確率統計と情報処理・演習（2007 年度後期）

2007 年 09 月 28 日

日本女子大学理学部数物科学科　　今野 良彦

September 25, 2007

この講義の目的と概要

目的

統計学とは，数量的なデータからそのデータの由来する現象に関する情報を科学的に取り出す方法とその理論体系である．また，近年，コンピュータの急速な普及により，あらゆる分野で統計的データ解析の重要性が認識されている．データ解析手法を使いこなすためには，統計学の背景にある科学的手法としての理論を理解することが重要である．このような観点からいくつかののテーマについて学び，演習時にコンピュータによる実習も行う.「確率統計と情報処理演習」をとること.

コンピュータとの関連

この講義では，コンピュータ環境 R を用いて実習を行い，統計学の背景となる理論の理解を深める.

教科書

垂水共之・飯塚誠也：

R/S-Plus による統計解析入門，共立出版（ISBN 4-320-01807-9）

この講義の計画

09 月 28 日	(1) ガイダンス　(2) R の起動と終了　(3) データの入力，修正，保存，読み込み
10 月 05 日	(1) 1 次元データの分布　(2) 代表値 — 平均値と中央値・最小値と最大値　(3) 箱ひげ図　(4) ばらつきの尺度 — 範囲・四分位範囲・平均偏差・分散・標準偏差
10 月 12 日	(1) ファイルからのデータ入力　(2) 1 次元データの分布　(3) 回帰直線とピアソンの相関係数　(4) ピアソンの相関係数の性質　(5) 順位相関係数　(6) 多変量データのグラフ表現
10 月 26 日	(1) 確率と確率分布　(2) 関数のグラフ　(3) 正規分布と一様分布
11 月 02 日	(1) χ^2 分布　(2) t 分布　(3) F 分布　(4) 2 変量正規分布　(5) 標本相関係数の分布
11 月 09 日	(1) 中心極限定理　(2) 一様分布の場合　(3) 種々の分布の場合
11 月 16 日	(1) 母集団と標本　(2) 点推定 — 不偏性・一致性・有効性・最尤法　(3) 正規分布の母平均の点推定　(4) 正規分布の母分散の点推定
11 月 30 日	(1) 信頼度と区間推定　(2) 正規分布の母平均の区間推定　(3) 正規分布の母分散の区間推定
12 月 07 日	(1) 検定の考え方　(2) 正規分布の母平均の検定　(3) 正規分布の母分散の検定
12 月 14 日	(1) 検出力について (2) 復習
12 月 21 日	予備日
01 月 11 日	試験

宿題の提出について

- 提出はメール：mtoukei(at)mp[dot]jwu[dot]ac[dot]jp
- メールの件名（subject）には，ローマ字の名前と苗字と学籍番号，締め切り日を書くこと．例：

  Mejiro Hanako 201600000 due 2007-10-05

- 解答はテキストファイルに記述し，それをメールに添付すること．ファイル名を

  ローマ字の名前-締め切り.txt　（例：mejiro-hanako-071005.txt)

  とすること．グラフは pdf file で添付し，それぞれの pdf file にはどのようなグラフがあるかをテキストファイルに解説すること．

- Word file では提出しないこと！

R について

- S 言語・環境を見本に開発されたフリーソフト（無償で使え，Windows, Mac, Linux）

  http://www.r-project.org/

- Graphics 機能の充実（画面または印刷物への出力）
- データを効率的に操作し、保管する機能
- データの配列機能と列の計算のための演算子のセット
- たくさんのデータ解析の手法
- 条件分岐，ループ，ユーザー定義の再帰的関数（プログラミングができる．C 言語に似ている）

R の基本操作 — データの入力と修正

起動するには

デスクトップの R のアイコンをクリック

データの入力

番号	身長	番号	身長
1	148	9	137
2	160	9	149
3	159	10	160
4	153	11	151
5	151	12	157
6	140	13	157
7	156	14	144

データの入力

```
> height<-c(140,150)
> height
[1] 140 150
>
> height<-c(
+ 140,
+ 150
+ )
> height
[1] 140 150
>
```

オブジェクトの検索

```
> weight<-c(40,50,54)
> weight
[1] 40 50 54
> weiht
エラー：オブジェクト "weiht" は存在しません
> ls()
 [1] "weight"
>
```

R の基本操作 — データの追加・削除・修正

データの追加・削除・修正

```
> h<-c(10,20,30,40,50)
> h
[1] 10 20 30 40 50
> newh<-append(h,c(60,70))
> newh
[1] 10 20 30 40 50 60 70
> h2<-c(newh[2:4],100,110)
> h2
[1]  20  30  40 100 110
```

データの保存

メニューの「ファイル」から「ディレクトーの変更」を選び、「Browse」からプロッピーディスクのあるディレクトリーを選択する。

```
> sink("result.txt")
> # result.tex という名前のファイルに以下の結果を書き込む
> height
> mean(height)
> max(height)
> sink()      # 書き込みを中止する。
```

result.tex の内容

メニューの「ファイル」から「ディレクトーの変更」を選び、「Browse」からプロッピーディスクのあるディレクトリーを選択する。

```
[1] 140 150
[1] 145
[1] 150
```

データと関数の保存

メニューの「ファイル」から「ディレクトーの変更」を選び、「Browse」から MO のあるディレクトリーを選択する．起動から終了までに作成したデータや関数が保存できる．

```
> cc<-c(10,20,39)
> save.image("intro.RData")
>
```

メニューの「ファイル」から「ディレクトーの変更」を選び、「Browse」からプロッピーディスクのあるディレクトリーを選択する。

[以前にセーブされたワークスペースを復帰します]

```
> load("intro.RData")
> cc
[1] 10 20 39
>
```

演習

以下の実行結果を提出せよ.

- オブジェクト名をつけて、そのファイルに身長のデータを入力せよ.
- 最後のデータを自分の身長のデータで置き換えよ.
- 締切り 2007 年 10 月 05 日（金）13:00

確率統計と情報処理・演習（2007 年度後期）

1 変量データ分析

2007 年 10 月 05 日

日本女子大学理学部数物科学科　　今野 良彦

September 25, 2007

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かるかを調べる．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かるかを調べる．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，四分位偏差

ヒストグラム

- 単峰型
  - 峰を中心に左右対称のもの
  - 左に偏ってもの（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.

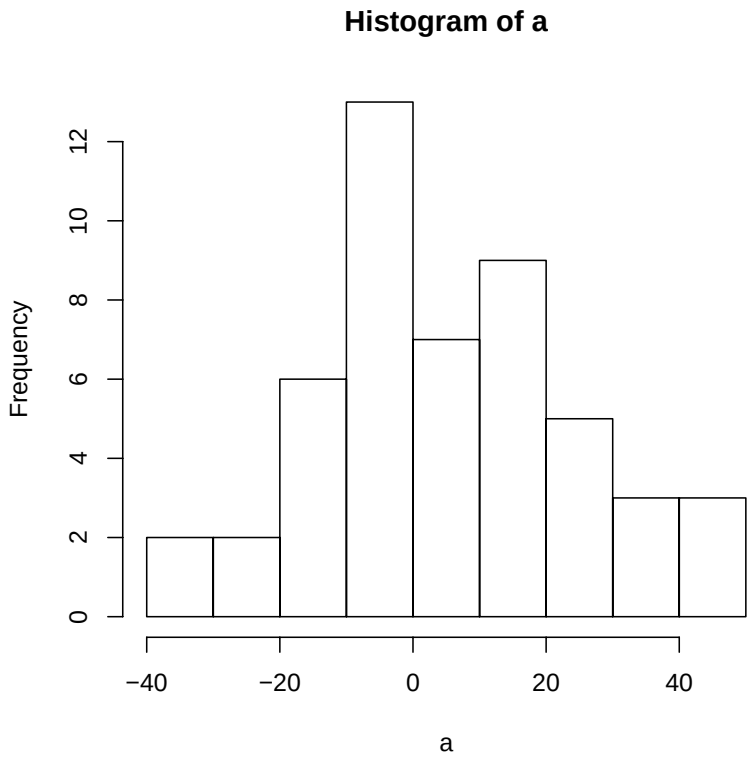

Figure 1: 作成されたヒストグラム

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)
>
```

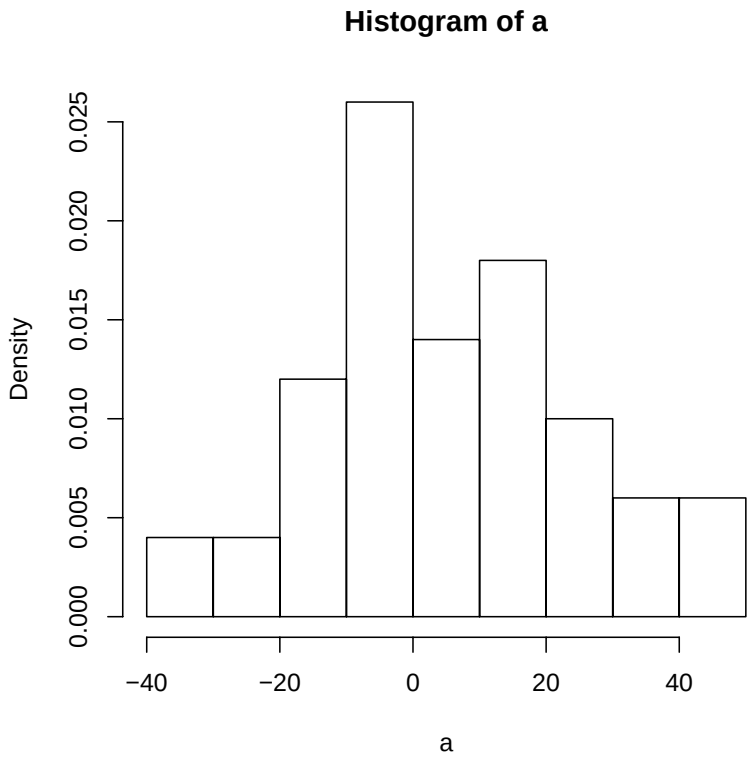

Figure 2: 作成されたヒストグラム

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+あたたの誕生日
>
> a2<-round(rt(50,10)*5)+あたたの誕生日
>
```

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 学籍番号下3桁-a1.pdf と 学籍番号下3桁-a2.pdf とせよ．さらに，ファイル ローマ字の名前-締め切り.txt

（例：mejiro-hanako-071012.txt)

代表値 — 平均値と中央値

n 個のデータの値を

$$x_1, x_2, \cdots, x_n$$

とする.

平均値

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

中央値

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

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