



# Encouraging use of R at NSTAC Japan

Ichiro Murata  
Assistant Director of National Statistics Center, Japan

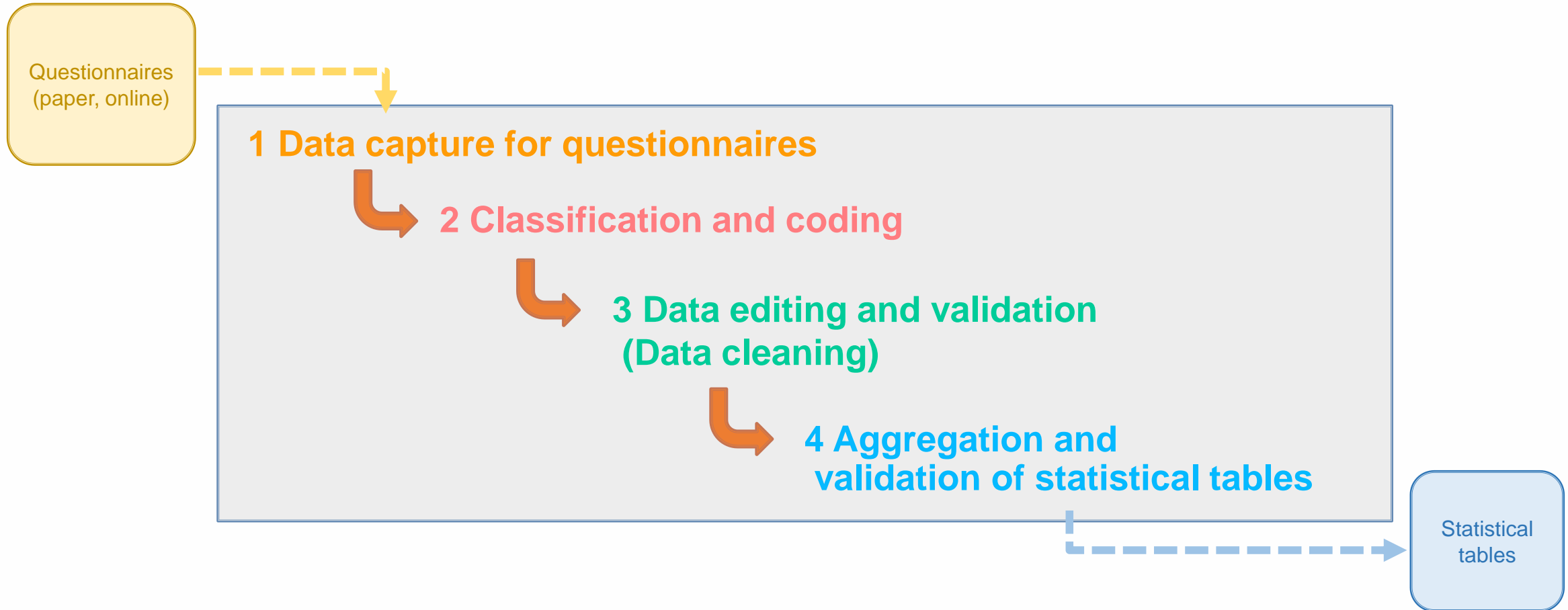
uRos2024, 27-29 November 2024, Piraeus

- ◆ Introduction - Motivation for use of R
- ◆ Infrastructure to support entering R
- ◆ Training and skill development

## ◆ Introduction

- Motivation for use of R

National Statistics Center (NSTAC) performs data conversion and data processing and tabulation of national survey / census questionnaires.



Our current environment:

- Windows PC for operations and several computing servers for batch processes
- Relational databases for data storage
- In-house development for main components
- Programming in Visual Basic .NET and C# (enhanced by in-house libraries)
- 400+ officers involved (including who engaged in the system development)

➡ Could our work be more efficient by leveraging R?

# i.e. The Awesome List of Official Statistics Software

The list includes many R packages, that frequently found on these processes:

- Sampling (GSBPM 4.1)
- Data integration and record linkage (GSBPM 5.1)
- Statistical data editing and imputation (GSBPM 5.3 | 5.4)
- Estimation and weighting (GSBPM 5.6 | 5.7)
- Output validation (GSBPM 6.2)
- Statistical disclosure control (GSBPM 6.4)

**Awesome official statistics software**

An item on this list is awesome because it is:

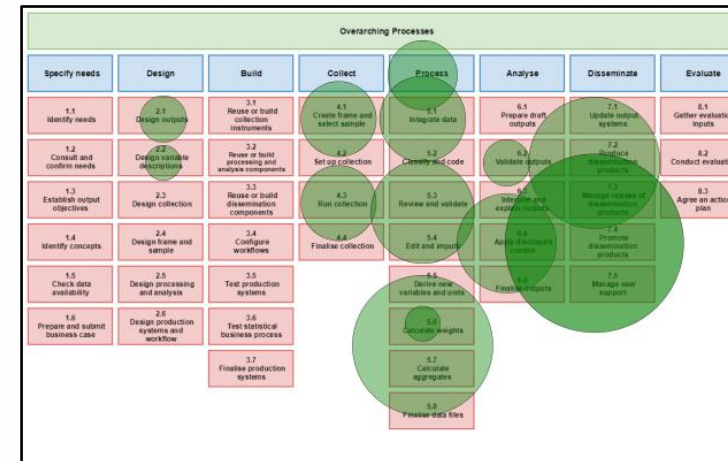
1. free, open source, available for download and
2. used in the production of, or provides access to, official statistics.

We prefer software that is easy to install and use and actively maintained. [Contributions](#) welcome.

News

- July 2024: The list promoted by UNECE HLG-MOS: [LinkedIn](#)
- Dec 2023: List presented at uRos2023: [abstract](#), [slides](#)
- Jun 2023: List highlighted in 71th CES conference

<https://github.com/SNStatComp/awesome-official-statistics-software>



Visualization by Olav ten Bosch

To take advantage of R, we need to start these things:

- Infrastructure and Tooling setup
- Training and Skill Development
- Identify Pilot Projects for Good Practices

# ◆ Infrastructure to support entering R

- Command Line Options

```
Rgui.exe --no-save --no-restore
```

↳ It will prevent users from unexpected Rdata handling

- Recommended R options

```
options(warnPartialMatchArgs=TRUE)  
options(warnPartialMatchAttr=TRUE)  
options(warnPartialMatchDollar=TRUE)  
options(showErrorCalls=TRUE)
```

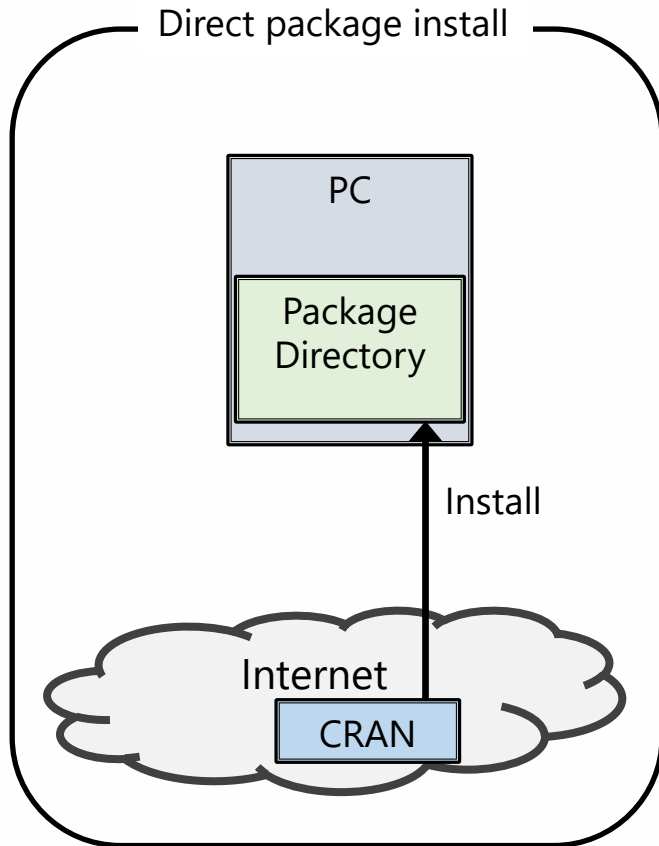
“warnPartialMatchArgs=TRUE” behavior

```
> round(5.55, digit=1)  
[1] 5.6  
Warning message:  
In round(5.55, digit = 1) :  
partial argument match of 'digit' to 'digits'  
>
```

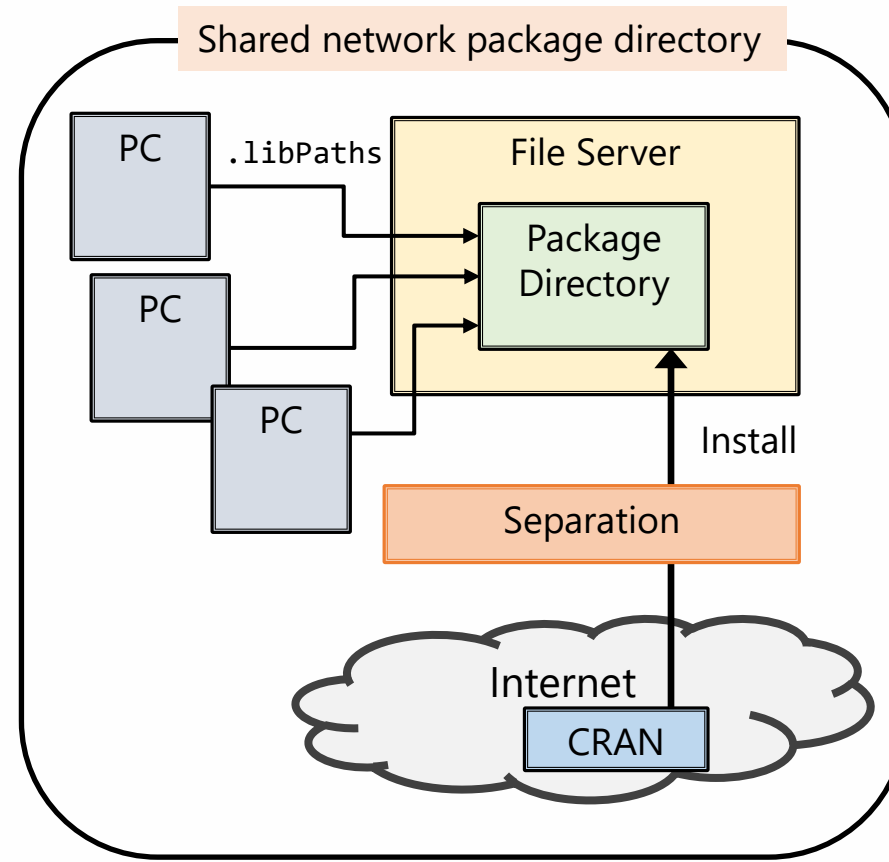
- The site-wide startup profile is a good place to implement for all employees (Providing the path to *R\_PROFILE* environment variable)
- Users can override the options through their own profile (Thanks to R startup sequence)

# Sharing the library folder

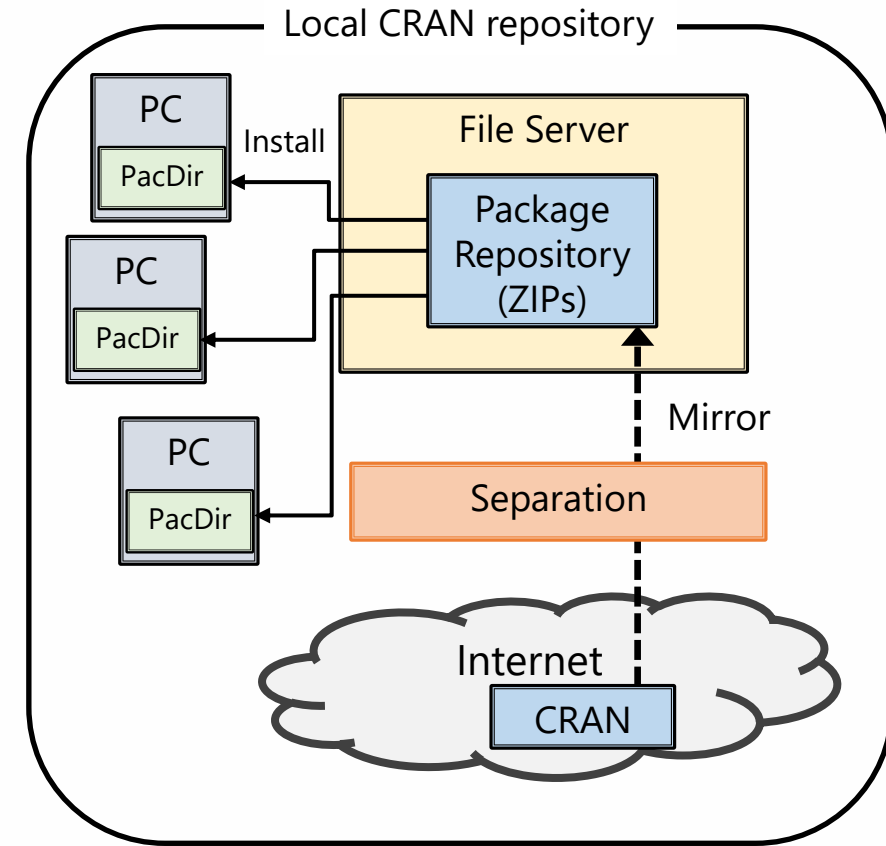
Considering how to use R packages in statistical organization...



× We are not allowed to connect the Internet directly



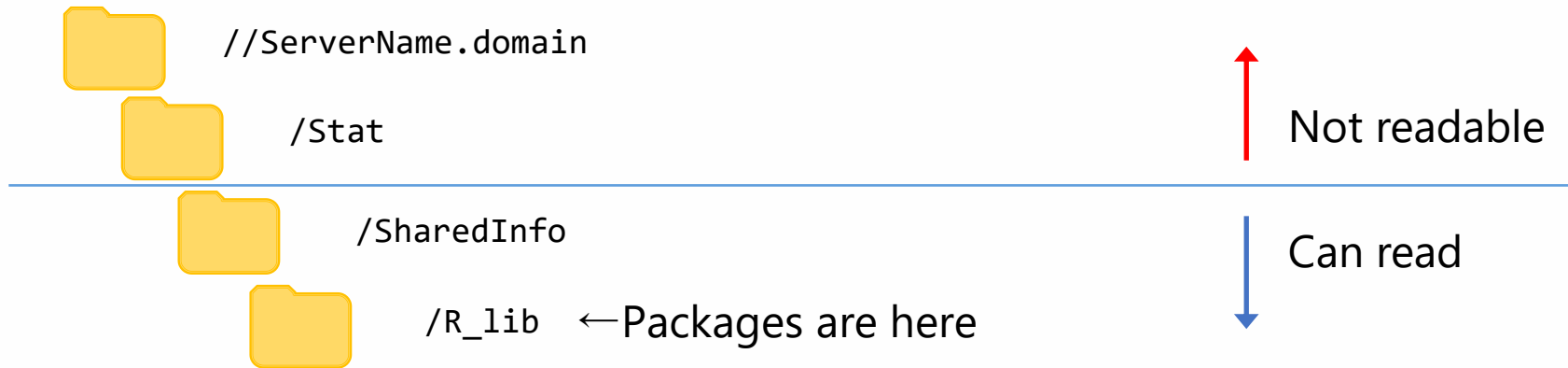
Our current implementation



...Under consideration

★ TIP: All users should be granted to access the shared package directory from **the root**.

- Error Case



```
> .libPaths("//ServerName.domain/Stat/SharedInfo/R_lib")
> library(dplyr)
Error in normalizePath(path.expand(path), winslash, mustWork) :
  path[1]="//ServerName.domain/Stat/SharedInfo/R_lib": Access Denied.
Calls: library -> normalizePath
>
```

■ Solution: Use network drive for shortcutting (in case of Windows OS)

```
> .libPaths("R:/SharedInfo/R_lib")
```

## ◆ Training and skill development

- Which data is adequate for R beginners?
  - “datasets” ([base](#))
  - “palmerpenguins” (known as an alternative to [iris](#))

These are easy to use, while we (NSTAC staffs) need more friendly and acceptable one (desirable if provided in Japanese)

- NSTAC has developed statistical datasets for education

One of the missions of NSTAC is to promote utilization of statistical data.

As a contribution for improvement of statistical literacy among people, NSTAC has developed a series of dataset from official statistics that is named:

SSDSE :  
Standardized Statistical Data Set for Education

SSDSE is basically composed of regional statistics

## Statistics (varies by types)

### Region

47 Prefectures  
or  
1741 Municipalities  
(as of 2024)

SSDSE-E-2024	Prefecture	A1101	A1102	A1301	A1302	A1303		L322102	L322109
年度		2022	2022	2022	2022	2022		2022	2022
地域コード	都道府県	総人口	日本人人口	15歳未満人口	15~64歳人口	65歳以上人口		住居費（二人以上の世帯）	教養娯楽費（二人以上の世帯）
R00000	全国	124947000	122031000	14503000	74208000	36236000		18645	26642
R01000	北海道	5140000	5098000	530000	2924000	1686000		24873	27234
R02000	青森県	1204000	1198000	123000	663000	419000		10541	20068
R03000	岩手県	1181000	1173000	125000	648000	408000		18814	25733
R04000	宮城県	2280000	2256000	258000	1363000	659000		22951	26516
R05000	秋田県	930000	926000	86000	484000	359000		13191	24327
R06000	山形県	1041000	1033000	113000	566000	362000		16140	22348
R07000	福島県	1790000	1776000	197000	1007000	586000		20888	25929
R08000	茨城県	2840000	2767000	321000	1655000	864000		18805	25789
R09000	栃木県	1909000	1865000	217000	1121000	572000		20869	24831
R44000	大分県	1107000	1092000	131000	600000	376000		18820	28260
R45000	宮崎県	1052000	1044000	136000	565000	352000		15433	21950
R46000	鹿児島県	1563000	1550000	201000	838000	523000		20822	22440
R47000	沖縄県	1468000	1446000	240000	884000	344000		25189	18429

From the document of SSDSE-E-2024

NSTAC has developed 6 type variations (SSDSE-A -- SSDSE-F)

Published at <https://www.nstac.go.jp/use/literacy/ssdse/> (Japanese only)

SSDSE (CSV / Excel)



- SSDSE-A (Social and Demographic Statistics by Municipality)
- SSDSE-B (Social and Demographic Time Series Statistics by Prefecture)
- SSDSE-C (Household Income and Expenditure by Prefecture)
- SSDSE-D (Time Use by Prefecture)
- SSDSE-E (Social and Demographic Statistics by Prefecture)
- SSDSE-F (Meteorological Observations by Prefecture)



Convert, select and rename

SSDSE (Rdata)



```
> load("SSDSE.Rdata")
> ls()
[1] "SSDSE_A" "SSDSE_B" "SSDSE_C" "SSDSE_D" "SSDSE_E" "SSDSE_F"
>
```

For the NSTAC staffs who are new to R, we have shared some documents for self-learning.

- Assuming no R knowledge
- Tabulation and visualization with corresponding simple command
- Pre-defined library path (through Rprofile.site)
- Executable on our own PC
- Using SSDSE for better understanding

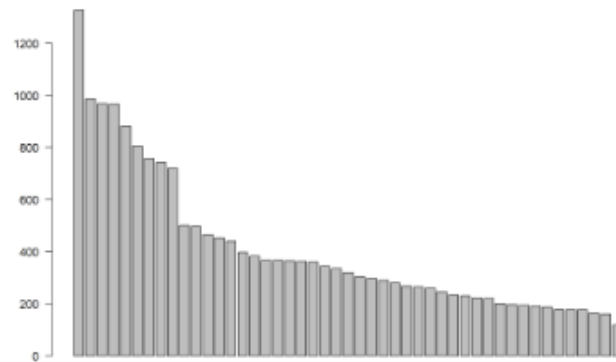
Our sharing documents composed by Quarto

## 統計センターにおけるRの始めかた: R in NSTAC

1.6 棒グラフ、ヒストグラム

小学校数 `E2101` の多い順に並べたものを、`barplot()` 関数を使って棒グラフで表示してみます。

```
barplot(schools_sorted$E2101, names.arg=schools_sorted$Prefecture, las=2)
```



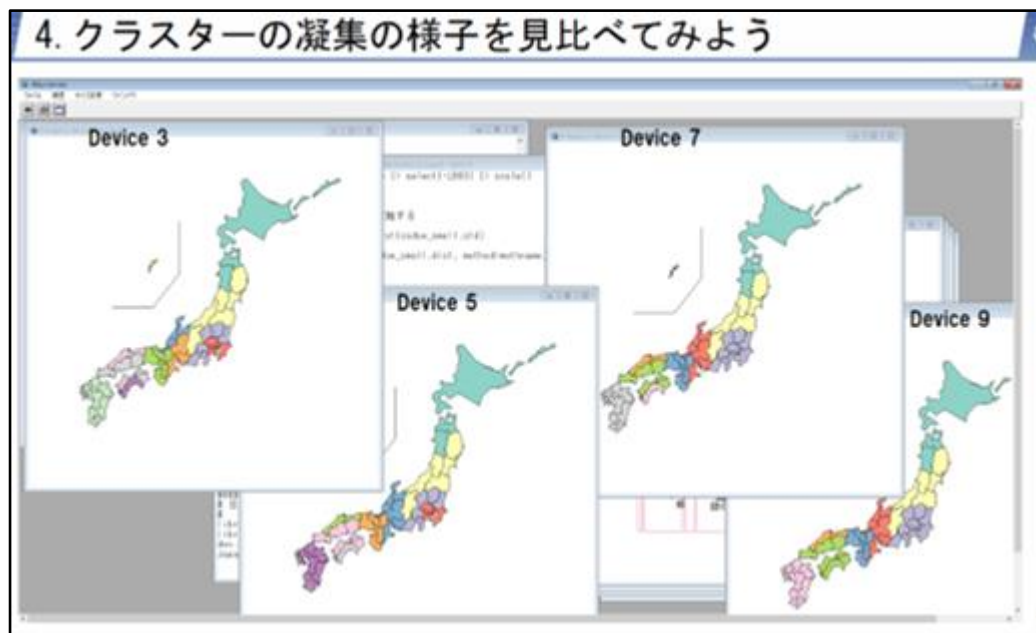
また、小学校数 `E2101` をいくつかの階級に分けて、それぞれに属する都道府県の数をヒストグラムで表示してみます。

```
hist(schools$E2101, xlab="小学校数", ylab="都道府県数")
```

目次

- 1.1 Rの起動
- 1.2 SSDSE (教育用標準データセット)
- 1.3 データの基本属性の確認
- 1.4 データの加工
- 1.5 基本統計量の確認
- 1.6 棒グラフ、ヒストグラム
- 1.7 集計
- 1.8 回帰分析
- 1.9 Rの終了

We have provided NSTAC staffs with multiple opportunities to take our R training course for recent years.



Training course:  
Regional food purchasing behavior characteristics (90 min)

### dplyr::select

`select()` は指定した列だけを取り出すときに使います。

```
select(SSDSE_E, 1:3) # 1~3列目だけを取り出す
select(SSDSE_E, AreaCode, Prefecture, A1101) # 列名を用いて1~3列目を取り出す
select(SSDSE_E, Prefecture:A4103) # 列Prefectureから列A4103までの全ての列を取り出す
```

列名の指定は上のように行うほかに、次のヘルパー関数を用いる方法も使えます。

- `starts_with()`
- `ends_with()`
- `contains()`
- `num_range()`
- `where()`

```
select(SSDSE_E, starts_with("H1")) # H1で始まる列名を取り出す
select(SSDSE_E, num_range("L3221", 1:2, width=2)) # L3221で始まり、1~2を0補足2桁でつけた
```

### dplyr::slice, dplyr::filter

`slice()` は行番号を指定して行を取り出すときに使います。

```
slice(SSDSE_E, 1:5) # 1~5行目だけを取り出す
slice(SSDSE_E, 20, 22, 24) # 20, 22, 24行目だけを取り出す
```

Training course:  
Data handling with dplyr (60 min)

- NSTAC produces official statistics
- We started encouraging use of R in our office
- To identify pilot projects is expected for the next step

Thank you for your attention.

uRos2024, 27-29 November 2024, Piraeus