無料で有意差を出す方法

～ 統計解析フリーソフト R の紹介 ～

本日のメニュー

- 統計解析と統計解析ソフト
- R とは？
- R と検定
- R とグラフィックス
- R とシミュレーション
- まとめ

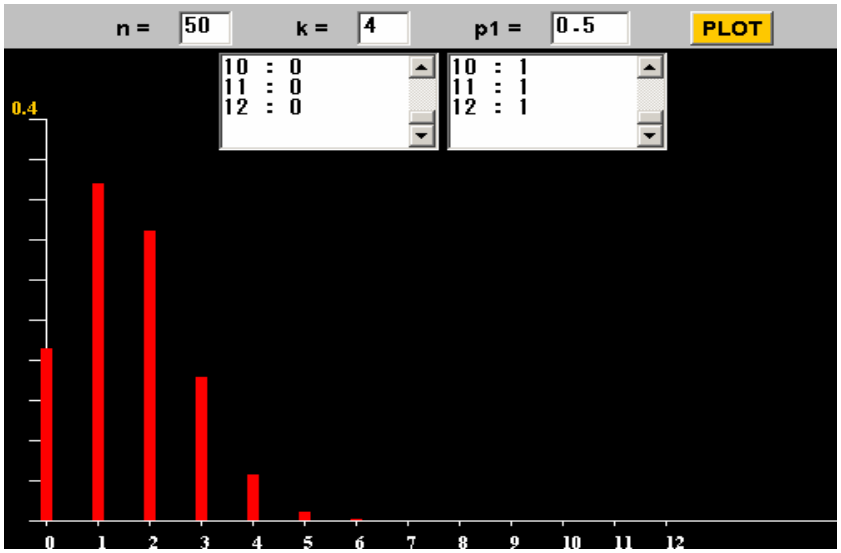
統計解析と統計解析用ソフト

統計解析と統計解析用ソフト

- 学校で統計を勉強するために必要なもの
  - 教科書や参考書
  - データを視覚的に見るためのグラフィックツール
  - 教科書に載っている計算手順を実際に計算するための計算ツール
- 会社でデータ解析を行うために必要なもの
  - データに関する資料
  - データを視覚的に見るためのグラフィックツール
  - データの要約を行うための計算ツール

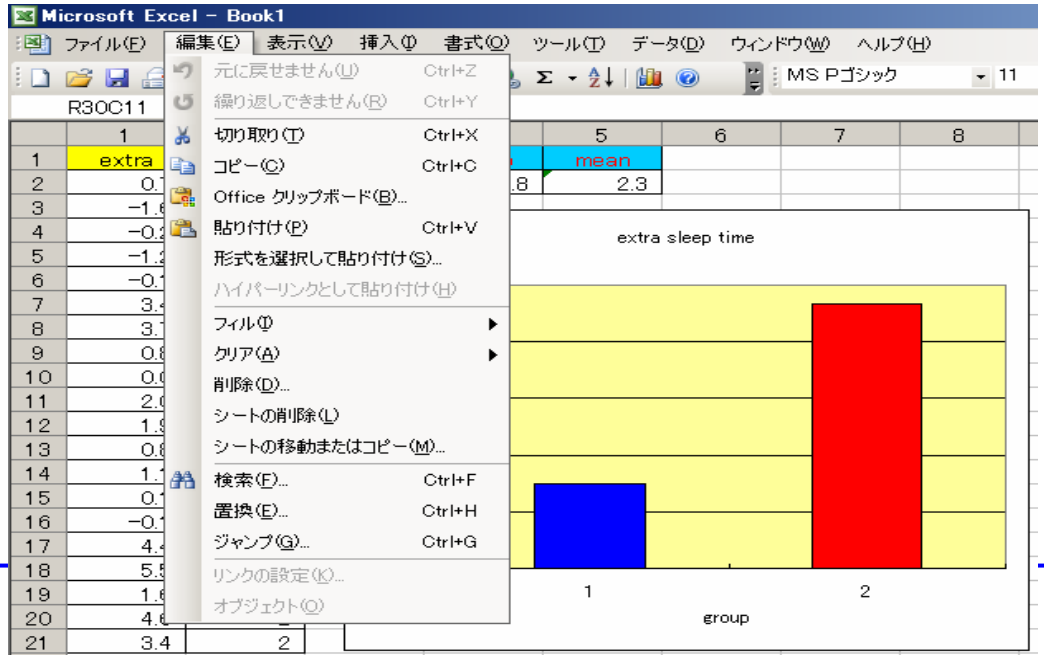
統計解析と統計解析用ソフト

- プログラム言語：C 言語や JAVA など
  - 計算速度が速い
  - プログラム作成が面倒（作成の手間，バグ取り）



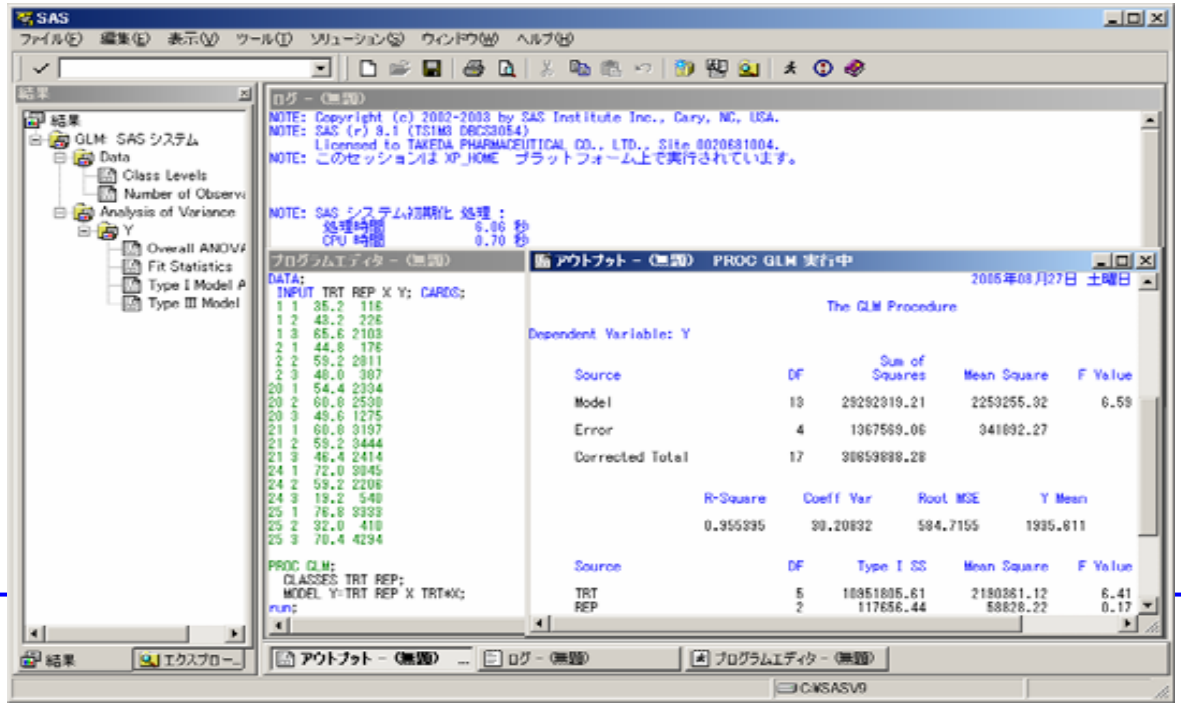
統計解析と統計解析用ソフト

- 表計算ソフト：EXCEL など
  - データを GUI 上で操作することが可能
  - 手軽にグラフィックス表示をすることが可能
  - データ加工やプログラム作成には不向き



統計解析と統計解析用ソフト

- 統計解析ソフト：SAS，SPSS など
  - データ操作に長け，グラフィックス表示可能，プログラム作成が出来る
  - 非常に高価（個人で買える値段ではない）



R とは？

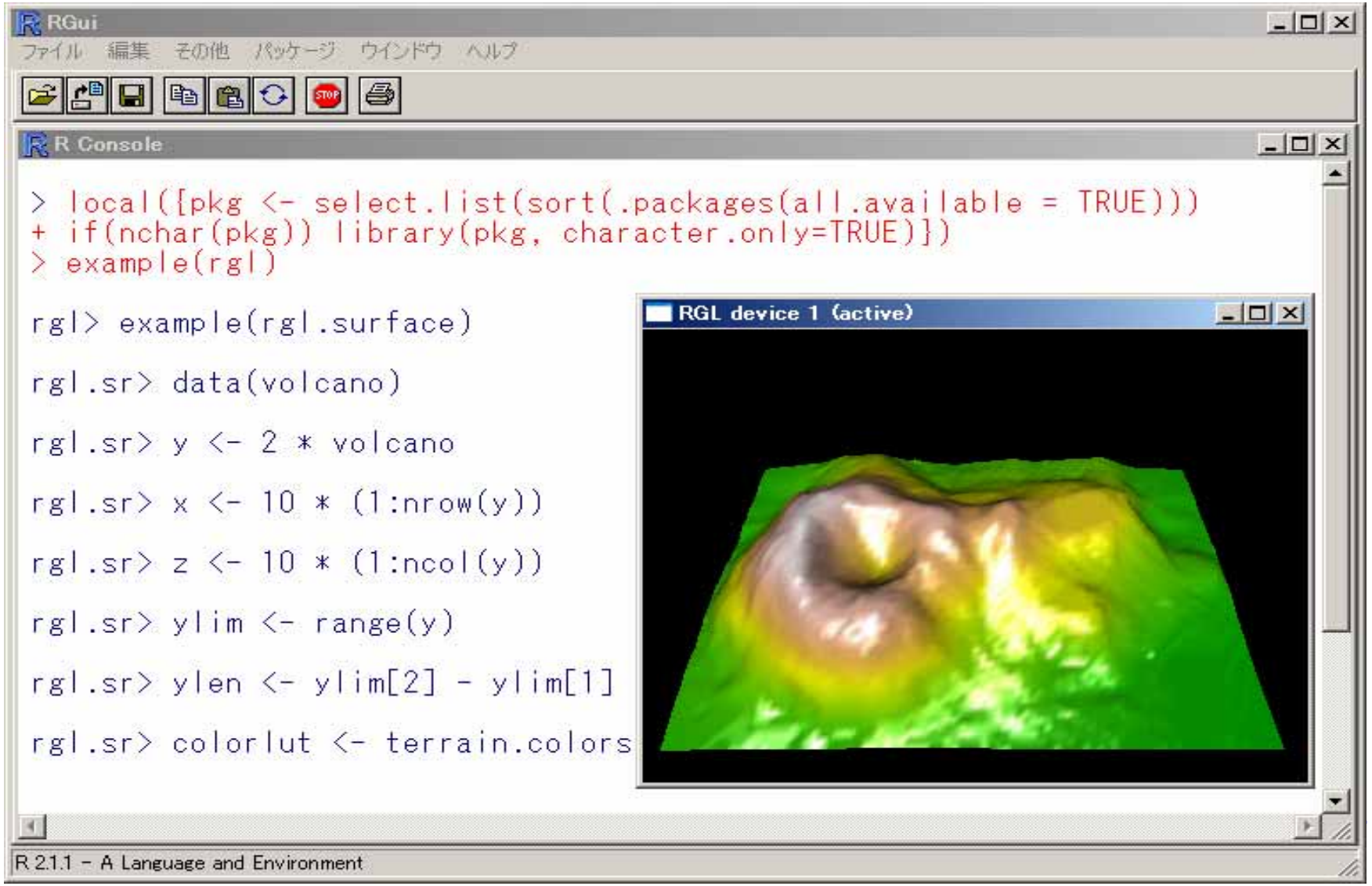
R とは？

- フリーの統計解析用ソフトウェア
- ベースは S システム（後の S言語，S-PLUS）

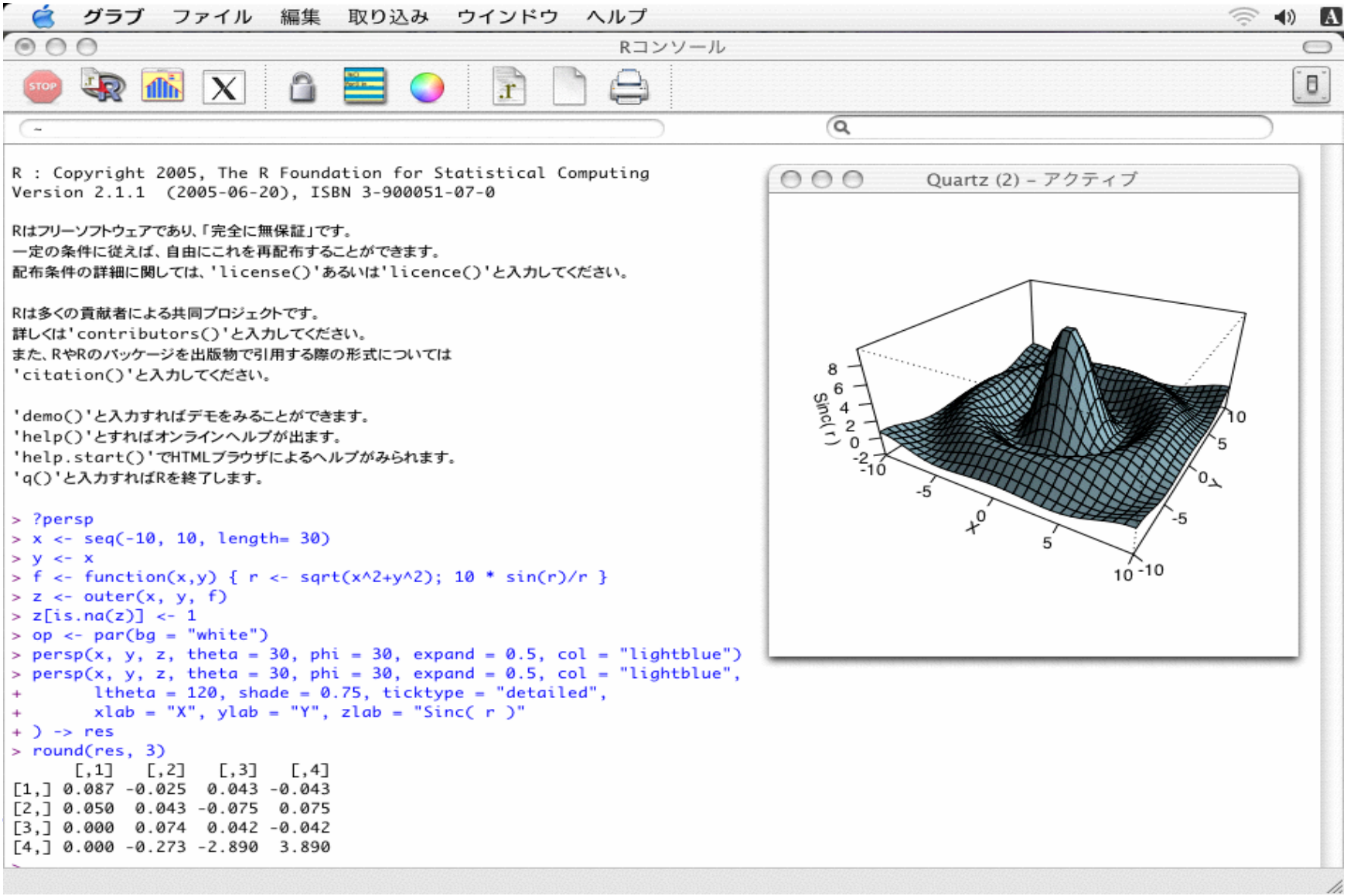
【長所】 関数電卓，数値計算，プログラミング，
統計解析，グラフィックスの機能がある
計算速度が速い
機能拡張が容易に行える

【短所】 EXCEL などの表計算ソフトに比べて
GUI の機能が劣っている・・・
大規模なデータを扱う場合は多少骨が折れる

Windows版 R


RGui
ファイル 編集 その他 パッケージ ウインドウ ヘルプ
R Console
> local({pkg <- select.list(sort(.packages(all.available = TRUE)))
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
> example(rgl)
rgl> example(rgl.surface)
rgl.sr> data(volcano)
rgl.sr> y <- 2 * volcano
rgl.sr> x <- 10 * (1:nrow(y))
rgl.sr> z <- 10 * (1:ncol(y))
rgl.sr> ylim <- range(y)
rgl.sr> ylen <- ylim[2] - ylim[1]
rgl.sr> colorlut <- terrain.colors
RGL device 1 (active)
R 2.1.1 - A Language and Environment

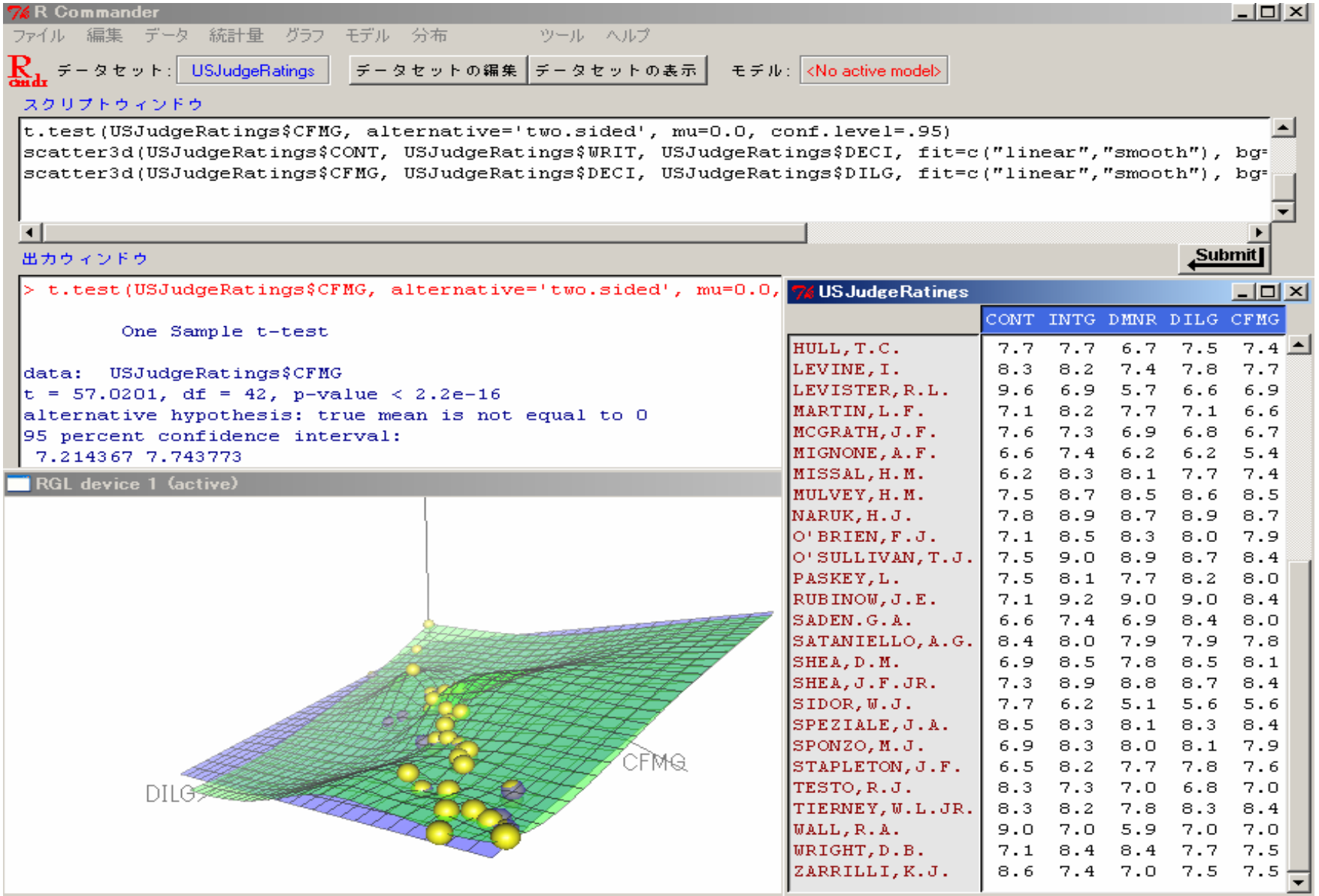
Mac OS X 版 R


グラブ ファイル 編集 取り込み ウインドウ ヘルプ
Rコンソール
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.1.1 (2005-06-20), ISBN 3-900051-07-0
Rはフリーソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()'あるいは'licence()'と入力してください。
Rは多くの貢献者による共同プロジェクトです。
詳しくは'contributors()'と入力してください。
また、RやRのパッケージを出版物で引用する際の形式については
'citation()'と入力してください。
'demo()'と入力すればデモをみることができます。
'help()'とすればオンラインヘルプが出ます。
'help.start()'でHTMLブラウザによるヘルプがみられます。
'q()'と入力すればRを終了します。
> ?persp
> x <- seq(-10, 10, length= 30)
> y <- x
> f <- function(x,y) { r <- sqrt(x^2+y^2); 10 * sin(r)/r }
> z <- outer(x, y, f)
> z[is.na(z)] <- 1
> op <- par(bg = "white")
> persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
> persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue",
+ ltheta = 120, shade = 0.75, ticktype = "detailed",
+ xlab = "X", ylab = "Y", zlab = "Sinc(r)"
+) -> res
> round(res, 3)
[,1] [,2] [,3] [,4]
[1,] 0.087 -0.025 0.043 -0.043
[2,] 0.050 0.043 -0.075 0.075
[3,] 0.000 0.074 0.042 -0.042
[4,] 0.000 -0.273 -2.890 3.890
Quartz (2) – アクティブ
Sinc(r)
X
Y

対応プラットフォーム

- Rは以下のバージョンがある
  - Windows 版
  - Mac OS X 版
  - Linux 版
    （Vine, Redhat, Debian, Mandrake, suse）
  - Unix 版
- Windows 版と Mac OS X 版はメニューやメッセージが日本語化されている

【参考】R Commander



	CONT	INTG	DMNR	DILG	CFMG
HULL,T.C.	7.7	7.7	6.7	7.5	7.4
LEVINE,I.	8.3	8.2	7.4	7.8	7.7
LEVISTER,R.L.	9.6	6.9	5.7	6.6	6.9
MARTIN,L.F.	7.1	8.2	7.7	7.1	6.6
MCGRATH,J.F.	7.6	7.3	6.9	6.8	6.7
MIGNONE,A.F.	6.6	7.4	6.2	6.2	5.4
MISSAL,H.M.	6.2	8.3	8.1	7.7	7.4
MULVEY,H.M.	7.5	8.7	8.5	8.6	8.5
NARUK,H.J.	7.8	8.9	8.7	8.9	8.7
O'BRIEN,F.J.	7.1	8.5	8.3	8.0	7.9
O'SULLIVAN,T.J.	7.5	9.0	8.9	8.7	8.4
PASKEY,L.	7.5	8.1	7.7	8.2	8.0
RUBINOW,J.E.	7.1	9.2	9.0	9.0	8.4
SADEN,G.A.	6.6	7.4	6.9	8.4	8.0
SATANIELLO,A.G.	8.4	8.0	7.9	7.9	7.8
SHEA,D.M.	6.9	8.5	7.8	8.5	8.1
SHEA,J.F.JR.	7.3	8.9	8.8	8.7	8.4
SIDOR,W.J.	7.7	6.2	5.1	5.6	5.6
SPEZIALE,J.A.	8.5	8.3	8.1	8.3	8.4
SPONZO,M.J.	6.9	8.3	8.0	8.1	7.9
STAPLETON,J.F.	6.5	8.2	7.7	7.8	7.6
TESTO,R.J.	8.3	7.3	7.0	6.8	7.0
TIERNEY,W.L.JR.	8.3	8.2	7.8	8.3	8.4
WALL,R.A.	9.0	7.0	5.9	7.0	7.0
WRIGHT,D.B.	7.1	8.4	8.4	7.7	7.5
ZARRILLI,K.J.	8.6	7.4	7.0	7.5	7.5

R による簡単な計算

- 四則演算はもちろん，累乗の計算や括弧が入った計算も行うことが出来る
- 関数電卓のような特殊な数学関数を使った計算も実行出来る

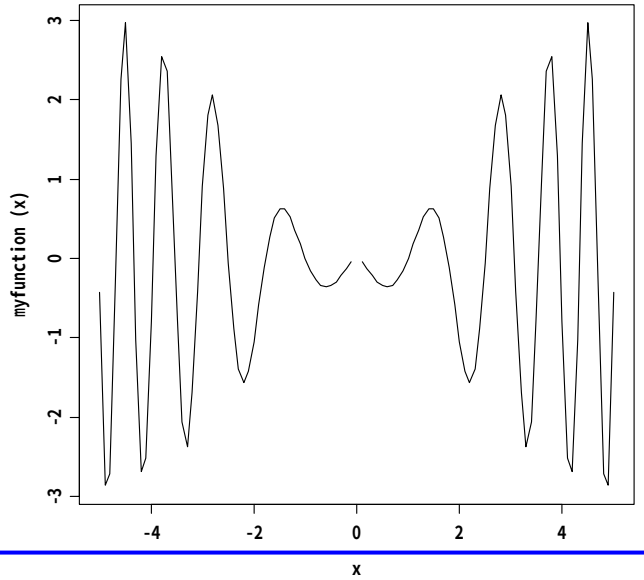
```
> (1 + 2 - 3 * 4) / 5^6
[1] -0.000576

> cos(pi) + exp(0) + sqrt(4)
[1] 2
```

R による簡単な計算

- 自分で関数を定義して，その関数を使って計算を行うことが出来る
- 関数の定義に慣れれば，シミュレーションが出来る

```
myfunction <- function(x) {
  return( sin(x^2)*log(x^2) )
}
curve(myfunction, -5, 5)
```

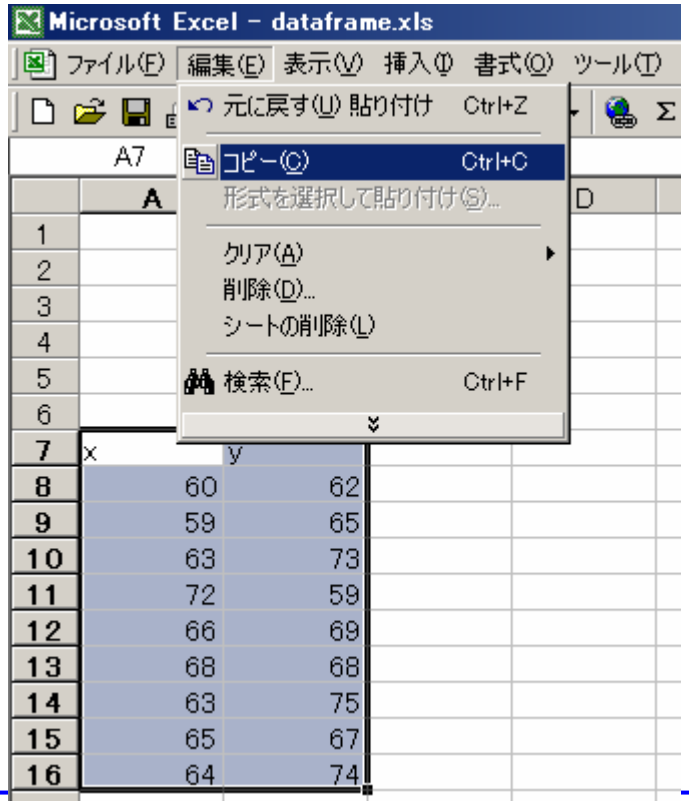


データの読み込みについて

- データを手で打ち込んで入力
- テキストファイルから読み込み
- EXCEL のセルをマウスでコピーして R に貼り付け
- 他の統計解析用ソフトのデータ形式ファイルの読み込み（EXCEL，SPSS，STATA，・・・）

EXCELのセルをコピー＆ペースト

■ Windows 版の場合

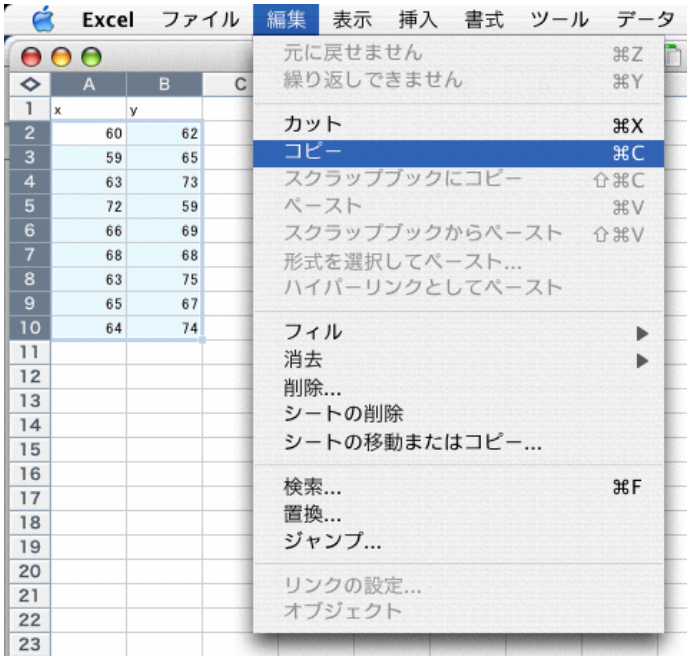



```
# 列名をコピーした場合
x <- read.delim("clipboard", header= T )

# 列名をコピーしなかった場合
x <- read.delim("clipboard", header= F )
```

EXCELのセルをコピー＆ペースト

Mac OS X 版の場合

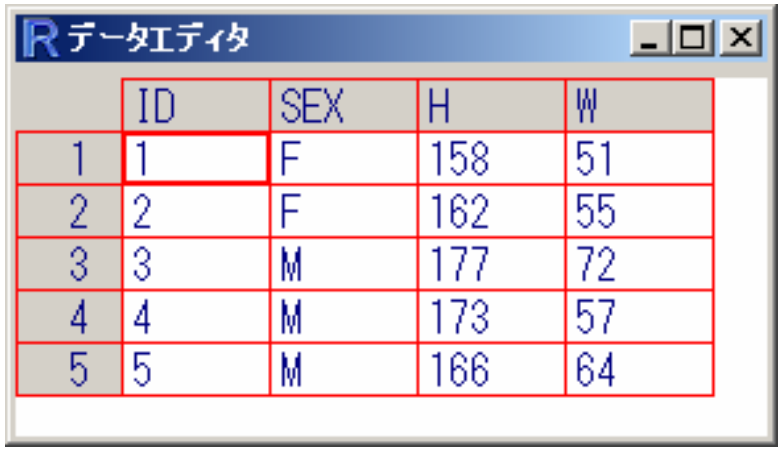



```
excel.mac <- function(...) {
  args <- c(...)
  temp <- matrix(scan(""), byrow=TRUE,
  ncol=length(args))
  data <- data.frame(temp)
  colnames(data) <- args
  return(data)
}
excel.mac("X", "Y")  # 列名を入力
1:                   # ペーストする
```

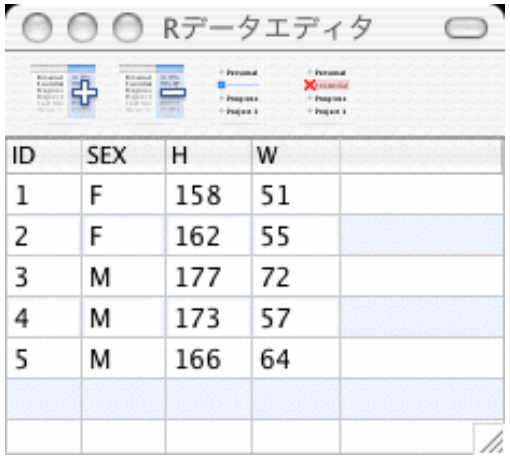
データの編集

- データをセル形式で見る場合は関数 edit(データフレーム名) を用いる

```
DF <- edit(DF)
```



Windows 版



Mac OS X 版

R と検定

R と検定

- さまざまな基本統計量を求めることが出来る

関数	機能	関数	機能
mean()	平均	max()	最大値
median()	中央値	min()	最小値
sum()	総和	range()	範囲
var()	分散	quantile()	分位点
sd()	標準偏差	IQR()	四分位偏差
cor()	相関係数	summary()	要約統計量

R と検定

- さまざまな検定を行うことが出来る

関数	機能
t.test()	t検定（1標本，2標本）
wilcox.test()	Wilcoxon検定（1標本，2標本）
var.test()	等分散の検定（F検定）
chisq.test()	χ^2 検定
fisher.test()	Fisherの直接確率検定
oneway.test()	一元配置分散分析
lm()	回帰分析，分散分析

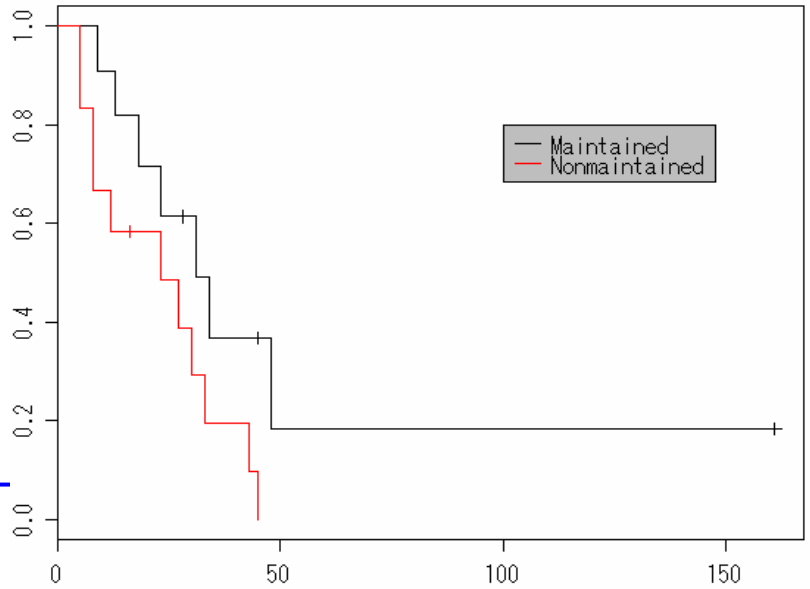
R と検定：(例)生存時間解析

■ ログランク検定（パッケージ：survival ）

```
Call:
survdiff(formula = time2 ~ x, data = MYDATA)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained     11        7    10.69      1.27      3.40
x=Nonmaintained 12       11     7.31      1.86      3.40
Chisq= 3.4  on 1 degrees of freedom, p= 0.0653
```

■ カプラン・マイヤープロット

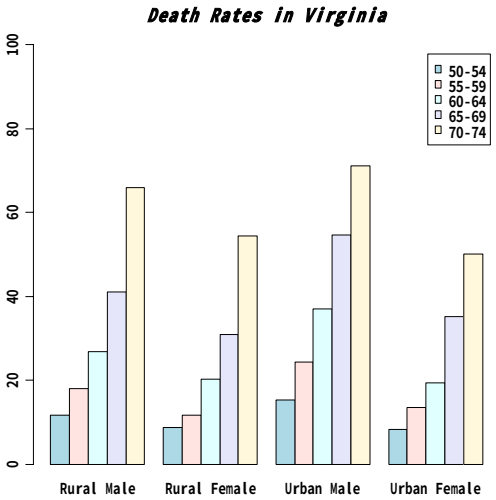


R とグラフィックス

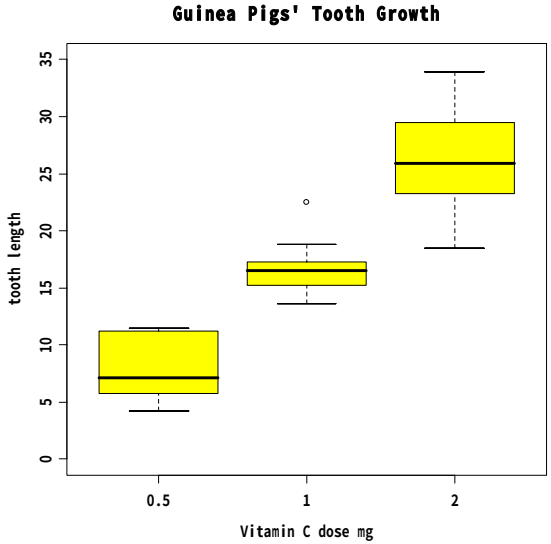
R とグラフィックス

- グラフィックスを作成するのが非常に楽！
  （他のソフトに比べると差が歴然！！）
- 簡単な命令ですぐに見栄えの良いグラフが得られる ⇒ 微調整は不要！
  - もちろん，カスタマイズも柔軟に行える
- すぐに画像ファイルとして保存できる
  - 手っ取り早くビットマップ形式でコピー＆保存可
  - さまざまな画像形式に対応
    （jpeg，png，emf，eps，pdf…）

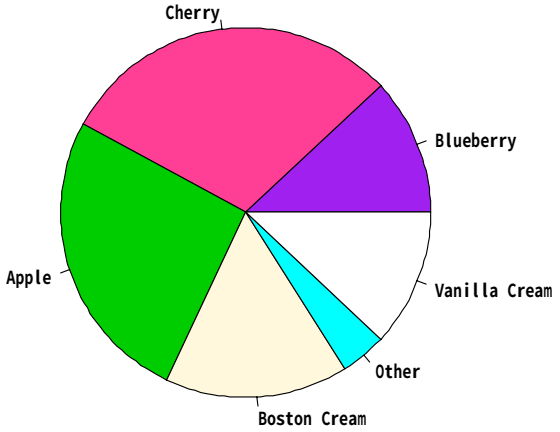
R とグラフィックス



棒グラフ

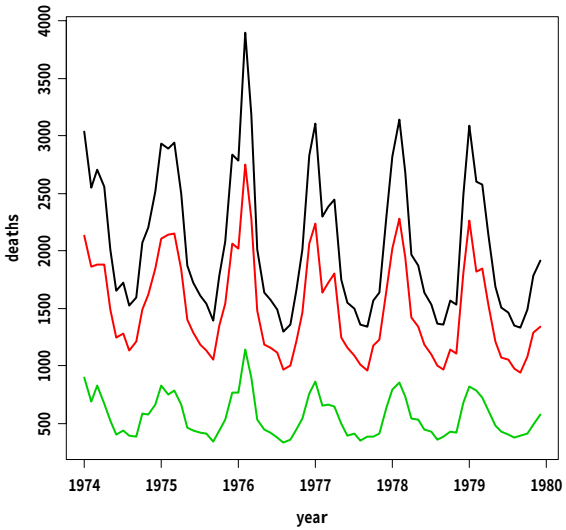


箱ひげ図

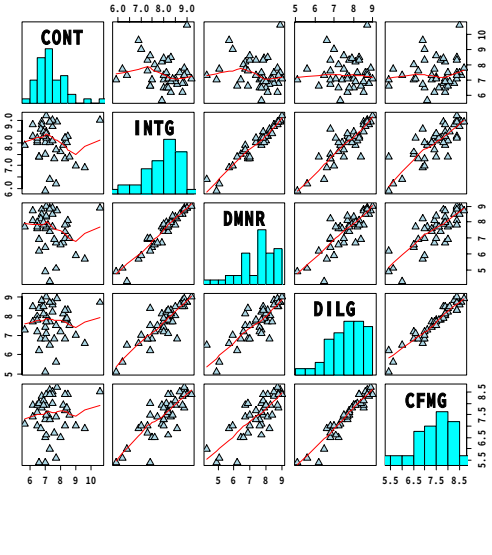


円グラフ

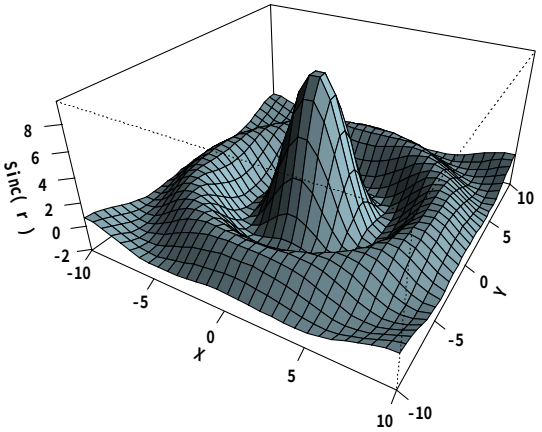
R とグラフィックス



時系列データのプロット



5変数データの散布図



2変数関数のグラフ

R とシミュレーション

R とシミュレーション

- プログラムの作成が比較的簡単なので，関数の定義に慣れれば，すぐにシミュレーションを実行することが出来る
  - 計算速度が速い
  - 乱数の質が高い
    （メルセンヌ・ツイスター法）
  - 結果をグラフィックスに出力することが出来る

例1：乱数生成法のあら探し

- 線形合同法：以下の式から一様乱数列を得る

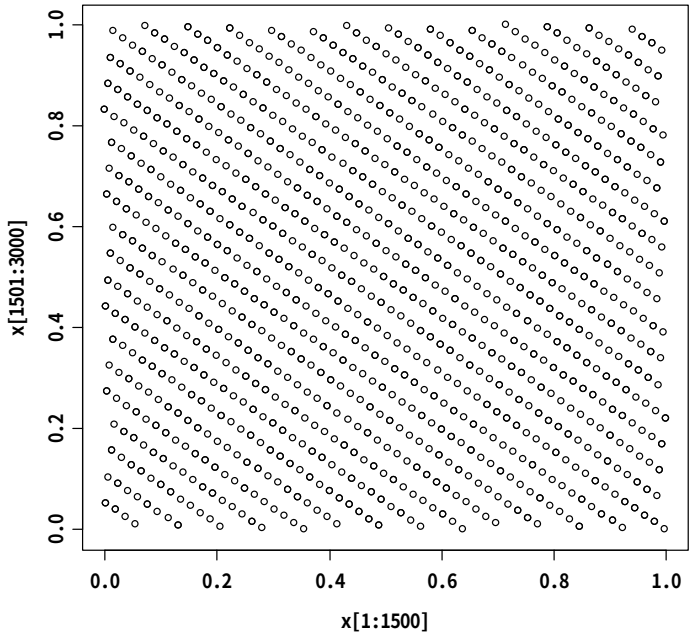
$$X_{n+1} = a*X_n + c \qquad (\mathrm{mod}\ M;\ n \geqq 1)$$

- 乗数 a と定数 c，法 M の選択によっては…
  - 乱数列の周期の長さが異なってくる
  - 生成した乱数の点列を組み合わたときに規則性（結晶構造）を生じることがある

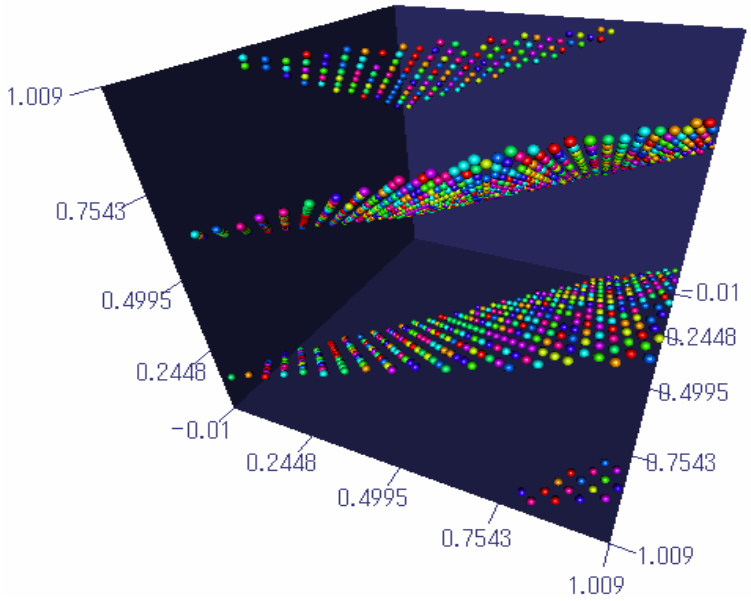
⇒ R で実験してみる

$X_{n+1} = 45X_n + 1 \ (\mathrm{mod}\ 2^{10})$ で乱数を3000個生成

例1：乱数生成法のあら探し



$(x_1, y_1), (x_2, y_2)$ … を1500組



$(x_1, y_1, z_1), (x_2, y_2, z_2)$ … を1000組

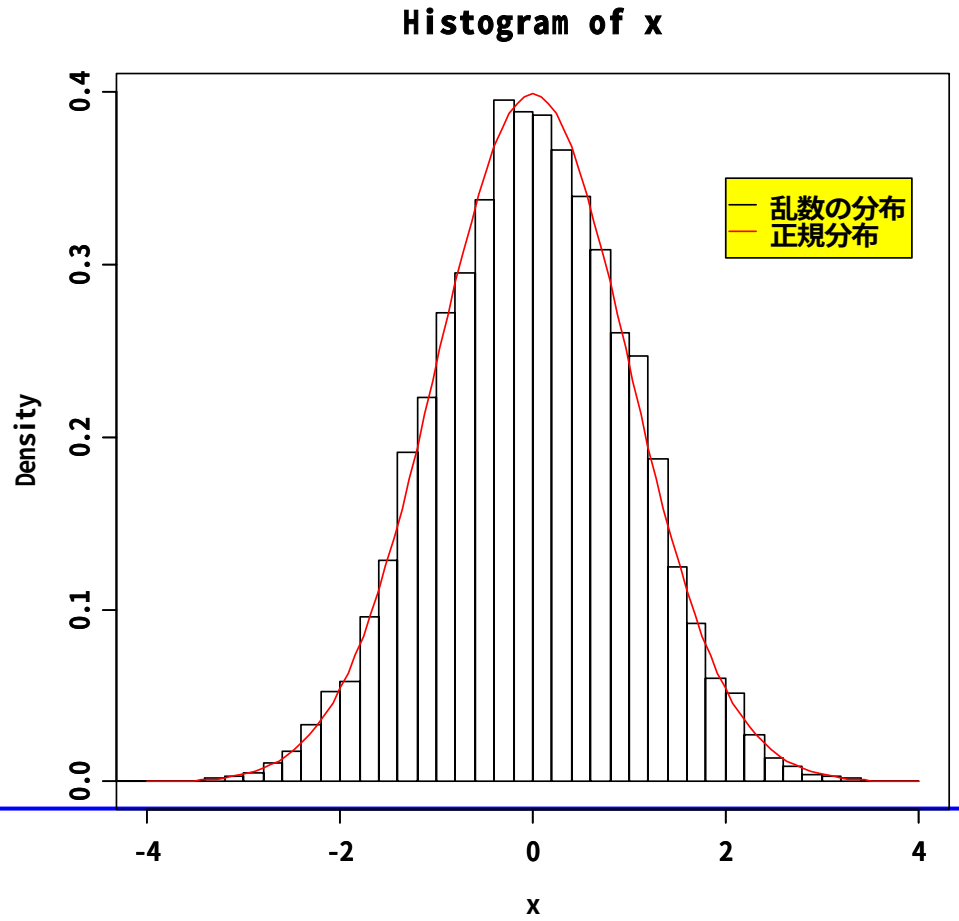
- ちなみに R はメルセンヌ・ツイスター法により乱数を生成している（周期 $2^{19937}-1$，623次元で一様分布 ）

例2：正規乱数の生成

- $U_1, \cdots, U_{12} \sim U(0, 1)$
  - $E(U_i) = 1/2$，$V(U_i) = 1/12 \quad (i = 1, \ldots, 12)$
- $X = U_1 + \cdots + U_{12} - 6 \sim N(0, 1)$
  - 中心極限定理より，X は正規分布に従う
  - 加算と減算で正規乱数を生成することが出来る
  - 一様乱数12個から生成しているので良質
  - この方法では，$|X| \geqq 6$ となるものが出来ない
    ⇒ 理論上の確率は数億分の1

例2：正規乱数の生成

- X を 10000 個生成したグラフ（実行時間1秒）



例2：正規乱数の生成

- R には特定の確率分布に従う乱数を生成する関数が多数用意されている

分布名	分布名
ベータ分布	多項分布
二項分布	負の二項分布
コーシー分布	正規分布
カイ二乗分布	ポアソン分布
指数分布	ウィルコクソンの符号付順位和統計量の分布
F分布	t 分布
ガンマ分布	一様分布
幾何分布	スチューデント化された分布
超幾何分布	ワイブル分布
対数正規分布	ウィルコクソン順位和統計量の分布
ロジスティック分布	

まとめ

本日のまとめ

- R とは？ ⇒ フリーの統計解析ソフト
- R で出来ること
  ⇒ 関数電卓，数値計算，プログラミング，
  統計解析，グラフィックスの機能がある
  ⇒ 機能拡張が容易に行える
- R とグラフィックス
  ⇒ グラフが簡単に作成できる
- R とシミュレーション
  ⇒ 計算速度が速い，プログラム作成が容易

参考文献

- 中澤 港『Rによる統計解析の基礎』（ピアソンエデュケーション）
- 間瀬 茂 他『工学のためのデータサイエンス入門』（数理工学社）
- 岡田 昌史 他『The R Book』（九天社）
- 荒木 孝治 他『フリーソフトウェア R による統計的品質管理入門』（日科技連）
- 『R-Tips（インターネット上）』 http://cse.naro.affrc.go.jp/takezawa/r-tips.pdf

無料で有意差を出す方法

～ 統計解析フリーソフト R の紹介 ～

完