A 「4.2.6 その他」の差し替え原稿 (本書 98 ページ)

最後に，社会的望ましさバイアスが生じる可能性のある以下の質問を考えてみましょう。

> 項目例 27:以下に示す就職先を選ぶ際の観点について，あなたが重要視する順にすべて並べ替えてください。
> A. 知名度　B. 給与水準　C. 福利厚生　D. 社会性

直接的に就職先を選ぶ際の観点の重要度を尋ねたとしても，社会的に望ましい態度をとる方が良いと判断されると，本音の回答を得ることは難しくなってしまいます。本当は知名度や給与水準を重要視していたとしても，社会貢献度や将来性などを重要視すると回答するかもしれません。このような問題に対処するために，間接的に観点の重要度を算出する方法を紹介します。

まず，典型的な企業を 20 から 30 社挙げ，調べたい観点を複数用意します。ここでは上述の四つを利用することにします。次に，各観点について，それぞれの企業を 5 段階評定で評価してもらいます[1]。加えて,「就職希望度」という観点からも各企業を評価してもらい，表 A.1 に示すように表形式にデータをまとめます。

表 A.1 は一人分の企業評価データです。評価者が複数いる場合には，表 A.1 のようなデータを評価者数分用意し，企業ごとに各観点についてすべての評価者の平均を計算します。例えば，三人の評価者が 20 社を評定した場合には，図 A.1 に示すように，三枚の企業評価データが得られます。この三枚の企業評価データから，企業 1 の知名度を求める場合には，三人のそれぞれの評定値を平均することによって，$4(=(5+4+3)/3)$ と求まります。

続いて，一人分もしくは複数人の評価をまとめた企業評価データに関して,「就職希望度」と各観点の間で一致度[2] を計算します。「就職希望度」と一致度の高い観点は，それだけ重要視されていることになりますので，一致度の高低によって重要度の評価を行います。

表 A.1: 企業の評価データ

	知名度	給与水準	福利厚生	社会性	就職希望度
企業 1	5	3	4	5	5
企業 2	4	3	1	4	2
企業 3	2	2	2	3	3
⋮			⋮		
企業 20	4	3	4	2	5

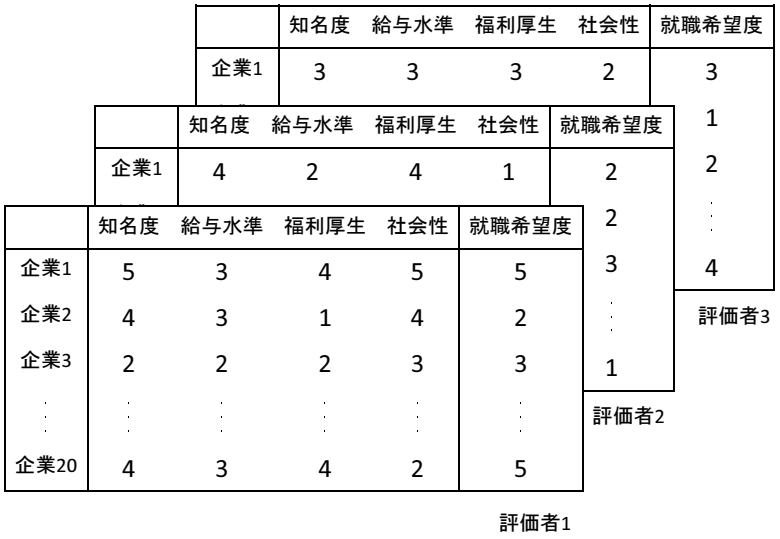

図 A.1: 三人の評定者の企業評価データ

[1]評定方法は 5 段階評定だけとは限りません。
[2]一致度の算出には，第 6 章で説明する相関係数などを利用します。

B 「6.2.5 残差分析」に関する補足 (本書 147 ページ)

残差分析は 2 つの項目が独立であると仮定したときのクロス表を用意し，これと実際に得られたクロス表との間のセルの度数の差をみることによって，回答結果の間に連関があるか評価する方法です。ここでは，連関の度合いの指標となる**残差**とそれを定義づける 2 つの項目が独立である場合のセルの度数である**期待度数**について説明します。

実際に得られた (i, j) セルの度数 n_{ij} と 2 つの項目が独立である場合の当該セルの期待度数 e_{ij} との差をさらに期待度数の平方根で除して定量化したものを**残差**といい，以下のように res_{ij} で表します。

$$res_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{B.1}$$

res_{ij} の絶対値が大きいほど，回答結果の間に連関が強いと解釈し，小さいほど連関が弱いと解釈します。ここで，2 つの項目が独立である場合の当該セルの期待度数 e_{ij} について考えてみましょう。

表 B.1: $a \times b$ のクロス表におけるセルの度数と周辺度数

		項目 B のカテゴリ					
		1	$\cdots$	j	$\cdots$	b	合計
項目 A のカテゴリ	1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1b}	$n_{1.}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ib}	$n_{i.}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	a	n_{a1}	$\cdots$	n_{aj}	$\cdots$	n_{ab}	$n_{a.}$
	合計	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.b}$	N

表 B.1 は $a \times b$ のクロス表における実際に得られた (i, j) セルの度数 n_{ij} と i 行および j 列の度数の合計である**周辺度数**$n_{i.}$，$n_{.j}$ を示したものです。項目 A と項目 B が独立である場合，項目 B の b 個のカテゴリへの度数の配分の割合は項目 A のどのカテゴリにおいても同様であることが期待されます。その上で，項目 B のカテゴリの周辺度数はそれぞれ $n_{.1} \cdots n_{.b}$ とそれぞれ違った値ですから，全体の度数 N を利用して，項目 A のそれぞれのカテゴリの周辺度数を項目 B のカテゴリの周辺度数の割合によって比例配分したものが当該セルにおいて期待される度数となります。すなわち，2 項目間に連関が全くない場合の (i, j) セルの期待度数 e_{ij} は以下によって与えられます。

$$\begin{aligned} e_{ij} &= n_{i.} \times \left(\frac{n_{.j}}{N} \right) \\ &= \frac{n_{i.} \times n_{.j}}{N} \end{aligned} \tag{B.2}$$

また，以下のように考えることもできます。表 B.1 のクロス表において，項目 A のみに注目したときに i 番目のカテゴリをとる確率は「i 行の周辺度数 ÷ 全体の度数」で表され，すなわち $n_{i.}/N$ です。これを項目 A の i 番目のカテゴリの**周辺確率**といいます。一方，項目 B の j 番目のカテゴリの周辺確率は「j 列の周辺度数 ÷ 全体の度数」で表され，$n_{.j}/N$ です。2 つの項目が独立である場合，カテゴリ i と j が同時に観察される確率である (i, j) セルの**同時確率**は独立性の定義より[1]，上述した 2 つの周辺確率の積で以下のように表されます。

$$\begin{aligned} \text{独立である場合の } (i,j) \text{ セルの同時確率} &= \frac{n_{i.}}{N} \times \frac{n_{.j}}{N} \\ &= \frac{n_{i.} \times n_{.j}}{N^2} \end{aligned} \tag{B.3}$$

この同時確率を用いて，全体の度数をそれぞれのセルに配分したものが，2 つの項目が独立である場合の期待度数 e_{ij} です。すなわち，

$$\begin{aligned} e_{ij} &= \frac{n_{i.} \times n_{.j}}{N^2} \times N \\ &= \frac{n_{i.} \times n_{.j}}{N} \end{aligned} \tag{B.4}$$

となり，(B.2) 式と同様の結果を得ます。

本節で用いた調整済み標準化残差は (B.1) 式の分母に以下で示す当該セルの標準偏差 $\sqrt{v_{ij}}$ を用いることで，厳密に標準化を施したものです。

$$\sqrt{v_{ij}} = \sqrt{e_{ij} \times \left(1 - \frac{n_{i.}}{N}\right)\left(1 - \frac{n_{.j}}{N}\right)} \tag{B.5}$$

■ 文献

金明哲 編 藤井良宜 著 (2010). カテゴリカルデータ解析, 共立出版

豊田秀樹 著 (1998). 調査法講義, 朝倉書店

南風原朝和 著 (2002). 心理統計学の基礎—統合的理解のために—, 有斐閣アルマ

[1] 変量 1 の周辺確率 × 変量 2 の周辺確率 ≡ 2 変量の同時確率が成り立つとき，またそのときのみ 2 つの変量は独立であるといいます。

C　段落テキストとチェックボックスのデータの処理

ここではGoogleフォームの段落テキスト形式とチェックボックス形式を利用した際の回答データの扱い方を紹介します。第2章で説明されている通り，段落テキスト形式は自由記述式項目，チェックリストは多肢選択複数回答項目に使われます。本章で使用するRスクリプトは以下の通りです。

```
source("myfunc/myfunc.R")
#段落テキスト形式
ondanka_data<-read.csv(file="data/温暖化データ.csv")
head(ondanka_data)
FAlist(ondanka_data[,1],sep="\n")
#チェックボックス形式
cake_data <- read.csv(file="data/ケーキデータ.csv")
head(cake_data)
MAtotal(cake_data[,1],sep=",")
```

段落テキスト形式およびチェックボックス形式のデータの例として，それぞれ第2章の図2.8と図2.14のフォームで収集したデータを利用します。段落テキスト形式の“「地球温暖化を防止するために私たちにできること」というテーマで，自らの考えを論述せよ。”という質問に対しては38人から，“以下に挙げるケーキの種類のうち，特に好きなものを3つ選んでください。”というチェックボックス形式の質問には83人から回答が集まりました。これら2つの質問への回答データ[1]をRで読み込み，head()で最初の6人分のデータを表示すると以下のようになります。

```
> head(ondanka_data)
                                                                                x
1                                          温暖化について学ぶ\n ちいさなことからコツコツやる
2                                        買い物かばんを持ち歩く\n クーラーの設定温度を変える
3 ビニール袋を使いすぎない\n 自転車でなるべく移動する\n クーラーの温度を上げる
4                                エアコンを使いすぎない\n 家電の電源をいれっぱなしにしない
5                                                                植物を家で育てる
6                                                                  わからない
> head(cake_data)
                                                x
1                    イチゴのショートケーキ，モンブラン，チーズケーキ
2                    イチゴのショートケーキ，シュークリーム，エクレア
3 イチゴのショートケーキ，シュークリーム，チーズケーキ
4                  モンブラン，シュークリーム，エクレア
5  イチゴのショートケーキ，モンブラン，シュークリーム
6              モンブラン，シュークリーム，チーズケーキ
```

段落テキスト形式およびチェックボックス形式では，1行につき1人の回答者からの回答が入力されていることがわかります。このままでは回答内容が見やすいとは言えず，データ処理も難しいため，それぞれのデータをRを使って整理してみましょう。

[1] 5.2.2項で説明した成形を施した後のデータです。

C.1　段落テキスト形式で収集したデータの加工

段落テキスト形式で収集したデータは，一文ずつに分けて表示することで，どのような回答結果が得られたか記述内容の一覧を作成することができます。この処理を行うためには，段落テキスト形式において一文一行で回答してもらう必要があります。一文ずつ分割するためには自作関数FAlist()を使用します。

```
FAlist(データ,sep="データ区切り文字")
```

sep=""という引数にはデータの区切り文字を指定します。あらかじめ一文一行で箇条書きにより回答してもらうことで，通常\n (改行を表す正規表現。\は¥でも可。) が区切り文字となるため，sep="\n"と指定します[2]。それでは先ほど説明したデータondanka_dataを実際にR上で処理してみましょう。

```
> FAlist(ondanka_data[,1],sep="\n")
 [1] "温暖化について学ぶ"                          "ちいさなことからコツコツやる"
 [3] "買い物かばんを持ち歩く"                      "クーラーの設定温度を変える"
 [5] "ビニール袋を使いすぎない"                    "自転車でなるべく移動する"
 中略
[65] "ビニール袋を使わない"                        "家電の電源を入れっぱなしにしない"
[67] "わからない、勉強する"                        "公共機関をつかう"
[69] "水の使いすぎに注意"
```

読み込んだデータを関数FAlist()に入れる際には，段落テキスト形式の項目に対応するデータの列番号を[,列番号]として指定してください。今回は1列目に入力されているためondanka_data[,1]とします。段落テキスト形式で収集したデータが一文ごとに区切られて表示されていることが確認できました。

C.2　チェックボックス形式で収集したデータの集計

チェックボックス形式で回答してもらったデータに関しては，それぞれの選択肢が選ばれた数を知ることができれば便利です。ここでは，自作関数MAtotal()を使って，多肢選択複数回答項目によるデータを集計してみましょう。関数

[2] 万が一エラーが生成された場合，データオブジェクトをprint()で確認し，適切な区切り文字をsep=で指定してください。その他に考えられる区切り文字としては，;や,などがあります。

MAtotal() では，まず前節で紹介した関数 FAlist() を使って選択肢ごとに区切られた形に加工した後，それぞれの選択肢が選ばれた数を集計する作業を 1 つの関数内で行っています。

```
MAtotal(データ,sep="データ区切り文字")
```

チェックボックス形式を利用して収集したデータは，選択肢と選択肢の間の区切り文字として通常，が使われているため，sep=","と指定しましょう。それでは先ほどのデータ cake_data を実際に集計してみます。

```
> MAtotal(cake_data[,1],sep=",")
                      回答数 割合
イチゴのショートケーキ     46 0.18
モンブラン                 51 0.20
チーズケーキ               54 0.22
シュークリーム             54 0.22
エクレア                   44 0.18
```

関数 FAlist() の場合と同様に，読み込んだデータを関数 MAtotal() に入れる際には，チェックボックス形式の項目に対応するデータの列番号を [, 列番号] として指定してください。以上のように，それぞれの選択肢が選ばれた数を集計することができました。「回答数」の列は選択された回数を,「割合」の列は全体に対する選択された割合を表します。ここではチーズケーキとシュークリームが 22%(0.22) で多く選択されていることがわかります。

D　質的変量のグラフ化

第 5 章 4 節では，量的変量の分布を見るために，度数分布表とヒストグラムを作成しました。度数分布表からでも特徴を読み取ることはできますが，ヒストグラムで視覚的に表現した方が，把握しやすくなりました。アンケート調査の結果を集計・分析する際に，図で要約を行うことはとても重要です。そこで，本稿では質的変量を図に要約する方法を説明します。

質的変量について，第 5 章 3 節では，変量ごとに各カテゴリの頻度や比率を表にまとめる方法を説明しました。また，第 6 章 2 節では，2 つの質的変量間の関係を把握するために，クロス表を作成する方法を説明しました。これらの表は，棒グラフを使用して図に表すことができます。

D.1　変量ごとの棒グラフ

まず，変量ごとに各カテゴリの頻度や比率を棒グラフに表してみましょう。以下に，本稿で使用する R コードを示しました。

```
source("myfunc/myfunc.R")
library(vcd)

college_data<-read.csv("college.csv")
barplot(table(college_data[,5]))
college_data<-read.csv("college.csv")
barplot(table(college_data[,5]))
barplot(table(college_data[,5]),col=heat.colors(5))

(cross<-xtabs(~V1+V5, data=college_data))
barplot(cross,legend.text=rownames(cross),beside=T)
barplot(cross,legend.text=rownames(cross),beside=F)

chirol<-read.csv("data/chirol_data.csv")
MA<-MAtotal(chirol[,3],sep=",")
barplot(MA[,1])
```

棒グラフは，関数 barplot()

```
barplot(height,legend.text, beside = F, horiz = F, col)
```

を使用して作成します。1 番目の引数 height には，グラフにしたい頻度や比率のベクトルや行列を指定します。2 番目の引数 legend.text を指定すると，図の右上に凡例を表示することができます。3 番目の引数 beside は，クロス表形式のデータを読み込んだ際に，グラフの種類を指定する部分です。beside=T の場合は，並列（集合）棒グラフ、beside=F の場合は積み上げ棒グラフとなります。4 番目の引数 horiz は，棒グラフの向きを指定する部分です。horiz=T とすると，棒を水平に描画します。5 番目の引数 col では色の指定を行います。R で指定されている色の表記を用いて，657 色の中から好きな色に変更することができます。白，黒，赤，緑，青，青緑，ピンク，黄色，灰色には，それぞれ 0〜8 の数字が割り当てられており，数字で指定することもできます。また，関数 rainbow()，heat.colors()，terrain.colors()，topo.colors()，cm.colors() を使用して，グラデーションを指定することもできます。

それでは，実際に棒グラフを作成してみましょう。ここでは，第 6 章 2 節で使用した「大学生の友人関係に関する調査」のデータを使用します。データを読み込んだ後，

```
barplot(table(college_data[,5]))
```

を実行します。すると，図 1 が描画されます [1]。図 1 より，他人から悩みなどを相談されることが多いかどうかについては,「まあまあ多い」と回答する人が最も多いことが簡単に読み取れます。また，グラフの色をグラデーションに変更したい場合には，

```
barplot(table(college_data[,5]),col=heat.colors(5))
```

のように，色を指定します。上記の図 2 の出力例では，関数 heat.colors() を使用していますが，自分でそれぞれの棒の色を指定したい際は，関数 c() を使用して，col=c(0:5) のように指定します。

D.2　並列（集合）棒グラフ・積み上げ棒グラフ

「大学生の友人関係に関する調査」では，項目 1 で性別の回答を得ています。そこで，次に，回答者の性別を考慮した棒グラフを描画することを考えます。このように別の項目で分類した棒グラフは，大きく並列（集合）棒グラフと積み上げ棒グラフに分けられます。並列（集合）棒グラフは，分類した項目ごとの

[1] グラフィックスの表示画面で横幅が狭い場合には，ラベルがうまく表示されない場合もあるため，適宜表示画面の大きさを調整してください。

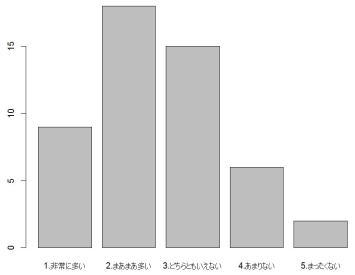

図 1: 「項目 5」の棒グラフ

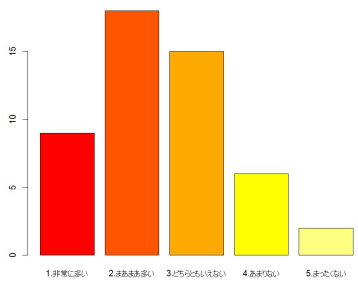

図 2: 「項目 5」の棒グラフ

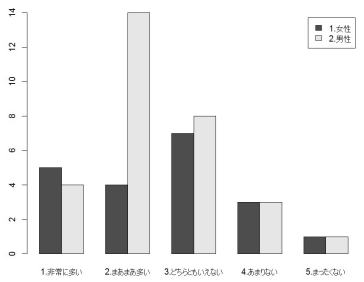

図 3: 男女別の項目 5 の棒グラフ

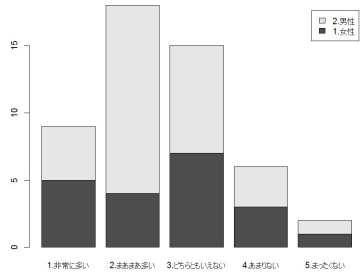

図 4: 項目 5 の積み上げ棒グラフ

頻度を比較する場合に適しており，積み上げ棒グラフは，分類した項目ごとの全体に対する割合と，全体の合計値を比較する場合に適したグラフです。

```
(cross<-xtabs(~V1+V5, data=college_data))
barplot(cross,legend.text=rownames(cross),beside=T)
```

を実行すると，図 3 のような集合棒グラフが描画されます。図 3 を見ると，男性は「まあまあ多い」と回答している回答者が多いのに対し，女性では「どちらともいえない」と回答している回答者が最も多くなっているため，男女で回答の傾向が異なることがわかります。また，図 4 のような積み上げ棒グラフは

```
barplot(cross,legend.text=rownames(cross),beside=F)
```

を実行すると作成することができます。図 4 を見ると,「まあまあ多い」以外の選択肢では，男女の回答の割合が約半分であるのに対し,「まあまあ多い」では圧倒的に，男性の回答者の割合が大きいことがわかります。

D.3　チェックボックス形式の回答の棒グラフ

最後に，チェックボックス形式で回答を得たデータで棒グラフを作成してみます。ここでは，第 8 章 3 節で使用した,「チロルチョコ」のデータを使用します。チェックボックス形式で得たデータは，第 8 章 3 節で説明した自作関数 `form2ca()` を使用するか，本稿 3 ページで説明した自作関数 `MAtotal()` を使用して，頻度を求めます。

```
MA<-MAtotal(chirol[,3],sep=",")
barplot(MA[,1])
```

上記のコードを実行すると，図 5 に示す「チロルチョコ」データの 3 列目の項目「チョコバナナ」のイメージの棒グラフが描かれます。図 5 から,「チョコバナナ」は，男性的で親近感があるという印象を持たれていることがわかります。

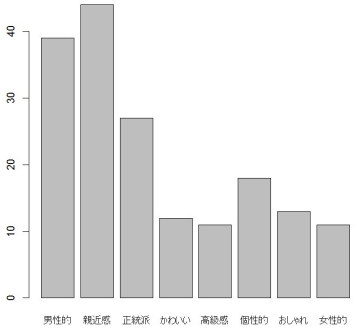

図 5: チョコバナナのイメージ