

# データ解析 01 イントロ Rに慣れる

東工大 情報科学科  
下平英寿

イントロダクション  
今日は雑談です

# 「データ解析」の課題より...

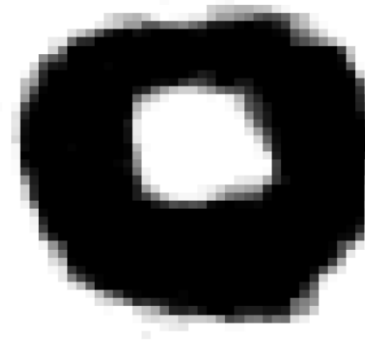
観測画像 Y



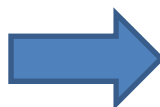
画像復元



推定画像 X



# 画像復元



1 = 0.0425   2 = 0.0456   3 = 0.0485   4 = 0.052   5 = 0.052   6 = 0.0532  
東 東 東 東 東 東

7 = 0.0563   8 = 0.0564   9 = 0.0567   10 = 0.0575   11 = 0.0585   12 = 0.0594  
東 東 東 東 東 東

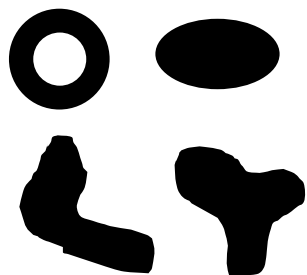
13 = 0.0605   14 = 0.0609   15 = 0.0615   16 = 0.0626   17 = 0.0632   18 = 0.0709  
東 東 東 東 東 東

19 = 0.0731   20 = 0.0736   21 = 0.0743   22 = 0.0757   23 = 0.0812   24 = 0.1094  
東 東 東 東 東 東

25 = 0.1133   26 = 0.1333  
東 東

# $3 \times 10^{752}$ 通りをどうやって計算？

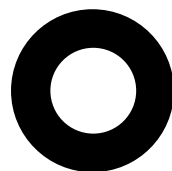
Xの事前分布



$P(X)$



真の画像 X



$P(Y | X)$

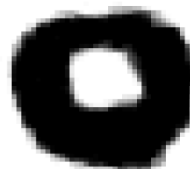


ノイズの確率モデル

観測画像 Y



推定画像 X



$$P(X | Y) = \frac{P(Y | X)P(X)}{\sum_{X'} P(Y | X')P(X')}$$

50x50=2500ピクセル

Xの組み合わせ =  $2^{2500} = 3 \times 10^{752}$  通り

ベイズ  
(1702-1761)



マルコフチェイン・モンテカルロ法 (MCMC法, ギブス・サンプラー)

ところで

「確率」と「統計」の違いは？

# 「統計」では「確率」も使います

- amazon.co.jpの「確率・統計」

本: 科学・テクノロジー > 数学 > 確率・統計

検索結果 845件中1件から12件までを表示 並び替え: 売れている順番

- 

マンガでわかる統計学 高橋 信、トレンドプロ (単行本 - 2004/7)  
新品: ¥ 2,100  
4新品 ¥ 2,100より 15中古品 ¥ 1,340より  
通常4~6日以内に発送  
★★★★☆ (53) Amazonプライム
- 

完全独習 統計学入門 小島 寛之 (単行本(ソフトカバー) - 2006/9/29)  
新品: ¥ 1,890  
18中古品 ¥ 1,269より  
通常4~6日以内に発送  
★★★★☆ (25) Amazonプライム
- 

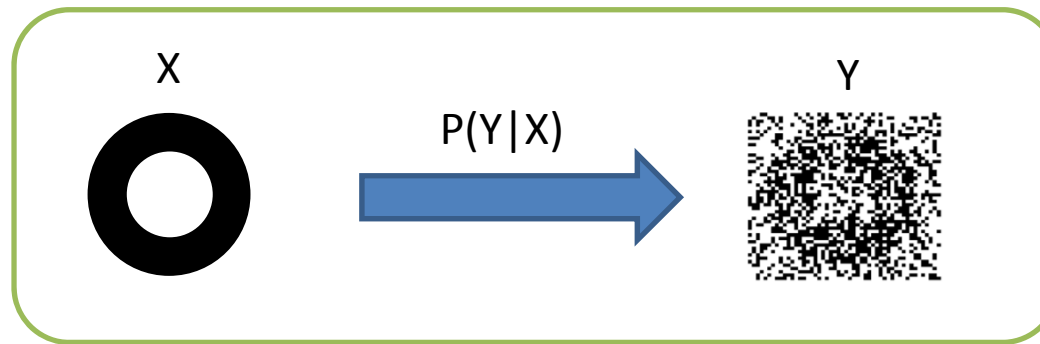
ビジネスマンのための「数字力」養成講座 (ディスカヴァー携書) 小宮 一麿 (新書 - 2008/2/27)  
新品: ¥ 1,050  
34中古品 ¥ 131より  
15点在庫あり。ご注文はお早めに。  
★★★★☆ (38) Amazonプライム
- 

はじめての統計学 鳥居 泰彦 (単行本 - 1994/11)  
新品: ¥ 2,345  
23中古品 ¥ 700より  
通常5~8日以内に発送  
★★★★☆ (46) Amazonプライム
- 

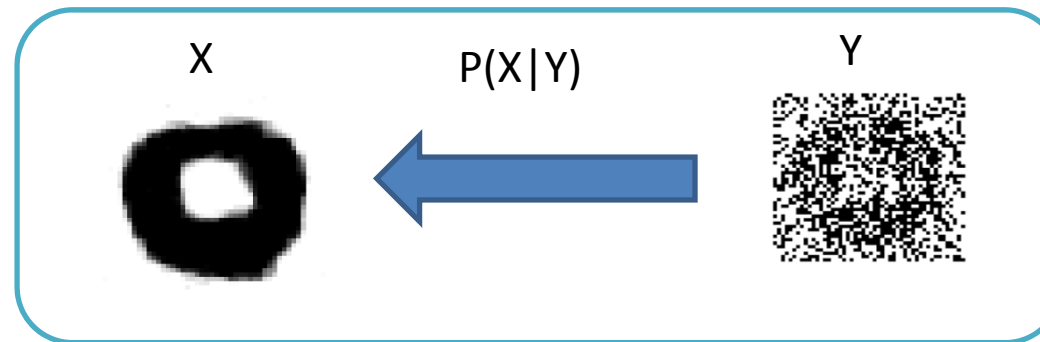
マンガでわかる統計学 回帰分析編 高橋 信、井上 いろは、トレンドプロ (単行本 - 2005/9)  
新品: ¥ 2,310  
3新品 ¥ 1,900より 9中古品 ¥ 1,800より  
通常4~6日以内に発送  
★★★★☆ (13) Amazonプライム
- 

SPSS 統計分析 小田 利勝 (単行本 - 2007/5)  
新品 ¥ 2,310  
2新品 ¥ 2,310より 2中古品 ¥ 3,447より

# 統計学は逆向きの思考



「確率」



「統計」

演繹と帰納



# 「データの科学」の歴史



- 1702-1761 ベイズ
- 1795 ガウス 最小2乗法, 正規分布
- 1890頃 ガルトン 相関, 回帰分析
- 1908 ゴセツト(スチューデント) t-分布
- 1920頃 フィッシャー 最尤法
- 1950頃? ロジスティック回帰
- 1970 ヘイスティングス MCMC法
- 1974 赤池 情報量規準
- 1979 エフロン ブートストラップ法



画像の課題で使う

下平研の研究

統計科学  
数学, コンピュータ  
機械学習  
生命科学

データから情報を抽出する  
アルゴリズム + コンピューティング

# MITの講義資料

## “Markov chain Monte Carlo”を検索

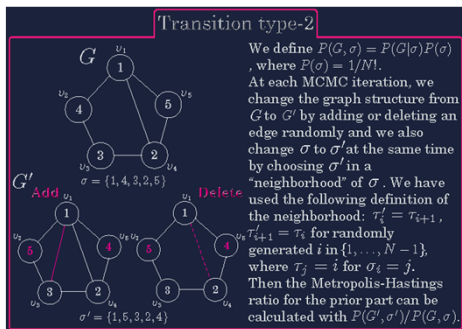
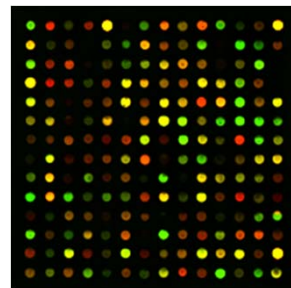
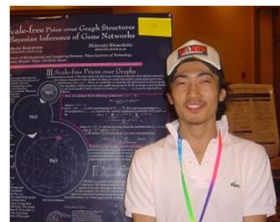
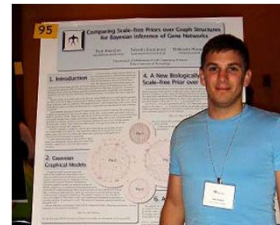
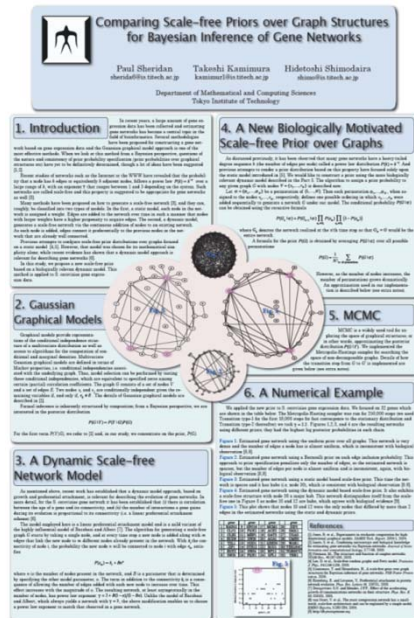
The screenshot shows the MIT OpenCourseWare website interface. At the top, there's a navigation bar with links like 'OCW HOME', 'COURSE LIST', 'ABOUT OCW', 'HELP', 'FEEDBACK', and a 'GIVE NOW' button. Below this, a search bar on the left contains the text 'Markov chain Monte' with a 'GO' button and a link to 'Advanced Search'. The main content area is titled 'Search Results' and shows 'Results 1 - 10 of about 23 for Markov chain Monte Carlo. Sort by: Date / Relevance'. Three search results are visible, each starting with a PDF icon and a link to a document. The first result is '18.366 Final Project: Using Bayesian Markov Chain Monte Carlo to Markov Chain Monte Carlo methods are used in a wide variety of applications, ranging for molecular modeling [1] to financial studies [2]. The appeal of the ocw.mit.edu/NR/rdonlyres/Mathematics/18-366Spring-2005/A5D4853A-4FA6-4265-9DD5-7BE6C23CE974/0/Allen\_proj05.pdf - 2006-08-16'. The second result is 'Sampling Good Motifs with Markov Chains Abstract Markov chain Monte Carlo (MCMC) techniques have been used with some success in bioinformatics [LAB + 93]. 1.2 The Markov Chain Monte Carlo Method ocw.mit.edu/NR/rdonlyres/Mathematics/18-417Fall-2004/372BB8F3-2BEE-450B-9F2B-5E4D38CC9551/0/cjp\_project.pdf - 2006-12-20 [ More results from ocw.mit.edu/NR/rdonlyres/Mathematics ]'. The third result is '16.412 Pset #1 Shuonan Dong 2/15/05 Part A: Topics of Fascination Algorithms such as Markov chain Monte Carlo and expectation-maximization, and models such as Kernel models and neural networks may be useful. ocw.mit.edu/NR/rdonlyres/Aeronautics-and-Astronautics/16-412JSpring-2005/3EB51BB4-486F-4993-9F5C-039112123A58/0/dongs\_pset1.pdf - 2006-06-26'. A fourth result is a text file: 'Bayes MCMC pred MR.m % % function g\_pred = Bayes MCMC pred MR Param); % % This routine computes the expectation of a vector g, given a % set of multi-response data, using Markov Chain Monte Carlo % (MCMC) simulation.'

マルコフチェーン・モンテカルロ法 : MIT(または東工大)の大学院講義ですね...

ちょっと難しいかも... とりあえずやってみる!

# 研究プロジェクト1

## MCMC法でグラフ構造を推定



(応用: DNAチップから  
遺伝子ネットワークを推定)

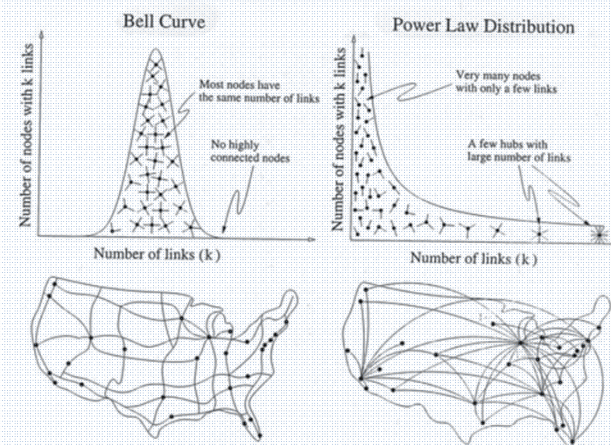
$k$  = "degree" : the number of edges  
or "links" of each node

$$P(k) \propto k^{-\gamma}$$



www  $\gamma = 2.1 \sim 2.7$   
 gene  $\gamma = 2.2$

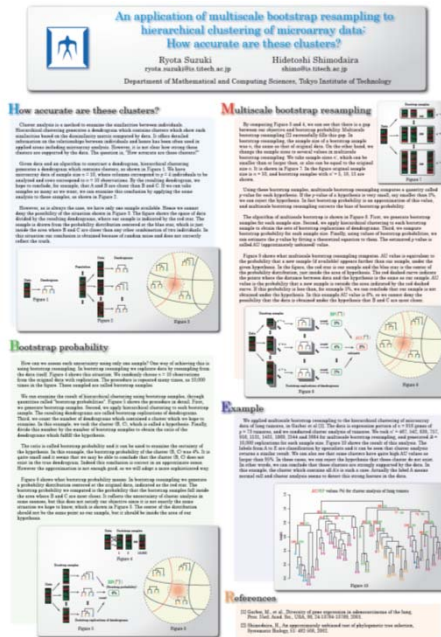
$k$ : 各ノードのリンク数



Barabasi, A., Albert, R., Emergence of scaling in random networks, Science, 1999.

# 研究プロジェクト2

## 「分類」のランダムネスを計算



鈴木了太くんの修論

「マルチスケール・ブートストラップ法」を分類の確率値計算のために実装した. 並列計算による高速化.

- Rの公式ライブラリ「pvclust」を作成
- 鈴木君の書いた学术论文: Bioinformatics誌に掲載
- 生物学のクローン研究などで利用されている

アルゴリズム+数学理論 (下平研の論文)

PNAS January 24, 2006 (online)  
(They used pvclust before publication of pvclust)

Brambrink et al.  
ES cells derived from cloned...

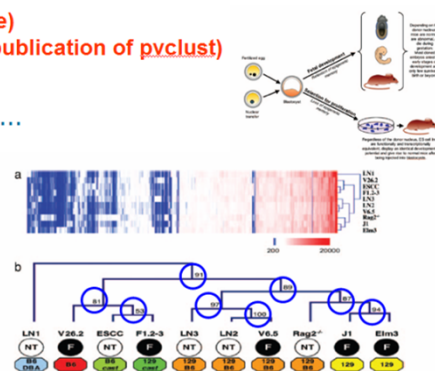


Fig. 2. Hierarchical clustering of individual ESC line expression profiles. (a) Heat map of clustering results (blue, no or very low expression; white, low expression; red, high expression). (b) Sample tree obtained from hierarchical clustering. ES cell line expression profiles cluster by genetic background (colored ovals) rather than by type of donor blastocyst (NT, cloned; F, fertilized; numbers next to nodes display multiscale bootstrap resampling probability based on 10,000 replications).

$\frac{1}{24}\phi^{pppp}, \bar{b}_0^c = -\frac{1}{2}d^{aa}\phi^{cpp} + d^{ab}\phi^{abc} - 3e^{aac}$  and  $\bar{b}_2^c = \frac{1}{2}\phi^{bpp}(2d^{bc} - \phi^{bcp}) - \frac{1}{4}\phi^{cpp}\phi^{ppp} + \frac{1}{6}\phi^{cPPP}$ . Then the distribution function of  $\hat{z}_\infty(y)$  is obtained immediately from Lemma 4 as shown below.

LEMMA 6. Let us consider a statistic

$$\hat{z}_q(y) \approx \hat{z}_\infty(y) + q_0 + q_1\hat{v} + q_2\hat{v}^2 + q_3\hat{v}^3 + \hat{u}_c g^c(\hat{v}),$$

where the coefficients are  $q_0 = O(n^{-1/2})$ ,  $q_1 = O(n^{-1})$ ,  $q_2 = O(n^{-1/2})$  and  $q_3 = O(n^{-1})$ , and  $g^c(\hat{v}) = O(n^{-1})$ ,  $c = 1, \dots, p-1$ , representing arbitrary polynomials of  $\hat{v}$ . The index  $q$  of  $\hat{z}_q$  indicates the coefficients. Assuming  $(\hat{U}, \hat{V}) \sim f(\hat{u}, \hat{v}; \lambda, 1)$ , the distribution function of  $\hat{z}_q(y)$  is expressed as

$$\Pr\{\hat{z}_q(Y) \leq x; \lambda\}$$

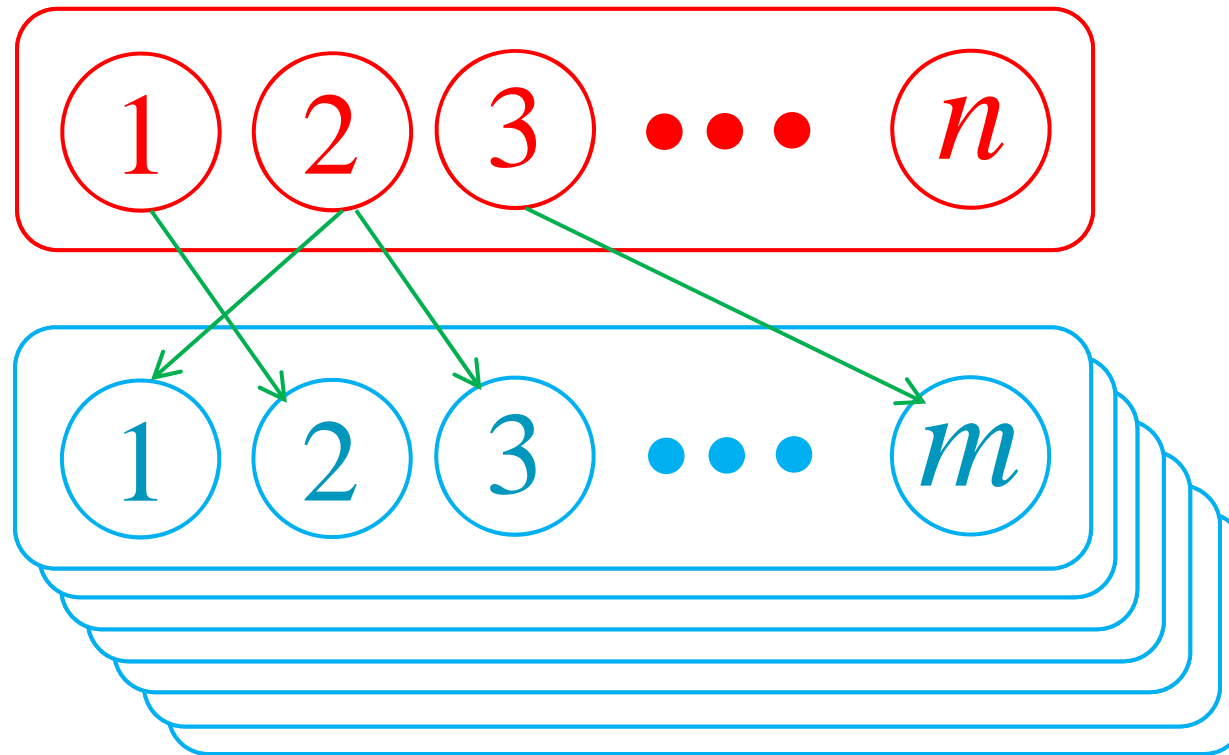
$$\approx \Phi[x - \lambda - q_0 - \frac{1}{3}\phi^{ppp}\lambda^2 + \frac{1}{6}\phi^{ppp}\lambda x - q_2x^2$$

$$+ \{(d^{ab})^2 + \frac{1}{8}(\phi^{app})^2 + \frac{7}{72}(\phi^{ppp})^2 - \frac{1}{24}\phi^{pppp} - \frac{1}{6}\phi^{ppp}q_0\}\lambda.$$

12



# サンプルサイズを $m$ に変更



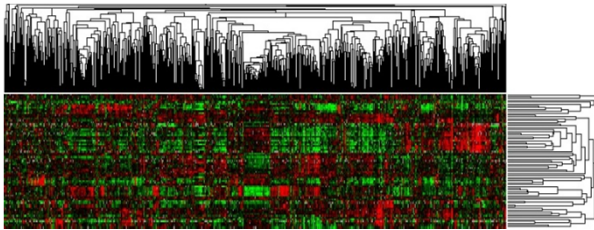
## ブートストラップ法

普通は当然  $m=n$

最新手法は???

# 並列計算でスピードアップ (下平研究室 2008/04/10)

問題: DNAチップから癌の診断



東工大のスパコン  
TSUBAME+TSUBASA

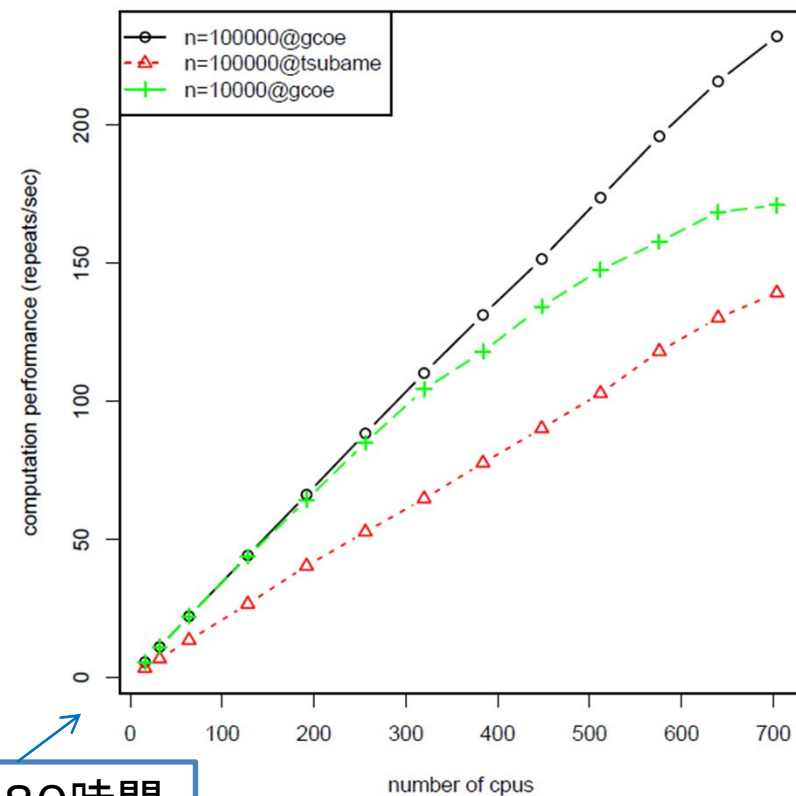


毎秒あたりの計算量

1個 = 80時間

700個 = 7分

bootstrap computation (pvclust)



計算プロセッサ(コア)の個数

## データ解析 2011年度

日時： 金曜3-4時限 10:45－12:15 西8号館W834講義室.

評価方法： レポート, ミニテストなど (出席はとりません. 期末試験は行いません.)

教員： 下平英寿 (数理・計算科学専攻 准教授, 情報科学科を担当します)

TA： 永田(M1), 金城(M1)

メールアドレス: [shimo-data2@is.titech.ac.jp](mailto:shimo-data2@is.titech.ac.jp) レポート添付ファイルの提出先や質問受付のために講義専用のメールアドレスを用意しています.

講義ウェブサイト(2011年度版)

<http://www.is.titech.ac.jp/~shimo/class/data2011/index.html>

2008年版の講義資料 (uda2008/main.tex 2008/05/02) を随時参照します. 各自でダウンロードしてください. たぶん印刷して持参する必要はないと思います.

講義や資料にないことは, ウェブなど利用して各自が自分で調べることが期待されています.

# レポート

- レポートボックス「データ解析」 西8W3階のエレベータ付近
- 課題を出した授業から1週間後の13:00
- 基本的に紙で提出



# ウェブサイト

<http://www.is.titech.ac.jp/~shimo/class/data2011/index.html>

## データ解析 2011年度

担当： 下平英寿 西8W-707 ([下平のウェブサイト](#), [下平研究室のウェブサイト](#))

日時： 火曜 3限, 4限 10:45-12:15 西8号館W834講義室.

評価方法： レポート, ミニテスト

講義で使ったスクリプトとデータは[このフォルダ](#)にあります

今年度から講義資料はOCWの「[理学部>情報科学科>データ解析](#)」

[講義と演習のページへ戻る](#)

The screenshot shows the OCW website for the Data Analysis course. The header includes the Tokyo Tech logo and the text "国立大学法人 東京工業大学". Below the header, there are statistics: "876 講義ノート公開中" and "10 音声・動画公開中". A search bar is present. The main content area is titled "データ解析 Data Analysis (下平 英寿)" and lists the course details: "火曜日3-4時限開講 W834". It also shows the update date "更新日: 2011年4月7日" and the access index "アクセス指標: ★★★★★". The course description mentions "数理統計学の基礎知識 (確率と統計第二) 理ソフトウェアであるRを用いて自分自身でデータ解析を行う". The course objectives and plan are also listed.

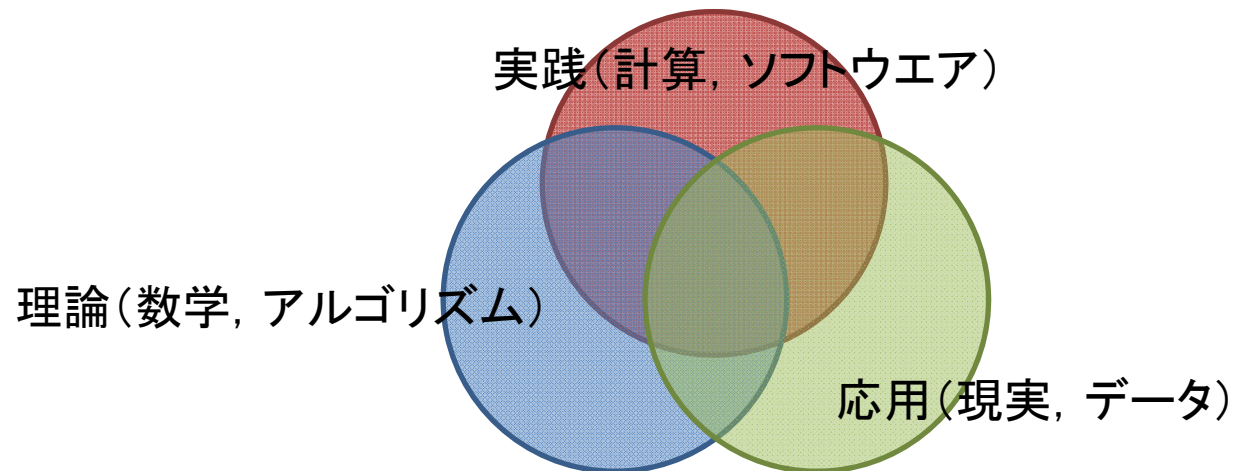
This screenshot shows the detailed course page for Data Analysis. It includes the course title "データ解析 Data Analysis (下平 英寿)" and the course details "火曜日3-4時限開講 W834". The syllabus is listed with the following items:

- 第1回 1 イン트로 (Lecture 1: Intro) - 平成23年04月12日(火) 3-4時限開講
- 第2回 2 回帰分析 (1) (Lecture 2: Regression Analysis (1)) - 平成23年04月19日(火) 3-4時限開講
- 第3回 3 回帰分析 (2) (Lecture 3: Regression Analysis (2)) - 平成23年04月26日(火) 3-4時限開講

Each lecture entry includes a "講義" (Lecture) button and a "添付資料" (Attachment) link with the file size and star rating.

# 今後の授業予定

1. イントロ
2. 回帰分析(その1) 実践編
3. 回帰分析(その2) 理論編
4. ロジスティック回帰 (スパムメール判別)
5. 検定・信頼区間 (理論)
6. ミニテスト?
7. . . .



# シラバス

講 義 名	データ解析 (Data Analysis)		
開 講 時 期	第 5 学期	単 位 数	2-0-0
担 当 教 員	下平 英寿 西 8 号館 (W) 7 階 707 号室 (内線 3219)		

【講義のねらい】

「Rを用いたデータ解析入門」、統計処理ソフトウェアであるRを利用して実践的なデータ解析ができるようになること（Rに含まれる関数を呼び出してデータ解析を実行する）、背後にある数学、統計学、アルゴリズムを理解すること（自分自身で関数を記述し、それを用いてデータ解析を行う）を目標とする。

【講義計画】

1. イントロダクション（社会人口統計データ、バイオインフォマティクス）

2. 期待値、大数の法則（ポートフォリオ、ヒストグラム）

3. モンテカルロ法（MCMC法、ギブスサンブラ）

4. ベイズの定理（画像復元）

5. 積率母関数、中心極限定理

6. 確率モデル（正規混合分布）

7. 判別問題、分類、識別（スパムメール判別）

8. パラメタ推定（最尤推定）

9. EMアルゴリズム（教師無し学習）

10. 最尤推定量の性質（クラメール・ラオの不等式、フィッシャー情報行列）

11. 検定と信頼区間

12. 線形回帰分析（ボストン住宅価格）

13. ロジスティック回帰分析（スパムメール判別のつづき、ニュートン法）

14. 主成分分析（ボストン住宅価格のつづき）

【成績評価】

レポート提出。

【テキスト等】

講義資料（PDF形式）を講義ウェブサイトから各自ダウンロードする。

【履修の条件】

理論的な理解を深めるためには確率と統計第一・第二を履修していることが望ましいが、履修していなくても可能。Rにシンタックスが類似の言語（Java等）の経験があればよい。Rプログラミングの詳細は講義中で説明せず、アルゴリズムの説明をとおして多少説明する程度である。詳細は講義ウェブサイトを参照

<http://www.is.titech.ac.jp/~shimo/class>

# 「データ解析」の講義目標

1. Rを利用してデータ解析する : Rに含まれる関数を使う
2. 背後にある数学, 統計学, アルゴリズムを理解する : 自分で関数を作る

Rってなんですか？

# ja.wikipedia.org/wiki/R言語

## R言語

**R言語**(アールげんご)は、オープンソースでフリーソフトウェアの統計解析向けプログラミング言語、及びその開発実行環境である。

R言語は、ニュージーランドのAuckland大学のRoss IhakaとRobert Gentlemanにより作られた。現在では、R Development Core Team (S言語開発者であるJohn M. Chambersも参画。 [R Project Contributors](#)) によって、メンテナンスと拡張がなされている。

なお、R言語仕様を実装した処理系の呼称名はプロジェクトを支援するフリーソフトウェア財団によれば **GNU R** (GNU R - Free Software Directory) だが、他の実装形態が存在しないため当記事では日本での慣用的呼称にならない仕様・実装をまとめて適宜“R言語”や“R”等と呼ぶ。

### 目次 [非表示]

#### 1 特徴

- 1.1 ベクトル処理言語
- 1.2 統計に適した解析環境
- 1.3 メルセンヌ・ツイスタによる乱数生成
- 1.4 高速な組み込み関数群
- 1.5 視覚化に優れたグラフ機能
- 1.6 データ互換性
- 1.7 ユーザープログラムを配信・利用できるCRANネットワーク機能
- 1.8 教育現場から実務・研究現場へ永続的に利用可能

#### 2 言語仕様

- 2.1 制御構造、サブルーチン
- 2.2 オブジェクト指向
- 2.3 データ型
  - 2.3.1 ベクトルとリスト
  - 2.3.2 データフレーム

#### 3 機能

- 3.1 データ入出力
- 3.2 データのプロット
- 3.3 ワークスペースの保存

#### 4 その他

### R言語



<b>パラダイム</b>	マルチパラダイム、関数型、オブジェクト指向、命令型
<b>登場時期</b>	1996年
<b>設計者</b>	Ross Ihaka、Robert Gentleman (共)
<b>開発者</b>	R Development Core Team
<b>最新リリース</b>	2.12.1 / 2010年12月15日
<b>型付け</b>	動的型付け
<b>主な処理系</b>	GNU R
<b>影響を受けた言語</b>	S言語、Scheme
<b>プラットフォーム</b>	クロスプラットフォーム
<b>ライセンス</b>	GNU GPL
<b>ウェブサイト</b>	<a href="http://www.r-project.org">www.r-project.org</a>

## S言語

**S言語** (えすげんご) とは、1984年、AT&Tベル研究所のJohn Chambers、Rick Becker、Allan Wilks らによって研究・開発された統計処理言語である。当初は「Sシステム」と呼ばれ、UNIX上における統計処理を行うソフトのコマンドの役割を果たす言語として開発された。その後、機能の拡張を続け1988年にプログラミング言語としてのS言語が策定された。さらに1991年にオブジェクト指向プログラミングの機能が追加され現在に至る。

言語の特長はChambersによって、*to turn ideas into software, quickly and faithfully*と説明されている。

GNUプロジェクトによりオープンソースで作成されているR言語はS言語の文法を取り入れており、ほぼ同等の機能をもつ。S言語自体の処理系としては商用版のS-PLUSが知られている。

S言語は1998年にACMのソフトウェアシステム賞を獲得している。

### S

<b>設計者</b>	Rick Becker, Allan Wilks, John Chambers
<b>開発者</b>	ベル研究所
<b>主な処理系</b>	S-PLUS
<b>影響を与えた言語</b>	R言語
<b>ウェブサイト</b>	<a href="#">S-PLUS</a>

表・話・編・歴



### プログラミング言語

>>他のプログラミング言語

■カテゴリ / ■テンプレート

Rをつかってみる

# Rのダウンロード

- CRANを検索して, <http://cran.r-project.org/>
- Windows => baseとたどって次をクリック
- [Download R 2.12.2 for Windows](#)

The screenshot shows the CRAN website with the following structure:

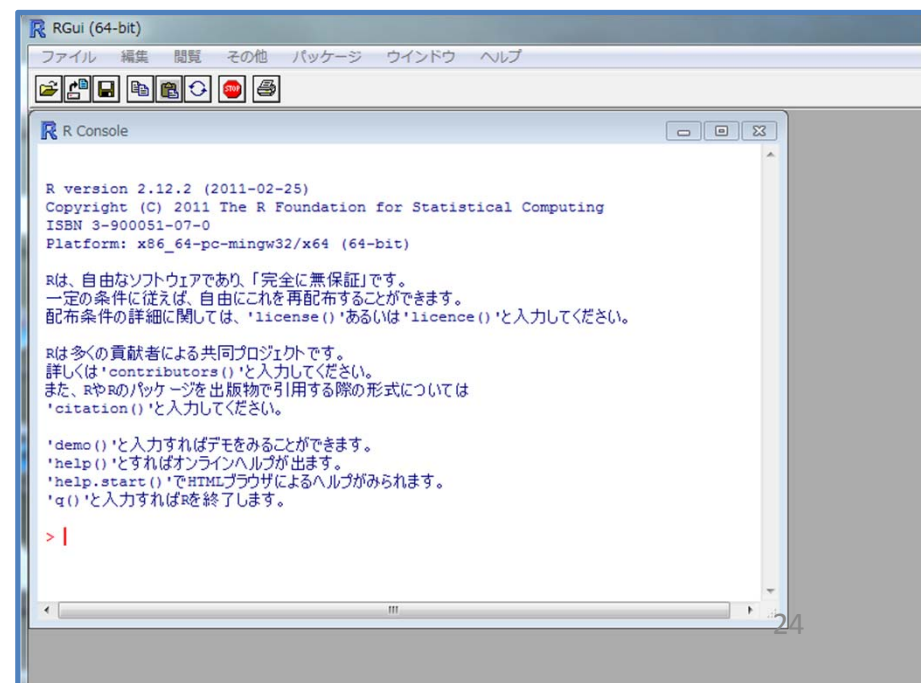
- CRAN logo and navigation links: [CRAN](#), [Mirrors](#), [What's new?](#), [Task Views](#), [Search](#), [About R](#), [R Homepage](#), [The R Journal](#), [Software](#), [R Sources](#), [R Binaries](#), [Packages](#), [Other](#), [Documentation](#), [Manuals](#), [FAQs](#), [Contributed](#).
- Subdirectories:
  - [base](#): Binaries for base distribution (managed by Duncan Murdoch)
  - [contrib](#): Binaries of contributed packages (managed by Uwe Ligges)
- R-2.12.2 for Windows (32/64 bit)
  - [Download R 2.12.2 for Windows](#) (37 megabytes, 32/64 bit)
  - [Installation and other instructions](#)
  - New features in this version: [Windows specific](#), [all platforms](#).
  - If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.
  - Frequently asked questions

# Rのインストール

- ダウンロードしたR-2.12.2-win.exeを実行
- 指示に従ってクリックしていけば完了
- アイコンをクリックすれば実行



(私の環境Windows7 64bitでは32bit版と64bit版の両方がインストールされました)





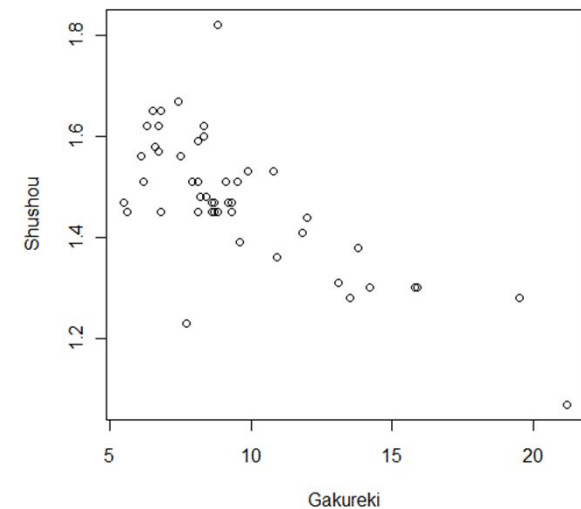
# 表形式データの読み込み

- `dat = read.table("gakureki-shushou.txt")`
- `plot(dat)`

## テキストファイルのデータ

"Gakureki" "Shushou"  
"Hokkaido" 7.7 1.23  
"Aomori" 5.5 1.47  
"Iwate" 6.1 1.56  
"Miyagi" 9.6 1.39  
.....

```
> dat = read.table("gakureki-shushou.txt")
> dat[1:5,]
      Gakureki Shushou
Hokkaido    7.7    1.23
Aomori      5.5    1.47
Iwate       6.1    1.56
Miyagi      9.6    1.39
Akita       5.6    1.45
> dim(dat)
[1] 47  2
> plot(dat)
> |
```



# ヘルプのみかた

## • help(plot)とタイプ

plot {graphics} R Documentation

Generic X-Y Plotting

Description  
Generic function for plotting of R objects. For more details about the graphical parameter arguments see [par](#).

Usage  
`plot(x, y, ...)`

Arguments

- x the coordinates of points in the plot. Alternatively, a single plotting structure, function or *any R object with a plot method* can be provided.
- y the y coordinates of points in the plot, *optional* if x is an appropriate structure.
- ... Arguments to be passed to methods, such as graphical parameters (see [par](#)). Many methods will accept the following arguments:

type  
what type of plot should be drawn. Possible types are

- "p" for **p**oints,
- "l" for **l**ines,
- "b" for **b**oth,
- "c" for the lines part alone of "b",
- "o" for both **o**verplotted,
- "h" for **'h**istogram' like (or 'high-density') vertical lines,
- "s" for stair **s**teps,
- "S" for other **s**teps, see 'Details' below,
- "n" for no plotting.

All other types give a warning or an error; using, e.g., type = "punkte" being equivalent to type = "p" for S compatibility. Note that some methods, e.g. [plot.factor](#), do not accept this.

main  
an overall title for the plot: see [title](#).

sub  
a sub title for the plot: see [title](#).

xlab  
a title for the x axis: see [title](#).

ylab  
a title for the y axis: see [title](#).

asp  
the y/x aspect ratio, see [plot.window](#).

```
plot(x, y, ...)
```

### Arguments

- x the coordinates of points in the plot. Alternatively, a single plotting structure, function or *any R object with a plot method* can be provided.
- y the y coordinates of points in the plot, *optional* if x is an appropriate structure.
- ... Arguments to be passed to methods, such as graphical parameters (see [par](#)). Many methods will accept the following arguments:

### Details

For simple scatter plots, [plot.default](#) will be used. However, there are `plot` methods for many R objects, including [functions](#), [data.frames](#), [density](#) objects, etc. Use `methods(plot)` and the documentation for these.

The two step types differ in their x-y preference: Going from (x1,y1) to (x2,y2) with x1 < x2, type = "s" moves first horizontal, then vertical, whereas type = "S" moves the other way around.

### See Also

[plot.default](#), [plot.formula](#) and other methods; [points](#), [lines](#), [par](#).

For X-Y-Z plotting see [contour](#), [persp](#) and [image](#).

### Examples

```
require(stats)
plot(cars)
lines(lowess(cars))

plot(sin, -pi, 2*pi)

## Discrete Distribution Plot:
plot(table(rpois(100,5))), type = "h", col = "red", lwd=10,
      main="rpois(100,lambda=5)")

## Simple quantiles/ECDF, see ecdf(), {library(stats)} for a better one:
plot(x <- sort(rnorm(47)), type = "s", main = "plot(x, type = 's')")
points(x, cex = .5, col = "dark red")
```

# ヘルプは全部読む必要なし！

- plot.default

plot.default {graphics}

R Documentation

The Default Scatterplot Function

Description

Draw a scatter plot with decorations such as axes and titles in the active graphics window.

Usage

## Default S3 method:  
plot(x, y = NULL, type = "p", xlim = NULL, ylim = NULL,  
     log = "", main = NULL, sub = NULL, xlab = NULL, ylab = NULL,  
     ann = par("ann"), axes = TRUE, frame.plot = axes,  
     panel.first = NULL, panel.last = NULL, asp = NA, ...)

Arguments

x, y

the x and y arguments provide the x and y coordinates for the plot. Any reasonable way of defining the coordinates is acceptable. See the function [xy.coords](#) for details. If supplied separately, they must be of the same length.

type

1-character string giving the type of plot desired. The following values are possible, for details, see [plot](#): "p" for points, "l" for lines, "o" for overplotted points and lines, "b", "c" for (empty if "c") points joined by lines, "s" and "S" for stair steps and "h" for histogram-like vertical lines. Finally, "n" does not produce any points or lines.

xlim

the x limits (x1, x2) of the plot. Note that x1 > x2 is allowed and leads to a 'reversed axis'.

ylim

the y limits of the plot.

log

a character string which contains "x" if the x axis is to be logarithmic, "y" if the y axis is to be logarithmic and "xy" or "yx" if both axes are to be logarithmic.

main

a main title for the plot, see also [title](#).

sub

a sub title for the plot.

xlab

a label for the x axis, defaults to a description of x.

ylab

a label for the y axis, defaults to a description of y.

ann

a logical value indicating whether the default annotation (title and x and y axis labels) should appear on the plot.

axes

a logical value indicating whether both axes should be drawn on the plot. Use graphical parameter "xaxt" or "yaxt" to suppress just one of the axes.

frame.plot

a logical indicating whether a box should be drawn around the plot.

panel.first

an expression to be evaluated after the plot axes are set up but before any plotting takes place. This can be useful for drawing background grids or scatterplot smooths.

panel.last

an expression to be evaluated after plotting has taken place.

asp

the y/x aspect ratio, see [plot.window](#).

...

other graphical parameters (see [par](#) and section 'Details' below).

- methods(plot)

```
> methods(plot)
[1] plot.acf*          plot.data.frame*   plot.decomposed.ts*
[4] plot.default       plot.dendrogram*   plot.density
[7] plot.ecdf          plot.factor*        plot.formula*
[10] plot.hclust*       plot.histogram*     plot.HoltWinters*
[13] plot.isoreg*       plot.lm             plot.medpolish*
[16] plot.mlm           plot.ppr*           plot.prcomp*
[19] plot.princomp*     plot.profile.nls*   plot.spec
[22] plot.spec.coherency plot.spec.phase     plot.stepfun
[25] plot.stl*          plot.table*         plot.ts
[28] plot.tskernel*     plot.TukeyHSD
```

Non-visible functions are asterisked

# 直線を当てはめてみる(単回帰分析)

- `kekka = lm(Shushou ~ Gakureki, dat)`
- `abline(kekka,col="red")`

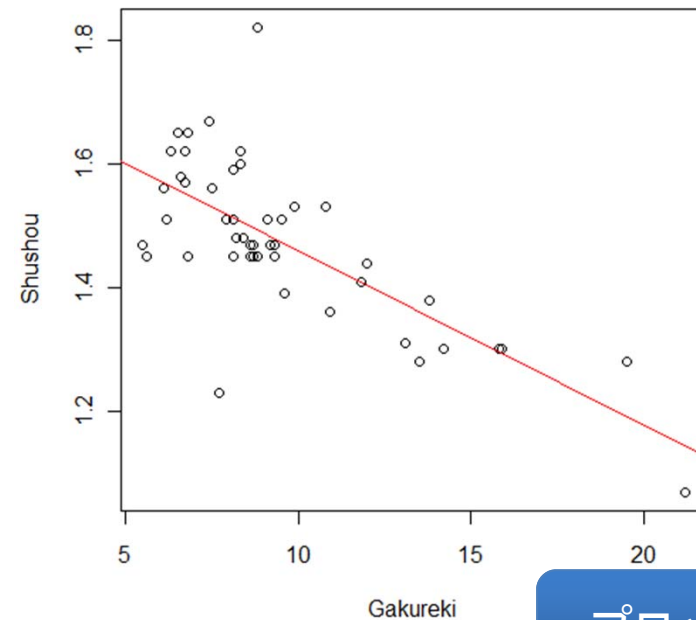
```
> dat = read.table("gakureki-shushou.txt")
> dat[1:5,]
      Gakureki Shushou
Hokkaido    7.7    1.23
Aomori      5.5    1.47
Iwate       6.1    1.56
Miyagi      9.6    1.39
Akita       5.6    1.45
> dim(dat)
[1] 47 2
> plot(dat)
> kekka = lm(Shushou ~ Gakureki, dat)
> kekka

Call:
lm(formula = Shushou ~ Gakureki, data = dat)

Coefficients:
(Intercept)    Gakureki
    1.74248     -0.02825

> abline(kekka,col="red")
> |
```

実行結果



プロットの図

# lmは線形モデルのあてはめ

- `kekka = lm(Shushou ~ Gakureki, dat)`

lm {stats}

R Documentation

## Fitting Linear Models

### Description

lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although [aov](#) may provide a more convenient interface for these).

### Usage

```
lm(formula, data, subset, weights, na.action,  
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,  
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

$$y = \beta_0 + \beta_1 x + \varepsilon$$

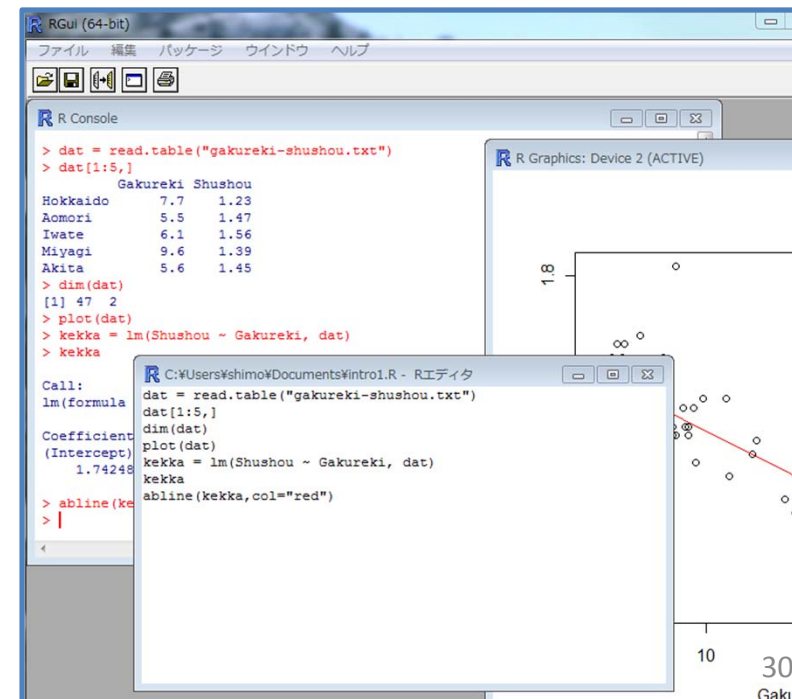
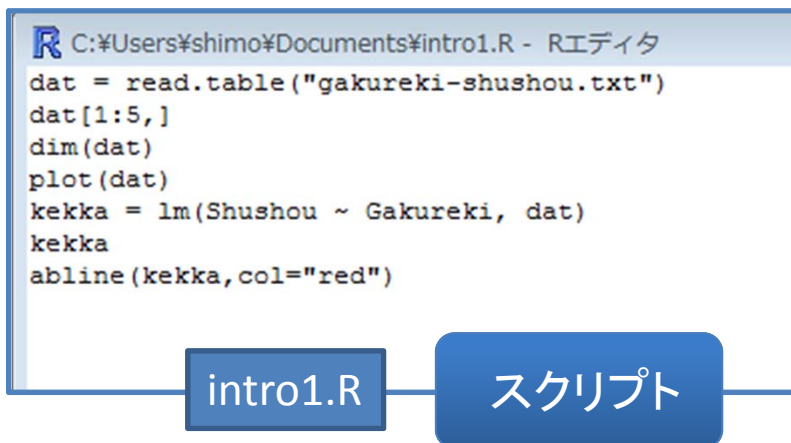
最小二乗法でパラメータを推定

Shushou

Gakureki

# スクリプトの作成, 保存, 実行

- ファイル=>新しいスクリプト
- Rエディタで作成
- ファイル=>保存
- 拡張子は . R
- ファイル=>スクリプトを開く
- 編集=>すべて実行

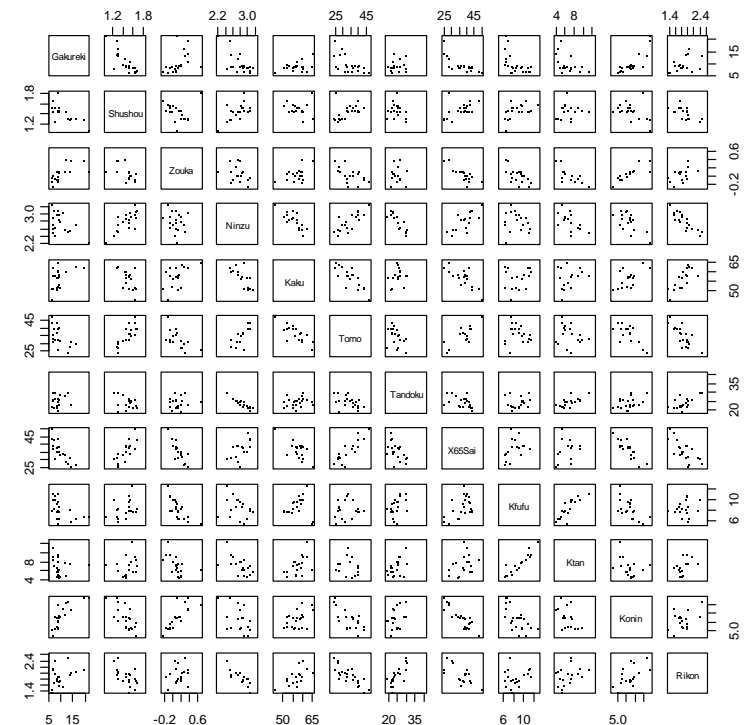


# 変数間の関係を調べるのは難しい

- “gakureki-rikon-12.txt” 47都道府県について12変量のデータ（学歴，出生率，増加率，世帯人数，など）

```
> dat2 <- read.table("gakureki-rikon-12.txt") # データの読み込み
> dim(dat2) # 行列の次元
[1] 47 12
> dat2[1:3,] # 最初の3行だけ表示
      Gakureki Shushou Zouka  Ninzu   Kaku   Tomo Tandoku X65Sai Kfufu Ktan
Hokkaido   7.7    1.23  0.04  2.42 60.54 26.54   29.95  30.50  9.90  7.39
Aomori     5.5    1.47 -0.02  2.86 54.20 34.38   24.08  38.99  7.45  6.61
Iwate      6.1    1.56 -0.07  2.92 50.87 38.82   24.47  42.42  7.87  6.05
      Konin Rikon
Hokkaido  5.77  2.40
Aomori    5.24  1.96
Iwate     5.14  1.48
> pairs(dat2,pch=".") # ペアごとの散布図
```

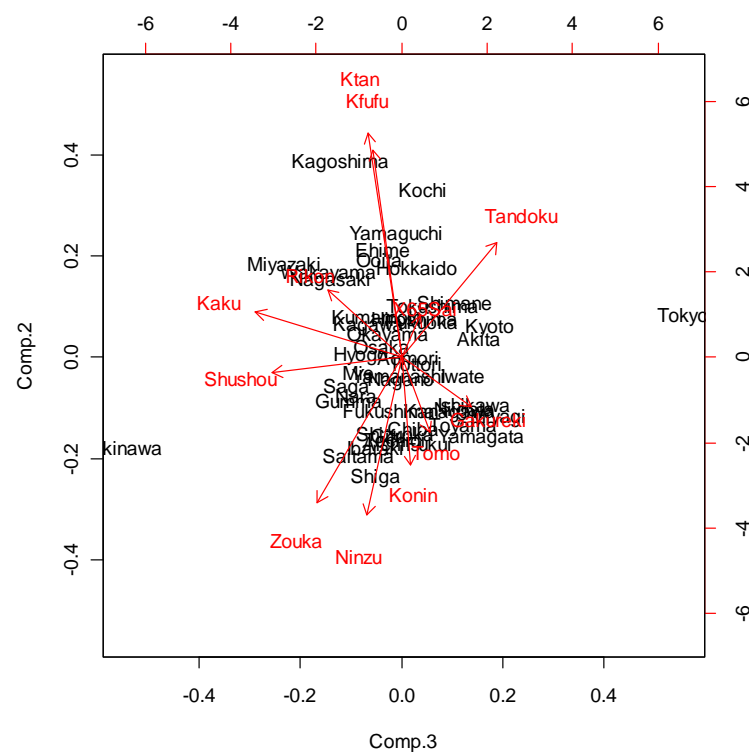
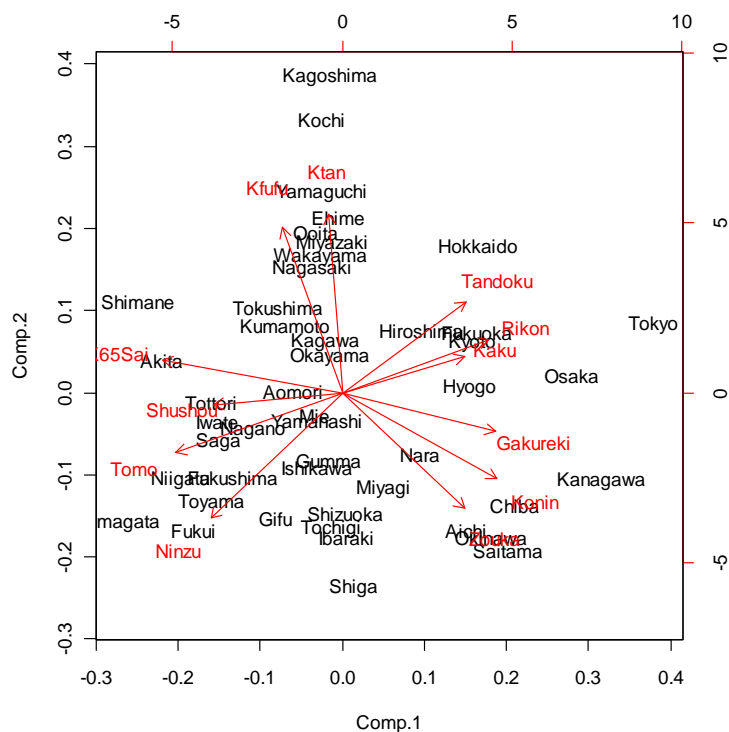
intro2.R



# 変量を合成してみる(主成分分析)

- princompとbiplotを使う

```
> kekka2 <- princomp(dat2,cor=T) # 主成分分析  
> biplot(kekka2) # バイプロット (第1、第2主成分)  
> biplot(kekka2,choi=c(3,2)) # バイプロット (第3、第2主成分)
```





# まとめ

- 「統計」ってなに？ 逆向き思考, 帰納推論
- 数学, 計算, データ
- 最近の研究動向, たくさん計算する
- メール [shimo-data2@is.titech.ac.jp](mailto:shimo-data2@is.titech.ac.jp)
- Rってなに？ インストール
- 「スクリプト」, 「実行結果」, 「プロットの図」
- 回帰分析(lm), 主成分分析(princomp)

## (参考) プロットのオプション

- `pch=2`, `pch="A"` などでマーク変更 (R plot `pch`などをウェブ検索！)
- `col=2`, `col="blue"` などで色変更 (R plot `col`でウェブ検索！)
- `lwd=2` などで線幅変更
- `lty=2` などで線種変更
- そのほかは`help(par)`で一覧できます
- プロットをワードにコピーしたりファイルに保存するには, プロットウィンドウで右クリックしてみる

# (参考) 社会・人口統計体系データ

- 総務庁統計局統計センターが公開
- <http://www.stat.go.jp/data/ssds/>
- 47 都道府県の様々な調査項目(2000 年度の 1182 項目)を下平がRで読み取れる形式に変換したものを自由に利用してください.
- X2000data.txt データ, X2000item.txt 変数名
- 詳細は2008年版講義資料の30ページ

# (参考) X2000の一部を取り出す方法

```
> ### データの読み込み
> X2000.data <- read.table("X2000data.txt") # X2000 データ本体です
> dim(X2000.data) # 行列の次元
[1] 47 1182
> X2000.data[1:5,500:503] # 一部だけ表示
      I1520301 I1520302 I1520401 I1520402
Hokkaido 6859.22 6898.59 121958 101431
Aomori 7064.20 7822.26 98232 93142
Iwate 7031.80 7857.32 104052 99401
Miyagi 6769.70 7170.04 97727 85584
Akita 7021.92 8142.30 94738 95532
> X2000.item <- read.table("X2000item.txt") # X2000 変数名, 単位, 全国平均, $
> dim(X2000.item) # 行列の次元
[1] 1182 4
> X2000.item[500:503,] # 一部だけ表示
      Imi      Tani      Zenkoku
I1520301 政府管掌健康保険受診率(被保険者千人当たり) 6831.72
I1520302 政府管掌健康保険受診率(被扶養者千人当たり) 6939.54
I1520401 政府管掌健康保険受診金額(被保険者1人当たり) (円:yen) 101121.00
I1520402 政府管掌健康保険受診金額(被扶養者1人当たり) (円:yen) 85331.00
      Bunrui
I1520301 J. 福祉・社会保障7)医療保険
I1520302 J. 福祉・社会保障7)医療保険
I1520401 J. 福祉・社会保障7)医療保険
I1520402 J. 福祉・社会保障7)医療保険
> ### 一部だけ取り出す方法
> dat <- X2000.data[,c("E09504", "A05203")] # 変数のID番号
> names(dat) <- c("Gakureki", "Shushou") # 分かりやすい名前をつけておく
> write.table(dat, "test.txt") # 表の書き出し
```

intro3.R

おまけ

# 参考文献？





2010/04/15

# 売れている順番@amazon.co.jp 「トレンドプロ マンガでわかる」

和書・「トレンドプロ マンガでわかる」	
検索結果21冊中1件から12冊までを表示 並び替え   売れている順番	
1.	<p>マンガでわかる統計学 高橋 信 トレンドプロ (単行本 - 2004/7)</p> <p>新品: ¥ 2,100</p> <p>5つ星 ¥ 2,100より 18つ星 ¥ 1,500より</p> <p>18時間以内に「お急ぎ便」でご注文いただくと、2010/4/16 金曜日までにお届けします。</p> <p>★★★★☆ (81) Amazonプライム</p>
2.	<p>マンガでわかる統計学 回帰分析編 高橋 信、井上 いろは、トレンドプロ (単行本 - 2005/9)</p> <p>新品: ¥ 2,310</p> <p>2つ星 ¥ 2,310より 14つ星 ¥ 1,746より</p> <p>通常2~5週以内にお届け</p> <p>★★★★☆ (8) Amazonプライム</p>
3.	<p>マンガでわかる統計学 因子分析編 高橋 信、井上 いろは、トレンド・プロ (単行本 - 2006/10/26)</p> <p>新品: ¥ 2,310</p> <p>2つ星 ¥ 2,310より 8つ星 ¥ 1,950より</p> <p>18時間以内に「お急ぎ便」でご注文いただくと、2010/4/16 金曜日までにお届けします。</p> <p>★★★★☆ (7) Amazonプライム</p>
4.	<p>マンガでわかるシーケンス制御 藤森 和弘、高山 ヤマ、トレンドプロ (単行本 - 2008/10)</p> <p>新品: ¥ 2,100</p> <p>2つ星 ¥ 2,100より 2つ星 ¥ 2,318より</p> <p>通常4~6日以内にお届け</p> <p>★★★★☆ (6) Amazonプライム</p>
5.	<p>マンガでわかる電気 藤森 和弘、マツダ、トレンドプロ (単行本 (ソフトカバー) - 2006/12)</p> <p>新品: ¥ 1,995</p> <p>2つ星 ¥ 1,995より 5つ星 ¥ 1,459より</p> <p>18時間以内に「お急ぎ便」でご注文いただくと、2010/4/16 金曜日までにお届けします。</p> <p>★★★★☆ (4) Amazonプライム</p> <p>後付の引用: "マンガでわかる統計学 マンガでわかるフリーエンジニア 基礎理論と実践" ... "</p>
6.	<p>マンガでわかるフリーエンジニア トレンドプロ、渋谷 道雄、晴瀬 ひろき (単行本 - 2006/3)</p> <p>新品: ¥ 2,520</p> <p>2つ星 ¥ 2,520より 8つ星 ¥ 1,599より</p> <p>18時間以内に「お急ぎ便」でご注文いただくと、2010/4/16 金曜日までにお届けします。</p> <p>★★★★☆ (12) Amazonプライム</p>
7.	<p>マンガでわかるデータベース 高橋 麻奈、あづま 蓮子、トレンドプロ (単行本 - 2005/12)</p> <p>新品: ¥ 1,995</p> <p>2つ星 ¥ 1,995より 7つ星 ¥ 1,647より</p> <p>18時間以内に「お急ぎ便」でご注文いただくと、2010/4/16 金曜日までにお届けします。</p> <p>★★★★☆ (7) Amazonプライム</p>
8.	<p>マンガでわかる物理 力学編 新田 英雄、高津 ケイタ、トレンドプロ (単行本 - 2006/11)</p> <p>新品: ¥ 2,100</p> <p>2つ星 ¥ 2,100より 8つ星 ¥ 980より</p> <p>18時間以内に「お急ぎ便」でご注文いただくと、2010/4/16 金曜日までにお届けします。</p> <p>★★★★☆ (6) Amazonプライム</p>
9.	<p>マンガでわかる線形代数 高橋 信、井上 いろは、トレンドプロ (単行本 - 2008/11)</p> <p>新品: ¥ 2,100</p> <p>2つ星 ¥ 2,100より 3つ星 ¥ 1,487より</p> <p>18時間以内に「お急ぎ便」でご注文いただくと、2010/4/16 金曜日までにお届けします。</p> <p>★★★★☆ (1) Amazonプライム</p>

21冊中



1位



2位



3位



「統計学」は大人気じゃないか！