に各年度の 1 人当たり県民所得の平均を追加したのが図 3 である。

図 3　1 人当たり県民所得の平均と標準偏差の推移



1975年～91年度の期間は標準偏差が大きくなっているが、平均も同様に増加している。経済成長に伴って、所得が増加した結果、標準偏差も大きくなったことに注意しなければならない。所得水準が異なる時点で相違すれば、各時点の所得のバラツキを評価する際、各時点の所得水準を調整した指標に基づかないと適切な比較ができないことを理解しよう。標準偏差は平均からのバラツキの度合いを求めるもので、その大きさは対象とするデータの平均で代表される水準によって異なった値となる。仮に、Y_i を円単位と万円単位の２種類の数値で標準偏差 s を計算すると、前者の s の値は後者の s の値の１万倍になることを容易に確認できるであろう。

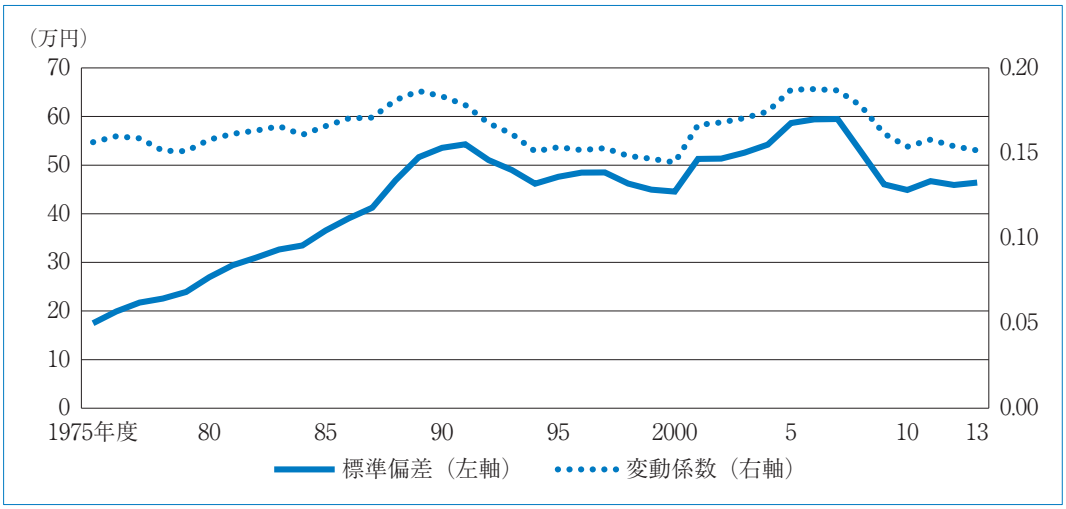
Q１：2013年度の１人当たり県民所得が万円単位で与えられたとき、その標準偏差は45.9である。１人当たり県民所得が円単位で示されたとき、その標準偏差はどうなるか？

１人当たり県民所得の標準偏差＝

バラツキを評価するのが同じデータ項目であっても、異なる時点や異なる属性のグループのバラツキを相互に比較する際、標準偏差を平均で割って求める変動係数が広く用いられている。**変動係数**は分母となる平均で水準を調整しているので、これを用いれば、より的確に１人当たり県民所得の各時点のバラツキを相互に比較することができる。

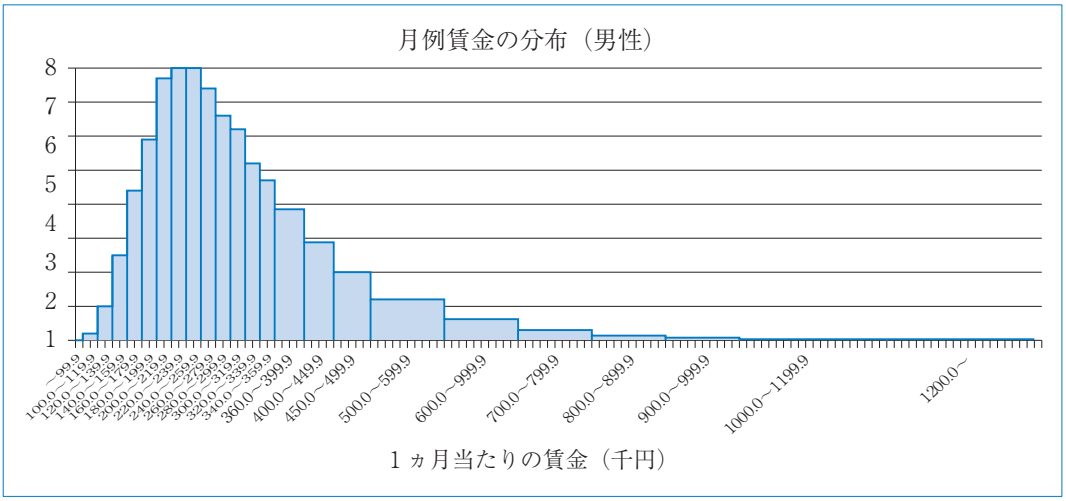
図４は、１人当たり県民所得の標準偏差と併せて変動係数を図示している。標準偏差が大きくなった1975～91年度の期間の変動係数はそれほど大きくなっていないことが分かる。変動係数の値は時期によって相違するが、2013年度は0.15であり、1975年度の0.16と比べてほぼ同様な水準にある。

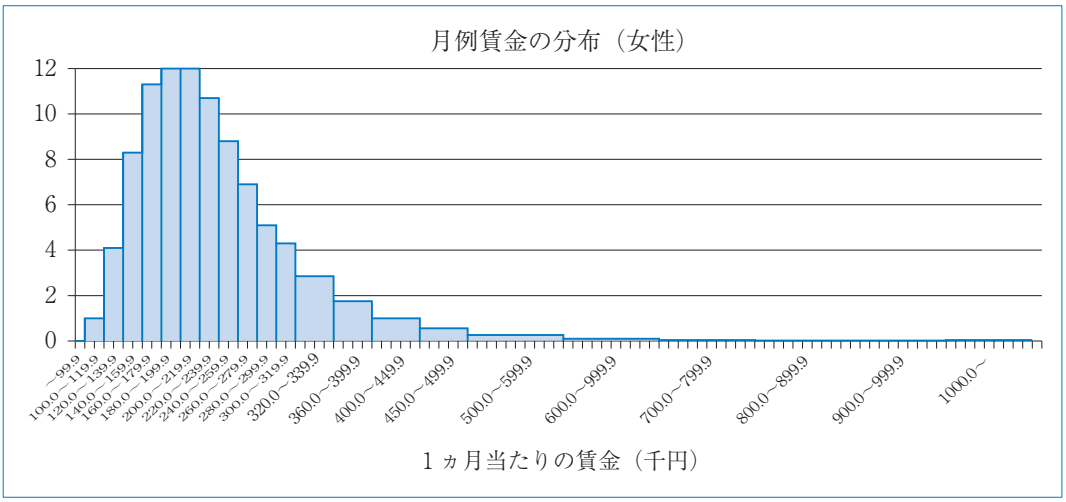
図４　１人当たり県民所得の標準偏差と変動係数



歪度（skewness）と尖度（kurtosis）

図５　賃金分布





資料：厚生労働省「平成27年賃金構造基本調査」

図５は2015年の男女別の賃金分布をヒストグラムで図示している。男性と女性のいずれの賃金分布も右裾が長くなっていて左右対称ではなく、平均が中央値より大きく乖離している。このような分布の歪みの程度を示す統計量が**歪度**（わいど）であり、$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$ で与えられる（$\bar{x}$ は平均、s は標準偏差）。分布が右に歪んでいるとき、歪度は算式の３乗項によって右裾にある少数のデータの影響を大きく受けて正の値をとる。他方、左に歪んでいる分布のとき、同様の理由で歪度は負の値をとる。分布が対称のとき、歪度は０となるので、歪度は分布の対称性を知る特性値となる。

その他、分布の尖り具合（裾の厚さ）の程度を示す統計量が**尖度**（せんど）であり、$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4$ で与えられる。第４部で紹介する正規分布の場合、尖度は３となるので、尖度が３より大きい分布のとき、正規分布より分布の裾が厚い、あるいは平均から大きく離れた値が多いと判断できる。また、外れ値の存在を知る指標ともなりうる。

このように、平均、分散（標準偏差）、歪度、尖度等の特性値を求めれば、分布の形状をある程度知ることができることが理解されるであろう。

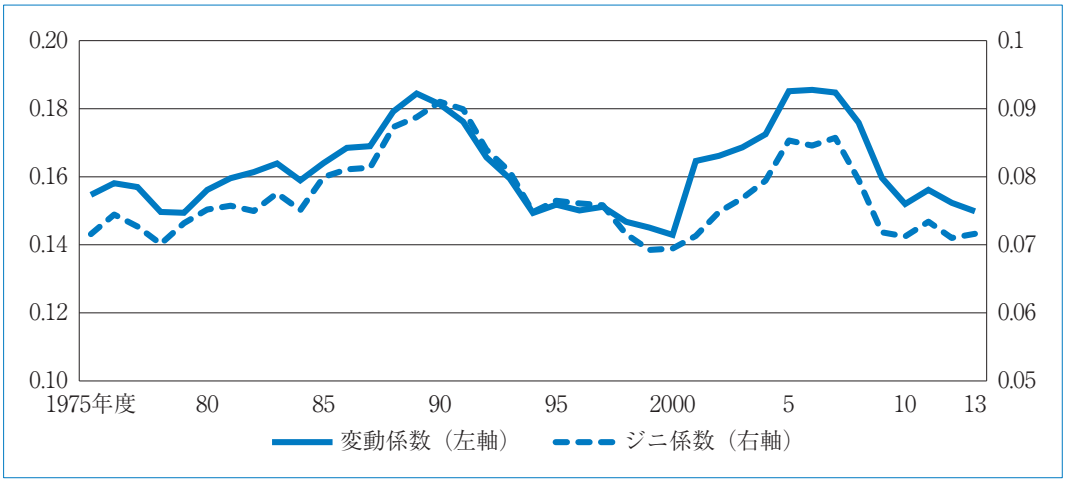
近年、所得の不平等を示す指標として、**ジニ係数**が広く一般に知られるようになってきており、昨今の国会論戦やニュースの中でもたびたび登場する。

ジニ係数（G）は、1936年にイタリアの数理統計学者のコッラド・ジニ（Corrado Gini）が社会における所得分配の不平等さを測る指標として考案したもので、

$G=\Sigma\Sigma|Y_i-Y_j|/2n^2\overline{Y}$として求められる。$0 \leqq G \leqq 1$であり、Gが０のとき完全平等を示し、１に近づくに従って不平等の程度が大きくなる。

図６にジニ係数と変動係数を併せて示す。

図６　１人当たり県民所得の変動係数とジニ係数



２つの指標はほぼ同じ変動を示しており、いずれに基づいても2013年度に至る約40年で都道府県間の所得格差が拡大したことは示していない。

ジニ係数とローレンツ曲線

所得分布の状況を図示したのが**ローレンツ曲線**で、米国の統計学者のマックス・ローレンツ（Max Lorenz）が1905年に考案しました。横軸に所得額の大きさの順に所得人数の累積百分比を、縦軸にそれに対応した所得金額の累積百分比をとって得られる曲線をいいます。所得分布が完全に平等であれば、曲線は対角線に一致し（均等分布線）、そこから下方に位置すればするほど不平等度が大きいという性質をもつので、所得分布の不平等度を測定することができます。

２つのグループの所得分布の不平等度を比較するとき、ローレンツ曲線が交わってしまうと、どちらがより不平等であるかが判断しづらくなります。ローレンツ曲線に基づいて考案されたのがジニ係数です。ジニ係数はローレンツ曲線と均等分布線によって囲まれた面積の正方形の面積に対する割合を２倍した値として求められます。均等分布線からの乖離が大きいほどジニ係数の値は大きくなるので、複数のグループの不平等度を数量的に比較できます。

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ 地域間格差は拡大しているか？

変動係数やジニ係数の約40年間の推移を見る限り、地域間で経済的な豊かさの格差が拡大したとはいえない。ただし、1980年代後半のバブル期に至る時期にかけてやや格差が拡大した。その後の景気低迷の中で、景気対策としての地方への公共事業の重点配分などの結果、地域間格差は低下した。景気対策の後遺症として、国と地方の財政赤字は膨れ上がり、小泉政権下での公共事業抑制や行政改革への転換に伴って、地域間格差は拡大した。その後の政策転換と2008年９月のリーマンショック後の世界不況のもとで地域間格差は縮小し、現在に至っている。バブル崩壊後やリーマンショック後の景気の大きな後退局面では、地域間格差が縮小している。

Q２：なぜ、景気後退期に地域間格差は縮小しているのだろうか？

〔本節の解答〕

Q１：１人当たり県民所得の標準偏差＝459,000

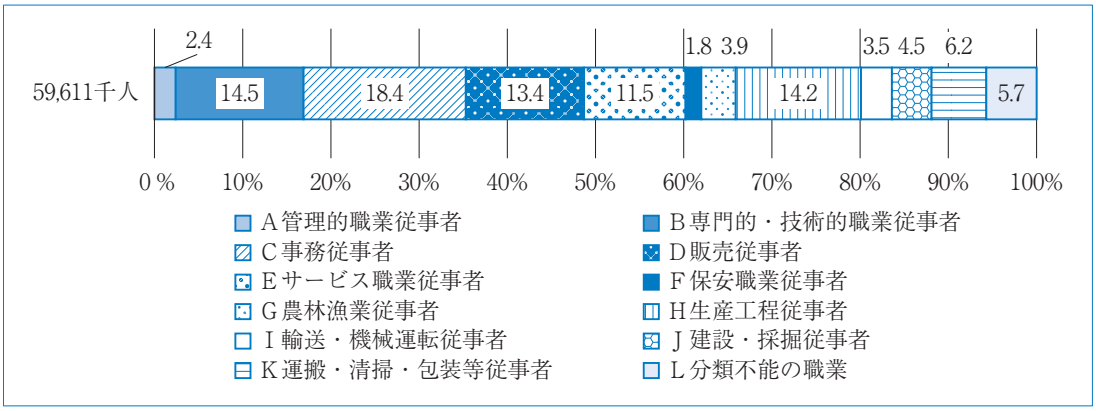
Q２：大都市圏から離れた地域では、地域経済のなかで農林水産業、建設業、公務の占める役割が大きいところが多い。

3　サービス経済化の状況とその背景を探る［関係の度合］

高校生の航平君は将来の進路を考えている。

「僕の父はIT企業、母は病院に勤務している。公介君の父は学校の先生、母は流通企業、健三君の父は金融業、……のように、周りの友達の親の勤務先はサービス業が多い。ところが、祖父は日本の高度経済成長が始まるころ就職したそうで、製造業が花形産業で友達の多くもそこで働いていたという。」

図１　就業者の職業構成



資料：総務省「平成22年国勢調査」

職業分類と産業分類

同じような仕事を一まとめにして統計データを有効に利用するために、職業分類が設けられています。図１の「就業者の職業構成」で表示されている職業は、総務省が定めている「**日本標準職業分類**」の大分類項目です。報酬を伴う仕事に就いている人の職業を大括りにしたもので、たとえば、「管理的職業従事者」には経営者や管理者等が分類されます。「専門的・技術的職業従事者」には学者、技術者、教員、医者、弁護士等が分類されます。それぞれの大分類の下に中分類、小分類が設けられていて、その分類項目からより詳細な職業に関するデータが利用できます。

同じような事業活動を行っている工場や事務所、営業所、店舗等を一まとめにしたのが産業です。職業分類と同様に、総務省が統計データを産業ごとに表すために「**日本標準産業分類**」を定めています。「農業、林業」、「建設業」、「製造業」、「情報通信業」、「卸売業、小売業」、「金融業、保険業」など、20の大分類項目が設けられていて、その下に、中分類、小分類、細分類があって、細分類は1460の項目数です。

STEP 1：Problem 問題　課題の設定

◇ なぜ近年はサービス業（第 3 次産業）で働いている人が多いのだろう？

サービス経済化といわれて久しいが、いま、どのような状況であるのか、また、その主な要因は何だろうか？

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ サービス経済化の状況とその背景を探る

サービス業で仕事をしている人が実際に増加して、全体の中で比率が上昇しているかを確認する。併せて、サービス業の生産活動についても高まっているかを確認する。次いで、サービス経済化を解明するための仮説として広く受け入れられている、**ペティ＝クラークの法則**を統計データに基づいて実証する。

ペティ＝クラークの法則

経済の発展につれて、国民経済に占める第 1 次産業の比重は次第に低下し、第 2 次産業、次いで第 3 次産業の比重が高まるという、産業構造の高度化を説明したものです。自然界から採取する農林業、漁業等の産業を第 1 次産業、その産出物を加工する製造業、建設業等の産業を第 2 次産業、それ以外の産業を第 3 次産業と大別したのは、ケンブリッジ大学の経済学者コーリン・クラーク（Colin Clark）で、統計学の始祖とも称されるウィリアム・ペティ（William Petty）が『政治算術』の中に記した内容を整理して1941年に提示しました。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ サービス経済化を就業者と GDP の産業別構成比から確認しよう！ サービス経済化の要因は？

産業別の就業者数は総務省「労働力調査」から1953年以降のデータが利用可能である。また、産業別の経済活動については、内閣府「国民経済計算」から利用できるが、SNA（国民経済計算）の基準改定によって1990年以降とそれ以前で経済活動の範囲が若干変更していることは留意しておこう。表 1 に1955年以降の産業別就業者数の推移を示す。

表１　産業別就業者数

年次	総数（万人）	第１次産業	第２次産業		第３次産業
			建設業	製造業	
1955	4090	1581	195	757	1557
56	4171	1539	197	805	1630
57	4281	1517	217	853	1694
58	4298	1453	223	898	1724
59	4335	1396	243	896	1800
60	4436	1383	253	946	1854
61	4498	1341	274	1011	1871
62	4556	1308	290	1066	1892
63	4595	1227	290	1108	1968
64	4655	1179	308	1129	2038
65	4730	1142	328	1150	2109
66	4827	1098	350	1178	2201
67	4920	1062	359	1252	2247
68	5002	1015	370	1305	2305
69	5040	970	371	1345	2348
70	5094	906	394	1377	2408
71	5121	834	414	1383	2484
72	5126	771	433	1383	2530
73	5259	718	467	1443	2619
74	5237	689	464	1427	2646
75	5223	677	479	1346	2712
76	5271	661	492	1345	2764
77	5342	653	499	1340	2839
78	5408	648	520	1326	2904
79	5479	625	536	1333	2976
80	5536	588	548	1367	3019
81	5581	567	544	1385	3073
82	5638	558	541	1380	3145
83	5733	541	541	1406	3229
84	5766	520	527	1438	3260
87	5911	497	533	1425	3432
88	6011	481	560	1454	3487
89	6128	470	578	1484	3566
90	6249	457	588	1505	3668
91	6369	433	604	1550	3752
92	6436	417	619	1569	3801
93	6450	389	640	1530	3862
94	6453	379	655	1496	3893
95	6457	373	663	1456	3939
96	6486	362	670	1445	3979
97	6557	357	685	1442	4039
98	6514	349	662	1382	4084
99	6462	341	657	1345	4078
2000	6446	331	653	1321	4102
01	6412	318	632	1284	4133
02	6330	301	618	1202	4158
03	6316	298	604	1178	4176
04	6329	290	584	1150	4236
05	6356	285	568	1142	4284
06	6389	275	560	1163	4320
07	6427	277	554	1170	4352
08	6409	273	541	1151	4370
09	6314	267	522	1082	4380
10	6298	258	504	1060	4411
11	6289	252	502	1049	4431
12	6270	243	503	1032	4430
13	6311	236	499	1039	4445
14	6351	233	505	1040	4474
15	6376	231	500	1035	4509

資料：総務省「労働力調査」

注：総数には就業先不詳を含む。1972年以前の数値には沖縄県は含まれていない。2002年以降の数値は日本標準産業分類改定を踏まえているが、それ以前は製造業とサービス業の間で若干の業種移動がある。

Q１：表１に基づいて、第３次産業就業者の比率を求めよう！

総数には就業先産業が不詳の就業者を含むので、第３次産業の就業者比率は第１次産業、第２次産業、第３次産業の就業者の合計から求めることが必要だよ！

図２に1955年以降の第３次産業就業者の比率の推移を示す。また、図３に第３次産業の総生産（GDP）の構成比の推移を示す。

図２　第３次産業就業率

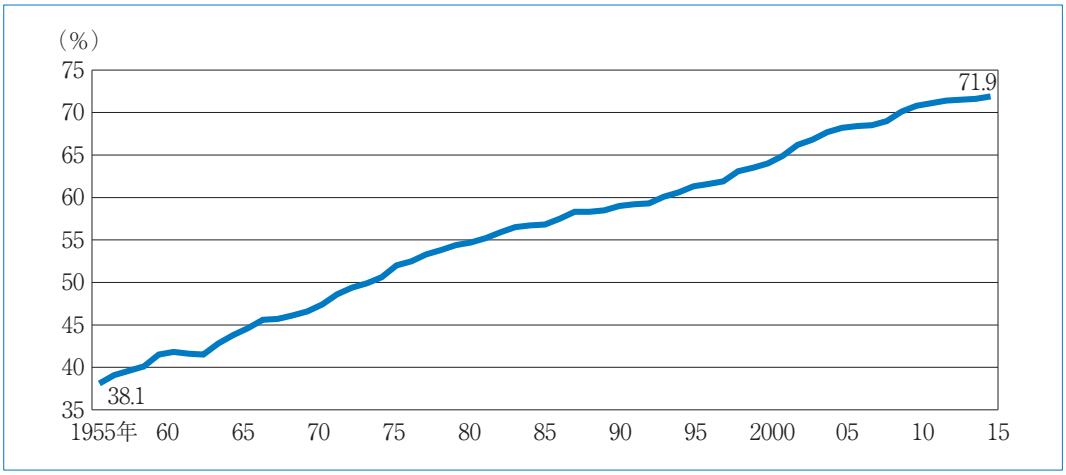
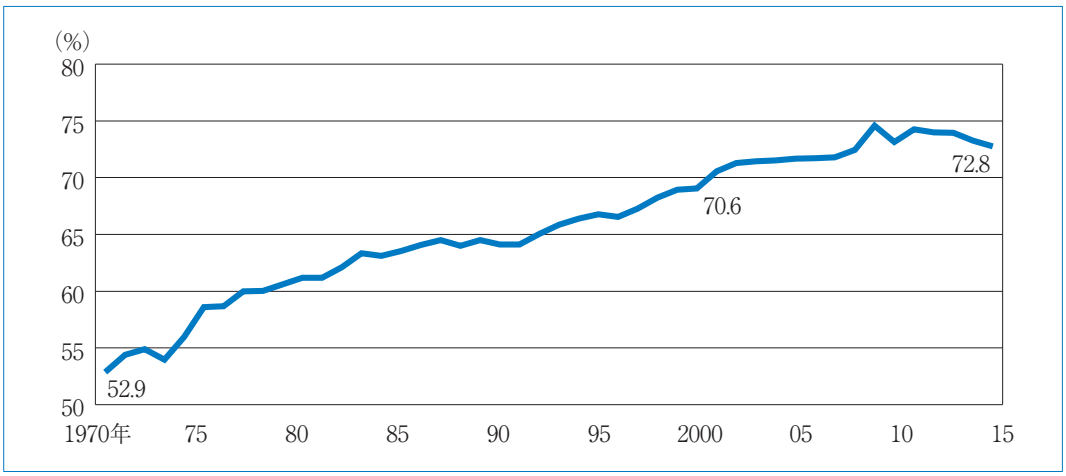


図３　第３次産業の GDP 構成比



資料：内閣府「国民経済計算」、総務省「労働力調査」

Q２：経済の発展を表す指標として、どのようなデータが適切であろうか？

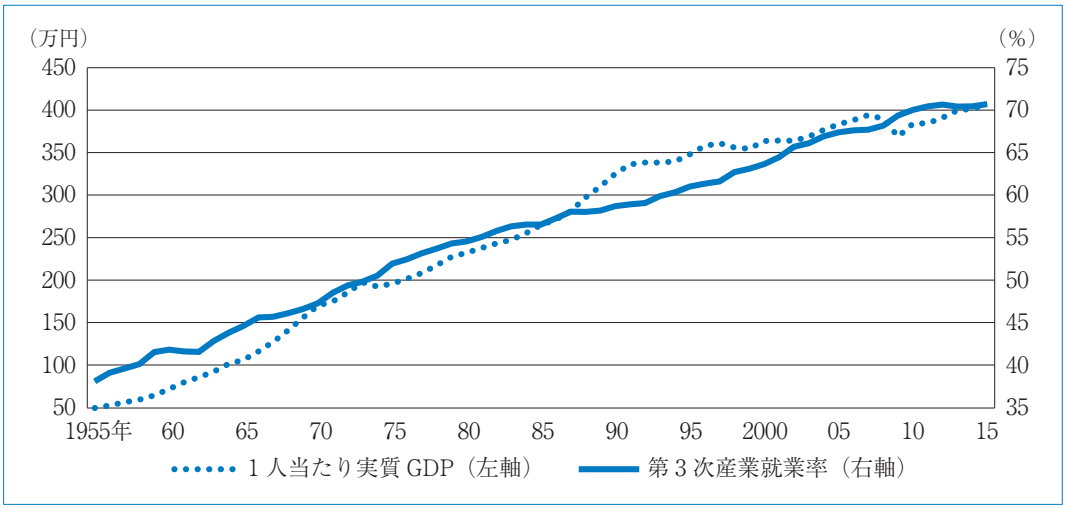
STEP 4：Analysis 分析　グラフや統計量で傾向を捉える

◇ サービス経済化の背景を探るため散布図を活用

第３次産業の就業者数は1955年の1557万人から2015年には4509万人へと約60年間に2.9倍に急増し、全産業に占める比率も38.1％から71.9％へ上昇の一途をたどっている。同様に、GDP についても、第３次産業の構成比は拡大しており、2001年には70.6％と70％を上回り、2015年は72.8％となった。

図４に１人当たり実質 GDP と第３次産業の就業者比率を併わせて図示すると、両者の推移が軌を一にしていることが分かる。

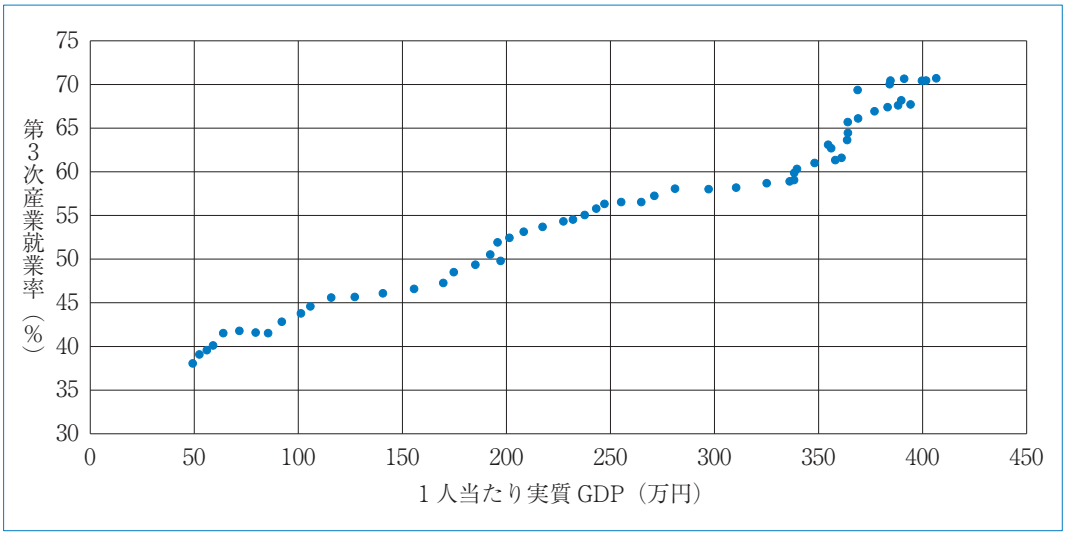
図４　１人当たり実質 GDP と第３次産業就業率の推移



資料：内閣府「国民経済計算」、総務省「労働力調査」

ペティ＝クラークの法則に従って、横軸に国の豊かさを表す指標として物価変動を調整した１人当たり**実質** GDP、縦軸にサービス経済化の指標として第３次産業就業率をとって**散布図**（図５）を描くと、右肩上がりの関係が明確に読み取れる。

図５　１人当たり実質 GDP と第３次産業就業率



1人当たり実質GDPをX、第3次産業就業率をYとしたとき、XとYの関係の度合いを**相関係数**rの値から量的に捉えることができる。XとYのn個のデータが与えられたとき、

rは $\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})/\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}$　$(i=1,2,\cdots,n)$ の式で求められる。

$-1 \leqq r \leqq 1$であり、$r=1$のとき完全に正の相関、$r=-1$のとき完全に負の相関といい、いずれもXとYのすべてのデータが直線上に位置する。また、rが0近傍のとき、XとY の間にはほとんど関係がない。rが± 1に近いほど相関の度合いが高い。

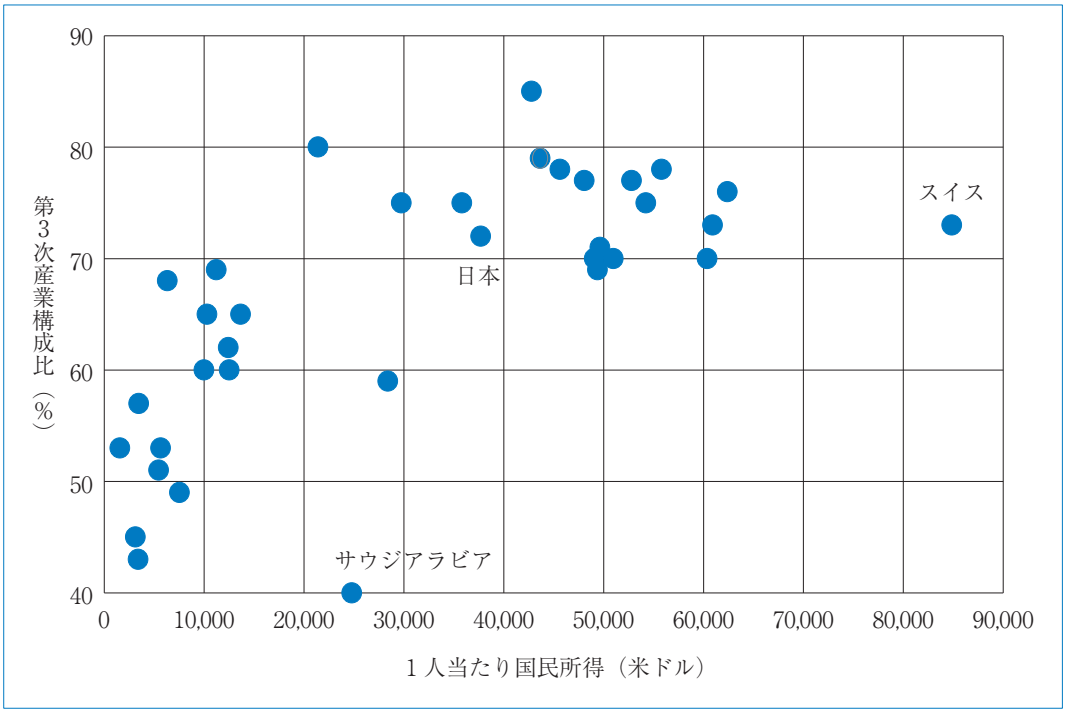
相関係数（correlation coefficient）

相関の概念を提唱し、相関係数rを最初に用いたのはゴールトン（Francis Galton）です。ゴールトンはダーウィンの従兄弟で、ダーウィンの「種の起源」に刺激を受け、親から子への遺伝の法則を探すために、親と子の身長や体重、骨格などを測定し、そこから相関係数の考えに至りました。ただ、ゴールトンは数学が得意ではなかったため、相関係数を定式化したのは弟子のカール・ピアソン（Karl Pearson）で、通常、相関係数といえば、ピアソンの名を冠して、ピアソンの積率相関係数と称されます。他に、分布に関する仮定を置かないノンパラメトリックな相関係数として、スピアマンの順位相関係数、ケンドールの順位相関係数などがあります。

日本において、経済の発展とサービス経済化には強い関係があることが分かったが、ペティ＝クラークの法則を世界各国のデータからも検証できるであろうか？

図6は2014年の各国の1人当たり国民所得（米ドル換算）と第3次産業のGDP構成比を図示している。サウジアラビアを除けば、全体として2つの指標の関連度はかなり高いことを確認できる。

図6　1人当たり国民所得と 第3次産業構成比



資料：日本統計協会「世界の統計」

STEP 5：Conclusion 結論　結論を導き、新たな課題を見出す

◇ サービス経済化は経済的豊かさの象徴

日本について、長期的な経済発展とともにサービス経済化が進展し、今日の状況にあることをデータに基づいて実証することができた。また、このような関係を示すペティ＝クラークの法則について、特定時点の世界各国の統計データからも、同様に実証することができた。

Q３：サウジアラビアは１人当たり国民所得（米ドル換算）と第３次産業の GDP 構成比についての世界各国で観察される関係から、何故、外れて位置するのだろうか？サウジアラビアを除くと関係の度合いはどのように変わるのだろうか？

〔本節の解答〕

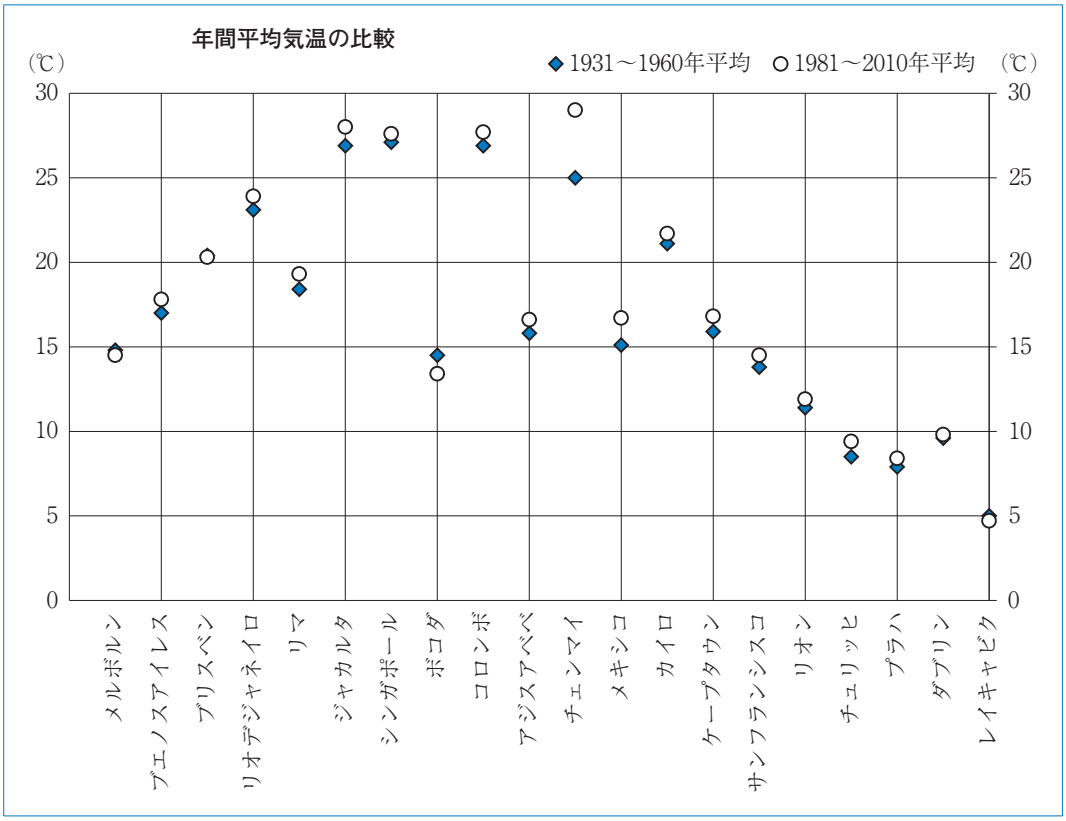
Q１：

1955年	38.1	65	44.6	75	51.9	85	56.5	95	61.0	05	67.4
	39.1		45.6		52.4		57.3		61.3		67.6
	39.6		45.7		53.1		58.1		61.6		67.7
	40.1		46.1		53.7		58.0		62.7		68.2
	41.5		46.6		54.3		58.2		63.1		69.4
60	41.8	70	47.3	80	54.5	90	58.7	2000	63.6	10	70.8
	41.6		48.5		55.1		58.9		64.5		71.1
	41.5		49.4		55.8		59.1		65.7		71.4
	42.8		49.8		56.3		59.9		66.1		71.5
	43.8		50.5		56.5		60.3		66.9		71.6
										15	71.9

Q２：１人当たり国民所得、１人当たり GDP、１人当たり所有資産等、人口１人当たりで求めることが適当である。

Q３：（前半の問いの解答は略）サウジアラビアを含めた相関係数は0.687であるが、サウジアラビアを除くと0.728となり、関係の度合いが高まることが見て取れる。

4　都市の平均気温と緯度はどんな関係？［散布図・相関分析による問題解決］



資料：国立天文台「理科年表」
注：数ヶ国で平均気温の対象年次が表記と異なっている

地球の温暖化の影響なのか、50年前と比べると主要な都市の平均気温は上昇している。ただし、各地域の平均気温の上昇よりも、地域別の平均気温の相違が著しい。地球上のさまざまな地域の年間平均気温は何によって決まっているのだろうか？

STEP 1：Problem 問題　課題の設定

◇ 自然現象に関わっている自然の法則を統計的に考察しよう！

先生：地球温暖化が進行していると言われています。既に避暑地の問題のときに議論したように、信州が、東京より夏に涼しいこともよく分かっています。そもそも私たちが住む地球は、赤道付近は暑いし、極地に近づくほど寒くなることは誰でも知っていますね。ここでは、世界のさまざまな地域の標準的な年平均気温はどのように決まっているのかといった問題を統計的に解決してください。

STEP 2：Plan 計画　どのようなデータ・統計資料を集めて分析するか

◇ 地域の年間平均気温に影響を与える要因を考えてみよう

先生：世界の各地域で年間平均気温は異なっています。各地域での年間平均気温に影響を与える原因にはどのようなものが考えられるか、皆で議論して、**特性要因図**を描いてください。その原因候補の中で、皆さんが比較的簡単にデータをとれるものを考えてみましょう

航平：地域の人間の活動が影響を与えているのではないかな？

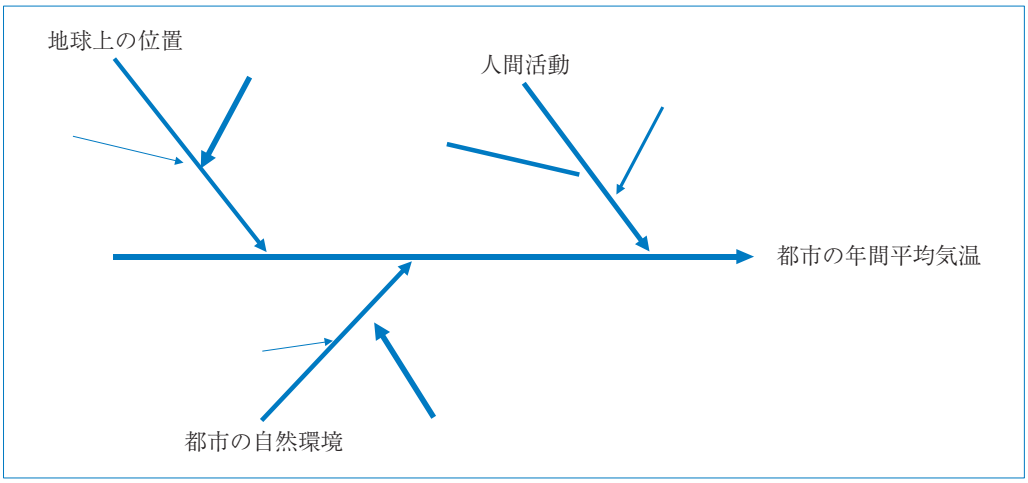
理恵：都市の自然環境も重要でしょう。軽井沢は涼しいって言ってたじゃない

航平：自然環境といっても、いろいろ考えられそうだね

公介：北海道が涼しいのは北にあるからだよね

先生：地域の地球上の位置、都市の自然環境、人間活動が地域の年間平均気温に影響を与える要因（原因候補）と考えられるようだね。それを特性要因図に描くと図 1 のようになります。これは大きな原因候補（大骨）を網羅しただけです。

図 1　特性要因図：都市の年間平均気温に影響を与える要因の網羅



Q 1：図 1 の特性要因図の大きな原因候補に関係し、データをとれる可能性のある具体的な原因候補を特性要因図の大骨上の対応する位置（中骨）に書き込んでください。

航平：大骨とか中骨って何か変な名前ですね

先生：特性要因図は、アメリカでは「魚の骨図（Fish Bone Diagram）」と呼ぶこともあるのです。原因の候補を魚の骨の大きな骨、それを具体化した原因を小さな骨に例えているイメージなのです。

理恵：特性要因図はアメリカで考案されたのですか？

先生：いいえ、日本の石川馨先生という方が考案したものです。海外でも石川ダイアグラムと呼ぶ人もいます。

STEP 3：Data 収集　必要なデータ・統計資料を集める

◇ 世界の都市の平均気温などを集めてみよう

世界各都市の平均気温は、理科年表や気象庁のホームページで調べることができる。航平君たちは、表１のようなデータを理科年表からとってきた。ただし、別途１か所だけ都市ではなく、南極の昭和基地のデータを調べた。

表１　世界の25都市の年間平均気温（℃）、緯度（北緯、南緯はーで表示）、標高（m）

地名	平均気温	緯度	標高	地名	平均気温	緯度	標高
昭和基地	−10.5	−69.00	18	ドーハ	27.0	25.15	11
メルボルン	14.5	−37.39	132	カイロ	21.7	30.06	116
ブエノスアイレス	17.8	−34.35	25	ケープタウン	16.8	33.58	46
ブリスベン	20.3	−27.23	4	東京	15.4	35.42	25
リオデジャネイロ	23.9	−22.55	5	サンフランシスコ	14.5	37.37	6
リマ	19.3	−12.01	12	北京	12.9	39.56	55
ジャカルタ	28.0	−6.11	8	サラエボ	10.4	43.52	630
シンガポール	27.6	1.22	5	リオン	11.9	45.43	197
ボゴダ	13.4	4.42	2547	チュリッヒ	9.4	47.22	555
コロンボ	27.7	6.54	7	プラハ	8.4	50.06	380
アジスアベベ	16.6	9.02	2354	ダブリン	9.8	53.26	68
チェンマイ	29.0	13.00	13	レイキャビク	4.7	64.08	54
メキシコ	16.7	19.24	2309				

資料：国立天文台「理科年表」

STEP 4：Analysis 分析　グラフから緯度と平均気温との関係を捉える

◇ 平均気温と関係があるデータを探ってみよう！

公介：データの間の関係性が強いかどうかは**相関係数**を計算すれば良いと習いましたので計算してみます
先生：それは本当かな？

Q２：緯度と平均気温の相関係数、標高と平均気温の相関係数を求めなさい。

緯度と平均気温の相関係数＝
標高と平均気温の相関係数＝

航平：どうも相関係数はあまり高くないようだな？

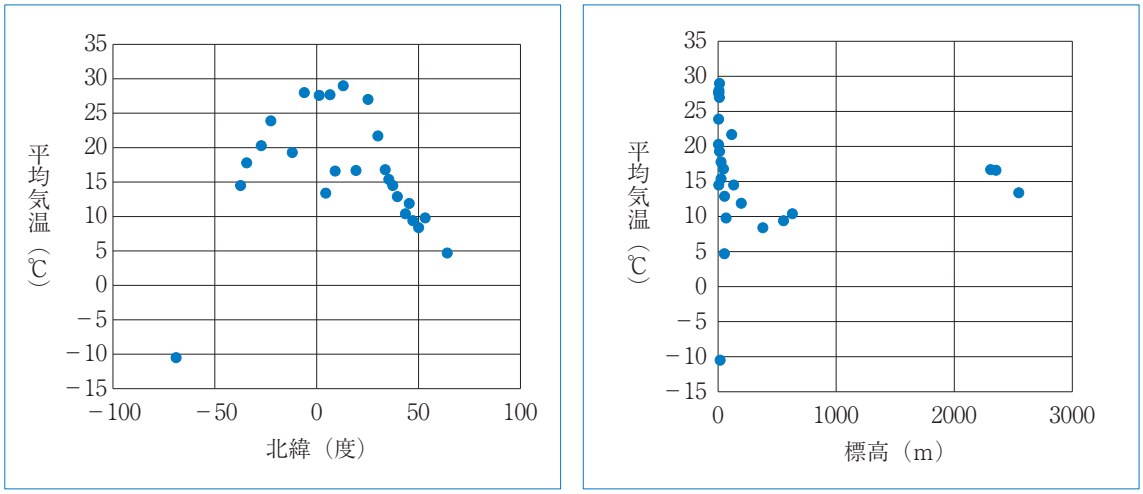
相関係数は、データの直線関係の強さを示します。散布図でデータが直線の上に乗っていれば、±１になるのですが、２次関数の上に乗っていても、必ずしも±１になるわけではありません。

◇ 散布図を注意深く眺めて関係性を想像しよう

先生：まずは、関係性が直線的かどうか、散布図というグラフで確認することが分析の第一歩です

表 1 のデータで緯度を横軸に、平均気温を縦軸にとった散布図と標高を横軸に平均気温を縦軸にとった**散布図**を図 2 に示す。

図 2　緯度・標高と平均気温の関係（散布図）



先生：図 2 の散布図から、どのような関数関係が想像できますか？

理恵：関数というのは、平均気温のデータをきちんと緯度で表現できる式のことですか？

先生：近似的に表現できる式ならば良いのです

公介：近似という意味が良く分かりません？

数学では変数 x と y の関係性は、関数 $y = f(x)$ を用いて表現する。一方、世界のさまざまな都市の平均気温 y と緯度 x との関係は、散布図を観察すれば、近似的にしか関係性は成立していない。実際、各都市のデータは、図 2 において想像された関数の上に正確に乗っているわけではない。散布図上の n 個の観測データの座標を (x_i, y_i), $i=1, \cdots, n$ とすれば、想像した関数に対して、その座標は

$$y_i = f(x_i) + e_i$$

と表現できる。e_i は、関数 $f(x_i)$ をデータに当てはめたときのハズレ（乖離）と考えれば良い。この e_i を統計では**残差（residual）**と呼ぶ。

このように、データ y は、関数で説明できる項と説明できない項の和に分けて表現される。

航平：緯度と平均気温の関係を示す図 2 の左側の散布図を見ると、緯度 0 度の赤道近辺を頂点として、緯度と平均気温は、上に凸の 2 次関数のような形状を示しているように思えます

理恵：南緯と北緯では、符号が違うけれど赤道について対称な関係のようです

公介：標高と平均気温の関係は、表 1 のデータが標高の低いところに固まっていて、その関係は思ったより無いね

◇ 関係性を定量的に表現してみよう

先生：それでは緯度と平均気温の関係をデータから示してみてください

一同：どうやれば良いのでしょうか？

データが近似的に示す関数関係を推察する簡単な方法は、条件付き平均値を求めることです。ある緯度帯を設定して、その緯度帯に入る都市を調べて、それらの都市の年間平均気温を平均や標準偏差を求めると良い。

Q３：緯度と平均気温との関係性を定量的に導きなさい

＊＊ヒント＊＊　表１の25都市データを緯度の絶対値の昇順に５都市ずつに並べ替え、その順番に５つの群に区分し、各群に属する都市を低緯度順に以下に示す。

群１　シンガポール、ボコタ、ジャカルタ、コロンボ、アジスアベバ
群２　リマ、チェンマイ、メキシコ、リオデジャネイロ、ドーハ
群３　ブリスベン、カイロ、ケープタウン、ブエノスアイレス、東京
群４　サンフランシスコ、メルボルン、北京、サラエボ、リオン
群５　チューリッヒ、プラハ、ダブリン、レイキャビク、昭和基地

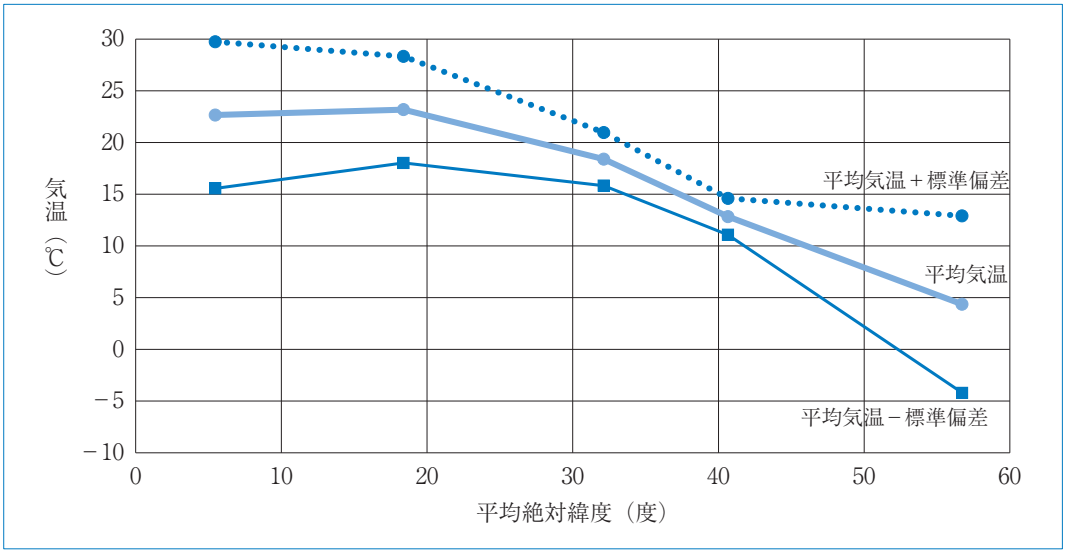
これから、各郡の緯度（の絶対値）の平均と平均気温の５都市平均・標準偏差を求め、表２に書き込みなさい。

表２　緯度の５層区分別の平均絶対緯度および平均気温の平均と標準偏差

群	平均 絶対緯度	平均気温の 平均	平均気温の 標準偏差
1			
2			
3			
4			
5			

５つの群について、表２で求めた平均絶対緯度を横軸として、縦軸に平均気温の平均と平均気温の平均±標準偏差を折れ線グラフに記すと図３となる。

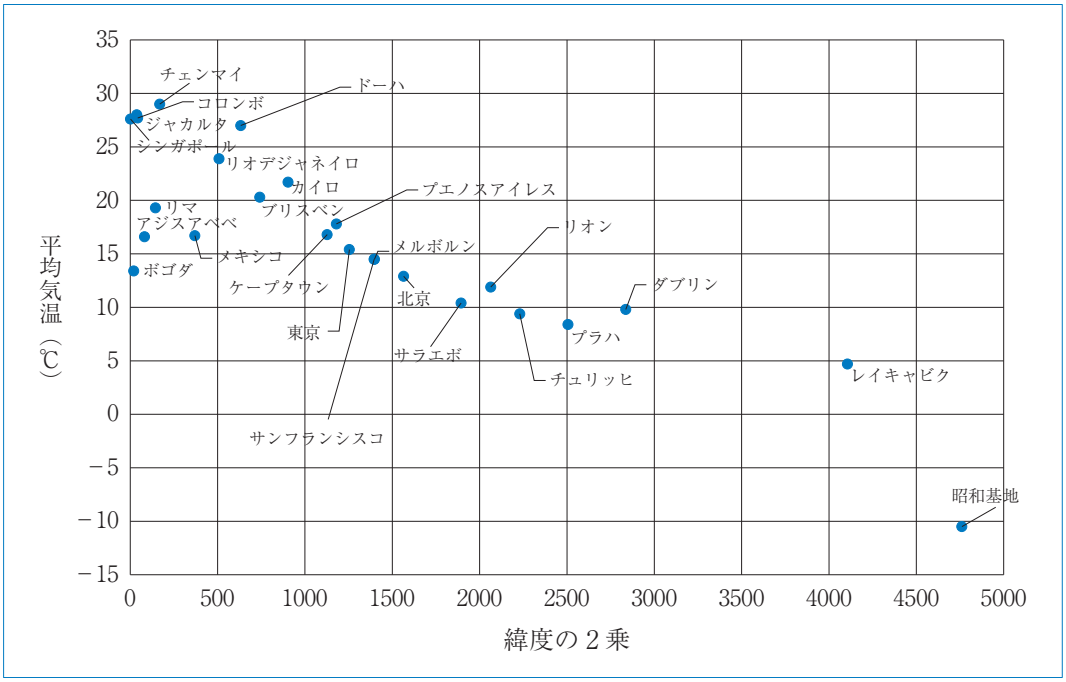
図３　平均緯度（平均）と平均気温、平均気温の平均±標準偏差



航平：図３をみると緯度と平均気温の関係は２次関数に見えますね
先生：２次関数だけが候補だとは思えないけれどね
理恵：三角関数みたいなものかもしれないわね
公介：ためしに、横軸に緯度の２乗、縦軸に平均気温をとった散布図を描いてみよう

公介君が描いた、散布図は図４のようになった。

図４　緯度の２乗と平均気温との散布図



先生：この散布図の関係性はどう見えますか
公介：直線関係に近いように見えます。この場合には相関係数を計算しても意味がありそうです。

Q４：緯度の２乗と平均気温の相関係数を求めなさい。

相関係数＝

航平：なぜ、緯度の２乗と平均気温との相関が高いのかな？
理恵：緯度は90度までしかないのに２次関数というのは奇妙よね。緯度100度があったら、もっと気温が下がるということですものね。もう少し、物理的裏付けのある緯度の変換はないかしら？

STEP 5：Conclusion 結論 1

Q５：次の文章の（　）の中に適当な語句を埋めなさい。

都市の平均気温は、緯度の（　）と直線的な関係にあることが分かる。
散布図から、おおよそどのような関係式になるだろうか？
地域の平均気温≒

理恵：直線関係に見える散布図から、一応、緯度の２乗と平均気温との関係式を導きましたが、もう少し数学的な方法はないのでしょうか？
先生：最小２乗法という方法で関係式を導くことができます。**回帰分析**とも呼ばれています。ただ、高校の数学の範囲を超えてしまいます

新たに見つかった課題解決のための、第２巡目の PPDAC サイクルに入る。

STEP 6：Problem 問題発見

◇ 関係性から外れているデータの統計的処理

航平：平均気温と関係のある緯度の２乗変換の散布図を眺めると、他の都市と異なって、直線関係上から少しずれている都市がいくつか見出せます
先生：なぜ、それらの都市が直線関係からずれるのかを考えてみましょう

多くのデータが示す散布図上の関係性とは少しだけ異なるように見えるデータは、**外れ値**と呼ばれる。外れ値が存在するということは、まだ結果の変化を説明する際、考慮されていない原因がデータに影響を与えているということが考えられる。

Q６：図４で外れの程度が大きい都市を外れの程度の大きい順に示しなさい。

外れの程度が大きい都市：

STEP 7：Plan 2

◇ 外れ値が生じている原因の考察

先生：外れている都市には、よく似たところがありませんか？もう一度、図１の特性要因図を眺めなおして、外れ値が生じる原因は何なのか考えてください。

公介：せっかくデータがあるのだから、外れている都市の標高を調べてみよう

Q７：先生と公介君の対話を参考に、外れ値が生じる要因を絞り込みなさい。

STEP 8：Data & Analysis 2

◇ 大きな外れ値を無くすデータの加工と解析

先生：都市の標高による平均気温への影響を調整すると、平均気温と緯度の関係はどうなるだろうか？

航平：表１の都市の平均気温の代わりに、都市の標高が０m だったとしたらどういう平均気温になるかを考えれば良いのでしょう

理恵：そうすると、（平均気温＋0.006×標高）を縦軸にした散布図を描いて、相関係数を求めれば良いのね

Q８：すべての都市の標高が０m だとしたら、その平均気温がどのようになるかを考えて、散布図を描き、相関係数を求めなさい。

相関係数＝

STEP 9：Conclusion 結論 2

Q９：都市の平均気温と緯度、標高の関係はどのようなものか、散布図から読み取りなさい。

都市の平均気温≒

重回帰分析

平均気温と緯度、標高との関係を計算で導く方法には、重回帰分析があります。

単回帰分析は**被説明変数**（目的変数）に対して**説明変数**が１つでしたが、説明変数を２つ以上に増やして回帰分析を行うことができます。これを重回帰分析といいます。

被説明変数を y、説明変数を x_1、x_2とした場合、重回帰分析から、切片 a、回帰係数 b_1、b_2を求めると、重回帰式は次のように表すことができます。

$y = a + b_1x_1 + b_2x_2$

重回帰式から y と x_1、x_2の２つの変数との関係を量的に求めることができるので、x_1と x_2の値から y を予測することができます。

第３巡目の PPDAC サイクルに向けて

上の結論で求めた関係から外れている都市はないだろうか？それは何故だろうか？

航平：依然として、リマやレイキャビクは少し、外れているように見える。リマやレイキャビクの気候について、Wikipedia で調べてみよう

航平君の調べた結果は、次のとおりである。

リマ：南緯12度と低緯度であるが沿岸を北流するペルー海流の影響によって気温は低く、最暖月の２月で22.5℃、最寒月の８月で15℃となる。 https://ja.wikipedia.org/wiki/ リマ より引用

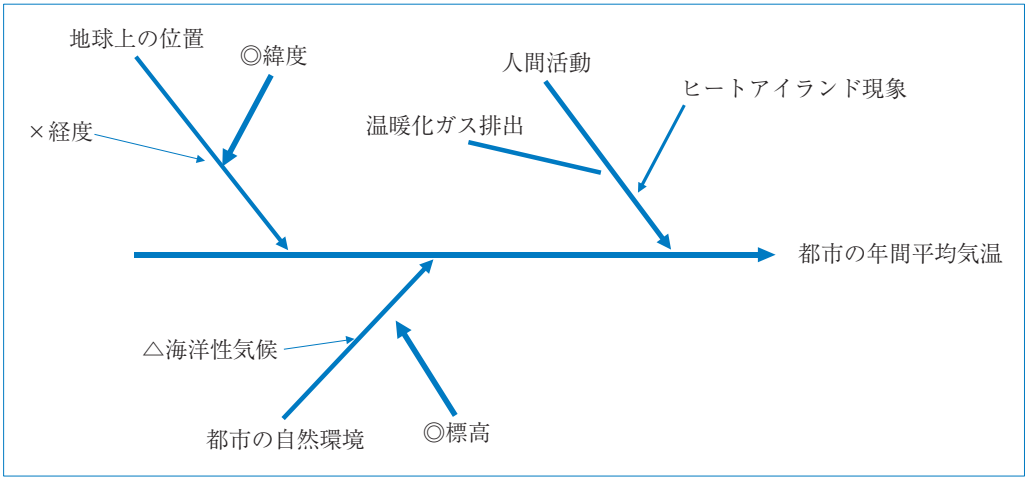
レイキャビク：アラスカのフェアバンクス、東シベリアのヤクーツクといった同緯度の地域と比べ、非常に温暖であることが最大の特徴である。ヤクーツクなど東シベリアでは内陸部中心に真冬ともなると－50℃前後の値を観測する事があるが、レイキャヴィークの最低気温は１年で最も寒い日でも－10℃程度にしかならない。この緯度のわりに温和な気候は、沖合を流れる暖流（北大西洋海流）、南から吹く偏西風に起因している。

https://ja.wikipedia.org/wiki/ レイキャヴィーク より引用

実際、リマは、図４の関係性から、リマは下側に、レイキャビクは上側にずれているように見え、ウィキペディアの記述と整合的である。したがって、図４に生じた外れ値は、海流などの影響と考えることができる。この海流や偏西風の影響の調整は、理科年表や気象庁から収集しているデータではできないので、今後の課題となる。

〔本節の解答〕

Ｑ１：特性要因図の一例



Ｑ２：緯度と平均気温の相関係数＝－0.045

標高と平均気温の相関係数＝－0.107

Q３：

群	平均 絶対緯度	平均気温の 平均	平均気温の 標準偏差
1	5.462	22.66	7.09
2	18.390	23.18	5.15
3	32.128	18.40	2.57
4	40.654	12.84	1.76
5	56.724	4.36	8.55

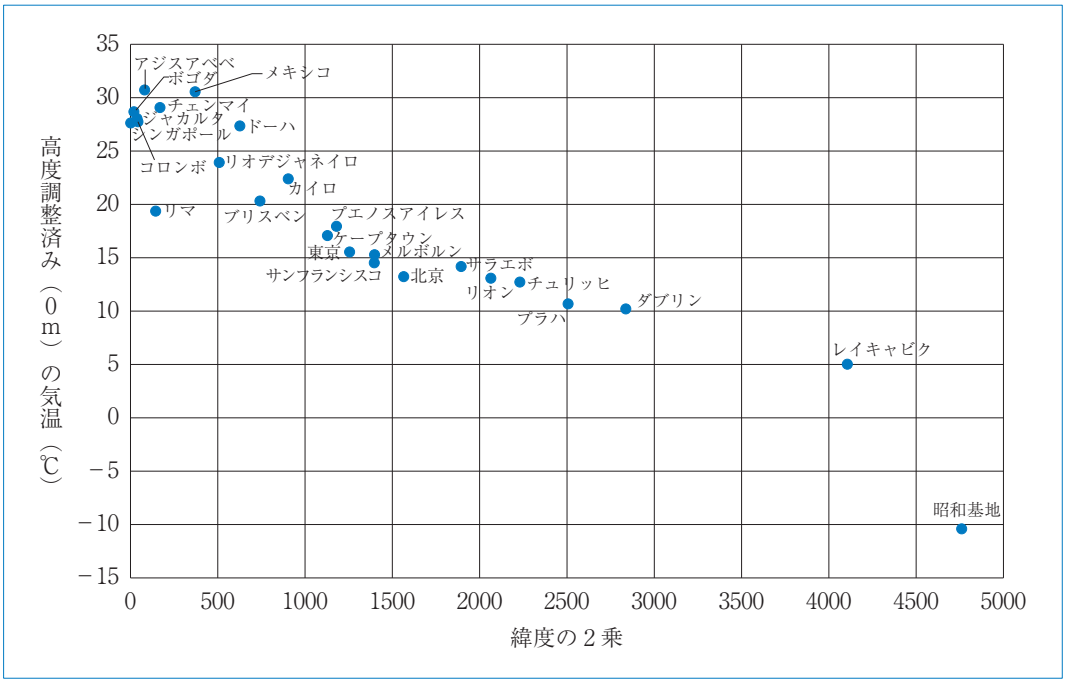
Q４：−0.877（４）２乗（コサインも正解）

Q５：おおよそ　$24-0.006\times$北緯^2、あるいは$-20+44\times\cos$（北緯）

Q６：図４で、他の都市が示す関係性から最も外れているのは、コロンビアの首都のボコダである。次に外れが大きいのは、メキシコの首都のメキシコ、エチオピアの首都のアジスアベバである。リマやレイキャビクもやや外れているように見える。

Q７：外れ具合が大きい都市は、すべて標高が2000m 以上である。このことは、図１で特性要因図に標高を入れることと整合的である。

Q８：



相関係数＝−0.943

Q９：都市の平均気温≒$28-0.007\times$緯度２$-0.006\times$都市の標高
あるいは、都市の平均気温≒$-24+\cos$(緯度)$-0.006\times$都市の標高

閑話休題　母集団と標本

第２部では、分析の対象とする個人、企業、地域など、集団の構成単位のすべてについて、データを利用できる状況を前提としていました。この場合、集団の特徴を度数分布、ヒストグラム、箱ひげ図等のグラフや平均、中央値、分散等の統計量から知ることができました。このような分析手法は、**記述統計学**（descriptive statistics）と呼ばれます。

一方、ある集団の構成単位のすべてに関する情報を得ることができず、集団の一部だけに関する情報を利用して、集団の特徴を推測したいことがあります。このような場合、知りたい（推測したい）集団全体を**母集団**（population）と呼び、取り出した一部を**標本**（sample）と呼びます。標本が母集団を正しく代表すると考えられるなら、その情報を利用することによって母集団に関して何らかの結論を導くことができるでしょう。母集団の特性を知ることを目的として、母集団から標本を選んで（**標本抽出、サンプリング** sampling）、それを統計的に分析し、母集団について推測する手法を、**統計的推測**（statistical inference）と呼びます。

統計的推測に基づいて有効な結論を導くためには、標本が母集団を適切に代表することが必要です。次の事例はそのことを十分に理解させてくれることでしょう。

～ 米国大統領選挙の予測の失敗～

無作為抽出の意義が広く認められるようになった事件があります。1936年の米国大統領選挙の選挙予測の**世論調査**で、出版社のリテラリー・ダイジェスト誌は回収数200万人超の大規模調査結果に基づいて共和党のアルフレッド・ランドンの勝利を予測しました。一方、世論調査会社のギャラップ社はわずか3000人を対象とした調査結果から民主党のフランクリン・ルーズベルトの勝利を予測しました。結果は後者の的中という大番狂わせでした。その理由は、両者の調査方法、特に標本抽出の方法にあります。リテラリー・ダイジェスト誌による調査は、自誌の購読者名簿や自動車・電話保有者名簿から対象を選んだため、結果として比較的高所得者に偏っていた可能性が指摘されています。これに対して、ギャラップ社は、有権者を性、年齢、社会階層、人種等の属性でグループに分け、それぞれのグループから規模に比例した割合で対象を抽出する**割当法**によることで、抽出された標本が母集団により近いものになったといわれています。

ギャラップ社が導入した割当法は、その後の世論調査で広く利用されるようになりましたが、12年後の1948年の大統領選挙では暗転しました。共和党のトマス・デューイが民主党のハリー・トルーマンに勝利するとの世論調査の予想結果はギャラップ社を含めてことごとく外れました。割当法において属性ごとの調査対象者数の割り当てを受けた調査員は、それに合致する人を必要数まで現地で選出する際、どうしても調査しやすい人を選びがちで、それによって調査結果に偏りを生じさせる可能性を残します。

世論調査ではこのような経験を踏まえて、偏りのない調査を行うため恣意的な要素を含まない無作為抽出法の重要性が認識されるようになりました。

記述統計から推測統計へ

標本が与えられたとき、平均やバラツキを計算したり、度数分布を作成する「記述統計」の手法を用いることによって、標本がもっている情報を整理することができます。それでは、母集団の分布や平均などを知るための統計的推測においては、標本の分布や平均をそのまま使うことができるでしょうか。標本が母集団を適切に代表しているのであれば、標本から得られた平均は母集団の平均に近いことが予想されますが、実際に、どの程度近いのでしょうか。別な問題として、ある政策に関する意識に関する世論調査の例では、標本における支持率は分かりますが、母集団における支持率は、標本の支持率とどの程度近いのでしょうか。いずれの例でも客観的な評価の方法が必要となります。

このような問題に答えるための方法の１つは、標本を客観的な基準で選ぶことです。たとえば、テレビなどで政策に関する意見として街頭のインタビューの結果を放送することがありますが、これでは適切な標本とは言えません。インタビューの時間帯にその場所にいる人には事務系の会社員は少ない、社会人より学生・生徒が多い、女性より男性が多い、などの偏りがあります。午前10時から午後４時までの電話調査についても、電話番号をどのように選ぶかの他、日中、自宅にいる人は会社員ではない、有権者に限ると主婦が多いなどの偏りがあります。ある学校で生徒の意識を聞く標本調査の場合に、男子と女子に分けてそれぞれ「代表的」と考える生徒を選んだとしても、それが適切な標本であるかどうかは、選んだ教員以外には判断が困難ですし、他の教員が選ぶ生徒は違う可能性が高いでしょう。このような場合に、どちらの結果の信頼性が高いかについては、客観的な判断の根拠が乏しいことになります。

客観性を保証する最も簡単な方法が無作為抽出と呼ばれるもので、選挙人名簿から同じ確率で有権者の標本を選ぶという方法です。さらに丁寧に調査するなら、性別、年齢階級別に選挙人名簿を分けて、それぞれのグループから無作為に有権者を選ぶ方法もあります。ただし、この場合には、母集団における性別、年齢別の有権者数の情報を利用する必要があるので、手間がかかります。したがって、母集団をその特徴に応じて分割するにしても、ある程度の限界があります。

無作為抽出という方法で標本抽出が適切に行われた場合、標本は、偶然によって選ばれますから、標本の平均や度数分布も確率的に変動します。そこで、標本から母集団に関する推論を行うためには、このような確率的な変動がどの程度の大きさとなるかを明らかにする必要があります。この意味で、確率に関する理解が必要となります。

統計的推測に先立って、第３部で確率の基本的知識、第４部で代表的な確率分布である二項分布、正規分布等を紹介します。