

rd-9. テーブルデータ 処理, 並べ替え (ソー ト), 集計・集約

データサイエンス演習
(R システムを使用)

<https://www.kkaneko.jp/cc/rd/index.html>

金子邦彦



アウトライン



9-1 データテーブル

9-2 選択, 射影, 自然結合, 直積

9-3 並べ替え (ソート)

9-4 集約

9-5 演算の組み合わせ



9-1 テーブルデータ

テーブルデータの例



科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

成績テーブル

```
library(dplyr)
d1 <- data_frame(
  科目=c("国語", "国語", "算数", "算数", "理科"),
  受講者=c("A", "B", "A", "B", "A"),
  得点=c(90, 80, 95, 90, 80) )
```

コンストラクタ

科目	教室
国語	101
算数	201
理科	301

教室テーブル

```
library(dplyr)
d3 <- data_frame(
  科目=c("国語", "算数", "理科"),
  教室=c("101", "201", "301") )
```

コンストラクタ



9-2 選択、射影、自然結合、 直積

選択



テーブルの中から、**選択条件**に合致する**レコード**のみを**選択**する。結果は、新しい**テーブル**になる

テーブル
成績

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

選択



科目	受講者	得点
国語	A	90
算数	A	95

新しいテーブル

結合条件は

「得点 \geq 90」

選択条件で用いる比較演算子



等しいか等しくないか

== 等しい

!= 等しくない

大小の比較

> より大きい

>= 以上

< より小さい

<= 以下

選択



科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

元データ

選択を行うテーブルのオブジェクト名	d1
選択条件	得点 >= 90

行いたいこと

	科目 (chr)	受講者 (chr)	得点 (dbl)
1	国語	A	90
2	算数	A	95
3	算数	B	90

結果

```
library(dplyr)
d1 <- data_frame(
  科目=c("国語", "国語", "算数", "算数", "理科"),
  受講者=c("A", "B", "A", "B", "A"),
  得点=c(90, 80, 95, 90, 80) )
d1 %>% filter(得点 >= 90)
```

テーブルの中の、必要なフィールド名リストを指定する。結果は、新しいテーブルになる

テーブル
成績

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

射影



科目	受講者
国語	A
国語	B
算数	A
算数	B
理科	A

新しいテーブル
フィールド名リストは
「受講者, 得点」

射影



科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

元データ

射影を行うテーブル
のオブジェクト名

d1

フィールド名リスト

科目, 受講者

行いたいこと

	科目 (chr)	受講者 (chr)
1	国語	A
2	国語	B
3	算数	A
4	算数	B
5	理科	A

結果

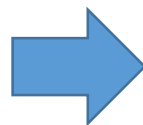
```
library(dplyr)
d1 <- data_frame(
  科目=c("国語", "国語", "算数", "算数", "理科"),
  受講者=c("A", "B", "A", "B", "A"),
  得点=c(90, 80, 95, 90, 80) )
d1 %>% select(科目, 受講者)
```

2つのテーブルの結合属性を用いて結合する。
結果は、新しいテーブルになる

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

科目	教室
国語	101
算数	201
理科	301

自然結合



科目	受講者	得点	教室
国語	A	90	101
国語	B	80	101
算数	A	95	201
算数	B	90	201
理科	A	80	301

新しいテーブル

自然結合



科目	受講者	得点	科目	教室
国語	A	90	国語	101
国語	B	80		
算数	A	95	算数	201
算数	B	90		
理科	A	80	理科	301

元データ

自然結合を
行うテーブ
ルのオブ
ジェクト名

d1,
d3

行いたいこと



	科目 (chr)	受講者 (chr)	得点 (dbl)	教室 (chr)
1	国語	A	90	101
2	国語	B	80	101
3	算数	A	95	201
4	算数	B	90	201
5	理科	A	80	301

結果

```
library(dplyr)
```

```
d1 <- data_frame(
```

```
  科目=c("国語", "国語", "算数", "算数", "理科"),
```

```
  受講者=c("A", "B", "A", "B", "A"),
```

```
  得点=c(90, 80, 95, 90, 80) )
```

```
d3 <- data_frame(
```

```
  科目=c("国語", "算数", "理科"),
```

```
  教室=c("101", "201", "301") )
```

```
inner_join(d1, d3)
```

直積は、2つのテーブルの全レコードの組み合わせ。結果は、新しいテーブルになる

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

科目	教室
国語	101
算数	201
理科	301

直積



	科目.x (chr)	受講者 (chr)	得点 (dbl)	科目.y (chr)	教室 (chr)
1	国語	A	90	国語	101
2	国語	A	90	算数	201
3	国語	A	90	理科	301
4	国語	B	80	国語	101
5	国語	B	80	算数	201
6	国語	B	80	理科	301
7	算数	A	95	国語	101
8	算数	A	95	算数	201
9	算数	A	95	理科	301
10	算数	B	90	国語	101
11	算数	B	90	算数	201
12	算数	B	90	理科	301
13	理科	A	80	国語	101
14	理科	A	80	算数	201
15	理科	A	80	理科	301

新しいテーブル

参考 Web ページ: <http://www.alfredo.motta.name/data-manipulation-primitives-in-r-and-python/>

直積



科目	受講者	得点	科目	教室
国語	A	90	国語	101
国語	B	80	算数	201
算数	A	95	理科	301
算数	B	90		
理科	A	80		

元データ

直積を行う
テーブルの
オブジェク
ト名

d1,
d3

行いたいこと

	科目.x (chr)	受講者 (chr)	得点 (dbl)	科目.y (chr)	教室 (chr)
1	国語	A	90	国語	101
2	国語	A	90	算数	201
3	国語	A	90	理科	301
4	国語	B	80	国語	101
5	国語	B	80	算数	201
6	国語	B	80	理科	301
7	算数	A	95	国語	101
8	算数	A	95	算数	201
9	算数	A	95	理科	301
10	算数	B	90	国語	101
11	算数	B	90	算数	201
12	算数	B	90	理科	301
13	理科	A	80	国語	101
14	理科	A	80	算数	201
15	理科	A	80	理科	301

結果

```
library(dplyr)
d1 <- data_frame(
  科目=c("国語", "国語", "算数", "算数", "理科"),
  受講者=c("A", "B", "A", "B", "A"),
  得点=c(90, 80, 95, 90, 80) )
d3 <- data_frame(
  科目=c("国語", "算数", "理科"),
  教室=c("101", "201", "301") )
d1$tmp = NA
d3$tmp = NA
full_join(d1, d3, by="tmp") %>% select(-tmp)
```



9-3 並べ替え (ソート)

並べ替え (ソート) の例



科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

並べ替え (ソート) 前



データを一定の規則で並べ替え。
並べ替えは行単位

	科目 (chr)	受講者 (chr)	得点 (dbl)
1	国語	B	80
2	理科	A	80
3	国語	A	90
4	算数	B	90
5	算数	A	95

得点で昇順

	科目 (chr)	受講者 (chr)	得点 (dbl)
1	算数	A	95
2	国語	A	90
3	算数	B	90
4	国語	B	80
5	理科	A	80

得点で降順

並べ替え (ソート)



- データを一定の規則（昇順または降順）で並べ替え
- 並べ替えはレコード単位
- 並べ替えの結果、新しいテーブルができる
- 並べ替え時に、「キーとなるフィールド名」と「順序（昇順または降順）」を設定する必要がある

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

並べ替え前



	科目 (chr)	受講者 (chr)	得点 (dbl)
1	国語	B	80
2	理科	A	80
3	国語	A	90
4	算数	B	90
5	算数	A	95

得点で昇順

昇順での並べ替え (ソート)



並べ替え (ソート) を行う テーブルのオブジェクト名	d1
キー	得点
順序	昇順

	科目 (chr)	受講者 (chr)	得点 (dbl)
1	国語	B	80
2	理科	A	80
3	国語	A	90
4	算数	B	90
5	算数	A	95

```
library(dplyr)
d1 <- data_frame(
  科目=c("国語", "国語", "算数", "算数", "理科"),
  受講者=c("A", "B", "A", "B", "A"),
  得点=c(90, 80, 95, 90, 80) )
d1 %>% arrange(得点)
```

降順での並べ替え (ソート)



並べ替え (ソート) を行う テーブルのオブジェクト名	d1
キー	得点
順序	降順

	科目 (chr)	受講者 (chr)	得点 (dbl)
1	算数	A	95
2	国語	A	90
3	算数	B	90
4	国語	B	80
5	理科	A	80

```
library(dplyr)
d1 <- data_frame(
  科目=c("国語", "国語", "算数", "算数", "理科"),
  受講者=c("A", "B", "A", "B", "A"),
  得点=c(90, 80, 95, 90, 80) )
d1 %>% arrange(desc(得点))
```

複数フィールドでの並べ替え (ソート)



並べ替え (ソート) を行う テーブルのオブジェクト名	d1
キー	得点, 受講者
順序	得点は降順 受講者は昇順

	科目 (chr)	受講者 (chr)	得点 (dbl)
1	算数	A	95
2	国語	A	90
3	算数	B	90
4	理科	A	80
5	国語	B	80

```
library(dplyr)
```

```
d1 <- data_frame(
```

```
  科目=c("国語", "国語", "算数", "算数", "理科"),
```

```
  受講者=c("A", "B", "A", "B", "A"),
```

```
  得点=c(90, 80, 95, 90, 80) )
```

```
d1 %>% arrange(desc(得点), 受講者)
```



9-4 集約

集約の例



科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

元データ



Aさんは 3科目
Bさんは 2科目受講した

A	3
B	2

集約の例

グループごとに、頻度や要約統計量を求める

- 頻度 (数え上げ)
種類ごとの数え上げ
- 要約統計量
平均 (mean)、標準偏差 (sd)、分散 (var)
中央値 (median)、四分位点 (quantile)、
最大値 (max)、最小値 (min)

集約では、グループの基準もいろいろ



グループの基準が
受講者

科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

元データ



A	88.33333
B	85

得点の平均

グループの基準が
科目

国語	85
算数	92.5
理科	80

得点の平均

集約の例

集約の例



① d1 %>% group_by(科目)
%>% summarise(n())

国語	2
算数	2
理科	1

データの**個数**

② d1 %>% group_by(受講者)
%>% summarise(mean(得点))

A	88.33333
B	85

得点の**合計**

③ d1 %>% group_by(科目)
%>% summarise(mean(得点))

国語	85
算数	92.5
理科	80

得点の**平均**

集約 ①



科目	受講者	得点	集約を行う テーブルのオ ブジェクト名	d1
国語	A	90	グループの基 準	受講者
国語	B	80		
算数	A	95	集約する フィールド名	得点
算数	B	90		
理科	A	80		

成績



	受講者 (chr)	min(得点) (dbl)	Q1 (dbl)	median(得点) (dbl)	mean(得点) (dbl)	Q3 (dbl)	max(得点) (dbl)
1	A	80	85.0	90	88.33333	92.5	95
2	B	80	82.5	85	85.00000	87.5	90

```
library(dplyr)
```

```
d1 <- data_frame(
```

```
  科目=c("国語", "国語", "算数", "算数", "理科"),
```

```
  受講者=c("A", "B", "A", "B", "A"),
```

```
  得点=c(90, 80, 95, 90, 80) )
```

```
d1 %>% group_by(受講者) %>% summarise(min(得点),  
Q1=quantile(得点, probs=0.25), median(得点), mean(得点),  
Q3=quantile(得点, probs=0.75), max(得点))
```

集約 ②



科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

成績

集約を行う テーブルのオブジェクト名	d1
グループの基準	科目
集約するフィールド名	得点



科目 (chr)	min(得点) (dbl)	Q1 (dbl)	median(得点) (dbl)	mean(得点) (dbl)	Q3 (dbl)	max(得点) (dbl)
1 国語	80	82.50	85.0	85.0	87.50	90
2 算数	90	91.25	92.5	92.5	93.75	95
3 理科	80	80.00	80.0	80.0	80.00	80

```
library(dplyr)
```

```
d1 <- data_frame(
```

```
  科目=c("国語", "国語", "算数", "算数", "理科"),
```

```
  受講者=c("A", "B", "A", "B", "A"),
```

```
  得点=c(90, 80, 95, 90, 80) )
```

```
d1 %>% group_by(科目) %>% summarise(min(得点), Q1=quantile(得点,  
probs=0.25), median(得点), mean(得点), Q3=quantile(得点, probs=0.75),  
max(得点))
```

ピボットテーブル（クロス集約表）の例



名前	性別	申し込み
A	男性	済
B	男性	未
C	女性	済
D	女性	未
E	男性	済
F	男性	未

元データ



	女性	男性
済	1	2
未	1	2

ピボットテーブル
（クロス集約表）の例

ピボットテーブル (クロス集計表)



名前	性別	申し込み
A	男性	済
B	男性	未
C	女性	済
D	女性	未
E	男性	済
F	男性	未

集約を行う
テーブルの
オブジェク
ト名

グループの
基準

d4

性別, 申
し込み



	性別 (chr)	申し込み (chr)	count (int)
1	女性	済	1
2	女性	未	1
3	男性	済	2
4	男性	未	2

※ 結果は縦長形式 (long-format)

```
library(dplyr)
```

```
d4 <- data_frame(
```

```
  名前=c("A", "B", "C", "D", "E", "F"),
```

```
  性別=c("男性", "男性", "女性", "女性", "男性", "男性"),
```

```
  申し込み=c("済", "未", "済", "未", "済", "未"))
```

```
d4 %>% group_by(性別, 申し込み) %>% summarise(count=n())
```

ピボットテーブル (クロス集計表)



名前	性別	申し込み
A	男性	済
B	男性	未
C	女性	済
D	女性	未
E	男性	済
F	男性	未

集約を行う
テーブルの
オブジェク
ト名

グループの
基準

d4

性別, 申
し込み



	申し込み	女性	男性
	(chr)	(int)	(int)
1	済	1	2
2	未	1	2

※ 結果は横長形式 (wide-format)

```
library(dplyr)
```

```
library(tidyr)
```

```
d4 <- data_frame(
```

```
  名前=c("A", "B", "C", "D", "E", "F"),
```

```
  性別=c("男性", "男性", "女性", "女性", "男性", "男性"),
```

```
  申し込み=c("済", "未", "済", "未", "済", "未") )
```

```
d4 %>% group_by(性別, 申し込み) %>% summarise(count=n()) %>%  
spread(性別, count)
```



9-5 演算の組み合わせ

演算の組み合わせの例



科目	受講者	得点
国語	A	90
国語	B	80
算数	A	95
算数	B	90
理科	A	80

成績テーブル
(オブジェクト名は d1)

```
d1 %>%  
  filter(得点 >= 90) %>%  
  select(科目, 受講者)
```

選択 + 射影

	科目 (chr)	受講者 (chr)
1	国語	A
2	算数	A
3	算数	B

```
d3 %>%  
  filter(教室 == 101) %>%  
  inner_join(d1) %>%  
  select(受講者)
```

選択 + 結合 + 射影

	受講者 (chr)
1	A
2	B

科目	教室
国語	101
算数	201
理科	301

```
d1 %>%  
  group_by(科目) %>%  
  summarise(Mean=mean(得点)) %>%  
  filter(Mean >= 85)
```

集約 + 選択

	科目 (chr)	Mean (dbl)
1	国語	85.0
2	算数	92.5

部屋テーブル
(オブジェクト名は d3)