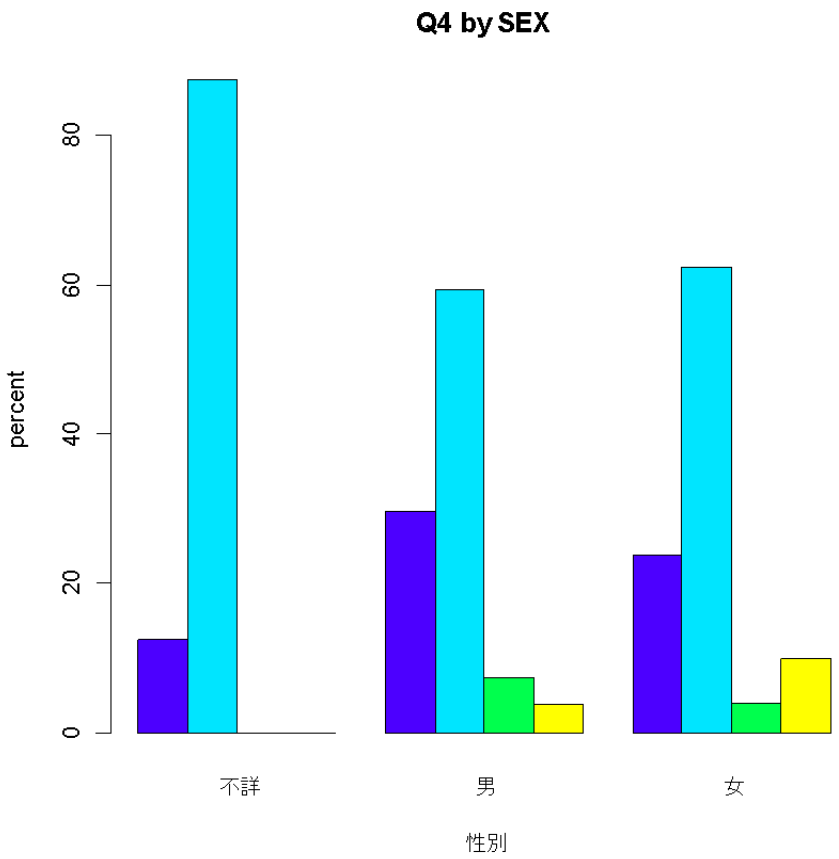
Ｒによる調査データの整理



barplot(100*prop.table(table(Q4, 性別), 2), main="Q4 by SEX",xlab="性別", names=c("不詳","男","女"), ylab="percent", col=topo.colors(4), beside=T)

Rによる調査データの整理 †

Rのインストール †

http://cran.md.tsukuba.ac.jp/bin/windows/base/R-2.15.0-win.exe ・・・　ftpサイトではないので、ブラウザ上でクリックするだけでセットアッププログラムをダウンロードできる。
http://www.r-project.org/ ・・・　Rプロジェクトのサイト
http://cran.md.tsukuba.ac.jp/　・・・　筑波大学のR関係サイト。ウィンドウズ版だけでなく、Linux版やMacOS X版も置かれている。

主な参考書

私がＲをまなぶ上で参考にしている文献のうちで主なものは下記の通り。

- 舟尾・高浪『データ解析環境「R」』（工学社、平成１７年）
- 岡田監訳『Rによる医療統計学』（丸善、平成１９年）
- U.リゲス（石田訳）『Rの基礎とプログラミング技法』（シュプリンガー・ジャパン、平成１８年）
- Ｐ．スペクター（石田基広・石田和枝訳）『Ｒデータ自由自在』（シュプリンガー・ジャパン、平成２０年）
- Braun and Murdock, *A First Course in Statistical Programming with R*, Cambridge, 2007.

Rの使い方(その１) †

Rの電卓的な使い方 †

+,-,/,*,^,sqrt()などの記号や関数が使える。

代入

変数への値の代入の記号は　「　<－　」である。
例：　x1 <- c(3,4,5,6,7)
x1 <- scan() を実行し、数字を順に入力してもよい。

```
> x1 <- c(3, 4, 5, 6, 7)
> x1
[1] 3 4 5 6 7
>
```

変数の確認（オブジェクトの一覧表示）

ls() 又は　objects()

関数

x1の平均は、mean(x1) 、標準偏差は、sd(x1)。なお、sdのdenominatorは、nではなく「 n - 1」。

ベクトル成分の取り出し

ベクトル名の後に「[」と「]」の記号を利用。　x[2]とすると、２番目の4が取り出される。
c(3,4,5,6,7)　は、コロン（「:」）を使って3:7としてもよい。c()は、値を並べるときに使う。

行列

- 行列の作成
  matrix(1:6, nrow=2, ncol=3)
  matrix(1:6, nrow=2, ncol=3, byrow=T) .....　byrow=Tとすると、左から右へ。
  m1 <- matrix(1:6, nrow=2) 行数・列数のどちらかを指定すればよい。

1	3	5
2	4	6

- 行列成分の取り出し
  m1[a,b] とすると、行列m1のa行b列の値が取り出される。
  m1[a,]はa行だけ、m1[,b]はb列だけが取り出される。

```
> m1
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> m1[1,2]
[1] 3
> m1[2,3]
[1] 6
> m1[1,]
[1] 1 3 5
> m1[2,]
[1] 2 4 6
> m1[,1]
[1] 1 2
```

- 周辺度数の追加
  m2 <- addmargins(m1,1) .....　列の計が付け加わる。

m2 <- addmargins(m1,c(1,2)) .....　列の計と行の計が付け加わる。

- 行や列ごとの計や平均の計算
  apply(m1, 2, sum) .....　各列の計
  colSums(m1) .....　上記と同じ結果
  apply(m1, 2, mean) .....　各列の平均
  apply(m1, 1, sum) .....　各行の計
  rowSums(m1) .....　上記と同じ結果

クロス表の検定（分割表における独立性の検定）

m3 <- matrix(c(24,12,12,18,9,5), 2) .....　最後の２は行数の指定（nrow=2）

24	12	9
12	18	5

この表の作成は、「scan()」を使い下記のようにして１行ずつ入力する方が簡単。
m3 <- matrix(scan(),ncol=3, byrow=T)

chisq.test(m3) .....　カイ２乗検定

```
> m3 <- matrix(c(24,12,12,18,9,5),nrow=2)
> m3
     [,1] [,2] [,3]
[1,]   24   12    9
[2,]   12   18    5
> chisq.test(m3)

        Pearson's Chi-squared test

data:  m3
X-squared = 5.1737, df = 2, p-value = 0.07526

>
```

chisq.test(m3)$expected .....　期待度数の表示
chisq.test(m3)$observed .....　観測度数の表示
fisher.test(m3) .....　フィッシャーの直接確率計算法、正確検定(exact test)

比率の差の検定

278/800と242/745の差を検定する。
prop.test(c(278,242), c(800, 745), correct=F, alternative="greater")
連続性の補正(continuity correction)、前者が後者「より大きい」(greater)という対立仮説を設定（つまり片側検定）。

```
> prop.test(c(278,242),c(800,745))

        2-sample test for equality of proportions
        with continuity correction

data:  c(278, 242) out of c(800, 745)
X-squared = 0.7891, df = 1, p-value = 0.3744
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02574150  0.07107707
sample estimates:
   prop 1    prop 2
0.3475000 0.3248322
```

プログラムの終了は　q()

作業スペースを保存するかどうかを選択できる。ファイルは「.RData」。
save.image("*****.RData")としてから終了してもよい。*****の部分は任意。
関連コマンド：getwd() ..... Rの作業ディレクトリの確認。

Rの使い方(その２)

エクセルデータの読み込み

エクセルの表からWindowsのクリップボードを経由してデータを読み込む方法（他に、read.csvを使う方法もある。）
data01 <- read.delim("clipboard", header=F)　.....　data01という部分は任意。変数名を読み込む場合は、「header=TRUE」とするか、省略。「TRUE」は「T」だけでもよい。
データの最初の６人分の表示　head(data01)　..... tail(data01)とすると最後の６人分。

```
> data01 <- read.delim("clipboard")
> head(data01)
  NO Q1.1 Q1.2 Q1.3 Q1.4 Q1.5 Q1.6 Q1.7 Q1.8 Q1.9 Q2 Q3 Q3S1
1  1    1    2    2    1    2    2    0    2    2  1  2    1
2  2    1    1    2    1    1    1    0    2    0  1  2    1
3  3    1    1    2    1    2    2    0    2    2  1  2    1
4  4    1    2    2    1    2    0    0    2    2  1  2    1
5  5    1    1    1    1    1    1    0    2    2  1  2    1
6  6    1    1    1    1    1    2    0    2    0  1  1    1
  Q6.2 Q7 Q8 Q8S1 Q8S2 Q8S3 Q9 Q10 Q10S1 Q10S2 Q10S3 Q11 Q11
1    2  1  1    1    0    0  2   1      1      0      0   2
2    2  3  2    0    0    0  2   3      0      0      0   1
```

集計

- 単純集計　table(data01$V1)　.....　V1,V2などの変数名は、自動的に付けられたもの。変数名はエクセルから読み込んでもよい。
- クロス集計
  t1 <- table(data01$V1, data01$V2)　.....　V1 by V2の２重クロス集計。
  t1 <- table(data01$V1, data01$V2, data01$V3) .....　V1 by V2 by V3の３重クロス集計。
  attach(data01) とすると、「data01$」を省略できる。　(attach, dettach)
  t1 <- table(V1, V2) .....　２重クロス集計
  t1 <- table(V1, V2, V3) .....　３重クロス集計

```
> attach(data01)
> table(Q4)
Q4
 1  2  3  4
33 86  6 11
> table(Q4,学科)
   学科
Q4   0  1  2  3
  1  1 10 10 12
  2  6 28 42 10
  3  0  1  5  0
  4  0  8  2  1
>
```

  周辺度数の表示　margin.table(t1, 2) .....　２は列方向の計を意味する。
- 行パーセント、列パーセントの計算
  100*prop.table(t1, 2) .....　列パーセント
  100*prop.table(t1, 1) .....　行パーセント
  以下のようにすると小数点以下第１位までになるので見やすい（そうしないと第６位まで表示される）。
  round(100*prop.table(t1, 2), digits=1)

データフレームの作成と編集 †

エクセルに似た表でデータを入力してもよい。
data02 <- data.frame()
fix(data02)

Rの使い方（その３） †

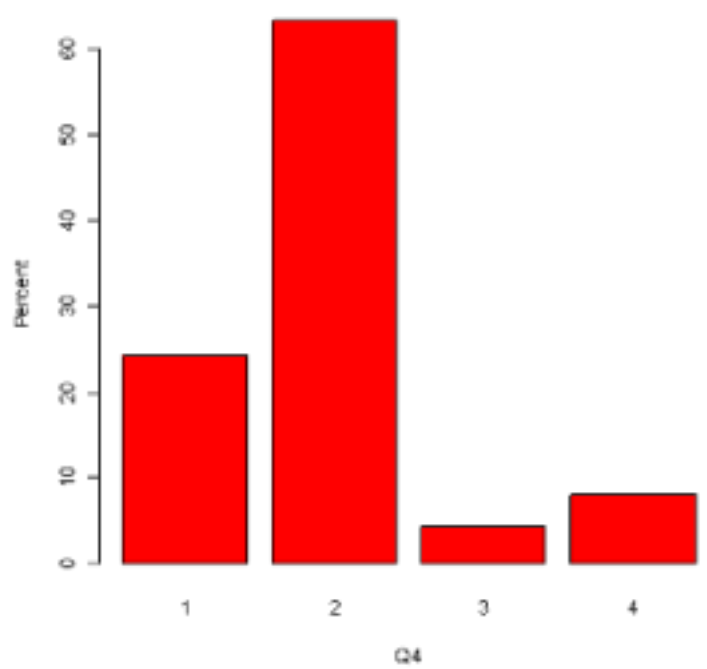
コードの変更（再コーディング） †

var1 <- c(1, 2, 3, 4, 5, 1, 3, 5, 5)
fvar1 <- factor(var1, levels=1:5) .....　ここまでは、データ例の作成――順序尺度のデータ。
levels(fvar1) <- factor(c(1, 1, 1, 2, 3)[fvar1]
1→1
2→1
3→1
4→2
5→3
変数の追加(コード変更した変数のデータフレームへの)
data02 <- transform(data01, V1b=fvar1)

別の方法：carパッケージのrecode関数の利用
var1b <- recode(var1, "c(1,2,3)=1; 4=2; 5=3")　.....「else」も使える。「5=3」は、「else=3」としてもよい。
社会調査のデータではなく成績評価に関連してであるが、つぎのようなこともできる。
d2 <- recode(d1,'0:59="不可";60:69="D";70:79="C";80:89="B";90:100="A"')

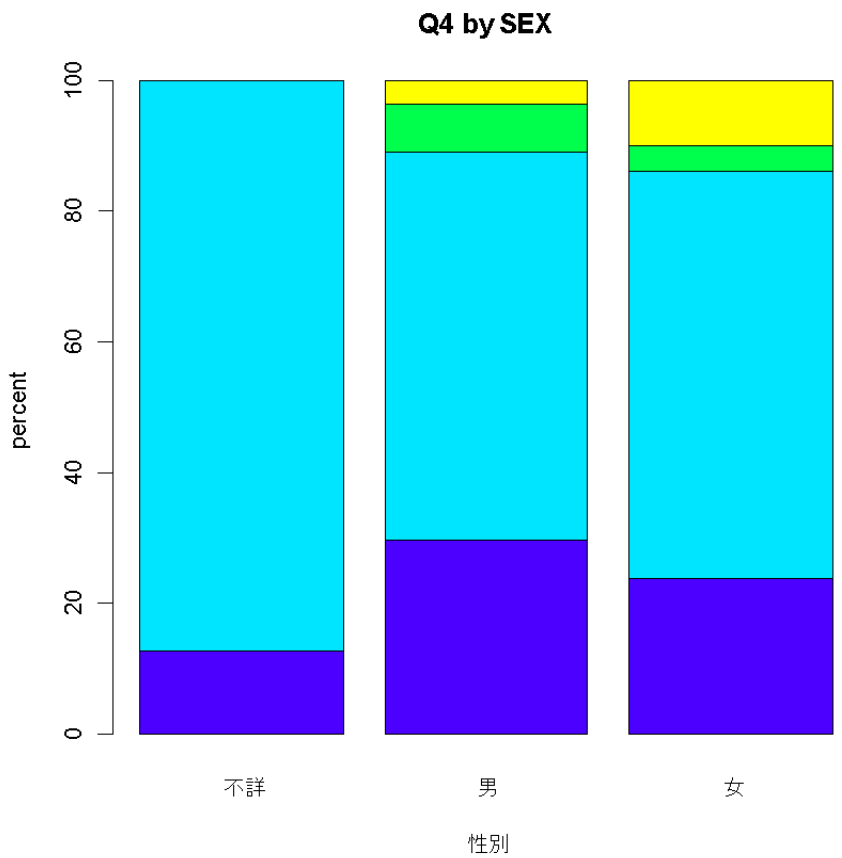
棒グラフの作成 †

- 単純集計結果のパーセント表示
  attach(data01) .....　括弧の中は、分析をしようとする任意のデータフレーム。
  table(data01$Q1) .....　単純集計（実数）。attach(data01)を実行してあれば「data01$」の部分は省略可。グラフを表示させるだけであればこの集計は不要。
  barplot(table(Q1)/sum(table(Q1))*100) .....　グラフの表示（単純集計の合計で除し１００を乗じている）

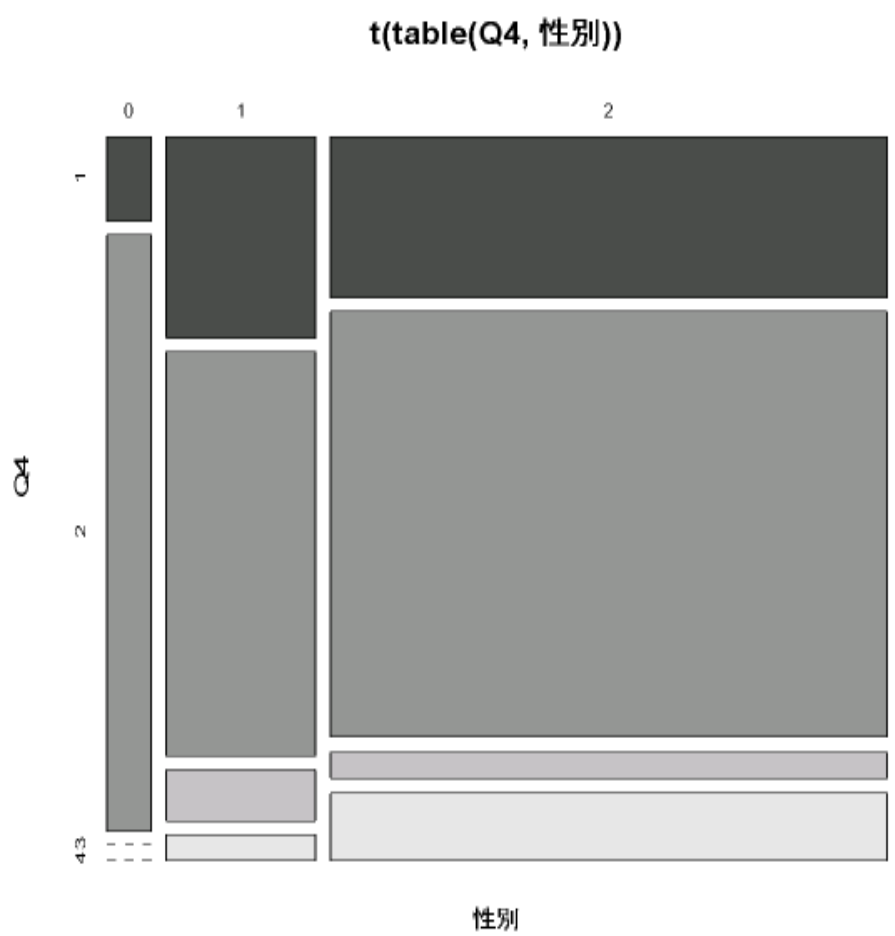


- 列パーセントの表示
  table(Q1, SEX) .....　問１の男女別集計（実数）、Q1, SEXの表記は任意。

barplot(100*prop.table(table(Q1, SEX), 2)) ..... パーセント表示の棒グラフ（「2」は列パーセントの計算を指示）
barplot(100*prop.table(table(Q1, SEX), 2), xlab="性別", names=c("男","女"), ylab="percent", col=("blue","red")) ..... 軸名等を付加、色の指定。



barplot(100*prop.table(table(Q4, 性別), 2), main="Q4 by SEX",xlab="性別", names=c("不詳","男","女"), ylab="percent", col=topo.colors(4))



mosaicplot(t(table(Q4,性別)),color=T)

クロス表からの無回答（コードは0）の削除

t1 <- table(Q1, SEX)
t1 <- t1[-1, -1] ..... 1行目と1列目が削除される。

Rの使い方（その４）

単純集計結果の一括表示

```
for ( i in 1:ncol(data01) ) {cat("***");cat(colnames(data01)
[i]);cat("***");print(table(data01[,i]));cat("\n")}
```

data01という名前のデータフレームの例。最後の"\n"は改行。

```
for ( i in 1:ncol(data01) ) {cat("***");cat(colnames(data01)
[i]);cat("***");print(round(table(data01[,i])/sum(table(data01[,i]))*100,digits=1));cat("\n")}
```

分析のための雛形

```
( ta <- table(Q1) ) .....   単純集計の結果をtaに格納する（実数）。
( ta <- ta[-1] )       .....   無回答「0」がある場合にこれを実行。
( tb <- 100*table(ta)/sum(ta)) ) .....   tbに単純集計結果のパーセント表示。
( tc <- table(Q1, SEX) ) .....   男女別Q1 ( Q1 by SEX)
( tc <- tc[-1,-1] ) .....   1 行目と 1 列目を削除する場合に実行。
100*prop.table(tc, 2) .....   列パーセント表示
chisq.test(tc) .....   カイ２乗検定（分割表における独立性の検定）
fisher.test(tc) .....   フィッシャーの直接確率計算法
```

クラマー係数を関数として定義

```
cr <- function(x) {
  if (dim(x)[1] > dim(x)[2] ) return(sqrt(chisq.test(x)$statistic/sum(x)/(dim(x)[2]-1)))
  else return(sqrt(chisq.test(x)$statistic/sum(x)/(dim(x)[1]-1)))
  }
```

ファイル名「cramer.R」で保存。source("cramer.R")で呼び出して使うことができる。

関数を定義してグラフを作成

- 単純集計結果（パーセント）のグラフ

```
table1 <- function(x,y){
xlabel <- colnames(y)[x]
barplot(table(y[x])/sum(table(y[x]))*100,col="red",xlab=xlabel,ylab="percent")
}
```

x：グラフの横軸の変数番号
y：データフレームの名前（data01など）

- クロス集計結果（パーセント）のグラフ

```
table2 <- function(x,y,z){
xlabel <- colnames(z)[y]
barplot(100*prop.table(table(z[,x],z[,y]),2),col=topo.colors(5),xlab=xlabel,ylab="percent")
}
```

x：グラフの縦軸の変数番号
y：グラフの横軸の変数番号
z：データフレームの名前（data01など）

変数番号の確認（上記の関数を実行するためにリストを作成）

for (i in 1:ncol(data01)) {cat(i);cat("--");cat(colnames(data01)[i]);cat("\n")}

実行例：
1--NO
2--Q1.1
3--Q1.2
（省略）
40--Q14
41--学科
42--学年
43--性別

Rの使い方（その５）

複数回答を許すの場合（多項選択）の処理

（以下では、データフレームの３列目から１０列目までを処理する。0か1が入っている。）

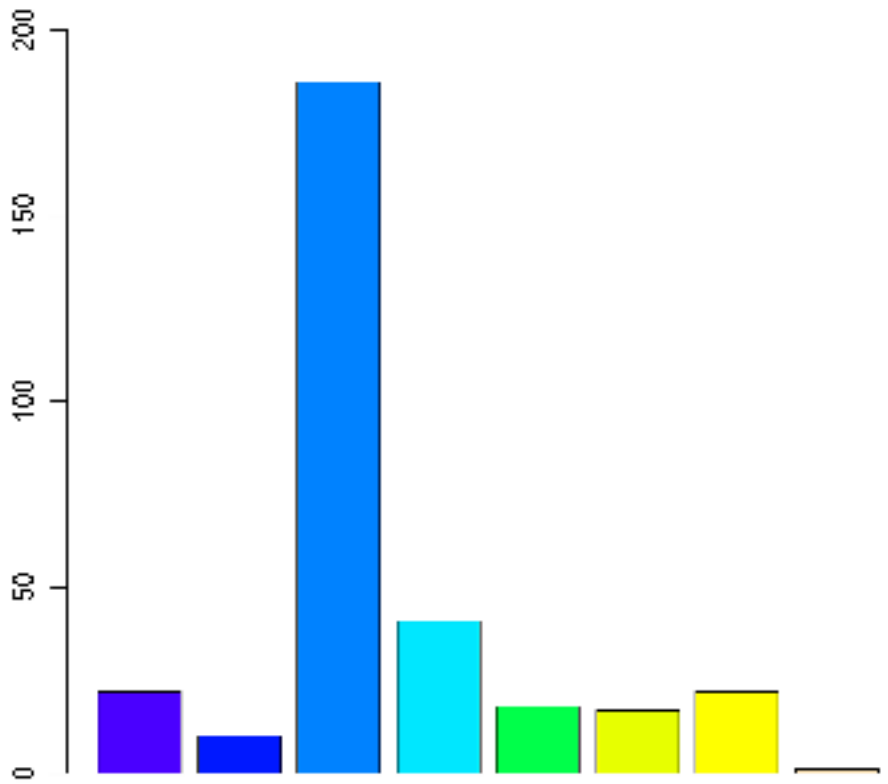
Q1 <- 1:8

Q1[1] <- table(data01[,3])[2]
Q1[2] <- table(data01[,4])[2]
（省略）
Q1[8] <- table(data01[,10])[2]

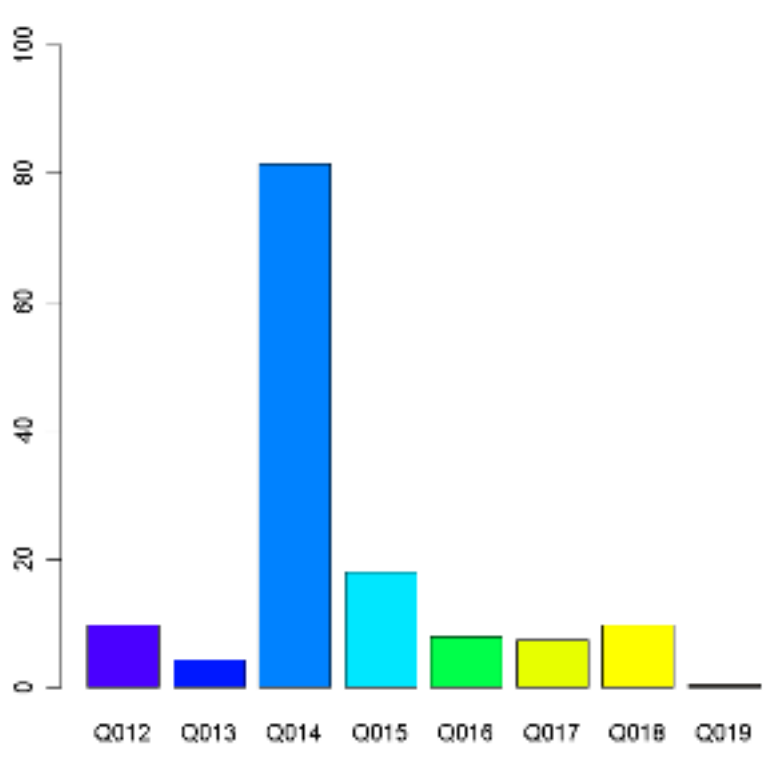
上記をまとめて実行するためには下記のようにする。
for (i in 3:10) {Q1[i-2] <- table(data01[,i])[2] }
しかし、以下のようにする方が簡単。
Q1 <- colSums(data01[,3:10])

クロス集計の例
m1 <- matrix(1:48,nrow=16)
for(i in 1:16){(m1[i,] <- table(data01[,i+19],data01[,40])[2,-4])}

グラフを作成
barplot(Q1,col=topo.colors(8),ylim=c(0,200))　.....　ylimで縦軸の範囲を指定。



パーセントで表示
barplot(Q1/228*100,col=topo.colors(8),ylim=c(0,100))



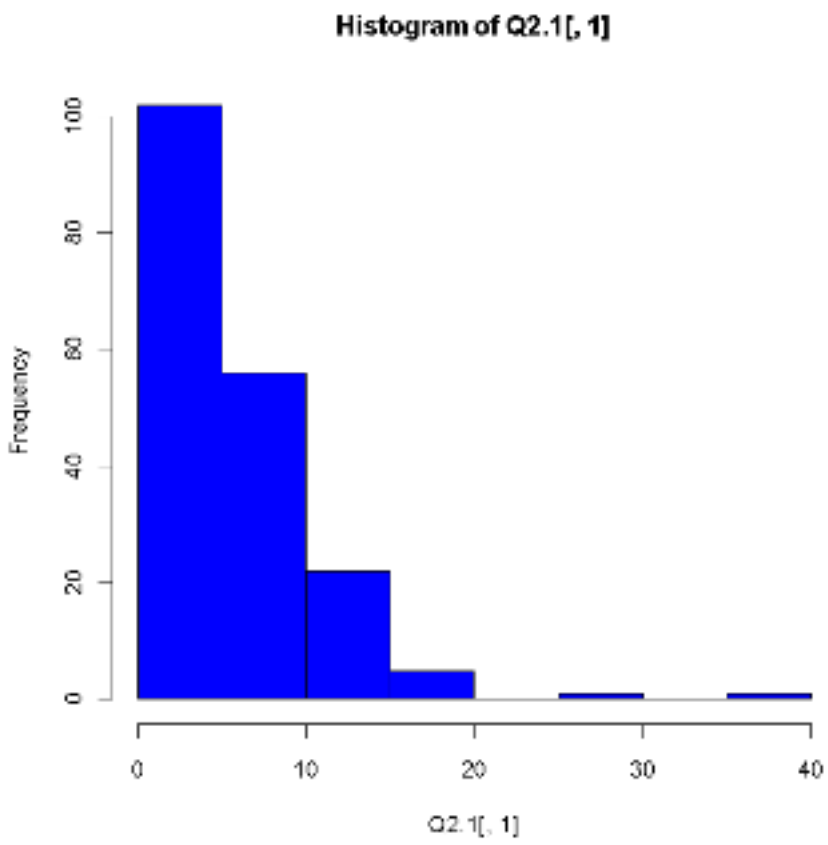
どの選択肢も選んでない人を調べるために以下のようにした。
q1 <- data01[,3:10]　.....　3列目から10列までのデータを取り出す。
a <- rowSums(q1)　.....　rowSumsは行の計、colSumsは列の計。
table(a) .....　aの単純集計をして、「0」の人数を見る。
データフレームdata01の３列目から１０列までに、まとめて扱いたい変数がある場合。
該当者は２名、選択数１が１７４名、選択数２が３２名などということがわかる。
sum(table(a))-table(a)[1]が228なので、これを、パーセントを計算する場合の分母にした。「多項選択」で１項目以上選択した人の数を分母にしている。

円グラフの作成 †

table(Q011)
pie(table(Q011),col=rainbow(2),labels=c("利用する","利用しない"))

ヒストグラムの作成 †

無回答（90）、非該当（99）を外すための操作を最初にしている。（バス停までの距離を分で回答してもらったときの例）
Q2.1 <- subset(data01,Q021<90,Q021)
hist(Q2.1[,1],col="blue")



summary(Q2.1)
sd(Q2.1)

パワーポイント版のＰＤＦファイル

- Ｒによる調査データの整理.pdf
- Ｒによる調査データの整理(つづき）.pdf

Last-modified: 2012-05-02 (水) 10:56:58 (1495d)

Site admin: Moteki Yutaka