4　1 変量データの分析

まず，データ ”height-weight” を呼び込む.

```
> x<-read.table("height-weight")
> x
    V1 V2
1  148 41
2  160 49
3  159 45
4  153 43
5  151 42
6  140 29
7  156 49
8  137 31
9  149 47
10 169 47
11 151 42
12 157 39
13 157 48
14 144 36
>
```

$V1$ が身長で $V2$ が体重なので，level をそのように変更する.

```
> colnames(x)<-c("height","weight")
> x
   height weight
1     148     41
2     160     49
3     159     45
4     153     43
5     151     42
6     140     29
7     156     49
8     137     31
9     149     47
10    169     47
11    151     42
12    157     39
13    157     48
14    144     36
>
```

つぎのようなデータの要約が可能である.

- 各変量のヒストグラムを作成する．hist(データ名)．また，幹葉表示をするには．”stem(データ名）” である.

```
> hist(x$height)
> hist(x$weight)
> # ヒストグラムのコマンドを調べる.
> x1<-x$V1
> x1
 [1] 148 160 159 153 151 140 156 137 149 169 151 157 157 144
> hist(x1)
> hist(x1,freq=F) # 相対頻度で書いたもの．ただし，横軸の値と区間をかけると相対度数.
>x1<-x$height
> # さらに，込み入った書き方.
```

```
> hist(x1,freq=F,xlim=c(130,175),ylim=c(0,0.1),xlab="",ylab="",main="")
>
> # 幹葉表示
> stem(x1)

  The decimal point is 1 digit(s) to the right of the |

  13 | 7
  14 | 0489
  15 | 1136779
  16 | 09

>
```

- 代表値 — 平均値・中央値

```
> mean(x$height)
[1] 152.2143
> x$height
 [1] 148 160 159 153 151 140 156 137 149 169 151 157 157 144
> rank(x$height)
 [1]  4.0 13.0 12.0  8.0  6.5  2.0  9.0  1.0  5.0 14.0  6.5 10.5 10.5  3.0
> sort(x$height)
 [1] 137 140 144 148 149 151 151 153 156 157 157 159 160 169
> median(x$height)
[1] 152
```

- ばらつきの尺度 — 分散・標準偏差

```
> var(x1)
[1] 71.41209
> sd(x1)
[1] 8.450567
> sqrt(var(x1))
[1] 8.450567
>
```

- 箱ヒゲ図 (Box plot) (1) 箱の中にはデータの 50% が含まれる．(2) 上のヒゲと下のヒゲは最大値と最小値.

```
> x1
 [1] 148 160 159 153 151 140 156 137 149 169 151 157 157 144
> boxplot(x1)
```

- 経験分布関数 $x_1, x_2, \ldots, x_n$ をデータとし，$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ を順序統計量とする．このとき,

$$F_n(x) = \begin{cases} 0 & (x < x_{(1)}, \\ \frac{k}{n} & (x_{(k)} < x \leq x_{(k+1)}, \\ 1 & x_{(n)} < x \end{cases}$$

を経験分布関数という．すなわち，確率変数 X の確率関数が

$$P(X = x) = \begin{cases} \frac{1}{n} & (x = x_i), \\ 0 & \text{その他} \end{cases}$$

で与えられたときの X の分布関数である.

```
> library(stepfun)  # Liberary の stepfun を呼び込む
> n50<-rnorm(50)    # 50 個の標準正規正規乱数を発生させる.
> ed<-ecdf(n50)     # その経験分布関数を計算させる.
```

```
> plot(ed, verticals=TRUE,do.p=FALSE)    # 経験分布関数のグラフを作図
> ed<-ecdf(x1)
> plot(ed, verticals=TRUE,do.p=FALSE)
> x1
 [1] 148 160 159 153 151 140 156 137 149 169 151 157 157 144
>  ed<-ecdf(x1)
> plot(ed, verticals=TRUE,do.p=FALSE)
```

演習 4　$x_1, x_2, \ldots, x_n$ としたとき，このデータの分散は

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2, \qquad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

である．しかし、これは

$$\frac{1}{n-1}\left(\sum_{i=1} x_i^2 - n(\bar{x}_n)^2\right)$$

と同じである．これらの式の値と ”var ” を用いた分散の値が一致することを身長と体重のデータについて確かめよ．ヒント：$n = 14$.

演習 5　データ $x_1, x_2, \ldots, x_n$ としたとき，の平均偏差は

$$\frac{1}{n}\sum_{i=1}|x_i - \bar{x}_n|$$

で定義される．身長と体重について平均偏差を求めよ．