統計データ処理

HI4 45 号 山口惺司

実施日：2024/04/24

2024/05/01

レポート提出日：2024/05/11

1. 実験目的

Rの統計処理に関するプログラミングを理解し,2変量(変数)データまでの統計処理ができる.

2. 課題

2.1. 課題1

Excelの CSV 形式のデータの入力例について,データの一部,スクリプト,実行結果を説明せよ.
Excelのデータを表1に示す.

表1 Excelサンプルデータ

Sex	ht	wt	high
F	170.4	66.8	high
F	171.3	66.8	high
F	159.1	58.1	low
F	145.9	49	low
M	171	83.3	high
M	175.8	78.3	high
M	170.1	55.2	high
M	165.7	71	low

ソースコード：

```
data <- read.csv("exampledata.csv")
print(data)
```

説明：

一行目で Excelのデータを読みこみ,二行目で出力している.

2.2. 課題2

テキスト2の6章~9章の課題1,2のそれぞれについて,スクリプト,実行結果を示し,説明せよ.ただし,6章の課題1については,**収縮期血圧,拡張期血圧,ヘモグロビン A1c** のヒストグラムとボックスプロットを作成せよ.

6章：

課題1.

demodata.csvのなかのデータの収縮期血圧：sbp, 拡張期血圧：dbp, ヘモグロビン A1c：ha1c,のヒストグラムとボックスプロットを描け.

ソースコード：

```
data <- read.csv("demodata.csv")
hist(data$sbp)
boxplot(data$sbp)
```

```
hist(data$dbp)
boxplot(data$dbp)
hist(data$ha1c)
boxplot(data$ha1c)
```

実行結果：

図 1~6 に示す.

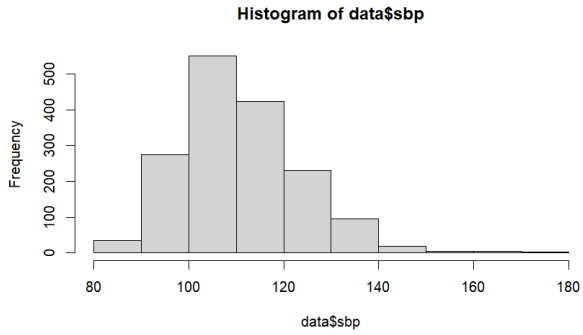

図 1 収縮期血圧:sbp のヒストグラム

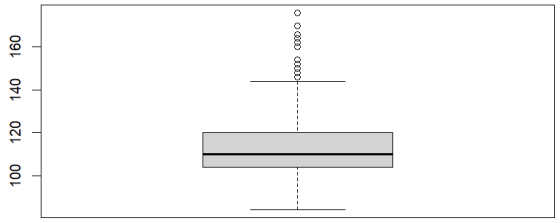

図 2 収縮期血圧:sbp のボックスプロット

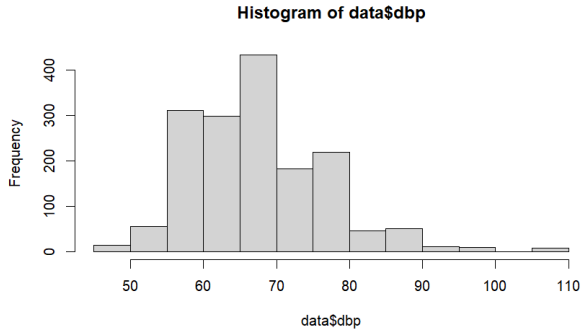

図 3 拡張期血圧:dbp のヒストグラム

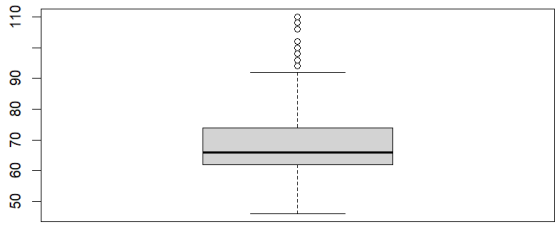

図 4 拡張期血圧:dbp のボックスプロット

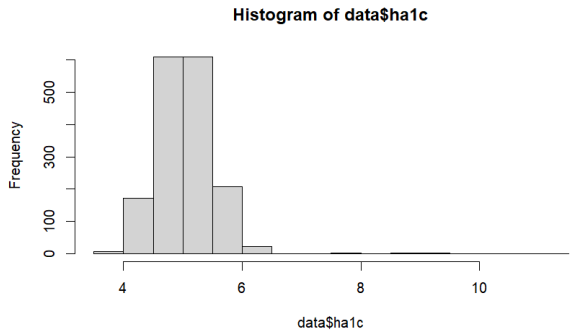

図 5 ヘモグロビン A1c:ha1c のヒストグラム

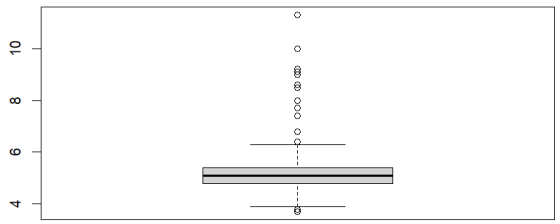

図 6 ヘモグロビン A1c:ha1c のボックスプロット

説明：

任意の要素について hist 関数でヒストグラム,boxplot 関数でボックスプロットをしている.

課題 2.

動脈硬化指数(AI)は以下のように定義される.この指数の要約統計量を求め,ヒストグラムとボックスプロットを描け.

$$\text{動脈硬化指数} = \frac{TC-HDL\_C}{HDL\_C}$$

ソースコード：

```
data <- read.csv("demodata.csv")

tc <- data$tc
hdlc <- data$hdlc

ai <- (tc - hdlc) / hdlc

hist(ai)
boxplot(ai)
```

実行結果：

図 7,8 に示す.

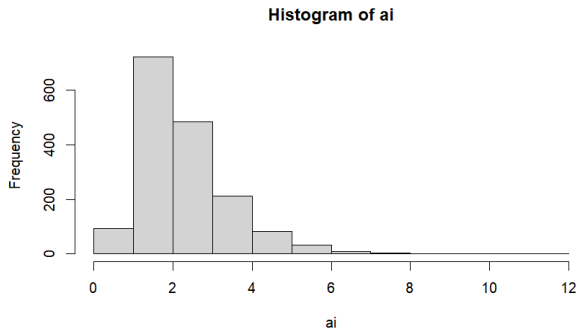

図 7 動脈硬化指数:AI のヒストグラム

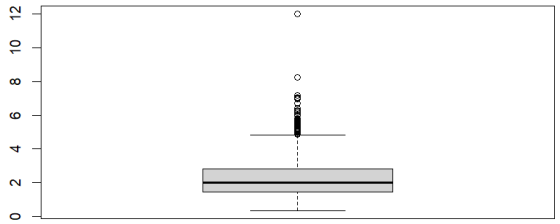

図 8 動脈硬化指数:AI のボックスプロット

説明：

変数 tc,hdlc に data の tc と hdlc を取り出し,代入している.

hist()関数でヒストグラム,boxplot()関数でボックスプロットをしている.

7 章：

課題 1.

x = c(1,2,3,4,5,6)のなかで,以下の条件式を満たす成分を取り出す式と結果を記せ.

(1) 3 より大きく,5 より小さい

式： x[c((3 < x) & (x < 5))]

結果：4

(2) 3 より小さいか,5 より大きい

式： x[c((3 > x) | (x > 5))]

結果：1 2 6

(3) 3 以下か,5 以上

式： x[c((x <= 3) | (x >= 5))]

結果：1 2 3 5 6

(4) 2 と 6 でない

式： x[c((x != 2) | (x != 6))]

結果：1 2 3 4 5 6

(5) 3 ではなく,かつ 1 以上 5 以下

式： x[c((x != 3) & (x >= 1) & (x <= 5))]

結果：1 2 4 5

説明：

<,>,<=,>=,!=,==,&,|などの演算子を用いて,条件式を満たす成分を取り出している.

8 章：

課題 1. minidata.csv を使って以下の問いに答えよ.

(1) 身長 150cm 未満の行データのみ抜き出す式を書け.

式： data[ht < 150,]

(2) 身長 150cm 以上，170cm 未満の行データのみ抜き出す式を書け.

式： data[ht >= 150 & ht < 170,]

(3) 身長 150cm 以上，170cm 未満で，女性のデータのみ抜き出す式を書け.

式： data[ht >= 150 & ht < 170 & sex == 'f',]

課題 2. demodata.csv を使って，以下の問いに答えよ.

(1) 男性のデータを変数”mdata”，女性のデータを変数”fdata”とするように式を書け.

式： mdata <- data[data$sex == "m",]

fdata <- data[data$sex == "f",]

(2) 男性の身長 ht,体重 wt のヒストグラムを描け.

ソースコード：

```
hist(mdata$ht)
hist(mdata$wt)
```

実行結果：

図 9,10 に示す.

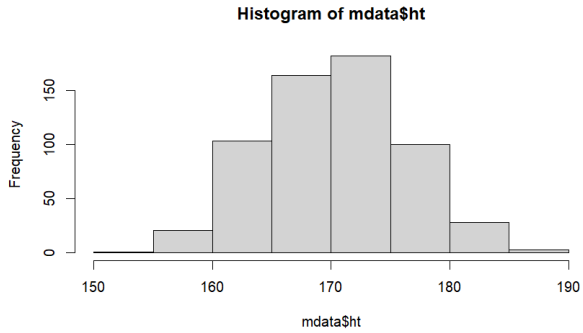

図 9 男性の身長 ht のヒストグラム

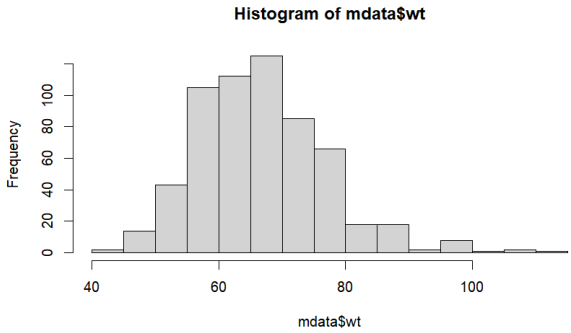

図 10 男性の身長 wt のヒストグラム

(3) 女性の身長 ht,体重 wt のヒストグラムを描け.

ソースコード：

```
hist(fdata$ht)
hist(fdata$wt)
```

実行結果：

図 11,12 に示す.

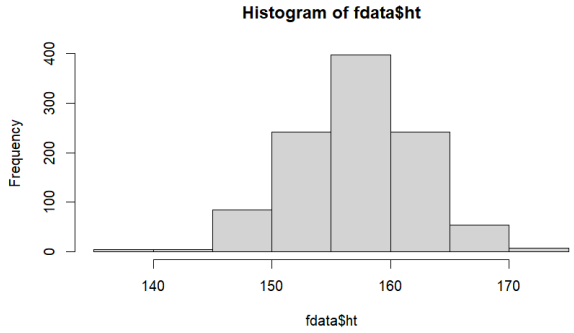

図 11 女性の身長 ht のヒストグラム

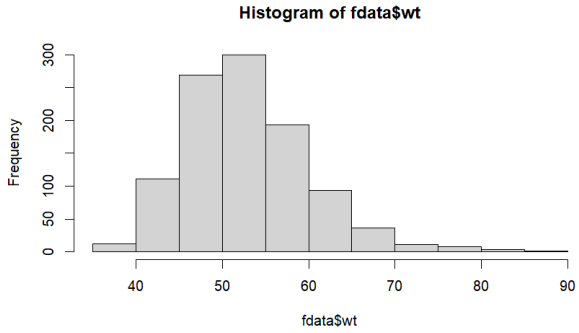

図 12 女性の身長 wt のヒストグラム

(4) 男性の身長 ht,体重 wt の要約統計量（平均・標準偏差・メジアン・四分位範囲）を求めよ.

ソースコード：

```
print(summary(mdata$ht))
print(sd(mdata$ht))
print(summary(mdata$wt))
print(sd(mdata$wt))
```

実行結果：

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  151.2   165.8   170.4   170.2   174.4   186.7
[1] 5.942523

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  43.90   59.42   66.10   66.92   72.80  110.30
[1] 10.25974
```

(5) 女性の身長 ht,体重 wt の要約統計量（平均・標準偏差・メジアン・四分位範囲）を求めよ.

ソースコード：

```
print(summary(fdata$ht))
print(sd(fdata$ht))
print(summary(fdata$wt))
print(sd(fdata$wt))
```

実行結果：

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  137.2   153.9   157.4   157.2   160.6   172.7
[1] 5.223774

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  36.30   47.70   52.05   52.92   57.00   85.60
[1] 7.238793
```

説明：

条件を満たす任意の要素を取り出し,hist()関数や boxplot()関数を用いてグラフにしている.
また,summary()関数,sd()関数を用いて要約統計量（平均・標準偏差・メジアン・四分位範囲）を求めている.

9 章：

課題 1： demodata.csv のデータについて以下の問いに答えよ.
関数 cut( )を使うと，量的変数を質的変数に変換することができる.
収縮期血圧 sbp を質的変数に置き換えて,sbpclass という変数に入れる.
sbpclass=cut(data$sbp, breaks=c(120,130,140,160,180), right=F)
同様に，拡張期血圧も質的変数に置き換えて，dbpclass という変数に入れる.
dbpclass=cut(data$dbp,breaks=c(0,80,85,90,100,110,Inf),right=F)

(1) こうしてできた 2 つの質的変数 sbpclass と dbpclass を要約せよ.

ソースコード：

```
print(table(sbpclass, dbpclass))
```

実行結果：

```
                dbpclass
sbpclass    [0,80) [80,85) [85,90) [90,100) [100,110) [110,Inf)
  [120,130)    198      58      11        5         0         0
  [130,140)     55      36      21       15         0         0
  [140,160)      3       9       4       12         4         0
  [160,180)      0       0       0        2         4         3
```

(2) 変数 sex と sbpclass を要約せよ.

ソースコード：

```
print(table(data$sex, sbpclass))
```

実行結果：

```
   Sbpclass
    [120,130) [130,140) [140,160) [160,180)
f       150        64        18         4
m       122        63        14         5
```

(3) 変数 sex と dbpclass を要約せよ.

ソースコード：

```
print(table(data$sex, dbpclass))
```

実行結果：

```
   dbpclass
    [0,80) [80,85) [85,90) [90,100) [100,110) [110,Inf)
f    957      47      15       14         3         2
m    479      75      22       20         5         1
```

課題 2： demodata.csv のデータについて以下の問いに答えよ.

(1) BMI (Body Mass Index)を表す新しい変数 bmi を定義する式を書け.

式：bmi <- data$wt / (data$ht/100)^2

(2) 変数 bmi と fat の散布図と相関係数を求めよ.

ソースコード：

```
plot(bmi, data$fat)
print(cor(bmi, data$fat))
```

実行結果：

```
[1] 0.7021726
```

出力した散布図を図 13 に示す.

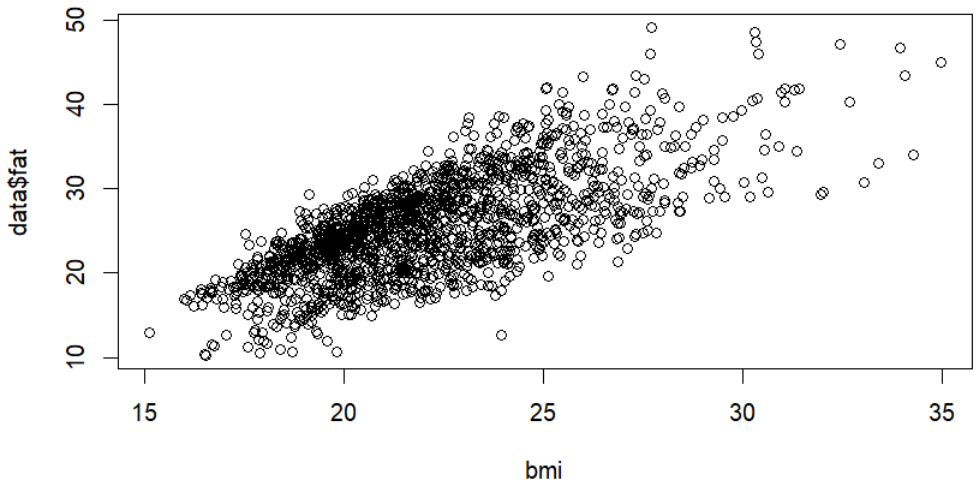

図 13 bmi と fat の散布図

(3) 変数 fat と tc の散布図と相関係数を求めよ.

ソースコード：

```
plot(data$fat, data$tc)
print(cor(data$fat, data$tc))
```

実行結果：

```
[1] 0.2163313
```

出力した散布図を図 14 に示す.

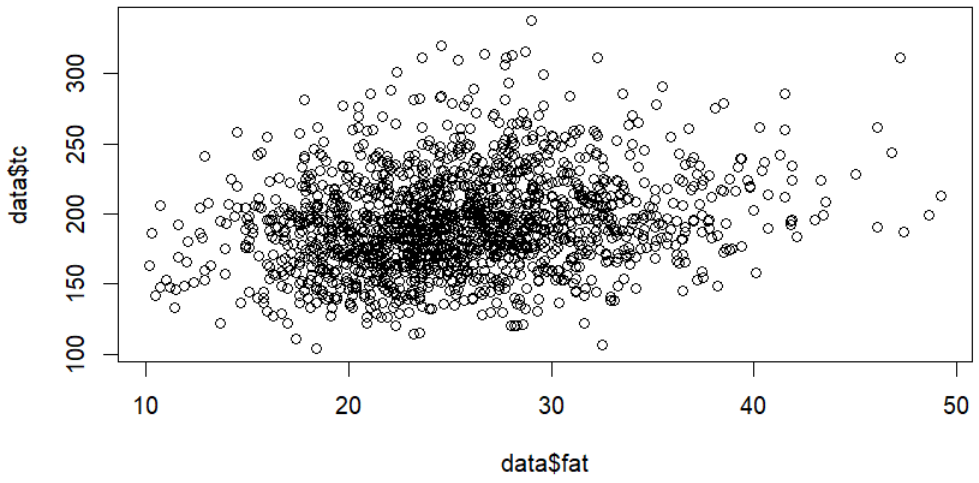

図 14 fat と tc の散布図

(4) 変数 fat と ggt の散布図と相関係数を求めよ.

ソースコード：

```
plot(data$fat, data$ggt)
print(cor(data$fat, data$ggt))
```

実行結果：

```
[1] 0.01587683
```

出力した散布図を図 15 に示す.

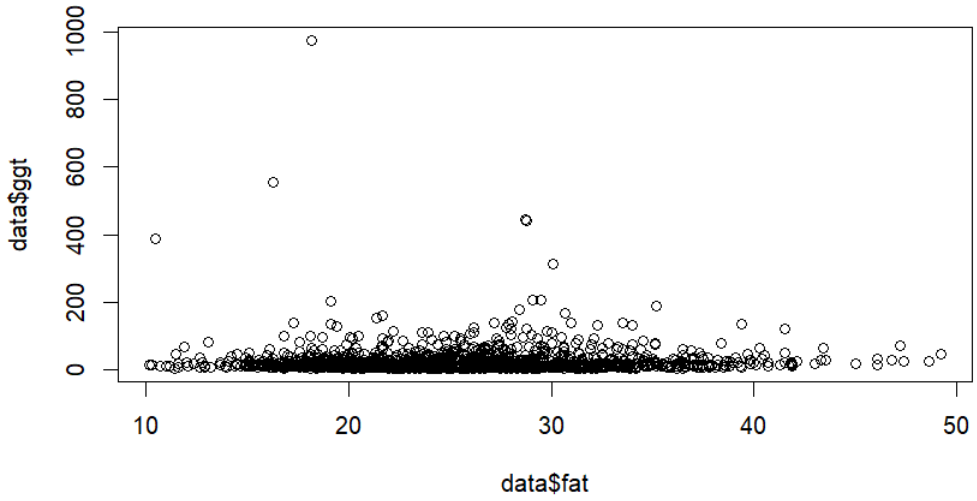

図 15 fat と ggt の散布図

説明：

cut()関数で sbp と dbp を区切り,それぞれ新しい変数 sbpclass,dbpclass に入れた.

table()関数にて,2 つのデータを要約している.

cor()関数で,相関係数を求めている.

plot()関数で,散布図を描いている.

2.3. 課題 3

R のデータセット iris(教科書 p.104, 105 参照)についてデータの要約を行い,その実行例について,スクリプト,実行結果を示し,説明せよ.

ソースコード：

```
print(summary(iris))
panel.pearson <- function(x, y, ...) {
   horizontal <- (par("usr")[1] + par("usr")[2]) / 2;
   vertical <- (par("usr")[3] + par("usr")[4]) / 2;
   text(horizontal, vertical, format(abs(cor(x,y)), digits=2))
}
plot(iris[1:4], main = "Edgar Anderson's Iris Data", pch = 21, bg =
c("red","green3","blue")[unclass(iris$Species)], upper.panel=panel.pearson)
```

実行結果：

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
```

```
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

出力したグラフを図 16 に示す.

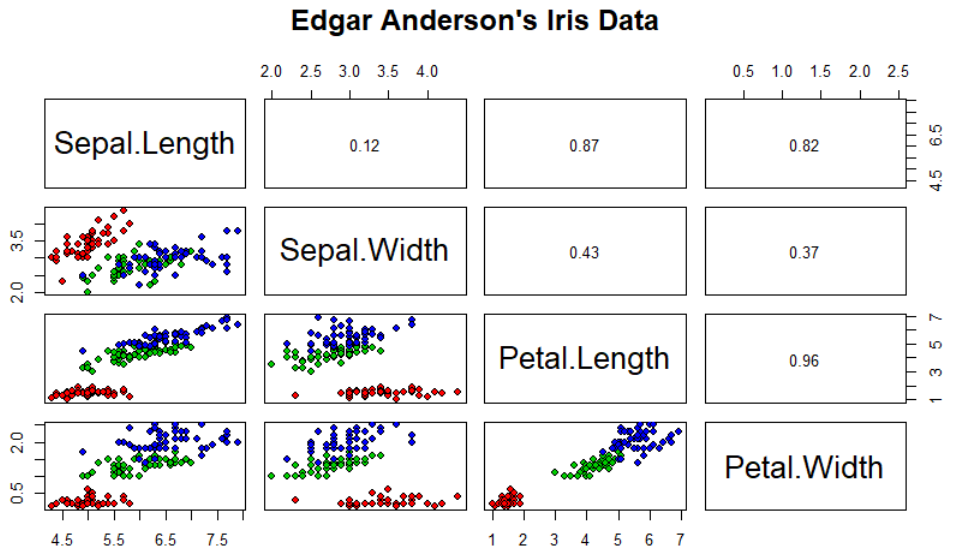

図 16 データセット iris の要約

説明：

summary()関数でデータの要約をしている.

グラフではウォーリック大学の「Plotting the Iris Data」という記事を参考にし,右上の方に相関係数,左下の方に散布図を表示させている.

2.4. 課題 4

demodata.csv の中の収縮期血圧 sbp,拡張期血圧 dbp を図 17 のようにカテゴリー化せよ.その際,「正常血圧」=bp1,「正常高値血圧」= bp2, 「高値血圧」=bp3,「I 度高血圧」=bp4,「II 度高血圧」=bp5,「III度高血圧」=bp6 と命名し,それぞれのカテゴリーに入る人を数えよ.

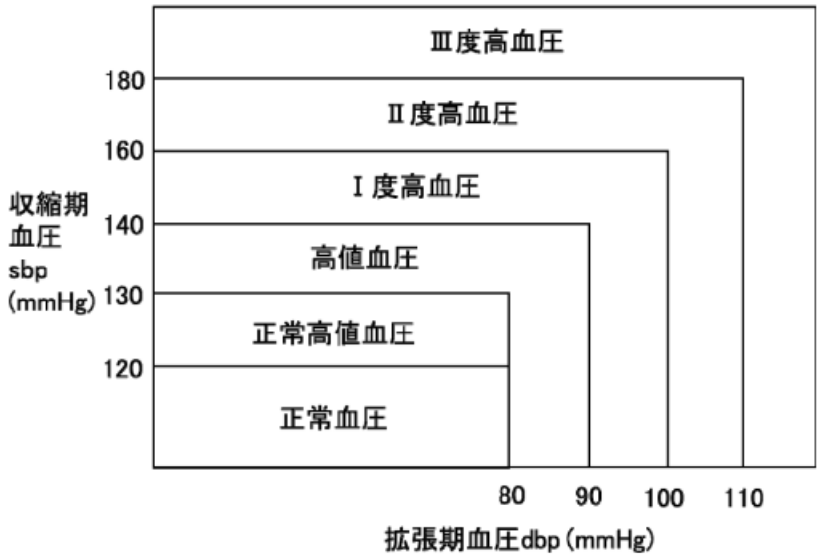

図 17 拡張期血圧と収縮期血圧のカテゴリー

ソースコード：

```
data <- read.csv("demodata.csv")
dbp <- data$dbp

sbp <- data$sbp
id <- data$id

bp1 <- id[(dbp < 80) & (sbp < 120)]
bp2 <- id[(dbp < 80) & (sbp < 130)]
bp3 <- id[(dbp < 90) & (sbp < 140)]
bp4 <- id[(dbp < 100) & (sbp < 160)]
bp5 <- id[(dbp < 110) & (sbp < 180)]
bp6 <- id[(110 <= dbp) | (180 <= sbp)]

bp5 <- setdiff(bp5, bp4)
bp4 <- setdiff(bp4, bp3)
bp3 <- setdiff(bp3, bp2)
bp2 <- setdiff(bp2, bp1)

print(length(bp1))
print(length(bp2))
print(length(bp3))
print(length(bp4))
print(length(bp5))
print(length(bp6))
```

実行結果：

```
[1] 1180
[1] 198
[1] 201
[1] 48
[1] 10
[1] 3
```

説明：

id をもとにカテゴリー化させた.

setdiff()関数は setdiff(a, b)のような使い方をし,これは a の要素から b の要素を取り除くというものである.

この関数を使い,重複する id を取り除いている.

3. 感想

今まで Python でグラフの作成をしていたが,R を使った方がより簡単にグラフを作成することができるため,驚いた.

データ処理に必要な関数がデフォルトで豊富に入っており,やはり R はデータ処理に適した言語なのだと改めて感じた.

4. 参考文献

University of Warwick Plotting the Iris Data

https://warwick.ac.uk/fac/sci/moac/people/students/peter_cock/r/iris_plots/