データ解析
第二回

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

今日の講義内容

- 関数定義
- for 文，if 文
- 乱数生成
- ヒストグラムによる可視化

関数定義

```
tmp <- function(x) {
         y <- x^2
         return(y)
       }
```

で

```
> tmp(2)
[1] 4
```

を得る.

関数定義を"hoge.R"なるファイルに書き込んで,

```
> source("hoge.R")
```

とすればファイル内で定義した関数を読み込める.

リストの返り値

```
> tmp <- function(x,y) list(x=2,sin.x=sin(x),y=y,log.y=log(y))
> z <- tmp(2,3)
> z
$x
[1] 2

$sin.x
[1] 0.9092974

$y
[1] 3

$log.y
[1] 1.098612
```

練習問題 1

$n \times d$ 行列 X と n 次元ベクトル y を受け取って，$(X^\top X)^{-1} X^\top y$ を返す関数を定義せよ．

for 文

```
for(x in 1:3){
    y <- x^2
    cat(y,fill=TRUE)
}
```

; を使って一行にまとめることも可能

```
for(x in 1:3){y <- x^2;cat(y,fill=TRUE);}
```

if 文

基本形

```
if(x < 0) -x else x
```

複数行・複数命令

```
if(x < 0){
    z <- x^2
    y <- x^3
}else{
    z <- x^3
    y <- x^2
}
```

練習問題 2

$n \times d$ 行列 X と n' 次元ベクトル y を受け取り，$n = n'$ なら $(X^\top X)^{-1} X^\top y$ を返し，$n \neq n'$ なら y 自身を返す関数を定義せよ．

乱数生成

r+(乱数名) で乱数生成
d+(乱数名) で確率密度関数
p+(乱数名) で累積分布関数
q+(乱数名) で分位点

例:正規分布 (norm)

```
> rnorm(3)
[1]  0.9372860  0.3960432 -0.5254500
> dnorm(1.4)    # X=1.4 における確率密度
[1] 0.1497275
> pnorm(1.4)    # P(X <= 1.4)
[1] 0.9192433
> qnorm(0.9192433)
[1] 1.400000
```

確率分布	乱数名
ベータ分布	beta
二項分布	binom
コーシー分布	cauchy
カイ二乗分布	chisq
指数分布	exp
F 分布	f
ガンマ分布	gamma
幾何分布	geom
超幾何分布	hyper
対数正規分布	lnorm
ロジスティック分布	logis
多項分布	multinom
負の二項分布	nbinom
正規分布	norm
ポアソン分布	pois
ウィルコクソンの符号付順位和統計量の分布	signrank
t 分布	t
一様分布	unif
スチューデント化された分布	tukey
ワイブル分布	weibull
ウィルコクソンの順位和統計量の分布	wilcox

ヒストグラムの表示

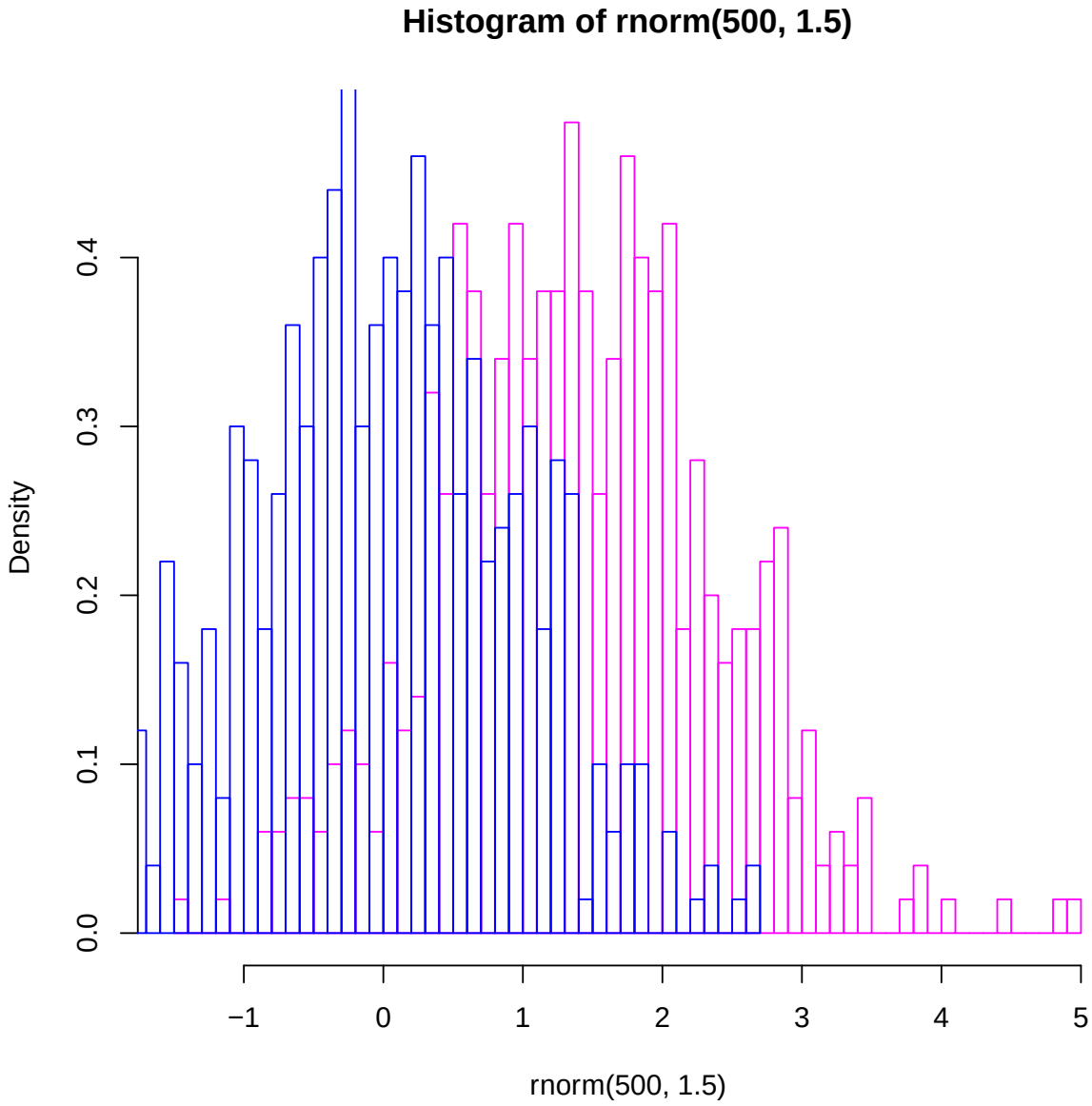
```
hist(rnorm(100))

hist(rnorm(100),breaks=20) #ビンの数を設定
```

２つのヒストグラムを重ねて表示.

```
hist(rnorm(500,1.5), col = "#ff00ff40", border = "#ff00ff",
breaks = 50, freq = FALSE)
hist(rnorm(500), col = "#0000ff40",
border = "#0000ff", breaks = 50, freq = FALSE, add = TRUE)
```

add = TRUE で重ね表示.
col でパネルの内側の色を指定，"#rrggbbtt" で 16 進数を使って RGB の強さと透過度を指定 (00 から ff まで)．最後の二桁は透過度.
border で枠の色を指定.
freq=FALSE で数ではなく密度を表示（各ビンに入ったサンプル数の割合）.

表示結果


Histogram of rnorm(500, 1.5)
Density
0.0
0.1
0.2
0.3
0.4
−1
0
1
2
3
4
5
rnorm(500, 1.5)

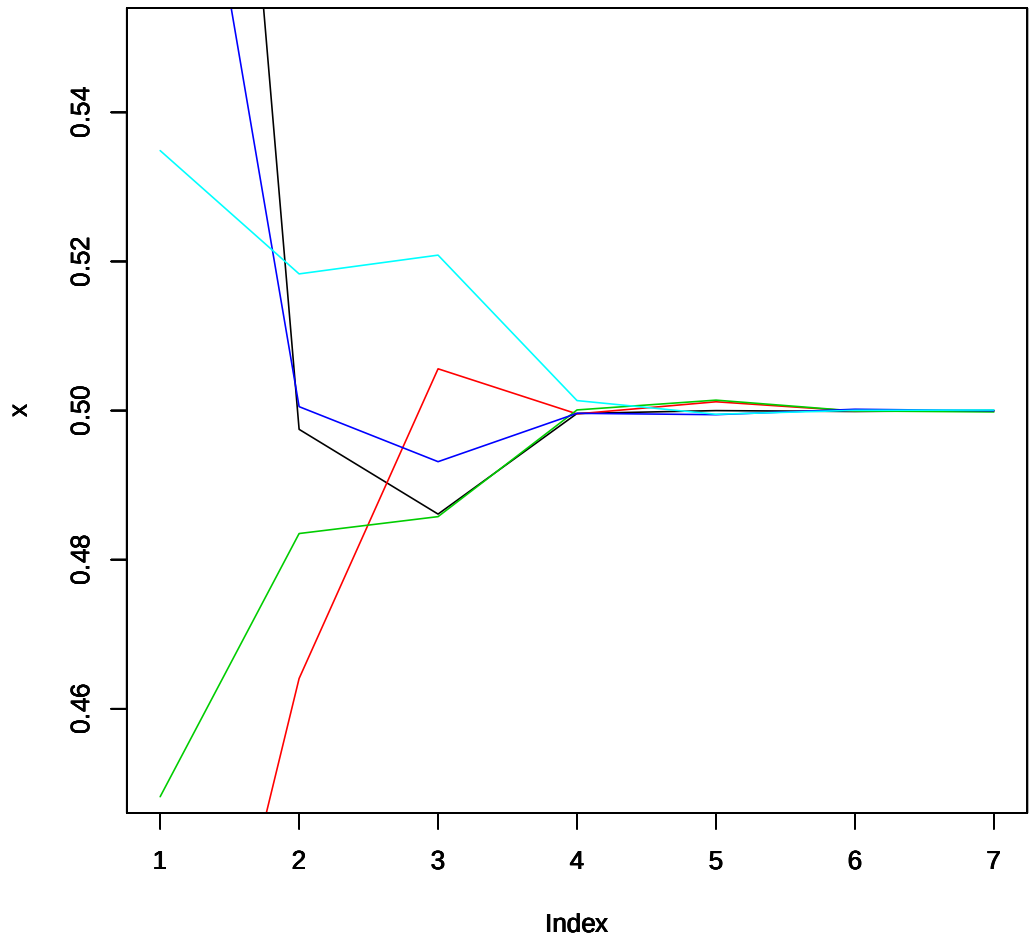
和，平均，分散，標準偏差

```
sum(runif(10))      #和
mean(runif(10))     #平均
var(runif(10))      #分散
sd(runif(10))       #標準偏差
```

大数の法則を確認

```
for(j in 1:5){
 for(i in 1:7) x[i]<-c(mean(runif(10^i)));
 plot(x,type='l',col=j,ylim=c(0.45,0.55));  #ylim でｙ軸の範囲を指定
 par(new=T);  #重ね書きモードにする
}
```

表示結果



レポート問題

1. モンテカルロ法
   $[0,1] \times [0,1]$ 上の一様分布 X を大量に (例えば 10^7 個) 発生させ，$\|X\| \leq 1$ なるサンプルの割合を求め，それによって半径 1 の 1/4 円の面積 $\pi/4$ の近似値を求めよ．
2. X を $[0,1]$ 上の一様分布として $\mathrm{E}[2\log(X)]$ はいくらか？また，$n = 100, 100^2, 100^3$ 個のサンプルを発生させて，$\frac{1}{n}\sum_{i=1}^{n} 2\log(X_i)$ を計算せよ．理論値には近づいているか？
3. 中心極限定理の再現
   一様分布から $n = 3, 10, 100$ 個のサンプル $\{X_i\}_{i=1}^{n}$ を発生させる試行をそれぞれ 1000 回ずつ行い，各試行で $\sqrt{n}(\sum_{i=1}^{n} X_i/n - \mathrm{E}[X])$ を計算し，中心極限定理が成り立っていることを確かめよ．確かめ方はどんな方法でも良い (例えばヒストグラムと正規分布の密度関数を同時にプロットしてみよ)．余裕があれば他の分布でも確かめてみよ．

レポートの提出方法

- 私宛にメールにて提出.
- 件名に 必ず「データ解析第一回レポート」と明記し，R のソースコードと結果をまとめたレポートを送付のこと.
- 氏名と学籍番号も忘れず明記すること.
- レポートは本文に載せても良いが，pdf などの電子ファイルにレポートを出力して添付ファイルとして送付することが望ましい (これを期に tex の使い方を覚えることを推奨します).
- 提出期限は講義最終回まで.

※相談はしても良いですが，コピペはダメです.

講義情報ページ

`http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/dataanalysis.html`