

Rを使ってみませんか

中西渉

watayan@watayan.net
名古屋高等学校

2010年7月18日

agenda

- 1 統計は必要
- 2 Rとは
- 3 R入門
 - 起動・終了
 - 基本
 - 例題
- 4 参考
 - 書籍
 - サイト
 - 勉強会

統計は必要

現教育課程 数学

科目	単元	キーワード	実態は...
数学 A	場合の数と確率	期待値	
数学 B	統計とコンピュータ	標準偏差 相関係数	
数学 C	確率分布 統計処理	標準偏差 正規分布 母平均の推定	

情報 C

表計算ソフトウェアなどの簡単な統計分析機能...

統計は必要

現教育課程 数学

科目	単元	キーワード	実態は...
数学 A	場合の数と確率	期待値	
数学 B	統計とコンピュータ	標準偏差 相関係数	履修しない 履修しない
数学 C	確率分布 統計処理	標準偏差 正規分布 母平均の推定	履修しない 履修しない 履修しない

情報 C

表計算ソフトウェアなどの簡単な統計分析機能...

「機能」だけ? その値の「意味」は?

数学で扱われる「統計」

- \sum の計算演習の延長
- 値の意味を学ばない

数学で扱われる「統計」

- \sum の計算演習の延長
 - 値の意味を学ばない
- 統計を知らない大人の出来上がり
- 平均だけを意識
 - デタラメなグラフ
 - 教員が案外統計を知らない

新教育課程

- 小学校からすべての学年で統計
- 高校... 数学Iで「データの分析」
分散，標準偏差，四分位数，箱ひげ図，散布図，相関係数，...

新教育課程

- 小学校からすべての学年で統計
- 高校... 数学Iで「データの分析」 ← 必履修!!
分散, 標準偏差, 四分位数, 箱ひげ図, 散布図, 相関係数, ...

- 統計解析ソフトウェア
- GPL で配布
- UNIX, Windows, Mac OS 上で稼働

どうして R?

- Excel では話にならない!!
 - Excel の統計関数のひどさには定評
青木繁伸氏 (群馬大) のまとめ
<http://aoki2.si.gunma-u.ac.jp/Hanasi/excel/>
 - 分析ツール... Excel 2008 にはないぞ!
 - どうやって「箱ひげ図」や「ヒストグラム」を描くの?
- SPSS では話にならない!!
 - 高すぎ!!

起動

- メニューから「システム」→「Konsole - ターミナルプログラム」
- R [Enter]

終了

- q() [Enter]
- Save workspace image? [y/n/c]:
と聞かれたら y か n で答える
y なら次に起動したときに続きから作業できる

- 四則計算などはそのまま
- `sqrt` や `abs` なんかも使える

入力してみよう

`3+2`

`3-2`

`3*2`

`3/2`

`3^2`

`sqrt(5)`

変数

- 代入は `<-`
- 変数は基本的にはベクトル
- `c(値 1, 値 2, ...)` でベクトルを作成

入力してみよう

```
x <- c(1,2,3,4,5)
```

```
y <- c(4,3,5,2,6)
```

```
x+y
```

```
x-y
```

```
x*y
```

```
2*x
```

```
sqrt(x)
```

```
x>3
```

統計関数

sum 和

length データの個数

mean 平均

median 中央値

max, min 最大, 最小

var 分散 ($n - 1$ で割る方)

sd 標準偏差 ($n - 1$ で割る方)

table 度数分布

summary 基本統計量

fivenum 5 値要約

cov 共分散 ($n - 1$ で割る方)

cor 相関係数

入力してみよう

```
x <- c(50,52,46,42,43,35,48,47,50,37)
y <- c(49,48,50,44,42,36,49,41,50,36)
mean(x)
median(x)
sd(x)
summary(x)
cor(x,y)
```

わからない関数は help(関数名)

テストの集計

- 数個のデータじゃつまらない
- でも入力めんどくさい
- CSV ファイルから読み込めばいい

入力してみよう

```
test <- read.csv("test.csv")  
summary(test)
```

- test.csv の 1 行名は項目名
- test はデータフレームというデータ形式
- ... 項目名つきの表だと思ってもらえばいい

テストの得点（続き）

- test の中身は以下のとおり

class	num	q1	q2	q3	q4	q5	point
A	1	16	6	16	18	14	70
A	2	14	10	19	16	16	75

...

- 各列は test\$項目名 で呼びだせる

入力してみよう

```
mean(test$point)
median(test$point)
stem(test$point)
table(test$point)
table(cut(test$point, seq(0,100,5)))
cor(test$q1, test$point)
by(test$point, test$class, mean)
```

テストの得点（続き）

グラフも簡単にかける

入力してみよう

```
hist(test$point)
hist(test$point, breaks=seq(0,100,5))
plot(test$q1, test$point)
boxplot(test$point)
boxplot(test$point ~ test$class)
```

投球データ

ある投手のあるシーズンの投球データが `pitcher.csv` に入っている

入力してみよう

```
p <- read.csv("pitcher.csv")
summary(p)
mean(p$Speed)
boxplot(p$Speed)
```

投球データ

ある投手のあるシーズンの投球データが `pitcher.csv` に入っている

入力してみよう

```
p <- read.csv("pitcher.csv")
summary(p)
mean(p$Speed)
boxplot(p$Speed)
```

あれ、彼は速球投手として知られているのだが...?

入力してみよう

```
hist(p$Speed)
```

投球データ（続き）

データが単峰か多峰かは重要なこと

入力してみよう

```
table(p$Type)
barplot(table(p$Type))
pie(table(p$Type))
by(p$Speed, p$Type, mean)
boxplot(p$Speed ~ p$Type)
```

こぼれ話...円グラフ

help(pie)にはこのように書いてある：

Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.

過去3年の体重・体脂肪率の記録が `weight.csv` にある

入力してみよう

```
w <- read.csv("weight.csv")  
summary(w)
```

このデータを元にしたグラフを作成してみよう

体重・体脂肪率（続き）

入力してみよう

```
plot(w$Date, w$Weight, type="l")
```

- 変動がはげしすぎて傾向がわからない
- 平滑化した曲線を重ねてみる（lowess 関数）

入力してみよう

```
lines(lowess(w$Date, w$Weight))
```

体重・体脂肪率（続き）

入力してみよう

```
plot(w$Date, w$Weight, type="l")
```

- 変動がはげしすぎて傾向がわからない
→ 平滑化した曲線を重ねてみる（lowess 関数）

入力してみよう

```
lines(lowess(w$Date, w$Weight))
```

- この結果はまずい! どうするべきか...

体重・体脂肪率（続き）

入力してみよう

```
plot(w$Date, w$Weight, type="l")
```

- 変動がはげしすぎて傾向がわからない
- 平滑化した曲線を重ねてみる（lowess 関数）

入力してみよう

```
lines(lowess(w$Date, w$Weight))
```

- この結果はまずい! どうするべきか...
- 別のデータを探ろう

体重・体脂肪率（続き）

体脂肪率についても同じことをやってみる

入力してみよう

```
plot(w$Date, w$FatRatio, type="l")  
lines(lowess(w$Date, w$FatRatio))
```

体重・体脂肪率（続き）

体脂肪率についても同じことをやってみる

入力してみよう

```
plot(w$Date, w$FatRatio, type="l")  
lines(lowess(w$Date, w$FatRatio))
```

体脂肪の量を調べてみてはどうか

入力してみよう

```
w$Fat <- w$Weight * w$FatRatio / 100  
plot(w$Date, w$Fat, type="l")  
lines(lowess(w$Date, w$Fat))
```

体重・体脂肪率（続き）

このグラフを PDF にして保存しよう

入力してみよう

```
pdf("weight.pdf")  
（さっきまでやったことを全部やる）  
dev.off()
```

体重・体脂肪率（続き）

このグラフを PDF にして保存しよう

入力してみよう

```
pdf("weight.pdf")  
（さっきまでやったことを全部やる）  
dev.off()
```

- もう 1 回やるの？ めんどくさすぎ!!

体重・体脂肪率（続き）

このグラフを PDF にして保存しよう

入力してみよう

```
pdf("weight.pdf")  
(さっきまでやったことを全部やる)  
dev.off()
```

- もう 1 回やるの? めんどくさすぎ!!
- 「こんなこともあるかと」この手順を `weight.R` というファイルに書いておいた

入力してみよう

```
pdf("weight.pdf")  
source("weight.R")  
dev.off()
```

- できあがった `weight.pdf` を見てみよう

体重・体脂肪率（続き）

入力してみよう

```
pdf("weight.pdf")  
source("weight.R")  
dev.off()
```

- できあがった weight.pdf を見てみよう
- weight1.R にはもっと細かい指定が書いてある

入力してみよう

```
pdf("weight1.pdf")  
source("weight1.R")  
dev.off()
```

- 最近 R 関係の本が増えた
- 何か一冊は手元にほしい（ヘルプが英語だから）
- 私の手持ちは
 - データ解析環境「R」 舟尾暢男・高浪洋平著，工学社
 - Rによるやさしい統計学 山田剛史・杉澤武俊・村井潤一郎著，
オーム社
- 本屋で気に入ったものを...

参考になるサイト

R project <http://www.r-project.org>

R の総本山

ダウンロードはここから

奥村晴彦先生のサイト

<http://oku.edu.mie-u.ac.jp/~okumura/stat/>

実際の作業上での細かい注意点がうれしい

R-tips <http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html>

RjpWiki <http://www.okada.jp.org/RWiki/>

R による統計処理 <http://aoki2.si.gunma-u.ac.jp/R/>

Nagoya.R

- 阪上辰也氏（名古屋大）が呼びかけ
- <http://corpus-study.info/nagoyar/>
- twitter @nagoyar
- 勉強会...「入門者講習」を毎回実施
- 次の勉強会は9月（予定）
- これまでの勉強会資料も見られる





Windows にインストールするときの注意

- インストール中に言語を聞かれたら、English を選ぶ
日本語だとインストーラが途中で文字化け
実行環境は日本語なので大丈夫
- 「編集」→「GUI プリファレンス」で Font を MS Mincho など
日本語用のものにする