R 入門（1）: 統計量の計算、ヒストグラム

1　おうちでインストール（初回のみ）

http://cran.md.tsukuba.ac.jp/へアクセスし、OS（Windows、Mac OS、Linux）を選ぶ
base をクリック→R-3.1.0-win.exe(最新版)をクリックし、デスクトップ上に保存
R-3.1.0-win.exe を double click して install する（install 後は R-3.1.0rc-win.exe を delete する）

デスクトップ上に¥R というディレクトリを作成しておく

2　データの作成（適宜）

(1) R で利用するデータを Excel で作っておく
　1行目　　変数名
　2行目以降　データ
(2) ファイルを[名前を付けて保存(A)]→[保存先]c:…¥R　[ファイル名]ex1　[ファイルの種類]CSV(カンマ区切り)とし、保存

4　R の起動（毎回）

(1) 作業 Directry の指定
　メニューから[ファイル]→[ディレクトリの変更]を選択
　[フォルダの参照]で[コンピュータ]－c ドライブ－ユーザー－…－デスクトップ－¥R
(2) プログラムの入力
　プログラムを copy&paste して[enter]を押す

```
data1 <- read.table("ex1.csv",header = TRUE, sep =",")
data1
mean(data1$height);
attach(data1)
mean(height); sd(height); summary(height)
hist(height); boxplot(height)
```

※グラフの保存
　R Graphics 窓上でマウスの右ボタンをクリック→メタファイル/ビットマップにコピー→Word に貼り付ける

演習　日吉の学生の通学時間(time)の分布の特徴について主な統計量とヒストグラムを作成したうえで、以下の観点から説明しなさい。
1)分布の範囲(最大と最小)　2)分布の中心　3)分布の散らばり　4)分布の歪み

R 入門（２）: 行列演算

１　準備

ホームページ上のデータ x.txt と y.txt を作業用フォルダにダウンロードしてください
R を起動して、作業フォルダを指定してください。
以下のコマンド部分を copy&paste してください

```
x <- matrix(scan("x.txt",sep=",",skip=1),10,2,byrow=T)
```

変数名など読み飛ばす行数を skip=で指定する
行列は，左の列から順に（上から下へ）埋められる。要素を上の行から順に（左から右へ）埋める場合は、引数 byrow=T を指定する

```
y <- matrix(scan("y.txt",skip=1),10,1,byrow=T)
```

行列は作ることもできる
単位行列　i <- diag(4)　　4×4 の単位行列
ゼロ行列　zero <- diag(0,3,4) 3×4 のゼロ行列　diag(0,3) 3×3 のゼロ行列
任意　　a <- matrix(c(1.1,2.1,3.1,1.2,2.2,3.2,1.3,2.3,3.3),3,3)

```
dim(x)
```

行列の次元の表示

```
x
```

行列の表示

```
b <- solve(t(x) %*% x) %*% (t(x) %*% y)
```

$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ の計算

転置行列　t(x)
行列の和、差、積　＋　－　%*%　　（　*　だと各要素の積になる）
逆行列　　solve(x)
x %o% y　外積　（outer(x,y）と同じ）
x %x% y　クロネッカー積
diag(3)　3×3 の単位行列を作る
eigen(X)　行列 X の固有値と固有ベクトルを求める.
qr(X)　行列 X の QR 分解を行う.
chol(X)　正値対称行列（エルミート行列）X のコレスキー分解を行う.
det(X)　行列 X の行列式を求める

```
yhat <- x%*%b
```

yhat=X'b　y の予測値

```
e <- y-yhat
```

e=y－X'b　残差

```
ss <- t(e)%*%e
```

e'e　残差２乗和

問題

１　消費関数の係数 **b** を求めなさい。
２　残差２乗和 **e'e** を求めなさい。
３　**e** と **X** の内積 **e'X** と **e** と **y**hat の内積 **e'y**hat はどんな行列になるか？

R 入門（３）

0　準備

ホームページ上のデータ cons.csv を z:¥R にダウンロードしてください。
R を起動して、作業フォルダを　z:¥R にしてください。

1　データの読み込み

```
d1 <- read.table("cons.csv",header = TRUE, sep =",")
d1
```

2　回帰分析

＞ **lm**(d1$cons~d1$income)　回帰分析
lm(d1$cons~0+d1$income)　切片なし回帰
lm(d1$cons~d1$income+ d1$year)　重回帰
lm(d1$cons~d1$income+ I(d1$income^2))　2 次関数
lm(d1$cons~d1$income+ I(d1$income*d1$year))　交差項
lm(log(d1$cons)~log(d1$income))　対数線形

※　lm(y~x1*x2)→y=a0+a1*x1+a2*x2+a3*x1x2　となるので注意！

＞ result<-**lm**(d1$cons~d1$income)　一番いい結果を result に保存する

＞ **summary**(result)　推定結果の表示

Residuals：残差の分布
Coefficients：係数の推定結果

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	係数の	標準誤差	t 値	P 値
Income	推定値 b	の推定値 s_b	$t=b/s_b$	

Residual standard error：残差 e の標準誤差 s
R-Squared：決定係数　Ajusted R-Squared：自由度修正済決定係数
F-statistic：すべての係数＝0 の検定統計量（F 値）

＞ yh=**predict**(result); e=**residuals**(result)　予測値と残差を保存
＞ **data.frame**(d1,yh,e)　データフレーム形式で表示
＞ **anova**(result)　分散分析表の表示

	自由度	２乗和	平均２乗和	F 値	P 値
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Income	k-1	ESS=$\Sigma(\hat{y}-\bar{y})^2$	ESS/(k-1)	$\frac{ESS/(k-1)}{RSS/(n-k)}$	
Residuals	n-k	RSS=Σe^2	RSS/(n-k)		

＞**plot**(d1$income,d1$cons)　X-Y プロット図
＞**abline**(result)　回帰直線を追加
＞**plot**(resid(result))　残差のプロット

演習

1. 消費関数の推計結果を以下の例を参考にしてまとめなさい
   消費＝38.2+0.815 所得　決定係数 R2=0.985
   (10.5) (3.25)　括弧内は t 値
   残差２乗和：582.4　標準誤差：0.254
2. **y**=$\hat{\mathbf{y}}$+**e** から **y'y**=$\hat{\mathbf{y}}'\hat{\mathbf{y}}$+**e'e** を導きなさい。Hint　$\hat{\mathbf{y}}$=**Py**、**e**=**y**－**Py**=**My**

R 入門（4）　モンテカルロ実験

100 組の乱数を作って回帰分析を行ったが、それを 1000 回繰り返すには LOOP を利用する必要がある。
※拡張機能のインストール
R を起動し、メニューから[パッケージ]－[パッケージのインストール]を選び、ダウンロードサイトを指定する（例えば Japan(Tsukuba)）→必要となるパッケージをインストール
Matrix(行列)　lmtest(線形回帰モデルの検定)　maxLik(最尤推定)　mlogit(多項ロジット)　systemfit(同時方程式推定パッケージ) etc

```
library(moments)                                    パッケージ moments の読み込み
m=100; n=1000
b=matrix(0,n,5)                                     0 を要素とする n ×5 の行列
for(i in 1:n){                                      n 回繰り返し(Loop)
x<-runif(m,0,10)                                    m 個の一様乱数
w<-rnorm(m,0,1)                                     m 個の正規乱数
y<-0.5+0.5*x+0.5*w                                  m 個の Y
s<-summary(lm(y~x))                                 回帰分析の結果を s に保存
b[i,1]<- s$coefficients[1,"Estimate"]               切片の推定値
b[i,2]<- s$coefficients[2,"Estimate"]               傾きの推定値
b[i,3]<- s$r.squared                                決定係数の推定値
b[i,4]<- s$coefficients[2,"t value"]                傾きの t 値
b[i,5]<- s$coefficients[2,"Std. Error"]             傾きの標準誤差
}
hist(b[,1],xlab="a",ylab="frequency",col="gray")    切片の分布
hist(b[,2],xlab="b",ylab="frequency",col="gray")    傾きの分布
summary(b[,2])                                      決定係数の基本統計量
sd(b[,2])                                           標準偏差
skewness(b[,2])                                     歪度
```

※lm で作られる結果のリストは str(s)で出力できます

※グラフを画像として保存したいときは

```
jpeg("graph1.jpg")
  plot(x,y)
dev.off()
```

のようにすれば良い
他にも pdf()　PDF ファイル、png()　PNG 画像、bmp()　BMP 画像、postscript()　PS 画像、pictex()　TeX ファイル　などがある。

実習
決定係数、傾きの標準誤差の分布はどのような形になるか？
ヒストグラムと主な統計量から①分布の形状　②分布の中心（平均、中央値）③分布の散らばり（最大、最小、標準偏差）④分布の歪み（歪度＞0 なら右スソが長い分布）の 4 点について述べなさい。

R 入門（補）　Rプログラムの実行（batch 処理）

Rプログラムの実行方法　①コマンドプロンプト上で逐次実行　②Rスクリプトの実行

(0) R を起動して、「ファイル」→「デイレクトリ変更」でデスクトップ上の¥R を指定しておく

1　Rスクリプトの作成
(1)「ファイル」→「新しいスクリプト」を選択すると, (2)「R Editor」が立ち上がります
(2) コマンドを書き込む
(3)R Editor 内をクリック後、「ファイル」→「保存」を選択し ファイル名(例えば ols.R) を入力し「保存(S)」をクリック
※(1)～(3)は notepad や秀丸等の editor を使っても構いませんが、ファイルの拡張子を##.R としてください

2　Rスクリプトの実行
(1)「ファイル」→「スクリプトを開く」を選択して, 実行したいファイルをダブルクリックする　→　R Editor が立ち上がります（この時点で修正も可能）
(2)R Editor 内をクリック後、「編集」→「全て実行」を選択する
　プロ r グラムにバグがあれば、書き直して、改めて実行する
※スクリプトの一部のみを実行したい場合は
　R Editor 内の実行したい部分をマウスでドラッグ（範囲指定）した後で「編集」→「カーソル行または選択中の R コードを実行」を選択する
(3)表示された結果を word 等に Copy&Paste する

3　作業の終了
(1)R Editor 内をクリック後、「ファイル」→「保存」を選択し、「保存(S)」をクリック
　うまくいったプログラムを保存する
(2)> の後に続けて「q()」をタイプしエンターキーを押す →『いいえ』をクリックする

4　すでに作成済みのRスクリプトを実行する
(0) 作業デイレクトリの変更でデスクトップ上の¥R を指定
(1) source('ols.R')

R 入門（５）　最尤法

テキスト　Poisson 分布の事例（サッカー10 試合のゴール数）

R で長いプログラムを実行する場合は **R Editor** を利用する
1)「ファイル」－「新しいスクリプト」で R Editor を起動
2) プログラムを入力する
3)「ファイル」－「保存」　mle.R という名前を付けて保存
4)「ファイル」－「スクリプトを開く」で mle.R を開く
5) 一括実行：「編集」－すべて実行

部分実行：実行したいコマンドをすべてドラッグして  をクリックする

```
y <- c(5,0,1,1,0,3,2,3,4,1)
n=10
lpoi<-function(p){
  lnL <- -n*p+log(p)*sum(y)-sum(log(factorial(y)))
return(-lnL);
}
est <- nlm(lpoi,0.5,hessian=TRUE)
est
```

関数
絶対値 abs()　平方根 sqrt()　総和 sum()　総積 prod()　階乗 factorial()
対数 log() log10()　指数 exp()　三角関数 sin() cos() tan()　逆関数 asin() acos() atan()

最小化問題を解く関数（Newton 法）
nlm(f, p, hessian = FALSE, iterlim=100)

引数:
f:最小化すべき関数、p:パラメータの初期値、iterlim:最大繰返し回数
hessian:もし 'TRUE' ならば最終段階での f の hessian が返される。

返り値：
minimum:f の最小値の推定値、estimate:f の最小値が得られたパラメータ値
gradient:最小値における gradient の値、iterations:実行される繰返し回数
hessian:f の最初値における hessian の値（要求された時だけ）
code:最適化過程の終了状態を示す整数値
1:gradient の相対値は零に近い。おそらく最小値が得られた。
2:引き続く繰返しが許容値以内。おそらく最小値が得られた。
3:最後の大局的ステップが推定値よりも低い点を見つけるのに失敗。
　推定値が近似的な局所的最小値であるか、steptol が小さ過ぎる。
4:繰返し回数が上限を越えた。
5:５回連続して最大ステップサイズ stepmax を超過。関数が下に非有界か、
　ある方向で有限値に上から漸近的に近付いているか、stepmax が小さ過ぎる。

(演習)

1　ある学生 i が傘を持ってくる(y_i=1)か持ってこない(y_i=0)はベルヌイ分布に従い、確率変数 y_i の確率関数は $f(y_i)=\theta^{y_i}(1-\theta)^{1-y_i}$ であるとする。無作為に選んだ N 人の学生に傘の有無を調査して($y_1, y_2,\ldots,y_n$)という結果を得たとき、この結果から θ の最尤推定量を求めよ。

(1)対数尤度関数 $\ln L(\theta; y_1,\ldots,y_n)$

(2) θ の最尤推定量を求めよ

(3)Fisher の情報量 $I(\theta)$から θ の MLE の分散を求めよ。

(レポート)　y_i は Weibull 分布 $f(y)=\alpha\beta y^{\beta-1}\exp(-\alpha y^{\beta})$ に従うとき

(1)　n 個の標本($y_1,\ldots,y_n$)の対数尤度関数を求めよ

(2)　α と β の最尤推定量を求めよ

```
y <- c(1.3043,0.49254,1.2742,1.4019,0.32556,0.29965,0.26423,
1.0878,1.9461,0.47615,3.6454,0.15344,1.2357,0.96381,
0.33453,1.1227,2.0296,1.2797,0.96080,2.0070)
n=length(y)                                   データ数
lweib<-function(p){
  a <- p[1]
  b <- p[2]
  lnL=n*log(a)+n*log(b)+(b-1)*sum(log(y))-a*sum(y^b)
  return(-lnL);
}
est <- nlm(lweib,c(0.5,0.5),hessian=TRUE)
eigen(est$hessian)                            固有値はすべて正（正定値符合行列）
im <- solve(est$hessian)                      情報行列
```

R 入門（６） 仮説検定

Cobb-Douglas 生産関数：$\ln Y=\beta_1+\beta_2 \ln L+\beta_3 \ln K+\varepsilon$

1 次同次性：$\beta_2+\beta_3=1$

Translog 生産関数：$\ln Y=\beta_1+\beta_2 \ln L+\beta_3 \ln K+\beta_4\{(\ln L)^2/2\}+\beta_5\{(\ln K)^2/2\}+\beta_6 \ln L \ln K+\varepsilon$

1 次同次性：$\beta_2+\beta_3=1$、$\beta_4+\beta_5+2\beta_6=0$

```
data<-read.table("http://www.stern.nyu.edu/~wgreene/Text/Edition6/TableF5-2.txt",header=T)
n=nrow(data)                                  行列の行数
lnv=log(data[,2]); lnl=log(data[,3]); lnk=log(data[,4])
lnl2=lnl^2/2; lnk2=lnk^2/2; lnlk=lnl*lnk
esttl <- lm(lnv ~ lnl + lnk + lnl2 + lnk2 + lnlk)      Translog 推定
estcd <- lm(lnv ~ lnl + lnk)                           Cobb-Douglas 推定
```

 # Wald Test（Linear Hypothesis Test：Rb=r）

```
library(car)                                  ライブラリ car の読み込み
r<-c(0,0,0)
R<-rbind(c(0,0,0,1,0,0),c(0,0,0,0,1,0),c(0,0,0,0,0,1))
lht(esttl,R,r)                                (1)      F 検定（デフォルト）
lht(esttl,R,r,test="Chisq")                   (1')     χ2 検定
# 1 次同次の検定
r<-c(1)
R<-c(0,1,1)
lht(estcd,R,r)                                         F 検定
```

問題 Translog で 1 次同次性 Ho: $\beta_2+\beta_3=1$, $\beta_4+\beta_5+2\beta_6=0$ の検定を行え

 # Fitted Test

```
eeu<-deviance(esttl); dfu=esttl$df.residual
eer<-deviance(estcd); dfr=estcd$df.residual
fv <- (eer-eeu)/(dfr-dfu)/(eeu/dfu)
pfv <- 1 - pf(fv,dfr-dfu,dfu)                          F 分布の Pr(F<fv)
pfv
anova(estu,estr,test="F")                     (2)
anova(estu,estr,test="Chisq")
```

(1) Cobb-Douglas の妥当性 Ho: $\beta_4=\beta_5=\beta_6=0$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	21	0.67993				
2	24	0.85163	-3	-0.17171	1.7678	0.1841

(1') Cobb-Douglas の妥当性

	Res.Df	RSS	Df	Sum of Sq	Chisq	Pr(>Chisq)
1	21	0.67993				
2	24	0.85163	-3	-0.17171	5.3033	0.1509

(2)1 次同次の妥当性

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	21	0.67993				
2	24	0.85163	-3	-0.17171	1.7678	0.1841

R 入門（7）　操作変数法

C：実質民間最終消費支出,G：実質政府支出,Y：実質ＧＤＰ,TW：世界貿易量,R：公定歩合

消費関数　$C=\alpha+\beta Y$　　操作変数：G, TW, R, T

まずは systemfit を install する
package→package の install→download site の選択「例えば japan(Tsukuba)」→systemfit の選択

では R を立ち上げて

```
cons<-read.table("http://www.econ.keio.ac.jp/staff/hk/ecm2/data/consinst.txt",skip=1,he
ader=T,sep=",")                          データの読み取り
attach(cons)                             cons$変数名としなくてもよい
# OLS の実行
ols<-lm(C~Y)                             普通の最小２乗法
# iv check (1)：Relevance 検定
cor(cons)                                相関係数
iv1<- lm(Y~G+TW+R)
anova(iv1)                               分散分析＝F 検定
```

[質問１]　G, TW, R は Y と高い相関があると言えるか？F test はクリアするか？

```
# 2 段階最小 2 乗法(library(sem);  tsls(C~Y,~G+TW+R))も可）
library(systemfit)
model<-C~Y                               モデルの指定、複数指定できる
inst1<-~G+TW+R                           操作変数の指定
tsls<-systemfit(model,"2SLS",inst=inst1)　２段階最小２乗法
tls<-summary(tsls)
```

[質問 2]　OLS と 2SLS の結果を比較せよ（係数、標準誤差、決定係数）

```
# 過剰識別検定
overid<-summary(lm(tls$residuals$eq1~G+TW+R))　　補助回帰
J<-3*overid$fstatistic[1]                　　J 検定
1-pchisq(J,3)                            　　P 値=1－Pr(χ2(3)<J)
```

[質問 3]過剰識別検定の結果から、選択した操作変数は適切であると言えるか？

```
# Hausman 検定
hausman.systemfit(tsls,ols)              Hausman 検定
```

[質問 4] Hausman 検定の結果を見て、Ho:E(X, ε)=0 を検定しなさい
　　Wu 検定を行い、Ho:E(X, ε)=0 を検定しなさい

R 入門（8）　HCSE、GLSE

クレジットカード支出
Avgexp = Avg. monthly credit card expenditure,　Income = Income, divided by 10,000 ,
MDR = Number of derogatory reports　Acc = Credit card application accepted (1=yes),
Age = Age in years+ 12ths of a year,　Ownrent = individual owns (1) or rents (0) home.
Selfempl = Self employed (1=yes, 0=no)

```
credit<-read.table("http://web.econ.keio.ac.jp/staff/hk/ecm2/data/F8-1a.txt",skip=1,head
er=T,sep=",")                                          データ読み込み
attach(credit)                                         変数名を簡略化
# OLS の実行
ols<-lm(Avgexp~Income)                                 OLS 推定
# 分散不均一性のチェック(1) Graph
plot(Income,residuals(ols),xlab="Income",ylab="residuals")
abline(h=0)
plot(fitted(ols),residuals(ols),xlab="Fitted",ylab="residuals")
abline(h=0)
# 分散不均一性のチェック(2) GQ test、BP test
library(lmtest)
gqtest(ols,order.by=Income)        Goldfeld-Quandt test
zlist<-~Income+I(Income^2)         変数リストの定義
bptest(ols,varformula=zlist)       BP test 被説明変数:student 化残差
                                   varformula を指定しないと、モデルの説明変数と同じ
```

※ White test は white.test()　in tseries
※ 系列相関のテストは、bgtest(ols,order=1,type="Chisq")、dwtest(ols)
[質問 1]　分散不均一性は存在するといえるか？（根拠も示せ）

```
# 分散不均一性への対処(1) 変数変換、変数追加
log<-lm(log(Avgexp)~log(Income))
gqtest(log,order.by=Income)
add<-lm(Avgexp~Age+Ownrent+Income+I(Income^2))
gqtest(add,order.by=Income)
```

[質問 2]　変数変換や変数の追加で分散不均一性は解消されたか？（根拠も示せ）

```
# 分散不均一性への対処(2) Robust Estimation(HCSE)
library(sandwich)
coeftest(add)
coeftest(add,vcov=vcovHC(add,type="HC0"))
```

※　HACSE は　vcovHAC　を利用する
[質問 3]　各係数の標準誤差 SE の推計値はどう変わったか？

```
# 分散不均一性への対処(3) GLS
library(nlme)
model<-Avgexp~Age+Ownrent+Income+I(Income^2)
gls(model,weights=varFunc(~Income))
gls(model,weights=varFunc(~Income+I(Income^2)))
```

[質問 4]　GLSE によって結果はどう変わったか？