R と RStudio

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 I E03(2019-06-07)

最終更新: Time-stamp: "2019-08-01 Thu 10:15 JST hig"

今日の目標

- R が扱いやすい CSV ファイルを C で出力できる
- R と RStudio による解析方法と, Excel による解析方法の違いが説明できる
- C や Excel と対比して R のデータフレームの基本的な計算方法を説明できる

http://hig3.net

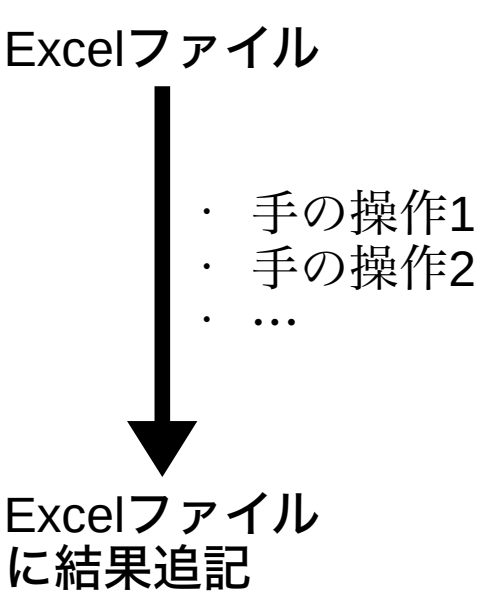
R とは?

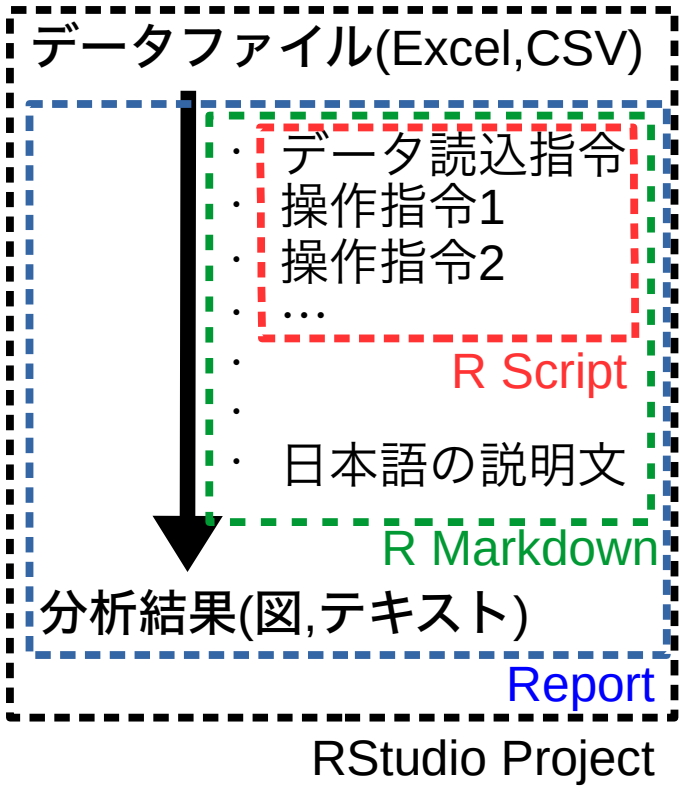
R は統計的計算と可視化のための言語と環境. 無料. オープンソース.
統計学分野で主流. データサイエンス分野で Python とシェアを二分.

`https://www.r-project.org`

`https://www.rstudio.com`

Excel と R の解析手順の比較





C:Visual Studio = R:RStudio

- C,R: 言語

C	R

.c
コンパイル言語
/*コメント*/
a=b;/*代入*/
a==b /*等式*/
a==b && c==d /*論理積*/
a==b || c==d /*論理和*/
	.R (R Script) .Rmd (R Markdown)
	インタプリタ言語, スクリプト言語,
	#コメント
	a<-b #代入
	a==b #等式
	a==b & c==d # 論理積
	a==b | c==d # 論理和

インタプリタとはコンパイルせずにそのまま実行できる処理系. 1 行ずつ実行してその結果を確認できる.

- Visual Studio, RStudio: 統合 (開発) 環境. ソースファイル, データファイル, 設定を, プロジェクトという単位で管理する.

この授業で R に与える CSV ファイルのフォーマット

CSV ファイルの読込方法

```
d<-read.csv("filename.csv",comment.char="#")
```

期待する CSV ファイルのフォーマットこの授業内では, 出力の仕様にそう書かれてる.

```
# (指定の)コメント0行以上
# (指定の)コメント0行以上
x,y,z #ラベル
# (指定の)コメント0行以上
# (指定の)コメント0行以上
1,3,2
2,9,4
4,5,6
# 途中にコメントが来てもよい
4,2,7
...
```

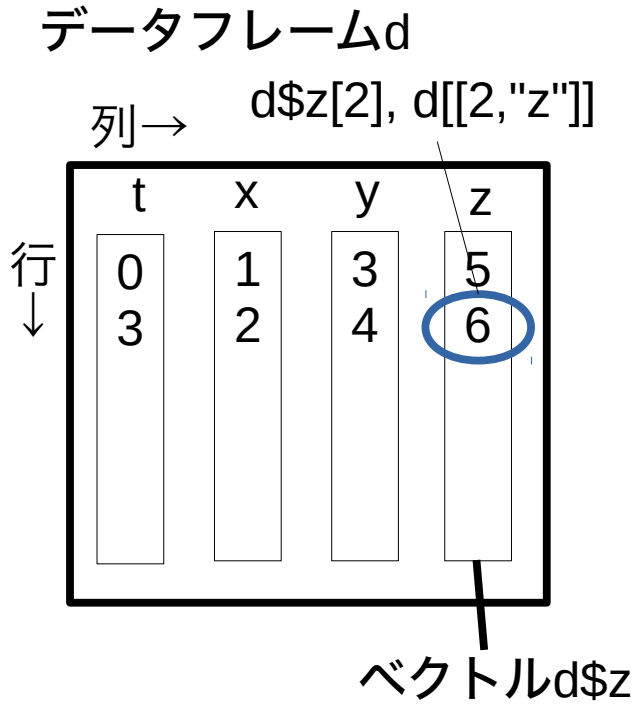
- コンマ改行区切り
- 全ての行には, 同じ個数の量.
- 同じ列には同じ意味の量 (同じ確率変数).
- 行方向は「繰り返し」特に 1 個のファイルが 1 個のサンプルを表すとき, 行の個数がサンプルサイズ.

R のデータフレームとオブジェクト

Excel の表 or CSV 形式

↓行, 列 →

t	x	y	z	(ラベル行)···
0	1	3	5	···
3	2	4	6	···
⋮				



- R のデータフレーム d に対して, d$列名, d[行指定，列指定]
- R でのデータフレーム中の列の指定 d$x は C の構造体 d.x に似てる. ただし, d$x はつねにベクトル (配列).

R の変数はすべてベクトル

R の

```
d$s<-d$a+d$b
d$t<-d$a*d$a*d$a
```

は, C で言えば

```
for(i=0;i<n;i++){
  s[i]=a[i]+b[i];
}
for(i=0;i<n;i++){
  t[i]=a[i]*a[i]*a[i];
}
```

規則的ベクトルの作成

```
v1<-c(1,5,-3) # 長さ1のベクトルの連結(concatenation)
v2<-1:10   # 1以上10以下, 公差1の等差数列をベクトルと見たもの
v3<-seq(1,10,3) # 1以上10以下, 公差3の等差数列をベクトルと見たも
```

1 変量の基本統計量/グラフを得る関数

データフレーム d の 列 r に対して,

- サンプルサイズ (列に含まれる行数) **length**(d**$**r)
- 標本平均値 **mean**(d**$**r)
- 標本期待値 $\phi(r)$ に相当する列を作って **mean**
- 標本比率 **ifelse** (条件, 真の返り値, 偽の返り値) などで標本期待値として計算するか, 条件を満たす行だけからなるベクトル d**$**r[d**$**r>2] を作って, **length** で行数を数える
- 不偏標本標準偏差 **sd**(d**$**r)
- 不偏標本分散 **var**(d**$**r)
- 四分位数 **summary**(d**$**r), **quantile**(d**$**r)
- ヒストグラム **hist** (d**$**r)