Commentary

# クロス集計表の視覚化
## Rを利用したグラフの紹介

中野康人
関西学院大学社会学部 教授

## 1 はじめに

クロス集計表である。社会調査士資格のカリキュラムでも，クロス集計表の作成やその解釈の仕方，そしてクロス集計表に基づく各種統計量や独立性の検定など，少なからぬ時間がクロス集計表に関連する事項に費やされている。クロス集計表は，社会調査に携わるものにとって欠くべからざる基礎的な分析技術の一つである。もちろん，本稿の読者の多くは，クロス集計表を作成したその先にある，より詳細な分析に主眼をおいているだろう。また，論文や詳細な調査報告を書く場合には，クロス集計表は重要なデータとして数値をそのまま掲載することが多い。実際，『社会と調査』を創刊号から概観すると，クロス集計表の情報が出る場合は，ほぼ「表」として掲載されている。

一方で，社会調査を実践するものとして，調査結果を速報や報告書として対象者に開示したり，プレゼンテーションの場で聴衆に説明したりすることがあるだろう。そのような場合に，数字を羅列したクロス集計表をそのまま使うのは，あまり得策とはいえない。読者や聴衆の関心を引きつけるには，視覚化が重要になる。本稿では，日頃まとめて語られることの少ない，クロス集計表の視覚化について具体的な事例を紹介していく。

なお，分析についてもいえることだが，視覚化についても使用するツール（ソフトウェア）が重要になる。多くの社会調査者にとって，普段自分が使っているツールでできることに自分ができることが制約されがちである。本稿では，統計解析のフリーソフトであるRでの事例を中心に紹介していく。なお、Rに関する解説は中村（2010）などを参照のこと。

## 2 クロス集計表

本題に入る前に，クロス集計表について確認しておこう。クロス集計表とは，原則として二つの離散変数について，各々の構成カテゴリをそれぞれ行と列に配置し，行と列がクロスするセルに該当するカテゴリの組み合わせにあてはまる統計量（頻度など）を配置した数表のことをいう。加えて，行，列，それぞれの合計の数値及び，全体の合計の数値を入れることが一般的である。頻度を表現するクロス集計表の場合，前者をセル度数，後者を周辺度数という。分析する変数が三変数以上になれば，三つ目（以上）の変数のカテゴリごとに，二変数のクロス集計表を入れ子状に作成することになるが，本稿では，基本的な二変数間のクロス集計表（二元クロス集計表）に限定して筆を進める。

頻度のクロス集計表は，集計の基本であるが，そこから直接変数間の関係を読み取るのは困難である。比率のクロス集計表，特に説明変数の周辺度数を基数とした％のクロス集計表から，説明変数のカテゴリごとに被説明変数の分布が異なるかどうかをみることによって，変数間の関係を読み取ることが容易になる。日本国内では，説明変数を行に，被説明変数を列に配

表1 サンプルデータのクロス集計表（頻度，行%，標準化残差）

| | まったく心配していない | 心配していない | どちらともいえない | 心配している | 非常に心配している | 合計 |
|---|---|---|---|---|---|---|
| 義務教育未了 | 339 | 497 | 783 | 698 | 707 | 3024 |
| 中学校卒業 | 1033 | 1812 | 4438 | 4123 | 3805 | 15211 |
| 高等学校卒業 | 490 | 1246 | 3533 | 3568 | 2967 | 11804 |
| 短大等卒業 | 156 | 513 | 1645 | 2287 | 1628 | 6229 |
| 大学卒業 | 149 | 590 | 1904 | 2836 | 2177 | 7656 |
| 合計 | 2167 | 4658 | 12303 | 13512 | 11284 | 43924 |

| | まったく心配していない | 心配していない | どちらともいえない | 心配している | 非常に心配している | 基数 |
|---|---|---|---|---|---|---|
| 義務教育未了 | 11.2 | 16.4 | 25.9 | 23.1 | 23.4 | 3024 |
| 中学校卒業 | 6.8 | 11.9 | 29.2 | 27.1 | 25 | 15211 |
| 高等学校卒業 | 4.2 | 10.6 | 29.9 | 30.2 | 25.1 | 11804 |
| 短大等卒業 | 2.5 | 8.2 | 26.4 | 36.7 | 26.1 | 6229 |
| 大学卒業 | 1.9 | 7.7 | 24.9 | 37.0 | 28.4 | 7656 |
| 合計 | 4.9 | 10.6 | 28.0 | 30.8 | 25.7 | 43924 |

| | まったく心配していない | 心配していない | どちらともいえない | 心配している | 非常に心配している |
|---|---|---|---|---|---|
| 義務教育未了 | 16.52 | 10.79 | − 2.69 | − 9.48 | − 3.01 |
| 中学校卒業 | 13.08 | 6.48 | 3.96 | − 12.09 | − 2.36 |
| 高等学校卒業 | − 4.59 | -0.20 | 5.43 | − 1.47 | − 1.61 |
| 短大等卒業 | − 9.56 | − 6.56 | − 3.04 | 10.99 | 0.87 |
| 大学卒業 | − 13.28 | − 9.06 | − 6.73 | 13.10 | 6.05 |

置し，行%を提示することが多い。

さらに本稿では，残差の情報にも注目したい。%の変化の度合いは，もともとのセル度数や周辺度数の大きさに依存する。同%の変化でも，それが意味する関係の強さは常に同じではない。残差は，ある基準値からの観測度数のずれ具合を数値で表現してくれる。基準となる値は，様々なモデルから算出される期待値が使われるが，もっとも一般的なのは独立モデル（二変数が完全に無関係である状態）から期待される値との差である。ピアソン残差は，その差を期待値の平方根で割ったもので，$\chi^2$値やクロス集計表に基づく統計量の基礎となる数値である。さらにそのピアソン残差を，各カテゴリ毎の頻度のばらつきの程度で調整した数値が標準化残差である。標準化残差は，二変数が独立である場合に近似的に標準正規分布に従うことが知られている。

以下では，これらの情報を視覚化するいくつかの方法を紹介していく。

## 3 具体例

以下では，ISSP（International Social Survey Program）の2010年調査「環境Ⅲ」から得られる，「環境問題に関する心配度」と「回答者の最終学歴」の二変数のクロス集計表（日本語版調査票のQ7とF6をクロス集計したもの。カテゴリは簡単のためにリコードしてある）を例として取り上げる（ISSP Research Group, 2012）。それぞれの頻度，行%，標準化残差のクロス集計表は表1の通りである。

## 4 視覚化

具体的な視覚化の方法を紹介する前に，一般的にデータを視覚化する際の注意点をまとめておきたい。

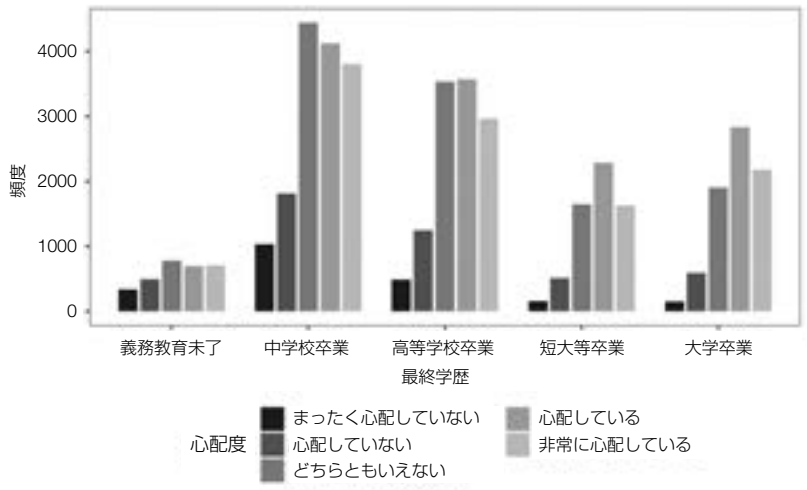


図1 棒グラフ

データの視覚化は，本来は数値であるものを数値以外で表現して「みてわかる」状態するものである。数値を数値以外で表現するために，どのようなもので代替するかというと，

形，長さ・高さ，大きさ，位置，色

などが，いまのところ紙・ディスプレイ・スクリーンなどの二次元的な表現装置で利用される要素である。また，それらを時間の流れの中で変化させることも要素の一つとなる。

これらの要素を数値の代替物として表現することになる。美しさや新奇さも人を引きつけるためには必要かもしれないが，データの内容を正しく伝えるという観点からは，データとの整合性，視認性，一覧性，解釈可能性が求められる。

## 5 クロス集計表の視覚化

**棒グラフ・帯グラフ**

一変量の離散変数の度数分布を視覚化する場合，単純な棒グラフがまずは第一候補となるだろう。二変量のクロス集計表の場合も，説明変数のカテゴリごとに分割した，被説明変数の棒グラフがよく使われる（図1）。行%の視覚化には，説明変数のカテゴリごとに被説明変数の比率を示す帯グラフが適している（図2）。いずれも，表計算ソフトや多くの統計ソフトで簡単に作成できる。Rでは，ベースグラフィックを含めて，棒グラフ作成の関数は複数ある。図1，図2は，Rのグラフ作成のパッケージとして広く使われているggplot2で作成したものである。ggplot2では，視覚化するデータを指定した後，geometryと呼ばれるオプションを付加することによって異なる形状のグラフを作成することができる。棒グラフ，帯グラフで指定するgeometryは，geom_bar()である。

棒グラフは，頻度を大雑把に把握するには適している。しかし頻度に基づく以上，ここから関係を読み取ることは困難である。帯グラフは%に基づくので，分布の違いを視覚化できるが，カテゴリ数が多い場合や，分布の変化が複雑な場合は，比較が容易でないため，解釈可能性が低くなる。

**リッカートプロット**

リッカートプロットは，帯グラフを改良した

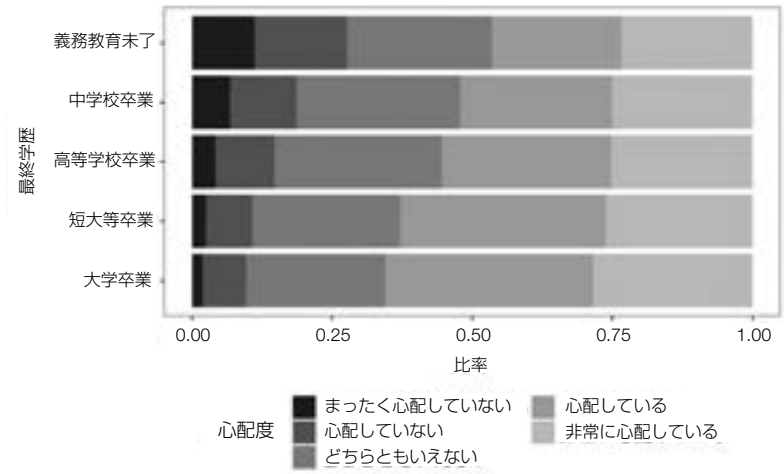


図2 帯グラフ

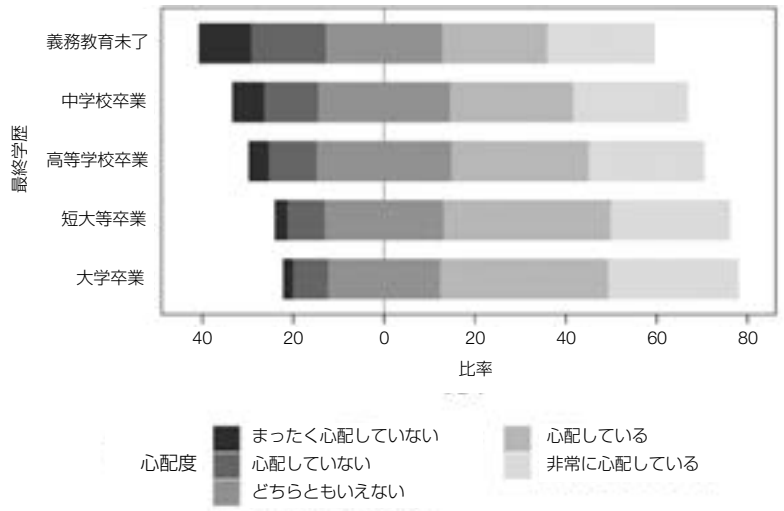


図3 リッカートプロット

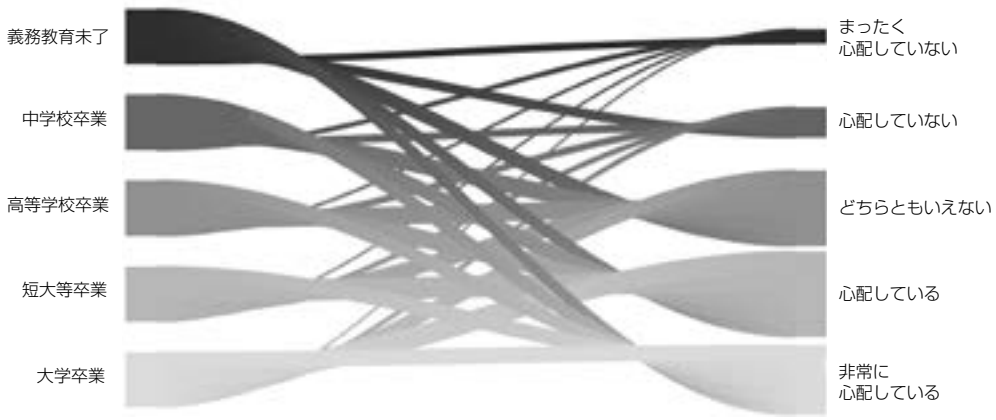


図4 リボンプロット

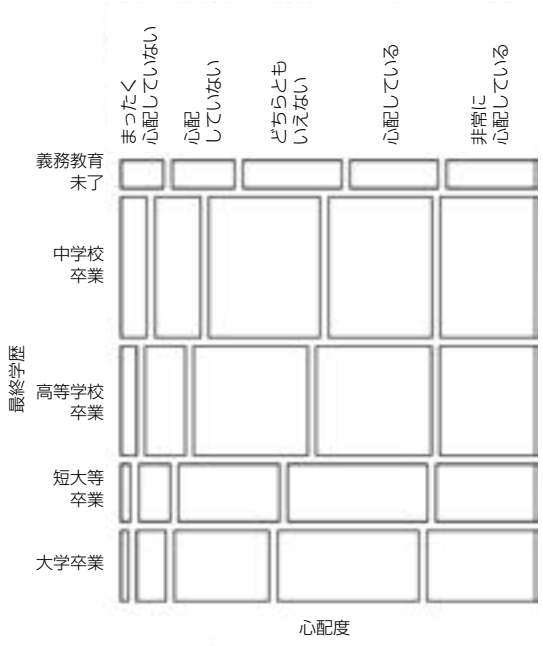


図5 モザイクプロット

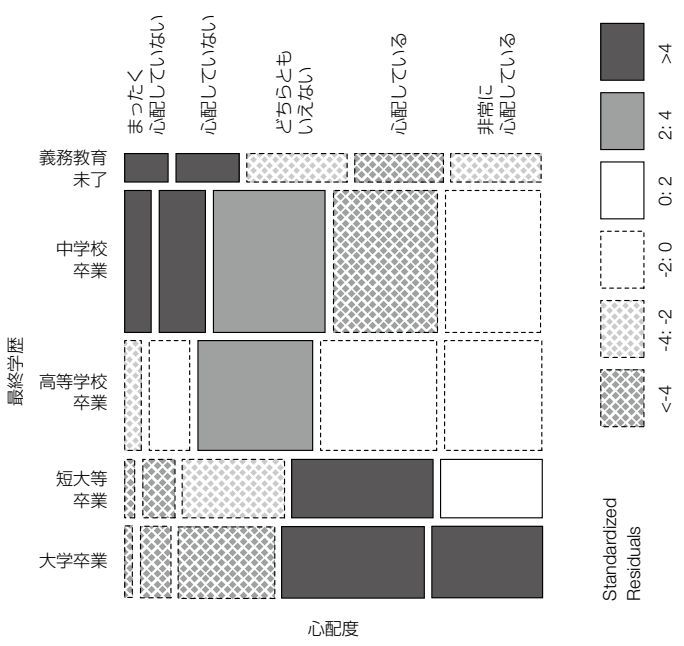


図6 モザイクプロット（標準化残差による塗り分け）

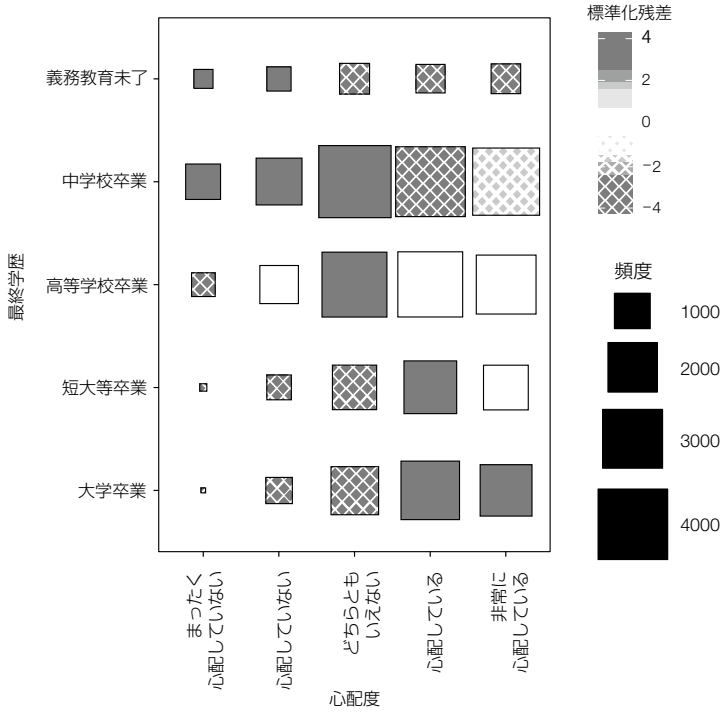


図7 フラクチュエーションプロット

グラフである（HeibergerandRobbins,2014）。被説明変数が順序尺度で測定されている場合，順序の中間点になるカテゴリを中心にして帯を並べることにより，説明変数のカテゴリごとに被説明変数のおおよその傾向（中間点を基準にしてポジティブな回答が多いのか，ネガティブな回答が多いのか）が掴める。プロットする帯は，頻度に基づくものでも比率に基づくものでも可能であるが，二変数間の関係を読み取ることを意図する場合は，比率のリッカートプロットを利用した方が解釈しやすい。

図3は，表1の行%に基づく比率のリッカートグラフで，HHやlikertなどのパッケージで簡単に作成できる。一覧性・視認性そして解釈可能性を高めたリッカートプロットであるが，順序尺度にしか適用できないというのが大きな欠点であろう。

### リボンプロット

もう一つ，帯グラフの変形版を紹介する。図4は，リボンプロット（リバープロット，サンキープロットともいう）である。左側に説明変数のカテゴリを，右側に被説明変数のカテゴリを配置し，説明変数の各カテゴリのリボンを該当する被説明変数の比率で分割した上で被説明変数のカテゴリごとに再びリボンとしてまとめている。図4では，行%のクロス集計表を視覚化しており，説明変数の各カテゴリを100%として，各被説明変数カテゴリに分割している。右側で左側から流入するリボンの太さを比較することによって，行%の分布の比較と同様の解釈が可能となる。頻度（もしくは全体%）に基づいてリボンの太さを視覚化すれば，セル度数だけでなく周辺度数の分布も一覧できるグラフとなる。Rでは，riverplotパッケージを利用すると簡単に描画できる。

行%の解釈プロセスを，左から右への視点の移動で実現するグラフで，直感的に解釈がしやすい。特に，カテゴリごとの分布の違いが明確である場合にはわかりやすい。とはいえ，現時点では一般的なグラフとはいえないので，初見の者がすぐに内容を理解するのは困難で，文章や口頭による説明が欠かせない。

### モザイクプロット

次に，帯グラフの完成型といってもよいであろうモザイクプロットを紹介する。

モザイクプロットは，Rのベースグラフィックでクロス集計表をプロットする際に出力される形式である（図5）。

説明変数の周辺度数の比率に応じて帯の太さが規定され，その上で被説明変数の各カテゴリの比率に応じて帯が分割されている。帯グラフの帯の太さを説明変数の周辺分布に応じて変化させたものと考えればよい。それぞれの分割されたタイルの面積は，セル度数もしくはセルの全体%を表現することになる。また，二変数が独立であれば，各行%は説明変数のカテゴリごとに異ならないことになるので，タイルは碁盤の目のようになる。ジクザグにずれてモザイク状になっていれば，独立状態から乖離していることを意味する。この状態で，頻度の情報と比率の情報を一覧できるグラフになっている。モザイクプロットの歴史や応用方法については，Friendly（1994,1999,2000,2002）やZeileisetal.（2007）が詳しい。Rのベースグラフィックの他には，VCDパッケージでも描画が可能である。Rでは，さらにここに残差情報を付加することができる。色（青が正，赤が負。ただし，印刷の都合上，ここでは正がアミかけ、負が格子柄）と枠線（実線が正，破線が負）によって，標準化残差の大きさと方向が一覧できる（図6）。

モザイクプロットは，頻度，比率，残差の三要素全てを視覚化できるグラフである。しかし，各変数のカテゴリ数が極端に多い場合，度数や行%に偏りがある場合には，どのタイルがどのカテゴリを参照しているのかが分かりにくくなり，視認性が著しく落ちるという欠点がある。

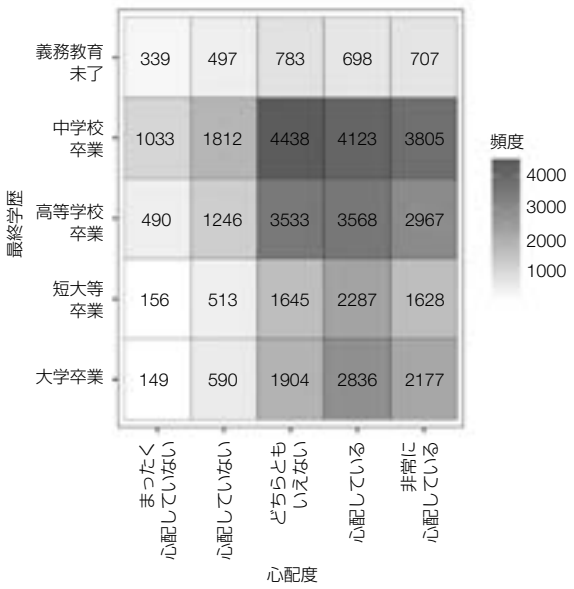


図8 ヒートマップ（頻度）

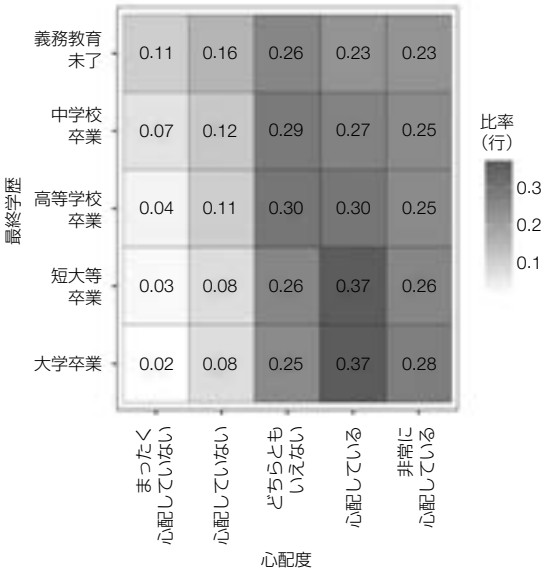


図9 ヒートマップ（行%）

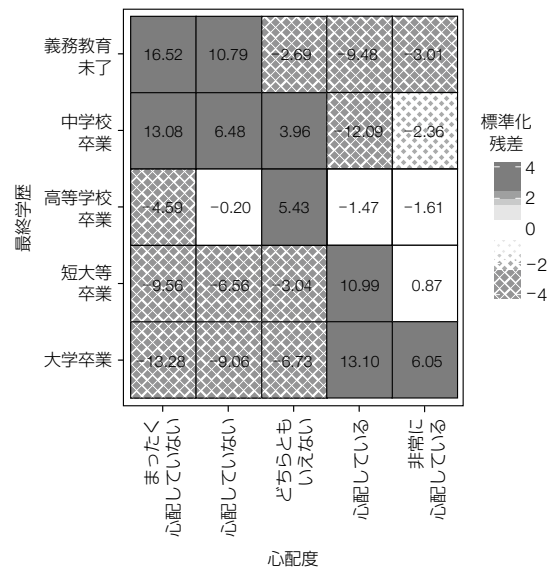


図10 ヒートマップ（標準化残差）

### フラクチュエーションプロット

モザイクプロットの視認性の悪さを克服する代替案の一つが，フラクチュエーションプロットである（図7）。クロス集計表と同様に，説明変数と被説明変数の各カテゴリをそれぞれ縦横に配置し，クロスさせて該当する部分に，頻度や%に応じた大きさのタイルを配置するグラフである。縦横の位置によってあてはまるカテゴリが確実に認識できる（WickhamandHofmann, 2011）。

Rでは，productplotsパッケージの関数prodplot()で簡単に作成できる。また，ggplot2では，geometryのgeom_point()を応用して描画可能である。この場合，タイルの大きさだけでなく，色についても情報を渡すことができる。頻度や%の情報を大きさに重ねる形で色として表現してもよいだろうし，標準化残差で色を塗り分ければ，完全にモザイクプロットの代替となる。

とはいえ，タイルのサイズの取り方と頻度の偏りによっては，視認性の問題は改善されない。極端に小さなタイルがあると，その色情報は識別しにくくなる。

### ヒートマップ

モザイクプロットの代替案のもう一つが，

ヒートマップである。モザイクプロットやフラクチュエーションプロットの視認性の悪さは，タイルのサイズに起因するが，ヒートマップでは大きさで情報を視覚化することを排除し，色と文字情報のみで視覚化している。つまり，クロス集計表を頻度や%や残差で色分けしたようなものである（図8, 図9, 図10）。

Rでは，ggplot2 パッケージで geometryの geom_tile() を利用することで作成可能である。

大きさという要素を諦めた分，一度に伝えられる情報は減るが，一覧性・視認性は確保されるので，特にカテゴリ数が多い場合に有効なグラフとなる。もっとも伝えたい情報を色で表現し，その他の情報は全てタイル内に文字情報として付記すれば，三要素全てを提示することも可能で，クロス集計表の代替として十分に活用できる。

## 6 まとめ

以上ここまで，クロス集計表を視覚化するいくつかの方法を紹介してきた。この他にも，例えば対応分析の出力なども直感的に理解しやすいグラフとなるだろう。いずれにも一長一短があり，一つの万能なグラフ形式があるわけではない。データの性質，伝えたい統計量，伝える相手に合わせて，適切なグラフを考えながら用意する必要がある。どのグラフについても，そこから二変数間の関係を読み取るにはある程度の統計リテラシーが必要になってくる。連続変数間の関係については，積率相関係数や散布図が高校数学の範囲で教育されるようになったが，離散変数間の関係についてはまだまだ一般的な知識とはいいがたい。視覚化と合わせて，関係を読み取る統計リテラシーの普及を期待するものである。

なお，本稿に掲載しているグラフは、紙面の都合上、色や形状に制約がある。グラフを作成するための Rスクリプト は URL（http://www.soc-nakano.net/asr18/）を参照されたい。

**文献**


Friendly, M., 1994, "Mosaic Displays for Multi-Way Contingency Tables," *Journal of the American Statistical Association*, 89(425): 190-200.

Friendly, M., 1999, "Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data", *Journal of Computational and Graphical Statistics*, 8(3): 373-395.

Friendly, M., 2000, *Visualizing Categorical Data*. SAS Institute, Carey, NC.

Friendly, M., 2002, "A Brief History of the Mosaic Display," *Journal of Computational and Graphical Statistics*, 11(1): 89-107.

Heiberger, R. M. and Robbins, N. B., 2014, "Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications," *Journal of Statistical Software*, 57(5): 1-32.

ISSP Research Group, 2012, International Social Survey Programme: Environment III - ISSP 2010. GESIS Data Archive, Cologne. ZA5500 Data file Version 2.0.0.

中村健太郎,2010「統計解析環境R言語の紹介」『社会と調査』5:104-108。

Wickham, H. and Hofmann, H., 2011, "Product plots," *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis '11)*, 17(12): 2223-2230.

Zeileis, A., Meyer, D., and Hornik, K., 2007, "Residual-based Shadings for Visualizing (Conditional) Independence," *Journal of Computational and Graphical Statistics*, 16(3): 507-525.