R による統計分析

以下では

1. R をインスツールする
2. データを読み込む
3. いくつかの統計量の計算
4. 偏差値の計算
5. それぞれのデータを取り出す
6. 散布図を描く

について説明する。

1. R をインスツールする

(これは家で自分のパソコンなどにインスツールする場合に参考にして下さい)。
統計ソフト R をインスツールするには
(Windows 版)) https://cran.r-project.org/ に行き、「Download R for Windows」→「base」→「Download R x.y.z for Windows」(ここで x, y, z は何か数字が入る。例えば 4.01.など)をクリックして、インスツーラーをダウンロード(どこかに保存して実行)。
(Mac 版)多少複雑。http://aoki2.si.gunma-u.ac.jp/R/begin.html を参照。

2. データを読み込む

以下では `read.table( )` 関数を使ってテキストファイル(拡張子が `.txt` のファイル)のデータの読み込み方を説明する。

2.1 データの用意

テキストファイルにデータを用意する。以下では exam.txt というファイルにある 3 列のデータを読み込む(ある講義の試験の結果)のデータ。exam.txt を開いてデータを確認すると

(exam.txt のデータ)

```
## Results of three exams for a class
mid1  mid2  final
56    46    9
28    16    13
41    31    11
81    47    18
```

となっている。1 列目は中間試験 1、2 列目は中間試験 2、3 列目は期末試験の結果である。このファイルを適当なディレクトリに保存する。保存したディレクトリを覚えておくように。

2.2 作業ディレクトリの変更

R を起動し R の画面のメニュー・バーから「ファイル」→「ディレクトリの変更」によってデータ (exam.txt) が置いてあるディレクトリを指定。確認のため

```
> dir()
```

と入力すると、現在の作業ディレクトリにあるファイルが全て表示されるので、そこに exam.txt があるか確認する。

2.3 read.table() 関数による読み込み

次のコマンドを実行する(以下を打ち込んでエンター・キーを押す)

```
> exam=read.table("exam.txt",head=TRUE,skip=1)
```

これは exam.txt にあるデータに exam という名前を付けて R に読み込みという命令を実行している。1 番目の引数は読み込むデータの拡張子込のファイル名、2 番目の因数 header = TRUE は実際に読み込むデータの最初の行に各データの名前が入っている事を R に知らせるためのものである(もし各データ系列の名前(変数名)がなく、数字のデータから始まっていれば header = FALSE とする)。3 番目の因数にはファイルのデータのうち最初の何行かを飛ばして読み込みを開始するときに使用する(skip = k 。で最初の k 行を読み込まないようにする事ができる)。

実際に読み込めたかどうかを確認するには head() 関数を使うとよい。データの最初の 5 行を読み込むには

```
> head(exam,5)
```

と打ち込んでエンター・キーを押すと

```
  mid1 mid2 final
1   56   46     9
2   28   16    13
3   41   31    11
4   81   47    18
5   72   51    18
```

のように表示される。データの読み込まれていることがわかる。

3. いくつかの統計量の計算

3.1 標本平均と標本中央値の計算

与えられたデータの標本平均を計算する。今 mid1 の列にあるデータの平均を計算するには

```
> mean(exam$mid1)
```

と入力する。同様に標本中央値を計算するには

```
> median(exam$mid1)
```

と入力する。

3.2 標本分散と全標本分散の計算

標本の大きさを n とすると、標本 $\{x_1, \ldots, x_n\}$ に対して、標本分散は

$$\text{標本分散} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

全標本分散は

$$\text{全標本分散} = \frac{1}{n}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

と定義される。ここで $\overline{x}$ は標本平均である。

標本分散は

```
> var(exam$mid1)
```

によって計算できる。上記は $n-1$ で割って計算した分散の値であるが、これを全標本分散の値に直すには、$(n-1)/n$ を掛けてあげればよい。まず標本の大きさ n を求めるには

```
> n=length(exam$mid1)
```

と入力する。`length()` は列ベクトルの長さを出力する関数である。この n を用いて

```
> ((n-1)/n)*var(exam$mid1)
```

と入力すれば全標本分散が求まる。 ここで “*” は積を表す。たとえば `2*3` は 2×3 を意味しており答えは 6 となる。

3.3 標本標準偏差と全標本標準偏差の計算

標本標準偏差は

```
> sd(exam$mid1)
```

によって計算できる。これは標本分散の平方根である。全標本標準偏差は全標本分散の平方根をとったものであるが、これはまず

```
> v=((n-1)/n)*var(exam$mid1)
```

によって全標本分散を計算し (`v` が全標本分散の値である)、その平方根を

```
> sqrt(v)
```

によって計算すれば求まる。ここで `sqrt()` は平方根を計算する R の関数である。

3.4 共分散と相関係数の計算

`mid1` と `mid2` データの共分散を計算するには

```
> cov(exam$mid1,exam$mid2)
```

と入力する。さらに mid1 と mid2 の相関係数を計算するには

```
> cor(exam$mid1,exam$mid2)
```

と入力する。

4. 偏差値の計算

偏差値は以下のように計算できる。まず標準(基準)化をする。

```
> m=mean(exam$mid1)
> s=sd(exam$mid1)
> nmid1=(exam$mid1-m)/s
```

nmid1 がそれぞれの点数が標準化されたものである。“ - ” は引き算、“ / ” は除算を意味している。m はスカラーだが R において、上記のようにベクトルからスカラーを引くと、ベクトルの各要素からそのスカラーの値を引いたものが計算される。以下に見るように足し算についても同様の事が成立する。

nmid1 を使って偏差値を計算するには

```
> smid1=10*nmid1+50
```

と入力する。偏差値の標本平均と標本分散を計算するには

```
> mean(smid1)
> sd(smid1)
```

と入力すればよい。50 と 10 になっているはずである。

5. それぞれのデータを取り出す

上記に分析において、まず exam というおおもとのデータを作り、その各列にある mid1 というデータを使用する際には、exam にあるデータであるという事を R にわからせるために exam$mid1 と打ち込まなければならなかった。しかしながら毎度毎度 exam$ を打ち込むのは面倒であるので、mid1 だけを取り出しておきたい。そこで新しく

```
> mid1=exam$mid1
```

というデータを作る。こうしておくと便利である。たとえば標本平均を計算するにでも、以前は mean(exam$mid1)と入力しなければならなかったが、今後は

```
> mean(mid1)
```

でまったく同じ結果を得ることができる。

練習問題

1. mid2、final データについて標本平均、標本標準偏差、を計算せよ

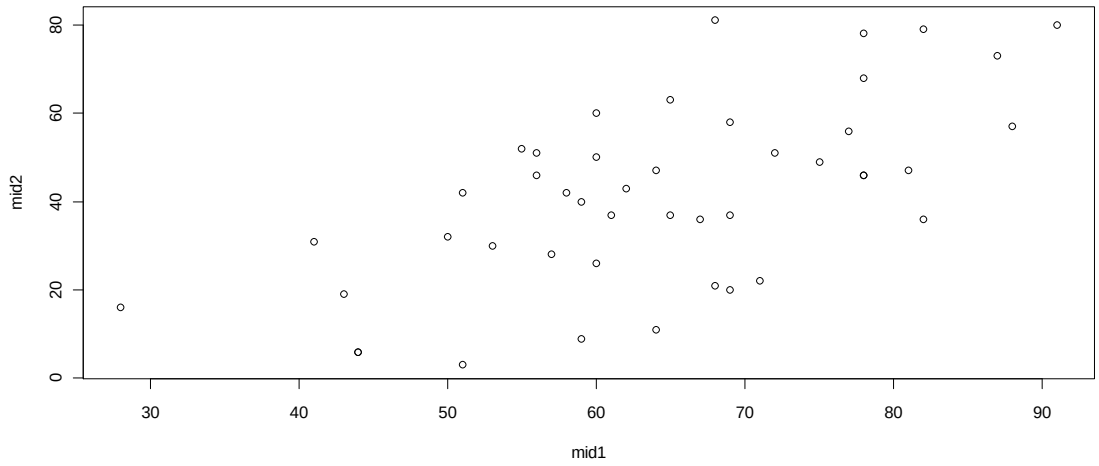
2. mid1 データの 56 の偏差値と mid2 データの 47 の偏差値を計算せよ。
3. mid1 と final データの相関、mid2 と final データの相関を計算せよ。

6. 散布図を描く

exam データにある mid1 と mid2 の散布図を描いてみよう。

```
> mid1=exam$mid1
> mid2=exam$mid2
> plot(mid1,mid2)
```

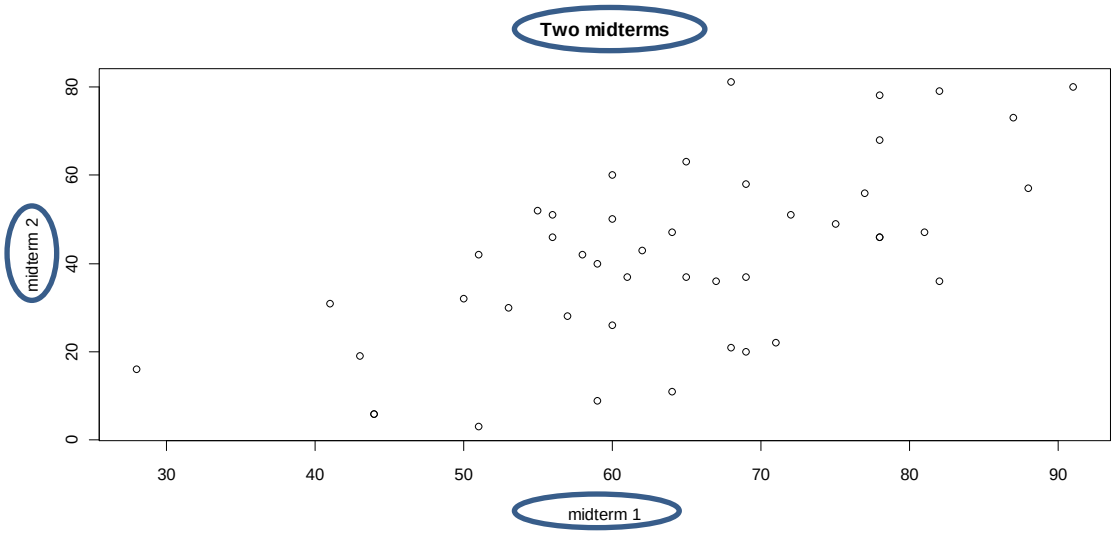
と入力すると mid1 を X 軸、mid2 を Y 軸とする以下のような散布図が描かれる。



6.1. 図にタイトル、X軸、Y軸にラベルをつける

以下のコマンドで、上記の散布図に図のタイトルと X 軸と Y 軸の名前を追加できる。

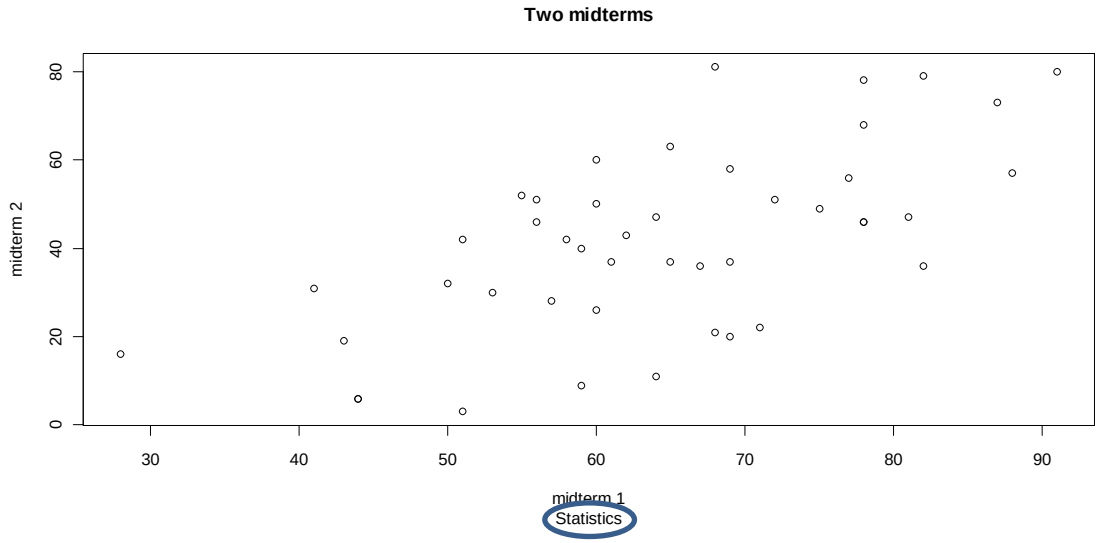
```
> plot(mid1,mid2,main="Two midterms",xlab="midterm 1",ylab="midterm 2")
```



また X 軸の下にサブタイトルを挿入するには

```
> plot(mid1,mid2,main="Two midterms",xlab="midterm 1",ylab="midterm 2",
  sub="Statistics")
```

と入力する。すると



というように X 軸の下にサブタイトルが挿入される。

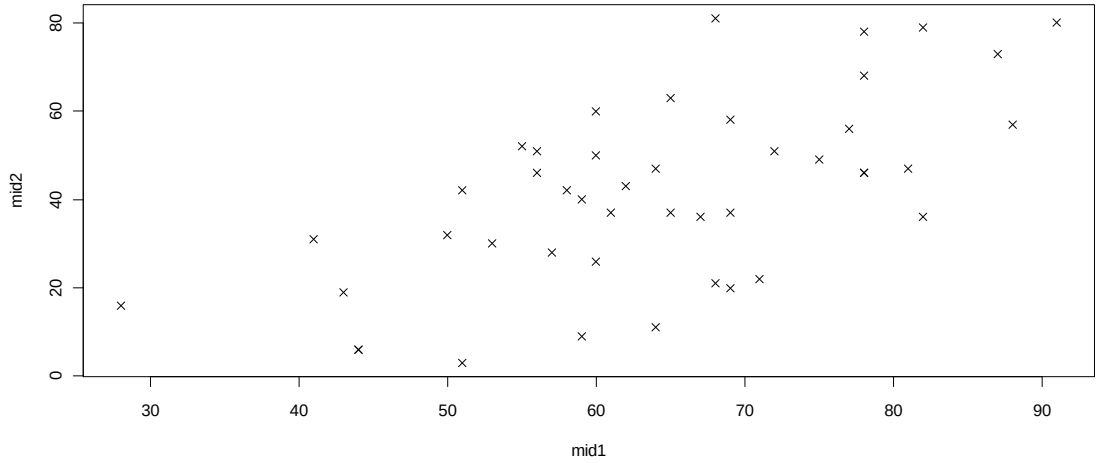
6.2. 図のマーカーの種類を変える

(以下では簡単化のためにタイトルや軸ラベルはつけていないが、`plot()` の中に上記の `main` や `xlab` を指定すればもちろんつけることができる)

図のマーカーの種類を変えるには pch の値を変える。例えば

```
> plot(mid1,mid2,pch=4)
```

と入力すると



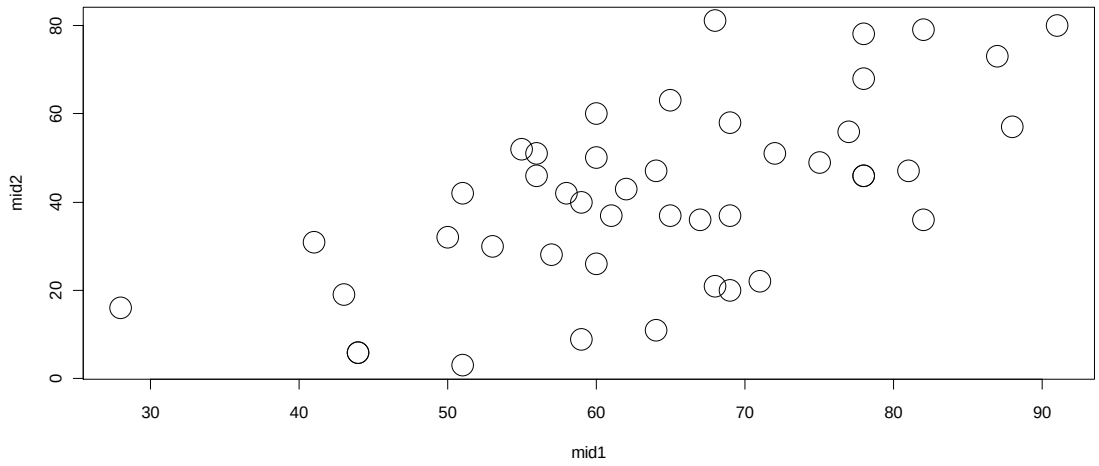
のような散布図を得る。上記の例では “○” から “×” に マーカーが変わっている。pch は1から

25までの数字をとり、それぞれの数字に対応してマーカーが変わる。主なものには 1: ○, 2:△, 3:+, 4:×, 22:□ などがある。

マーカーを大きくするには cex を用いる。たとえばマーカーの大きさを 3 倍にしたい場合は

```
> plot(mid1,mid2,cex=3)
```

と入力すれば



のような図が表示される。またマーカーの色を変更することもできる。これには col を用いる。たとえばマーカーの色を赤色にしたい場合には

```
> plot(mid1,mid2,col="red")
```

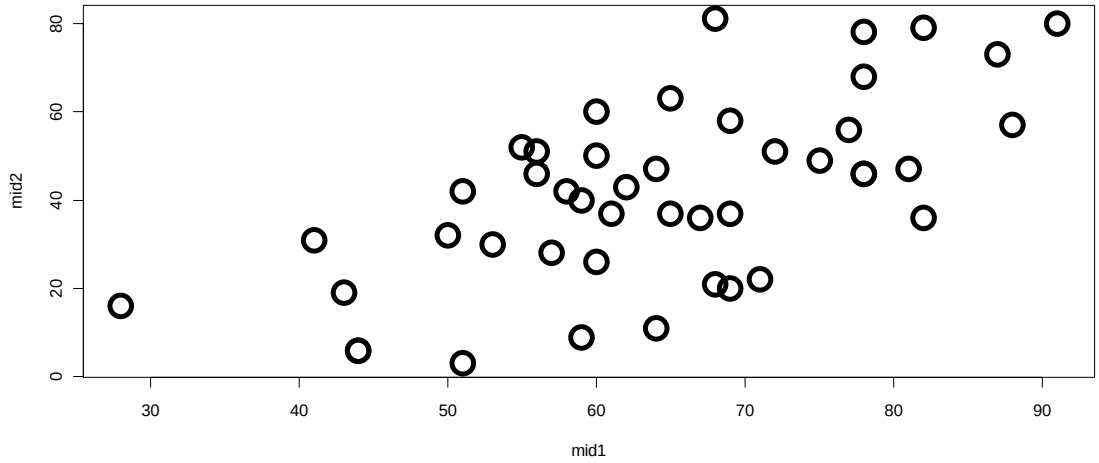
とする。主なものに 青: blue, 緑: green, 黄色: yellow などがある。どのような色が使用可能かどうかは

```
> colors()
```

またマーカーの線の太さを変えたいときには lwd を用いる。たとえば

```
> plot(mid1,mid2,lwd=5)
```

と入力すれば



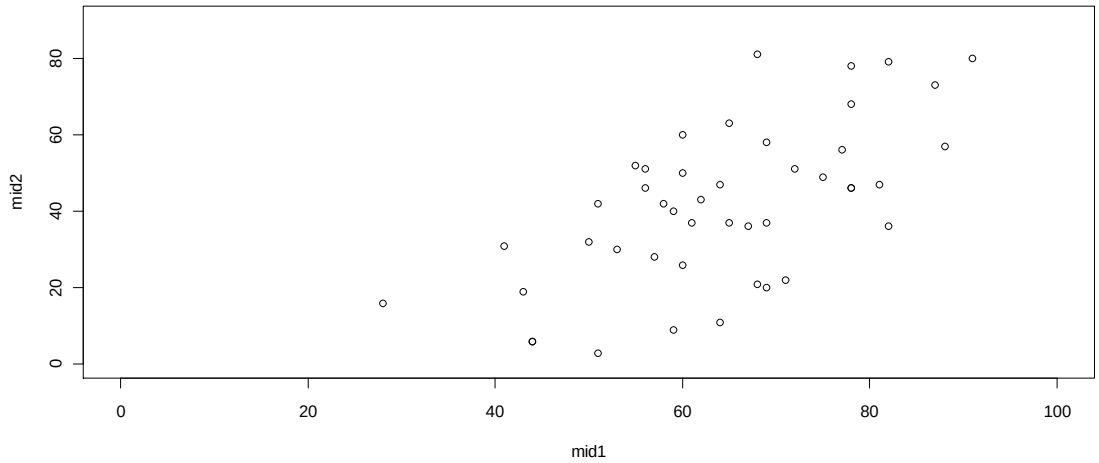
のようにマーカーの線の太さを変更することができる。太さに応じて数字を大きくする。

6.3 図の軸の範囲を変える

通常 `plot()` 関数によって散布図を描くと、X 軸と Y 軸の範囲は R が自動的に設定する。これを変えるには `xlim` および `ylim` を用いる。例えば、X 軸の範囲を 0 から 100、Y 軸の範囲を 0 から 90 にするには

```
> plot(mid1,mid2,xlim=c(0,100),ylim=c(0,90))
```

と入力すると以下の図を得る。



練習問題

1. {mid2, final} と {mid1, final} の散布図を□ というマーカーで描いてみる.
2. (1)で描いた散布図のマーカーの大きさ、線の太さと色をそれぞれ 3,5,緑に変えてみよう。
3. (1)、(2)で描いた散布図に適当なタイトルと軸ラベルを挿入してみよう。