MEIJI UNIVERSITY

先端メディアゼミナールⅡ

菊池浩明

1章　データマイニングとは

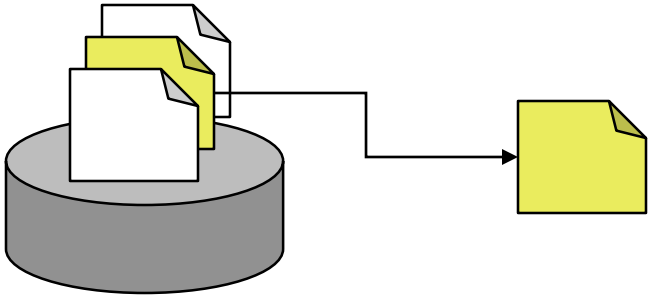
春

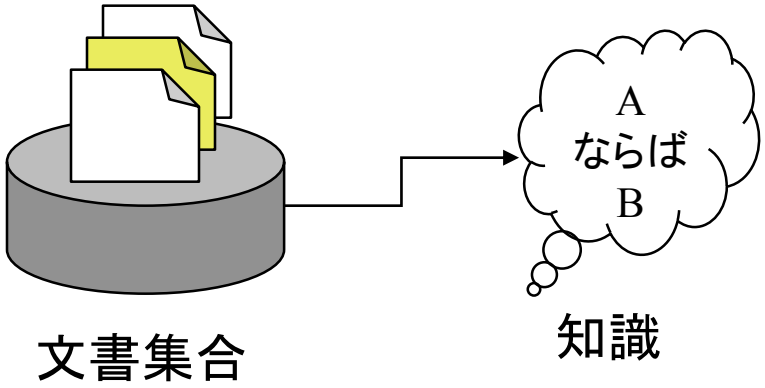
FMSセミナーⅡ（2年次春学期）

- データマイニング入門（2012年春）
  - 豊田秀樹（心理学者）
  - 「輪講」（担当者が調べて発表．質疑応答）
  - 身近なデータで例題が豊富
  - 統計解析 オープンソースR

データマイニングと情報検索

- 情報検索
  - Information Retrieval
  - 文書集合から適切な文書を探すこと
- データマイニング
  - Data mining
  - 大規模なデータから知識（パターン）を見つけること



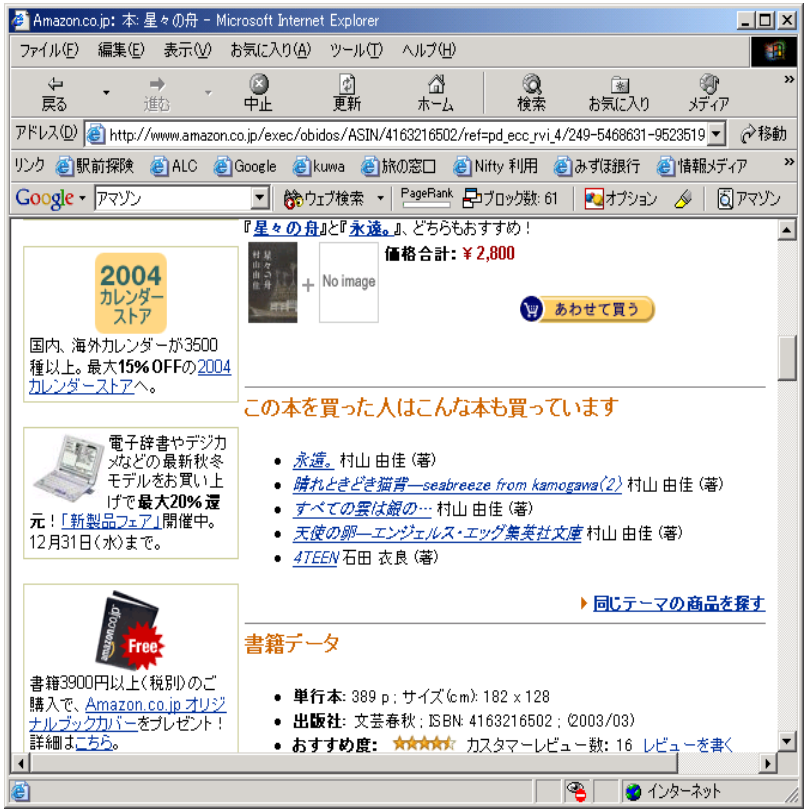


「ビッグデータ」

- Big Data
  - 巨大なデータを扱う課題，および，そのための技術の総称
  - Volume (テラ，ペタの量)，Variety (多様化，非構造化)，Velocity (頻度，リアルタイムデータ)
- 背景
  - 電子化の推進：スマートフォン，RFIDの普及，M2M
  - 消費者行動の予測やマーケティングへの応用の期待，医療分野への応用，個人情報保護

Amazon.co.jp

- 検索「**星々の舟**」
  こんな本も買っています
  - *永遠。* 村山 由佳 (著)
  - *晴れときどき猫背 —seabreeze from kamogawa〈2〉* 村山 由佳 (著)
  - *すべての雲は銀の…* 村山 由佳 (著)
  - *天使の卵—エンジェルス・エッグ集英社文庫* 村山 由佳 (著)
  - *4TEEN* 石田 衣良 (著)



「コンビニ購買記録」

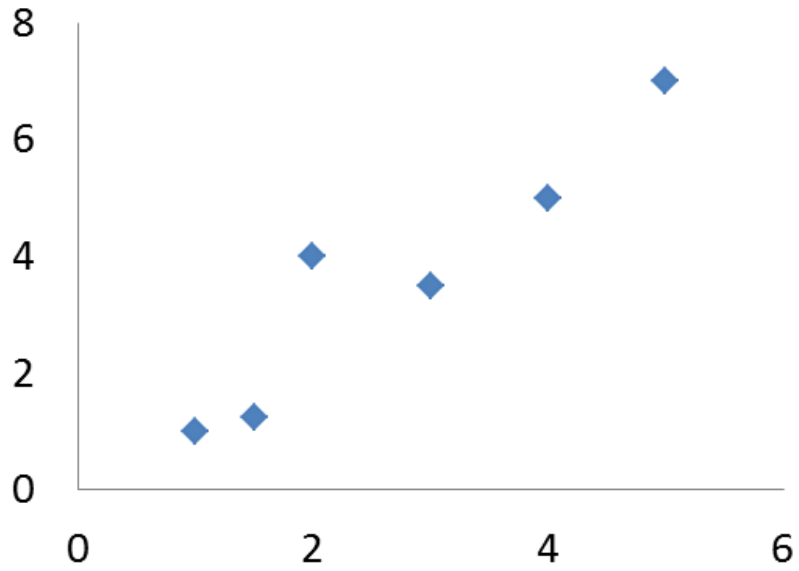
	牛乳	卵	ビール	おむつ	新聞
2003/12/1 10:00	1			1	
2003/12/1 10:21	1	1	1		
2003/12/1 11:04			1		1
2003/12/1 12:30				1	
2003/12/1 14:01		1			1
2003/12/1 15:42				1	1
2003/12/2 11:01	1	1		1	
2003/12/2 12:20	1	1			

問題：項目間の相関規則

- 相関ルール
  - 粉ミルクを買う人はビールをよく買う
  - おにぎりとお茶を買う人は肉まんを買う
  - セキュリティの本を買う人はミステリーをよく買う
  - 膨大なデータベースから意味のある（意外性のある）相関ルール（知識）をどのように抽出すればよいか

問題

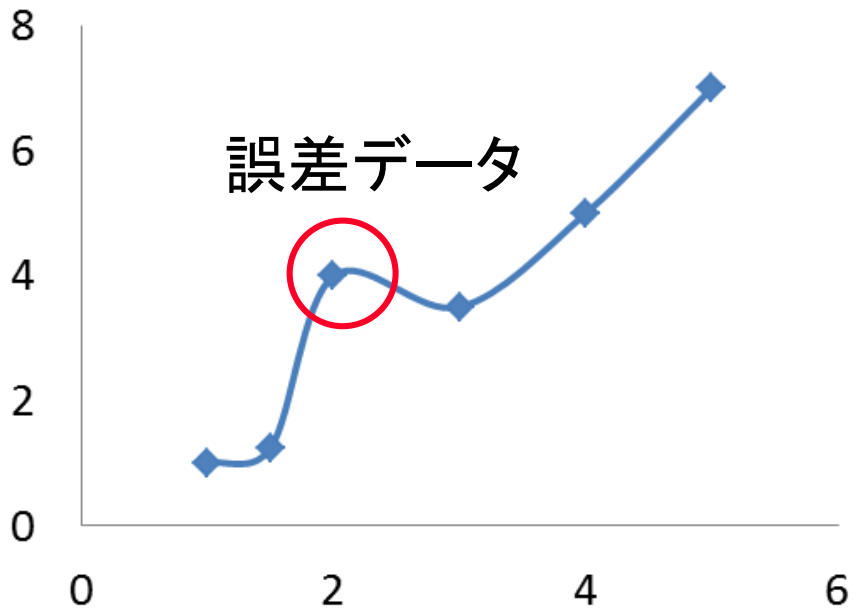
- このデータをどの様に表したらよいか？



x	y
1	1
1.5	1.25
2	4
3	3.5
4	5
5	7

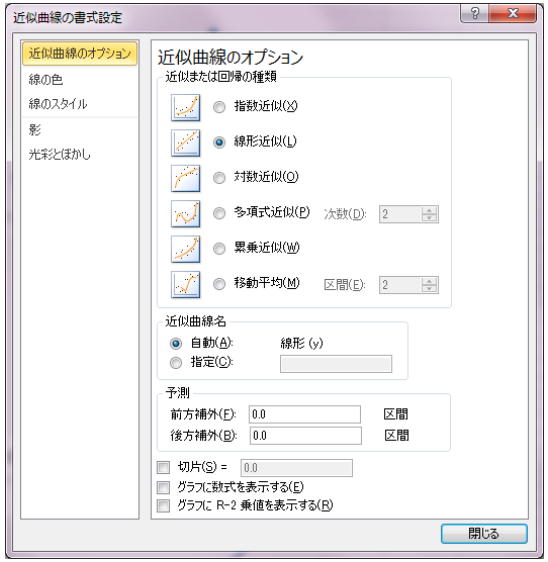
悪い例

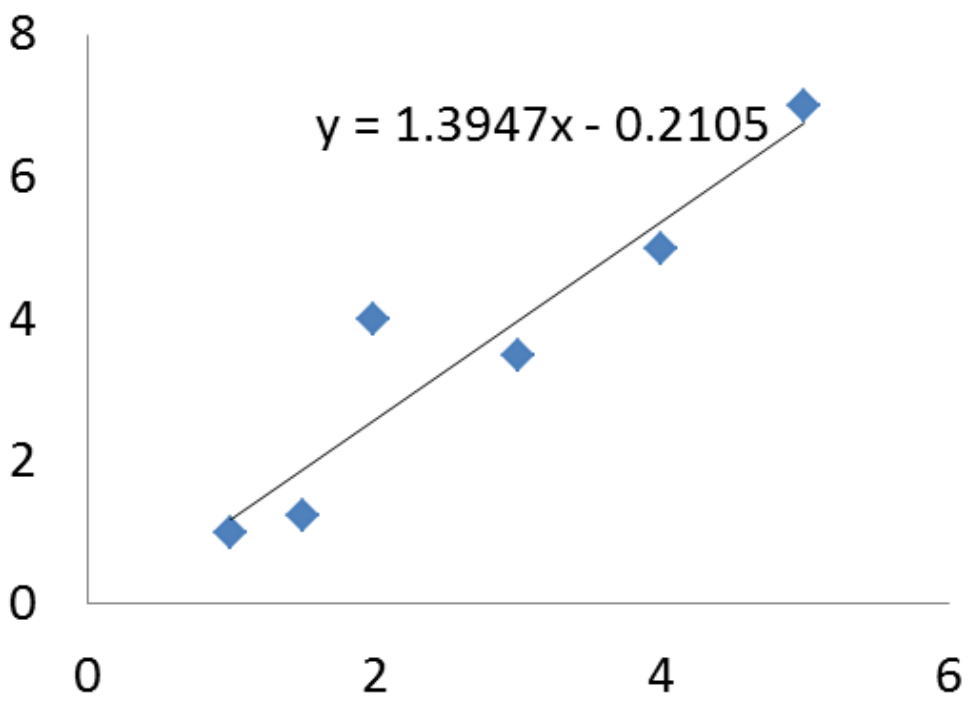
- 過学習 (over fitting)



線形近似

- 最小二乗法
  - 「近似曲線の追加」

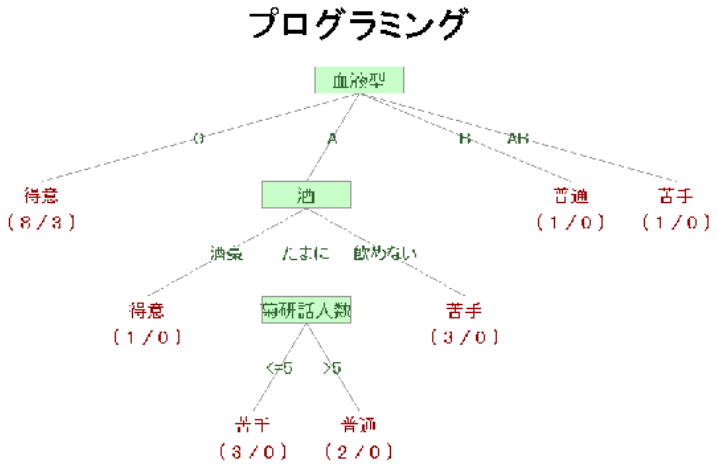


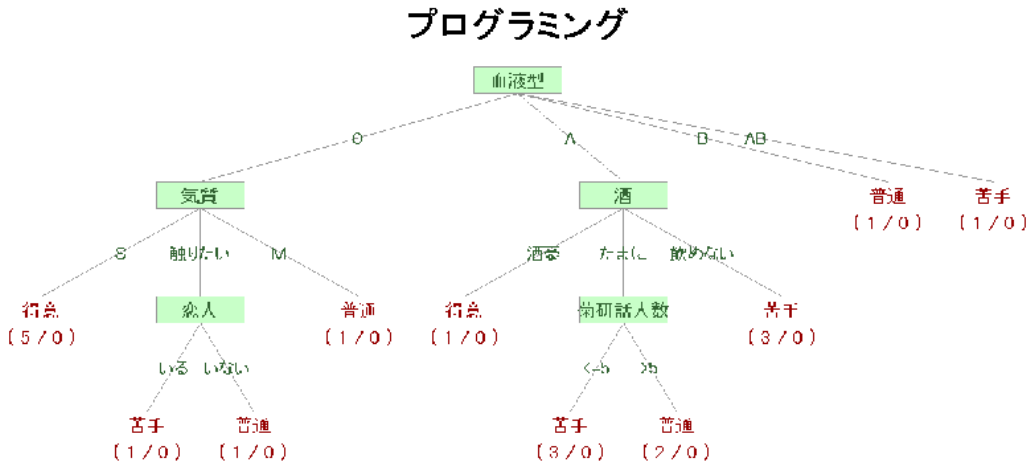


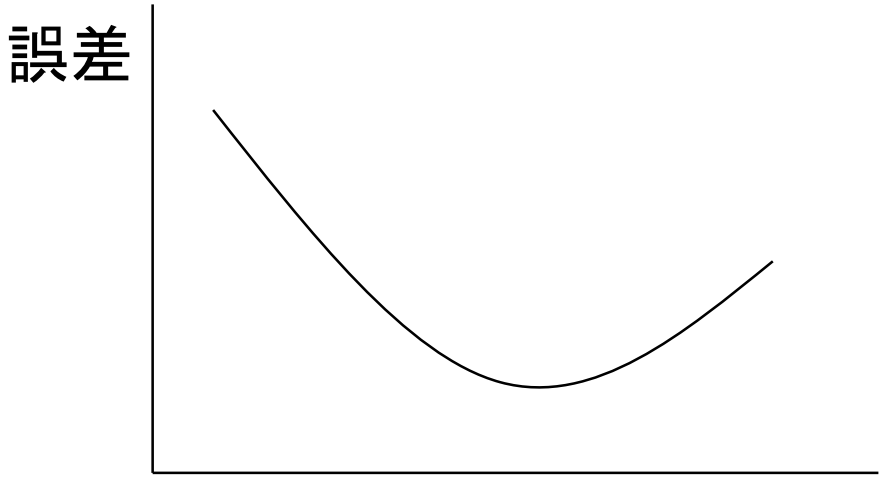
最小二乗法

- アルゴリズム
  - 入力: (x1,y1),…,(xn, yn)
  - あてはめ多項式: $f(x) = a+bx$
- 誤差の総和 $S = \sum(f(xi) - yi)^2$
  - $dS/da = 0$
  - $dS/db = 0$ の連立方程式を解く.

決定木の過学習

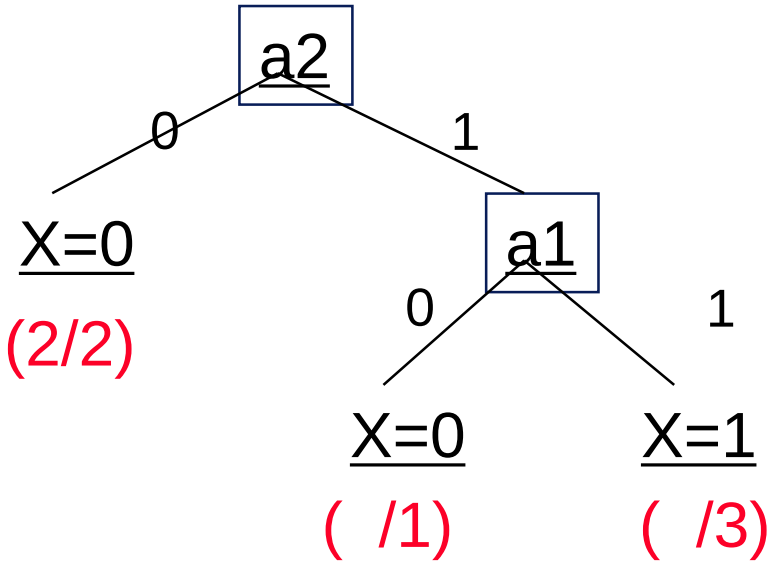

プログラミング
血液型
O
A
B
AB
得意
（8／3）
酒
普通
（1／0）
苦手
（1／0）
酒豪
たまに
飲めない
得意
（1／0）
麻雀人数
苦手
（3／0）
<=5
>5
苦手
（3／0）
普通
（2／0）


プログラミング
血液型
O
A
B
AB
気質
酒
普通
（1／0）
苦手
（1／0）
S
触りたい
M
得点
（5／0）
恋人
普通
（1／0）
酒豪
たまに
飲めない
得点
（1／0）
麻雀人数
苦手
（3／0）
いる
いない
<=5
>5
苦手
（1／0）
普通
（1／0）
苦手
（3／0）
普通
（2／0）


誤差
木のサイズ

誤差の定義

- 決定木　T
  - (正解数/分類列数)

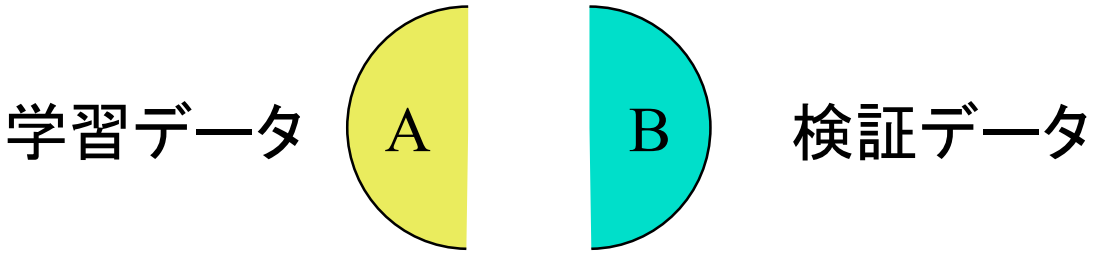


a1	a2	a3	X	T
1	1	1	1	1
1	0	0	0	0
0	1	0	0	0
1	1	0	0	1
0	0	1	0	0
1	1	0	1	1

エラー率 E(T) = 誤り識別数/N = 1/6

クロスバリデーション

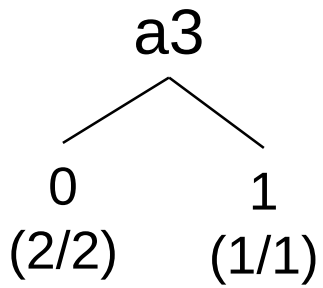
- 予測精度



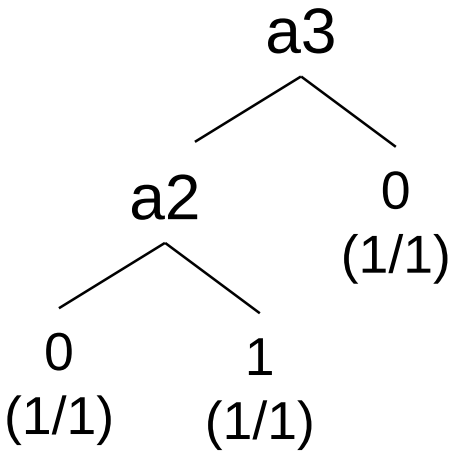
Aで学習	Bで検証	誤差率 $E_1(T_A)$
Aで検証	Bで学習	誤差率 $E_2(T_B)$
		誤差率 $E_1(T_A)+E_2(T_B)$

例）

- A



- B



a1	a2	a3	X	T
1	1	1	1	1
1	0	0	0	0
0	1	0	0	0

1	0	0	0	1
0	1	1	0	0
1	1	0	1	1

Nフォールドクロスバリデーション

- 手続き
  - 1. 大きさの等しいN個にランダムに分割
  - 2. N-1個を学習データにして決定木を作る
  - 3. 残りの1個を検証データとして，2の精度を求める.
  - 4. 2と3をN回繰り返し，平均を求める

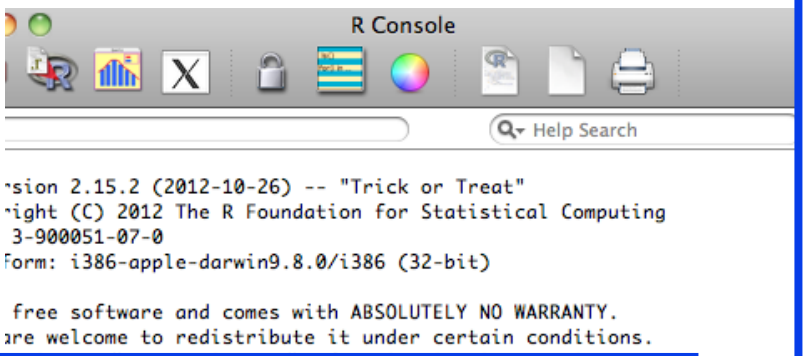
1.7 Rことはじめ

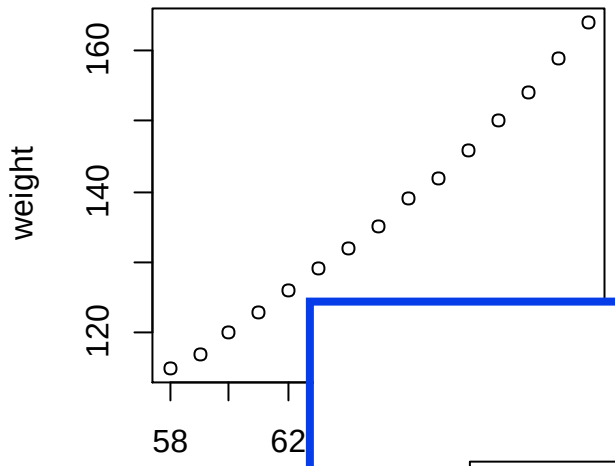
R言語のインストール

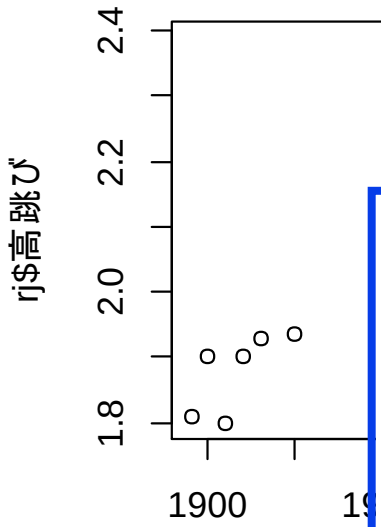
フレームの操作

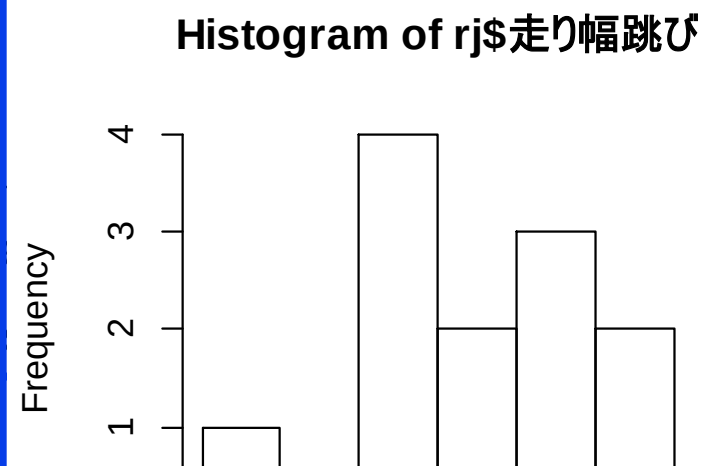
基本統計量

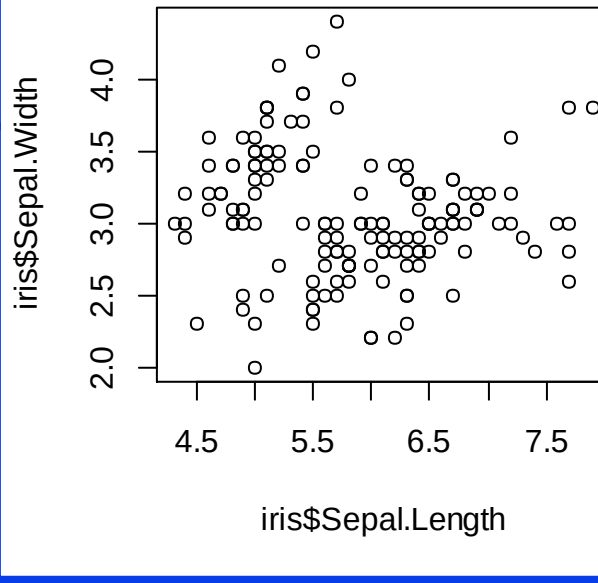
統計ツール R


R Console
Help Search
rsion 2.15.2 (2012-10-26) -- "Trick or Treat"
right (C) 2012 The R Foundation for Statistical Computing
3-900051-07-0
Form: i386-apple-darwin9.8.0/i386 (32-bit)
free software and comes with ABSOLUTELY NO WARRANTY.
are welcome to redistribute it under certain conditions.


weight
160
140
120
58
62

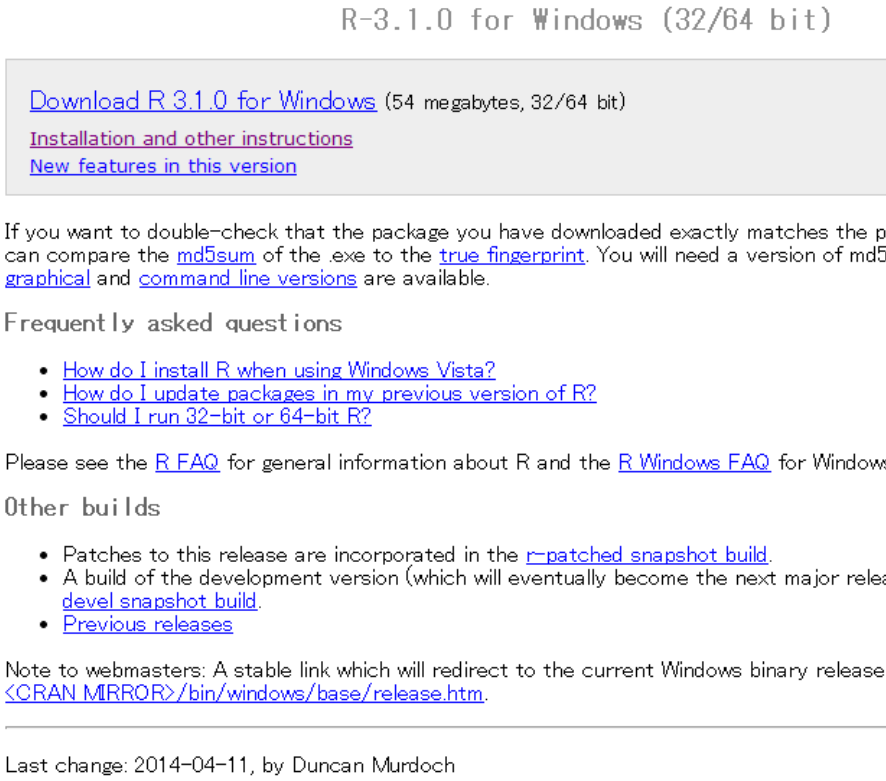

rj$高跳び
2.4
2.2
2.0
1.8
1900


Histogram of rj$走り幅跳び
Frequency
4
3
2
1


iris$Sepal.Width
4.0
3.5
3.0
2.5
2.0
4.5
5.5
6.5
7.5
iris$Sepal.Length

Rのインストール

- R 3.1.0
- Windows
  - http://cran.md.tsukuba.ac.jp/bin/windows/base/
- Mac
  - http://cran.md.tsukuba.ac.jp/bin/macosx/



R-3.1.0 for Windows (32/64 bit)

Download R 3.1.0 for Windows (54 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded exactly matches the p
can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5
graphical and command line versions are available.

Frequently asked questions

- How do I install R when using Windows Vista?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows

Other builds

- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major relea
devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release
<CRAN MIRROR>/bin/windows/base/release.htm.

Last change: 2014-04-11, by Duncan Murdoch

Rコンソール

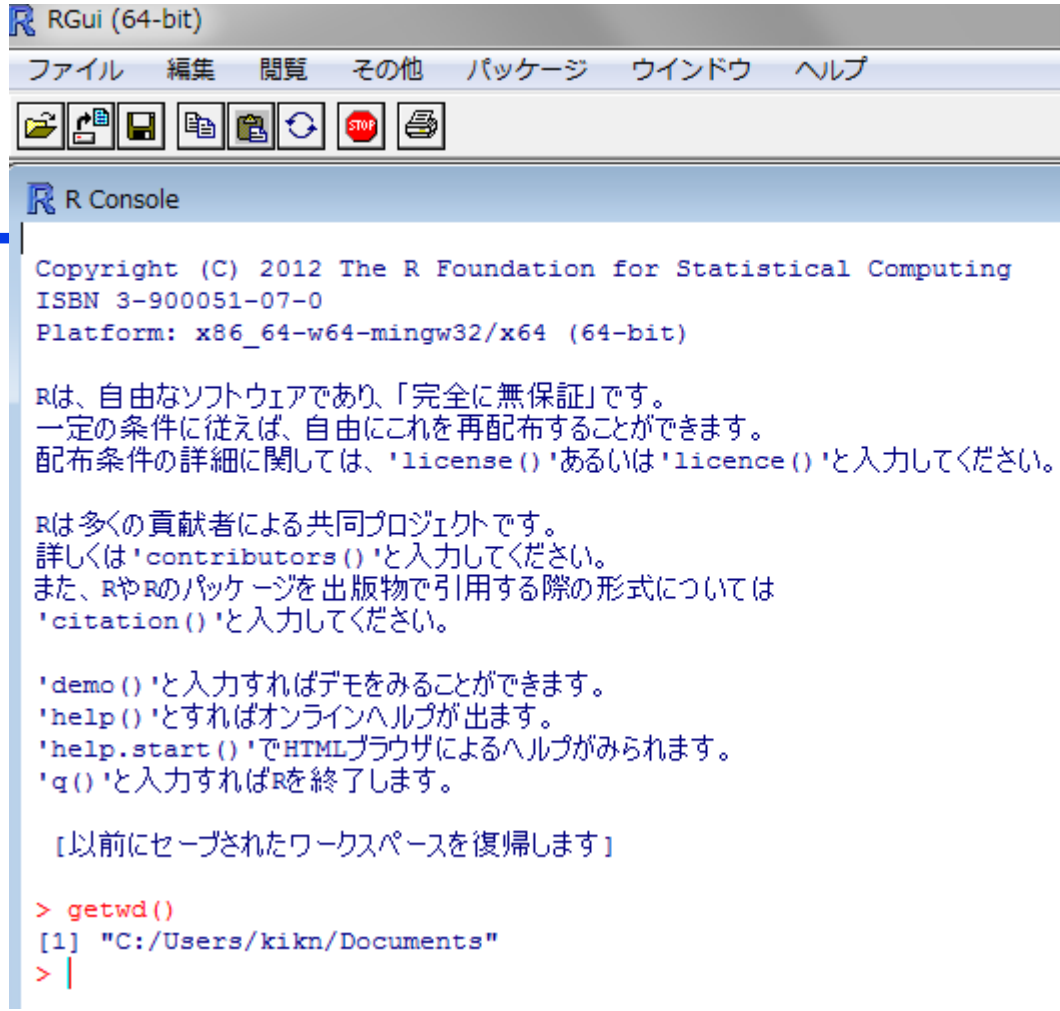
- コマンドベース
  - プロンプト

    >
  - 現在のパス

    > getwd()

    [1] "C:/Users/kikn/Documents“
  - 変更

    > setwd("../Dropbox/Share/DL02045/Program/chap1/")



データの読み方

- サンプルデータ
  - Share/DL02045/Program/chap1/陸上.csv
  - CSVファイル（カンマで区切られたテキスト）
- 読み込み

```
> data=read.csv("陸上データ.csv“)
> data
  高跳び X100m走 走り幅跳び 開
1  1.810   12.00     6.350  189
2  1.900   11.00     7.185  190
3   1.800   11.00     7.340  19
```

	A	B	C	D	E
1	高跳び	100m走	走り幅跳び	開催年	区分
2	1.81	12	6.35	1896	戦前
3	1.9	11	7.185	1900	戦前
4	1.8	11	7.34	1904	戦前
5	1.9	10.8	7.48	1908	戦前
6	1.93	10.8	7.6	1912	戦前
7	1.936	10.6	7.15	1920	戦前
8	2.12	10.5	7.83	1956	戦後
9	2.18	10	8.03	1964	戦後
10	2.38	9.92	8.72	1988	戦後
11	2.34	9.96	8.67	1992	戦後
12	2.39	9.84	8.5	1996	戦後
13	2.36	9.85	8.31	2004	戦後

データフレームの操作

- 行と列の取出し

```
> data$開催年                    列名での列の取出し
 [1] 1896 1900 1904 1908 1912 1920 1956 1964 1988 1992 1996 2004
> data[2,]                       2行目の取出し
  高跳び X100m走 走り幅跳び 開催年 区分
2    1.9      11      7.185   1900 戦前
> data[,2]                       2列目の取出し
 [1] 12.00 11.00 11.00 10.80 10.80 10.60 10.50 10.00  9.92  9.96  9.84
9.85
> data[2:4,]                     2行目から4行目まで取出し
  高跳び X100m走 走り幅跳び 開催年 区分
2   1.90    11.0      7.185   1900 戦前
3   1.80    11.0      7.340   1904 戦前
4   1.90    10.8      7.480   1908 戦前
```

データの加工

- 値の置換

```
> data[2,1]
[1] 1.9
> data[2,1]=2.5
> data[2,1]
[1] 2.5
```

- データフレームの書き出し

```
>  write.csv(data,"a.csv")
```

データフレーム

- データフレームの構造 (structure)

> **str**(data)

‘data.frame’:　12 obs. of　5 variables:　　12行5列

＄高跳び　: num　1.81 2.5 1.8 1.9 1.93 ...　　実数値変数

＄X100m走　: num　12 11 11 10.8 10.8 10.6 10.5 10 9.92 9.96 ...

＄走り幅跳び: num　6.35 7.18 7.34 7.48 7.6 ...

＄開催年　: int　1896 1900 1904 1908 1912 1920 1956 1964 1988 1992 ...　　整数値変数

＄区分　: Factor w/ 2 levels “戦後”,“戦前”: 2 2 2 2 2 2 1 1 1 1 ...　　名義的変数（2値）

集計

基本統計量

```
> summary(data)
     高跳び          X100m走          走り幅跳び          開催年          区分
 Min.   :1.800   Min.   : 9.84   Min.   :6.350   Min.   :1896   戦後:
 1st Qu.:1.900   1st Qu.: 9.95   1st Qu.:7.301   1st Qu.:1907   戦前:
 Median :2.028   Median :10.55   Median :7.715   Median :1938
 Mean   :2.087   Mean   :10.52   Mean   :7.764   Mean   :1945
 3rd Qu.:2.345   3rd Qu.:10.85   3rd Qu.:8.357   3rd Qu.:1989
 Max.   :2.390   Max.   :12.00   Max.   :8.720   Max.   :2004
```

- Median = 中央値, Mean = 平均値
- 問） 次のデータのメジアンと平均を求めよ. 1，1，1，2，8，8，8

集計応用編

- 区分ごとの集計

```
> by(data, data$区分, summary)
data$区分: 戦後
     高跳び          X100m走          走り幅跳び        開催年        区分
 Min.   :2.120   Min.   : 9.840   Min.   :7.830   Min.   :1956   戦後:6
 1st Qu.:2.220   1st Qu.: 9.867   1st Qu.:8.100   1st Qu.:1970   戦前:0
 Median :2.350   Median : 9.940   Median :8.405   Median :1990
 Mean   :2.295   Mean   :10.012   Mean   :8.343   Mean   :1983
 3rd Qu.:2.375   3rd Qu.: 9.990   3rd Qu.:8.627   3rd Qu.:1995
 Max.   :2.390   Max.   :10.500   Max.   :8.720   Max.   :2004
------------------------------------------------------------
data$区分: 戦前
     高跳び          X100m走          走り幅跳び        開催年        区分
 Min.   :1.800   Min.   :10.60   Min.   :6.350   Min.   :1896   戦後:0
 1st Qu.:1.833   1st Qu.:10.80   1st Qu.:7.159   1st Qu.:1901   戦前:6
 Median :1.900   Median :10.90   Median :7.263   Median :1906
 Mean   :1.879   Mean   :11.03   Mean   :7.184   Mean   :1907
 3rd Qu.:1.923   3rd Qu.:11.00   3rd Qu.:7.445   3rd Qu.:1911
 Max.   :1.936   Max.   :12.00   Max.   :7.600   Max.   :1920
```

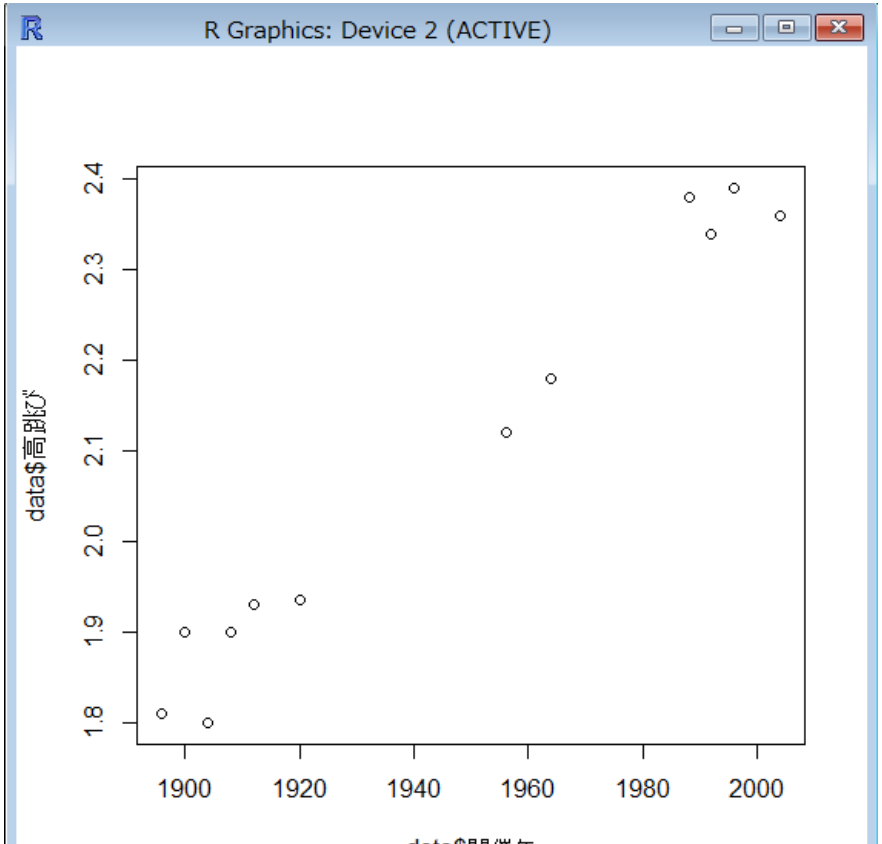
  - By(データフレーム，区分インデックス，適用する関数)

演習1.

- Rをインストールして，陸上データ.csvを読み込み.
  - 高跳びの平均値を求めよ.
  - 100m走が10秒代の時代における，高跳びの平均値を求めよ.
  - 次のデータのメジアンと平均を求めよ．1，1，1，2，8，8，8

散布図

- plot(X座標ベクトル，Y座標ベクトル)
  - plot(data$開催年, data$高跳び)



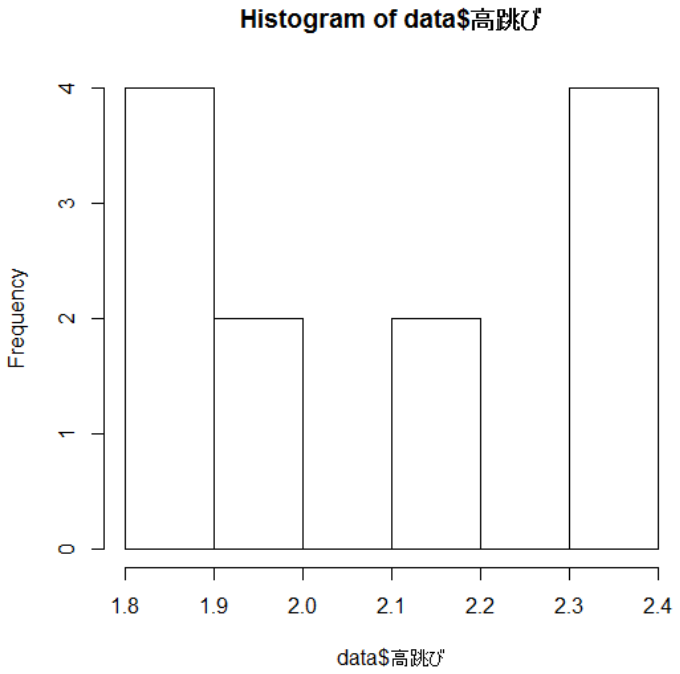
ヒストグラム

- ヒストグラム描画（度数分布図）

> data$高跳び

[1] 1.810 1.900 1.800 1.900 1.930 1.936 2.120 2.180 2.380 2.340 2.390 2.360

>**hist**(data$高跳び)

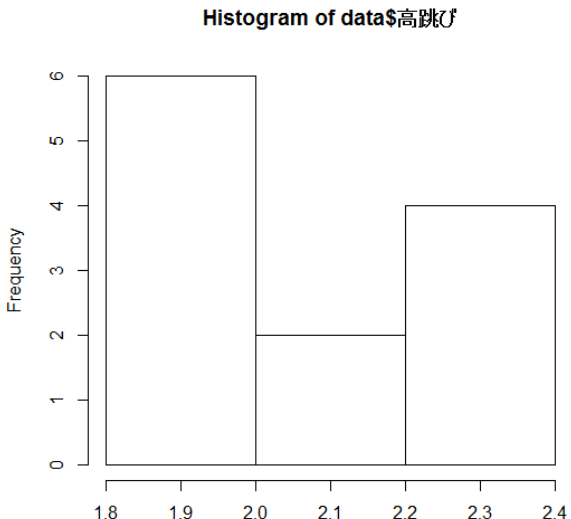


もっと凝ったことがしたい

- help(関数)

  >help(hist)

  >hist(data$高跳び, breaks=3)



127.0.0.1:31797/library/graphics/html/hist.html

アプリ 明治大学 Gmail ~kikn Google

これは 英語 のページです。翻訳しますか？ 翻訳 いいえ 英語を翻訳しな

hist {graphics}

Histograms

Description

The generic function hist computes a histogram of the given data values. If plo
plotted by plot.histogram, before it is returned.

Usage

```
hist(x, ...)

## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = NULL, border = NULL,
     main = paste("Histogram of " , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, ...)
```

Arguments

x	a vector of values for which the histogram is desired.
breaks	one of:

- a vector giving the breakpoints between histogram cells,
- a function to compute the vector of breakpoints,

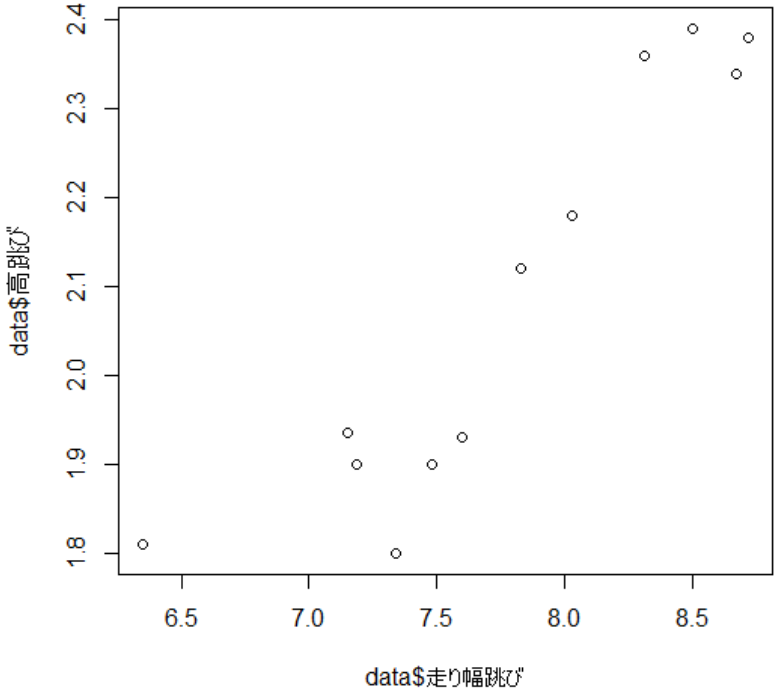
R-3.1.0-win.exe

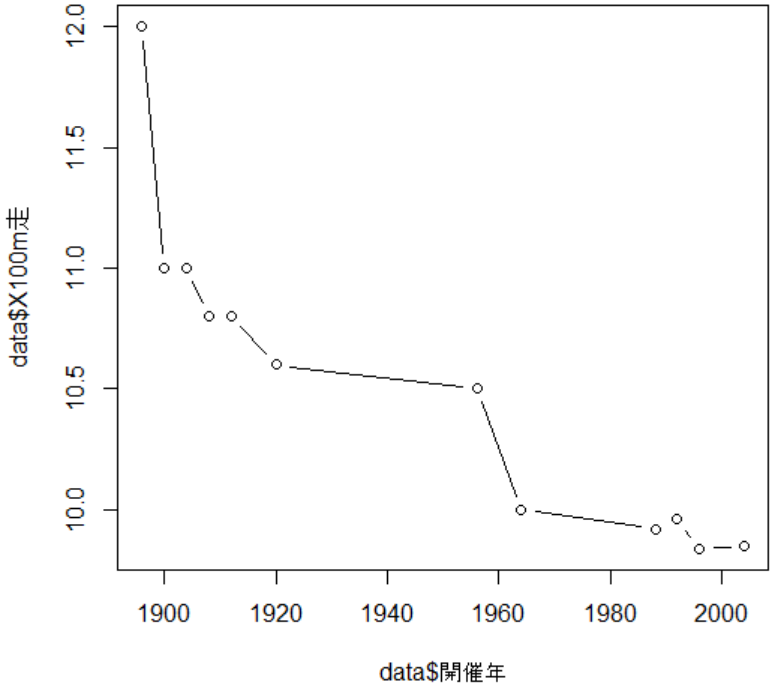
もっと色々なデータを試したい

- 組み込みデータフレーム
  - data()　一覧表示
  - AirPassengers　Monthly Airline Passenger Numbers 1949-1960
  - BJsales　Sales Data with Leading Indicator BJsales.
  - BOD　Biochemical Oxygen Demand
  - CO2　Carbon Dioxide Uptake in Grass Plants
  - ChickWeight　Weight versus age of chicks on different diets
  - iris　あやめの種類

演習2.

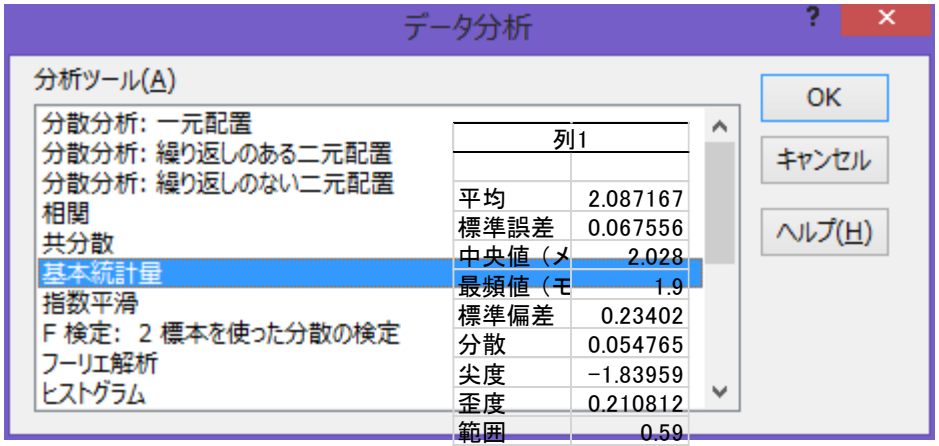
- 次のグラフを求めよ.





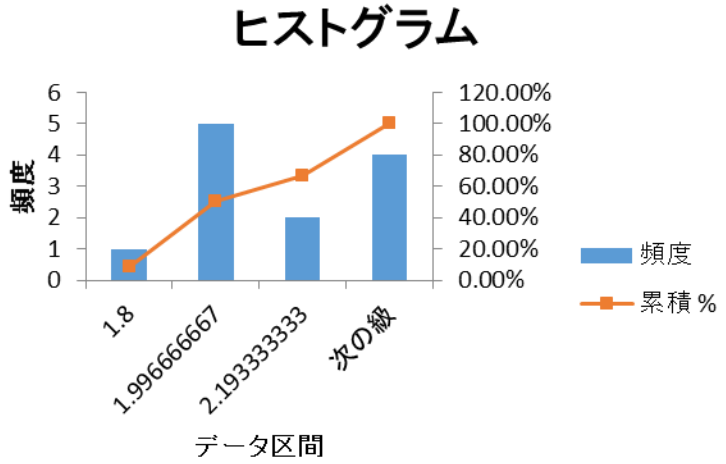
参考) Excelで同じことを行う

- 分析ツール
  - ファイル>オプション>アドイン>分析ツール
  - データ>分析>分析ツール>基本統計量



	列1
平均	2.087167
標準誤差	0.067556
中央値（メ	2.028
最頻値（モ	1.9
標準偏差	0.23402
分散	0.054765
尖度	-1.83959
歪度	0.210812
範囲	0.59
最小	1.8
最大	2.39
合計	25.046
標本数	12
最大値(1)	2.39

- ヒストグラム
  - データ>分析ツール>ヒストグラム

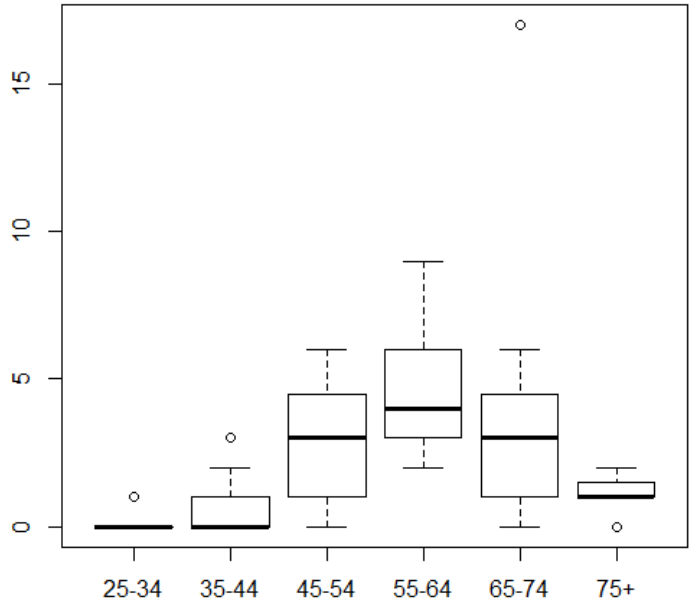


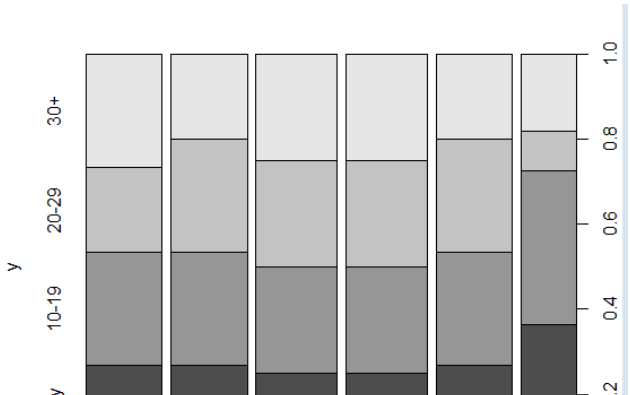
まとめ

- データマイニングとは，多量のデータを分析する有効な手法の集合体.
- 統計学に密接に関係しており，多学習や交差妥当化などの課題や評価法がある.
- Rはデータマイニングにも使えるオープンソース. テキストベースでコマンドを入力して操作する. 行と列を持つデータフレームを基本構造として,
- 基本統計量 (summary, by)，可視化 (drow, hist)

宿題

- 組み込みデータフレームから食道がん患者と喫煙，飲酒の関係のデータを探し，
  - (1) 右のグラフを求めよ.
  - (2) 煙草を日に9本までしか吸わない人と30本以上吸う人のがん患者数を求めよ.





感想

- R言語はテキストベースで使いやすい．ヒストリー機能などを効果的に使える．
- データフレームは，まとまった処理をするのに適している．Excelとの相性もよい．
- 多機能過ぎて，マニュアルを見ても適切な機能を探すのが大変．マニュアルを引くためのコマンドを知っておく必要があり．
- Drawのグラフがデータに応じて勝手に変わる仕組みがよく分からない．

担当者の仕事

- テキストを読み，理解する
  - ウェブに頼らない．自分の理解した内容を自分の言葉で説明．
- 内容に沿った演習
  - テキストに沿った内容．2，3回，ゼミ中に行う
- まとめと感想
- 宿題
  - 2題程度，学んだ内容を復習

参考文献

- 豊田「データマイニング入門」東京出版
  - 貸出用数冊．持ち帰り禁止
- 金「Rによるデータサイエンス」，森北出版
- 他に買ってほしい本があれば購入します