計量社会学 Ⅰ（2014年度春学期 担当:保田（やすだ））
2014.4.11

第1回「導入：なぜ社会を数値にするのか」

■全体的な目標

計量社会学（quantitative sociology）とは、社会を知るために積極的に数値（統計データ）を活用する社会学の一分野である。社会へのアプローチ方法によって分類した呼び方で、理論によるアプローチ（理論社会学）や歴史によるアプローチ（歴史社会学）と対比される。家族や組織、教育など、対象とする社会現象の領域は問わない。

この講義では、ⅠとⅡを合わせて計量社会学の基本的な考え方を使いこなせるようになることをめざす。大きく考えると、Ⅰでは**記述統計**（descriptive statistics）の活用を、Ⅱは**推測統計**（inferential statistics）の活用を学修する。合わせて修得することが望ましいが、一方だけでも理解できるように講義する。

記述統計……データがもつ情報を要約して記述する統計的方法
　例）関大生100人の調査を集計すると、1ヶ月の読書冊数は平均10.2冊だった
推測統計……一部のデータから調べてもいない全体を推し測る統計的方法
　例）関大生100人の調査から、大学全体でバイトをしているのは55～65%と予想される

計量社会学Ⅰの具体的な目標は以下の3点である。

1）基本的な記述統計の数値を算出し、その意味を読み取れるようになる
2）関心に即して、調査データの集計方法を立案できるようになる
3）計量社会学の意義を理解する

ただ単に「○○を算出しなさい」と言われて計算できるのではなく、置かれている状況に応じてどんな数値を整理すべきか自分で考え、他人にその意味を説明できることを求める。

逆に、（Ⅱも含めて）この講義を終えても、以下の点は限界として残ることを了承してほしい。あくまで「考え方」を身につけてもらう。

1）数学的な理解は最小限に留まる
2）逆に、実際的な統計分析ソフトの操作を練習するわけでもない
3）データの集め方（社会調査の方法）については解説しない

※1）については、関心があれば授業外で教える。
　2）については、「社会学研究法a」（2年生以上配当）で、ある程度触れる。
　3）については、「社会調査方法論」「社会調査論」で学べる。
　「社会調査演習」「社会調査実習」（2年生以上配当）では全体を深く経験できる。
　以上の科目＋社会学研究法bが社会調査士資格の取得のために必要な科目（社会学研究法a,bは一応どちらか一方でも可だが、両方の履修を強く奨める）。

■計量社会学の意義

今回は、はじめに「なぜ社会を数値にするのか」、つまり「なぜ社会学に統計を使わなければならないのか」ということについて、簡単に解説する。

大雑把にいえば、社会学に関心のある人々の中で数字を扱うことが好きな人は、そう多くはない（というか、相当に少ない）。皆さんの中には、統計というと難しそうで、自分の手に負えるようなものではない、と感じている人もいるだろう。また、数値で示されるような薄っぺらい内容には興味がもてない、と否定的な印象をもつ人もいるだろう。

にもかかわらず、社会学部の科目として計量分析や統計的調査に関する科目が多く設けられているのはなぜだろうか。そして、その多くが「1年生の配当科目になっている」のはなぜだろうか。それはもちろん役立つからではあるのだが、いろいろな分野で役立つ統計学を、とくに社会学に活用することには「特別な意義」がある。ここでは、次の2つの意義に注目しよう。

・数値を使えば、社会に実態を与えることができる
・数値を使えば、他人と協力できる

これらの意義があるからこそ、自らは理論的考察や質的調査（観察や聞き取りによるフィールドワーク）に取り組む研究者であっても、計量社会学の取り組みを軽視することはない。また、その意義があるからこそ、計量社会学からは、ただの技術を超えた学問的なおもしろさが感じられる（はず）。

■数値で社会に実態を与える

それぞれ、もう少しきちんと説明しよう。社会学はいろいろな現象を扱う学問だが、ともかく「社会」（人間関係の集まり）を対象にしている。ところが、社会を科学的に扱おうとしたとき重大な問題にぶつかる。当たり前のことであるが、社会は目に見えない。科学の基本姿勢は「まず観察し、次に観察された不思議なことを説明すること」であるが、その第一歩である「観察」ができないのである。「いや、私は社会で暮らしている人々を見たり、その人たちから話を聞いたりすることができる」と思う人もいるかもしれないが、そこで見ているのは社会の影響を受けた（あるいは社会を作り出している）人々の様子であって、社会そのものではない。また、聞くことのできる話は、その人が感じている社会のあり方であって、やはり社会そのものではない。

この難しさを克服するために、社会学者は観察可能な情報から理論的に社会のあり方を予想したり、関心のある社会集団に深く関わっている人々の話に深く耳を傾けたり、あるいはその社会の中に自ら飛び込んだり（参与観察）と、実にさまざまな手段でアプローチする。社会学の方法が何でもありになることの一因は、この「社会が観察困難」ということへのチャレンジの結果なのである。

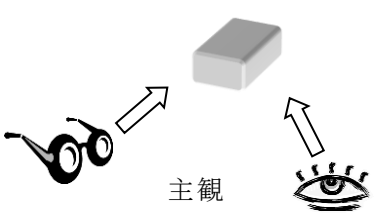
その中で、計量社会学のアプローチは、見えるもの（測定できる個人レベルの情報）を集計すれば、見えない社会も見えるようになるはずだ、というものである。たとえば「日本社会で夫婦別姓に賛成の人は50％です」という統計は、1人ひとりが夫婦別姓に賛成している、あるいは反対している、という観察可能な情報を集めて、「賛成の割合」という社会の数値を作ることで、社会に実態を与えているわけである。

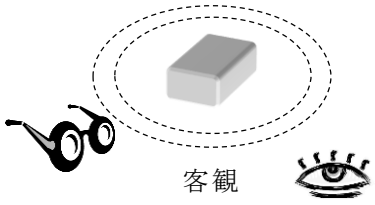
このアプローチがもつとりわけ強力な点は、その社会について誰も知らない新たな事実を「発見できる」ということにある。インタビューの結果は、当事者にとっては自明ですでに知っていることである（一般の人には知れわたっていないかもしれないが）。また、研究者の理論的な考察は、その研究者が頭の中で知っている事実にもとづいている（甚大な苦労の末にたどり着いたものではあるが）。これに対して、数値で表される社会の様子は、ときに、本当に世の中の誰一人として考え及ばなかった意外な事実を教えてくれる。計量社会学者はよく「データに語らせる」という言い方をするが、まさに人工的に実態を与えられた社会が、自分のことをしゃべりだすわけである。この未知の発見が、計量社会学の第一の意義、魅力である。

例）夫婦別姓について「平成24年度 家族の法制に関する世論調査」（内閣府 2012）
　　渡辺（2011）　p.18 事実婚・同棲の割合の国際比較　　p.29 生涯未婚率の推移

■数値にすれば協力できる

数値によって表現された社会は、通常、ほかの手段よりも客観的なものである。客観的であることは何となくよいことと感じられるだろうが、実際には、客観的な情報よりも主観的な助言の方が、人の心を深く打ったり、より役に立ったりすることが多い。そもそも客観性とは何だろうか。**主観**（subjectivity）が観察をする側をメインにしているのに対して、**客観**（objectivity）は観察される側がメインになっている状態を指す。つまり、主観的な観察は見る人によって見え方が違う（それゆえに、より適切な観察に近づける可能性を秘めているともいえる）が、客観的な観察は誰が見ても同じということである。





誰が見ても同じ数値であるという事実は、ひとりよがりではない、といった消極的な利点を超えて非常に重要な意味をもっている。すなわち、誰が行っても同じということは、無限に多くの研究の間で協力することができるということを意味している。1980年代に「新人類」と呼ばれた若者がどのような価値観を持っていたのか数値化した研究があったとする。このとき、同じ方法で現在の若者を数値化すれば、2つの若者社会を時空を超えて比較研究できる。誰が見ても同じであるから、すでにこの世にいない研究者とも協力できる。多様で変化の激しい社会現象を研究する上で、この無限の協力は強い武器となる。

> ※もちろん、実際には「同じ方法で数値化」することが、そんなに容易なわけではないが、その問題は調査法の課題なので、この講義では追求しない。社会科学における客観性の利点と問題点については、竹内（1971）が深く考察している。

例）片桐（2009）の5年おきの学生調査
　　極旨醤油らーめん一刻堂 お客様アンケート

計量社会学のこれらの利点は、当たり前のように感じられるかもしれないが、我々凡人が社会学という難しい課題に立ち向かうためには、極めてありがたい。計量社会学は、捉えがたい社会の姿を直接的に観察することを可能にし、薄っぺらい数値を（他人といっしょに）無数に積み重ねることで重厚な社会認識に地道に近づくことを可能にする。やや長い道のりになるが、計量社会学の考え方を1つでも多く身につけて、その共同作業に参加してほしい。そして、皆さん自身の「社会学」の役に立ててほしい。

今日のポイント
①計量社会学は、研究対象ではなく、アプローチ法による社会学の分類
②数値を使って社会学をすることの意義
・数値を使えば、社会に実態を与えることができる
・数値を使えば、他人と協力できる

■授業の予定

1．導入：なぜ社会を数値にするのか	
2．計量社会学で扱うデータ	
3～4．分布の読み方	（1） 度数分布 （2） 代表値とばらつき
5～7．関係の読み方	（1） 散布図とクロス表 （2） 相関係数 （3） クロス表の連関係数
8～10．記述の実践	（1） PPDACサイクル （2） 比較のプランと作表 （3） グラフの描き方
11～12．因果関係への注意	（1） 相関と因果 （2） 見せかけの関係の追求
13～14．経年変化への注意	（1） 白書と政府統計 （2） 変化の意味
15．まとめ：発見を共有する	
学期末試験	

■事務連絡

・第3回以降、毎回、√の計算できる電卓を持参のこと。
・成績評価について
学期末の試験のみで評価（持ち込み全て可）、出席による加点・減点なし
60点以上で合格（60～69点＝C可、70～79点＝B良、80～89点＝A優、90～100点＝S秀）
ただし、事前の4回の小テストで60％得点していない者は学期末試験を受験できない
小テストは、A4用紙1枚を持ち込み可。最終日には小テストの追試もおこなう
・質問は授業後か、研究室（C605）、メール（tyasuda@zf7.so-net.ne.jp）で
・テキストは用いないが、岩井・保田（2007）などで自学することもできる（と思う）

＜文献＞

岩井紀子・保田時男 2007『調査データ分析の基礎』有斐閣.
片桐新自 2009『不安定社会の中の若者たち：大学生調査から見るこの20年』世界思想社.
竹内啓 2013[1971]『社会科学における数と量 増補新装版』東京大学出版会（とくに第1、2章）.
保田時男 2014「計量社会学の考え方」永井良和・間淵領吾・大和礼子編『基礎社会学 新訂第3版』世界思想社, pp.43-54（4章）.
渡辺淳一 2011『事実婚 新しい愛の形』集英社新書.

第2回「計量社会学で扱うデータ」

■社会学のデータは多様

前回解説したとおり、社会学の対象である「社会」は直接見たり触ったりすることができない。そのため、社会学者はありとあらゆる手段で、社会を知るための根拠、すなわち「データ」を集めようとする。社会学でいうデータには、数値で整理される統計的なデータだけではなく、人々を観察したりインタビューで話を聞いたりした記録や、日記などの歴史的な資料など、幅広いものが含まれる。大量の対象について一定の単純な方法で測定を繰り返して集めるいわゆる統計データのことを、一般に**量的データ**（quantitative data）と呼ぶ。一方、少量の事例について会話や映像、文章やなど比較的自由度の高い方法で集められたデータを**質的データ**（qualitative data）と呼ぶ。

計量社会学では、量的データを分析して利用するが、質的データの重要性も忘れてはならない。大切なことは、困難に立ち向かうためにあらゆる手段を尽くすという姿勢であり、逆に言えば、量的データは使わないという拒絶もあってはならない。

量的データの例

```
001 2 31 3 2 2 2001 4
002 1 29 2 2 2 2000 3
003 1 33 1 1 2 1998 2
004 1 30 2 1 1 2003 4
005 2 28 3 2 1 2003 4
006 2 35 2 1 2 1999 1
007 1 30 1 1 1 2002 1
……
```

質的データの例

2012年10月23日 13:00からのインタビュー
校長「私は子どもが何を求めているのかは突き詰めると大人にはわからないものだと思ってるんですよ。そういうと誤解されるかもしれませんが」
調査者「もう少し詳しくその考えを聞かせてください」
校長「私が言いたいのは子どもの世界には子どもの世界のルールがあって、大人のものとは違う。それを大人が知ろうとしても子どもは明かしてはくれない……

■計量社会学で扱うデータ

次の表は、計量社会学で扱われる典型的な量的データを例示している。ある大学の学生120人について、性別、やる気、家庭学習時間の違いが、ある科目の成績にどのような影響を与えるのかを調べようとしている。1行1行に対して1人1人の学生の情報が対応している。性別、IQ等は、それぞれの生徒がさまざまな値をとるので、データの**変数**（variable）と呼ばれる。それぞれの変数に対して1つの決まった値を持つ単位を**ケース**（case）と呼ぶ。ここでは、1人1人の学生がケースである。それぞれのケースに対して、それぞれの変数の値が記されているものがデータである。通常、社会調査のデータでは、変数は個々の質問項目に対応し、ケースは1人1人の回答者に対応することが多い。

このようなデータを集計して、たとえばクラス別の平均値をまとめたような情報もデータと呼ぶことがある。区別のために、1ケースごとの細かい情報が揃っているデータを**素データ［ローデータ］**（raw data）と呼び、一定のグループで情報をまとめたデータを**集計データ**（aggregate data）と呼ぶ。

	性別	やる気	家庭学習時間	成績
1人目	女	非常に強い	4時間	優
2人目	女	やや強い	5時間30分	秀
3人目	男	やや弱い	2時間	可
4人目	女	やや弱い	4時間	可
⋮	⋮	⋮	⋮	⋮
119人目	女	ふつう	2時間	不可
120人目	男	非常に弱い	4時間30分	良

⇩

	A_i	B_i	C_i	D_i
i=1	2	5	4.0	3
i=2	2	4	5.5	4
i=3	1	2	2.0	1
i=4	2	2	4.0	1
⋮	⋮	⋮	⋮	⋮
i=119	2	3	2.0	0
i=120	1	1	4.5	2

いずれにしても、統計データはまず複数の数値情報でなければならない（dataはdatumの複数形）。たとえば、「山田君の身長は150cm」という情報や「中学2年生男子の平均身長は159.9cm」という集計値は、単独ではデータではない。また、1つのケースについて様々な事柄を調べて多くの数値情報を集めているのではなく、同じ事柄（変数）について、複数のケースから情報を集めていることが重要である。そうでなければ、統計的に扱うことができない。だから、まずデータは縦に長くなければならない。

通常、あらゆるデータは統計学で扱いやすいように、すべて記号と数字に置き換えて扱われる。上の場合、家庭学習時間という変数をCという記号で表した。C_iは特にi番目の学生の家庭学習時間を表し、iに具体的な数値を入れると、それは特定の値を表すようになる。たとえば、C_2は2番目の学生の家庭学習時間を表し、C_2=5.5と書ける。

もともと数字で表されていなかったデータも数字に置き換えて扱われる。たとえば性別A_iは男を1、女を2で表すことにした。同じように成績D_iは｛秀,優,良,可,不可｝をそれぞれ｛4,3,2,1,0｝で表している。

■質的変数と量的変数の区別

このように全ての変数のデータを数字にしてしまうと、全ての変数を同じように扱えるような気分になってしまうが、それは誤りである。ある変数の数字がもともとどのように作られたのかによって、その変数の扱いは変える必要がある。特に、質的変数と量的変数の区別は非常に重要である。**質的変数［カテゴリー変数］**（qualitative variable; categorical variable）とは、数量的な特色がないため計算ができない変数を指す。これに対して、**量的変数**（quantitative variable）は、数量的な計算が可能な変数である。

※テキストによっては、質的変数／量的変数という用語の代わりに、質的データ／量的データという用語を使っている。このような表現は、データといえば統計的なデータに決まっているような（いわゆる理系の）分野を前提とする場合によく使われる。我々にとっては紛らわしいので、この用法は避けた方がよい。

たとえば、先のデータでは性別や成績は質的変数であり、家庭学習時間は量的変数である。成績は量的変数じゃないのか、と思うかもしれないが、不可が可になること（0→1）と可が良になること（1→2）は、どちらも差が1であるが、全然意味が違うので数量として計算は成り立っていない。ということは、本来、成績の平均値を出すようなことはでない。統計的な視点からは、推薦入試の「評定平均4.0以上」とか「GPA3.2」という計算は不適切である。この計算が適切になるような成績の付け方をしているという前提が必要になる。

質的変数と量的変数の区別は、どのような統計的分析が可能かを決定する重要な別れ目である。当然のことながら、ふつうは計算ができる方が分析しやすい。質的変数と量的変数をしっかりと区別して、可能であれば質的変数ではなく量的変数にすることができないか考えることが重要である。データの集め方を変更して量的変数にできないか、あるいは集めた後でデータを加工して量的変数を作り出すことはできないか、という発想が必要になる。

ところで、もう1つの変数「やる気」が質的変数か量的変数かはやや大切な問題なので、授業の最後に改めて考える。「やる気」のように、5段階や4段階で意見や意識の強さを測る尺度をとくに評定尺度（rating scale）と呼ぶ。（例：5 非常に賛成、4 賛成、3 どちらともいえない、2 反対、1 非常に反対）

■測定尺度

ある変数が質的変数か量的変数かは、その変数の数値がどのようなものさしで測定されたものであるかによって判断される。もう少し細かくこの辺りの事情を見てみよう。

スティーブンス（Stanley S. Stevens）は1946年に測定のものさし、つまり**測定尺度**（measurement scale）の水準を名義、順序、間隔、比率の4段階に分類することを提案しているが、現在もこの考え方は有効である。一般に、名義、順序尺度により測定された変数を質的変数、間隔、比率尺度により測定された変数を量的変数と呼ぶ（この辺りのことは多くの入門書に記されているが、小田（2009）や轟・杉野（2010）などがわかりやすい）。

測定尺度の4つの水準

名義尺度（nominal scale）	数字は名札替わりの記号として使っているだけで、まったく計算はできない変数 （例：性別、学科、職業）
順序尺度（ordinal scale）	1より2が大きいなど、数字の順序・大小関係には意味があるが、実際的にはほとんど計算のできない変数 （例：学年内の成績順位、）
間隔尺度（interval scale）	数字の間隔（差）が同じなら同じ数量とみなせるので、平均を出すなど、ほとんどの計算ができる変数 （例：気温、5点満点の意識評定）
比率尺度[比例尺度] （ratio scale）	数字が2倍なら、数量も2倍とみなせるので、どんな計算でもできる変数 （例：体重、年収、通勤時間）

※測定尺度の違いは、かなりの程度、絶対的な基準により判断される。しかし、測定尺度の水準が必ずしもはっきりしない場合もあるので注意は必要である（例：教育年数）。

質的変数と量的変数の区別は最も基礎的な区別として重要であるが、ある変数に対してある統計的な手続きを当てはめることができるかどうかを、より細かく判断するためには、4つの測定尺度の違いに注意しなければならない。

■離散変数と連続変数

量的変数は、測定尺度とは別の視点から**離散変数**（discrete variable）と**連続変数**（continuous variable）に分類できる。離散変数とは、取りうる値がいくつかの点で定まっており、間の値を取りえない変数である。たとえば、家族の人数は、3.5人のような値は取りえないので、離散変数である。これに対して、連続変数は理論上、無限に細かい測定が可能である。たとえば、家の広さ（㎡）は連続変数である。家族の人数も家の広さも、測定尺度の視点からは、比率尺度による量的変数で変わりはない。

離散変数と連続変数の区別は、当面取り組むデータの整理・要約の視点からはあまり重要でないが、確率論との結びつきを考える際には重要となるので、概念としては覚えておこう。

今日のポイント
①計量社会学で扱う量的データ（統計データ）は、同じ変数について、多くのケースから情報を集めて積み重ねたもの
②計算できる「量的変数」と計算できない「質的変数」の区別は重要
（より細かくは、測定尺度の4段階［名義・順序・間隔・比率］にも注意）

（問題）

1. 次のような変数は、名義・順序・間隔・比率のどの尺度で測られた変数だろう？
（1）4年間の取得単位数
（2）好きなスポーツ選手（1＝イチロー、2＝浅田真央、3＝……）
（3）オリンピックでの国別メダル獲得数の順位（1位＝アメリカ、2位＝ロシア、……）
（4）西暦○○年生まれ

2. 評定尺度を順序尺度とみなすか、間隔尺度とみなすかは、社会調査のデータ分析では非常に重大な問題である。どちらで考えるべきか、自分の意見をまとめてみよう。

＜文献＞

小田利勝 2009 『社会調査法の基礎』 プレアデス出版.
轟亮・杉野勇編 2010 『入門・社会調査法: 2ステップで基礎から学ぶ』 法律文化社.

※過去の配付資料はwebに置いています。欠席時は各自で補充を。
http://www2.itc.kansai-u.ac.jp/~tyasuda/

第3回「分布の読み方（1） 度数分布と代表値」

■度数分布表

調査データの分析の第一歩は通常、それぞれの変数に対してそれぞれの値を取るケースの数、つまり**度数**（frequency）を数えることから始まる。非常に単純な作業であるが、ある側面から見てどのような人々が何人いるかという度数分布は、その社会の姿をもっとも端的に表しておりばかにできない。

表1 2014年度計量社会学 I 履修者の「数字の好き嫌い」

	度数	%
1 大嫌い	6	7.2
2 まあ嫌い	24	28.9
3 ふつう	23	27.7
4 まあ好き	27	32.5
5 大好き	3	3.6
計	83	100.0

それぞれの変数の集計結果を上のような**度数分布表**（frequency distribution table）にまとめておくと、分布状態が大まかに分かるので、便利である（表1）。度数分布表では、人数そのもの（度数）に加えてパーセント（%）などを示すことがよくある。%は全体を100人に統一した場合の相対的な人数を示すので、**相対度数**（relative frequency）と呼ばれる。犯罪被害率など出現頻度の低い現象についは、1000人あたりの人数（パーミル‰）や10万人あたりの人数など、全体を100にしない相対度数も用いられる。相対度数は必要に応じて付け加えたり省いたりしてもかまわないが、あくまで調査結果の基本は度数だ、ということを忘れてはならない。たとえば同じ相対度数50%でも、600人中300人の場合と4人中2人の場合では結果の読み取りが当然異なる。だから、基本となる度数が不明になるような表（%のみの表）は、通常作成してはならない。少なくとも全体のケース数は明記しなければならない。全体の人数は「n＝103」のように、「n」で表記する約束になっている。

■取りうる値が多い場合の度数分布表の作り方

上の例のように、扱う変数で選択肢の限られている場合には、そのままそれぞれの値ごとにケース数を数えればよい。しかし、取りうる値の数が多い場合には、全ての値について度数分布表を作っても、ほとんど役に立たない（例：身長142.6cm 1人、142.7cm 1人、142.8cm 2人、……）。一定の範囲の**階級**（class）を作成し、各階級の範囲に入る回答の数を数えるのが一般的である。

それぞれの階級について、級間の中心の値を**階級値［中心点］**（midpoint）と呼ぶ。中心点を示しておくとグラフを作成する際や、平均などの統計値を計算する際に便利である。

表2　通勤時間の度数分布表（第2回全国家族調査 NFRJ03若年データ）

	中心点	度数	%
7分以下	－	344	13.6
約15分（8～22分）	15	636	25.2
約30分（23～37分）	30	319	12.6
約45分（38～52分）	45	177	7.0
約60分（53～67分）	60	182	7.2
約75分（68～82分）	75	54	2.1
約90分（83～97分）	90	49	1.9
98分以上	－	28	1.1
計		1789	70.9

階級の幅を自分で設定するのは意外と難しい。厳密な規則はないが、次の3点くらいに注意しながら、5～10個程度の階級にわけることが原則である。

1）全てのケースがいずれか1つの階級に収まるように、階級幅は互いに排他的（exclusive）で、全体として包括的（exhaustive）に定めなければならない。2つの階級にまたがらないように、「以上」「未満」を用いるなどする。
2）それぞれの階級幅は等しくする。幅が異なると、分布が把握しにくい。ただし、一番上や一番下の階級の幅は等しくできないことが多い。
3）キリのよい数値の扱いには注意する。社会調査のデータでは、例えば通勤時間の分布が「15分」「30分」などキリのよい値に集中することがあるので、階級をキリのよい数値で区切ると分布が歪んで表れることがある（表2）。

■集計結果の図示

集計の結果は表ではなく、図（グラフ）で表した方が、分布の状態がよく分かることがある。グラフを作成すると見栄えがよくなることが多いが、見栄えをよくすることが作図の目的ではない。比較したい統計量を視覚情報に置き換えることで、直感的な判断ができるようにすることが目的なので、次の2点をはっきりと意識する必要がある。1）どんな統計量を比較しようとしているのか。2）比較のためにどんな視覚情報を利用しているのか。

逆の言い方をすれば、1）比較したいもの以外の余分な情報は排除する（例えば、2次元で表現できる図を立体化するなど、余分な情報を加えない）、2）錯覚による誤解を誘う表現をしない、といったことが重要になる。

例えば、代表的な5種類のグラフの特徴は表3のようにまとめられる。グラフ作成の詳細は、後の回で改めて触れる。

表3　代表的なグラフのポイント

	比較の対象	利用する視覚情報
棒グラフ	ある数量の大きさ	棒の長さ
折れ線グラフ	ある数量の連続的な変化	線の傾き
円グラフ	全体に占める構成比	パイの面積
帯グラフ	グループ別の構成比	帯の面積
ヒストグラム	連続した階級の度数	柱の面積

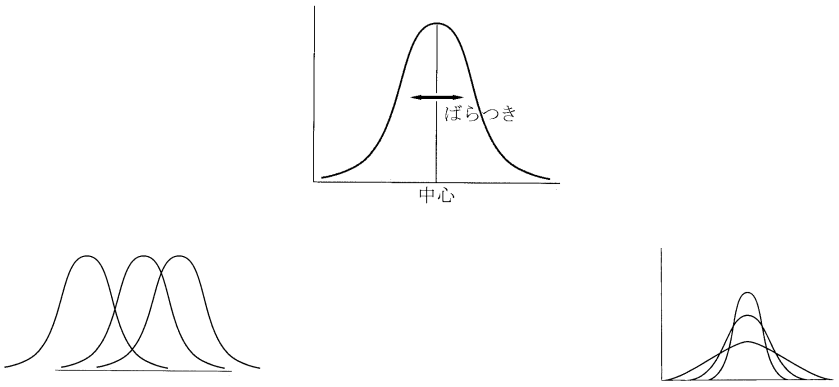
■**基本統計量**

度数分布表は、データのおおまかな分布を知るために作成するものであった。いろいろなデータの度数分布表を作ってみれば分かることであるが、多くの量的変数は、どこかの点を中心にして多くの度数が分布し、中心から離れるとだんだん度数が少なくなるという形で分布する。したがって、

1）中心がどの辺りにあるのか
2）中心からどの程度ばらついているのか

さえ数値で表せば、度数分布表を作成する手間をかけることなく、およその分布を把握できる（図1)。

中心を表現する一連の統計量を**代表値［中心傾向］**(average; measure of central tendency)、ばらついている程度を表現する一連の統計量を**ばらつき［散らばり、散布度］**（variability; measure of dispersion）と呼ぶ。代表値とばらつきはまとめて**基本統計量［要約統計量、記述統計量］**(basic statistics; summary statistics; descriptive statistics）などと呼ばれる。代表値もばらつきも、具体的な計算方法（統計量）は複数のやり方がある。



ばらつきは同じで、中心傾向の異なる分布　　中心傾向は同じで、ばらつきの異なる分布

図1　代表値とばらつき

■**さまざまな代表値**

今回は代表値についてのみ解説する（ばらつきについては次回)。代表値としては、以下の3つがよく使用される。データの分布がきれいに左右対称である場合には、これらはいずれも同じ値を取る。しかし、実際の分布には多かれ少なかれ歪みがあるので、これらの3つの代表値は異なった値になる。代表値の種類によって、捉えることのできる特性が異なるので、場合によって使い分けなければならない。

最頻値（mode）…… もっとも度数の多い測定値または階級

中央値（median）…… 測定値を大きさの順に並べたとき真ん中番目にくる値
（ケース数が偶数のときは $\frac{n}{2}$ 番目と $\frac{n}{2}+1$ 番目の数値の平均）

平均値（mean）…… $\bar{x}=\frac{1}{n}\sum x_i$

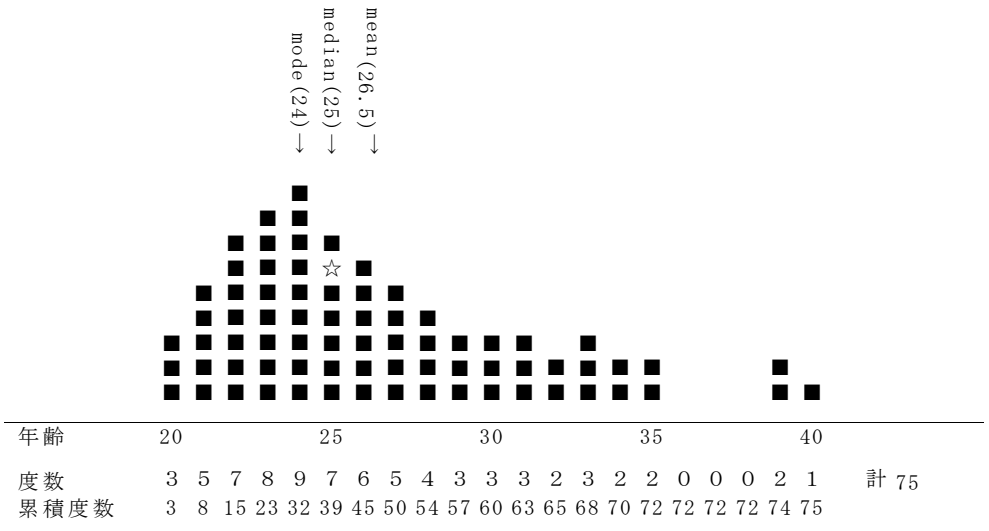


図2　代表値の模式図

もっともよく用いられる代表値は平均値であり、数学的に非常に扱いやすい。ただし、平均は**はずれ値**（outlier）の影響を受けやすい（図2)。中央値ははずれ値の影響を受けにくく、情報が完結していない場合でも算出できる（例：半数が死亡した時点で寿命の中央値は確定する)。しかし、それは逆にデータの全情報を代表していないとも言える。最頻値は他のカテゴリーの分布について情報が全く繁栄されていないが、一方で「多数を占めるものが中心」という日常的な代表性感覚に見合う。

また、測定尺度の水準によって、用いることのできる代表値の限界があることにも、注意が必要である。たとえば、中央値は順序尺度でも算出できるが、平均値は数値の間隔が一定でなければ意味がないので、間隔尺度か比率尺度でなければ算出できない。それぞれの意味と限界を正確に理解して、用いる代表値を選ぶことが肝要である。

今日のポイント
①調査データの分析は、まず各変数の度数分布をよく観察すること
　度数分布表の基本ルールは守ろう（nの提示、階級の区切り方）
②度数分布の概要は、基本統計量（代表値とばらつき）で示せる
③代表値の種類（平均値、中央値、最頻値）は、長所と短所を考えて使い分ける

（問題）

1. バイト時給のデータ｛820,900,850,1100,2300,870｝について、平均値と中央値を示そう（すべて1ケースずつなので、最頻値は出せない)。
2. 表1のデータを間隔尺度とみなして、平均値、中央値、最頻値を示そう。
3. 結婚年齢の平均値の代わりに、中央値や最頻値を大きく報道すれば、人々の結婚行動に何らかの社会的影響があるだろうか（あるいは、ないだろうか)。自分の予想を論じてみよう。

第4回「分布の読み方（2） ばらつき」

■さまざまなばらつき

基本統計量は、代表値とばらつきという2つの数値で、度数分布のおおまかな状態を表現するものであった。今回は、分布の裾野がどの程度広がっているのか、つまり分布のばらつきの程度を示す統計量について解説する。量的変数のばらつきの指標としては、一般に次の5つがよく用いられる。

範囲 R＝最大値－最小値

四分領域 $Q=\frac{Q_3-Q_1}{2}$

分散 $s^2=\frac{1}{n}\sum\left(x_i-\bar{x}\right)^2$ ただし、一般には不偏分散 $s^2=\frac{1}{n-1}\sum\left(x_i-\bar{x}\right)^2$ を用いる

標準偏差 $s=\sqrt{s^2}$

変動係数 $C.V.=\frac{s}{\bar{x}}$

■範囲

範囲（range）の意味はすぐ分かるであろう。最大値と最小値の間の幅は、もっとも直感的にデータのばらつきの程度を示している。たとえば、「先月、何日アルバイトをしたか」という学生調査で下のようなデータAが得られたとすると、範囲R＝21－5＝16である。

データA｛5,8,12,19,21｝ （単位：日）

一方、下のデータBであれば、範囲R＝24－2＝22で、こちらの方がアルバイト日数のばらつきが大きいことを1つの数値でわかる。ちなみに、どちらのデータも平均値は13.0、中央値は12である。2つのデータは分布の中心が同じで、ばらつき具合だけが異なる。

データB｛2,7,12,20,24｝ （単位：日）

範囲はもっとも単純なばらつきの指標なので、もっとも単純な代表値である最頻値とセットで用いられることが多い。代表値とばらつきの種類の中で何を用いるかは、基本的に図1のような対応がある。長所と短所も、対応する代表値と同様と考えてよい。

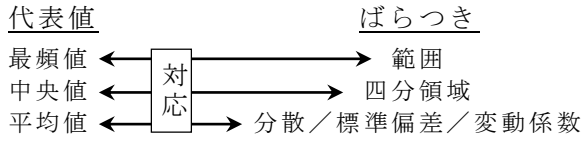
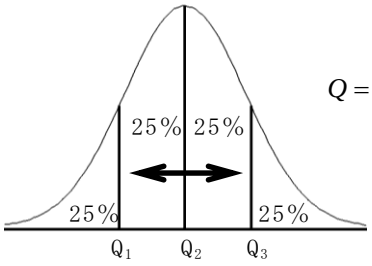


図1 代表値とばらつきの指標の対応

■四分領域

中央値とセットで用いられるのは**四分領域**（quartile range; semi-inter-quartile range）である。中央値は分布全体を二等分する点であるが、全体を4等分する3つの点を**四分位数**（quartile）と呼び、小さい方から第1四分位数（Q_1）、第2四分位数（Q_2）、第3四分位数（Q_3）と呼ぶ。25パーセンタイル点、50パーセンタイル点、……も同じ意味である（図2）。

四分領域は、全体の分布をケース数で4等分に分割した場合に、1番目の区切り点である第1四分位数（Q_1）と3番目の区切り点である第3四分位数（Q_3）との間の幅を2で割ったものである。つまり、中央値（第2四分位数）を中心と考えた場合に、中心からどの程度離れれば、分布の端までの半分に至るかということで、中心からの標準的なばらつきの程度を表している。



$$Q = \frac{Q_3 - Q_1}{2}$$

Q_1……第1四分位数＝25パーセンタイル点
Q_2……第2四分位数＝50パーセンタイル点＝中央値
Q_3……第3四分位数＝75パーセンタイル点

図2　四分位数と四分領域

※四分領域と同じものを四分位偏差と呼んだり、四分偏差と呼んだりすることもある。また、Q_3-Q_1を2で割らない値を四分位範囲（inter-quartile range）という指標で用いることもある。quartile関連の用語、訳語はやや混乱しがちなので注意しよう。

（問題1）

2009年の第3回全国家族調査（NFRJ08）のデータを使って、働いている40歳の人々の通勤時間を男女で比較してみた（自営を除く）。その結果は、以下のとおりである。

	男性	女性
ケース数（*n*）	44	36
平均値	28.7分	17.3分
中央値	20分	15分
最頻値	20分	10分
最小値	3分	0分
最大値	90分	45分
第1四分位数（Q_1）	15分	10分
第2四分位数（Q_2）	20分	15分
第3四分位数（Q_3）	40分	25分
分散	475.7	148.8
標準偏差	21.8	12.2

(1) 男女別に通勤時間の「範囲」を求めてみよう。

(2)男女別に通勤時間の「四分領域」を求めてみよう。

(3)これらの数値で男女の通勤時間についてどのような違いがわかるのか。「範囲」や「四分領域」という用語を知らない人に説明してみよう。

■分散・標準偏差・変動係数

残りのばらつきの指標である分散、標準偏差、変動係数は一連のものである。平均を中心と考えると、各ケースのばらつきは平均との偏差$x_i-\bar{x}$で表せる。ばらつきの大きさを示す上で、偏差の正負には意味がないので、偏差を2乗して符号を消してやる。その上で全ケースを合計すれば、全体的なばらつきの量が1つの数字になる。この合計を全体のケース数nで割って平均化した値が**分散**（variance）である。ただし、一般にはnの代わりに$n-1$で割ることが多い（特に区別する場合には、$n-1$で割る方を不偏分散と呼ぶ）。$n-1$で割る理由は全く数学的な都合のためである。現時点でその理由を理解する必要はない。実際的には、扱うケース数が大きければ、nで割る結果とn-1で割る結果はほとんど変わらない。

上の5ケースのデータAでは、平均$\bar{x}=13.0$なので、

$$\text{不偏分散}s^2=\frac{\sum(x_i-\bar{x})^2}{n-1}=\frac{(5-13)^2+(8-13)^2+(12-13)^2+(19-13)^2+(21-13)^2}{5-1}=47.5$$

と計算できる。

同じようにデータBについて不偏分散を計算すると（やはり平均$\bar{x}=13.0$なので）、

$$\text{不偏分散}s^2=\frac{\sum(x_i-\bar{x})^2}{n-1}=\frac{(2-13)^2+(7-13)^2+(12-13)^2+(20-13)^2+(24-13)^2}{5-1}=82.0$$

となる。データBの方がばらつきの程度が大きいことが数値に反映されている(82.0＞47.5)。

ただし、分散は計算の過程で単位も2乗されているので、数値の大きさが具体的に何を意味するのかわかりにくい（データAの分散は「82.0日^2」など）。そこで分散の正の平方根を取ることで単位を戻してわかりやすくしたものが**標準偏差**（standard deviation; SD）である。たとえばデータAの標準偏差は$s=\sqrt{s^2}=\sqrt{47.5}\fallingdotseq6.89$と計算できる。このとき、単位は「6.89日」ととなり、標準的には平均値±標準偏差、つまり13.0±6.89日（6.11～19.89日）くらいの間に多くの人々がばらついていることが、具体的にわかる。標準偏差はもっともポピュラーに用いられるばらつきの指標である。

感覚的にはわかりやすい標準偏差も、目的によっては欠点を持っている。例えば、幼稚園児の身長の標準偏差が4.5cmで、20歳の成人の身長の標準偏差が5.0cmであったとする。この場合、絶対的な量としては成人の方が身長のばらつきが大きい。しかし、幼稚園児は成人よりもはるかに平均身長が低いにもかかわらず、4.5cmもの標準偏差を示しており、相対的には、成人よりもむしろ大きくばらついている。このようなときに用いるのが**変動係数**（coefficient of variation）である。変動係数は平均的な規模の違いを相殺するために、標準偏差を平均値で割った値を用いる。仮にいまの例で幼稚園児の平均身長が100cm、成人の平均身長が165cmであったとすると、それぞれの変動係数は、4.5÷100≒0.045、5.0÷165≒0.030...と算出され、幼稚園児の方が相対的にはばらつきが大きいことが示される。これらの数値はつまり、幼稚園児は平均身長の4.5%程度の幅でばらついているのに対して、成人は平均身長の3.0%程度の幅でしかばらついていない、という意味である。

■Σの計算

分散などの計算では、記号「Σ」（シグマ）が用いられる。Σはアルファベットの「S」に当たるギリシャ文字で、「合計」を表す英単語「sum」の頭文字を示している。その由来から分かるように、Σの意味は「計算結果を合計する」という意味で、統計学ではほとん

ど1つの使い方しかしない。すなわち、「すべてのケースについて同じ計算を行い、その結果を全員について合計する」という意味である。この用い方しかしないので、Σの上下の表記は通常、省略される。

例）

	A_i	B_i	C_i	D_i
i=1	2	105	4	4
i=2	2	110	5.5	5
i=3	1	97	2	3
i=4	2	115	4	3

(n=4)

$$\sum(C_i+5) = \{(4+5)+(5.5+5)+(2+5)+(4+5)\} = 35.5$$

$$\frac{1}{n}\sum B_i = \frac{1}{4}(105 + 110 + 97 + 115) = 106.75$$

$$\frac{1}{n}\sum(B_i-100)^2 = \frac{1}{4}\left\{(105-100)^2+(110-100)^2+(97-100)^2+(115-100)^2\right\}$$
$$= \frac{1}{4}(25+100+9+225) = 89.75$$

Σを用いた分散の計算式がしっくりこない場合には､「すべてのケースについて同じ計算をする」という過程を下のように表にしてみるとよい。

	x	$(x_i-\bar{x})^2$
1 人目	25	
2 人目	29	
3 人目	32	
4 人目	25	
5 人目	21	

合計 ⇩

$\Sigma(x_i-\bar{x})^2=$ □ ÷(n−1) ⇨ 不偏分散 $s^2=$ □

（問題2）

上のデータ｛25, 29, 32, 25, 21｝は、ある調査で5人の女性に理想の結婚年齢を尋ねた結果である。

(1) 平均値と中央値を算出しよう（復習）。

(2) 表を使って、分散（不偏分散の方）を計算してみよう。

(3) 標準偏差を算出してみよう。

(4)算出した標準偏差をデータと照らし合わせて､計算がおよそ間違いないか確認しよう。

（問題3）

「問題1」（40歳の男女別の通勤時間）の表を参照。

(1) 男女別に、通勤時間の変動係数を算出しよう。

(2) 変動係数は比率尺度の変数にしか使えない（間隔尺度の変数ではダメ）。なぜか。

(3) 表に示されている統計量や、これまでに算出したばらつきの統計量から、男女の通勤時間の分布を、およそのグラフで描いてみよう。

(4) 40歳の男女で、なぜこのような通勤時間の違いが出るのか、その社会的な理由を予想してみよう。

■歪度と尖度

代表値やばらつきと比べれば重要度は落ちるものの、分布を要約する指標として歪度と尖度という統計量が存在する。実際に数値を計算することはあまりないが、分布の形について議論をするために知っておかなければいけない概念である。

歪度（skewness）は、分布の形がどの程度左右対称に近いか（あるいは左右対称からかけ離れているか）を示す統計量である。左右対称の場合には0となる。右（値が大きい方）に裾を引いている場合には正の値となり、左（値が小さい方）に裾を引いている場合には負の値となる。

尖度（kurtosis）は分布の尖り度合いを表す統計量である。きれいなベル型の正規分布の場合には、尖度はちょうど3となる。より尖っている場合には3より大きな値を取り、尖り方が緩やかな場合には3より小さな値をとる。正規分布を基準として考えるために、初めから3を引いた値を尖度として表すこともある。

$$\text{歪度} = \frac{\frac{1}{n}\sum (x_i - \bar{x})^3}{s^3} \qquad \text{尖度} = \frac{\frac{1}{n}\sum (x_i - \bar{x})^4}{s^4}$$

今日のポイント
①ばらつきの各指標は、それぞれ代表値の種類と対応している。
②ばらつきの各指標は、それぞれ計算できるようになっておこう（とくに標準偏差）。
③基本統計量の数値から、具体的な分布の形が想像できるようになろう。

※次回（5／9）の授業初めに 1 回目の小テスト

小テストは、A4 用紙 1 枚を持ち込み可。

第 1～4 回の内容について、基本統計量の計算や語句の意味などを確認。

√が計算できる電卓必須。小テストでは携帯電話の電卓機能でもよい（学期末試験では不可）。

第5回「関係の読み方（1） 散布図とクロス表」

■変数間の関係を読む

これまで、度数分布表や基本統計量の解説においては、1つの変数の分布について考えることを前提に話を進めてきた。しかし、社会学的に意味のあるデータの読み取りをするには、2つ以上の変数の分布を同時に観察し、その関係性を捉えることが有効であることが多い。2つ以上の変数を同時に考慮するもっとも基本的な方法は、**クロス集計表［クロス表、分割表］**（cross tabulation; cross table; contingency table）を作成することである。

クロス表は非常によく目にするもので、基本的な作り方も簡単である。例えば、次のような質問によって捉えられる「三世代同居への賛否」が、「性別」によってどう異なるのか、に関心を持っているとしよう。

問 あなたは一般に、三世代同居（親・子・孫の同居）は望ましいことだと考えますか。

1 望ましい　　2 望ましくない

この場合、下のような「性別」と「三世代同居への賛否」のクロス表を作成する（表1）。条件が交差（クロス）したマスの中にそれぞれの度数を書き入れるので、クロス表と呼ばれる。クロス表の1つ1つのマスは**セル**（cell）と呼ぶ。例えば、左上のセルの「927」という数値は「男性」で、かつ三世代同居に「賛成」というケースが927人いたことを示す。通常は周りに合計の人数を書き入れるが、この部分を**周辺度数**（marginal frequency）と呼ぶ。周辺度数は場合によっては省略する。

表1 男女別の三世代同居への賛否

	賛成	反対	計
男性	927	366	1293
女性	950	600	1550
計	1877	966	2843

注：JGSS-2000のデータから作成

表1のクロス表をよく見れば、「男性の方が三世代同居に賛成しやすく、女性の方が反対しやすい」という傾向がわかるはずである。つまり、性別と三世代同居の賛否は無関係ではなく、2つの変数には関係がある。ここで、「男性も女性も、反対より賛成の方が多いのだから性別は関係なかった」と読んではならない。統計的な社会調査データは、常に相対的な視点から読み取る。つまり、「比較的○○だ」という読み方が重視される。男性では反対よりも賛成が約2.5倍もいるのに対して、女性では約1.5倍しかいない。女性の方が相対的に賛成しにくい（反対しやすい）という関係は明らかである。

計量社会学でこのような相対的な見方が重視されるのは、調べている変数の分布に絶対

的な意味がないことが多いためである。たとえば、全体的に見ると三世代同居に賛成している人は反対の2倍くらいいるが、この結果から「*日本人は三世代同居を支持　反対の2倍！*」といった見出しの新聞記事を書くことはおかしい。なぜならば、これは「三世代同居は望ましいことだと考えますか」という聞き方をしたらそうなっただけで、「三世代同居はすばらしいと思いますか」とか、「三世代同居を積極的に支持しますか」といった別の聞き方で基準が変われば、簡単に数値が違ってくるからである（おそらく賛成が減る）。一方で、聞き方によって基準が変わっても、「男性の方が女性よりも三世代同居に賛成である」という関係性には、違いが出ないはずである。

■3つのパーセント

さて、いまの例の場合はかなり男女の違いがはっきりしていたが、もう少し微妙な傾向を即座に判断したいときには、やはり相対度数（%）を併記することが望ましい。ただし、クロス集計表には、%の算出の仕方が複数ありうる。1行1行を100%としたときの相対度数である**行%**（row percent）、1列1列を100%としたときの**列%**（column percent）、全体を100%としたときの**全体%**（total percent）の3つである（図1）。

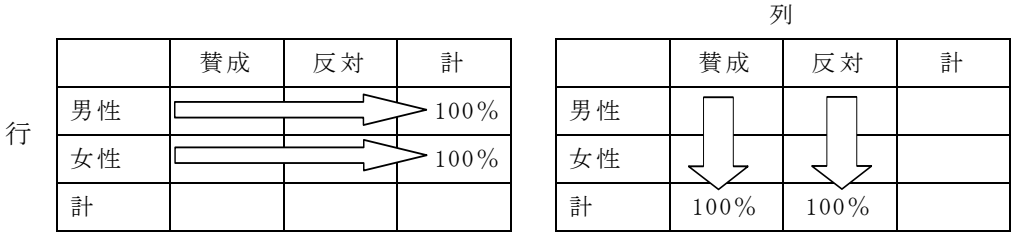


図1　行%と列%

3つの%をすべて併記してクロス表を作ってみると、下のようになる（表2）。

表2　3種類の%付きのクロス表

		三世代同居への賛否		
		賛成	反対	計
男性	度数	927	366	1293
	行%	71.7	28.3	100.0
	列%	49.4	37.9	45.5
	全体%	32.6	12.9	45.5
女性	度数	950	600	1550
	行%	61.3	38.7	100.0
	列%	50.6	62.1	54.5
	全体%	33.4	21.1	54.5
計	度数	1877	966	2843
	行%	66.0	34.0	100.0
	列%	100.0	100.0	100.0
	全体%	66.0	34.0	100.0

しかし、実際にはこのようなクロス表は作成しない。3種類の％の意味を考えて、必要とされるものだけを残し、不要なものは省くべきである。このクロス表の場合、それぞれの％は以下の情報を表している。

行 ％：男性の中での賛否の分布と、女性の中での賛否の分布を比べる

列 ％：賛成の人の中での男女の分布と、反対の人の中での男女の分布を比べる

全体％：全回答者の中での性別と賛否の組み合わせの分布（各割合を比べる）

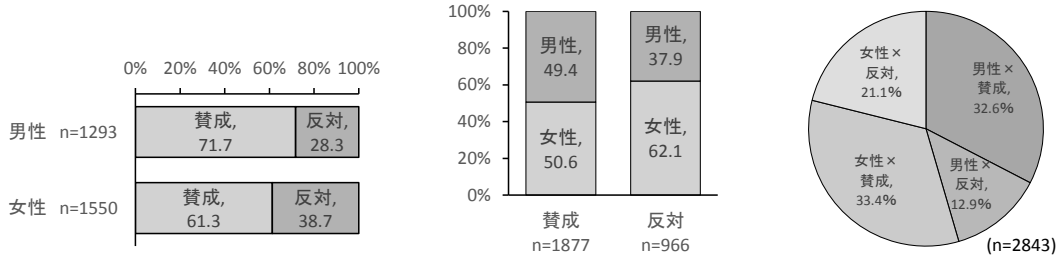
いまここでクロス表を作っている目的を思い出してみると、三世代同居への賛否の分布が男女でどう違っているのかを確かめることであった。つまり、男性の中での賛否の分布と女性の中での賛否の分布を比較して違いを見つけたいわけである。すると当然、必要な％の種類は行％であり、それ以外の列％、合計％は不要である。結局、例えば次のような形でクロス表を作成することが適切ということになる（表3)。

表3　男女別の三世代同居への賛否

	賛成	反対	計
男性	927 (71.7%)	366 (28.3%)	1293 (100%)
女性	950 (61.3%)	600 (38.7%)	1550 (100%)
計	1877 (66.0%)	966 (34.0%)	2843 (100%)

注：JGSS-2000のデータから作成

どの％が適切かピンときにくい場合は、その％からできあがるグラフを考えてみるとわかりやすい。この場合、図2のように比べてみると、行％のグラフこそ知りたい情報であることが理解できるのではないだろうか。



（a）行％の場合　　（b）列％の場合　　（c）全体％の場合

図2　それぞれの％に対応するグラフ表現

なお、一般的には、列％ではなく行％を書き入れるように想定して、2つの変数を配置する方がわかりやすいクロス表になる。つまり、分布に関心のある中心的な変数（ここでは三世代同居への賛否）を列側に配置して、グループ分けのための変数（ここでは性別）を

行側に配置し、行%を比較することで関係性を読み取る。レイアウトなどの都合で特別な事情がない限り、この配置の方が自然に数値を読み取ることができる。言い方を変えれば、最終的に大事な「結果」の変数を列側に、その分布を左右する「原因」の変数を行側に配置して、行%を記すことがふつう、ということである（後の回で触れるが、原因・結果という言い方は、統計データを見る際にはやや語弊があるが、考える際にはこの方がわかりやすい）。

また、クロス表の表現の仕方は、細かく見れば千差万別であるが（図3）、表面的な違いに惑わされず、示すべき%を示すことに注意を払おう。また、度数分布表と同様に相対度数（%）は副次的な統計量に過ぎないので、基本となる度数を必ず示すこと（または再現可能であること）も重要な注意点である。図3（d）のように、各セルの度数は示さずに行%だけを示す表現も有効であるが、それぞれの100%に相当する合計ケース数（n）を記しておかなければならない。

これはクロス表をもとにしてグラフを作成する際にも同じである。100%に相当する合計ケース数（n）だけは明記しなければならない。

(a)

三世代同居 / 性別	賛成	反対	計
男性	927	366	1293
女性	950	600	1550
計	1877	966	2843

(b)

	三世代同居への賛否		
性別	賛成	反対	合計
男性	927 *71.7 %*	366 *28.3 %*	1293 *100.0 %*
女性	950 *61.3 %*	600 *38.7 %*	1550 *100.0 %*
合計	1877 *66.0 %*	966 *34.0 %*	2843 *100.0 %*

(c)

	賛成	反対	合計
男性	927 (71.7 %)	366 (28.3 %)	1293 (100.0 %)
女性	950 (61.3 %)	600 (38.7 %)	1550 (100.0 %)

(d)

	賛成	反対	n
男性	71.7%	28.3%	1293
女性	61.3%	38.7%	1550

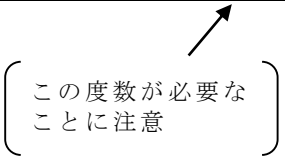


図3　クロス表のいろいろな表現

（問題）

下の表は、「婚姻状態（既婚／未婚）」と「欲しい子どもの性別（男の子／女の子）」のクロス表である（JGSS-2000のデータ）。このクロス表を（1）～（4）の目的で作っているとすると、それぞれの場合について望まれる％の種類は行％、列％、全体％のいずれか。また、実際に％を算出して、それぞれの疑問に回答せよ。

		欲しい子ども		
		男の子	女の子	計
婚姻状態	既婚	992	1359	2351
	未婚	219	211	430
	計	1211	1570	2781

（1）男の子を欲しい人と女の子を欲しい人で、既婚者の割合が高いのはどちらなのか。
（2）既婚者と未婚者で欲しい子どもの性別に違いがあるのだろうか。
（3）全体に占める未婚で女の子を欲しがっている人の割合はどのくらいなのか。
（4）女の子を欲しがっている人の割合が高いのは、既婚者なのか、未婚者なのか。

■散布図

2つの変数の間の関係性を調べるためにクロス表の作成について学習したが、量的変数の場合は同じ目的でしばしば**散布図**（scatter plot; scattergram; scatter diagram）が作成される。散布図は、2つの変数をそれぞれX軸、Y軸として1人1人の回答を対応する座標に点で記した図である（図4）。

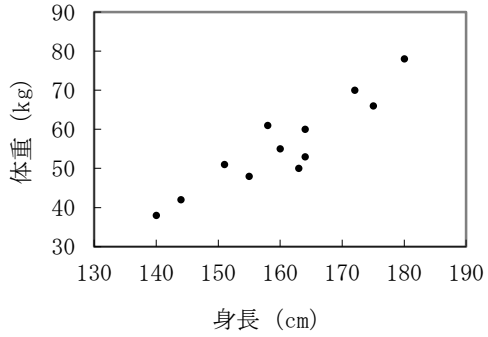


図4　散布図の例

散布図は、クロス表よりも直感的に2つの変数の関係性を理解でき便利なものであるが、残念ながら計量社会学で扱う調査データでは散布図を活用できる機会は多くない。なぜならば、社会調査のデータに含まれる変数は、多くの場合、回答選択肢の数が少なく、散布図を描くのに適していないからである（5段階の評定尺度など）。そのため、やはり関係性

を表わす場合の基本は、クロス表になる。クロス表の作成、%の使い分けを確実に理解しよう。

散布図がその力をもっとも発揮するのは、十分に多くの値を取り得るときである。年齢など多様な値を取り得る変数、複数の変数から作成した合計得点、集計データにおける平均値や比率など、多様な値を取り得る変数を分析する際には、散布図が非常に有効であろう。

（問題）

以下のようなことが調べたいとき、どのようなクロス表を作成すればよいか、表の枠組みを提案しなさい。また、仮にこのクラスで調査をすれば、おそらくこのような結果になるという架空の度数を各セルに記入し、必要なパーセントを計算しなさい。その上で、その結果が仮説を支持する結果なのか、支持しない結果なのかを明記しなさい。

（1）男子学生と女子学生では、男子学生の方が一人暮らしをしている割合が高いだろう。

（2）アルバイトをしている比率が大きいのは、1年生よりも2年生以上の方だろう。

今日のポイント

①2変数間の関係性の分析は、クロス表の%を相対的に比べることが基本。

②目的に応じて3つの%（行%、列%、全体%）を使い分ける。

③使える場面は限定的だが、散布図でも2変数間の関係性が読み取れる。

第6回「関係の読み方（2） 相関係数」

■復習

数回にわたって統計操作の説明が積み重なってきたので､ポイントを整理しておこう(表1)。計量社会学で扱うデータは、複数の変数について多くのケースを調べた統計データである（第2回)。

いろいろと複雑な分析技法も存在するが、まず大切なことは各変数（各調査項目）の分布をよく観察することである｡度数分布表やグラフを用いて1変数の分布を観察する際の注意をまず学習した（第3回)。しかし、実際には多くの度数分布表を素朴に観察することは大変である。そこで、分布の中心と散らばり具合だけを基本統計量で要約する方法を学習した（第4回)。

次に、2つ以上の変数の関係を読み取る話である。2変数の関係は、散布図やクロス表で読み取る（第5回)。クロス表では、適切な%を算出して比較しなければならないが、関心の中心となる変数を列側に、比較するグループを示す変数を行側に配置して、行%を読むことが基本である。やや不適切な表現だが、原因と考える変数を行に配置し、結果と考える変数を列に配置するといってもよい。

表1 いま学習していること

	素朴な観察	統計量による要約
1つの変数の 分布を調べる →	度数分布表 単純なグラフ	基本統計量 代表値（最頻値、中央値、平均値） ばらつき（範囲、四分領域、分散・標準偏差・変動係数）
2つの変数の 関係を調べる →	クロス表 散布図	関係性を表わす統計量 相関係数 連関係数（ユールのQ、ファイ係数、オッズ比など） 順序相関係数（ガンマ、ロー、タウなど）

■「2変数の関係性」をさらに比較する

さて、ではここで「2変数の関係性をさらに比較する」という状況を考えてみよう。たとえば､「授業への出席率が高いほど成績がよい」という関係があるとして、1年生の場合と2年生の場合では、このような関係性の強さに違いが出るのか、といった疑問が浮かぶことがあるかもしれない。たとえば、1年生のときの出席は義務感で出ているだけで、2年生の出席の方がやる気が反映されているので､2年生の方が､出席率と成績の関係が強くなるのではないか、といった仮説が考えられる。このことを確認するためには、1年生と2年生で別々に､「出席率と成績のクロス表（または散布図)」を作成して、比較すればよい。

ところが、2学年ならまだよいが、4つの学年で比べてみようとか、13個の学部で違いを調べようとか考えると、クロス表や散布図を読み取るだけでも大変である。そこで、自然な発想として、2変数の関係性の強さや方向性を1つの数字に要約することができれば、比較が簡単になるはずだ、という考えが思い浮かぶ。度数分布表を読み取る代わりに、平均

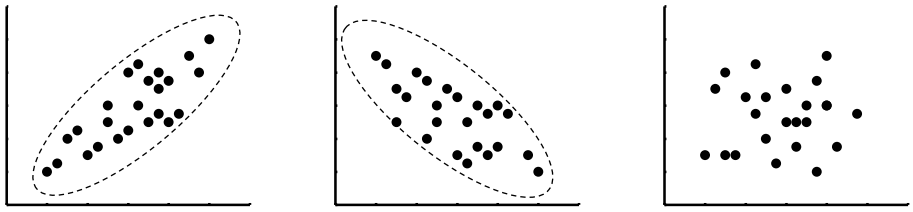
や標準偏差といった数値（基本統計量）に要約したのと同じことである。

クロス表を要約する統計量は、よく使われるものが複数あり、やや複雑である。一方、散布図を要約する統計量では、圧倒的によく使われるものが1つだけある。今回は、散布図を要約する、ピアソンの**相関係数**（correlation coefficient）に絞って、2変数の関係性を要約する意味を学習しよう。クロス表を要約する統計量は、次回解説する。

■相関係数の意味

ピアソンの相関係数(ふつう､単に相関係数といえばピアソンの相関係数のことである)は、量的変数同士の関係性について、散布図に現れる関係性の方向性と強さを1つの数値に要約する。社会学で扱う調査データには質的変数が多いものの、相関係数の考え方は全体の基礎として確実に理解しなければならない。

2つの量的変数XとYの間で､一方の変化に対して他方が比例的に変化する傾向をもつとき、2つの変数は**相関**（correlation）する、という。散布図で描けば、図1（a）（b）のように、直線傾向の関係をもつ場合が相関である。（a）はXが増えればYも増え、Xが減ればYも減るので、2つの変数が同じ方向に動く。この場合を正の相関と呼ぶ。一方、（b）は、XとYが逆方向の動く（Xが増えればYは減り、Xが減ればYは増える）ので、負の相関と呼んで区別する。たとえば、読書量と成績は正の相関をもつ、とか、仕事へのやる気と疲労感は負の相関を示す、とかいう使い方をする。



（a）正の相関
r＝0.804

（b）負の相関
r＝－0.697

（c）無相関
r＝0.066

図1　相関関係

さて、関係の方向性を無視すると、（a）と（b）ではどちらの方が強い相関であろうか。ピアソンの**相関係数**（correlation coefficient）を用いれば、一見しただけでは判断しにくい関係の強さを数値で比較できる。相関係数は一般に記号「r」で表記され、必ず－1から＋1の間の値をとる。正の相関が強いほど＋1に近い値になり、負の相関が強いほど－1に近い値になる。相関関係がない場合には0に近い値になる。図1の場合、（a）と（b）では（a）の方がr＝0.804とサイズが大きいので、より強い相関ということになる。かりに（b）がr＝-0.9であれば、（b）の方が相対的に強い相関である。

絶対量として相関係数の大きさがどの程度あれば、「強い」相関と考えればよいのかは、一概には言えない。ただ、社会学的なトピックの場合、およそ次のようにみなされる。±0.2を越えると弱い相関があると見られることが多い。さらに±0.4を越えていれば、はっきりと相関があると見られる。±0.7を越えていると、かなり強い相関と見られる。

■相関係数の計算

2つの変数XとYの相関係数の計算式は、次のとおりである。

$$\text{相関係数 } r = \frac{\dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}}{\sqrt{\dfrac{\sum(X-\bar{X})^2}{n-1}}\sqrt{\dfrac{\sum(Y-\bar{Y})^2}{n-1}}} \qquad \begin{pmatrix}\bar{X}\text{は}X\text{の 平 均}\\ \bar{Y}\text{は}Y\text{の 平 均}\\ n\text{は 全 回 答}\end{pmatrix}$$

数学的な理解はこの講義の目的ではないが、それほど複雑なことを考えているわけではない。相関係数の分子は共分散と呼ばれる数値で、2つの変数での2次元の散らばり具合を示している。平均を中心にして右上や左下への散らばりが大きいほど、大きなサイズの正の値になり、右下や左上への散らばりが大きいと、大きなサイズの負の値になる。

$$\text{分散 } s^2 = \frac{\sum(X-\bar{X})^2}{n-1} \quad \left(= \frac{\sum(X-\bar{X})(X-\bar{X})}{n-1}\right) \quad \leftarrow\text{似ている}\rightarrow \quad \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}$$

共分散自体を相関の指標とすることもできるが、共分散はXとYの各変数がもつそもそもの散らばり具合が大きければ、大きなサイズの値になってしまう。そこで、共分散をXとYの標準偏差で割ってやり、各変数の散らばりの影響をキャンセルして純粋に相関の強さだけを示すようにしたものが相関係数である。

$$\text{相関係数} r = \frac{X\text{と}Y\text{の共分散}}{X\text{の標準偏差}\cdot Y\text{の標準偏差}} = \frac{s_{xy}}{s_x s_y}$$

例）

右のデータから相関係数を算出したい。（高齢者の友人関係についての仮想データ）

	X＝年齢（歳）	Y＝友人との会話時間（hour）
1人目	50	4.2
2人目	55	4.5
3人目	62	3.3
4人目	60	4.0
5人目	68	3.5

①XとYの基本統計量を算出

Xの平均＝59　Xの標準偏差＝6.86

Yの平均＝3.9　Yの標準偏差＝0.49

②XとYの共分散を算出

$$s_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}$$

$$= \frac{1}{5-1}\{(50-59)(4.2-3.9)+(55-59)(4.5-3.9)+(62-59)(3.3-3.9)+(60-59)(4.0-3.9)+(68-59)(3.5-3.9)\}$$

$$= \frac{1}{4}(-2.7-2.4-1.8+0.1-3.6) = -2.6$$

③相関係数を算出

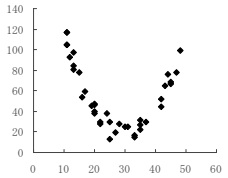
$$r = \frac{-2.6}{6.86\times 0.49} = -0.77$$

④意味を読み取る

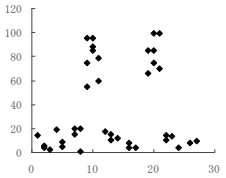
「r＝－0.77なので、2つの変数は強い負の相関を示している。つまり、年齢が高いほど友人との会話時間は短くなる傾向がある」

■相関係数の注意点

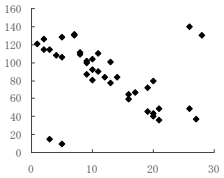
相関係数は非常に頻繁に用いられるが、万能ではないことに注意が必要である。相関係数は2つの変数の間の直線的な関係性しか表していない。規則的ではあるけれども、直線的でない関係性（曲線的な関係など）には反応しない（図2のa、b)。



（a） r ＝-0.32



（b） r ＝0.15



（c） r ＝-0.36
※外れ値がなければ、-0.91

図2　相関係数に反映されない関係性のパターン

もう1つの注意点は、外れ値の影響を非常に受けやすいということである（図2のc)。これは、平均値が持っていた欠点と同様であり、データが持つすべての情報を利用するタイプの統計量が持つ宿命のようなものである。

（問題）

問1. ある大学生の調査で、アルバイトの量（時間／月）と読書冊数（冊／月）の相関係数を調べると、r＝－0.55だったという。この結果の正しい読み取りすべてに○を付けなさい。
（　　）アルバイトが多いほど読書が多い傾向がある
（　　）アルバイトが多いほど読書が少ない傾向がある
（　　）アルバイトが少ないほど読書が多い傾向がある
（　　）アルバイトが少ないほど読書が少ない傾向がある

問2. ある研究で、若い女性が「どのくらい趣味にお金を使うか」を調べている（1ヶ月当たりの教養娯楽費の支出額で測定する)。いくつかの事柄と関係性が強いのではないかと考えて、相関係数を調べてみた結果が下の表である（仮想データ)。相関係数から読み取れることを文章で整理しよう。関係の方向性（±）と強さ（数値のサイズ）に注意すること。
例）……な女性ほど、趣味に費やすお金が多い。一番関係が強いのは……である。

	相関係数
世帯収入（税込みの年収）	0.540
労働時間（1週間の平均時間）	-0.228
テレビ視聴時間（1週間の平均時間）	0.044
結婚していること（0=結婚していない、1=結婚している）	-0.656
親と同居していること（0=同居していない、1=同居している）	0.352

今日のポイント
①「2変数の関係性」をさらに比較するためには、クロス表or散布図を1つの数値に要約できれば便利である。
②散布図を要約する統計量は、相関係数。
+1に近いほど正の相関。-1に近いほど負の相関。0に近いほど無相関。

第7回「関係の読み方（3） クロス表の連関係数」

■2×2のクロス表における3つの連関係数

ピアソンの相関係数は、量的変数同士の関係性を表わす散布図を要約した数値である。しかし、社会調査のデータには質的変数が多く含まれ、変数間の関係性はクロス表で表されることが多い。クロス表に示される関係性も、相関係数と同じように1つの統計量で表すことができる。このような統計量にはいくつかの種類があるが、総称して**連関係数**（association coefficient; coefficient of association）、関連性の指標、関連性の統計量などと呼ばれる。

クロス表の基本は2×2の配置である。2×2のクロス表の各セルの度数を下のようにa、b、c、dで表すならば、よく用いられる連関係数は次のように算出される。

a	b
c	d

ユールのQ　$Q=\dfrac{ad-bc}{ad+bc}$

ファイ係数　$\phi=\dfrac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$

オッズ比　$OR=\dfrac{ad}{bc}$

図1　2×2のクロス表における連関係数

■ユールのQとファイ係数

いずれの連関係数でも、2つの変数の間に関連がまったくない状態の定義は共通している。2つの変数の間に関連がない状態とは、一方の変数の値が違っても他方の変数の分布に変動がない状態のことである。つまり、1行目のグループでも2行目のグループでも、もう一方の変数の分布に違いがない。このとき、a:b=c:dで、変形するとad=bcとなる。すなわち、2つの変数にまったく関連がない状態とは「a×d」と「b×c」が一致するクロス表である。

ユールのQ（Yule's Q）と**ファイ係数**（phi coefficient）の式に注目すると、分子がad-bcなので、関連がまったくない場合には値が0になることがわかる。また、aやdが大きい関連では＋の値、bやcが大きい関連では－の値を取る。相関と同じように、前者を正の関連、後者を負の関連と呼ぶ※。さまざまな例で確認するとわかるが、ユールのQもファイ係数も－1～＋1の値しか取らない。つまり、いずれも相関係数とまったく同じ読み方ができる（＋1に近いほど正の関係が強く、－1に近いほど負の関係が強い）。非常に簡単である。

※質的変数では、「賛成/反対」のようにどちらがプラス側なのかはっきりしている変数もあるが、「男性/女性」のようにどちらがプラス側なのかはっきりしない変数も多い。この場合も便宜的にセルaやセルdが多いことを正の関連と呼ぶことにする。

少し前の回であげた「性別」と「三世代同居への賛否」のクロス表で、ユールのQとファイ係数を算出してみよう（表1）。程度は強いとはいえないが、いずれも正の値なので、クロス表に見られる正の関係性を適切に反映している。

表1　男女別の三世代同居への賛否

	賛成	反対	計
男性	927 (71.7%)	366 (28.3%)	1293 (100%)
女性	950 (61.3%)	600 (38.7%)	1550 (100%)
計	1877 (66.0%)	966 (34.0%)	2843 (100%)

注：JGSS-2000のデータから作成

$$Q = \frac{ad - bc}{ad + bc} = \frac{927 \times 600 - 366 \times 950}{927 \times 600 + 366 \times 950} = 0.231$$

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{927 \times 600 - 366 \times 950}{\sqrt{1293 \times 1550 \times 1877 \times 966}} = 0.109$$

読み取り方が同じなのに、ユールのQとファイ係数で数値が異なるのはなぜだろうか。データによっては、そのサイズがもっと大きく異なるときもある（たとえば、ユールのQでは0.8なのに、ファイ係数では0.4しかない、といったように）。このような違いが出るのは、両者の間で「最大の関連」の定義が異なるからである。ファイ係数では2つの変数の値が1対1に対応することが最大の関連とみなす。たとえば、男性はこの法案に全員賛成するが、女性は全員反対といった場合である。そのため、ファイ係数は、b=c=0のときが最大の正の関連で「＋1」となり、a=d=0のときが最大の負の関連で「-1」となる。これに対してユールのQでは最大の関連をもっと緩やかに考える。男性は法案に全員賛成しているが、女性は賛否が分かれているという場合でも、ユールのQは性別と賛否の間に最大の関連があると考える（男性は全員賛成なのだから、性別の関連は最大）。つまり、b=0またはc=0のとき「＋1」となり、a=0またはd=0のとき「－1」になる。

これはどちらが正しいという問題ではないが、社会調査で扱われる変数は、多くの場合、「相対的な」測定の結果にすぎない。その意味からは、2つの選択肢の間に絶対的な断絶を認めないユールのQの方がふさわしい場面は、自然科学に比べれば多いといえる。

■オッズ比

別の統計量である**オッズ比**（odds ratio）は、「オッズ」という概念に基づいている。オッズとはあることが起こる「見込み」のことであり、正確に記すと、「あることが起こらない確率に対して、あることが起こる確率が何倍あるか」を表わす。少し前の回であげた「性別」と「三世代同居への賛否」のクロス表で考えよう（表1）。男性グループに注目すると、三世代同居に賛成する確率は$\frac{a}{a+b}$であり、賛成しない確率は$\frac{b}{a+b}$である。したがって、三世代同居に賛成するオッズは$\frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b} = \frac{927}{366} = 2.53$と算出できる。つまり、男性は、三世代同居に反対する確率に比して賛成する確率が2.53倍ある（男性の賛成オッズは2.53）。

同じように、女性グループでは、三世代同居に賛成するオッズが$\frac{c}{d} = 1.58$である。これ

ら2つのオッズの比$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{2.53}{1.58} = 1.60$が、オッズ比である。つまり、女性に比べて男性は、1.6倍ほど三世代同居に賛成する見込み（オッズ）が大きいことを示す。オッズ比の式は、結局、$\frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$と非常に簡単なものに整理できる。変数間にまったく関連がなければ$ad = bc$なので、オッズ比は$\frac{ad}{bc} = 1$になるが、これは2つのオッズに違いがなければ、当然、その比が1になることからもわかる。正の関連が強いほどオッズ比は1より大きくなり、負の関連が強いほど1より小さくなる。

それぞれの連関係数は、クロス表がもつ完全な情報を削ぎ落として、関連性の一側面を1つの数値に要約している。どの要約が分析の目的に見合うかを考えて利用する統計量を選択しなければならない。オッズ比は「見込みが〇倍」という具体性をもつのでわかりやすい。しかし、ユールのQやファイ係数は最大の関連が±1で、プラス側とマイナス側で対称になるという抽象的なわかりやすさをもつ。さらに、ユールのQとファイ係数の間では、その抽象的な最大関連の定義が異なるので、この点に注意して使い分ける。

■連関係数の値を比較する

連関係数は単独の数値だけで読むのではなく、複数のクロス表で関係性の方向や強さを相対的に比較するためのものである。多くのクロス表を比較するときにこそ、連関係数が真価を発揮する。たとえば、表1では男性の方が三世代同居に賛成し、女性の方が反対するという傾向が確かめられたが、「この関係性は、若者でも中年でも高齢者でも同じなのだろうか」という疑問をもったとしよう。

このことを確認するためには、人々を年齢層で分けて、複数のクロス表を作成し、各クロス表の行%から2変数の関係性を慎重に読み取ればよい。しかし、表が多くなってくると、その読み取りも簡単ではない。そこでユールのQ等の連関係数を算出して、これを比較すれば、より確実で簡便にクロス表間の比較ができる（図2）。

sexa 性別 と op2gnr Q12 三世代同居観 と age10 年齢10歳刻み のクロス表

度数

age10 年齢10歳刻み			op2gnr Q12 三世代同居観 1望ましい	op2gnr Q12 三世代同居観 2望ましくない	合計
1 20代	sexa 性別	1 男	120	61	181
		2 女	122	83	205
	合計		242	144	386
2 30代	sexa 性別	1 男	126	53	179
		2 女	128	100	228
	合計		254	153	407
3 40代	sexa 性別	1 男	166	55	221
		2 女	151	110	261
	合計		317	165	482
4 50代	sexa 性別	1 男	205	75	280
		2 女	209	135	344
	合計		414	210	624
5 60代	sexa 性別	1 男	184	74	258
		2 女	164	108	272
	合計		348	182	530
6 70代	sexa 性別	1 男	103	44	147
		2 女	127	53	180
	合計		230	97	327
合計	sexa 性別	1 男	904	362	1266
		2 女	901	589	1490
	合計		1805	951	2756

→

	ユールのQ
20代	0.14
30代	0.30
40代	0.37
50代	0.28
60代	0.24
70代	-0.01

→

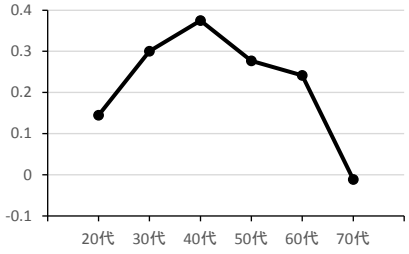


図2　年齢層別に、「性別と三世代同居の関連性」を比較した結果

注：JGSS-2000のデータから作成。ユールのQが大きいほど「男性の方が同居に賛成し、女性の方が反対する傾向」が強いことを意味する。

（問題）

（1）図2のようなユールのQの比較から、何がわかったことになるのか、連関係数の意味を知らない人にも通じるように、結果を読み取りなさい。

（2）なぜこのような結果になったのか、その社会的な理由を自分なりに解釈してみよう。

■大きなクロス表における連関係数

2×2より大きなクロス表での連関係数としては、**クラメールのV**（Cramer's V）がもっともよく利用される。

$$\text{クラメールの } V = \sqrt{\chi^2 / n(\mathrm{k}-1)}$$

※k はクロス表の行数と列数のうち小さい方の数。χ^2（カイ二乗値）は推測統計でよく用いられる統計値。この授業では追求しないことにする。

大きなクロス表では関係性の方向が多様にありえるので（正の関係、負の関係と要約できない)、その方向を特定せずに関係の大きさのみを示す。クラメールのVは0～1の値を取り、1に近いほど関係性が強い。ただし、この数値では、クロス表のどの部分の度数が大きいことを表わしているのか、関係性の中身がまったくわからないことに注意が必要である。

■順序尺度の変数同士の関連性の指標

ピアソンの相関係数は平均値を基準にした指標なので、平均値が計算できる量的変数(つまり、間隔尺度か比率尺度の変数）に対してしか用いることができない。しかし、数値の間が等間隔でなくとも、少なくとも順序尺度の変数であれば、**順序相関係数［順位相関係数］**（rank correlation coefficient）と総称される似たような統計量を用いることができる。つまり、順序が決まっている選択肢（例：これまでに海外に行ったことが「1 まったくない」「2 一度はある」「3 何度もある」）でできている変数同士であれば、大きなクロス表をそのまま用いて－1～＋1の値で関係性を要約することができる。

比較的よく用いられる順序相関係数には次のようなものがある。**スピアマンのρ（ロー）**（Spearman's rho）は値を全ケースの中での順位に変換してからピアソンの相関係数を求める。**グッドマンとクラスカルのγ（ガンマ）**（Goodman-Kruskal's gamma）は、あらゆるケースのペアから、2変数の大小関係が一致するペアの数Pと一致しないペアの数Qを求め（いずれかの変数の値が同じになるペアは集計から除く)、γ=(P－Q)/(P+Q)とその相対比を指標とする。γで集計から除いていた「値が同じになるペア」も分母に加えると**ケンドールのτ_a（タウ）**（Kendall's Tau-a）になり、別のやり方で取り除くとケンドールのτ_bになる。社会調査のデータ分析では順序相関係数はしばしば有効なので、細かな計算方法はともかくとして、値の読み方とそれが用いられる理由は理解しておこう。

（問題）

問1. 次の統計量の中から、相関係数と同じ読み取り方ができるものをすべて選びなさい（－1～＋1の値を取り、＋1に近いほど正の関係が強く、－1に近いほど負の関係が強い）。

ア ユールのQ　**イ** ファイ係数　**ウ** オッズ比　**エ** スピアマンの ρ　**オ** グッドマンとクラスカルの γ　**カ** ケンドールの τ_a

問2. 右のクロス表は、20代の若者に「新聞を信頼するか」「テレビを信頼するか」を尋ねた調査結果である（JGSS-2000,2005,2010）。

2つの変数の関係性について、順序相関係数（グッドマンとクラスカルの γ）を算出すると、調査年ごとに、
2000年……0.528
→2005年……0.911
→2010年……0.815　であった。

調査年	新聞への信頼＼テレビへの信頼	とても信頼している	少しは信頼している	ほとんど信頼していない	計
2000	とても信頼している	21	51	16	88
	少しは信頼している	9	188	56	253
	ほとんど信頼していない	0	3	16	19
	計	30	242	88	360
2005	とても信頼している	12	24	0	36
	少しは信頼している	7	118	27	152
	ほとんど信頼していない	0	3	18	21
	計	19	145	45	209
2010	とても信頼している	14	27	4	45
	少しは信頼している	4	115	25	144
	ほとんど信頼していない	0	4	26	30
	計	18	146	55	219

(1) 順序相関係数（γ）から読み取れることとして正しいものすべてに○をつけなさい。
（　　）2000年よりも2010年の方がテレビを信頼する若者が増えた
（　　）2000年から2005年にかけて新聞を信頼する若者が増えたが、2010年にはやや減った
（　　）どの年でも、新聞を信頼する人の方がテレビも信頼する傾向がある
（　　）新聞を信頼しない人ほどテレビも信頼しないという傾向が一番強いのは2005年だ
（　　）2000年には、新聞を信頼する人ほどテレビは信頼しないという関係性があった

(2) 新聞・テレビを信頼することについて、若者の間でなぜこのような時代的変化が生じたのか、理由を解釈してみよう。

今日のポイント
①2×2のクロス表では、2変数の関係性を要約するために連関係数を使う。
　主な連関係数は、ユールのQ、ファイ係数、オッズ比
　　└→相関係数と同じ読み方　└→関連がないとき値が1
②大きなクロス表では、クラメールのVが有名。
　0（無関連）～1（最大関連）の値を示す
③順序尺度の変数同士の場合には、特殊な順序相関係数も有効。
　スピアマンの ρ、グッドマンとクラスカルの γ、ケンドールの τ_a など
　いずれも、読み取り方は相関係数と同じ

※次回（5／30）の授業初めに 2 回目の小テスト

小テストは、A4 用紙 1 枚を持ち込み可。

第5～7回の内容について、クロス表の作り方と読み方、相関係数や順序相関係数の読み取り、各種の連関係数の読み取りと計算、語句の意味などを確認。

第8回「小休止」

■難しかった？

前回尋ねた「これまでの授業で難しかった点」で多かった意見。

・相関係数や連関係数（ユールのQなど）が何を表わしているのか？
・相関係数の計算
・連関係数の計算
・連関係数の使い分け（どんなときにユールのQで、どんなときにファイ係数？）
・順序相関係数の読み方
・相関係数や連関係数を比較するという意味（前回の図2は何だ？）
・行％、列％、全体％の使い分け
・ばらつきの指標（標準偏差など）の計算
・ばらつきの指標の使い分け
・数式が（いっぱい）出てくるとわからない
・用語がいっぱい出てくると混乱する
・数値の意味を言葉にすること
・計量社会学の知識がない人に説明する、という問題
・聞き逃したところがわからない

（以下、比較的少数意見）
・質的変数と量的変数の区別
・間隔尺度等の変数の種類
・変動係数の意味
・はずれ値という概念
・結果から社会的な理由を解釈すること
・連関係数はΣが出てこないからよくわからない

■行％、列％、全体％（p.22の問題 再掲）

下の表1は、「婚姻状態（既婚／未婚）」と「欲しい子どもの性別（男の子／女の子）」のクロス表である（JGSS-2000のデータ）。このクロス表を（1）～（4）の目的で作っているとすると、それぞれの場合について望まれる％の種類は行％、列％、全体％のいずれか。また、実際に％を算出して、それぞれの疑問に回答せよ。

		欲しい子ども		
		男の子	女の子	計
婚姻状態	既婚	992	1359	2351
	未婚	219	211	430
	計	1211	1570	2781

（1）男の子を欲しい人と女の子を欲しい人で、既婚者の割合が高いのはどちらなのか。
（2）既婚者と未婚者で欲しい子どもの性別に違いがあるのだろうか。
（3）全体に占める未婚で女の子を欲しがっている人の割合はどのくらいなのか。
（4）女の子を欲しがっている人の割合が高いのは、既婚者なのか、未婚者なのか。

もしも世界全体が100人の村だったら……と考えたい
→全体％

もしも世界が「既婚者ばかりの100人の村」と「未婚者ばかりの100人の村」でできていたら……と考えたい
（既婚者村と、未婚者村で、ほしい子どもの違いを比べたい）
（1行目と2行目のグループを100人ずつ調べて、○○の分布の違いを比べたい）
→行％

もしも世界が「男の子がほしい100人の村」と「女の子がほしい100人の村」でできていたら……と考えたい
（男の子ほしがり村と、女の子ほしがり村で、婚姻状態の違いを比べたい）
（1列目と2列目のグループを100人ずつ調べて、○○の分布の違いを比べたい）
→列％

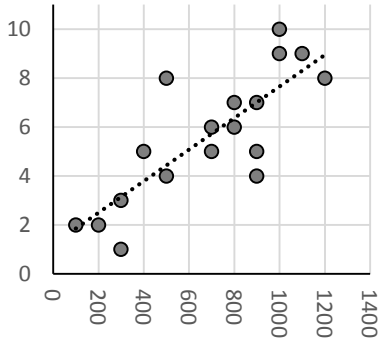
自分でクロス表を作るときには、行％を出せばいいようにすることが基本
①回答の分布を知りたい、関心の中心となる変数　→列側に配置
②比べやすいように、100 人ずつに統一するグループを表わす変数　→行側に配置
③グループ間で行％を比較

■「関係」を要約するとは？

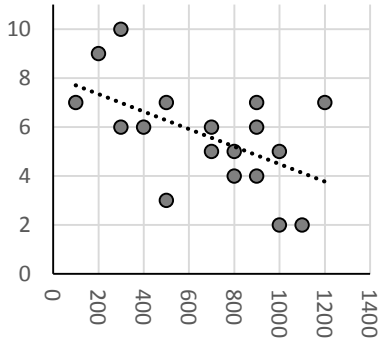
相関係数（r）は、2つの変数の関係性を「方向性」と「強さ」に絞って要約する。

①関係の方向性（→±で表わす）

Xが増えれば、Yは増えるのか、それとも減るのか



r＝0.80
（正の相関）



r＝－0.55
（負の相関）

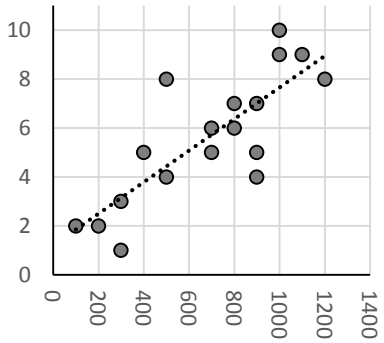
仮想データ

Xが年収（万円）
Yが幸福感（10点満点）

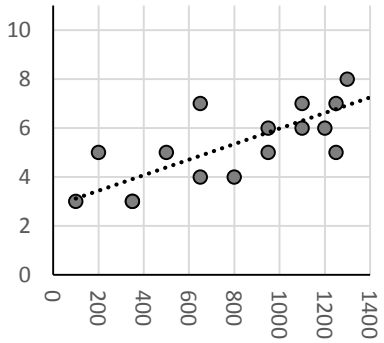
②関係の強さ（→数値のサイズが±1にどれだけ近いかで表わす）

Xの値によって、Yはどれだけはっきり予測できるのか

Xが1増えたときにYがどれだけ多く増えるのか、ではない

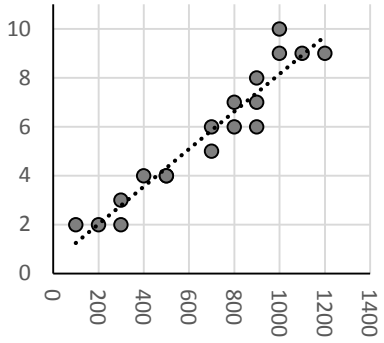


r＝0.80

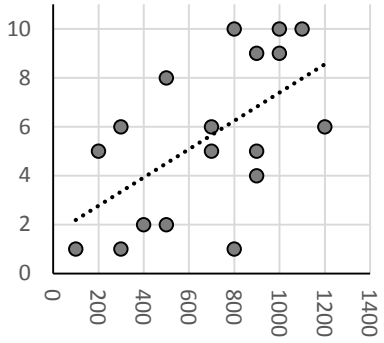


r＝0.81

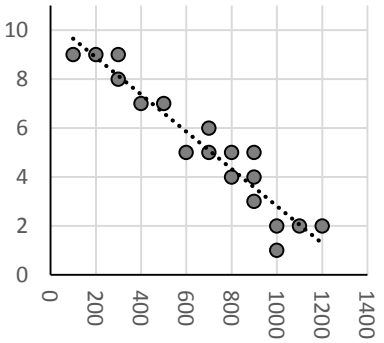
はっきり予測できるというのは、比例関係（直線）にどれだけ近いかということ



r＝0.96
（直線に非常に近い）



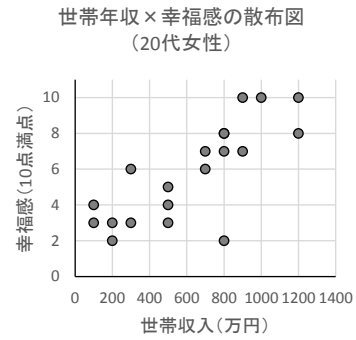
r＝0.56
（直線からややずれている）



r＝－0.95
（直線に非常に近い）

■関係を比べるとは？

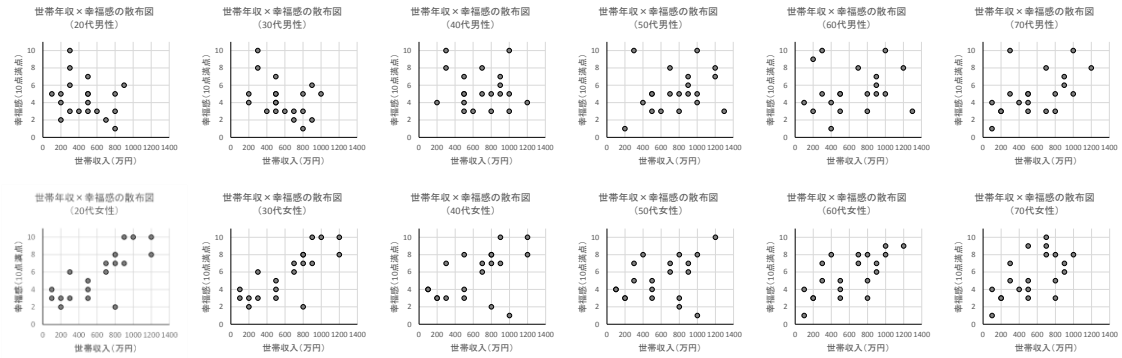
散布図で 2 つの変数（世帯年収と幸福感）の関係はわかる。



↓

では、世帯年収と幸福感の関係は、性別や年齢層によってどう違うのか？

こんなのを「20 代男性の場合」「30 代女性の場合」……といくつも見比べるのは大変。



↓

世帯年収と幸福感の関係性だけを 1 つの数値（相関係数）に要約する

↓

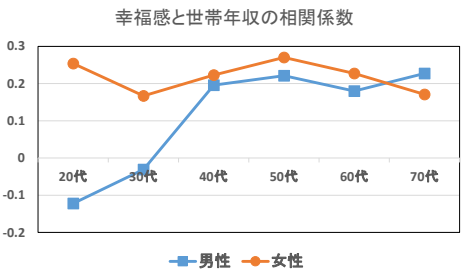
相関係数を見比べれば、どのグループで幸福感と世帯年収の関係が強いのか一目瞭然

	20 代	30 代	40 代	50 代	60 代	70 代
男性	-0.122	-0.031	0.196	0.221	0.180	0.227
女性	0.254	0.167	0.223	0.270	0.227	0.171

注：JGSS-2010 の実際の分析結果

↓

さらにグラフにすれば、全体的にパッと見て比較できる



■連関係数の場合も考え方はまったく同じ

ファイ係数は、クロス表における2つの変数の関係性を「方向性」と「強さ」に絞って要約する。(※実は、クロス表の選択肢を0と1の値で表わして相関係数を無理やり計算すると、ファイ係数と一致する。ファイ係数は相関係数のクロス表版で、数学的にも扱いやすい)

①関係の方向性（→±で表わす）

Xがポジティブならば、Yはポジティブなのか、それともネガティブなのか

	週末に出かける	出かけない	計
恋人がいる	35（70.0%）	15（30.0%）	50（100%）
いない	50（33.3%）	100（66.6%）	150（100%）
計	85	115	200

ファイ係数＝0.32

（正の関連）

（恋人がいる方が週末に出かけやすい）

	週末に出かける	出かけない	計
恋人がいる	10（20.0%）	40（80.0%）	50（100%）
いない	120（80.0%）	30（20.0%）	150（100%）
計	130	70	200

ファイ係数＝－0.54

（負の関連）

（恋人がいる方が週末に出かけにくい）

②関係の強さ（→数値のサイズが±1にどれだけ近いかで表わす）

Xの値によって、Yはどれだけはっきり予測できるのか

	週末に出かける	出かけない	計
恋人がいる	50	0	50
いない	0	150	150
計	50	150	200

ファイ係数＝1.00

（完全に予測できる）

	週末に出かける	出かけない	計
恋人がいる	0	50	50
いない	150	0	150
計	150	50	200

ファイ係数＝－1.00

（こちらも完全に予測できる）

ただし、社会調査の回答は質問文が違えば、容易に動くなど、測定の曖昧さがある。

例)「週末に出かける予定がありますか」を「週末にでかけようと思いますか」に変更

→「出かける」という回答が増える

ユールのQであれば、こうした影響を受けにくい。

回答が絶対的なものでない場合、ユールのQの方が関連性の強さを妥当に表せることが多い。

	週末に出かける	出かけない	計
恋人がいる	50	0	50
いない	0	150	150
計	50	150	200

ファイ係数＝1.00

ユールのQ＝1.00

	週末に出かける	出かけない	計
恋人がいる	50	0	50
いない	40	110	150
計	50	165	200

ファイ係数＝0.64

ユールのQ＝1.00

クロス表の関連性を具体的に「見込み（オッズ）が何倍」と表わしたいならばオッズ比。

（問題）

高齢女性の医療不安について分析している。一人暮らしの女性の方が、将来の医療に不安を感じているのではないか、という仮説を考えて、下のようなクロス表を作成した。

	不安がある	不安がない	計
一人暮らし	18	14	32
一人暮らしでない	56	73	129
計	74	87	161

注：JGSS-2008のデータから70代女性のみ抽出して集計。質問文は「ご自身やご家族の将来のことを考えたとき、「必要なときに医療を受けられない」という不安をどのくらい感じますか」

(1) 回答の分布を知りたい、関心の中心となる変数は？

→｛一人暮らしかどうか・不安があるかどうか｝

比べやすいように、100 人ずつに統一するグループを表わす変数は？

→｛一人暮らしかどうか・不安があるかどうか｝

(2) このとき、必要なパーセントは行％か列％か。

(3) 実際に必要な％を計算して、「仮説は正しい」といえるか結果を読み取りなさい。

(4) ファイ係数、ユールの Q、オッズ比をそれぞれ算出しなさい。

第9回「記述の実践（1） PPDACサイクル」

■個別の技術をつなげる

ここまでに、計量社会学で必要になるデータ記述について、基本的な方法を学習し終わった。すなわち、数値を用いることで社会に客観的な形を与えるための方法として、

1）1つの変数の分布の示し方（度数分布表、基本統計量［代表値とばらつき］）
2）2つの変数の関係の示し方（散布図、クロス表、相関係数、連関係数）

を学習した。より高度な分析技法も多く存在するが、ここまでで学習してきた基本的な方法をうまく組み合わせて駆使するだけでも、大部分の目的は十分に果たすことができる。

この授業の後半部分は、個別に学習してきたことを使って、統計データによる社会の記述の「実践」に可能な限り触れてもらう。それぞれの作業を全体の目的とつなげて理解することを意識してもらいたい。

■PPDACサイクルとは？

計量社会学に限らず、統計的な証拠に基づいて何らかの問題解決を探る手順を**PPDACサイクル**と呼ぶ※。PPDACサイクルとは、ニュージーランドの統計教育学者が90年代後半に提唱した考え方で（Wild & Pfannkuch 1999）、ニュージーランドでは小中学生のうちから、下のようなポスターでその枠組みが叩き込まれているという。

※似たような言葉に、経営学や品質管理で用いるPDCAサイクルがあるが、別ものである。

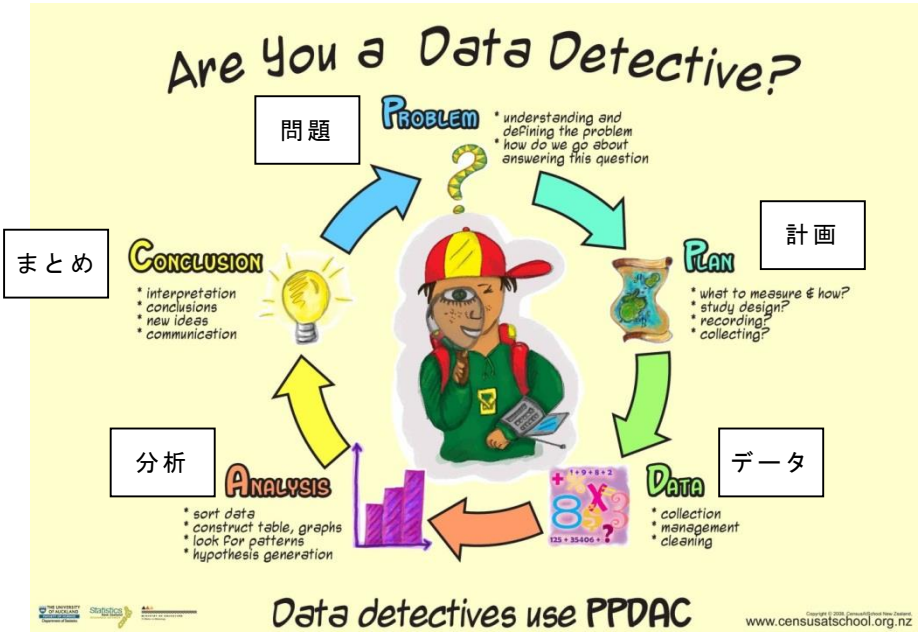


図1 PPDACサイクル

P、P、D、A、Cは、それぞれ**Problem（問題）**、**Plan（計画）**、**Data（データ）**、**Analysis（分析）**、**Conclusion（まとめ）**の頭文字である。簡単にその手順を追ってみよう（より細かくは表1のとおり）。

[P]最初の大事なステップは、自分が取り組もうとしている問題・疑問が何なのか、はっきりとさせることである。問題があいまいなまま調査を始めても、けっしてうまくいくことはない。

[P]第2のステップは、どうすれば疑問が解けるのか、計画を立てることである。どこからどのようなデータを取ってきて、どう並べるのか、大雑把な全体像を描く。

[D]第3のステップは、計画に沿ったデータ収集である。社会調査によって新しいデータを集めるべき場合もあれば、すでに存在するデータをインターネットなどで集める方が有効な場合もある。

[A]第4のステップでは、収集したデータを計画どおりに分析する。分析といっても大げさに考える必要はない。集めた数値をわかりやすい表に並び替えることや、グラフに整理することも、分析に含まれる。

[C]第5のステップでは、分析によってわかったことをまとめて、最初に設定した疑問への解答を示す。ここでは、自分の答えを間違いなく他人に伝えるコミュニケーションの技術も重要となる。

こうして当初の問題の解答が得られると､それによって新たな疑問が生じることがある。あるいは、当初の計画通りに疑問が解き明かされずに、問題の一部が取り残されてしまうこともある。いずれの場合も、その問題に取り組むため、最初のPに立ち戻るサイクルとなる。PPDACサイクルを回し続けることで、状況の改善や理解が進むと期待できる。

■文章・グラフ・表の選択

計量社会学の初学者は、PPDACサイクルの最後の段階にあたる「まとめ（Conclusion)」を明確にイメージすることが、まず大切である。この段階で意識すべきことは、他者への伝達、広い意味でのプレゼンテーションの仕方である。統計情報は、扱いを誤るとむしろ伝わりにくい種類の情報である。単に自分が正しい情報を得るだけではなく、それを適切に伝達することに腐心しなければならない。

統計的な分析結果は、**文章・グラフ・表**のいずれかで表現される。どれを用いても表現できるが､3つの中からもっとも状況に適したツールを「自覚的に」選択することが大切である。大まかに以下のような点に留意して判断するとよい。

・伝達したい数値はいくつあるのか？
・伝達時間はどのくらいあるのか？
・正確な値を伝える必要があるか？

伝達したい数値が2、3個しかないのであれば、図表は大げさで、文章の中に数値を含めた方がよい。多くの数値を表現したいときには図表を用いるが、グラフと表の役割は大きく異なる。短い時間で多くの情報が伝わるのはグラフである。また、1つひとつの正確な値を伝える必要がなく、大まかなパターンを伝えたい場合にはグラフの方が適切である。1つずつの値を正確に伝えたい場合は表を用いる。このような側面から総合的に判断する。

（問題）

次のそれぞれの統計表は、「表のまま」「文章の中に埋め込む」「グラフにする」のうち、どれがもっとも適切だろうか。理由といっしょに考えてみよう。

（1）野球部のピッチャーAについて、球速の推移を部内で検討中。

各投球回の平均急速（km/h）

	1回	2回	3回	4回	5回	6回	7回	8回	9回
第1試合	135	133	134	130	121	124	120	118	120
第2試合	130	130	128	131	122	120	118	119	118
第3試合	132	131	130	129	128	129	131	127	125
第4試合	134	132	131	130	115	121	115	120	118
第5試合	133	130	131	128	115	121	120	121	117

（2）各工場での不良品率を、部長に報告中。

	商品A	商品B	商品C	商品D
吹田工場	2.54‰	1.31‰	0.15‰	1.44‰
堺工場	2.77‰	1.29‰	0.22‰	1.56‰
松坂工場	3.10‰	1.44‰	0.98‰	1.89‰

（3）学園祭の入場者の内訳を、実行委員会で報告中。

関大生	68%
他学の学生	12%
一般の人	20%

n=7,250

（4）営業のため、パワーポイントで商品の購買層を説明中。

	新商品を選択	従来品を選択	計
男性客（n=352）	71.3%	28.7%	100%
女性客（n=198）	44.6%	55.4%	100%

■パターンの要約

分析結果をただ図表などで提示するだけでは、プレゼンテーションとしては不十分である。その図表で数値のどのようなパターンを記述しているつもりなのか、必ず「言葉で」説明する必要がある。

記述したい情報は、ほとんどの場合、何らかの変数間の「関係性」である。変数間の関係性は、非常に広い内容を指す。たとえば、「国際化」という時系列的なトレンド（時代の変化）は、「時間（年次）」という変数と「国際性の重要度」という変数の間の関係性である。したがって、関係性の記述に慣れることは、一般に重要である。

よくある悪い例は「死亡率は年齢と関係する」というように関係の有無にだけ言及してしまう記述である。**関係の方向性（±）と強さ（サイズ）**を示さなければ、十分な記述ではない。これらは両方そろって示さなければならない。たとえば、「年齢が上がるにつれて死亡率も上がる」は、関係の方向性は示しているが強さを示していない。適切な記述は「年齢が5歳上がるごとに死亡率はほぼ倍増する」といったように方向性と強さを含んだ上で、なるべく簡潔な表現である。

■GEEアプローチ

また、関係性を記述するといっても、分析の結果はそれほどきれいに一貫したパターンを示すわけではない。このときよくある間違いは、細かな点を一つずつ並べて記述してしまい、結局まとめになっていないというものである。また逆に、注目したいパターンのみを抜き出して記述して、都合のよい数値だけを紹介しているようになってしまうこともある。

複雑になりがちなパターンをバランスよく要約するには**GEEアプローチ**（GEE approach）が効果的である（Miller 2004）。まず、細かいことは無視して図表の一番大きなパターンを記述する（一般化 generalization）。次に、そのパターンが具体的に図表のどこからどのように読み取られたのか、いくつかの数値で例を示す（例示 example）。最後に、そのパターンが当てはまらない箇所が図表の中にある場合には、その箇所について言い訳をする（例外 exception）。この枠組みを意識すれば、正確な情報をわかりやすく伝えやすい。

（GEEアプローチによる記述の例）

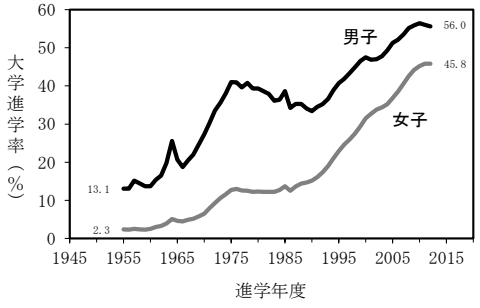


図2　男女別大学への進学率の推移（短期大学は除く）
出典：学校基本調査（文部科学省 1955～2012）

図2のように、男子も女子も大学進学率は急上昇しており、10年ごとに約7ポイントのペースで進学が増えている[G]。その結果、最新の2012年調査では、男子は56.0%、女子も45.8%と半数程度が大学に進学していることになる[E]。ただし、1975～90年ごろは例外で、進学率の上昇が停滞しており、男子の進学率はむしろ低下していた[E]。

男女の差に注目すると、男子の方が10ポイント程度進学率が高いという傾向は全体的に変わっていない[G]。たとえば、1955年で10.8ポイントの差だったのが、2012年調査でも10.2%の差が維持されている[E]。例外はやはり1975～90年ごろで、この時期は男子の進学率だけが急上昇したため、男女差が最大で約30ポイントに広がっていた[E]。

（問題）

(1)「関係の方向性と強さ」という視点から、次の記述の悪い点を指摘しよう。

・アンケートの結果、食堂の満足度は値段と関係することがわかりました。

・この大学生調査から、飲酒の翌日はケガをしやすいことが明らかになった。

(2)「GEEアプローチ」という視点から、次の記述をよりよいものに書き換えよう。

通路に置くゴミ箱の数を増やせばゴミのポイ捨てが減るか、実験してみました。ゴミ箱を5個にした月曜日は、ゴミのポイ捨てが25か所でみつかりました。ゴミ箱を6個にした火曜日は22か所で、ゴミ箱7個の水曜日は20か所、ゴミ箱8個の木曜日は10か所、ゴミ箱9個の金曜日は11か所でした。ゴミ箱の数とポイ捨ての量が関係することがわかります。

	月	火	水	木	金
ゴミ箱の数	5	6	7	8	9
ポイ捨ての数	25	22	20	10	11

今日のポイント
①統計的な問題解決は、データの収集・分析の技術があるだけではだめ
PPDACサイクルを意識しよう
②分析結果の表現では、以下の点にとくに気をつけよう
・文章／表／グラフのどれを使うのが一番よいか、自覚的に判断する
・変数間の関係は、関係の方向性（±）と強さ（サイズ）を両方とも示そう
・複雑なパターンは、GEEアプローチで文章を整理しよう

＜文献＞

C.J. Wild and M. Pfannkuch. 1999. “Statistical Thinking in Empirical Enquiry,” *International Statistical Review*, 67(3):223-265.

渡辺美智子. 2007. 「統計教育の新しい枠組み：新しい学習指導要領で求められているもの」 『数学教育学会論文誌』 48(3,4):39-51.

Miller, Jane E. 2004. *The Chicago Guide to Writing about Numbers*. The University of Chicago Press. (＝長塚隆監訳. 2006. 『数を表現する技術: 伝わるレポート・論文・プレゼンテーション』 オーム社.)

第10回「記述の実践（2） 比較のプランと作表」

■統計分析≒作表

前回、PPDACサイクルという考え方に触れ、とくに最後の段階（Conclusion: まとめ）の注意点について解説した。プレゼンテーションしたい分析結果をどのように表現するかという話である。

しかし、当然ながら、表現方法を考える以前に、まず表現すべき結果を分析で出さなければならない。統計的な分析技法は、さまざまに存在するが、基本的に考えるべきことは、**作表**（tabulation）である。つまり、どんな分析をするかを考えるということは、突き詰めると「どんな表を作るかを考えること」といってよい。最終的に「表」ではなく「グラフ」や「文章」で数値を表現するとしても、その元は「表」だからである。

PPDACサイクルの「Plan」の段階では、作表のイメージを中心に考えるとよい。目的を果たすためにはどのような表が必要か、その表を作るためにはどのような変数群が必要か、それらの変数はどのような質問項目で測定できるか、といったことをさかのぼって考える。

■基本：まず度数分布表とクロス表

一口に作表といっても、いろいろな種類の表があるが、私たちは少なくとも度数分布表とクロス表の作成について学んでいる。まずは、これらを確実に作れるようになろう。下のように、目的に沿った表のイメージから1人で実践できなければならない。

「学園祭には何年生が多く来ているのだろうか」（目的）
→参加者を調査して「学年の度数分布表」を作ろう（抽象的な作表イメージ）
→具体的にはこんな形の表で、たとえばこんな数値が入るはずだ（作表イメージの具体化）

「自宅生より下宿生の方が学園祭に参加していそうだが、本当にそうだろうか」（目的）
→在学生を調査して「住居×学園祭参加のクロス表」を作ろう（抽象的な作表イメージ）
→仮説どおりならば、クロス表にこんな数値（%）が入るはずだ（作表イメージの具体化）

これらに十分理解した上で、さらに一変数の分布を要約した基本統計量（平均値や標準偏差）を整理した作表や、二変数の関係を要約した相関係数や連関係数を整理した作表にも慣れてほしい。

もちろん、実際に作表をするためには、SPSSなど何らかの統計分析ソフトを使用しなければならない（使用しないと大変である）。しかし、どんな表が作りたいかということが手書きでもはっきりとイメージできていれば、ソフトの操作はまったく難しくない（楽に集計をするためのソフトなのだから、難しいわけがない）。実際に集計をしなくても練習はできる。こういうことを知りたいとすると、どんな表を作ればよいことになるのか、まずはコンピューターやデータを離れてイメージする力を鍛えよう。

■補足：比較の重要性を再び

改めて強調しておくが、計量社会学のデータから適切に意味を読み取るには、比較の視点が大切になる。単純な度数分布表やクロス表を作成しているときも、どんな人々とどんな人々のグループを比べているのか（何を比較の軸にしているのか）をはっきりと意識しよう。たとえば、「未婚男性の生活満足度が低い（5点満点中、平均2.2点）」と分析したとき、それは既婚の男性と比較しているのか、未婚女性と比較しているのか、はたまた数年前の未婚男性と比較しているのか、外国の未婚男性と比較しているのか、比較対象をはっきりさせなければ、数値には意味がない（1変数の度数分布表を読み取る場合も、ある選択肢の度数を他の選択肢の度数と比較している。）。このため、どのような目的の分析であっても1つの数値だけを算出することは、ほとんど考えられず、複数のグループについて、同じ種類の数値を算出して比較するはずである。したがって、必然的に、分析結果は複数の数値を併記した「作表」になる。作表のプランは、同時に比較のプランなのである。

■補足：クロス表の縮約

10個の選択肢からなる度数分布表はわりと簡単に読み取ることができるが、10×10のクロス表は100個のセルがあるので読み取りにかなり骨が折れる。また、しばしばそのような表を作ること自体が無意味である。社会調査のデータで作るクロス表は、調査項目の選択肢をそのまま用いるのではなく、行や列の数を減らして、縮約したクロス表を作ることが必要になってくる場合が意外と多い（表1、2が縮約の例）。

縮約したクロス表を作らなければならない状況は、大きく2通りある。1つは、確認したい事柄に対して、選択肢の数が不必要に多い場合である。たとえば、内閣を支持するかしないかというYES／NOの区別だけが関心の対象であるのに、「強く支持する」「ある程度支持する」「どちらかといえば支持する」……など、支持・不支持の程度が細かく分かれているときには、単純に支持と不支持に二分して縮約したクロス表を示す方が分かりやすい。

もう1つの状況は、できあがった表の中に度数が0であったり非常に小さかったりするセルが多く、すかすかのクロス表になってしまう場合である。ない袖は振れないので、いくつかの似たような選択肢をまとめて、可能な範囲で意味のある集計をする方が賢明である。

表1　世帯収入と貧富解消政策への賛否（縮約前）

		Q47 貧富解消政策への賛否						
		賛成	どちらかといえば賛成	どちらともいえない	どちらかといえば反対	反対	無回答	合計
		度数	度数	度数	度数	度数	度数	度数
Q05 世帯収入のレベル	平均よりかなり少ない	111	71	88	11	7	5	293
	平均より少ない	319	224	296	58	24	9	930
	ほぼ平均	305	308	522	120	46	5	1306
	平均より多い	53	70	114	39	21	2	299
	平均よりかなり多い	4	7	9	3	2	1	26
	無回答	9	8	12	1	1	8	39

注：データはJGSS-2000

表2　世帯収入と貧富解消政策への賛否（縮約後）

賛否 / 世帯収入	賛成	どちらともいえない	反対	計
少ない	725 *60.0%*	384 *31.8%*	100 *8.3%*	1209 *100%*
ほぼ平均	613 *47.1%*	522 *40.1%*	166 *12.8%*	1301 *100%*
多い	134 *41.6%*	123 *38.2%*	65 *20.2%*	322 *100%*
計	1472 *52.0%*	1029 *36.3%*	331 *11.7%*	2832 *100%*

今日のポイント
①どんな分析でも、とにかく、最終的に作る「表の形」をイメージしよう
②データを集める「前に」作表のプランを立てることが大切。トレーニングしよう

（問題）

右のような質問紙調査を90名の大学生に対して行ったとする。

次のようなことを知りたいときに、どのような表を作成すればよいか。それぞれイメージする表を作成して、数値は予想で書き入れなさい。

(1) この学生たちは「お金」をどのくらい重要と考えているか？

(2) 男子と女子では、どちらの方が大阪を「住みやすい」と感じているのだろう。

＜実習用アンケート＞

Q1　充実した大学生活のために、次のことはどのくらい重要だと思いますか。また、現在の状態について、どのくらい満足していますか。それぞれに○を付けてください。

1 2 3 4 5　極めて重要 / とても重要 / ある程度重要 / 少しは重要 / あまり重要でない

1 2 3 4 5　満足 / どちらかといえば満足 / どちらともいえない / どちらかといえば不満 / 不満

	重要度					満足度				
(a) 目標を立てること	1	2	3	4	5	1	2	3	4	5
(b) 授業での勉強	1	2	3	4	5	1	2	3	4	5
(c) 授業外の勉強	1	2	3	4	5	1	2	3	4	5
(d) 家族からの支援	1	2	3	4	5	1	2	3	4	5
(e) 十分な睡眠	1	2	3	4	5	1	2	3	4	5
(f) よい食事	1	2	3	4	5	1	2	3	4	5
(g) お金	1	2	3	4	5	1	2	3	4	5
(h) 趣味	1	2	3	4	5	1	2	3	4	5
(i) 資格の取得	1	2	3	4	5	1	2	3	4	5
(j) アルバイト	1	2	3	4	5	1	2	3	4	5
(k) 一人の時間	1	2	3	4	5	1	2	3	4	5
(l) 友人関係	1	2	3	4	5	1	2	3	4	5
(m) 就職の見込み	1	2	3	4	5	1	2	3	4	5
(n) 部活・サークル	1	2	3	4	5	1	2	3	4	5

Q2　次のうち、「大阪」のイメージに合うと思うもの<u>すべてに○</u>をつけてください。
1 ごみごみしている　2 好ましい　3 活気がある　4 怖い　5 楽しい
6 住みやすい　7 華々しい　8 息苦しい　9 安らか　10 かっこいい
11 悲しい　12 すばらしい　13 忙しい　14 さみしい　15 恥ずかしい

Q3　次のうち、「東京」のイメージに合うと思うもの<u>すべてに○</u>をつけてください。
1 ごみごみしている　2 好ましい　3 活気がある　4 怖い　5 楽しい
6 住みやすい　7 華々しい　8 息苦しい　9 安らか　10 かっこいい
11 悲しい　12 すばらしい　13 忙しい　14 さみしい　15 恥ずかしい

Q4　あなたは男性ですか、女性ですか。
1 男性　2 女性

(3) 大阪びいきな人は東京を目の敵にすることがあると聞く。たとえば、大阪を「楽しい」と主張する人は、東京を楽しくないと主張する傾向があるのか？

(4) 結局の所、学生は全体的に見て大学生活の何を重要視しているのかを要約したい。a～nの中で、重要度が高い項目はどれなのか、教えてほしい。

(5) 誰でも同じくらい満足している項目もあれば、人によって満足・不満が大きく分かれる項目もある。a～nの中で、満足度の格差が大きい項目がどれなのかを知りたい。

(6) 自分が重要視している事柄ほど、力を入れているので満足しているとも考えられるし、逆に要求水準が高まって不満を抱えているとも考えられる。たとえば、「趣味」が重要と考えている人は、そうでない人よりも自分の趣味への満足度が高いのか、低いのか？

(7) 重要に思っていることと満足していることがマッチしている項目とマッチしていない項目（重要だけど満足できていないなど）を知りたい。a～nのそれぞれについて、重要度と満足度の間の関係が強い項目、弱い項目はどれなのか？

(8) 男子と女子では東京のイメージがいくらか違うだろうが、どの選択肢について、とくにイメージが違っているのか、男女差が大きいベスト3を特定したい。

(9) 自由な分析視点から、このデータを使ってできる面白い「作表」を提案てほしい。

第11回「記述の実践（3） グラフの描き方」

■データ分析を実践するということ

繰り返すが、統計的なデータ分析の基本は作表である。「社会についてどのようなことを知りたいか？」→「どのような表を作ればそれがわかるか？」という発想が自然にできれば、計量社会学は間違いなく楽しい。自分がイメージした表の中に入る数値さえ調べれば、知らなかったこと、予想しなかったことが次々に「自分の手元で」明らかになるからである。誰かが本に書いていたことでもない。誰かが教えてくれたことでもない。いままさに、自分が社会（社会調査データ）と直接対話して得られた情報である。

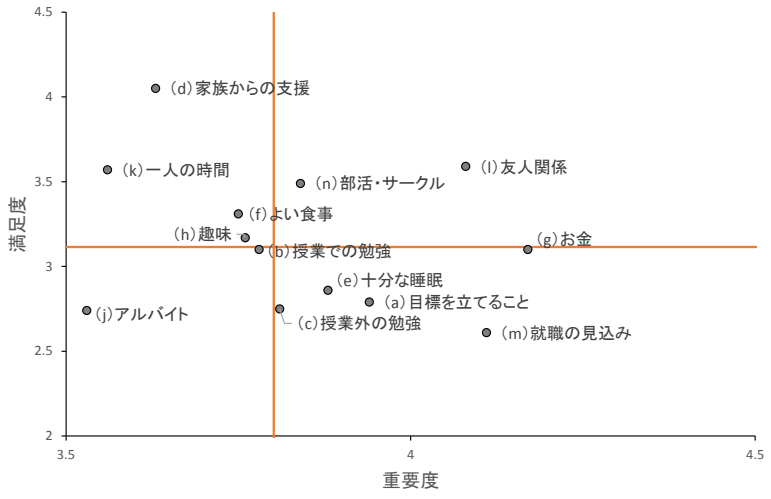
作表のための道具は、まずは簡単なものでよい。実際、我々が知りたいことの多くは、度数分布表とクロス表だけで知ることができる。とにかく、1つの変数（調査項目）の分布が知りたいときは度数分布表、2つの変数の関係が知りたいときはクロス表である。平均値や標準偏差、相関係数など要約のための統計量を使えば、複数の度数分布表、クロス表の結果をまとめて1つにできるので、作表の幅はさらに大きく広がる。

工夫をすれば、これらのわずかな道具立てだけで、本当に多様なことを知ることができる。前回練習に使ったデータを用いて例を示そう。表1は、大学生活に関する14個の項目それぞれについて、重要度と満足度の分布を平均値で要約して、横に並べたものである。つまり、14×2＝28個の度数分布表のそれぞれを要約した数値（平均値）を改めて1つの表に作表し直した。これを見ると、同じ項目の重要度と満足度を比較できる。グラフにした図1を見るともっと意味がわかりやすいだろう（このような図示を重要度－満足度分析と呼ぶことがある）。(e) 十分な睡眠や (m) 就職の見込み、(a) 目標を立てることについては、重要だと思っているが、現状に満足していないということなどが一目瞭然になる。この作表の元になっているのは、度数分布表（およびそれを要約した平均値）だけである。クロス表すら使っていない。

表1 大学生活の重要度と満足度の比較（平均値）

	重要度	満足度
(a)目標を立てること	3.94	2.79
(b)授業での勉強	3.78	3.10
(c)授業外の勉強	3.81	2.75
(d)家族からの支援	3.63	4.05
(e)十分な睡眠	3.88	2.86
(f)よい食事	3.75	3.31
(g)お金	4.17	3.10
(h)趣味	3.76	3.17
(i)資格の取得	3.38	2.48
(j)アルバイト	3.53	2.74
(k)一人の時間	3.56	3.57
(l)友人関係	4.08	3.59
(m)就職の見込み	4.11	2.61
(n)部活・サークル	3.84	3.49

図1 大学生活の重要度―満足度分析

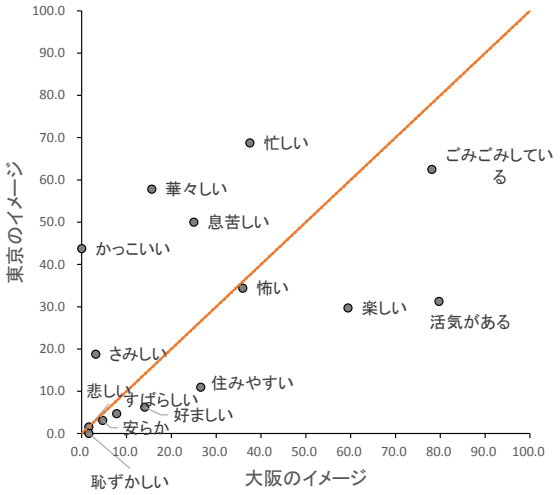


同じように、大阪と東京のイメージに当てはまるものを複数回答で尋ねた結果は、表2のように比較できる。これも15×2＝30個の度数分布表（それぞれに○があったかなかったかが1つの変数になるので、項目の数だけ度数分布表ができる）の結果を単純にまとめたものである。図2のように図示すると、対角線で区切って右下は大阪の方がイメージの強い項目、左上は東京の方がイメージの強い項目であることがすぐにわかる（このような図示を競合分析と呼ぶことがある）。これも度数分布表を元にしただけだが、十分に興味深い。

表2　大阪と東京のイメージの比較（選択割合）

	大阪	東京
ごみごみしている	78.1	62.5
好ましい	14.1	6.3
活気がある	79.7	31.3
怖い	35.9	34.4
楽しい	59.4	29.7
住みやすい	26.6	10.9
華々しい	15.6	57.8
息苦しい	25.0	50.0
安らか	4.7	3.1
かっこいい	0.0	43.8
悲しい	1.6	1.6
すばらしい	7.8	4.7
忙しい	37.5	68.8
さみしい	3.1	18.8
恥ずかしい	1.6	0.0

図2　大阪と東京の競合分析



まずは簡単な道具立てだけでよい。多くの高度な分析技法を学習するよりも、「自分で作表するのだ」という姿勢でさまざまなデータに臨もう。アンケート用紙を見たら、どのような表が作れそうか想像しよう。素データが扱えるならば、実際に作表してみよう。新聞や雑誌で表やグラフを見たら、自分で同じものを作るときの手順を思い浮かべてみよう。

■グラフの必要性

先の例からもわかるように、グラフによる提示はしばしば強力である。とくに、多くの数値からパターンを読み取る場合には、表のままよりも情報が伝わりやすい。グラフの元になる表を適切に作ることがもっとも大切であるが、その上でグラフの適切な作り方を知ることは非常に役立つ。何より、情報が視覚化されるグラフ作りは単純に楽しい。

第3回で少し触れたが、グラフ作成の際に気をつけなければならない基本原則は次の2点である。

・グラフは何らかの数値を比較する。
・グラフはそのために何らかの視覚情報を利用する。

これらは当たり前のように思えるかもしれないが、どの「種類」のグラフがどのような数値の比較をするために、どのような視覚情報を利用しているのかは、意外と意識されていない。表3は代表的な5種類のグラフについて、これらの情報をまとめている。

表3　代表的なグラフのポイント（第3回の再掲）

	比較の対象	利用する視覚情報
棒グラフ	ある数量の大きさ	棒の長さ
折れ線グラフ	ある数量の連続的な変化	線の傾き
円グラフ	全体に占める構成比	パイの面積
帯グラフ	グループ別の構成比	帯の面積
ヒストグラム	連続した階級の度数	柱の面積

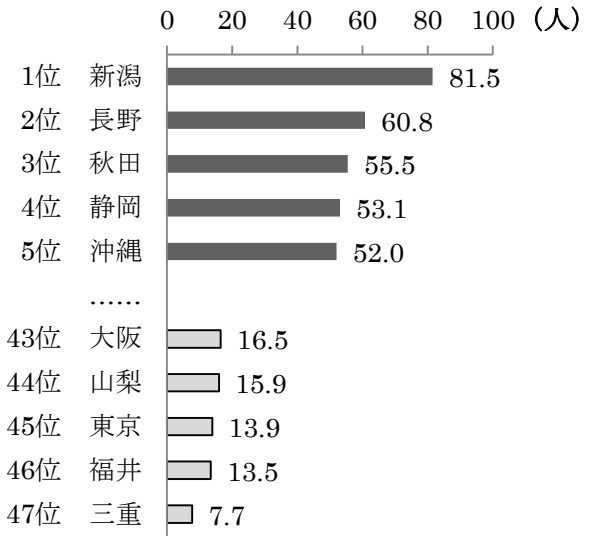
棒グラフは何らかの数量の大きさを比較するために、棒の長さでその数量の大きさを表したものである。比較するものは、度数の他に相対度数（%）や比率尺度の変数の*注 平均値など、その絶対的な大きさに意味があるものであれば何でもよい（図3)。〔※注：間隔尺度の変数は数値の絶対量を比べられないので、棒グラフはおかしいことに注意〕

一方、折れ線グラフで比較すべきなのは、それぞれの頂点の高さではない。比較すべき単位は、頂点と頂点を結ぶそれぞれの線分である。線分の傾き方を比較することで、変化の傾向が読み取れる（図4)。

円グラフと帯グラフは両方とも、全体に占めるそれぞれのカテゴリーの構成比を示す。帯グラフは、特にその構成比をグループ間で比較するのに向いている（図5、6)。

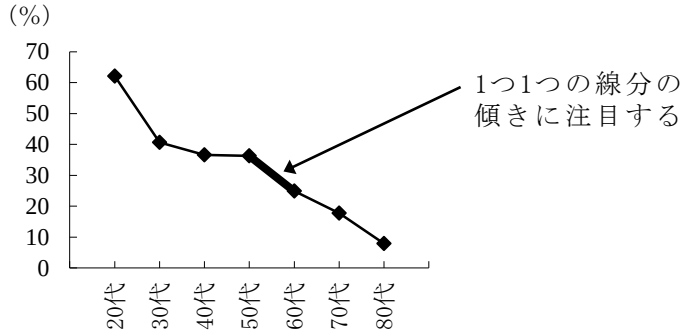
ヒストグラムは、棒グラフの棒と棒の間の隙間をなくしただけに見えるが、その意味合いは全く異なる。棒グラフがその長さに意味があるのに対して、ヒストグラムはその「面積」に意味がある。ヒストグラムの柱と柱がくっついているのは、隣の区分と連続的に繋がっているからである。したがって、隣あった柱の面積を合わせて、より広い範囲の度数を一目で把握することもできる（図7)。

図3 人口10万人あたりのバスケ競技者人口



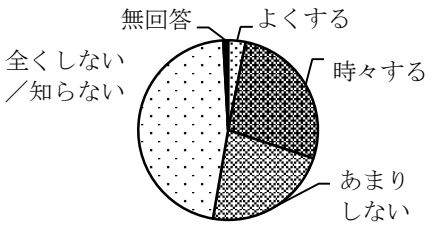
出所：2009年度「バスケットボール競技者登録者数」（財団法人日本バスケットボール協会）　都道府県別人口は、「平成21年3月31日現在、住民基本台帳に基づく人口、人口動態及び世帯数」（総務省）による

図4 世代によるカラオケをする割合の変化



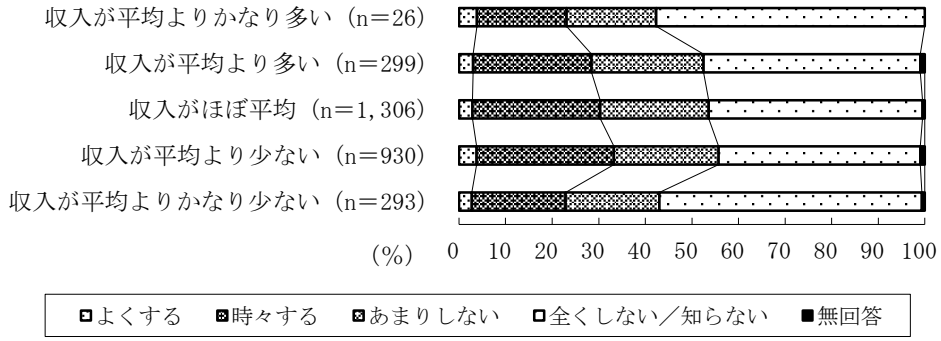
出所：JGSS-2000

図5 宝くじはどのくらいの人が買っているの？



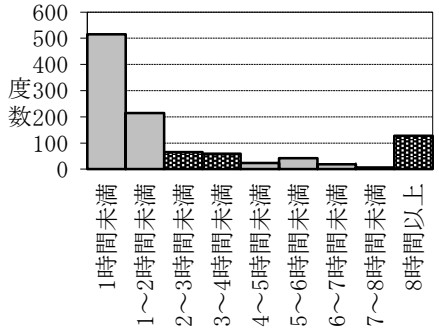
出所：JGSS-2000

図6 ふつうの収入の人が宝くじを買う（収入と宝くじ購入頻度の関係）



出所：JGSS-2000

図7 ペットを飼っている人が1日にペットと過ごす時間

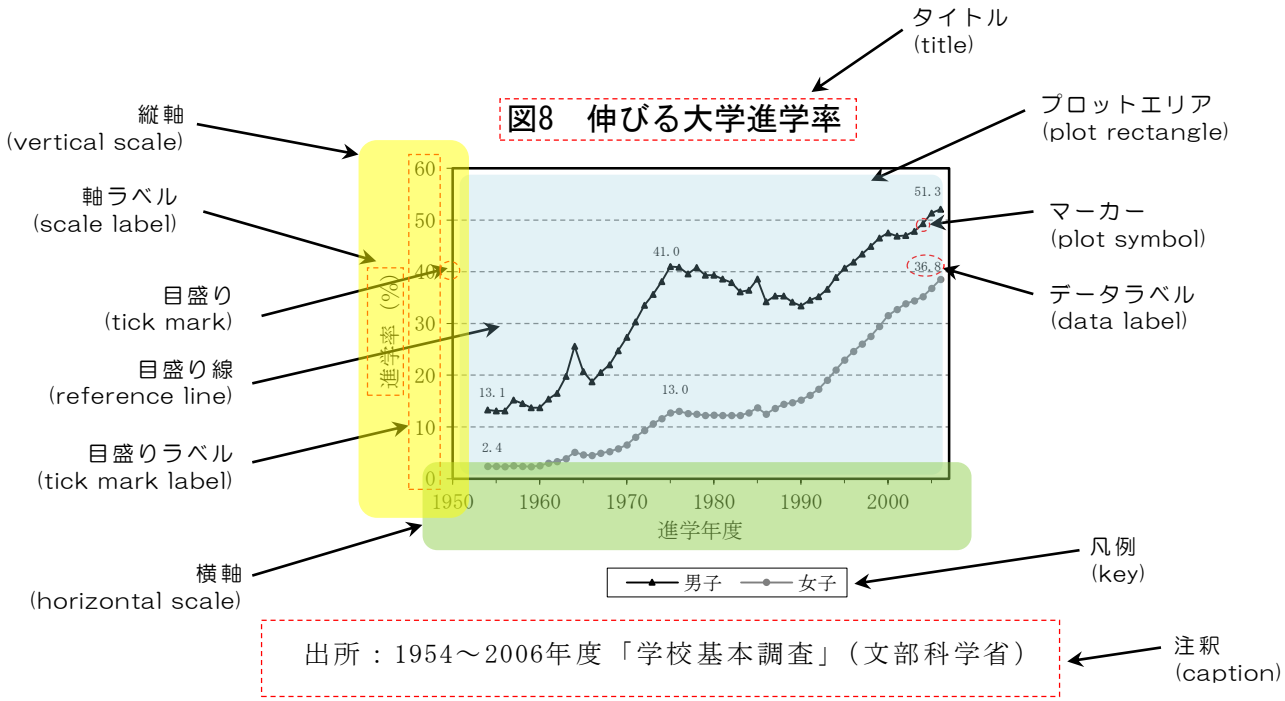


出所：JGSS-2000

■グラフのパーツ

グラフのパーツに注目した場合、グラフ作りの基本原則として以下のような点があげられる。

・必ずタイトルを付ける。(図表番号を含む)
・どこかからデータを取った場合、出所を示す。(何年に誰がした何という調査か)
・軸には必ず軸ラベル、目盛りラベルを付ける。
・プロットエリアには、極力プロット以外(凡例など)を含めない。
・1つのグラフで多くのことを表そうとしない。
・不必要な装飾は避ける。



■グラフの誤用

比較のために利用する重要な視覚情報を混乱させるようなグラフは作成してはならない。例えば、図9のように目盛りが0から始まっていない棒グラフが不適切なのは、数値の大きさを表すはずの「棒の長さ」を混乱させるからである。折れ線グラフであれば目盛りが0から始まっていないことに問題はない。折れ線グラフではむしろ、目盛りの範囲を調整して折れ線のおよその角度が45度になるようにするのが適切とされている(人間は45度付近の角度をもっとも敏感に感知できる)。

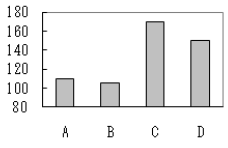
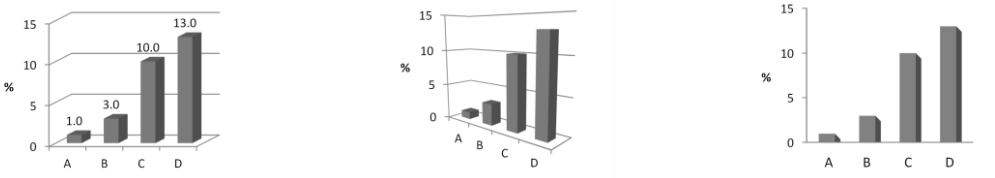


図9　不適切な棒グラフ(打ち切り)

また、図10のような立体の棒グラフが不適切とされるのも、棒の長さがわかりにくくなるからである。その意味で、(a) よりも (b) の方が混乱が大きくなる。これに対して (c) の立体棒グラフにはほとんど問題はない (棒の長さがわかりやすいため)。



(a)　　(b)　　(c)

図10　不適切な棒グラフ (立体)

視覚情報の混乱ではなく、そもそもそのグラフで比較できない数値をグラフ化してしまうことにも注意しなければならない。例えば、間隔尺度の変数の平均値を棒グラフで比較している誤りをよく見かける。0からの距離に意味がない間隔尺度の平均値はそのサイズに意味がない。直感に訴えるグラフは強い力を持つだけに扱いに注意を要する。

■文献紹介

通り一遍のことが知りたければ、山本 (2005) がコンパクトである。上田 (2005) は基本を押さえつつも、グラフの研究者としてマニアックな指摘もあり、おもしろい。ジョーンズ (2007＝2008) は一見するとただのビジネス書だが、意外と内容がしっかりしている。

実践的なExcelによるグラフ作成の本は、早坂清志のものが圧倒的によい。ハウツーとして優れているだけでなく、統計学的な視点をふまえて適切なグラフ作成を促している。基本的なものは早坂 (2009)、マニアックなものは早坂 (2008) で解説されている (早坂 (2011) は、たぶん早坂 (2008) の改訂版。内容を未確認)。

〈文献〉

ジェラルド・E・ジョーンズ著、夏目大訳 2008 『チャート・図解のすごい技』 日本実業出版社. (原著2007年刊行)

早坂清志 2008 『Excelの極意 4 「魅せるグラフ」を極める』 毎日コミュニケーションズ.

早坂清志 2009 『達人が教えるExcelグラフテクニック101』 毎日コミュニケーションズ.

早坂清志 2011 『Excelの極意 2　グラフ』 毎日コミュニケーションズ.

上田尚一 2005 『統計グラフのウラ・オモテ』 ブルーバックス.

山本義郎 2005 『グラフの表現術』 講談社現代新書.

今日のポイント
①単純な集計で作表のプランを立てることにひたすら慣れよう
②グラフは「何の数値を比較するのか」「どんな資格情報で比較するのか」に注意
③基本の5つのグラフから、意識的に最適なグラフを選ぼう

（前回の問題の模範解答）

※当然、他のやり方もある。数値は実際の調査結果。

（1）表Aのように、お金の重要度は非常に高く評定されている。半数近くの人が5点満点での「4点」という回答で、「1点」や「2点」という人は5%程度しかいない。

表A　お金の重要度の度数分布表

	度数	%
重要度　1	1	1.6
2	2	3.1
3	7	10.9
4	29	45.3
5	25	39.1
計	64	100.0

（2）表Bのクロス表でわかるとおり、女子の方がわずかに大阪が「住みやすい」と○している割合が高い。ただし、6%程度の違いで差は大きくない。

表B　性別と「大阪は住みやすい」のクロス表

	住みやすいに○あり	○無し	計
男子	8 23.5%	26 76.5%	34 100%
女子	9 30.0%	21 70.0%	30 100%
計	17 26.6%	47 73.4%	64 100%

（3）表Cのとおり、大阪が楽しいと思う人は、東京も楽しいと思う割合が相対的に高いので、仮説は否定される。ただし、大阪は楽しいが東京は楽しくないという人が22人いるのに対して、逆に東京だけが楽しいという人が3人しかいないことには注目すべきである。

表C 「大阪は楽しい」と「東京は楽しい」のクロス表

	東京:楽しいに○あり	○無し	計
大阪:楽しいに○あり	16 42.1%	22 57.9%	38 100%
○無し	3 11.5%	23% 88.5%	26 100%
計	19 29.7%	45 70.3%	64 100%

（4）表Dは各項目の重要度の高さを平均値で要約して、数値が高い順に並べ直したものである。お金～友人関係までが4点以上の高い平均値を示している。

表D　学生生活の各項目の重要度の平均値

	重要度の平均値
(g)お金	4.17
(m)就職の見込み	4.11
(l)友人関係	4.08
(a)目標を立てること	3.94
(e)十分な睡眠	3.88
(n)部活・サークル	3.84
(c)授業外の勉強	3.81
(b)授業での勉強	3.78
(h)趣味	3.76
(f)よい食事	3.75
(d)家族からの支援	3.63
(k)一人の時間	3.56
(j)アルバイト	3.53
(i)資格の取得	3.38

（5）表Eは各項目の回答のばらつき具合を標準偏差で要約して、高い順に項目を並べたリストである。数値が高いことは、人によって満足・不満の回答が分かれやすいことを意味している。極端な違いはないが、よい食事、アルバイト、趣味などで、満足度の格差が大きい。

表E　学生生活の各項目の満足度の標準偏差

	満足度の標準偏差
(f)よい食事	1.14
(j)アルバイト	1.13
(h)趣味	1.09
(d)家族からの支援	1.05
(n)部活・サークル	1.05
(g)お金	1.03
(i)資格の取得	1.01
(e)十分な睡眠	0.98
(c)授業外の勉強	0.97
(k)一人の時間	0.91
(l)友人関係	0.89
(b)授業での勉強	0.86
(a)目標を立てること	0.85
(m)就職の見込み	0.73

(6) 表Fは趣味の重要度によって満足度がどう異なるかクロス集計したものである。選択肢が5つで煩雑なので、点数が高いグループと低いグループに分割し直した。趣味を重要と思っている学生の方が、趣味への満足度も高いことがわかる。単純にパーセントで見ると40%程度の差があり、関係が非常に強い。

表F　趣味の重要度と満足度のクロス表（縮約）

	趣味の満足度 高い(4･5)	低い(1･2･3)	計
趣味の重要度 高い(4･5)	19 51.3%	18 48.7%	37 100%
低い(1･2･3)	3 12.0%	22 88.0%	25 100%
計	22 35.5%	40 64.5%	62 100%

(7) 表Gは各項目の重要度と満足度の関係性を相関係数で要約したものである。値が大きい順に並べ替えている。つまり、部活・サークルや趣味では、正の相関が強いので、重要と考えている人ほど満足度も高く、両者がマッチしている。一方、お金や家族からの支援では負の相関であり、お金を重要と思っている人ほど、満足度が低いということである。

表G　学生生活の各項目の重要度と満足度の相関係数

	重要度と満足度の相関係数
(n)部活･サークル	0.419
(h)趣味	0.335
(k)一人の時間	0.312
(l)友人関係	0.246
(j)アルバイト	0.138
(a)目標を立てること	−0.042
(c)授業外の勉強	−0.080
(i)資格の取得	−0.090
(e)十分な睡眠	−0.095
(f)よい食事	−0.109
(m)就職の見込み	−0.123
(b)授業での勉強	−0.146
(g)お金	−0.181
(d)家族からの支援	−0.246

(8) 性別と東京のイメージの各項目で15個のクロス表を作り、各表で男女の選択割合を比較すれば、イメージの違いを特定できる。男女差が大きかった順に並べ直したリストが表Hである。活気がある、住みやすいなどのイメージは男子の方が強く、さみしい、忙しい、怖いといったイメージはやや女子に強いことがわかる。

ただし、選択率で比較すると、性別とは関係なくそもそも選択率が高い項目では男女差も大きくなりやすく、選択率が低い項目では男女差が小さくなりやすくなってしまう。このことを問題と考えるならば、各クロス表での関係性を連関係数で要約して比較する方がよい。表IはユールのQで比較した結果である。結果は表Hと似通っている。

表H　東京のイメージの男女差（選択率で比較）

東京は……	男子の選択率	女子の選択率	男女差(男−女)
3 活気がある	38.2	23.3	14.9
6 住みやすい	17.6	3.3	14.3
1 ごみごみしている	67.6	56.7	11.0
8 息苦しい	52.9	46.7	6.3
12 すばらしい	5.9	3.3	2.5
7 華々しい	58.8	56.7	2.2
15 恥ずかしい	0.0	0.0	0.0
9 安らか	2.9	3.3	−0.4
2 好ましい	5.9	6.7	−0.8
11 悲しい	0.0	3.3	−3.3
10 かっこいい	41.2	46.7	−5.5
5 楽しい	26.5	33.3	−6.9
14 さみしい	11.8	26.7	−14.9
13 忙しい	61.8	76.7	−14.9
4 怖い	20.6	50.0	−29.4
n	34	30	

表I　東京のイメージの男女差（ユールのQで比較）

東京は……	性別と各項目の関連性(ユールのQ)
6 住みやすい	0.723
3 活気がある	0.341
12 すばらしい	0.289
1 ごみごみしている	0.230
8 息苦しい	0.125
7 華々しい	0.044
9 安らか	−0.065
2 好ましい	−0.067
10 かっこいい	−0.111
5 楽しい	−0.163
13 忙しい	−0.341
14 さみしい	−0.463
4 怖い	−0.588
11 悲しい	−1.000
15 恥ずかしい	−−

（問題）

上記の（1）～（8）の表をグラフ化するとすれば、基本の5種類のグラフ（棒グラフ・折れ線グラフ・円グラフ・帯グラフ・ヒストグラム）の中でどれが最適か。その理由も説明しなさい。

※次回（6／27）の授業初めに3回目の小テスト

小テストは、A4用紙1枚を持ち込み可。

第9～11回の内容について確認。結果を伝える文章の書き方、必要な作表の判断、グラフの適切な使用など。

第12回「因果関係への注意（1） 相関と因果」

■シンプソンのパラドックス

1つの調査データの中で、次のような矛盾するような結果が得られることは、ありえるだろうか。

1) 男子学生の中で、自宅生と下宿生でどちらの方が自分で料理をしているかを調べると、（当然であるが）下宿生の方が料理をしていた。
2) 女子学生の中で調べても、やはり下宿生の方が料理をしていた。
3) ところが、男女を合わせた全体でみると、自宅生の方が料理をしていた。

結論を言ってしまうと、このようなパラドックス（逆説）は起こりうる。下のようにやや極端な数値で例をあげてみれば、そのことはすぐわかるであろう。

表1 男女別のクロス表

		自分で料理をするか		
		する	しない	計
男性	自宅生	3(10%)	27	30
	一人暮らし	20(20%)	80	100
	計	23	107	130
女性	自宅生	70(70%)	30	100
	一人暮らし	27(90%)	3	30
	計	97	33	130

表2 男女を合わせたクロス表

	自分で料理をするか		
	する	しない	計
自宅生	73(56%)	57	130
一人暮らし	47(36%)	83	130
計	120	140	260

このように、集団に分けた場合と全体で観察した場合で認められる関連性が大きく異なる現象を、**シンプソンのパラドックス**（Simpson's paradox）と呼ぶ。統計的な調査で非常によく見られる現象で、解釈を誤りやすいので、確実にその意味を理解する必要がある。

■シンプソンのパラドックスの原理

この一見すると奇妙な現象は、言葉で書けば次のように説明できる。全体として見たときに自宅生に料理をする人が多くなっているのは、ただ単に女子学生に自宅生が多いためである。女子学生の方が男子学生よりも料理をしているので、集計上は、自宅生に料理をしている人が多いことになる。

もう少しシステマティックには、3つの変数の関係図式から理解できる。もともと観察している2つの変数をＸとＹ、集団に分けるための変数をＺとする。集団に分けた3重クロス表で見えているＸとＹの関係性は、図1（a）の太線の部分のみを純粋に表している。これに対して、変数Ｚで分けずに全体で観察しているＸとＹの関係性は、純粋なＸＹ間の関係性に加えて、ＸＺ間の関係性とＹＺ間の関係性が折り重なって見える関係性が、いっしょくたに混ざったものを表していることになる（b）。

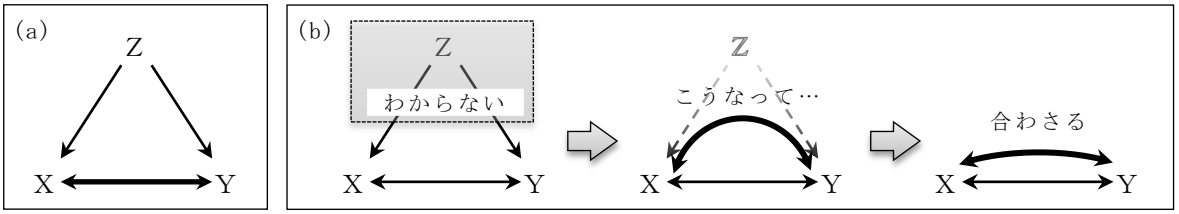


図1　シンプソンのパラドックスの原理

■見せかけの関係

このとき混ざりあった関連性の組み合わせによって､いろいろと不思議な現象が起こる。この現象を正しく解釈するためのもっとも重要なキーワードが、**見せかけの関係［擬似相関］**（spurious relation; spurious correlation）である。見せかけの関係とは、適切なグループ分けをしないで全体を見ると､2つの変数の間にあたかも重要な関係があるかのように見えるが､それは共通の原因である第3の変数によって引き起こされているにすぎない、という場合を指している。このとき、本質的には意味がない歪んだ関係が観察されることになる。最初にあげた例は､「性別」という共通原因によって､「自宅生であること」と「料理をすること」の間に、見せかけの正の関係が発生して、本来の負の関係を覆い隠してしまったのである。ここでは質的変数（カテゴリー変数）によるクロス表で例を示したが、量的変数であっても、考え方はまったく変わらない。

この現象は、計量社会学にとって極めて重要な問題を示唆している。我々が統計的な調査データから知りたいことは、ほとんどの場合、何らかの**因果関係**（causal relation）の有無やその大きさである。統計は、その因果関係を客観的に示す、と多くの人々が信じている。つまり､「自宅生の方が料理をしている」という統計データは､「自宅生であることが料理をすることを引き起こす」証拠である、と考えてしまう。ところが、見せかけの関係が存在する以上、ただ単に2つの変数（XとY）の関係を統計的に調べても、それで因果関係がわかるわけではない。一般に、この事実は「**相関と因果は異なる**」という戒めとして徹底的に注意される（ここで用いられる「相関」は、相関係数に表される直線的な関係に限定せずに､統計データの表面的な関係全般を指す広義の相関である)。この戒めを忘れると、完全に間違ったデータ解釈を次々におこなってしまうことになる。

■共通の原因への注目

一方で、この問題を回避する方法は難しいわけではない。先の例からもわかるように、問題を引き起こす第3の変数さえ自覚していれば､その変数でグループ分けした上で､もともと関心のあった2つの変数の関係を調べればよい。もし、見せかけの関係であれば、グループ別の観察では関係性が見られなくなるはずであるし､見せかけの関係でないのならば、グループ分けしても同様の関係性が残るはずである。

具体例を示そう。表3は、実際の調査データでの見せかけの関係の例である。「子どもを1人だけもつとしたら、男の子がほしいか、女の子がほしいか」を尋ねている。表3（a）からは､「タバコを吸う人の方が男の子をほしがる傾向が強い」ということがわかる。この関係性は客観的な事実であるが、このことから「タバコを吸えば、男の子がほしい気持ちが引き起こされる」､つまり因果関係がある、と解釈することは思考が飛躍している。少し考

えればわかるように、これは性別という共通の原因による見せかけの関係である。一般に、現代日本人は自分と同性の子どもをほしがる傾向があるので、男性は男の子をほしがり、女性は女の子をほしがりやすい。また、男性の方が喫煙率が高い。このことから、本質的な因果関係がない2つの変数の間に見せかけの関係が観察されることになる。

そこで、本当に見せかけの関係かどうかを確認するために、表3（b）のように男女別にして集計をやり直してみると、「喫煙」と「ほしい子どもの性別」の間にはほとんど何の関係もなくなった。同じ性別の中では、何の関係性も観察されないという結果が、性別が重要な共通原因であったことを示している。もし、男女別でもまだ関係性が観察されるならば、性別が引き起こす見せかけの関係以外の意味が残されていることを意味する（本質的な因果関係かもしれないし、また別の原因による見せかけの関係かもしれない）。

表3　実際の見せかけの関係の例（喫煙×ほしい子どもの性別：JGSS-2000）

（a）グループ分けしない場合

	男の子がほしい		女の子がほしい		計
喫煙する	479	54.8%	395	45.2%	874
喫煙しない	729	38.5%	1164	61.5%	1893
計	1208		1559		2767

（b）性別でグループ分けした場合→「喫煙」と「ほしい子どもの性別」の関係が消滅

		男の子が欲しい		女の子が欲しい		計
男性	喫煙する	411	65.2%	219	34.8%	630
	喫煙しない	384	61.3%	242	38.7%	626
	計	795		461		1256
女性	喫煙する	68	27.9%	176	72.1%	244
	喫煙しない	345	27.2%	922	72.8%	1267
	計	413		1098		1511

このように見せかけの関係を引き起こす共通原因のことを、**先行変数［交絡変数］**（antecedent variable; confounding variable）と呼ぶ※。

※ 本来の用語の意味からは、「交絡変数」の方が正確な用語であるが、社会学では当初この考え方が紹介されたときに、「先行変数」の呼び方が広まってしまったので、伝統的にこちらをよく用いる。先行変数は、本来、ある変数よりも先に起こると想定される変数のことを指す。だから、正確には、先行変数の一部が交絡変数として見せかけの関係を引き起こす、といえる。

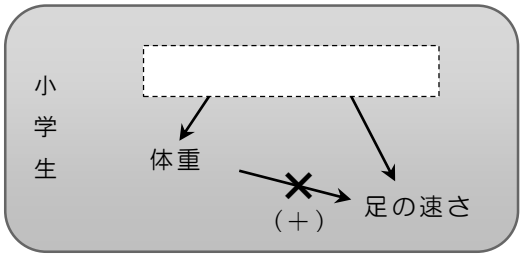
とにもかくにも重要なことは、社会現象を観察するときに、積極的に第3の変数による見せかけの関係の可能性を考慮することである。統計調査の結果を用いて新聞等でなされる主張の中には、見せかけの関係を示しているにすぎない可能性が高いものが頻繁に見受け

られる（例：別資料の「コーヒーと肝がん」「朝食と成績」）。もちろん、本当に見せかけの関係かどうかは、データによって検証しなければはっきりとした結論を下すことはできない。しかし、大部分の過ちは、慎重な思考だけで十分に看破できる。常に、見せかけの関係の可能性を疑って、先行変数［交絡変数］を頭の中で探すクセを付けることである。それだけで一段階も二段階も上の水準で社会現象について考えることができる。

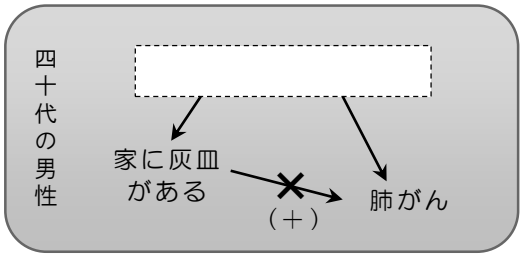
（問題）

1．次のような2変数について調査データで関係性を調べると、まず間違いなく強い関係性が観察される。しかし、この関係性は見せかけの関係の可能性がある。どのような共通原因が見せかけの可能性を引き起こすと考えられるか、先行変数を想像してみよう。

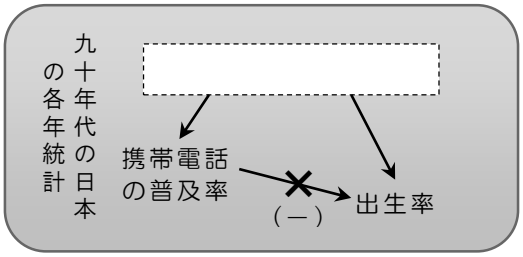
（1）



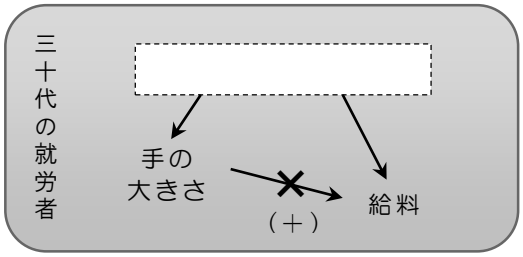
（2）



（3）



（4）



2．身近なことで、見せかけの関係が観察されるであろう現象を、何か1つ想像し、共通原因を含めた3つの変数の関係を図示しなさい。矢印には正の関係か負の関係かがわかるように＋－の記号を付けること。

今日のポイント
①統計でわかるのは相関関係。因果関係とは違う
②見せかけの関係（疑似相関）にだまされないためには、関係を引き起こす共通原因（先行変数、交絡変数）を想像することが大切

〈文献〉

ボーンシュテッド&ノーキ著 海野道郎・中村隆監訳 1990 『社会統計学』 ハーベスト社.
保田時男 2014 「見せかけの関係にだまされない」 関西大学 WEB 版模擬講義 http://www.kansai-u.ac.jp/koudai/movie/index.html（2014 年 6 月 13 日取得）（iTunesU でも配信）

第13回「因果関係への注意（2） 見せかけの関係の追求」

■相関関係と因果関係は異なる（復習）

前回、「相関関係と因果関係は異なる」ということを学習した。つまり、クロス表や散布図、あるいはそれを要約した相関係数や連関係数で2つの変数に関係性があることがわかったとしても、それはそのまま因果関係が存在することの証明にはならない。たとえば、「友人が多い学生の方が、大学生活に満足している」ということが調査でわかったとしても、それは「友人の数」という原因が「大学生活の満足」という結果を引き起こす因果関係を示すことにはならない（友人の少ない学生に強制的に友人を作らせても、大学生活の満足度の分布が上昇しない可能性がある）。

その理由は、2つの変数の相関関係が共通の原因（先行変数）による見せかけの関係である可能性があるからであった。たとえば、「部活やサークルに入った」ということが、友人を増やし、同時に大学生活の満足度を高めているのかもしれない。あるいは、「適応力の高い性格」が共通の原因なのかもしれない。

■因果関係は証明できるのか

相関関係は因果関係の存在を保証してはくれない。では、因果関係の存在を証明するためには、どうすればよいのか。この辺りの事情について詳しい書籍としては、久米（2013）をお勧めする。政治学の例が中心だが、社会科学全般に通用する優れたテキストである。

結論を述べてしまうと、究極的には統計データから因果関係を証明することは、不可能である。なぜならば、統計データからは社会で起こっていることについて、何らかの原因が何らかの効果を「引き起こしていることそのもの」を観察することができないからである。観察できるのは、何らかの原因（と考えられるもの）と何らかの結果（と考えられるもの）がよくいっしょに発生しているという事実に過ぎない。したがって、相関関係の存在は示せるが、因果関係の存在は証明できない。

したがって、因果関係の存在を主張するために満たさなければならない最低限の条件（因果関係の必要条件）に注意を払いながら、常に間違えている可能性を頭に置いておかなければならない。あらためて因果関係が成立するための必要条件を整理すると、以下の3点にまとめられる。

条件① 統計的関係性の存在
条件② 時間順序が正しい
条件③ 見せかけの関係でないこと

まず、2つの変数の間に統計的な関係性が存在しなければならない。これは当たり前のことであって、クロス表や散布図でまったく何の関係性も見られない2つの変数の間に因果関係があると考えることはできない（ただし、本来の因果関係とは逆方向の見せかけの関係

が存在することで、両者が打ち消し合って何の関係性も見出せなることが理論上はありえる。現実的には、そのような偶然は滅多に発生しない)。

次に、時間順序を考えたときに、原因の方が結果に先行していなければならない。前回は注目しなかったが、因果関係の誤解として、単純に原因と結果を逆に考えてしまう、という可能性もある。たとえば、友人が多いから大学生活に満足しているのではなく、大学生活に満足しているからよく学校に脚を運び、友人が増えているのかもしれない。この条件のポイントは、「時間順序が分からなければ、因果関係をはっきりさせることはできない」ということである。先ほども例にあげたとおり、「友人が多い学生ほど、大学生活に満足している」ということが観察されても、友人が多いことと、大学生活に満足していることのどちらが時間的に先行しているのかわからない。そのため、この情報だけでは、どちらが原因かを特定して因果関係を定めることはできない。

もし、5月に友人が増えた学生ほど6月に大学生活への満足度が上昇していた、といったデータであれば、時間順序がはっきりしているので、この条件を満たすことになる。このような理由から、因果関係に関心の強い調査では同じ人を何回かの時点で繰り返し調査する方法（パネル調査と呼ぶ）が好まれる。あるいは、データ上の保証がなくても、理論的に考えてどちらが時間的に先か自明だ、と考える場合もある。たとえば、「天気が晴れている方が、来客が多い」という相関関係は、データ上どちらが先かわからなくても、天気の方が先であることは自明だろう（ある店の来客者数によって天気が変わるわけがない)。ただし、本当に自明なのか、判断を間違えないように慎重な注意は必要である。

くりかえすが、3つ目の条件を理解することは、もっとも重要である。たとえ、2つの変数XとYの間に統計的な関係性が存在し、時間順序が確認されたとしても、XとYの関係性が、「XとYに共通の原因」によってもたらされたものであってはならない。もし、共通の原因が存在するのならば、XとYの関係は見せかけの関係（擬似相関）にすぎないことになる。そうでないことを確認するためには、その共通原因（先行変数）でグループ分けをした上でもう一度2つの変数の関係性を調べればよい。

しかし、実際的に考えると、どこまで確認すれば「見せかけの関係ではない」ということを示したことになるのであろうか。AI研究者のジュディア・パールは(Pearl 2000＝2009)は、そのような統計的な条件をはじめて体系的に整理した。これは大変興味深い研究であるが、同時に、現在の社会調査のデータは、因果関係の特定に必要とされる精巧なデータから大きくかけ離れていることを示している。残念ながら、我々は統計データだけでは社会事象の因果関係を特定できないと考えた方がよいだろう。そのため、統計的なデータだけでなく質的調査（観察やインタビュー）や理論的な考察にも取り組むことが非常に大切になる。

前回からの繰り返しになるが、見せかけの関係に惑わされないためには、常識的な知識や理論的な考察をもとに、2つの変数には「共通の原因があるかもしれない」と常に注意を払うことが、もっとも大切である。統計的なデータ分析の結果を待つまでもなく、大部分の見せかけの関係は頭の中だけで駆逐できる。新聞や雑誌、インターネットには、調査結果をもとにして因果関係を示唆する記事がよく掲載されている。それは本当に因果関係なのか。因果が逆の可能性、見せかけの関係である可能性に常に注意を払い、批判的に検討する姿勢を日々訓練しよう。

（問題1）

「家族といっしょの方が自殺する？」

高齢者の自殺というと一人暮らしの孤独な老人というイメージを持ちがちだが、上野（2007＝2011）によると、**高齢者の自殺率は、意外なことに一人暮らしの老人よりも同居家族がいる老人の方が高い**。上野はこのことを一人暮らしの老人が「さみしい」わけではない証拠としている。ここで根拠としている調査データは明記されていないが、福島県精神保健福祉センターの調査や秋田県の調査などいくつかのデータで、このような事実が確認されているので、「一人暮らしの高齢者よりも、家族と同居している高齢者の方が、自殺率が高い」ことは安定的な客観的事実のようである。

（1）この事実から、次のような述べることは適切か、それぞれ○×を付けなさい。

（　　）家族と同居している老人は、一人暮らしに変えた方が自殺の可能性が減る
（　　）いま家族と同居している老人は、いま一人暮らしの老人よりも自殺する可能性が高い
（　　）家族との同居は、老人が自殺する原因の一つである
（　　）「家族と同居すること」と「自殺」は、因果が逆の可能性がある

（2）「家族との同居」と「自殺」の間には、どんな見せかけの関係が発生している可能性があるか。（できれば複数の可能性を考えよう）

ヒント①自殺は女性より男性に圧倒的に多い（7割が男性）。
ヒント②現在の日本社会では、経済的に許されれば一人暮らしをする老人が多い。

（問題2）

あなたの友人が新聞記事「父親と長く過ごすほど我慢強い子に」（別資料）を読んで、次のように主張している。見せかけの関係の視点から、できるだけ簡単な言葉で（中学生でもわかる程度の言葉で）この主張を批判しなさい。

「新聞で見たけど、赤ちゃんの時に父親と過ごす時間が長かった子どもは、大きくなってから我慢強かったり、集中力が高かったりするらしいよ。ていうことは、法律で強制的に『父親は週に○○時間以上子どもと過ごすこと』とか決めれば、我慢強い子どもが増えるってことだよね。日本の将来を考えたら、そのくらいやっちゃった方がいいんじゃないかな。国が何年もかけてやった調査でわかったことなんだから、活かさないと。」

■補論：何でもグループ分けすればよいのか

見せかけの関係による混乱を避けるために、とにかく何でもかんでも細かくグループ分けして集計すればよいのかというと、それは間違いである。

ある関係が見せかけの関係である、という場合に大切なことは、第3の変数ＺがＸに因果関係上で先行していることである（図1のa）。Ｚ→Ｘという方向の因果だからこそ、Ｘの値を人為的に操作したとしても、Ｙの値が変化することはない（Ｘ→Ｚ→Ｙという流れはできないので）。一方で、Ｘの方がＺに因果関係上で先行しているときには、Ｘの値が変わればＺの値の変化を介してＹの値も変化する（図1のb）。したがって、（a）は見せかけの関係

だが、(b) は見せかけの関係ではない。第3の変数Ｚを加えることで、ＸとＹの関係の道筋をより詳しく示したことになる。2つの変数の共通の原因として見せかけの関係を作っている変数のことを先行変数［交絡変数］と呼ぶのに対して、2つの変数の間に入って関係を仲介する変数のことを**媒介変数**（intervening variable）と呼んで区別する。

ここで重要なことは、ＺとＸの因果の方向が逆であっても（Ｚが先行変数であっても、媒介変数であっても)、統計データが示す3重クロス表の形はまったく変わらない、ということである。つまり、媒介変数でグループ分けしても、先行変数でグループ分けした場合と同じように、元の2変数の関係性は消滅する。原因Ｘは媒介変数Ｚを変化させることを介して結果Ｙに影響するわけであるから、強引に媒介変数Ｚが同じ人々だけでグループを作れば、関係性が観察できなくなることは当然である。

したがって、何でもかんでも細かくグループ分けしてしまうと、見せかけの関係だけでなく、意味のある関係（媒介関係）までも、見えなくなってしまう。統計は、本質的に因果の方向を考えようとしない（むしろ積極的に避ける)。社会現象の因果を考えるための材料は、積極的に統計の外（理論や日常の観察）から持ち込まなければならない。

（a） Ｚ→Ｘの場合　　（b） Ｘ→Ｚの場合

図1　先行変数と媒介変数

■補論：実験と調査

一般的に、いわゆる「理系」では見せかけの関係への注意は大きな問題にならない。見せかけの関係は特に「文系」で問題になる。それは、理系の統計データが主に実験によって収集されるのに対して、文系の統計データが主に調査によって収集されるからである。

なぜ、実験だと見せかけの関係が問題にならないのか。実験では、何らかの効果を発揮すると仮定される刺激について、一方のグループにはその刺激を与え（実験群と呼ぶ)、もう一方のグループには刺激を与えない（統制群と呼ぶ)。これら2つのグループを比較することで、その刺激の効果を計測する。たとえば、ある薬が特定の病気に効果をもつかどうかを調べるために、一方のグループにはその薬を与え、もう一方のグループには与えない（偽薬を与える)。

このとき重要なことは、誰をどちらのグループに割り当てるかはランダム（無作為）に決められる、ということである。つまり、「X→Y」における「X」には、偶然以外の何者も影響を及ぼさない。したがって、XとYに共通の原因は存在せず、見せかけの関係は起こりえない。

これに対して、調査は人工的な刺激を与えるのではなく、人々のあるがままの現状を調べる。したがって、「X→Y」における「X」は、その人の自由意思や社会経済的な制約などから様々な影響を受け、見せかけの関係が発生する危険性に満ちあふれている。この面で

は、文系の計量社会学は、理系の実験統計よりも明らかに困難な問題に立ち向かわなければならない。

（問題3）

「出席と成績の関係」

(1) ある授業で各学生の出席回数と成績の関係を調べると、出席回数が多い学生ほど成績がよいことがわかった。つまり、「出席」と「成績」の間には正の相関がある。このことから、「成績を上げるためには、とにかく出席させることが一番大切だ」という意見に対して、別の人が「それは元々の学習意欲の違いによる見せかけの関係ではないか？」と疑問を唱えた。どういう意味か「見せかけの関係」という言葉を知らない人にもかわるように、具体的に説明しなさい。

(2) 実際の社会では、見せかけの関係と本当に意味のある因果関係が混じり合っていて、非常にややこしい。計量社会学の授業について、学生の「出席」「成績」「意欲」「理解」を調べたとすると、どんな関係性が現れると思うか、図式（矢印と＋－）を描いた上で、その図式で何を表したつもりか、文章で説明しなさい。

今日のポイント
①因果関係を証明する十分条件はないが、必要条件はある
・統計的関係性の存在
・時間順序が正しい
・見せかけの関係でないこと
②見せかけの関係と媒介関係を混同しないように注意
③見せかけの関係は、調査データを使う限り逃れられない問題
⇒とにかくいつも意識しなければならない

〈文献〉

久米郁男 2013 『原因を推論する: 政治分析方法論のすゝめ』 有斐閣.
Pearl, Judea 著、黒木学訳 2009 『統計的因果推論』 共立出版.（原書、*Causality*, 2000年出版）
上野千鶴子 2007 『おひとりさまの老後』 法研.（文庫版、2011、文春文庫）

※次回（7／11）の授業の終わりに最後の小テスト

小テストは、A4 用紙 1 枚を持ち込み可。相関と因果の違い、見せかけの関係の理解が中心。

4 回の小テストの合計点が 60 点以上ない場合、学期末試験を受験できない。
小テストが 60 点に満たなかった者は 15 回目の授業後に小テストの追試を受けること。（一部の小テストを受験できなかった者も含む）
小テストの合計が 85 点以上の場合は、学期末試験の得点を少しだけ加算する。

第14回「白書と政府統計」

■既存の統計資料の利用

計量社会学を実践するためには、当然、目的に見合った統計データを手に入れなければならない。データを得るためには自らが社会調査をして一次データを集める以外に、他人が集めたデータを再利用する方法もある。他人が集めたデータを**二次データ**（secondary data）と呼び、その分析を**二次分析**（secondary analysis）と呼ぶ。とくに、政府調査などの既存統計を二次データとして利用することは有益である。自ら調査をすることに比べれば極めてわずかな労力で信頼性の高いデータが利用できる。うまく活用しよう。

■内閣府の世論調査

一昔前まで、既存統計を利用するためには、図書館で分厚い冊子をめくり、必要な統計表を探し、たくさんの数字を書き写さなければならなかった（図書館のリファレンスコーナー）。しかし、現在は多くの統計資料がインターネットで公開されており、Excelデータでそのまま利用できるものも多い。

非常に便利な世の中だが､逆に､どこから手を付ければいいのかわからないこともある。初めて統計資料を探索する者は、まず「内閣府の世論調査」を眺めてみるとよいだろう。比較的身近なテーマについての短いアンケートデータが、大雑把な集計で公開されている（ほとんどの場合、単純な度数分布表のまま)。調査テーマは多岐にわたるので、いくつか興味のあるデータが見つかるに違いない。

○内閣府の世論調査
http://survey.gov-online.go.jp/

■基幹統計

内閣府の論調査は、親しみやすくおもしろいものの、かなり荒い集計データなので､突っ込んだ分析にはむいていない。より深い情報を手に入れるためには、もう少し「固い」統計資料を探したい。たとえば､**国勢調査**は5年に一度､日本に住むすべての人々を対象に行われる、もっとも固い統計資料である。固い統計資料は他にもたくさんあるが、特に重要な統計資料は**基幹統計**（2009年施行の統計法改正で指定統計から改名）と呼ばれ、国民はその作成に協力することが法律で義務付けられている。基幹統計は、ほぼ同じ調査内容で毎年（あるいは数年おきに）データが集められる**繰り返し横断調査［反復横断調査］**（repeated cross-sectional surveys）である。

↑国勢調査イメージキャラクター センサスくん。平成2年調査から使用。（※センサスcensusとは、国勢調査のように原則全戸を対象とする公的人口調査のこと）

基幹統計一覧（平成26年4月現在、55種）

内閣府	国民経済計算
総務省	国勢統計　住宅・土地統計　労働力統計　小売物価統計　家計統計 個人企業経済統計　科学技術研究統計　地方公務員給与実態統計 就業構造基本統計　全国消費実態統計　社会生活基本統計　経済構造統計 産業連関表
財務省	法人企業統計
国税庁	民間給与実態統計
文部科学省	学校基本調査　学校保健統計　学校教員統計　社会教育調査
厚生労働省	人口動態統計　毎月勤労統計　薬事工業生産動態統計 医療施設統計　患者統計　賃金構造基本統計　国民生活基礎統計　生命表 社会保障費用統計
農林水産省	農林業構造統計　牛乳乳製品統計　作物統計　海面漁業生産統計 漁業構造統計　木材統計　農業経営統計
経済産業省	工業統計　経済産業省生産動態統計　商業統計 ガス事業生産動態統計　石油製品需給動態統計　商業動態統計調査 特定サービス産業実態統計　経済産業省特定業種石油等消費統計 経済産業省企業活動基本統計　鉱工業指数
国土交通省	港湾統計　造船造機統計　建築着工統計　鉄道車両等生産動態統計 建設工事統計　船員労働統計　自動車輸送統計　内航船舶輸送統計 法人土地・建物基本統計

このような固い統計資料は、政府統計の総合窓口サイト「**e-Stat（イー・スタット）**」から入手できる。ただし、膨大な統計表があるため、慣れないと目的の情報のありかを探すだけで一苦労である。また、古い資料にはアクセスできない場合がある。

○政府統計の総合窓口　「e-Stat（イー・スタット）」

http://www.e-stat.go.jp/

■どんな既存統計があるのかを、知るためには？

e-Statは非常に便利であるが、そもそもどんな統計資料が存在するのかを知らなければ、目当てのものを見つけることは難しい。代表的な既存統計を知るための1つの方法は、**白書**を読むことである。白書は、官公庁のそれぞれが担当分野の動向をまとめて毎年発行する冊子である。白書には実にさまざまな統計資料が利用されており、何度も出てくるような統計は、その分野の代表的な統計資料であることがわかる。近年の白書は電子版がインターネットで公開されている。

○首相官邸から白書へのリンク　「資料集」→「白書」

http://www.kantei.go.jp/

○内閣府から白書へのリンク　「活動・白書等」→「白書、年次報告書等」

http://www.cao.go.jp/

また、国立国会図書館の「リサーチ・ナビ」は、もっと直接的に、代表的な既存統計を教えてくれる。いくらかは統計資料に慣れていないと統計の内容が想像しにくいが、非常によくまとめられているので、自分の関心のある分野について、じっくりと取り組んでみるとよい。

○国立国会図書館　「調べ方案内」→「リサーチ・ナビ」→「統計」

http://www.ndl.go.jp/

■素データの利用

二次データとして利用できるのは、ほとんどの場合、集計データであるが、素データのまま公開利用できるものもある。社会学では、2000年から1、2年おきに行われている繰り返し横断調査の**JGSS（日本版総合的社会調査）**などが学生でも利用できる（指導教員を通した申請が必要）。

素データとして公開利用できるデータは、ふつう**データアーカイブ**という機関を通して利用できる。調査の実施者は自分が集めたデータを広く有効活用してもうために、データアーカイブにデータを預け､データを必要とする利用者は､データアーカイブに申請して、データを貸してもらう。日本の社会科学分野での最大のデータアーカイブは、東京大学の**SSJデータアーカイブ**である。一部のデータは、学生でも利用できる。また、素データが利用できない場合でも全体の集計データは公開されている。一度、データを探索してみるとよい。

○JGSS

http://jgss.daishodai.ac.jp/

○SSJデータアーカイブ

http://ssjda.iss.u-tokyo.ac.jp/

■その他

ここで紹介した以外にも、世の中には多くの既存統計があふれている。市町村が行った調査や、大学、民間団体が行った調査もある。インターネットで検索できるデータもあれば、紙媒体だけで手に入るデータや、調査実施者だけが持っているデータもある。いずれにしても、自ら一次データを集めることに比べれば、既存統計を探すことの手間は、非常に小さい。テーマに合ったおもしろいデータがないか、よく探索してみることである。

おまけ：小学生～高校生向けの統計学習サイト「なるほど統計学園」
統計を利用する流れがわかりやすく、わりと使えるサイト

○統計局 統計学習サイト 「なるほど統計学園」
http://www.stat.go.jp/naruhodo/

今日のポイント
①基幹統計など信頼できるデータは積極的に二次分析に利用すべき
②データアーカイブを利用すれば、素データを自由に分析できる

第15回「まとめ」

■計量社会学とは

・計量社会学……積極的に数値（統計データ）を活用する社会学の一分野
　記述統計……データが持つ情報を要約して記述する（計量社会学Ⅰ）
　推測統計……一部のデータから調べてもいない全体を推し測る（計量社会学Ⅱ）
・数値を使う意義
　①数値を使えば、社会に実態を与えることができる（←誰も知らない社会をデータが語る）
　②数値を使えば、他人と協力できる（←客観的だから）

■計量社会学のデータ

・社会学のデータ＝量的データ＋質的データ
・計量社会学のデータ＝変数×ケース
　集めたままの細かいデータ＝素データ［ローデータ］
　グループでまとめたデータ＝集計データ
・測定尺度による変数の分類
　名義尺度……数字は名札代わり
　順序尺度……数字の順序だけに意味がある　→質的変数（計算できない変数）
　間隔尺度……数字の間隔が量を表す
　比率尺度……数字が2倍なら量も2倍　→量的変数（計算できる変数）
・確率論からの変数の分類
　離散変数……取りうる値がいくつかの点で決まっており、間はありえない変数
　連続変数……理論上、無限に細かい測定ができる変数

■記述統計の基本的な道具

	素朴な観察	統計量による要約
1つの変数の 分布を調べる　→	度数分布表 単純なグラフ	基本統計量 代表値（最頻値、中央値、平均値） ばらつき（範囲、四分領域、分散・標準偏差・変動係数）
2つ以上の変数の 関係を調べる　→	クロス表 散布図	関連性の統計量 相関係数 連関係数（ユールのQ、ファイ係数、オッズ比など） 順序相関係数（ガンマ、ロー、タウなど）

■1つの変数の分布を表わす（度数分布表）

・度数分布表は度数が重要。相対度数のみではダメ（少なくとも全体のnは示す）。

・階級の分け方の原則

①排他的で包括的

②階級幅は等しくする

③キリのよい数値の扱いに注意

■基本統計量の利用

・基本統計量……1つの変数の分布を要約する統計的な数量

代表値＋ばらつき

・どの代表値を用いるかは、長所と欠点をよく考えること（はずれ値の影響など）。

└→ 最頻値（モード）……とにかく度数の多いもの
└→ 中央値（メディアン）……ちょうど真ん中
└→ 平均値（ミーン）……全部足してケース数で割る

・どのばらつきの統計量を用いるかも、それぞれの意義をよく考えること。

└→ 範囲……最大値－最小値
└→ 四分領域……中央値から第3四分位までの距離と第1四分位までの距離の平均
└→ 分散……平均との偏差を平方したものの平均
└→ 標準偏差……分散の正の平方根
└→ 変動係数……標準偏差を平均で割ったもの

・補足的な基本統計量
- 歪度……左右対称からのゆがみ具合
- 尖度……きれいなベル型と比べたとがり具合

・Σの計算は「すべてのケースで同じ計算をして、結果を足し合わせる」だけ。

■2つの変数の関連性を表わす（クロス表、散布図）

・2変数の関連性を探るときには、クロス表が基本（全体をグループに分けて集計）。

・クロス表の相対度数は、適切なものを選ぶことが重要。

└→行％/列％/全体％がありうる

・量的変数同士の関係は、散布図でも読める。

■関連性の統計量の利用

・2つの変数の関連性も1つの数値で表せれば便利（基本統計量と同じ発想）。

・相関係数……散布図に表わされる量的変数同士の関係性を－1～＋1で表わす。

r＞0→正の相関（2つの変数が同じ方向に増減する）

r＜0→負の相関（2つの変数が別々の方向に増減する）

・連関係数……クロス表に表わされる質的変数同士の関係性を表わす統計量の総称。

（当然、量的変数もクロス表にすれば使える）

- 2×2のクロス表の場合→ユールのQ、ファイ係数、オッズ比
- より大きなクロス表の場合→クラメールのV
- 順序尺度変数の場合→スピアマンのρ（ロー）、グッドマンとクラスカルのγ（ガンマ）、ケンドールのτ_a（タウ）

■統計的な記述の実践

・PPDACサイクル……統計的に問題を解決する際のステップ。

Problem, Plan, Data, Analysis, Conclusion

問題、計画、データ、分析、まとめ

・「文章・グラフ・表」の選択を自覚的に。

・発見したパターンを文章にする際の注意。

変数間の関係性を記述することが基本。関係性の方向性（±）と強さを両方示す。

複雑な記述はGEEアプローチ（一般化、例示、例外の順序）に留意。

・統計分析≒作表

どんな分析をするかを考えることは、どんな表を作るか考えること。

作表を考えるためには、比較の軸を意識しなければならない。

度数分布表、基本統計量、クロス表、相関係数など単純な道具だけで十分効果的。

・実際のクロス表は縮約する必要がある場合が多い。

・グラフ作成の原則

①グラフは数値を比較する
②グラフは視覚情報を利用する
→代表的グラフで、どんなデータを比較するために、どの視覚情報を利用しているのか、注意

※そのグラフの大事な視覚情報を軽視すると、誤解を招くグラフを作成してしまう。

■見せかけの関係

・シンプソンのパラドックス

……2つの集団に分けた場合と全体で見た場合で関連性のあり方が異なる現象

・相関と因果は異なる

⇒「見せかけの関係」の仕組みを確実に理解する。

先行変数と媒介変数を区別。

・因果関係は証明できない（最低限の必要条件があるのみ）。

└→①統計的関係の存在
②時間順序が正しい
③見せかけの関係でない

■既存の統計資料の利用

・基幹統計を中心に、二次分析できそうなデータの雰囲気を知っておくこと。

・データアーカイブで素データの分析も可能なことを知っておくこと。

〈試験について〉

7月25日、60分間の試験

持ち込みすべて可（ただし、頭に入っていないと時間が足りなくなるはず）

電卓は携帯電話以外で（小テストと異なるので注意）