

確率統計と情報処理・演習（2010 年度後期）

1 変量データ分析

2010 年 10 月 01 日

日本女子大学理学部数物科学科 今野 良彦

September 20, 2010

今日の講義の目的と概要

目的

- （１）データが入力されたら，とりあえずデータにどんな値がどれくらいあるかを調べる．この様子をあらわしたものを「分布」という．表現法として，度数分布表や図（ヒストグラム）がある．
- （２）分布を説明するために，分布の中央あたりの値を「分布の中心」として「分布の代表値」とする．
- （３）つぎに，データのばらつきを表現する尺度を議論する．

今日の講義の目的と概要

目的

（１）データが入力されたら，とりあえずデータにどんな値がどれくらいいるかを調べる．この様子をあらわしたものを「分布」という．表現法として，度数分布表や図（ヒストグラム）がある．

（２）分布を説明するために，分布の中央あたりの値を「分布の中心」として「分布の代表値」とする．—— 平均値と中央値．さらに，最小値と最大値，ボックスプロット（箱ひげ図）

（３）つぎに，データのばらつきを表現する尺度を議論する．—— 分散と標準偏差，平均偏差，範囲，四分位範囲^{しぶんい}

ヒストグラム

- 単峰型
 - － 峰を中心に左右対称のもの
 - － 左に偏ったもの（下段左） — 山の裾の部分に注目している！
 - － 右に偏ったもの（下段右）
- 双峰型・多峰型 — 異種のデータが混在している場合が多い.

ヒストグラムの作成

ヒストグラムの作成

```
> a<-rnorm(100,10,10)
> hist(a)
> a<-round(rnorm(50,10,20))
> a
 [1]  -1  39  -4 -13  -2  40 -17   1  20  14
[11]   3  27   6  47  47 -23   6  -8  -5  16
[21] -17  -5  15   6   0   0  33 -16  23 -39
[31]  29  22   7 -10  26  16 -10  13  -1  -1
[41]  42 -23  -4   9  18 -34  14  13  -4   0
> hist(a)
```

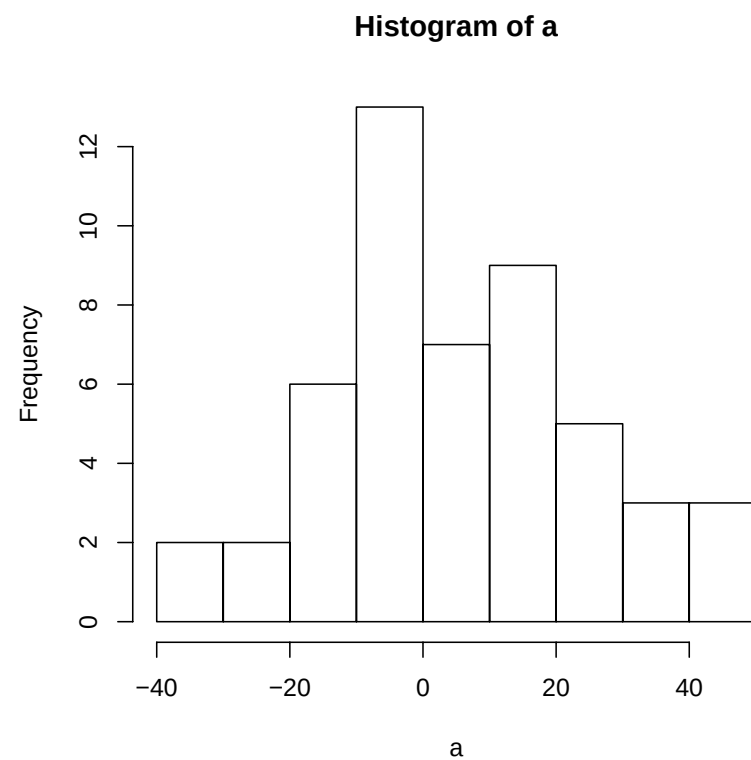


Figure 1: 作成されたヒストグラム

ヒストグラムの作成

```
>  
> # コマンド hist のオプションを調べる  
> ?hist  
> # 割合で表示  
> hist(a,freq=F)  
>
```

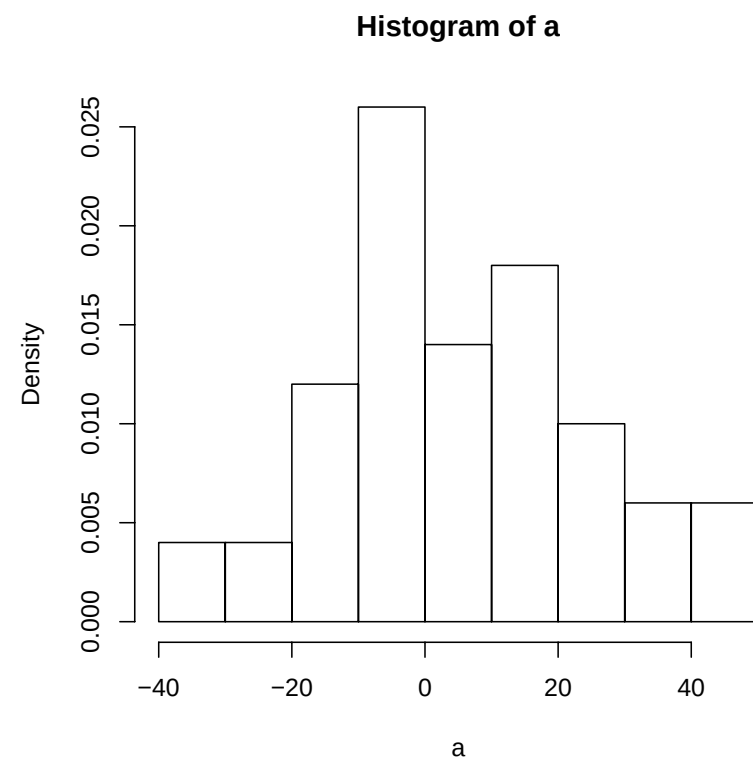


Figure 2: 作成されたヒストグラム

問題 1

- 以下のようにデータを発生させ、そのヒストグラムを作成し、そこからわかることを述べよ.

ヒストグラムの作成

```
>  
>  
> a1<-round(rnorm(500,0,1))*5+あなたの誕生日  
>  
> a2<-round(rt(50,10)*5)+あなたの誕生日  
>
```

オブジェクト a1 と a2 のヒストグラムを作成し、それを pdf file で保存する．ファイル名は 学籍番号-a1.pdf と 学籍番号-a2.pdf とせよ．さらに、ファイル 学籍番号-名前-締め切り.txt

（例：20916***-目白花子-101008.txt）

締め切り：2010 年 10 月 08 日（金）13 時

代表値 — 平均値と中央値

n 個のデータの値を

$$x_1, x_2, \dots, x_n$$

とする.

平均値

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

中央値

$x_1, x_2, \dots, x_n$ を昇順に並び替えたものを $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ としたとき,

$$Me = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ が奇数} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & n \text{ が偶数} \end{cases}$$

たとえば, $x_1 = 20, x_2 = 10, x_3 = 0$ のとき,

$$x_{(1)} = 0, x_{(2)} = 10, x_{(3)} = 20$$

となる. したがって, $n = 3$ は奇数だから

$$Me = x_{(2)} = 10$$

また, たとえば, $x_1 = 20, x_2 = 10, x_3 = 0, x_4 = 30$ のとき,

$$x_{(1)} = 0, x_{(2)} = 10, x_{(3)} = 20, x_{(4)} = 30$$

となる. したがって, $n = 4$ は偶数だから

$$Me = \frac{x_{(2)} + x_{(3)}}{2} = \frac{10 + 20}{2} = 15$$

平均値の意味

$g(a) := \sum_{i=1}^n (x_i - a)^2$ とおく．このとき， $g(a)$ を最小にする a の値は何か？

$$\begin{aligned} g(a) &= \sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - a)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x}_n)^2 + \sum_{i=1}^n (\bar{x}_n - a)^2 - 2(a - \bar{x}_n) \sum_{i=1}^n (x_i - \bar{x}_n) \\ &= \sum_{i=1}^n (x_i - \bar{x}_n)^2 + \sum_{i=1}^n (\bar{x}_n - a)^2 \geq \sum_{i=1}^n (x_i - \bar{x}_n)^2 \end{aligned}$$

よって， $a = \bar{x}_n$ で最小．

絶対偏差

では,

$$f(a) := \sum_{i=1}^n |x_i - a|$$

とおく. このとき, $f(a)$ を最小にする a の値は統計的意味をもつか?

すわなち, 分布の位置を表す代表値として用いることができるか?

答えは Yes! しかし, 明示的な表現はできない!

平均値と中央値の違い

- ★ 平均値と中央値が近い場合 → 概ねその値を中心として左右対称であることが多い.
- ★ 平均値と中央値が離れている場合 → 対称性が崩れて右ないし左に歪んでいるか、離れた「**外れ値**」があることがおおい.

注意：ヒストグラムなどで確認できること！

垂水共之・飯塚誠也 著「R/S-PLUSによる統計解析入門」（共立出版，2006年4月25日のサポートのページ

<http://www.mikawayu.to/appstat/>

の `boxplot_interactive.r` を使い説明せよ！

最大値と最小値

n 個のデータの値を

$$x_1, x_2, \dots, x_n$$

としたとき,

$x_{(1)}$: 最小値, $x_{(n)}$: 最大値

—— 最大値と最小値 ——

```
> height
[1] 166 154 165 169 155 158 168 154 162 153
> min(height)
[1] 153
> max(height)
[1] 169
>
```

箱ひげ図（ボックスプロット）

もとのデータ $x_1, x_2, \dots, x_n$ を並び替えたものを

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

とする.

- 第1 ^{しぶんい}四分位点 — データの下から $n/4$ 個の値
- 中央値（第2四分位点） — データの下から $n/2$ 個の値
- 第3四分位点 — データの下から $3n/4$ 個の値

summary と boxplot.stats

```
> x<-1:12
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   3.75   6.50   6.50   9.25   12.00
>
#
> boxplot.stats(x)
$stats
[1] 1.0 3.5 6.5 9.5 12.0
$n
[1] 12
$conf
[1] 3.76336 9.23664
$out
numeric(0)
# stats のみ出力
> boxplot.stats(x)$stats
[1] 1.0 3.5 6.5 9.5 12.0
> boxplot.stats$conf
NULL
> boxplot.stats(x)$conf
[1] 3.76336 9.23664
```

boxplot：箱ひげ図の出力

```
> x<-1:12  
> boxplot(x)  
>
```

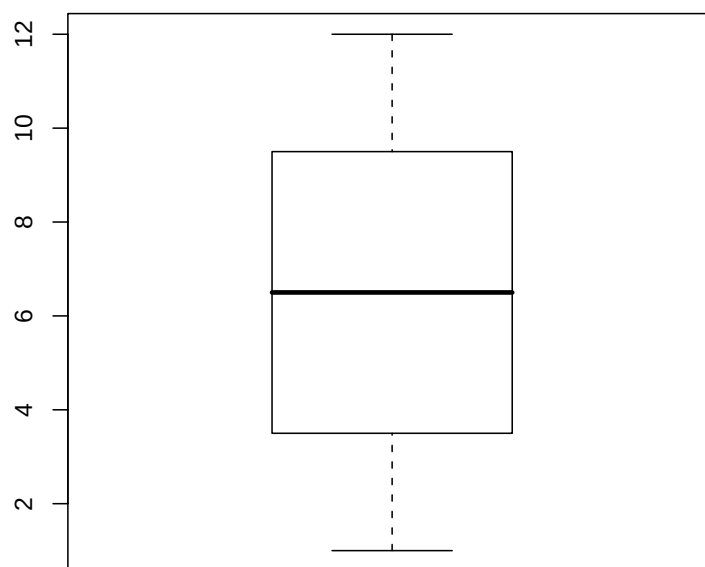


Figure 3: 箱ひげ図

ばらつきの尺度：分散と標準偏差

2つのデータの比較

```
> height
[1] 148 160 159 153 151 140 156 137 149 160 151 157 157 144
> height2
[1] 138 162 158 151 145 134 160 137 151 163 152 163 158 147
> mean(height)
[1] 151.5714
> mean(height2)
[1] 151.3571
> median(height)
[1] 152
> median(height2)
[1] 151.5
> boxplot(height,height2,names=c("height","height2"))
```

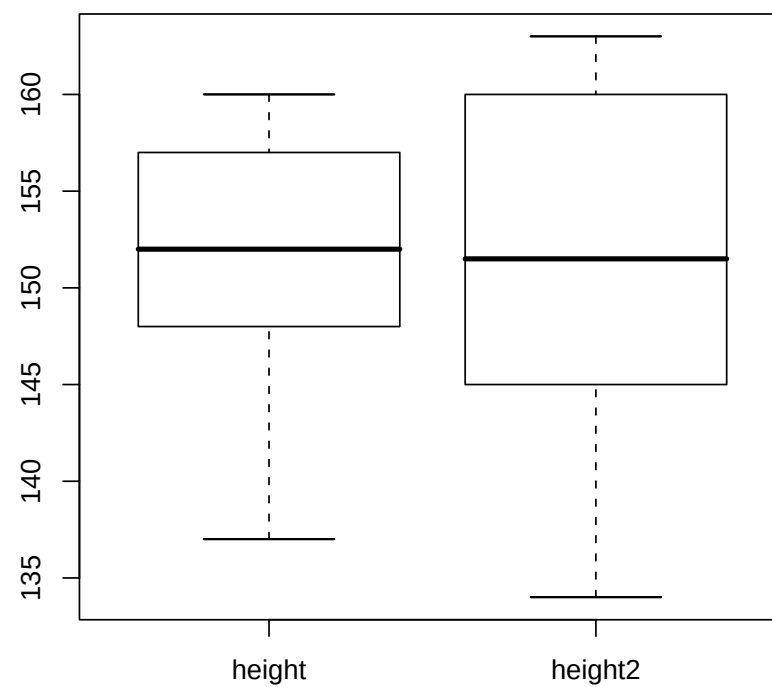


Figure 4: 2つの箱ひげ図

代表値（分布の位置）は似ているが、値のばらつき具合が違う！

- 範囲 — `diff(range(x))`

$$x_{(n)} - x_{(1)}$$

- 四分位範囲^{しぶんい} — `diff(quantile(x,c(0.25,0.75),names=F))`

$$x_{(3n/4)} - x_{(n/4)}$$

- 平均偏差 — `sum(abs(x-mean(x)))/length(x)`

$$(1/n) \sum_{i=1}^n |x_i - \bar{x}_n|$$

- 分散 — `var(x)` または

`sum((x-mean(x))^2)/(length(x)-1)`

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

- 標準偏差 — `sd(x)` または

`sqrt(sum((x-mean(x))^2)/(length(x)-1))`

これらの値は大きいほどばらつく！

問題 2

co2 と入力すると Mauna Loa Atmospheric での 1959 年から 1997 年までの毎月の CO₂ の濃度のデータが利用できる.

```
> cc<-seq(1,length(co2),by=12)
> cc
[1] 1 13 25 37 49 61 73 85 97 109 121 133 145 157 169 181 193 205 217
[20] 229 241 253 265 277 289 301 313 325 337 349 361 373 385 397 409 421 433 445
[39] 457
> # オブジェクト y にあたる誕生月のデータを入力
> apr<-co2[cc+(あなたの誕生月マイナス1を入力)]
>
> apr<-co2[cc+3]
> apr
[1] 317.56 318.87 319.31 320.42 321.22 321.40 321.97 323.54 324.25 324.86 326.50
[12] 327.97 327.62 329.56 331.33 332.48 333.14 334.41 335.90 337.59 338.71 340.60
[23] 342.33 343.39 344.77 346.88 348.17 349.37 350.80 353.41 355.26 356.04 358.48
[34] 359.07 359.41 361.25 363.48 364.76 366.40
>
```


- (1) 誕生月のデータの平均と中央値を求めよ.
- (2) 箱ひげ図を作成せよ.
- (3) 範囲, 四分位範囲, 平均偏差, 分散, 標準偏差を求めよ.

テキストファイルに実行文と結果を貼り付け, 箱ひげ図は pdf file (学籍番号-boxplot.pdf) で添付すること.

テキストファイル: 20916***-目白花子-101008.txt

締め切り: 2010 年 10 月 08 日 (金) 13 時