

データの整理と集計

村田 昇

講義の内容

- データフレームの操作
- ファイルの取り扱い
- データの集計

データフレームの操作

R に用意されているデータ構造

- 下記は基本的なもので標準環境で利用できる
 - ベクトル (vector)
 - 行列 (matrix)
 - リスト (list)
 - データフレーム (data frame)
 - 配列 (array)

データフレームからの項目の抽出

- 添字の番号を指定
- 要素の名前で指定
- 除外: マイナス記号 (-) をつけて指定
- 論理値で指定
 - TRUE: 要素の選択
 - FALSE: 要素の 除外
- 欠損値 NA の扱いは状況依存なので注意
 - NA: 値が得られていないことを表すスカラー値

データ例

- `datasets::airquality`

New York Air Quality Measurements

- Description: Daily air quality measurements in New York, May to September 1973.
- Format: A data frame with 153 observations on 6 variables.
 - * [,1] Ozone numeric Ozone (ppb)
 - * [,2] Solar.R numeric Solar R (lang)
 - * [,3] Wind numeric Wind (mph)
 - * [,4] Temp numeric Temperature (degrees F)

- * [,5] Month numeric Month (1-12)
- * [,6] Day numeric Day of month (1-31)
- help(airquality) で詳細を確認
- datasets は R の標準パッケージ
- **パッケージ名:: オブジェクト** という書き方で同名のオブジェクトを区別できる

行の抽出

- 行番号による指定

```
## 行番号のベクトルで指定して抽出
airquality[1:10,] # 1-10 行を抽出
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10

- 条件の指定

```
## 条件の指定の仕方
airquality[1:15,]$Ozone>100 # 条件に合致する行は TRUE (NA は欠損値)
airquality[1:15,]$Ozone>100 & airquality[1:15,]$Wind<=5 # 条件の AND
with(airquality[1:15,], Ozone>100 & Wind<=5) # 上と同じ (短い書き方)
with(airquality[1:60,], Ozone>100 | Wind<=5) # 条件の OR
```

```
[1] FALSE FALSE FALSE FALSE NA FALSE FALSE FALSE
[9] FALSE NA FALSE FALSE FALSE FALSE FALSE
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[9] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[9] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[1] FALSE FALSE FALSE FALSE NA FALSE FALSE FALSE
[9] FALSE NA FALSE FALSE FALSE FALSE FALSE FALSE
[17] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[25] NA NA NA FALSE FALSE TRUE FALSE NA
[33] NA NA NA NA NA FALSE NA FALSE
[41] FALSE NA NA FALSE NA NA FALSE FALSE
[49] FALSE FALSE FALSE NA TRUE TRUE NA NA
[57] NA NA NA NA
```

- 条件に合致する行番号の抽出

```
## 関数 which() で TRUE の番号を抽出
which(with(airquality, Ozone>100 & Wind<=5)) # 全データから TRUE を抽出
```

```
[1] 62 99 117 121
```

- 条件に合致する行の抽出

```
## 条件を指定して行を抽出
airquality[which(with(airquality, Ozone>100 & Wind<=5)), ]
```

	Ozone	Solar.R	Wind	Temp	Month	Day
62	135	269	4.1	84	7	1
99	122	255	4.0	89	8	7
117	168	238	3.4	81	8	25
121	118	225	2.3	94	8	29

列の抽出

- 列番号による指定

```
## 列番号のベクトルで指定して抽出
airquality[which(with(airquality, Ozone>100 & Wind<=5)), c(1,5,6)]
```

	Ozone	Month	Day
62	135	7	1
99	122	8	7
117	168	8	25
121	118	8	29

- 列名による指定

```
## 複数の列の場合
airquality[which(with(airquality, Ozone>100 & Wind<=5)),
             c("Month", "Day")]
```

	Month	Day
62	7	1
99	8	7
117	8	25
121	8	29

- 列名による指定 (1つの場合)

```
## 1つの列の場合は以下でも良い (ただしデータフレームではなくベクトルになる)
airquality[which(with(airquality, Ozone>100 & Wind<=5)),]$Month
airquality[which(with(airquality, Ozone>100 & Wind<=5)), "Month"] # 上と同じ
## データフレームとして抽出したい場合は drop=FALSE を指定する
airquality[which(with(airquality, Ozone>100 & Wind<=5)), "Month", drop=FALSE]
```

```
[1] 7 8 8 8
[1] 7 8 8 8
      Month
62      7
99      8
117     8
121     8
```

関数 subset()

- 複合的な条件を指定してデータを整理する関数
- 基本書式

```
subset(x, subset, select, drop=FALSE)
## x: データフレーム
## subset: 抽出する行の条件
## select: 列の選択 (未指定の場合は全ての列)
## drop: 結果が 1 行または 1 列の場合の扱い、ベクトル (TRUE)・データフレーム (FALSE)
```

関数 subset() の例

- 前出の例の書き換え

```
### 関数 subset() の使い方
subset(airquality,
       subset = Ozone>100 & Wind<=5,
       select = c(1,5,6))
subset(airquality,
       Ozone>100 & Wind<=5, # 順序通りなら引数の名前は省略可
       c(Month,Day)) # 名前は$の後と同じ扱いで "" は不要
```

```
      Ozone Month Day
62      135     7   1
99      122     8   7
117     168     8  25
121     118     8  29

      Month Day
62         7   1
99         8   7
117        8  25
121        8  29
```

- いろいろな記述の仕方 (!, is.na(), %in%)

```
## Ozone に欠測 (NA) がなく、かつ Day が 5 か 10 のデータの Wind から Day までの列を抽出
subset(airquality,
       subset = !is.na(Ozone) & Day %in% c(5,10),
       select = Wind:Day)
```

```
      Wind Temp Month Day
41    11.5   87     6  10
66     4.6   83     7   5
71     7.4   89     7  10
97     7.4   85     8   5
128    7.4   87     9   5
133    9.7   73     9  10
```

- いろいろな記述の仕方 (|, -)

```
## Ozone が 120 以上か、または Wind が 3 以下の Temp 以外の列を抽出
subset(airquality,
       subset = Ozone>120 | Wind <= 3,
       select = -Temp)
```

```
      Ozone Solar.R Wind Month Day
53      NA      59  1.7     6  22
62     135     269  4.1     7   1
99     122     255  4.0     8   7
117     168     238  3.4     8  25
121     118     225  2.3     8  29
126     73     183  2.8     9   3
```

演習

練習問題

- `datasets::airquality` に対して以下の条件を満たすデータを取り出さない。
 - 7月のオゾン濃度 (Ozone)
 - 風速 (Wind) が時速 10 マイル以上で、かつ気温 (Temp) が華氏 80 度以上の日のデータ
 - オゾン (Ozone) も日射量 (Solar.R) も欠測 (NA) でないデータの月 (Month) と日 (Day)

ファイルの取り扱い

データファイルの読み書き

- 実際の解析においては以下の操作が必要
 - 収集されたデータを読み込む
 - 整理したデータを保存する
- R で利用可能なデータファイル
 - CSV 形式 (comma separated values): テキストファイル
 - RData 形式: R の内部表現を用いたバイナリーファイル
 - Excel 形式: RStudio の読み込み機能が利用可能
- データフレームを対象とした扱いを整理する

作業ディレクトリ

- R は **作業ディレクトリ** で実行される
 - ファイルは作業ディレクトリに存在するものとして扱われる
 - それ以外のファイルを扱う場合はパスを含めて指定する
- 作業ディレクトリの確認の仕方
 - コンソールの上部の表示
 - 関数 `getwd()`
- 作業ディレクトリの変更の仕方
 - **Session** メニューの **Set Working Directory** で指定
 - * 読み込んだファイルの場所を選択
 - * Files Pane の場所を選択
 - * ディレクトリを直接選択
 - 関数 `setwd()`

関数 `getwd()/setwd()` の例

- コンソール / R Script からの作業ディレクトリの操作

```
## 作業ディレクトリの確認 (環境によって実行結果が異なる)
getwd()
## 作業ディレクトリの移動 (環境によって指定の仕方も異なる)
setwd("~/Documents") # ホームディレクトリ下の「書類」フォルダに移動
```

- 作業ディレクトリはコンソールのタブにも表示されている

関数 write.csv()

- データフレームを CSV ファイルへ書き出す関数
- 基本書式

```
write.csv(x, file="ファイル名")
## x: 書き出すデータフレーム
## file: 書き出すファイルの名前 (作業ディレクトリ下, またはパスを指定)
```

– 他にも細かい指定ができるので詳しくはヘルプを参照

関数 write.csv() の例

- CSV ファイルの書き出し

```
## 関数 write.csv() の使い方 (CSV ファイルの操作)
(myData <- subset(airquality,
                  subset = Ozone>120,
                  select = -Temp)) # データフレームの作成
dim(myData) # データフレームの大きさを確認
## 作業ディレクトリの中に data というフォルダを用意しておく
write.csv(myData, file="data/mydata.csv") # csv ファイルとして書き出し
```

```
      Ozone Solar.R Wind Month Day
62      135     269  4.1      7   1
99      122     255  4.0      8   7
117     168     238  3.4      8  25
[1] 3 5
```

関数 read.csv()

- CSV ファイルからデータフレームを読み込む関数
- 基本書式

```
read.csv(file="ファイル名", header=TRUE,
          row.names, fileEncoding)
## file: 読み込むファイルの名前 (作業ディレクトリ下, またはパスを指定)
## header: 1行目を列名として使うか否か
## row.names: 行名の指定 (行名を含む列番号/列名, または行名の直接指定が可能)
## fileEncoding: 文字コードの指定 (日本語の場合, 主に使うのは "utf8", "sjis")
```

– 他にも細かい指定ができるので詳しくはヘルプを参照

– 関数 read.table(), scan() など参照

関数 read.csv() の例

- CSV ファイルの読み込み

```
## 関数 read.csv() の使い方 (CSV ファイルの操作)
(newdata <- read.csv(file="data/mydata.csv", # 前の例のファイル
                     row.names=1)) # 1列目を行名に指定
dim(newdata) # 正しく読み込めたか大きさを確認
```

```
      Ozone Solar.R Wind Month Day
62      135     269  4.1      7   1
99      122     255  4.0      8   7
117     168     238  3.4      8  25
[1] 3 5
```

関数 `save()`

- RData ファイルへ書き出す関数
- 基本書式

```
save(..., file="ファイル名")
## ...: 保存するオブジェクト名 (複数可, データフレーム以外も可)
## file: 書き出すファイルの名前 (作業ディレクトリ下, またはパスを指定)
```

– CSV 形式と異なり, **複数** のデータフレームを 1 つのファイルに保存することができる

関数 `save()` の例

- RData ファイルの書き出し

```
### 関数 save() の使い方 (RData ファイルの操作)
(myDat1 <- subset(airquality, Temp>95, select=-Ozone))
(myDat2 <- subset(airquality, Temp<57, select=-Ozone))
dim(myDat1); dim(myDat2) # 大きさを確認
save(myDat1, myDat2, file="data/mydata.rdata") # RData 形式で書き出し
```

```
      Solar.R Wind Temp Month Day
120      203  9.7   97      8  28
122      237  6.3   96      8  30
      Solar.R Wind Temp Month Day
5         NA 14.3   56      5   5
[1] 2 5
[1] 1 5
```

関数 `load()`

- RData ファイルから読み込む関数
- 基本書式

```
load(file="ファイル名")
## file: 読み込むファイルの名前 (作業ディレクトリ下, またはパスを指定)
```

– 同じ名前のオブジェクトがあると上書きされるので注意

関数 `load()` の例

- RData ファイルの読み込み

```
## 関数 load() の使い方 (RData ファイルの操作)
(myDat1 <- subset(airquality, Ozone > 160)) # 新たに作成
load(file="data/mydata.rdata") # RData 形式の読み込み
myDat1 # save したときの名前で読み込まれ上書きされる
myDat2
```

```
      Ozone Solar.R Wind Temp Month Day
117    168      238  3.4   81      8  25
      Solar.R Wind Temp Month Day
120      203  9.7   97      8  28
122      237  6.3   96      8  30
      Solar.R Wind Temp Month Day
5         NA 14.3   56      5   5
```

演習

練習問題

- 以下のデータを読み込み、操作を行ってみよう。
 - <https://www.e-stat.go.jp> より取得したデータ (地域から探す / 全県を選択 / 項目を選択してダウンロード)
 - データファイル (文字コード: utf8)
 - * jpdata1.csv: 県別の対象データ
 - * jpdata2.csv: 対象データの内容
 - * jpdata3.csv: 県別と地域の対応関係
 - 作業ディレクトリの data 内に置いて読み込む

```
myData <- read.csv(file="data/jpdata1.csv", fileEncoding="utf8", row.names=1)
myItem <- read.csv(file="data/jpdata2.csv", fileEncoding="utf8")
myArea <- read.csv(file="data/jpdata3.csv", fileEncoding="utf8")
```

- 日本語に問題がある場合は英語版を読み込む

```
myDataEn <- read.csv(file="data/jpdata1-en.csv", row.names=1)
myAreaEn <- read.csv(file="data/jpdata3-en.csv")
```

データの集計

集約のための関数

- データを集約するための基本的な関数は用意されている
 - 関数 `sum()`: 総和
 - 関数 `mean()`: 平均
 - 関数 `max()`: 最大値
 - 関数 `min()`: 最小値
 - 関数 `summary()`: 基本統計量
- これ以外にも集約を行なう関数は沢山ある

関数の例

- 練習問題のデータの集計を行う

```
myData <- read.csv(file="data/jpdata1.csv",
                  row.names=1, fileEncoding="utf8")
## 一度読み込んでいれば上の行は不要
sum(myData$人口) # 全国の総人口 (列名で選択)
mean(myData[,4]) # 面積の平均値 (行列として列を選択)
median(myData[[4]]) # 面積の中央値 (リストとして列を選択)
min(myData["若年"]) # 若年人口の最小値 (列名で選択)
with(myData, max(老人)) # 老年人口の最大値 (関数 with() を利用)
```

```
[1] 126708000
[1] 793554.5
[1] 609719
[1] 72000
[1] 3160000
```



```
### 日本語に不具合がある場合
myDataEn <- read.csv(file="data/jpdata1-en.csv", row.names=1)
## 一度読み込んでいれば上の行は不要
sum(myDataEn$population) # 全国の総人口 (列名で選択)
mean(myDataEn[,4]) # 面積の平均値 (行列として列を選択)
median(myDataEn[[4]]) # 面積の中央値 (リストとして列を選択)
min(myDataEn["young_population"]) # 若年人口の最小値 (列名で選択)
with(myDataEn, max(old_population)) # 老年人口の最大値 (関数 with() を利用)
```

関数 apply()

- 列あるいは行ごとの計算を行う関数
- 基本書式

```
apply(X, MARGIN, FUN)
## X: データフレーム
## MARGIN: 行 (1) か列 (2) かを指定
## FUN: 計算すべき統計量の関数
```

– 総和や平均は専用の関数も用意されている

```
* rowSums()/colSums()
* rowMeans()/colMeans()
```

関数 apply() の例

- 抽出したデータの集計を行う

```
### 関数 apply の使い方
x <- subset(myData, select=婚姻:失業) # 抽出
colMeans(x) # 各列の平均
apply(x, 2, max) # 列ごとの最大値
sapply(x, max) # 上と同じ (help(sapply) を参照)
## 自作関数の適用 (関数に名前を付けずに利用できる)
apply(x, 2, function(z){sum(z>mean(z))}) # 平均より大きいデータ数
```

```
      婚姻      離婚      失業
4.437021 1.631064 4.221277
婚姻 離婚 失業
6.19 2.41 6.30
婚姻 離婚 失業
6.19 2.41 6.30
婚姻 離婚 失業
      20      22      25
```

```
### 日本語に不具合がある場合
x <- subset(myDataEn, select=marriage:unemployed) # 抽出
colMeans(x) # 各列の平均
apply(x, 2, max) # 列ごとの最大値
sapply(x, max) # 上と同じ (help(sapply) を参照)
apply(x, 2, function(z){sum(z>mean(z))}) # 平均より大きいデータ数
```

関数 aggregate()

- 各行をグループにまとめて統計量を計算する関数
- 基本書式

```
aggregate(x, by, FUN)
## x: データフレーム
## by: 各行が属するグループを指定するベクトルをリストで与える (複数可)
## FUN: 求めたい統計量を計算するための関数
```

– 同様な目的に関数 `tapply()` も利用可

関数 `aggregate()` の例

- 同じ値を持つグループごとの平均値を求める

```
### 関数 aggregate の使い方
## 人口から面積まで地方ごとの平均値を計算
x <- subset(myData, select=人口: 面積)
aggregate(x, by=list(地方=myArea$地方), FUN=mean)
```

	地方	人口	若年	老人	面積
1	関東	6178286	737000.0	1564000.0	463329.3
2	近畿	3204429	395714.3	898714.3	473223.6
3	九州	1795000	243875.0	511000.0	556395.0
4	四国	947000	112250.0	305750.0	470091.5
5	中国	1473800	186400.0	448600.0	638433.4
6	中部	2372889	302888.9	667555.6	742297.6
7	東北	1472667	169333.3	452666.7	1115790.7
8	北海道	5320000	588000.0	1632000.0	7842078.0

- 代入せずにまとめて書くことも可能

```
aggregate(subset(myData, select=人口: 面積),
          by=list(地方=myArea$地方),
          FUN=mean)
```

	地方	人口	若年	老人	面積
1	関東	6178286	737000.0	1564000.0	463329.3
2	近畿	3204429	395714.3	898714.3	473223.6
3	九州	1795000	243875.0	511000.0	556395.0
4	四国	947000	112250.0	305750.0	470091.5
5	中国	1473800	186400.0	448600.0	638433.4
6	中部	2372889	302888.9	667555.6	742297.6
7	東北	1472667	169333.3	452666.7	1115790.7
8	北海道	5320000	588000.0	1632000.0	7842078.0

- 以下も同じ結果を返す

```
y <- data.frame(x, 地方=myArea$地方)
aggregate(. ~ 地方, data=y, FUN=mean)
```

	地方	人口	若年	老人	面積
1	関東	6178286	737000.0	1564000.0	463329.3
2	近畿	3204429	395714.3	898714.3	473223.6
3	九州	1795000	243875.0	511000.0	556395.0
4	四国	947000	112250.0	305750.0	470091.5
5	中国	1473800	186400.0	448600.0	638433.4
6	中部	2372889	302888.9	667555.6	742297.6
7	東北	1472667	169333.3	452666.7	1115790.7
8	北海道	5320000	588000.0	1632000.0	7842078.0

- まとめて書くことも可能

```
aggregate( . ~ 地方, # 右辺で条件付けて左辺 (右辺以外) を計算
           data=data.frame(subset(myData,select=人口:面積),
                           地方=myArea$地方),
           FUN=mean)
```

	地方	人口	若年	老人	面積
1	関東	6178286	737000.0	1564000.0	463329.3
2	近畿	3204429	395714.3	898714.3	473223.6
3	九州	1795000	243875.0	511000.0	556395.0
4	四国	947000	112250.0	305750.0	470091.5
5	中国	1473800	186400.0	448600.0	638433.4
6	中部	2372889	302888.9	667555.6	742297.6
7	東北	1472667	169333.3	452666.7	1115790.7
8	北海道	5320000	588000.0	1632000.0	7842078.0

- 複数の条件でグループ分け

```
## 地方と、人口が中央値以下か否かでグループ分けして平均値を計算
aggregate(x, by=list(地方=myArea$地方,
                    過疎=with(myData, 人口<=median(人口))),
           FUN=mean)
```

	地方	過疎	人口	若年	老人	面積
1	関東	FALSE	6178285.7	737000.0	1564000.0	463329.3
2	近畿	FALSE	4681250.0	573750.0	1305500.0	517317.2
3	九州	FALSE	3436000.0	456000.0	957500.0	619800.0
4	中国	FALSE	2368000.0	305500.0	688000.0	779697.5
5	中部	FALSE	3510200.0	451400.0	973200.0	994346.8
6	東北	FALSE	2102500.0	250000.0	600000.0	1053306.0
7	北海道	FALSE	5320000.0	588000.0	1632000.0	7842078.0
8	近畿	TRUE	1235333.3	158333.3	356333.3	414432.0
9	九州	TRUE	1248000.0	173166.7	362166.7	535260.0
10	四国	TRUE	947000.0	112250.0	305750.0	470091.5
11	中国	TRUE	877666.7	107000.0	289000.0	544257.3
12	中部	TRUE	951250.0	117250.0	285500.0	427236.0
13	東北	TRUE	1157750.0	129000.0	379000.0	1147033.0

- 別の書き方

```
aggregate( . ~ 地方 + 過疎, FUN=mean, # + で条件を追加
           data=data.frame(subset(myData,select=人口:面積),
                           地方=myArea$地方,
                           過疎=with(myData, 人口<=median(人口))))
```

	地方	過疎	人口	若年	老人	面積
1	関東	FALSE	6178285.7	737000.0	1564000.0	463329.3
2	近畿	FALSE	4681250.0	573750.0	1305500.0	517317.2
3	九州	FALSE	3436000.0	456000.0	957500.0	619800.0
4	中国	FALSE	2368000.0	305500.0	688000.0	779697.5
5	中部	FALSE	3510200.0	451400.0	973200.0	994346.8
6	東北	FALSE	2102500.0	250000.0	600000.0	1053306.0
7	北海道	FALSE	5320000.0	588000.0	1632000.0	7842078.0
8	近畿	TRUE	1235333.3	158333.3	356333.3	414432.0
9	九州	TRUE	1248000.0	173166.7	362166.7	535260.0
10	四国	TRUE	947000.0	112250.0	305750.0	470091.5
11	中国	TRUE	877666.7	107000.0	289000.0	544257.3
12	中部	TRUE	951250.0	117250.0	285500.0	427236.0
13	東北	TRUE	1157750.0	129000.0	379000.0	1147033.0

```

### 日本語に不具合がある場合
## 人口から面積まで地方ごとの平均値を計算
x <- subset(myDataEn,select=population:area)
aggregate(x, by=list(region=myAreaEn$region), FUN=mean)

aggregate(subset(myDataEn,select=population:area),
           by=list(region=myAreaEn$region),
           FUN=mean)

y <- data.frame(x,region=myAreaEn$region)
aggregate( . ~ region, data=y, FUN=mean)

aggregate( . ~ region, # 右辺で条件付けて左辺 (右辺以外) を計算
           data=data.frame(subset(myDataEn,select=population:area),
                           region=myAreaEn$region),
           FUN=mean)

## 地方と、人口が中央値以下か否かでグループ分けして平均値を計算
aggregate(x, by=list(region=myAreaEn$region,
                     depop=with(myDataEn, population<=median(population))),
           FUN=mean)

aggregate( . ~ region + depop, FUN=mean,
           data=data.frame(subset(myDataEn,select=population:area),
                           region=myAreaEn$region,
                           depop=with(myDataEn, population<=median(population))))

```

演習