計量社会学 Ⅰ（2019年度春学期 担当:保田（やすだ））
2019.4.12

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（2014）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） 度数分布と代表値 （2） 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 2014『不透明社会の中の若者たち』関西大学出版部.
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 2018「計量社会学の考え方」酒井千絵・永井良和・間淵領吾編『基礎社会学 新訂第4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；**変量**［variate］もほぼ同じ意味）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きず、統計的な視点からは、「評定平均4.0以上」とか「GPA3.2」というような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」は質的変数なのか量的変数なのか。この変数のように、5段階や4段階で意見や意識の強さを測るやり方をとくに評定尺度（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）。評定尺度が質的変数か量的変数かは、やや大切な問題なので、後で改めて考えてほしい(問題2)。

■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（2017）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、きょうだいの中での生まれ順）
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、IQ）
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 2017 『入門・社会調査法〔第3版〕』 法律文化社.

※過去の配付資料はwebに置いています。紛失等は各自で補充を。
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 計量社会学 I 履修者の「数字の好き嫌い」

（a）2019年度

	度数	%
1 大嫌い	13	13.7
2 やや嫌い	30	31.6
3 ふつう	22	23.2
4 やや好き	26	27.4
5 大好き	4	4.2
計	95	100.0

（b）2018年度

	度数	%
1 とても嫌い	4	5.3
2 やや嫌い	19	25.0
3 ふつう	40	52.6
4 やや好き	9	11.8
5 とても好き	4	5.3
計	76	100.0

（c）2017年度

	度数	%
1 大嫌い	13	10.2
2 まあ嫌い	26	20.5
3 ふつう	40	31.5
4 まあ好き	41	32.3
5 大好き	7	5.5
計	127	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ては、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、その幅の中心の値を**中心点［階級値］**（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	19.2
約15分（8～22分）	15	636	35.6
約30分（23～37分）	30	319	17.8
約45分（38～52分）	45	177	9.9
約60分（53～67分）	60	182	10.2
約75分（68～82分）	75	54	3.0
約90分（83～97分）	90	49	2.7
98分以上	－	28	1.6
計		1789	100.0

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）。

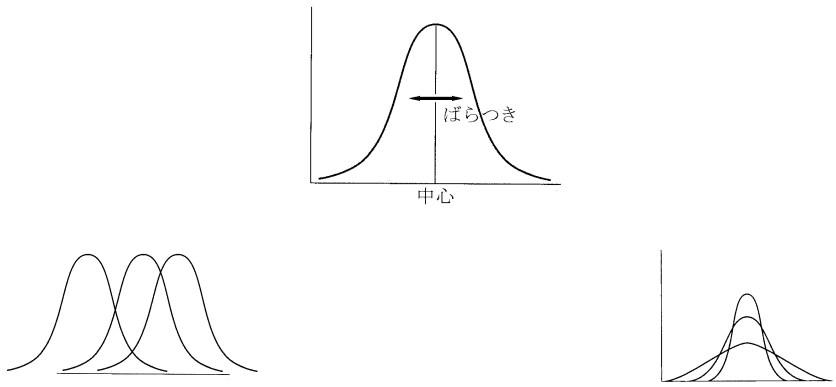
■基本統計量

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）…… 測定値を大きさ順に並べてケースをちょうど半々に分割する値
（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）…… $\bar{x}=\frac{1}{n}\sum x_i$

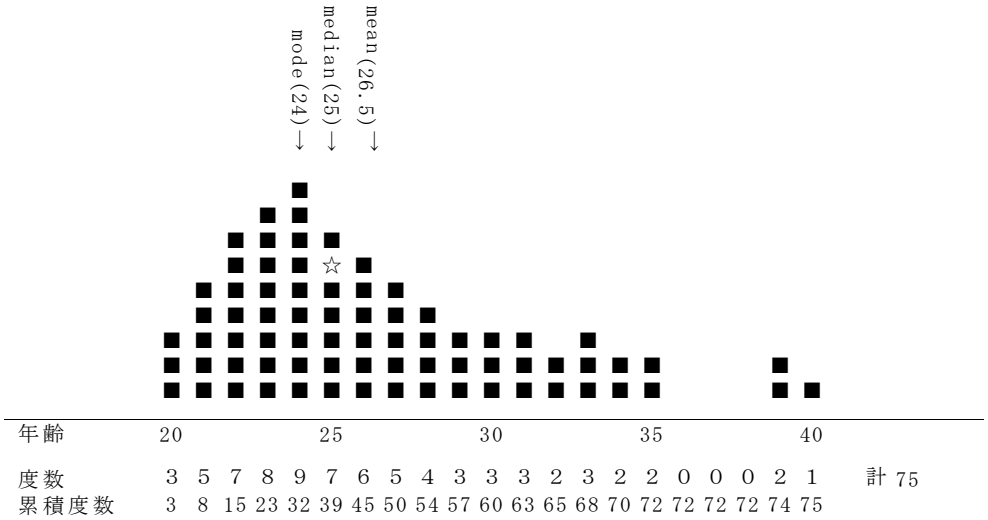


図2　代表値の模式図

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。

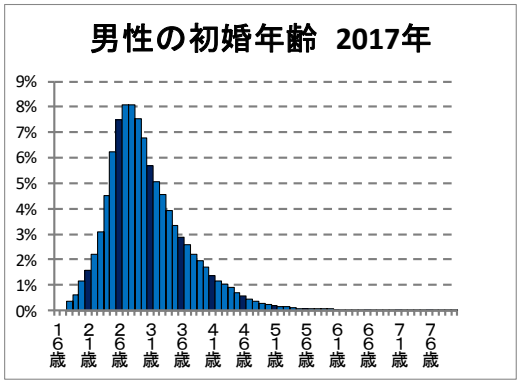
また、測定尺度の水準によって、用いることのできる代表値の限界があることにも、注意が必要である。たとえば、中央値は順序尺度でも算出できるが、平均値は数値の間隔が一定でなければ意味がないので、間隔尺度か比率尺度でなければ算出できない。それぞれの意味と限界を正確に理解して、用いる代表値を選ぶことが肝要である。

今日のポイント
①調査データ分析の第一歩は、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に注意（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

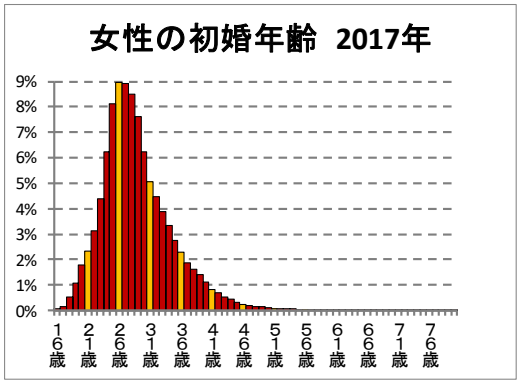
（問題）

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（a)（b)（c）のデータを間隔尺度とみなして、それぞれの年度の、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、人々の結婚行動にどのような社会的影響があるだろうか（あるいは、影響がないだろうか)。自分の予想を論じてみよう。

（参考資料：厚生労働省「平29年人口動態調査」から作成）



平均値 _____歳
中央値 _____歳
最頻値 _____歳



平均値 _____歳
中央値 _____歳
最頻値 _____歳

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

四分位偏差　$Q=\frac{Q_3-Q_1}{2}$

分散　$s^2=\frac{1}{n-1}\sum\left(x_i-\bar{x}\right)^2$　（nで割る場合もあるが、標本調査の分散は通常n-1で割る）

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｛5,8,12,19,21｝（単位：日）というデータが得られたとすると、範囲R＝21－5＝16である。

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。

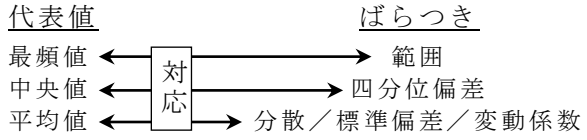


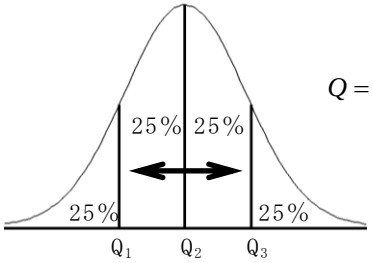
図1　代表値とばらつきの指標の対応

■四分位偏差

中央値とセットで用いられるのは**四分位偏差**（qrartile deviation）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）。

四分位偏差は、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　四分位数と四分位偏差

※四分位偏差と同じものを四分偏差や四分領域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（quartile range; 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」を求めなさい。

(2)男女別に通勤時間の「四分位偏差」を求めなさい。

(3)これらの指標で男女の通勤時間についてどのような違いを読み取ることができるのか。「範囲」や「四分位偏差」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。ただし、ばらつきの大きさを示す上で偏差の＋－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つの数値になる。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よるもので、現時点でよく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。

先ほどの5ケースのデータ｛5,8,12,19,21｝では、平均$\bar{x}=13.0$なので、

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

と計算できる。同様に、データが｛2,7,12,20,24｝だとすれば、平均は同じく$\bar{x}=13.0$になるが、分散を計算すると、

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

となる。後のデータの方がばらつきが大きいことが数値に反映されている（47.5＜82.0）。

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（アルバイト日数の分散は「47.5日^2」など）。そこで分散の正の√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。たとえば最初のデータの標準偏差は$s=\sqrt{s^2}=\sqrt{47.5}\fallingdotseq6.89$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。

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) ⇨ 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なさい（復習）。

(2) 上の表を使って、分散（n-1 で割る不偏分散の方）を計算しなさい。

(3) 標準偏差を算出しなさい。

(4) 算出した標準偏差をデータと照らし合わせて、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なさい。

(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。理由を説明しなさい

(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、男女それぞれの通勤時間の分布を、およそのグラフで描きなさい。

(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。

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（5／17）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） 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

■相対的に読む

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。社会調査のデータは、通常、相対的な視点からの読み取りに意味がある。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。男性の方が相対的に賛成しやすい（女性の方が反対しやすい）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　3人に2人が賛成！*」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、「三世代同居を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（ワーディングと呼ぶ）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」という相対的な関係性には、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）。

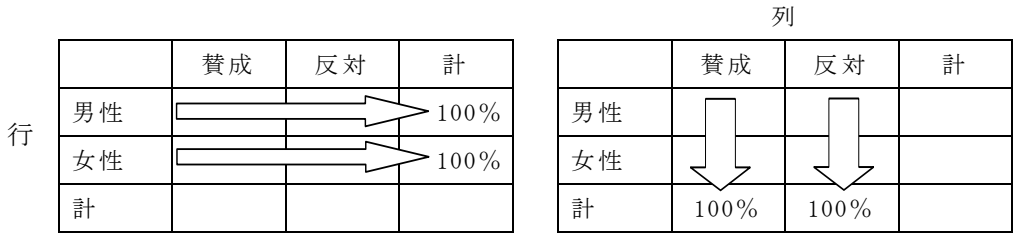


図1　行%と列%

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）

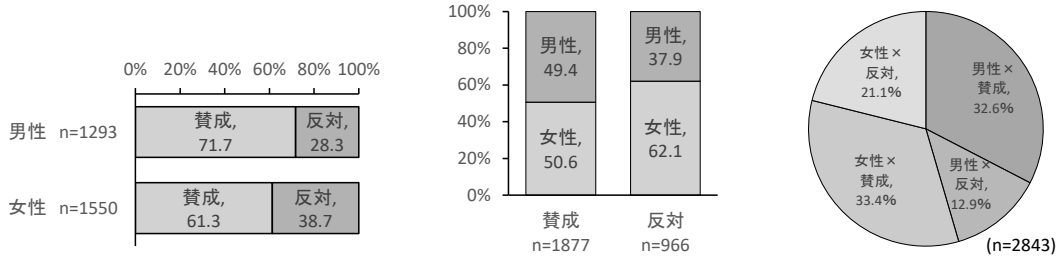
いまここでクロス表を作っている目的を思い出してみると、三世代同居への賛否の分布が男女でどう違っているのかを確かめることであった。つまり、男性の中での賛否の分布と女性の中での賛否の分布を比較して違いを見つけたいわけである。すると当然、必要な%の種類は行%であり、それ以外の列%、合計%は不要である。結局、例えば次のような形でクロス表を作成することが適切ということになる（表3)。

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、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、ここでは深入りしない)。

また、相対度数（%）は副次的な統計量にすぎないので、基本となる度数を必ず示すことも重要な注意点である。何らかの理由で各セルの度数を示さない場合でも、それぞれのグループの100%に相当する合計ケース数（n）だけは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）だけはグラフ脇に明記する。

（問題）

1. 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。

2. 以下の仮説を検証したいとき、どのようなクロス表が作成できればよいか、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通常は、原因を行側に配置し、結果を列側に配置することで、行%を比較する。

第6回「関係の読み方（2） 散布図と相関係数」

■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1）。

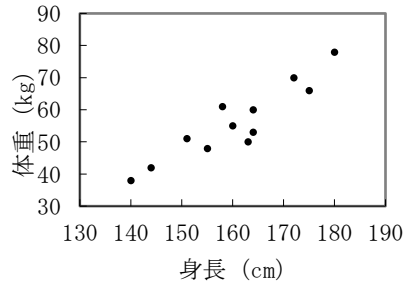


図1 散布図の例

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るが、残念ながら計量社会学における活用機会は限られる。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）。あくまでクロス表が基本と考えた上で、十分に多様な値を取り得る変数の場合（年齢、取得単位数など、あるいは複数の項目の合計得点、集計データにおける平均値や比率など）には、散布図も積極的に活用する、というぐらいの姿勢が適切であろう。

■「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年生の場合でも2年生の場合でも同じ程度の関係性が見られるのか、といった疑問が浮かんだとする。このことを確認するためには、1年生と2年生で別々に、散布図を作成して比較すればよい。

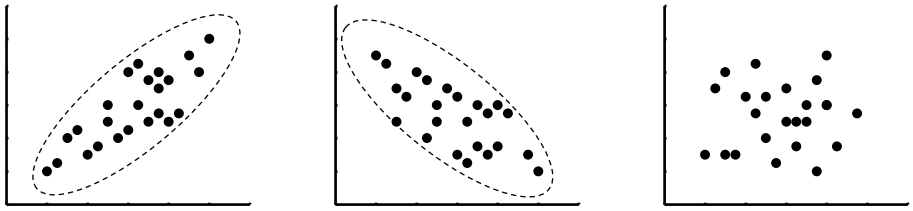
ところが、2学年ならまだよいが、4学年×13学部＝72個のグループで違いを調べようとか考えると、散布図を比較して読み取るだけでも大変である。そこで、自然な発想として、2変数の関係性の程度を「1つの数字」に要約できれば、比較が簡単になるはずだ、という考えが思い浮かぶ。度数分布表を読み取る代わりに、平均や標準偏差といった数値（基本統計量）に要約したのと同じことである。

そこで、使用される統計量が**相関係数**（correlation coefficient）である。2つの量的変数の関係を要約する統計量は他にも多数存在するが、相関係数が圧倒的によく用いられる（厳密にはピアソンの積率相関係数）。相関係数は、一般常識のレベルの統計量であり、その利便性と欠点を十分に理解していなければならない。

■相関係数の意味

相関係数は、2つの量的変数の関係性について、その「方向性」と「強さ」を1つの数値に要約する。理由にかかわらず2つの変数間に何らかの規則的な関係が見られるとき、2つの変数の間に**相関**（correlation）がある、という。「理由にかかわらず」というのは、その関係が本質的に意味のある因果関係かどうかとか、その関係にどんな意味を見出すかとか、そういったことをまったく問わずに、ただ単に客観的に2変数の間に統計的な関係が観察される、という意味で使う用語だということである。

相関係数では、相関の中でもとくに代表的な関係である直線的な関係傾向を数値で表す。つまり、量的変数XとYの間で、図2（a）（b）のような傾向の関係をもつ場合である。（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2　相関関係

相関係数は、通常、記号「r」で表し、必ず－1から＋1の間の値をとる。相関係数は、関係の「方向性」を値の±でそのまま表現し、正の相関を持つ場合には＋の値（r＞0）、負の相関を持つ場合には－の値（r＜0）となる。さらに、その関係の「強さ」を値のサイズで比較できる。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。ほとんど関係が見られない場合には0に近い値になる。図2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。

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。

（問題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

（問題2）

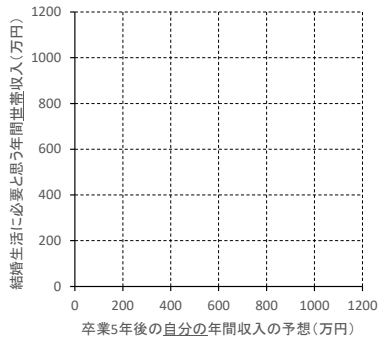
(1)次のアンケートを男女10人ずつに行ってみて､2変数の散布図を男女別に描きなさい。

Q1：卒業の5年後、あなたは自分の年間収入が何万円になっていると予想しますか。
（税金を抜く前の額面通りの収入）

Q2：結婚生活を始めるためには「夫婦合わせて」年間どのくらいの世帯収入が必要だと思いますか

(2) 男女それぞれについて、散布図から相関係数の値を予想しなさい。

(3) 実際に、男女それぞれの相関係数を計算しなさい。



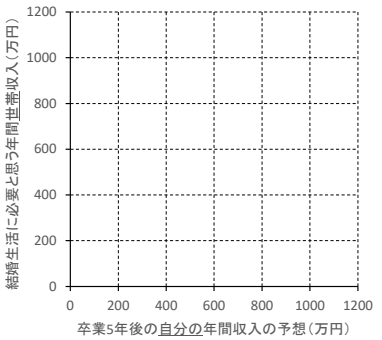
【男性】

(予想)

r＝＿＿＿

(実際)

r＝＿＿＿



【女性】

(予想)

r＝＿＿＿

(実際)

r＝＿＿＿

（問題3）

下の表は複数の変数の間での相関係数をまとめたものである（相関表と呼ばれる)。「理想の恋人として何を重視するか」という質問に対する各項目の得点をX_1～X_5で表している。この表から読み取れる事柄として正しいものに○、誤っているものに×を付けなさい。

（　　）顔の良さを重視している人ほどスタイルも重視している傾向がある
（　　）性格を重視することと顔の良さを重視することは、ほとんど関係がない
（　　）スタイルを重視しない人は、価値観を重視する傾向がある
（　　）X_1とX_2の相関がもっとも強く、X_3とX_4の相関が2番目に強い

	X_1 顔の良さを重視	X_2 スタイルを重視	X_3 頭の良さを重視	X_4 性格の良さを重視	X_5 価値観を重視
X_1 顔の良さを重視		.389	-.005	-.273	-.252
X_2 スタイルを重視	.389		.044	-.137	-.328
X_3 頭の良さを重視	-.005	.044		.046	-.102
X_4 性格の良さを重視	-.273	-.137	.046		-.009
X_5 価値観を重視	-.252	-.328	-.102	-.009	

■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。

$$\text{相関係数 } r = \frac{\dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1}}{\sqrt{\dfrac{\sum(X-\bar{X})^2}{n-1}}\sqrt{\dfrac{\sum(Y-\bar{Y})^2}{n-1}}} = \frac{X\text{とY}\text{の共分散}}{X\text{の標準偏差}\cdot Y\text{の標準偏差}}$$

$\left(\begin{array}{l}\bar{X}\text{は}X\text{の平均値}\\ \bar{Y}\text{は}Y\text{の平均値}\\ n\text{は全回答者数}\end{array}\right)$

一見複雑だが、それほどややこしい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つの変数の2次元的な散らばり具合を示す。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。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、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。

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

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

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

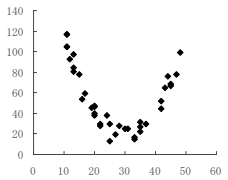
$$=\frac{1}{4}(-2.7-2.4-1.8+0.1-3.6)=-2.6$$

③相関係数を算出

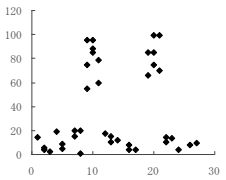
$$r=\frac{-2.6}{6.86\times0.49}=-0.77$$

■相関係数の注意点

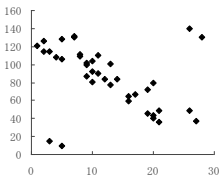
相関係数は非常に頻繁に用いられるが、万能ではない。変数間の直線的な関係性しか表していないので、曲線的な関係など別の規則性には適切に反応しない（図3のa、b)。また、外れ値の影響を非常に受けやすい（図3のc)。外れ値の影響を受けやすい欠点は、平均値を利用するタイプの統計量が持つ宿命である。



（a） r ＝-0.32



（b） r ＝0.15



（c） r ＝-0.36
※外れ値がなければ、-0.91

図3　相関係数に反映されない関係性のパターン

今日のポイント
①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。
②散布図に表される関係性を1つの数値に要約するのが相関係数。
+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） 小休止（復習と補足）」

■2つの質的変数の間で関係（相関）を読む

クロス表を作って、適切な%を読み取ることが基本。

行%、列%、全体%

（p.20の問題1の再掲）

		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」だったら……
（既婚者村と、未婚者村で、ほしい子どもが違うか比べたい）　→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」だったら……
（男の子ほしがり村と、女の子ほしがり村で、婚姻状態が違うか比べたい）　→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

■2つの量的変数の間で関係（相関）を読む

散布図を作って読み取ることが基本。

正の相関、負の相関の区別。

■量的変数の「関係（相関）」を1つの統計量に要約する

相関係数に要約することが基本。

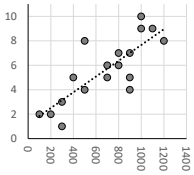
相関係数（r）は、2つの変数の関係（相関）を「方向性」と「強さ」に絞って要約する。

①関係の方向性（→±で表わす）

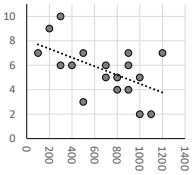
Xが増えれば、Yは増えるのか、それとも減るのか

仮想データ

Xが年収（万円）
Yが幸福感（10点満点）



r＝0.80
（正の相関）

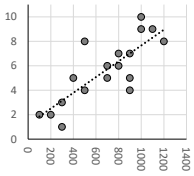


r＝－0.55
（負の相関）

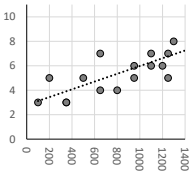
②関係の強さ（→数値のサイズが±1にどれだけ近いかで表わす）

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（傾きの角度）、ではない

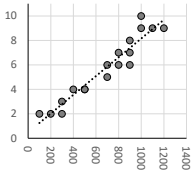


r＝0.80

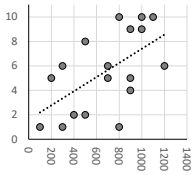


r＝0.81

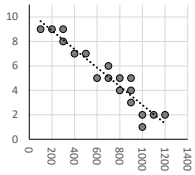
はっきり予測できるというのは、比例関係（直線）にどれだけ近いかということ



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）

相関係数の計算はやや面倒だが、難しいわけではない。

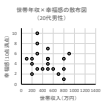
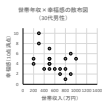
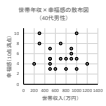
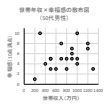
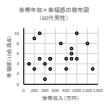
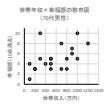
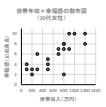
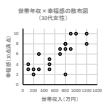
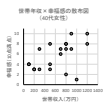
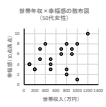
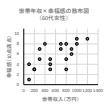
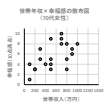
計算を間違えないことよりも、「XとYの共分散」を「Xの標準偏差」「Yの標準偏差」で割っていることの意味を理解することの方が大切。

■関係（相関）を比べる

相関係数に要約すれば、関係（相関）の違いを簡単に比較することができる。

例）一般に、「年収が高い方が幸福感が高い」という相関がある。この年収と幸福感の相関は、性別や年齢層によってどう違うのだろうか？

「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。

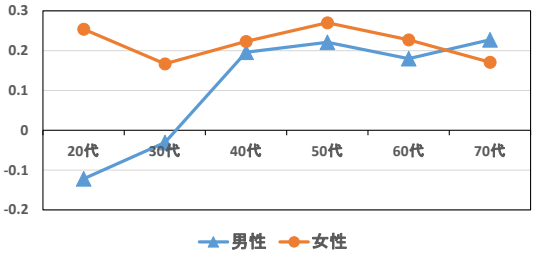
	20 代	30 代	40 代	50 代	60 代	70 代
男性						
女性						

※この散布図はイメージです

↓

各グループで、年収と幸福感の相関を相関係数に要約すれば、違いが一目瞭然

	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（世帯年収と幸福感の相関係数：男女・年齢層別の変化）

（問題）

これまでの話の流れから予想すると、次回の講義では何が扱われると予想されるでしょう？

第8回「関係の読み方（4） クロス表の連関係数」

■大きな話の流れ

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)。まず大切なことは各変数（各調査項目）の度数分布表をよく観察することである（第3回)。しかし、実際には多くの度数分布表の観察は大変なので、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3、4回)。

次に、2つ以上の変数の関係（相関）を読み取る話である。社会調査のデータでは、2変数の関係は基本的にクロス表で読み取る（第5回)。一方、利用場面は限られるが、量的変数同士の関係は散布図でも読み取ることができる。さらに、散布図に表れる関係は「相関係数」という1つの統計量に要約できる。相関係数を比較すれば、「2変数の関係性の比較」が容易になる（第6回)。これは代表値やばらつきの指標で「1変数の分布の比較」が容易になるのと同じことである。

ここで､当然の発想として､「クロス表に見られる関係性も何らかの統計量で要約できるはずだ」と考なければならない。

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	基本統計量 代表値（最頻値、中央値、平均値） ばらつき（範囲、四分位偏差、分散・標準偏差・変動係数）
2つの変数の 関係（相関）を 調べる→	クロス表 散布図	関係性（相関）を表わす統計量 連関係数（ユールのQ、ファイ係数、オッズ比など） 相関係数

■2×2のクロス表における3つの連関係数

クロス表に表れる2変数の相関は、association（「連関」「関連性」と訳す）と表現されることが多い。クロス表の相関（連関、関連性）を1つの統計量に要約したものを総称して**連関係数**（association coefficient; coefficient of association）あるいは関連性の統計量と呼ぶ。連関係数には複数の種類があるが、もっとも基本的な2×2のクロス表についてはとくによく考えられており､次の3つがよく用いられる。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、それぞれ下のように算出される（図1)。

a	b
c	d

ユールのQ $\quad Q=\dfrac{ad-bc}{ad+bc}$

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

オッズ比 $\quad OR=\dfrac{ad}{bc}$

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性（相関）がまったくない状態の定義は共通している。それは、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から、a:b=c:dのときといえる。この式を変形すると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。散布図や相関係数と同じように、前者を正の相関（関連）、後者を負の相関（関連）と呼ぶ※。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、相関係数とまったく同じ読み方ができる。

※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が多いことを正の相関と呼ぶことにする。

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2）。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2　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

$$\text{ユールのQ} \qquad Q = \frac{ad - bc}{ad + bc} = \frac{927 \times 600 - 366 \times 950}{927 \times 600 + 366 \times 950} = 0.231$$

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

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。これは両者の間で「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。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。

（問題1）

クラス内のアンケートで何らかの2×2のクロス表を作成し、ユールのQとファイ係数を算出し、意味を説明しなさい。

			計
	（　　%）	（　　%）	
	（　　%）	（　　%）	
計			

ユールのQ　　Q＝

ファイ係数　ϕ＝

■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。いまの例では、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）。同様に、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$\frac{2.53}{1.58}=1.60$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。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。

オッズ比の長所は「見込みが○倍」という具体性をもつことである。一方、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点で、抽象的だが扱いやすい。用途に応じてこれらを使い分けなければならない。

■連関係数の値を比較する

相関係数と同様に、連関係数も複数の数値を相対的に比較するときにこそ意味がある。たとえば、いま例にあげている三世代同居への賛否について、「男性の方が賛成しやすい」

ことがわかったが、この関係性はどの年齢層でとくに強く見られるのか、相対的に比較してみよう。改めて、年齢層別に複数のクロス表を作成して、ユールのQやオッズ比などの連関係数を比較すればよい。複数のクロス表をただ慎重に読み取るより、確実で簡便である。

（問題2）

実際に上の関心を満たすためにクロス表を作成した（JGSS-2000のデータ）。これをもとにして20～70代のそれぞれについてユールのQを算出し、どの年齢層で「男性の方が同居に賛成する」傾向が強いのか確かめなさい。

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

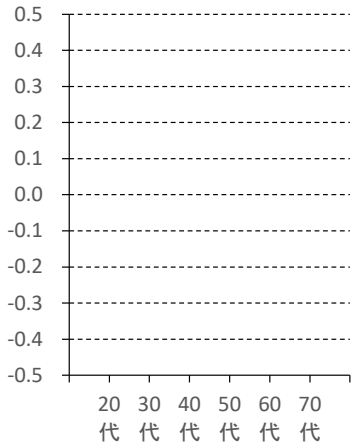
度数

age10 年齢10歳刻み			op2gnr Q12 三世代同居観 1 望ましい	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	
30代	
40代	
50代	
60代	
70代	

↓



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使い分けに注意

具体性を取るか（オッズ比）、抽象性を取るか（ファイ係数、ユールのQ）

最大の関連を厳しく捉えるか（ファイ係数）、緩やかに捉えるか（ユールのQ）

※次回（6／14）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 5～8 回の内容について、クロス表の作り方と読み方、相関係数・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。

計量社会学 I（2019年度春学期 担当:保田（やすだ））
2019.6.14

第9回「記述の実践（1） 比較のプランと作表」

■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（クロス表、散布図、連関係数、相関係数）

を学習した。これらの方法を駆使すれば、たいていの分析目的は果たすことができる。

この授業の後半は、個別に学習してきた技術を組み合わせて、統計データによる社会の記述の「実践」に触れてもらう。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。

■統計分析≒作表

実践的な分析作業の一番大切な枠組みは、**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」で表現するとしても、その元となる肝は「表」だからである。

作表のイメージは調査設計段階でも重要である。調査を設計するときには、「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で測定できるか」といったように作表からさかのぼって考える。つまり、作表のイメージがなければ、そもそもデータを集めることもできない。

■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、これらを確実に扱えるようになろう。下のように、目的に沿った作表のイメージ化を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）

「自宅生は学園祭にあまり参加しないイメージだが、本当にそうだろうか」（目的）
→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、

二変数の関係を相関係数や連関係数で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の表で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）。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・トレーニングを積んでほしい。

■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なければならない。たとえば、「未婚男性の生活満足度が低い」という分析結果は、1つの数値で示せると思うかもしれない（未婚男性の生活満足度は5点満点で平均値が2.2点、等）。しかし、1つの数値では何と比較して「低い」と判断しているのか不明なため、メッセージの説得力は弱い。

実際、比較の視点は多様に考えられ、比較対象によってメッセージはまったく異なってくる。未婚「女性」と比べて低い、「既婚」男性と比べて低い、「10年前の」未婚男性と比べて低い、「外国の」未婚男性と比べて低いなどである。いずれにしても、1つの数値だけではなく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必要がある。複数の数値を比較のために並べると表になる。つまり、必然的に「作表」につながる。

■補足：クロス表の縮約

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）。たとえば、選択肢が4つや5つの評定尺度は一般的であるが、そうして得られた2つの変数の関係性を4×4～5×5のクロス表で表現すると、多くのセルで読み取りに骨が折れる。たとえば、2×2～3×3のクロス表に縮約することを考えるべきである。

表1　女性の仕事の子どもへの影響×仕事への賛否のクロス表（縮約前）

		夫に十分な収入がある場合には、妻は仕事をもたない方がよい				
		賛成	どちらかといえば賛成	どちらかといえば反対	反対	計
母親が仕事をもつと、小学校へあがる前の子どもによくない影響を与える	賛成	6 *23.1%*	8 *30.8%*	8 *30.8%*	4 *15.4%*	26 *100.0%*
	どちらかといえば賛成	11 *13.3%*	36 *43.4%*	28 *33.7%*	8 *9.6%*	83 *100.0%*
	どちらかといえば反対	6 *5.6%*	30 *28.0%*	61 *57.0%*	10 *9.3%*	107 *100.0%*
	反対	0 *0.0%*	6 *9.4%*	33 *51.6%*	25 *39.1%*	64 *100.0%*
	計	23 *8.2%*	80 *28.6%*	130 *46.4%*	47 *16.8%*	280 *100.0%*

注：データはJGSS-2010の20～30代男性

表2　女性の仕事の子どもへの影響×仕事への賛否のクロス表（縮約後）

		夫に十分な収入がある場合には、妻は仕事をもたない方がよい		
		賛成	反対	計
母親が仕事をもつと、小学校へあがる前の子どもによくない影響を与える	賛成	61 *56.0%*	48 *44.0%*	109 *100.0%*
	反対	42 *24.6%*	129 *75.4%*	171 *100.0%*
	計	103 *36.8%*	177 *63.2%*	280 *100.0%*

注：データはJGSS-2010の20～30代男性

今日のポイント
①分析作業の枠組みは、どんな「作表」をするかに集約される
②データを集める「前に」作表のプランを立てるイメージ・トレーニングが大切

（問題）

右のような質問紙調査を120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 あまり重要でない　2 少しは重要　3 ある程度重要　4 とても重要　5 極めて重要

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

	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すべてに○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すべてに○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10回「記述の実践（2） グラフの描き方」

■グラフの必要性

視覚に訴えるグラフは、数値が持つ情報を伝えるための強力な武器になる。とくに、多くの数値からパターンを読み取る場合には、表のままよりも格段に情報が伝わりやすい。何より、情報が視覚化されるグラフ作りは単純に楽しい。

グラフ作成にあたっては、次の2つの目的を見失わないようにしなければならない。

・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1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1 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

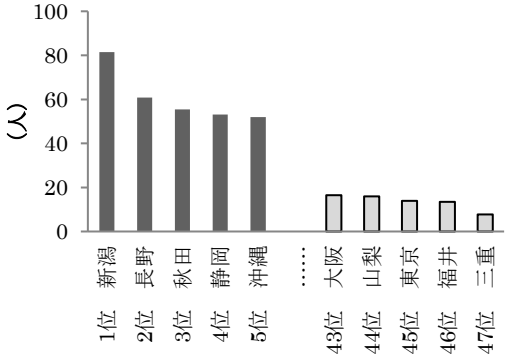
棒グラフは何らかの数量の大きさを比較するために、棒の長さでその数量の大きさを表したものである。比較するものは、度数の他に相対度数（%）や比率尺度の変数の*注 平均値など、その絶対的な大きさに意味があるものであれば何でもよい（図1）。〔※注：間隔尺度の変数は数値の絶対量を比べられないので、棒グラフはおかしいことに注意〕

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の傾きである。傾きの角度を比較することで、変化の傾向が読み取れる（図2）。

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3、4）。

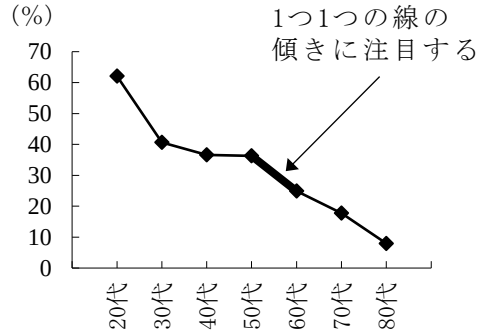
ヒストグラムは、棒グラフの棒と棒の間の隙間をなくしただけに見えるが、その意味合いは全く異なる。棒グラフがその長さに意味があるのに対して、ヒストグラムはその「面積」に意味がある。ヒストグラムの柱と柱がくっついているのは、隣の区分と数量が連続的に繋がっているからである。したがって、隣あった柱の面積を合わせて、より広い範囲の度数を一目で把握することもできる（図5）。

図1 人口10万人あたりのバスケ競技者人口



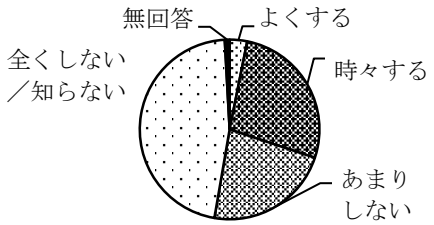
出所：2009年度「バスケットボール競技者登録者数」
（財団法人日本バスケットボール協会）

図2 世代によるカラオケをする割合の変化



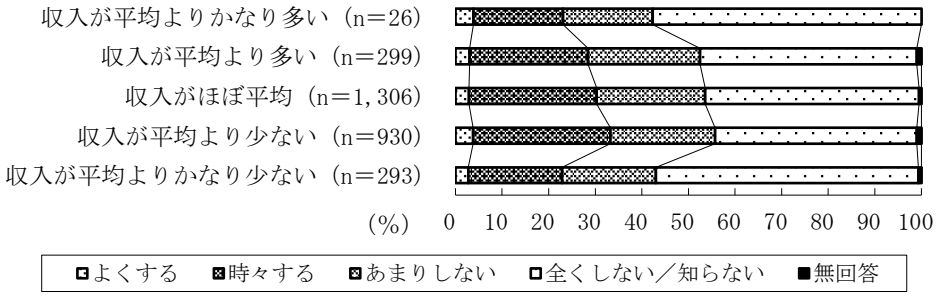
出所：JGSS-2000

図3 宝くじを買う頻度？



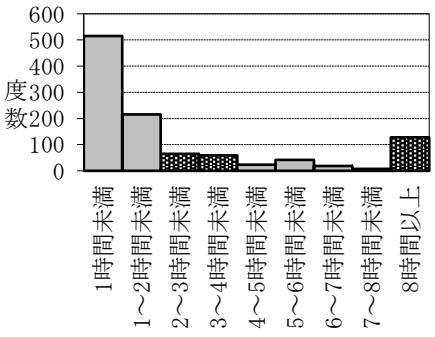
出所：JGSS-2000

図4 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5 ペットを飼っている人が1日にペットと過ごす時間

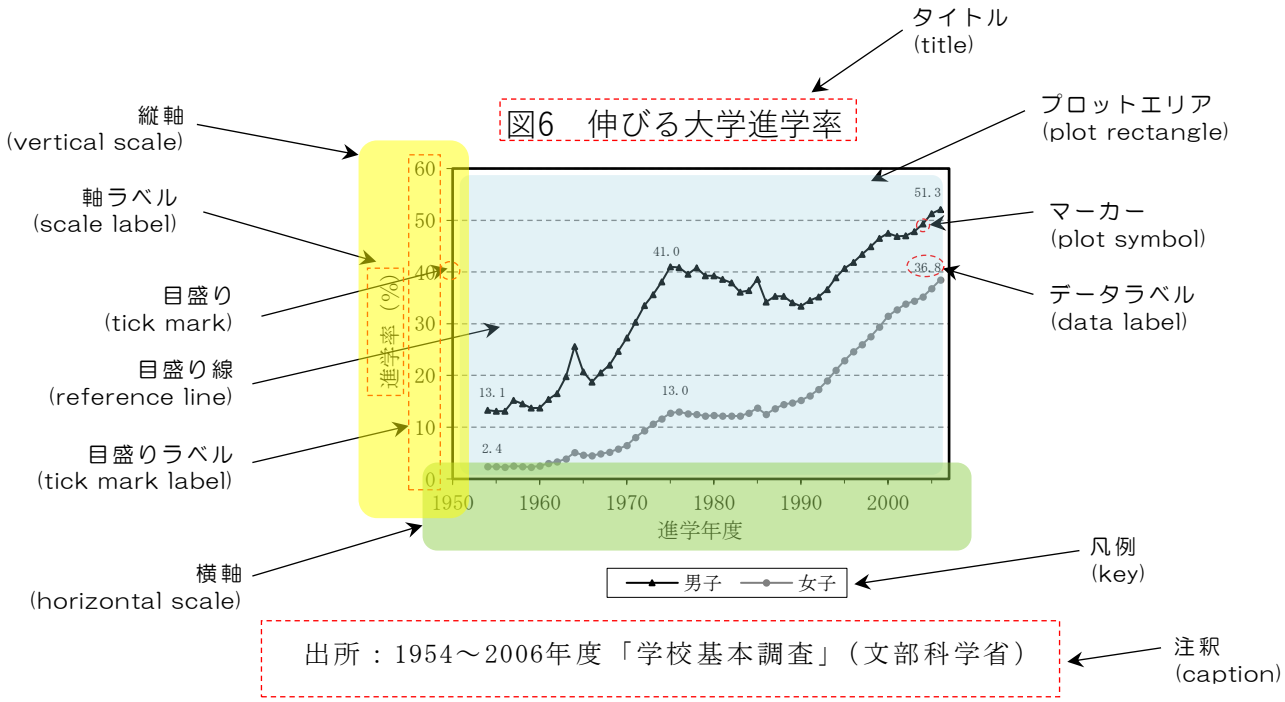


出所：JGSS-2000

■グラフのパーツ

パーツに注目すると、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。

・必ず図表番号（図4、表5など）とタイトルを付ける。
・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、できるだけグラフ本体以外のもの（凡例など）を含めない。
・1つのグラフの中に複数の比較軸を複雑に持ち込まない。
・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。たとえば、図7のように目盛りが0から始まっていない棒グラフは不適切である。なぜならば、棒グラフにとって数値の大きさを表す命であるはずの「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から始まっていなくても問題はない。折れ線グラフの命は線の高さではなく「線の傾き（の相対的比較）」だからである。折れ線グラフではむしろ、全体的な傾きがおよそ45度程度になるように目盛りを調整すると、線の間で傾きを比べやすく望ましいとされている（人間は45度付近の角度をもっとも敏感に感知できる）。

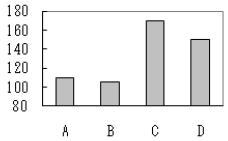
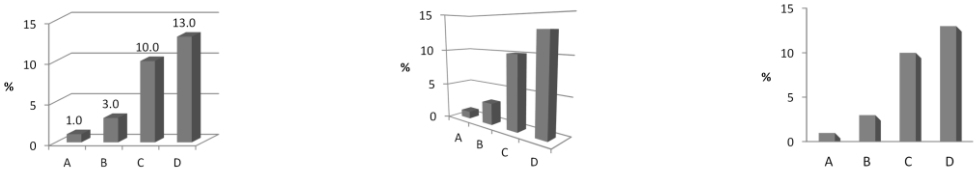


図7　不適切な棒グラフ（打ち切り）

また、図8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8　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。早坂 (2015) はExcel 2013までの対応だが、Excel 2016でも基本的には同じである。

〈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 2015 『EXCELグラフ作成 ［ビジテク］ データを可視化するノウハウ』 翔泳社.

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」「どんな視覚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）

※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。7割近くの人が5点満点での「5 極めて重要」という回答で、「1」や「2」という人はいない。

表A　お金の重要度の度数分布表

	度数	%
重要度 1	0	0.0
2	0	0.0
3	7	8.3
4	19	22.6
5	58	69.0
計	84	100.0

（2）表Bのクロス表でわかるとおり、大阪を「住みやすい」と考える割合は男子学生の方がやや高い。男子学生も女子学生も、大阪が「住みやすい」と○を付けている割合は非常に高いものの、約13%の差があり、相対的な違いはある。

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	34 79.1%	9 20.9%	43 100%
女子	27 65.9%	14 34.1%	41 100%
計	61 72.6%	23 27.4%	84 100%

（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。単純に大都市を楽しいと感じるかどうかで両方の評価が決まっているように見える。

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	29 85.3%	5 14.7%	34 100%
○無し	20 40.0%	30 60.0%	50 100%
計	49 58.3%	35 41.7%	84 100%

（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。1位～5位までの「お金」～「目標を立てること」までは、いずれも4.3以上の高い平均値で、他との差が大きい。

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	4.61
(l)友人関係	4.43
(e)十分な睡眠	4.36
(m)就職の見込み	4.33
(a)目標を立てること	4.32
(f)よい食事	4.05
(k)一人の時間	4.04
(h)趣味	4.00
(d)家族からの支援	3.96
(c)授業外の勉強	3.75
(b)授業での勉強	3.72
(j)アルバイト	3.60
(i)資格の取得	3.54
(n)部活・サークル	3.33

（n=84）

（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。項目間で極端な違いはないが、十分な睡眠や部活・サークルについて、満足度の格差がもっとも大きい。

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
(e)十分な睡眠	1.26
(n)部活・サークル	1.26
(h)趣味	1.21
(f)よい食事	1.19
(i)資格の取得	1.16
(j)アルバイト	1.16
(g)お金	1.13
(l)友人関係	1.12
(m)就職の見込み	1.09
(a)目標を立てること	1.04
(c)授業外の勉強	1.01
(k)一人の時間	0.99
(d)家族からの支援	0.95
(b)授業での勉強	0.94

（n=84）

(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、趣味への満足度が高いという明らかな傾向が読み取れる。趣味が重要と考えている学生は、7割以上が現状の趣味に満足している。

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)	42 72.4%	16 27.6%	58 100%
低い(1･2･3)	11 42.3%	15 57.7%	26 100%
計	53 63.1%	31 36.9%	84 100%

(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、「家族からの支援」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、「目標を立てること」は弱いものの負の相関であり、それ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
(d)家族からの支援	0.498
(h)趣味	0.382
(l)友人関係	0.335
(n)部活･サークル	0.305
(j)アルバイト	0.293
(b)授業での勉強	0.242
(f)よい食事	0.238
(c)授業外の勉強	0.159
(m)就職の見込み	0.063
(e)十分な睡眠	0.031
(g)お金	0.008
(k)一人の時間	0.004
(i)資格の取得	−0.098
(a)目標を立てること	−0.179

(n=84)

(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。「ごみごみしている」「息苦しい」というイメージは男子の方が強く、「好ましい」「楽しい」「華々しい」といったイメージは女子に強いことがわかる。

ただし、選択率で比較してしまうと、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も小さくなりやすくなってしまう。

その意味では、各クロス表での関係性を純粋に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と似通っているが、「悲しい」の評価など、異なる面もある。

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)
1 ごみごみしている	69.8	43.9	25.9
8 息苦しい	58.1	41.5	16.7
3 活気がある	53.5	43.9	9.6
10 かっこいい	39.5	34.1	5.4
12 すばらしい	18.6	14.6	4.0
11 悲しい	2.3	0.0	2.3
14 さみしい	9.3	7.3	2.0
15 恥ずかしい	0.0	0.0	0.0
9 安らか	2.3	2.4	−0.1
4 怖い	41.9	43.9	−2.0
13 忙しい	72.1	75.6	−3.5
6 住みやすい	11.6	17.1	−5.4
2 好ましい	18.6	29.3	−10.7
5 楽しい	34.9	48.8	−13.9
7 華々しい	51.2	68.3	−17.1
n	43	41	

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(男子)と各項目の関連性(ユールのQ)
11 悲しい	1.000
1 ごみごみしている	0.494
8 息苦しい	0.325
3 活気がある	0.190
12 すばらしい	0.143
14 さみしい	0.130
10 かっこいい	0.115
9 安らか	−0.024
4 怖い	−0.042
13 忙しい	−0.091
6 住みやすい	−0.220
5 楽しい	−0.280
2 好ましい	−0.288
7 華々しい	−0.346
15 恥ずかしい	−

(n=84)

（問題）

1. 上の表A～I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。

2. 右の表を折れ線グラフで表現しなさい。ただし、グラフの各パートが完全に整っている隙のないグラフを描くこと。

表2　健康関係の重要度の平均値の推移

調査年 / 項目	2013	2014	2015	2016	2017	2018	2019
十分な睡眠	3.76	3.88	4.34	4.13	4.09	4.50	4.36
よい食事	3.6	3.75	4.11	3.87	3.84	4.28	4.05

注：計量社会学1の「実習用アンケート」から作成（2013～2019年）

第11回「記述の実践（3） PPDACサイクル」

■PPDACサイクルとは？

前の2回で計量的な記述においては、①作表を意識した分析プランと、②適切なグラフ表現が重要なことを学習した。また、この講義の前半では③各種の分析の道具立て（統計量）について学習した。いま、改めてこれらが統計データを用いた一連の問題解決の流れの中でどのように結びついてくるのかを確認しよう。

統計的な証拠に基づいて何らかの問題解決を探る手順は、**PPDACサイクル**という枠組みで整理できる※。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）、基本的な流れが端的にまとめられている（図1）。

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

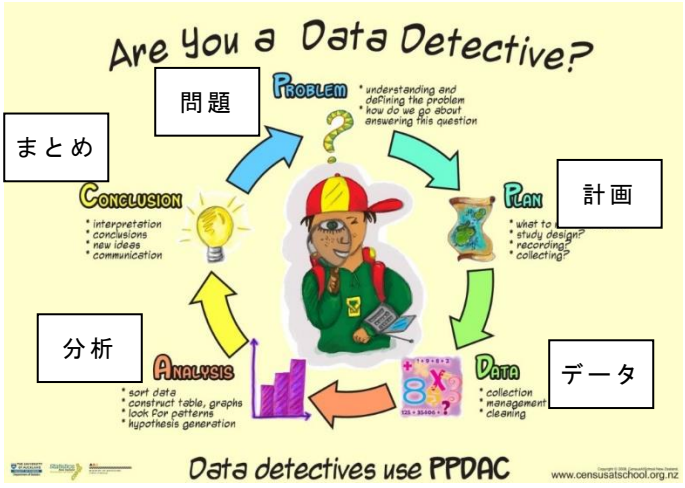


図1 PPDACサイクル

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。大雑把に言えば次のような流れになる。

[P]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。

[P]どうすればその疑問が解けるのか、計画を立てる。どこからどのようなデータを取ってきて、どのような作表を目指すのか、大雑把な全体像を描く。

[D]分析に必要なデータ収集を行う。あるいは、すでに存在するデータから二次利用できるものを入手する。

[A]データを計画どおりに分析して、数値の比較やパターンの読み取りをおこなう。

[C]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。

■最初と最後を意識する

一連の流れを一回限りの問題解決で終わらせずに「サイクル」として継続させることは非常に大切である。つまり、C（まとめ）の段階で解答が出なかった点や新たに生じた疑問を改めてP（問題）として探求を続けるということである。とくに計量社会学では「他人と協力できる」という特徴を生かすために、この意識が殊更に重要になる。

そのためにはPPDACサイクルの最初（問題）と最後（まとめ）に意識的に力を入れなければならない。具体的には、まず「問題」を疑問文の形で明確に特定すべきである。たとえば、「研究のテーマは〈大学周辺のゴミのポイ捨て問題〉です」といっただけでは曖昧である。「大学前の通りでは、いつポイ捨てが多く発生しているのか？」「男子学生の方が多くポイ捨てをしているのはなぜなのか？」といった具体的な**リサーチ・クエスチョン**（research question）を疑問文の形で明確にする。

一般的に言えば、扱う問題について「実態」「原因」「解決策」のどの段階に注目するのかをはっきりさせることが有効である。たとえば、ゴミのポイ捨て問題の場合、実際にどのくらいポイ捨てがあるのか（実態）、なぜポイ捨てがなくならないのか（原因）、ポイ捨て対策としてゴミ箱を増やすことは有効なのか（解決策）といった水準がある。

また、サイクルの最後（まとめ）の段階においても、計量的研究には独特の困難がある。端的に言ってしまうと、多くの人々は数字を見るのが嫌いであり、ただ数値を示すだけではメッセージは伝わらない。客観的な情報を損なわないようにしつつ、メッセージが自然に頭に入りやすくするような特段の工夫が必要になる。具体的には、統計的なメッセージを発するときには、以下のような3つのポイントについて気を配るべきである。

1）文章・グラフ・表の選択
2）関係の方向性と強さを明示
3）GEEアプローチ

（問題1）

「中学生のSNSといじめ問題」というおおまかな範囲で、問題の実態についてのリサーチ・クエスチョンの例を1つ考えなさい。同様に、問題の原因、解決策についてのリサーチ・クエスチョンの例を1つずつ考えなさい。なるべく細かい具体的な疑問文を目指すこと。

■文章・グラフ・表の選択

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。どれでも表現できる場合でも、大まかに以下のような点に留意して最適なツールを選択すべきである。

・伝達したい数値の個数は多いのか少ないのか？
・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。

■関係の方向性と大きさ

分析結果をただ図表などで提示するだけではなく、必ず「言葉で」記述する必要がある。1変数の分布を表現することは多くの人ができる（「○○と答えた人は××%でした」等）。一方で、2変数の「関係性」を正しく言葉で表現することは、意外とできていない。

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。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。「年齢が上がるにれて死亡率は上昇し、5歳ごとにほぼ倍増する」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を心がける。

■GEEアプローチ

また、関係性を記述するといっても、グラフ等が複雑なパターンを示すことがある。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。

複雑なパターンをバランスよく言語化するには**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（だいたい generalization）。次に、そのパターンが具体的に図表のどこから読み取られたのか、いくつかの数値で例を示す（数値例 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、その箇所について断りの文言を記す（但し書き 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。

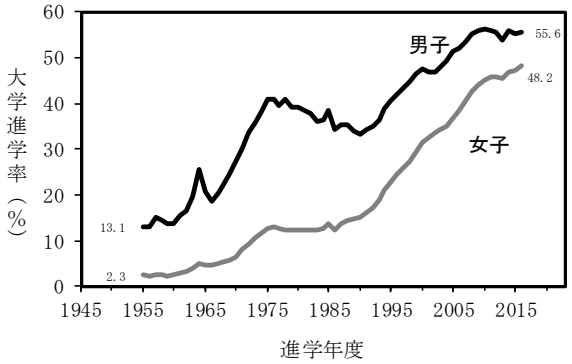


図2　男女別大学への進学率の推移（短期大学は除く）
出典：学校基本調査（文部科学省 1955～2016）

（GEEアプローチによる記述の例）図2のように、男子も女子も大学進学率はこの60年で急上昇しており、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]**。最新の2016年調査では、男子55.6%、女子48.2%と約半数もの人々が大学に進学している**[E]**。ただし、1975～90年ごろは例外的に進学率が停滞している**[E]**。

男女の差に注目すると、男子の方が常に進学率が高く、約10ポイントの差がある**[G]**。たとえば、1955年では10.8ポイント差、2016年調査でも7.4ポイント差と大きな違いはない**[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高まり、男女差が最大30ポイントまで広がった特殊な時代である**[E]**。

（問題2）

(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なさい。

・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なさい。

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リサーチ・クエスチョンは、実態・原因・解決策のどの段階の疑問なのかを明確に意識しよう
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.

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.）

※次回（7/5）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。必要な作表の判断、グラフの適切な使用、結果を伝える文章の書き方など。

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、自宅生と一人暮らしの場合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）一人暮らしの方が料理をしていた。
2）女子学生の中で調べても、やはり一人暮らし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）。

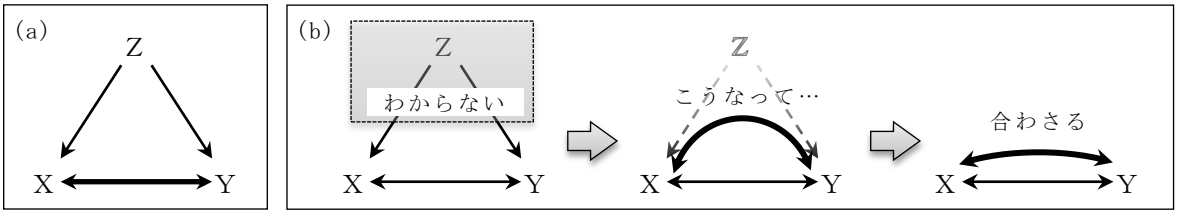


図1　シンプソンのパラドックスの原理

■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」という因果の証拠である、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）の相関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､と解釈することは明らかにおかしい。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は、男女別にして集計をやり直してみればよい。表3（b）のように、「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、という結果から、性別が重要な共通原因であったことがわかる。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、**交絡変数［先行変数］**（confounding variable; antecedent 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。

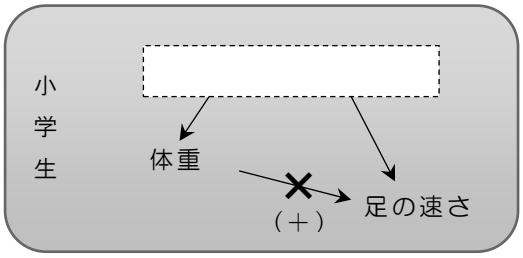
とにもかくにも重要なことは、社会現象を観察するときに、積極的に第3の変数による見せかけの関係の可能性を考慮することである。統計調査の結果を用いて新聞等でなされる主張の中には、見せかけの関係を示しているにすぎない可能性が高いものが頻繁に見受け

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、交絡変数［先行変数］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。

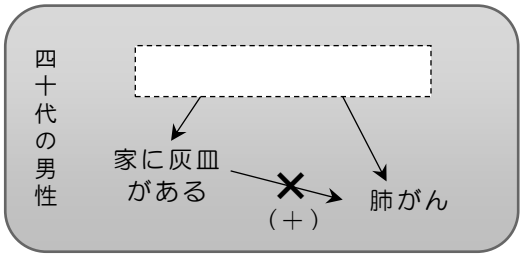
（問題）

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、交絡変数を想像してみよう。

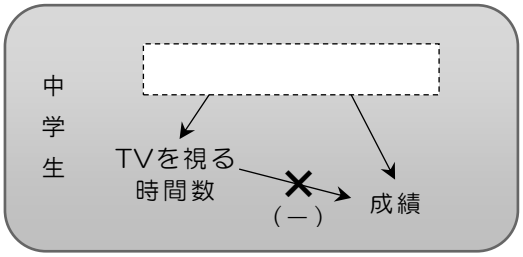
（1）



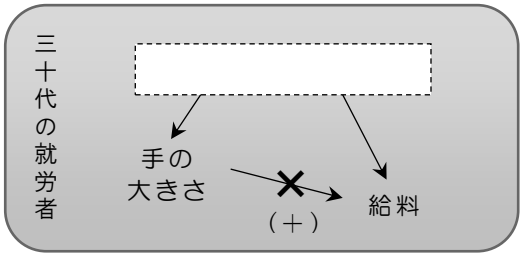
（2）



（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（交絡変数、先行変数）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.

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、という意味）。

その理由は、2つの変数の相関関係が共通の原因（交絡変数）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、統計データから因果関係を証明できるような究極的な手段（因果関係を証明する十分条件）は存在しない。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。我々に可能なことは、因果関係を主張するために最低限満たしていなければならない条件（因果関係を証明する必要条件）に注意を払うことである。一般的には次の3点があげられる。

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。

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、「時間順序がはっきり分からなければ、因果関係もはっきりとは分からない」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不明である。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。

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、共通の原因による見せかけの関係かもしれない。

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。共通の原因の可能性に気づくことさえできれば、その変数を考慮した統計分析、あるいは質的なアプローチ（インタビューや観察）から、その検討を行うことはそう難しいことではない。

新聞や雑誌、インターネットには、統計的な相関関係をもとにして因果関係を主張する記事がよく見られる。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**。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」ことはある程度安定的な客観的事実のようである。

（1）この事実から、次のように主張することは適切か、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」の間の見せかけの関係は、どのような交絡変数（共通の原因）によって発生している可能性があるか、考えてみよう。

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望む老人の方が多い。

（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、集中力が高い子どもに育つらしいよ。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：交絡変数と媒介変数の違い

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→Ｚという矢印の方向であれば、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交絡変数［先行変数］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。

このように矢印の方向は重要であるが、統計データから矢印の方向を知ることはできない。そのための材料は、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　交絡変数と媒介変数

■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りにくい。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」は完全に偶然によって決まるものなので、その交絡変数（ＸとＹに共通の原因）は存在するはずがなく、したがって見せかけの関係も起こりえない。

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）一般に、たいていの講義科目では学生の出席回数と成績の間に正の相関関係が見られる。つまり、出席回数が多い学生ほど成績がよい。このことから、「成績を上げるためには、とにかく出席することが大切だ」という意見をよく聞く。この意見に対して、見せかけの関係の視点から批判を加えなさい。

（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＋－）を描きなさい（矢印は実際に意味のある因果関係だけを記し、見せかけの関係矢印で記さないこと）。その上で、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因果関係を主張するための必要条件に注意を払う

- 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

〈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.

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）

※次回は

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。

19日に「小テストの追試」を実施する。60点に満たなかった者は受けること。

第14回「白書と政府統計＋まとめ」

■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**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）を用いて収集されている。

↑国勢調査イメージキャラクター センサスくんとみらいちゃん。センサスくんは平成2年調査から使用。みらいちゃんは平成27年調査にオンライン回答が可能になった際に導入。

基幹統計一覧（平成28年10月31日現在、56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　人口推計
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	学校基本統計　学校保健統計　学校教員統計　社会教育統計
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 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　「リサーチ・ナビ」→「経済・社会・教育」→「統計を調べる」
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らうために、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://csrda.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

「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	基本統計量 代表値（最頻値、中央値、平均値） ばらつき（範囲、四分位偏差、分散・標準偏差・変動係数）
2つの変数の 関係（相関）を 調べる→	クロス表 散布図	関係性（相関）を表わす統計量 連関係数（ユールのQ、ファイ係数、オッズ比など） 相関係数

■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）……ケースを50%ずつにわける真ん中
└→ 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
└→ 四分位偏差…中央値から第1,第3四分位までの距離
└→ 分散……平均との偏差の2乗、の平均
└→ 標準偏差……分散の正の√
└→ 変動係数……標準偏差を平均で割ったもの

・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。

2×2のクロス表の場合→ユールのQ、ファイ係数、オッズ比

■統計的な記述の実践

・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、連関係数など単純な道具で十分。

・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。

・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。

■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。

交絡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。

〈学期末試験について〉

7月26日に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）