# 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

- 確率統計の復習
- 関数定義
- for 文，if 文
- 乱数生成
- ヒストグラムによる可視化

# 構成

# 確率変数

確率変数とは集合 $A$ に対して，その集合に $X$ が含まれる確率 $P(X \in A)$ が矛盾なく定まっている変数のことである．

# より正確な定義

標本空間 $\Omega$ 上に $P(A)$ が定まっている集合の族として $\sigma$-加法族 $\mathcal{F}$ を考える．$\sigma$-加法族とは以下の性質を満たす集合の族である．

- $\Omega \in \mathcal{F}$.
- $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.
- $A_n \in \mathcal{F}\ (n = 1, 2, \dots) \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

標本空間 $\Omega$ とその上に定まった $\sigma$-加法族 $\mathcal{F}$ の組 $(\Omega, \mathcal{F})$ を可測空間という．
$A \in \mathcal{F}$ に対して確率 $P(A)$ を返す関数を確率測度と言う．確率測度は次の性質を満たさなくてはいけない．

- $P(A) \geq 0 \quad (\forall A \in \mathcal{F})$.
- $A_n \in \mathcal{F}\ (n = 1, 2, \dots)$ が互いに排反ならば，$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.
- $P(\Omega) = 1$.

$(\Omega, \mathcal{F}, P)$ を確率空間という．

# より正確な定義

位相空間上（この講義では $\mathbb{R}^p$ しか考えない）で，全ての開集合を含む最小の $\sigma$-集合族のことを**ボレル集合族**という．

$\mathbb{R}^p$ 値確率変数 $X$ とは，$X : \Omega \to \mathbb{R}^p$ なる**可測関数**である．つまり，$\mathbb{R}^p$ 上のボレル集合族を $\mathcal{B}$ として，$\forall B \in \mathcal{B}$ に対し $X^{-1}(B) \in \mathcal{F}$ を満たす関数である．

言い替えると全ての集合 $B \in \mathcal{B}$ に対して，$X$ が $B$ の中に値を取る確率 $P(X \in B)$ が $P(X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\})$ として矛盾なく定まっているということである．

通常は $\Omega = \mathbb{R}^p$，$\mathcal{F} = \mathcal{B}$ として，$X : \mathbb{R}^p \ni \omega \mapsto \omega \in \mathbb{R}^p$ として考える．

# 確率密度関数と分布関数 ( 1 変量)

$X$ を実確率変数とする.

- 確率分布関数

$$F(x) = \mathrm{P}(X \leq x)$$

- 確率密度関数 (連続確率変数)

$$p(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}.$$

$F$ が微分可能でなければ密度関数は存在しない．確率密度関数が存在すれば

$$P(X \in A) = \int_A p(x)\mathrm{d}x$$

である．$F(x)$ が連続な確率変数を連続確率変数とよび，微分可能な場合には絶対連続確率変数とよぶ．連続でも絶対連続とは限らない（カントールの階段関数）.

- 確率質量関数 (離散確率変数)

$$p(a) = P(X = a).$$

離散確率変数が $a$ に値をとる確率を確率質量関数とよぶ.

# 多変量確率変数

$X = (X_1, \ldots, X_d)$ を $d$ 次元確率変数とする.

- $X = (X_1, \ldots, X_d)$ の同時分布: $P(X \in A)$.

$$P(X \in A) = \int_A f(x_1, \ldots, x_d) \mathrm{d}x_1 \ldots \mathrm{d}x_d$$

と書ける時, $f$ を同時確率密度関数と言う.

- $X_j$ の周辺分布: $P(X_j \in A)$.

$$f_j(x_j) = \int \ldots \int f(x_1, \ldots, x_d) \mathrm{d}x_1 \ldots \mathrm{d}x_{j-1} \mathrm{d}x_{j+1} \ldots \mathrm{d}x_d$$

の時, $f_j$ を周辺密度関数と言う.

- 独立性:

$$X_1 \text{ と } X_2 \text{ が独立} \Leftrightarrow P(\{X_1 \in A_1\} \cap \{X_2 \in A_2\}) = P(X_1 \in A_1) P(X_2 \in A_2).$$

# 各種統計量

- 平均

$$\mu = \mathrm{E}[X] = \int x p(x) \mathrm{d}x.$$

- 分散 ($X$ は一次元の確率変数とする)

$$V = \mathrm{E}[(X-\mu)^2] = \int (x-\mu)^2 p(x) \mathrm{d}x.$$

  分布がどれだけ「広がっているか」の指標.

- 標準偏差

$$\sigma = \sqrt{V}.$$

# 各種統計量

- 共分散 ($X, Y$ をそれぞれ一次元の確率変数とする)

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \int (x - \mu_X)(y - \mu_Y) p(x, y) \mathrm{d}x\mathrm{d}y.$$

  ただし，$\mu_X$, $\mu_Y$ はそれぞれ $X, Y$ の平均.

- 相関 ($X, Y$ をそれぞれ一次元の確率変数とする)

$$\mathrm{Corr}(X, Y) = \mathrm{E}\left[\frac{(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}\right] = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

  ※ 相関が 0 でも独立であるとは限らない．反例を作ってみよ.

- 分散共分散行列 ($X \in \mathbb{R}^p$ は多次元の確率変数とする)

$$\Sigma = \mathrm{E}[XX^\top] = \int (x - \mu)(x - \mu)^\top p(x) \mathrm{d}x \in \mathbb{R}^{p \times p}.$$

  分散の多次元版．各成分の分散を対角成分に，成分間の共分散を非対角成分に持つ.

# 各種統計量 (続き)

- 歪度 ($X$ は一次元の確率変数とする)

$$\beta_1^{1/2} = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

  分布がどれだけ対称でないか（歪んでいるか）の指標.

- 尖度

$$\beta_2 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3.$$

  分布がどれだけ尖っているか.
  (−3 をしないで定義する場合もある．3 は標準正規分布の尖度で，標準正規分布と比べて尖っているかどうかの指標になる．)

# 大数の法則

## Theorem (大数の弱法則)

$X$ を $\mathbb{R}^p$ 値確率変数とする．$\mu = \mathrm{E}[X]$ に対し ($\|\mu\| < \infty$ とする)，$X$ と同じ分布に従う *i.i.d.* 確率変数列 $X_i$ の平均は $\mu$ に確率収束する: $\forall \epsilon$

$$\lim_{n\to\infty} P\left(\left\|\frac{\sum_{i=1}^n X_i}{n} - \mu\right\| \geq \epsilon\right) = 0. \qquad \text{（確率収束）}$$

実は，上と同じ条件でより強い「大数の強法則」が成り立つ．

## Theorem (大数の強法則)

$X$ を $\mathbb{R}^p$ 値確率変数とする．$\mu = \mathrm{E}[X]$ に対し，$X$ と同じ分布に従う *i.i.d.* 確率変数列 $X_i$ の平均は $\mu$ に概収束する:

$$P\left(\lim_{n\to\infty} \frac{\sum_{i=1}^n X_i}{n} = \mu\right) = 1. \qquad \text{（概収束）}$$

# 大数の弱法則の略証 (分散存在時)

$X$ が一変量で分散 $V$ $(< \infty)$ が存在する場合に証明する (多変量の場合も全く同様にして示せる). このとき,

$$\begin{aligned} \mathrm{E}\left[\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)^2\right] =& \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{E}\left[(X_i - \mu)^2\right] \\ =& \frac{V}{n} \to 0 \end{aligned}$$

である. あとは, マルコフの不等式より

$$\begin{aligned} & \mathrm{P}\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| \geq \epsilon\right) \\ =& \mathrm{P}\left(\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)^2 \geq \epsilon^2\right) \\ \leq& \frac{1}{\epsilon^2} \mathrm{E}\left[\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)^2\right] \to 0. \end{aligned}$$

より詳細は伊藤清著『確率論』を参照のこと.

# 中心極限定理

## Definition (分布収束 (弱収束，法則収束))

ある実数値確率変数 $X$ が 確率密度関数を持つ時，ある確率変数の列 $S_n$ が，$X$ に分布収束するとは，

$$\mathrm{P}(S_n \leq R) \to \mathrm{P}(X \leq R) \quad (\forall R \in \mathbb{R})$$

となることと定義される．$X_n \xrightarrow{\mathrm{d}} X$ と書く．

## Theorem (中心極限定理)

$X_i$ $(i = 1, \ldots, n)$ を $\mathbb{R}$ 上 *i.i.d.* 確率変数とし，それらの平均と標準偏差がそれぞれ $\mu$，$\sigma$ であるとする．このとき，

$$\frac{\sum_{i=1}^{n} X_i - \mu}{\sigma\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, 1).$$

# 分布収束の一般化

先のスライドでは分布収束先として絶対連続確率変数を仮定したが，任意の $\mathbb{R}^p$ 上ボレル確率測度に対して分布収束の概念は一般化できる．

## Definition

$\mathbb{R}^p$ 値確率変数 $X_n$ が確率変数 $X$ に分布収束するとは，任意の**有界連続関数** $f$ に対して

$$\mathrm{E}[f(X_n)] \to \mathrm{E}[f(X)]$$

が成り立つこととして定義される．このとき，$X_n \xrightarrow{\mathrm{d}} X$ と書く．

さらにこれは以下の条件と同値であることが知られている (Portmanteau の定理)

- 任意の開集合 $G \subseteq \mathbb{R}^p$ に対して $\liminf \mathrm{P}(X_n \in G) \geq P(X \in G)$.
- 任意の閉集合 $F \subseteq \mathbb{R}^p$ に対して $\limsup \mathrm{P}(X_n \in F) \leq P(X \in F)$.
- $P(X \in \partial B) = 0$ なる任意のボレル集合 $B$ に対して，

  $\mathrm{P}(X_n \in B) \to \mathrm{P}(X \in B)$. ただし，$\partial B$ は $B$ の境界つまり $\bar{B} - \overset{\circ}{B}$ (閉包-内点集合) である．

# 中心極限定理（多変量版）

## Theorem (中心極限定理)

$X_i$ $(i = 1, \ldots, n)$ を $\mathbb{R}^p$ 上 *i.i.d.* 確率変数とし，それらは平均 $\mu$，分散共分散行列 $\Sigma$ を持つとする．このとき，

$$\frac{\sum_{i=1}^n X_i - \mu}{\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, \Sigma).$$

任意のボレル集合 $A$ に対して

$$\mathrm{P}\left(\frac{\sum_{i=1}^n X_i - \mu}{\sqrt{n}} \in A\right) \to \mathrm{P}_{N(0,\Sigma)}(A)$$

である，とも言い替えられる．

# パラメータ推定問題

パラメータ $\theta \in \Theta$（母数）で特徴付けられた分布の族 $P_\theta$ があった時，これを「統計モデル」と言う:

$$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}.$$

やりたいこと: データ $X$ があるパラメータ $\theta_0$ に対応する分布 $P_{\theta_0}$ から生成されているとき，未知のパラメータ $\theta_0$ をデータ $X$ から推定したい．
（例：正規分布の平均と分散，二項分布の平均）

# 最尤推定

各 $P_\theta$ に対し，密度関数 $p_\theta(x)$ が存在しているとする（離散分布においては確率質量関数とする）.

尤度関数: $p_\theta(X)$
対数尤度関数: $\log p_\theta(X)$

## 最尤推定量

$n$ 個の観測データ $X = (X_1, \ldots, X_n)$（i.i.d.）が与えられたとき，最尤推定量は次のように与えられる.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log p_\theta(X) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_\theta(X_i).$$

# 最尤推定量の一致性

**Theorem (一致性)**

（ある条件のもと）最尤推定量は 一致性 を持つ:

$$\lim_{n \to \infty} \mathrm{P}(\|\hat{\theta} - \theta_0\| \geq \epsilon) = 0 \quad (\forall \epsilon).$$

※ サンプル数無限大の極限で真のパラメータに近づく.

# 最尤推定量の漸近正規性

**Fisher 情報行列**

$$
\begin{aligned}
F(\theta) &:= \mathrm{E}_\theta \left[ \nabla \log(p_\theta(X)) \nabla^\top \log(p_\theta(X)) \right] \\
&= \left( \mathrm{E}_\theta \left[ \frac{\partial \log(p_\theta(X))}{\partial \theta_i} \frac{\partial \log(p_\theta(X))}{\partial \theta_j} \right] \right)_{i,j}.
\end{aligned}
$$

## Theorem (漸近正規性)

（ある条件のもと）最尤推定量は 漸近正規性 を持つ:

$$
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathrm{d}} N(0, F^{-1}(\theta_0)).
$$

漸近分散が $F^{-1}(\theta_0)$ であることを，漸近有効性と言う（漸近的に最適）.

# 漸近正規性の略証

$\hat{\theta}$ は尤度関数を最大化しているので,

$$\nabla \log p_\theta(X)|_{\theta=\hat{\theta}} = 0$$
$$(= \sum_{i=1}^{n} \nabla \log p_\theta(X_i)|_{\theta=\hat{\theta}})$$

である．$\theta_0$ のまわりで左辺をテイラー展開して $\frac{1}{\sqrt{n}}$ 倍することで,

$$0 = \frac{1}{\sqrt{n}} \nabla \log p_\theta(X)|_{\theta=\theta_0} + \frac{1}{\sqrt{n}} \nabla\nabla^\top \log p_\theta(X)|_{\theta=\theta_0} (\hat{\theta} - \theta_0) + o_p(1)$$

である．よって,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \underbrace{\left( \frac{-1}{n} \nabla\nabla^\top \log p_\theta(X)|_{\theta=\theta_0} \right)^{-1}}_{\xrightarrow{p} F(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}} \nabla \log p_\theta(X)|_{\theta=\theta_0}}_{\xrightarrow{d} N(0, F(\theta_0))} + o_p(1)$$
$$\xrightarrow{d} N(0, F(\theta_0)^{-1}).$$

# 構成

# 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
```

# 構成

# 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．

# 構成

# 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 |

# 構成

# ヒストグラムの表示

```
hist(rnorm(100))
```

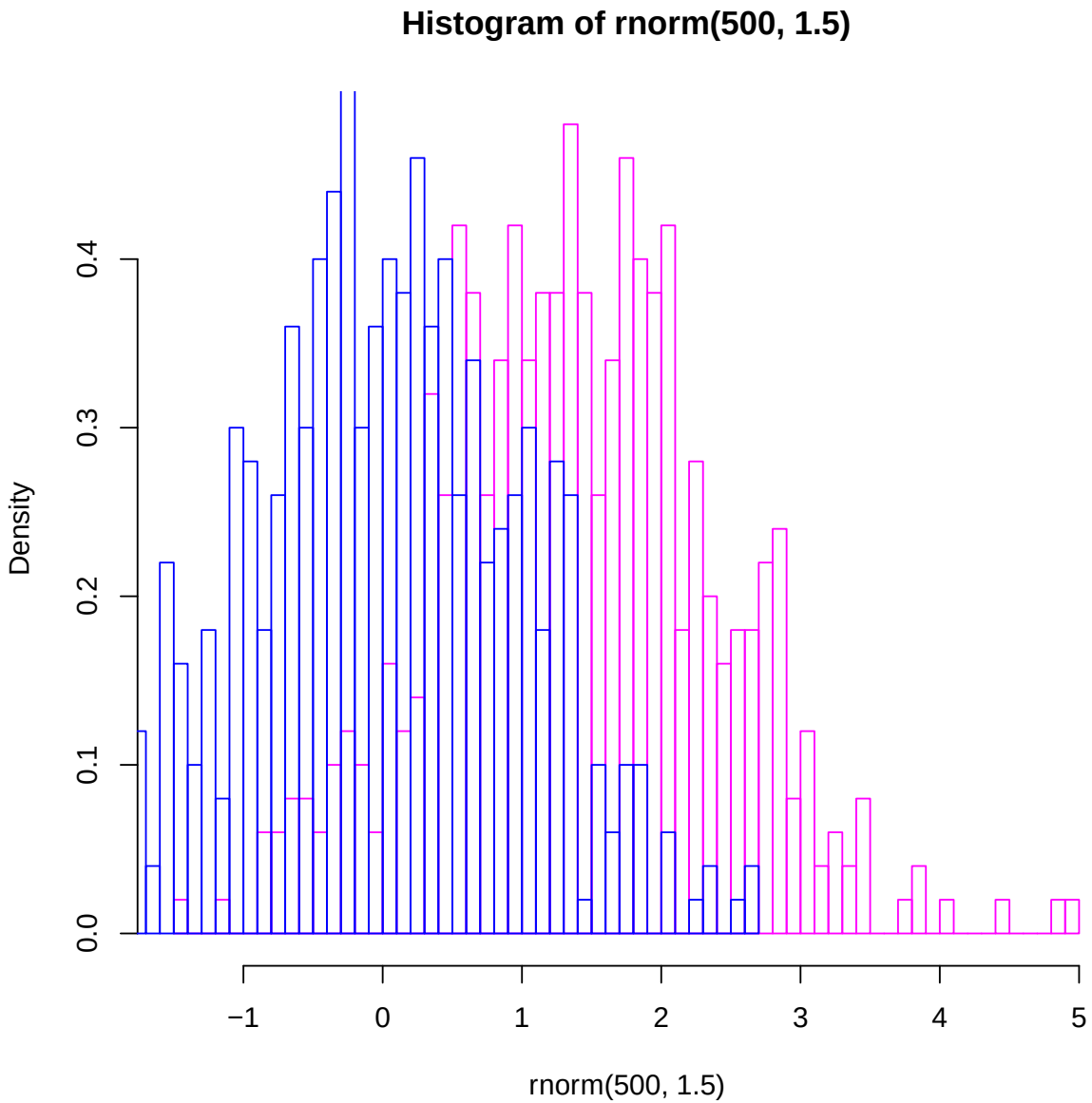
```
hist(rnorm(100),breaks=20) #ビンの数を設定
```

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)

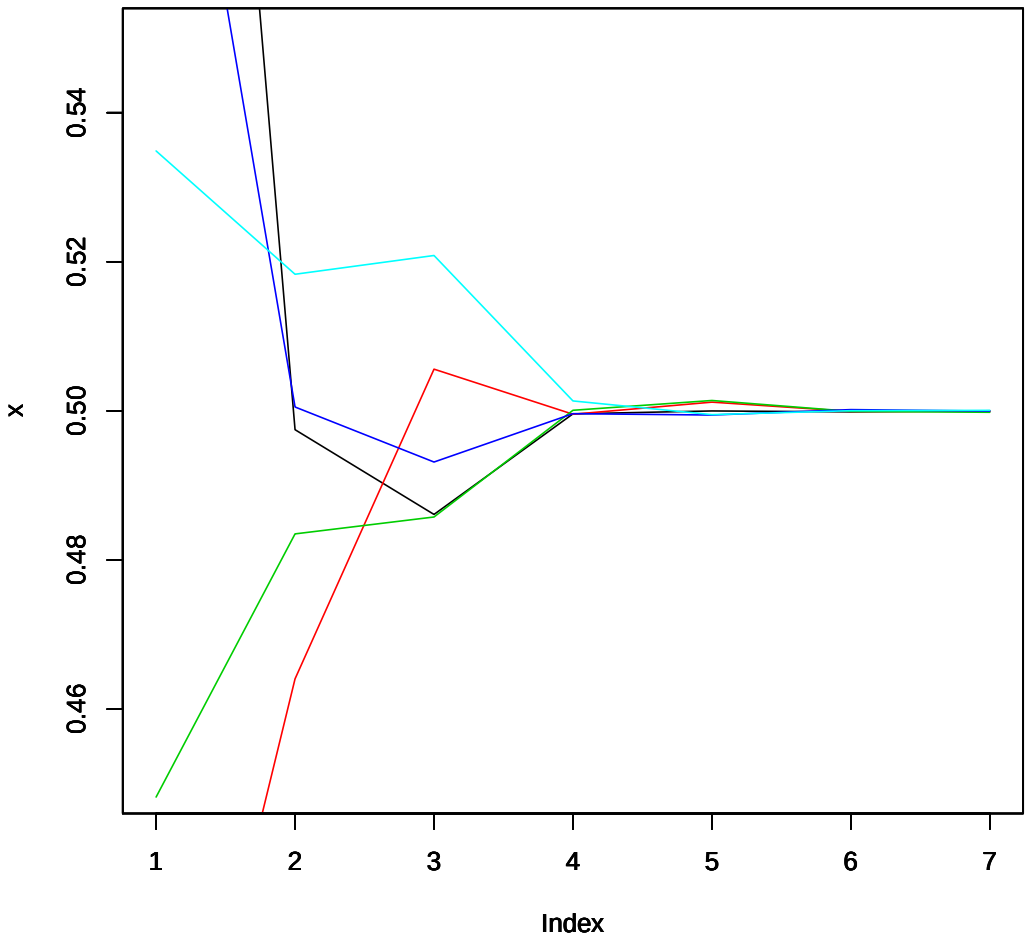
# 和，平均，分散，標準偏差

```
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



# 構成

# 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 $\sqrt{n}(\sum_{i=1}^{n} X_i/n - \mathrm{E}[X])/\sigma$ 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/2015/dataanalysis/dataanalysis.html`