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

# **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.

# **課題**

* 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.

* 1. 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.

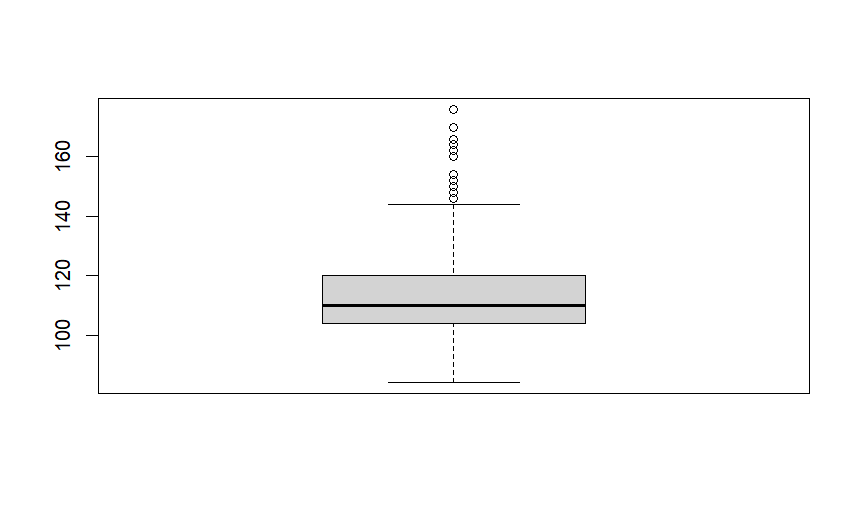
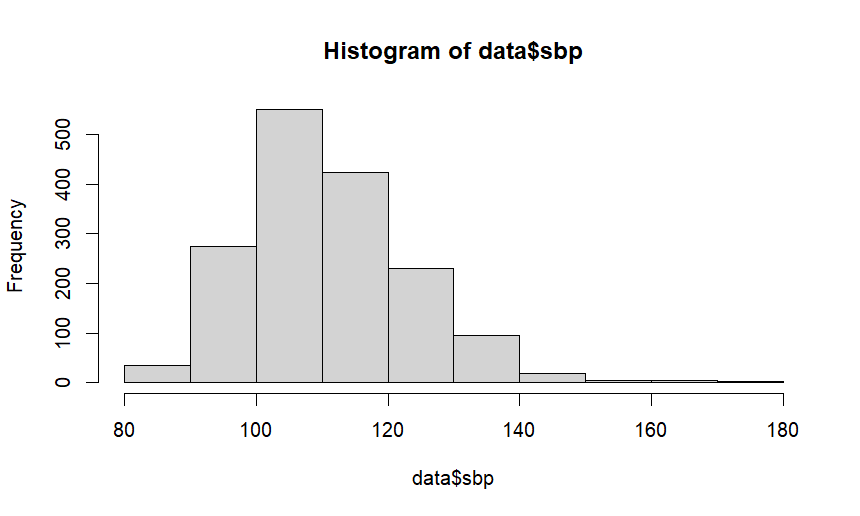


図1 収縮期血圧:sbpのヒストグラム 図2 収縮期血圧:sbpのボックスプロット

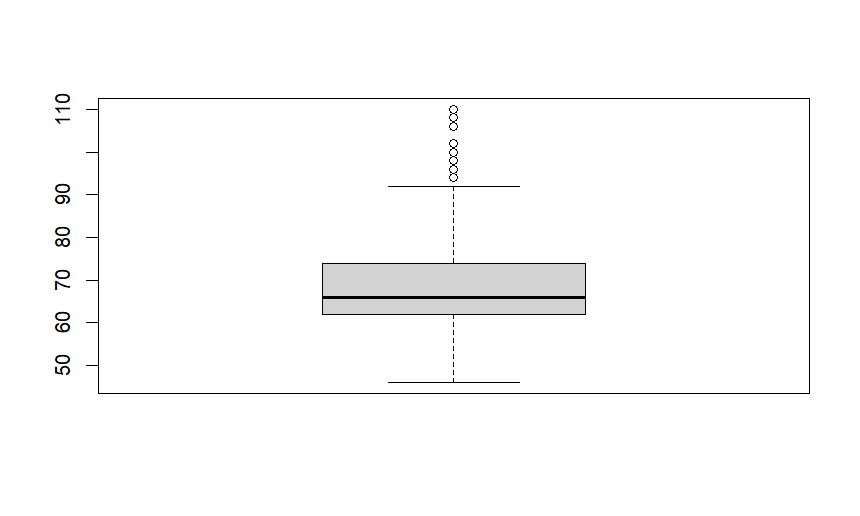
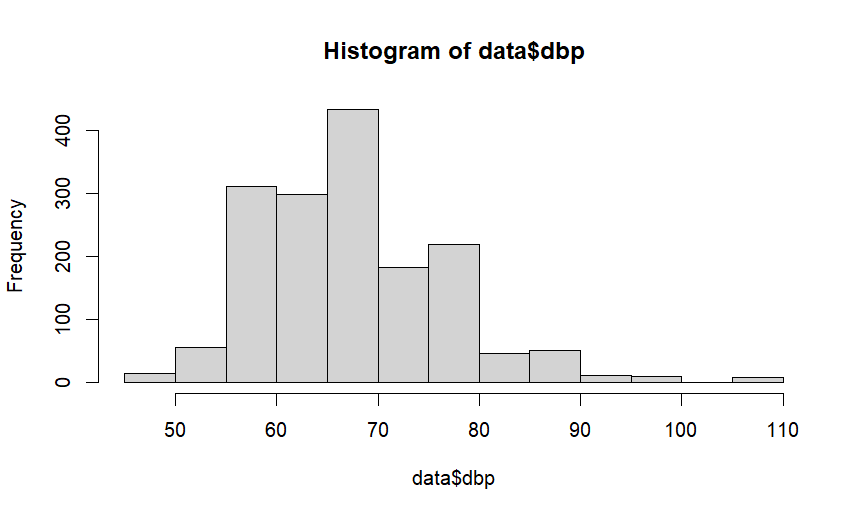


図3拡張期血圧:dbpのヒストグラム 図4 拡張期血圧:dbpのボックスプロット

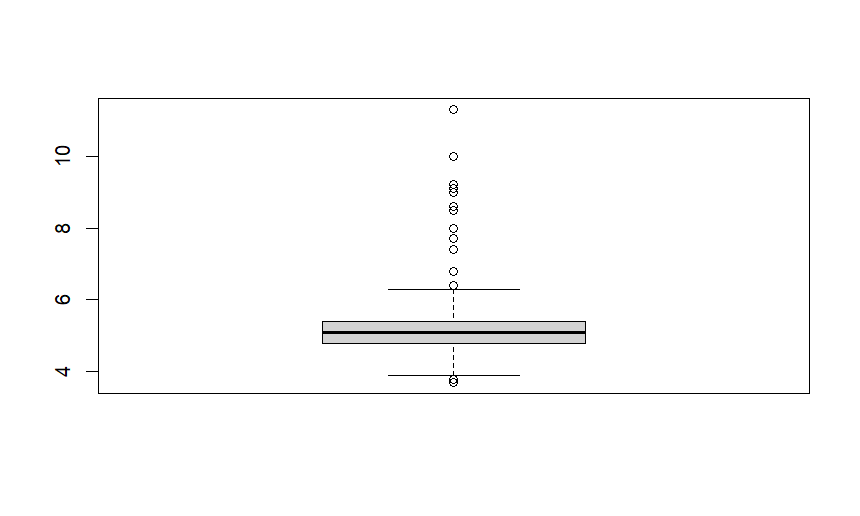
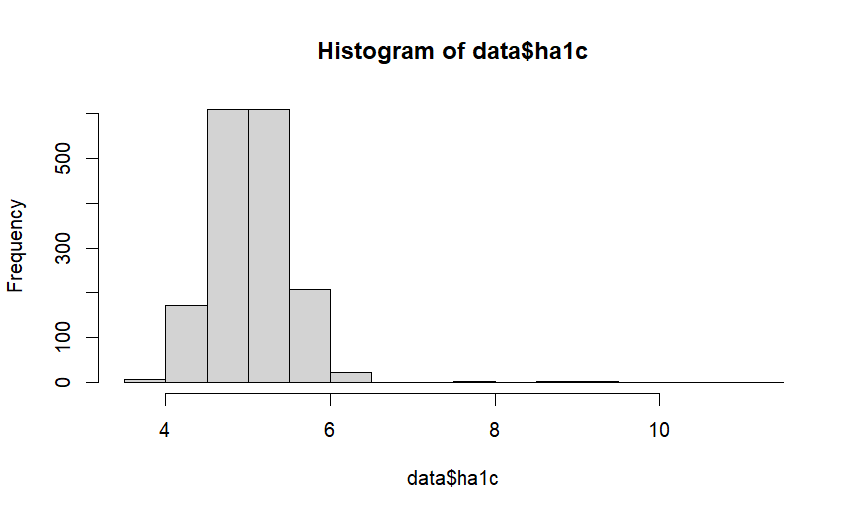


図5 ヘモグロビンA1c:ha1cのヒストグラム 図6 ヘモグロビンA1c:ha1cのボックスプロット

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.

動脈硬化指数

ソースコード：

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

tc <- data$tc

hdlc <- data$hdlc

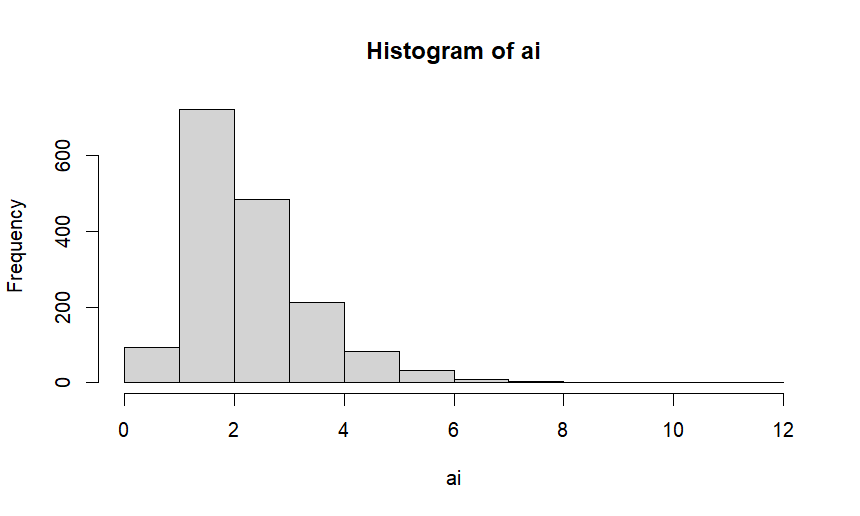
ai <- (tc - hdlc) / hdlc

hist(ai)

boxplot(ai)

実行結果：

図7,8に示す.

ダイアグラム が含まれている画像

自動的に生成された説明

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

1. 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

1. 3以下か,5以上

式： x[c((x <= 3) | (x >= 5))]

結果：1 2 3 5 6

1. 2と6でない

式： x[c((x != 2) | (x != 6))]

結果：1 2 3 4 5 6

1. 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,]

1. 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,]

1. 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",]

1. 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.

グラフ, ヒストグラム

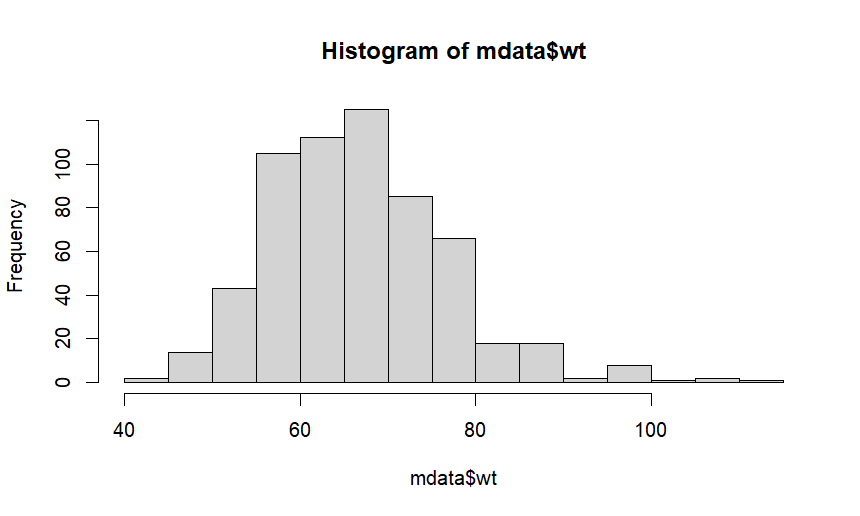
自動的に生成された説明

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

1. 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.

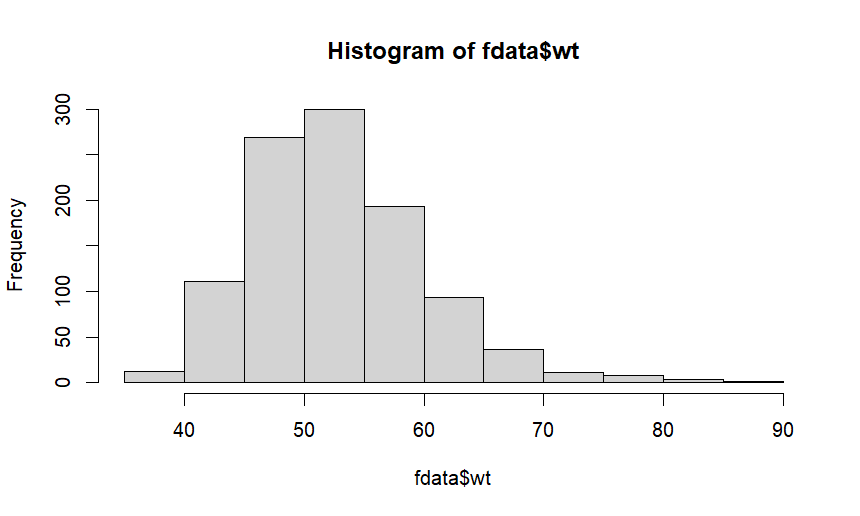
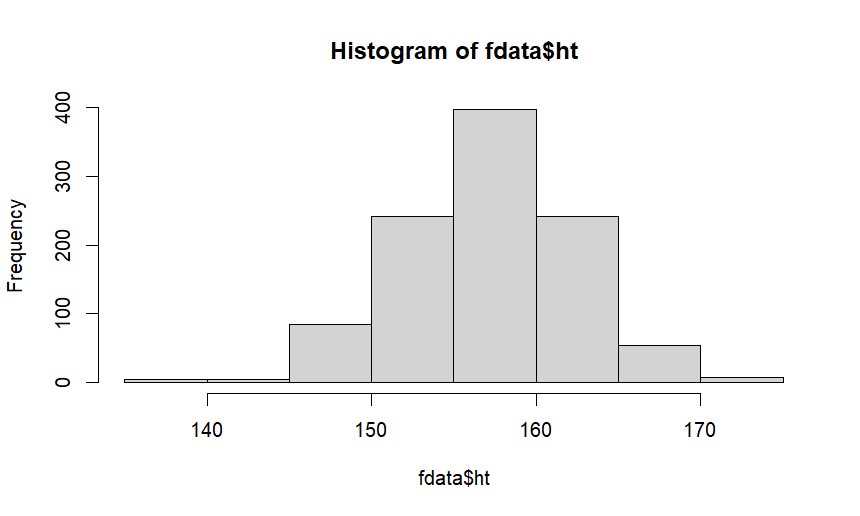


図11 女性の身長htのヒストグラム 図12 女性の身長wtのヒストグラム

1. 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

1. 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

1. 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

1. 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

1. 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.

グラフ, 散布図

自動的に生成された説明

図13 bmiとfatの散布図

1. 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.

グラフ, 散布図

自動的に生成された説明

図14 fatとtcの散布図

1. 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.

グラフ, 散布図

自動的に生成された説明

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.

* 1. 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.

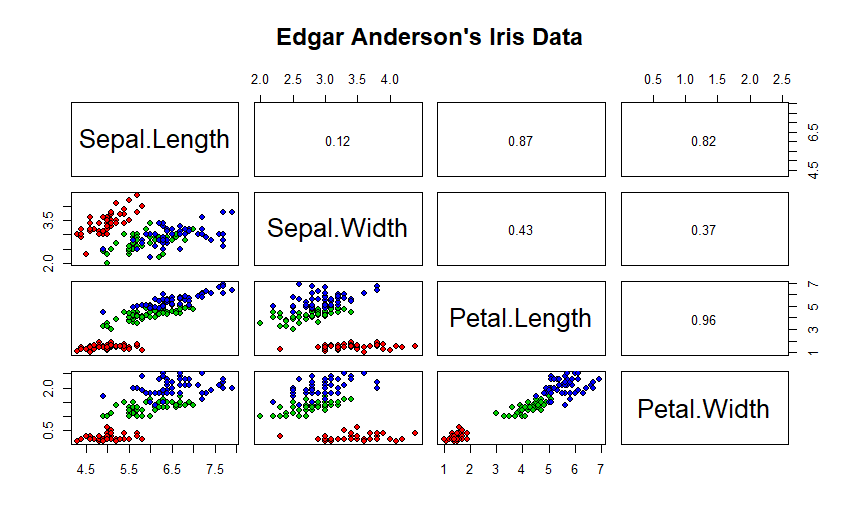


図16 データセットirisの要約

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.

* 1. 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.

グラフ

自動的に生成された説明

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.

# **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.

# **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/>