広島大学

データサイエンス基礎

Fundamental Data Science

第5回　R によるデータの視覚化

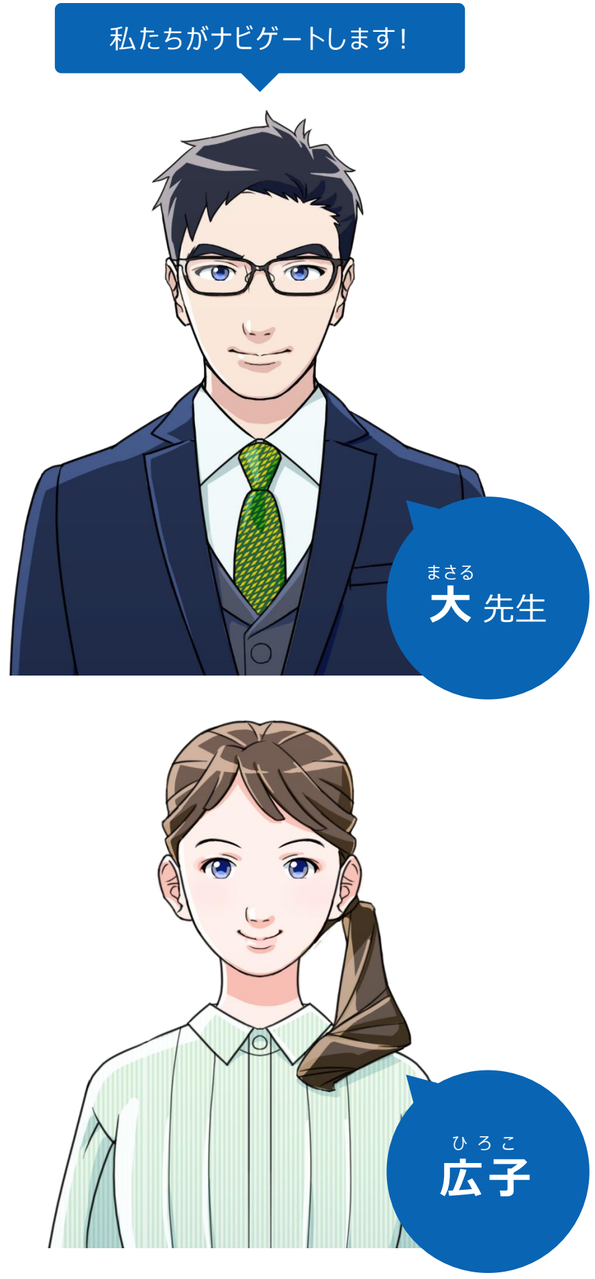
広島大学　数理統計学グループ

今日の内容

RStudio (クラウド版) を使って

- **RStudio の使い方**
- **データの読み込み, 抽出,
  要約統計量の計算**
- **グラフ (ヒストグラム, 箱ひげ図, 散布図)
  の描き方**

**※ インストール版 RStudio もクラウド版と使い方は基本同じ
違う箇所 (2か所) は適宜説明**

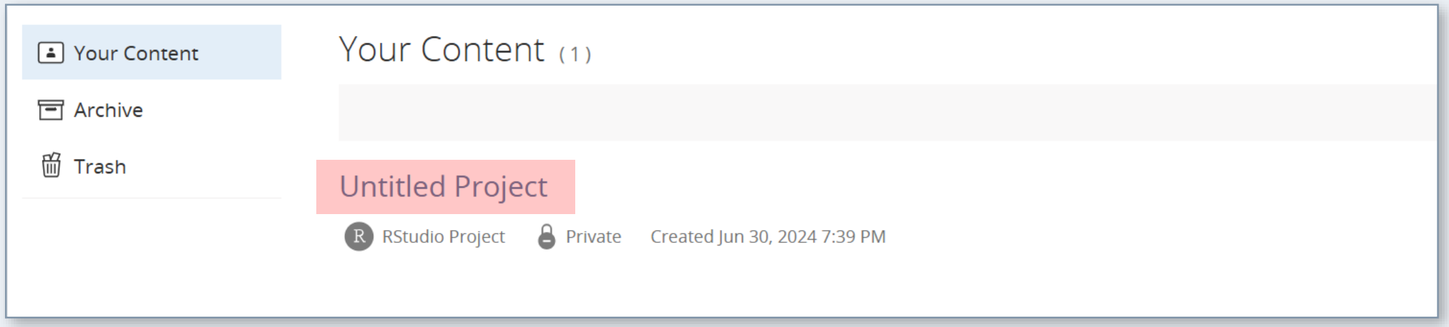


RStudio (クラウド版) の使い方
起動と計算

Posit Cloud の使用

Your Workspace 内の作成済みのプロジェクトをクリック

https://posit.cloud/



※ インストール版 RStudio を使う場合

Windows

スタートメニュー

⇒「RStudio」と検索

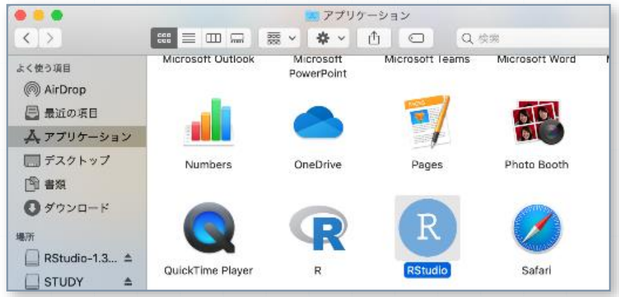
⇒ RStudio のアイコンをクリック

Mac

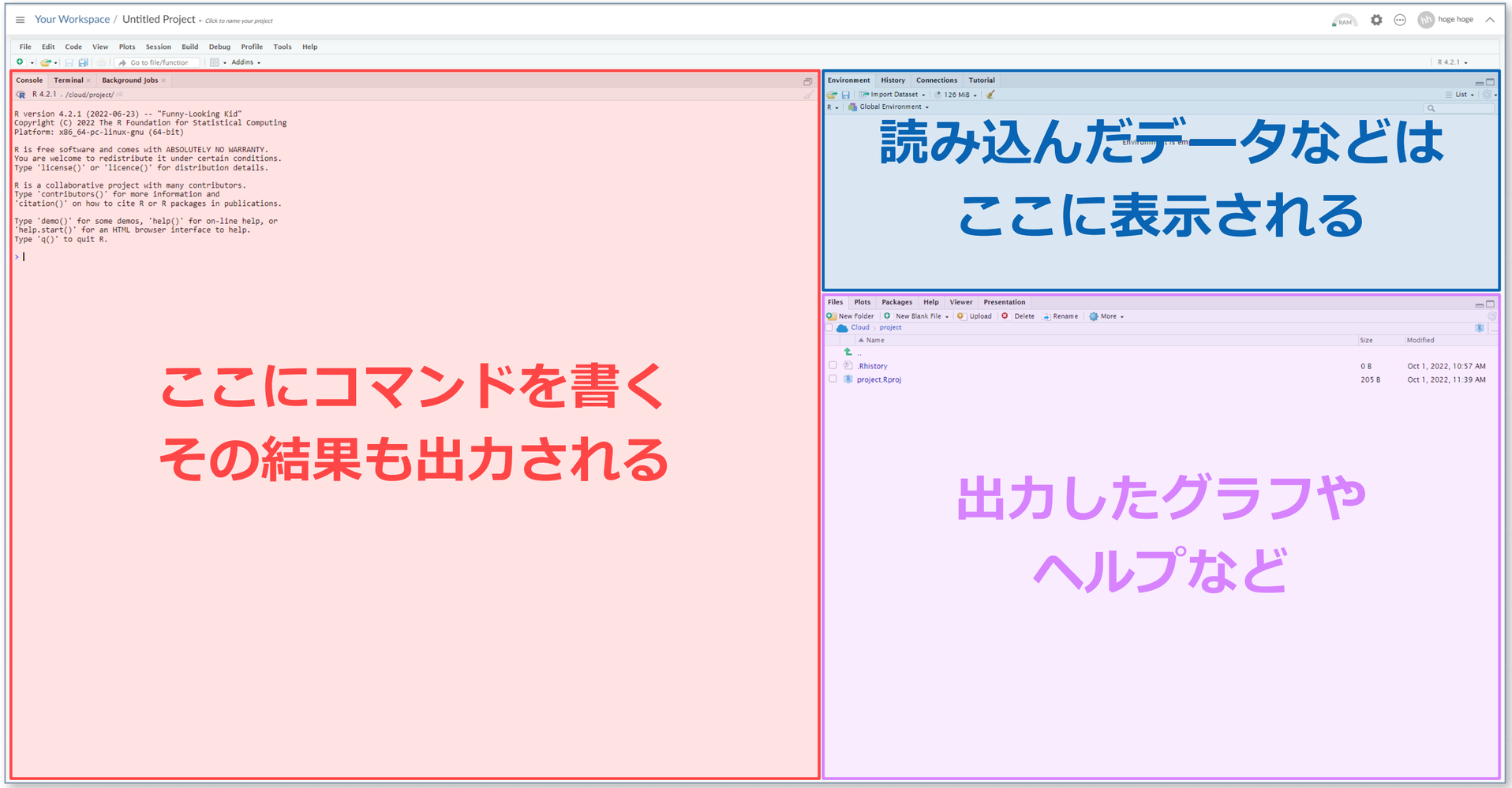
Finder

⇒ アプリケーション

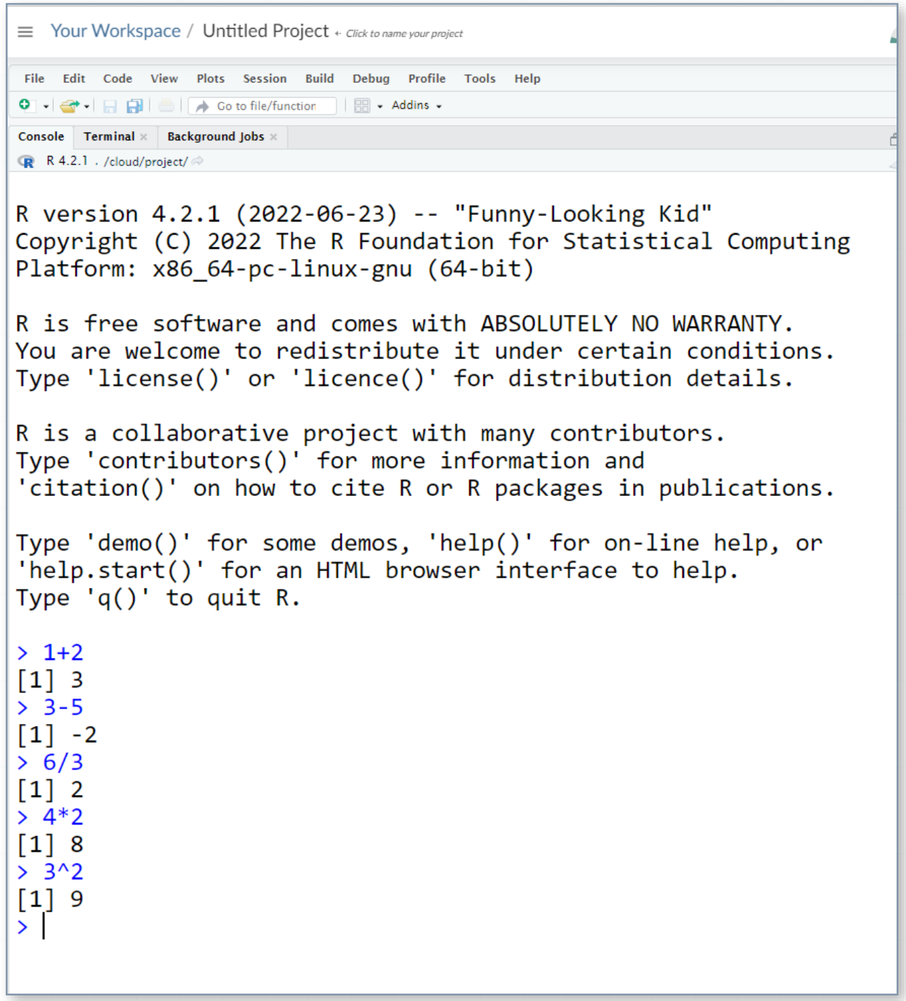
⇒ RStudio のアイコンをクリック



画面


Your Workspace / Untitled Project
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Background Jobs
R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
ここにコマンドを書く
その結果も出力される
Environment History Connections Tutorial
読み込んだデータなどは
ここに表示される
Files Plots Packages Help Viewer Presentation
.Rhistory
project.Rproj
出力したグラフや
ヘルプなど

基本計算




```
> 1+2
[1] 3
> 4*2
[1] 8
> 3-5
[1] -2
> 3^2
[1] 9
> 6/3
[1] 2
```

実行したいコマンドを半角でかいてEnter で実行

練習問題 (1)

Q.1 20 – 40 を計算せよ

Q.2 2の3乗を計算せよ

Q.3 100^1/2 と 100^(1/2) を計算してその違いを確かめよ

解答例: 練習問題 (1)

```
> 20-40
[1] -20
> 2^3
[1] 8
> 100^1/2
[1] 50
> 100^(1/2)
[1] 10
>
```

1

2

3

3

計算するとき,
() の位置に注意!!

100^1/2 は
100^1 を2で割る
という計算になります

100^(1/2) は
$\sqrt{100}$ と同じ計算ですね

データの読み込む前の下準備

成績データ

「seiseki.csv」

Data **中学2年生の成績データ** (個体数=166, 杉山, 藤越, 小椋, 2014, より抜粋).

1. ID number
2. 国語 (点)
3. 社会 (点)
4. 数学 (点)
5. 理科 (点)
6. 音楽 (点)
7. 美術 (点)
8. 体育 (点)
9. 技家 (点)
10. 英語 (点)

	A	B	C	D	E	F	G	H	I	J
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
2	1	30	43	51	63	60	66	37	44	20
3	2	39	21	49	56	70	72	56	63	16
4	3	29	30	23	57	69	76	33	54	6
5	4	95	87	77	100	77	82	78	96	87
6	5	70	71	78	67	72	82	46	63	44
7	6	67	53	56	61	61	76	70	66	40
8	7	29	26	44	52	37	68	33	43	13
9	8	56	54	37	59	35	64	53	67	7
10	9	45	21	7	44	16	52	34	46	3

データの整理

- すべての変数に半角英数字で変数名をつける
- 変数名は1行目に1行だけでつける
- データのタイトルなど，変数名とデータ以外の行はすべて削除する

※ これらをしないと, データが読み込めてもその後エラーが起こることが多い

RStudio で読み込むために
データの下準備を
覚えてください

各変数名

	A	B	C	D	E	F	G	H
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku
2	1	30	43	51	63	60	66	3
3	2	39	21	49	56	70	72	5
4	3	29	30	23	57	69	76	3

データの整理

おすすめの例

	A	B	C	D	E	F	G	H	
1	ID	kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gik
2	1	30	43	51	63	60	66	37	
3	2	39	21	49	56	70	72	56	
4	3	29	30	23	57	69	76	33	

- 変数名がすべて半角文字でついている
- 変数名とデータ (数値) 以外の行がない

おすすめしない例

	A	B	C	D	E	F	G	H		
	成績データ									
		kokugo	shakai	sugaku	rika	ongaku	bijutu	taiiku	gika	eigo
	1	30	43	51	63	60	66	37	44	20
4	2	39	21	49	56	70	72	56	63	16
5	3	29	30	23	57	69	76	33	54	6

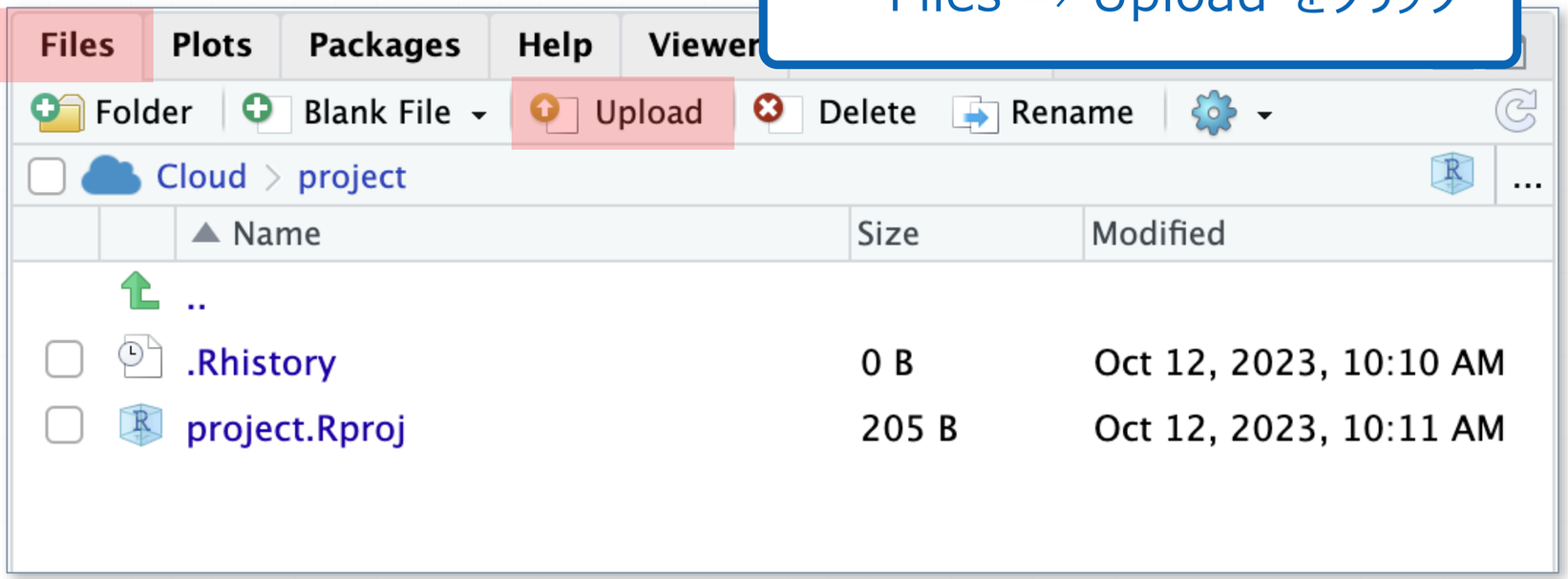
変数名とデータ (数値) 以外の行がある

変数名がついていない

RStudio (クラウド版) での下準備

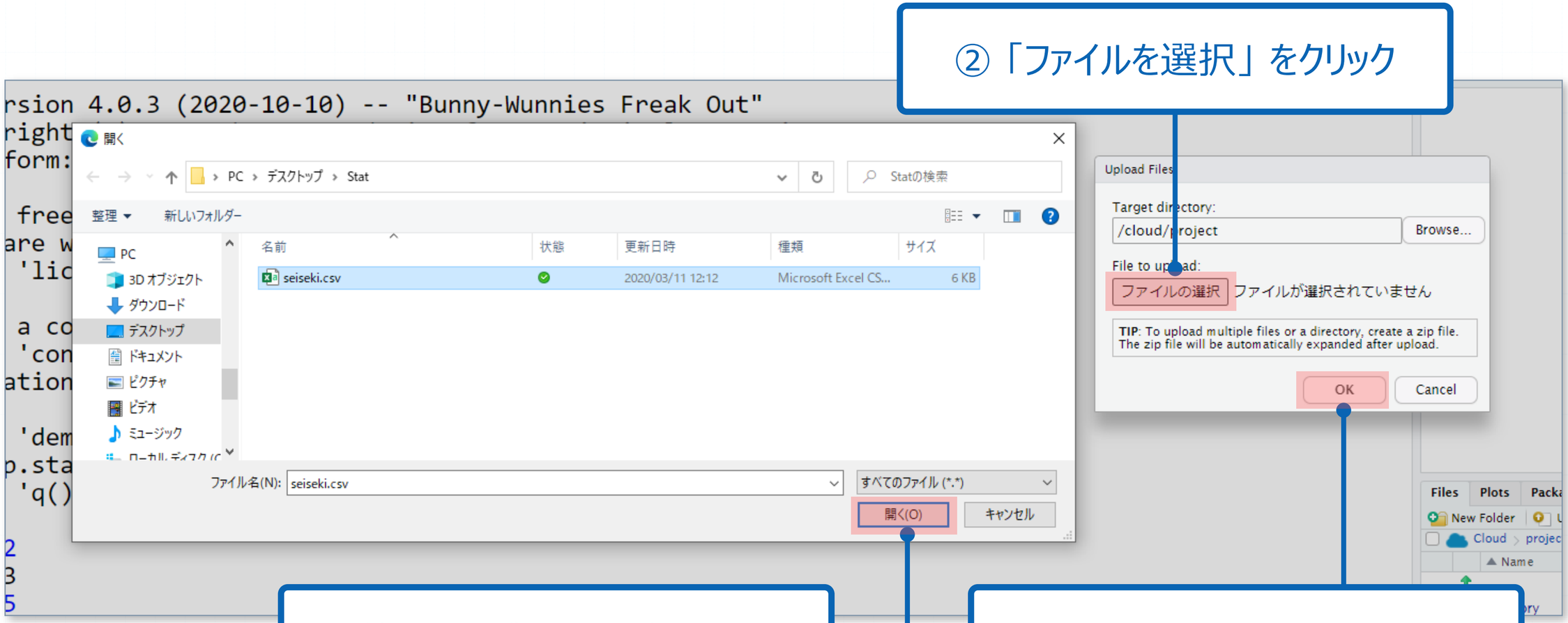
使いたいデータをクラウド上に Upload する
⇒ クラウド上でそのデータが使用可能に!!

① 画面右下
Files ⇒ Upload をクリック

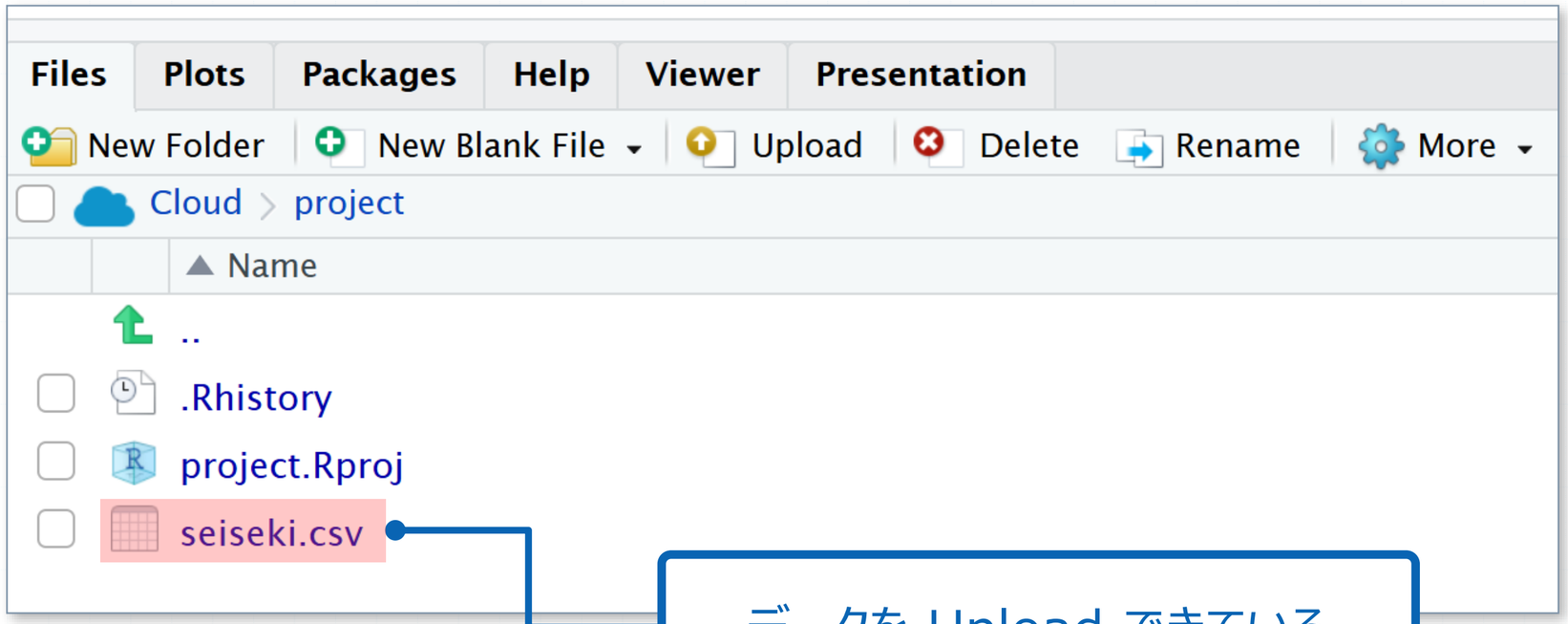


では，実際に
使うデータはどうやって
読み込むんですか?

RStudio (クラウド版) での下準備



RStudio (クラウド版) での下準備



データを Upload できている

ここまでがデータを
読み込むための下準備です

一度 Upload したデータはブラウザを閉じても残る
Upload は1回だけでよい

RStudio (インストール版) での下準備

ディレクトリの変更を行う!!

RStudio で**作業する場所 (ディレクトリ)** を指定する

読み込みたいデータが保存されている場所のこと.

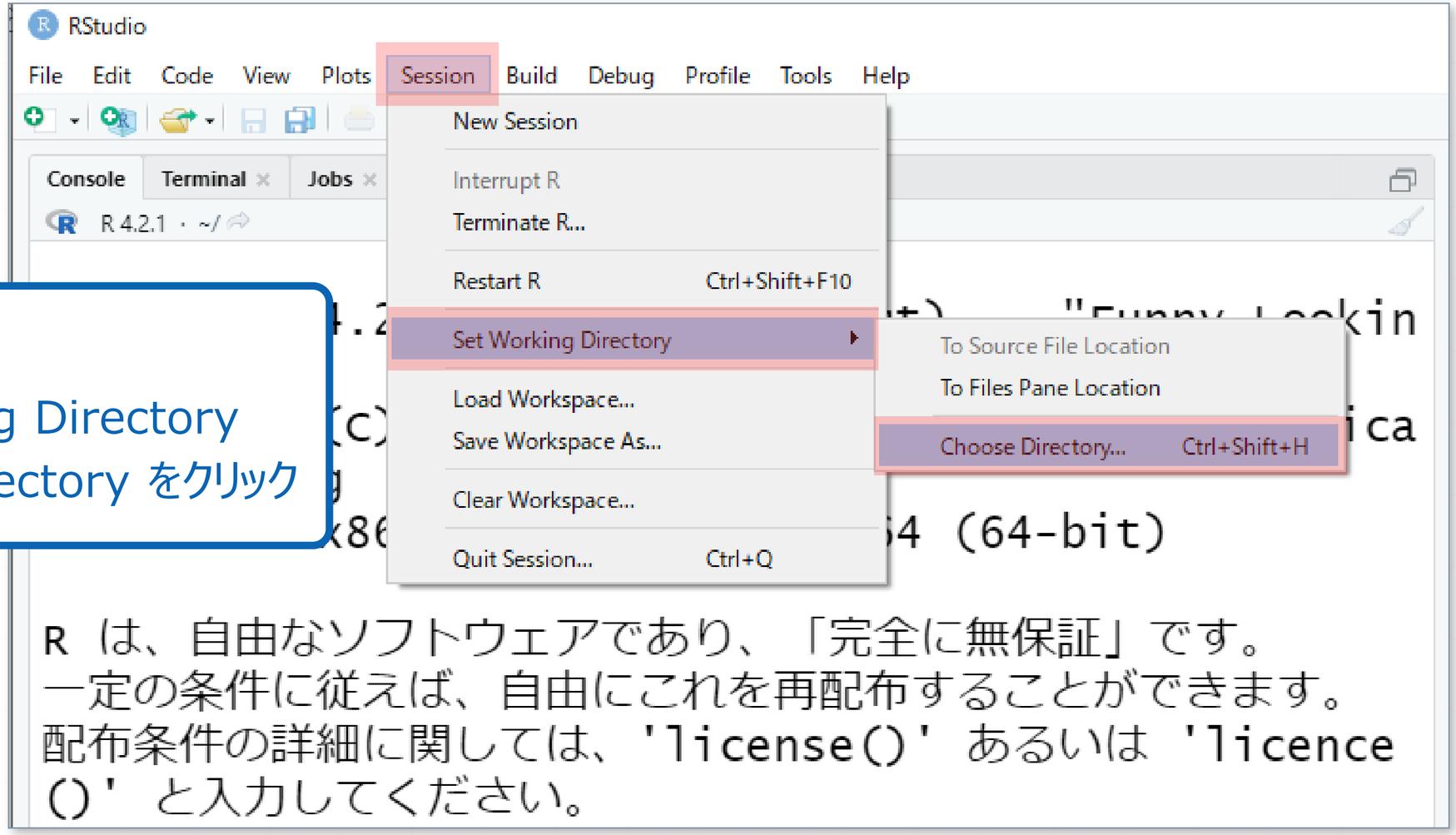
デスクトップ上に保存しているならデスクトップを,
あるフォルダに保存しているならそのフォルダを指定する!!

ここからは
インストール版の RStudio
でのやり方の解説です

- これをしないと, データが読み込めないことが多い
- RStudio を起動するたびに指定しないといけない

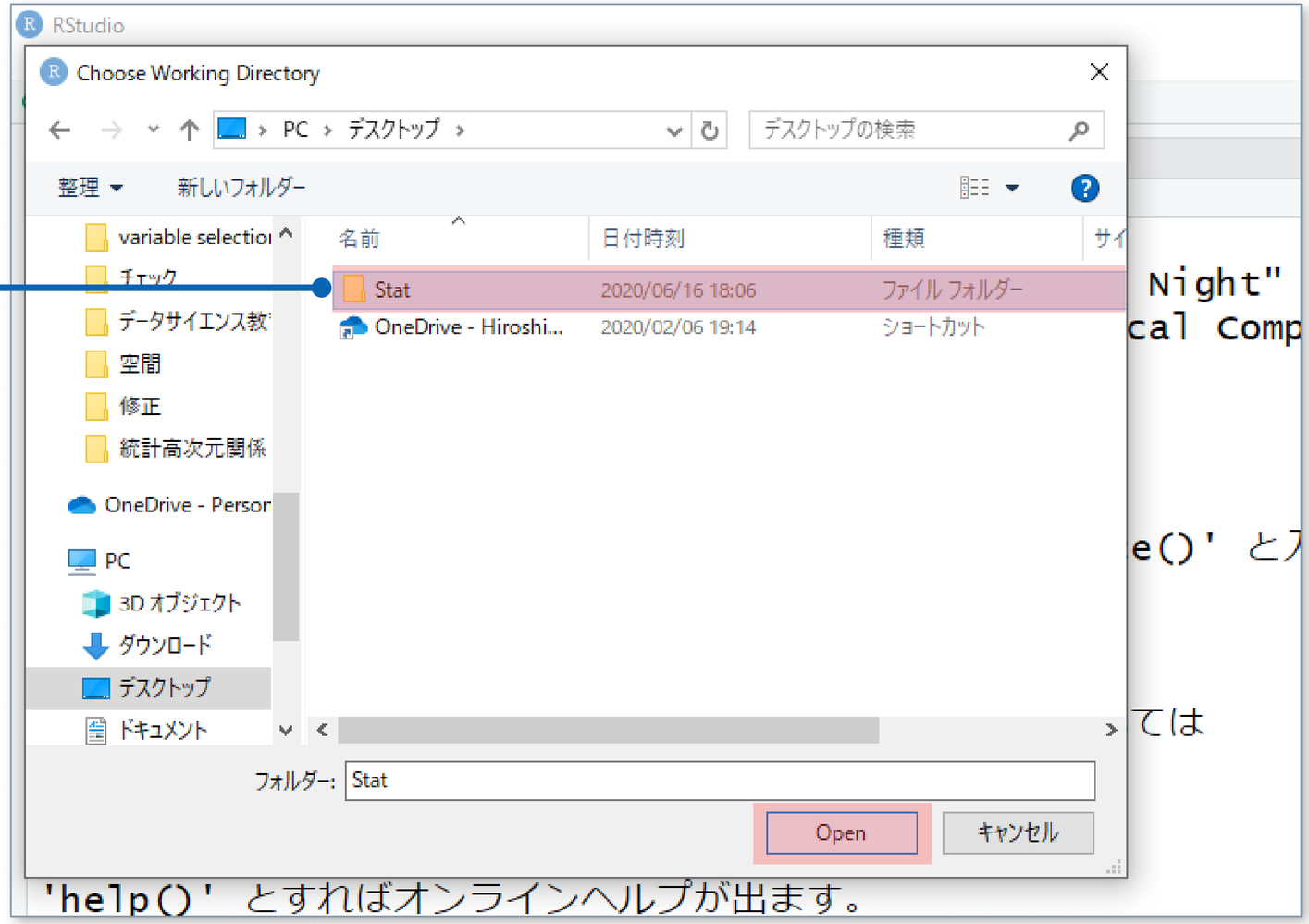
RStudio (インストール版) での下準備

① Session
⇒ Set Working Directory
⇒ Choose Directory をクリック



RStudio (インストール版) での下準備

② csvファイルが保存されている
フォルダを指定
⇒ Open



データの読み込みと抽出
要約統計量の計算

データの読み込み

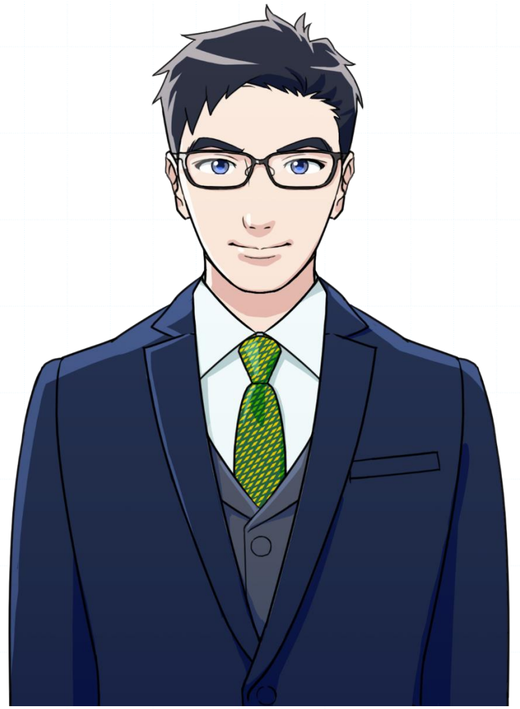
```
# seiseki.csv を読み込んでオブジェクト X に保存する.
X <- read.csv("seiseki.csv")
```

```
# X の先頭を確認する
head(X)
```

```
# X の変数名を確認する
colnames(X)
```

- #以降はコマンドとして認識されない
- もし, テキストファイルを読み込みたいなら,
  `read.table("ファイル名.txt")` とする

それでは
実際に成績データを
読み込んでいきますよ



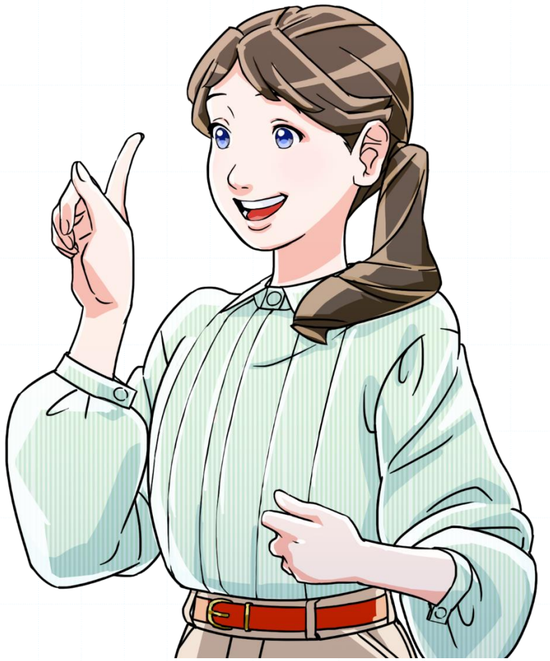
データの抽出

```
# X 内の指定した変数のデータを抽出する
X$kokugo
# 赤字の箇所は変数名をかく．kokugo のデータを取り出す．
```

```
X$shakai # shakai のデータを取り出す．
```

! 指定する変数名は colnames で確認すること

ここまでは
データの読み込みと
一部のデータを取り出す
方法でしたね



要約統計量の計算

```
# 平均値
mean(X$kokugo)
```

```
# 中央値
median(X$kokugo)
```

```
# 最小値，四分位点（ヒンジ法），最大値
fivenum(X$kokugo)
```

```
# ; を使えば一行でかける
mean(X$kokugo); fivenum(X$kokugo)
```

矢印キー (↑, ↓) でこれまで入力したコマンドを再利用できる

他にもいろんな
コマンドがあるので、
調べてみるのもいいですね

練習問題 (2)

Q.1 「seiseki.csv」のデータを読み込みオブジェクト X に保存せよ

Q.2 X の変数名を確認し, 数学データを抽出せよ

Q.3 数学データの平均値と中央値はどちらが大きいか

Q.4 国語データの四分位範囲を `fivenum` を利用して計算せよ

解答: 練習問題 (2)

Q.1 「seiseki.csv」のデータを読み込みオブジェクト X に保存せよ

Q.2 X の変数名を確認し, 数学データを抽出せよ

Q.3 数学データの平均値と中央値はどちらが大きいか

Q.4 国語データの四分位範囲を `fivenum` を利用して計算せよ

グラフを描く

※ 以下のようにデータを X に保存しておくこと

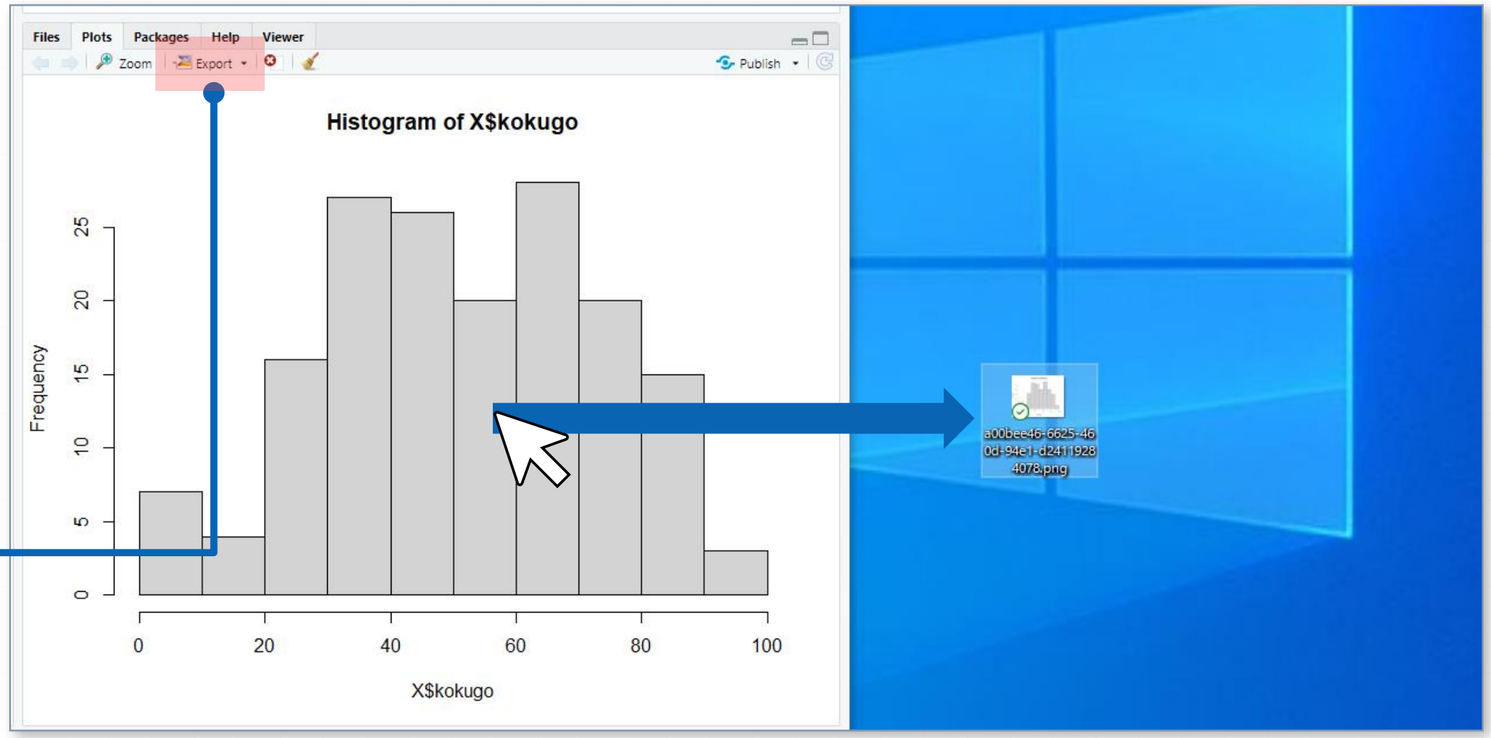
```
X <- read.csv("seiseki.csv")
```

ヒストグラムを描く

```
# ヒストグラムを描く
hist(X$kokugo)  # 国語データのヒストグラム
```

図をドラッグして
ドラッグ先に保存

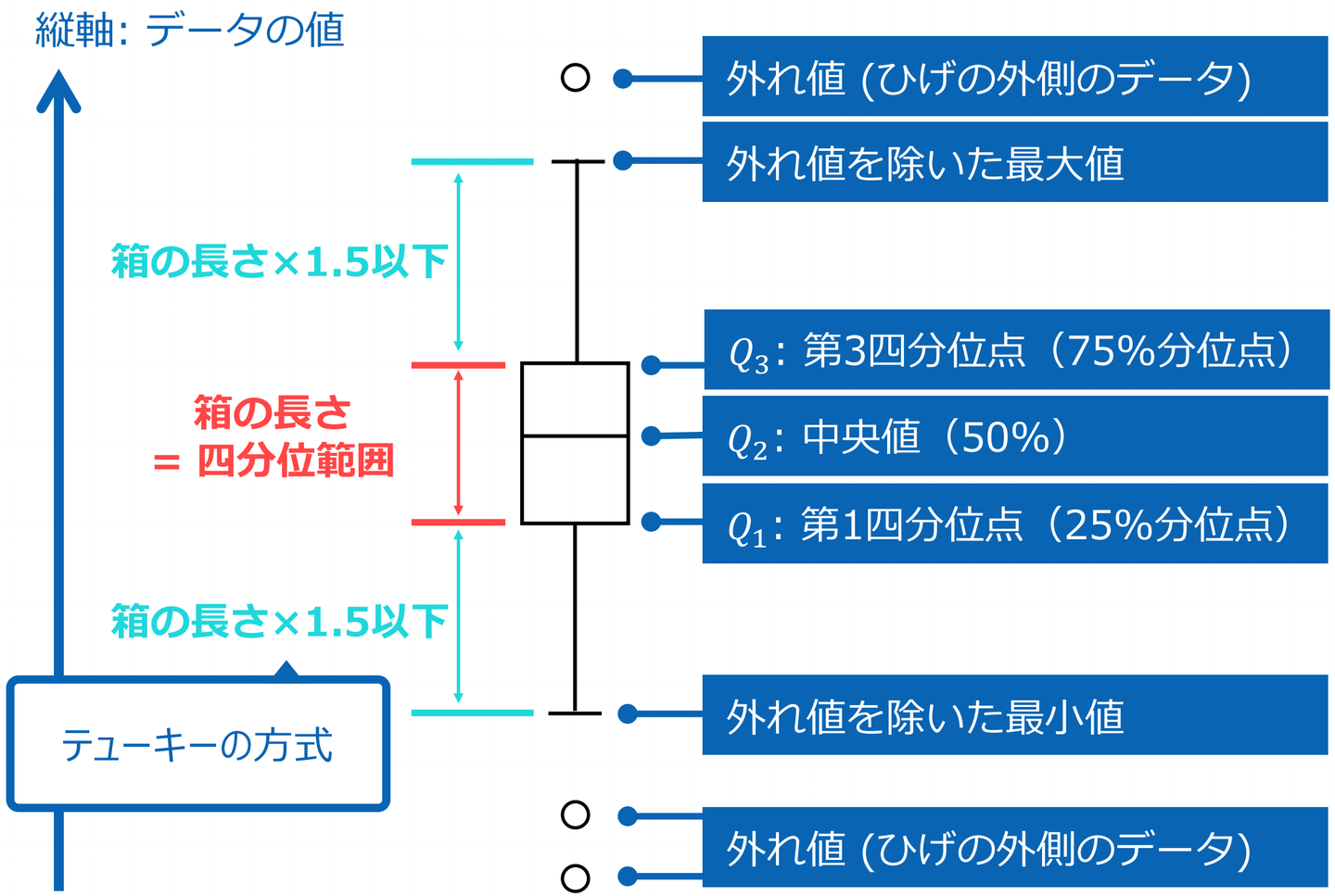
「Export」
→「Save as Image…」
　でも保存は可能



箱ひげ図

箱ひげ図
データの要点を
簡潔に表したグラフ

箱の長さ
＝ 散らばり具合



箱ひげ図を描く

```
# 箱ひげ図を描く
boxplot(X$kokugo)
```

```
# 箱ひげ図を並べて描く
boxplot(X$kokugo, X$shakai)
# 左から国語,社会の順で箱ひげ図が並ぶ
```

```
# 3つ以上も並べて描ける
boxplot(X$kokugo, X$shakai, X$sugaku)
# 描いた順と箱ひげ図の並びは同じ
```

実際に箱ひげ図を
かいてみましょう

練習問題 (3)

Q.1 理科データと英語データのヒストグラムをそれぞれ描け

Q.2 左から順に数学, 英語, 技家の箱ひげ図を並べて描き,
3科目のうち最もバラつきが大きそうな科目はどれか?

解答: 練習問題 (3)

Q.1 理科データと英語データのヒストグラムをそれぞれ描け

解答: 練習問題 (3)

Q.2 左から順に数学, 英語, 技家の
箱ひげ図を並べて描き,
3科目のうち最もバラつきが
大きそうな科目はどれか?

A

散布図

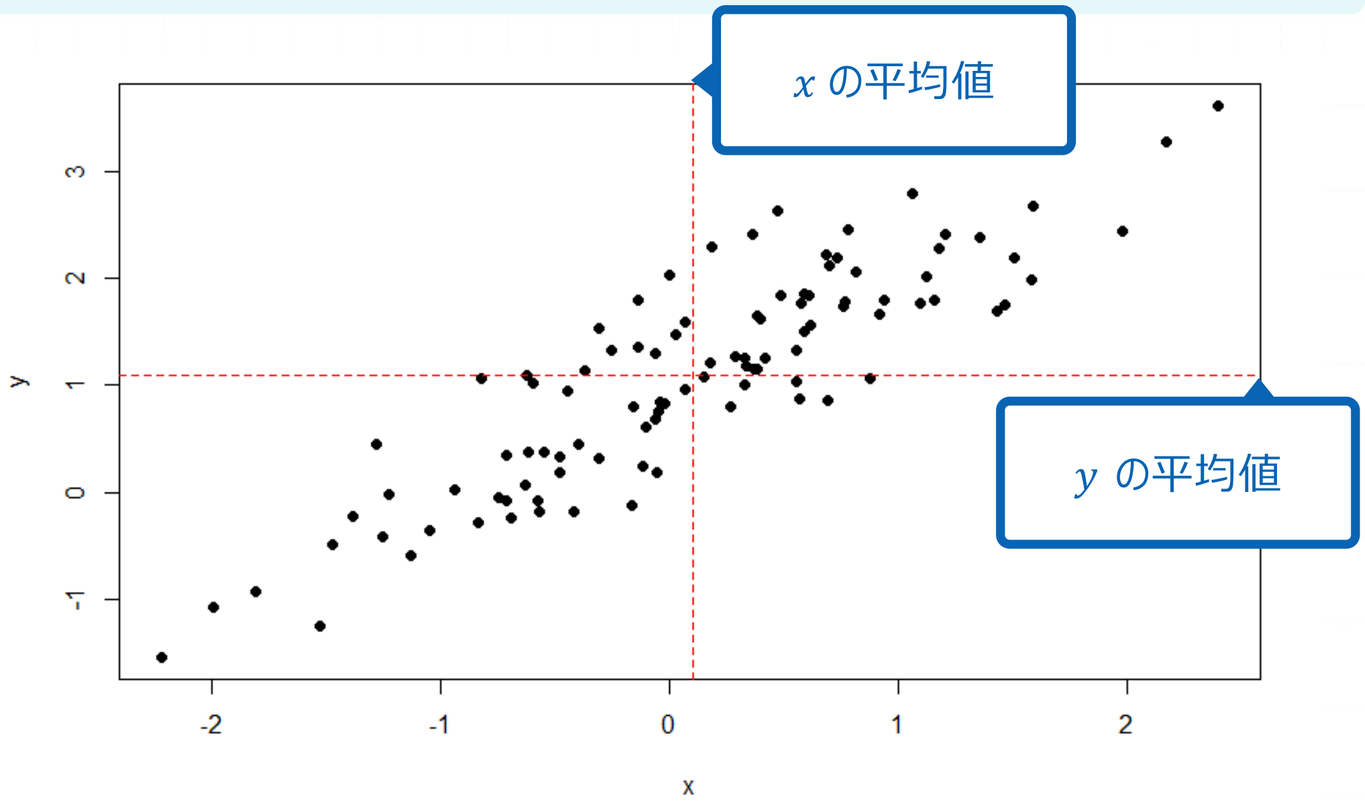
散布図

2つの変数 (x, y) の関係性を可視化したグラフ

❶ 2つの変数の平均値の直線を
引いて4つの区画にわける

❷ データ点が**右上, 左下**に多い
→ **右上がりの (直線) 傾向**

右下, 左上に多い
→ **右下がりの (直線) 傾向**



散布図を描く

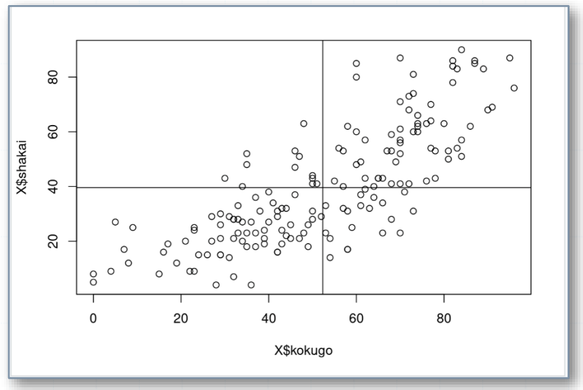
```
# 国語，社会データの散布図を描く
plot(X$kokugo, X$shakai) # 先に書いたほうが横軸．
つまり国語が横軸，社会が縦軸となる
```

```
# 散布図を描く
plot(X$kokugo,X$shakai)

# 散布図に横軸の平均値の直線を追加
abline(v=mean(X$kokugo)) # v= で横軸に垂直な直線を描く

# 散布図に縦軸の平均値の直線を追加
abline(h=mean(X$shakai)) # h= で縦軸に垂直な直線を描く
```

! abline は必ず plot の後に入力すること



データサイエンス基礎, R によるデータの視覚化: P.33

練習問題 (4)

Q.1 以下の2つの散布図を平均値の直線も追加して, それぞれ描け:

横軸が数学, 縦軸が英語の散布図
横軸が数学, 縦軸が体育の散布図

Q.2 1で描いた各散布図から, データの傾向を視覚的に判定せよ.

解答: 練習問題 (4)

横軸が数学, 縦軸が英語の散布図

A

解答: 練習問題 (4)

横軸が数学, 縦軸が体育の散布図

A

補足: オプション (マーカー, 色変更)

```
# 散布図のマーカーを変更
plot(X$sugaku,X$taiiku,
     pch=16 # マーカーを黒丸に指定 (デフォルト 1: 白丸)
     )

# 散布図に横軸の平均値の直線を追加
abline(v=mean(X$sugaku),
     col="red", # 直線の色を変更
     lty=2 # 直線を破線にする (デフォルト 1: 実線)
     )

# 散布図に縦軸の平均値の直線を追加
abline(h=mean(X$taiiku), # h= で縦軸に垂直な直線を描く
     col="red",
     lty=2
     )
```

今までの図だと
少し見づらいですよね…
見栄えを良くしましょう

今日のまとめ

- **RStudio の使い方**
- **データの読み込み, 抽出,
要約統計量の計算**
- **グラフ (ヒストグラム, 箱ひげ図, 散布図)
の描き方**
  - ヒストグラム: `hist`
  - 箱ひげ図: `boxplot`
  - 散布図: `plot`

※ オプション追加で色や線の太さなど変更可能
ネットで検索すれば, いろいろ出てきます

実践的になってきましたね!
次回も頑張りましょう