

自分で触ってよくわかる
部分抽出の話：
条件にあう対象者を抽出する

神戸市立医療センター中央市民病院
臨床研究推進センター

宮越 千智

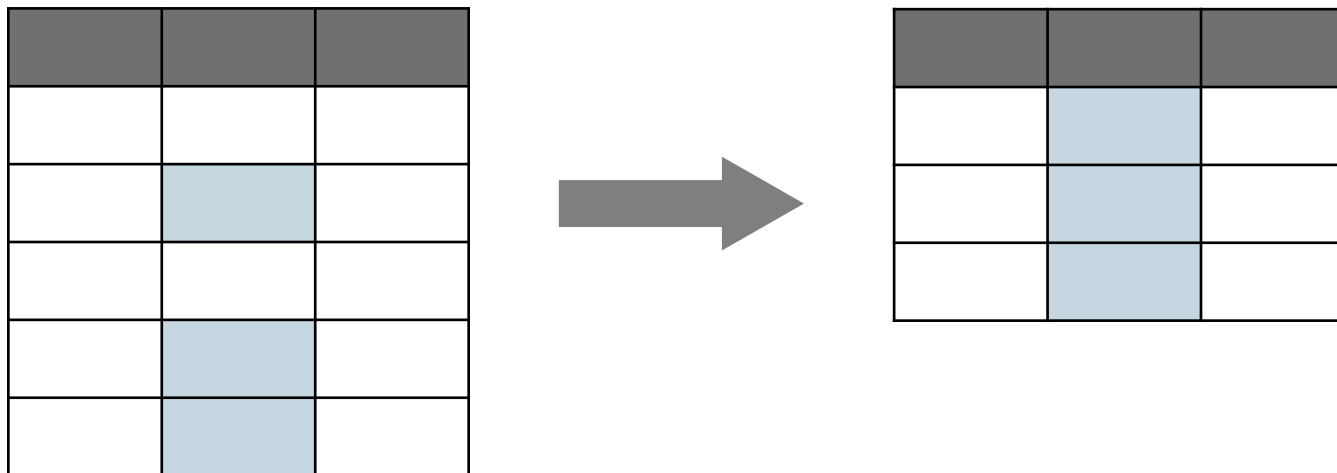
今回の学習目標

- ✓ データの中で条件に当てはまるものだけを抽出することができる
- ✓ ある変数について大きい順・小さい順にデータの一部を抽出することができる

復習：データ整形に必要な操作

必要な行・列のみ取り出す

条件に該当する行のみ抽出



条件指定によく使う演算子

		R	Python
等価	等しい	==	
不等価	等しくない	!=	
未満	～より小さい	<	
超過	～より大きい	>	
以下	～以下	<=	
以上	～以上	>=	
範囲	a以上b以下	between(a, b)	a <= x <= b
論理積	かつ	&	
論理和	または		

Rを使いたい人 

Rを使いたい人:

条件に合致する集団を抽出する

1. survivalパッケージ、tidyverseパッケージを読み込み、
pbcデータが使える状態にしておく

```
library(survival)
library(tidyverse)
data(pbc)
```

2. filter()関数で条件を指定

```
pbc %>% filter(age>=60 & sex=="f")
```

- ✓ 変数名には引用符不要
- ✓ カテゴリー変数の値は引用符で括る

Rを使いたい人:

上位・下位の一部を抽出する

1. survivalパッケージ、tidyverseパッケージを読み込み、
pbcデータが使える状態にしておく

```
library(survival)
library(tidyverse)
data(pbc)
```


2. slice_max()関数またはslice_min()関数で参照する変数と抽出数を指定

✓ 例1: ageが大きい人から10人

```
pbc %>% slice_max(age, n=10)
```

✓ 例2: ageが小さい人から10%

```
pbc %>% slice_min(age, prop=0.1)
```

Pythonを使いたい人 



Pythonを使いたい人: 条件に合致する集団を抽出する

1. survialパッケージのpbcデータが使える状態にしておく

```
import pandas as pd
import statsmodels.api as sm
dataset = sm.datasets.get_rdataset("pbc", "survival")
df = dataset.data
```

2. .query()メソッドで条件を指定

```
df.query('age>=60 & sex=="f"')
```

- ✓ 条件全体を引用符で括る
- ✓ カテゴリ変数の値は引用符で括る



Pythonを使いたい人: 上位・下位の一部を抽出する

1. survialパッケージのpbcデータが使える状態にしておく

```
import pandas as pd
import statsmodels.api as sm
dataset = sm.datasets.get_rdataset("pbc", "survival")
df = dataset.data
```

2. .nlargest()メソッドまたは.nsmallest()メソッドで参照する変数と抽出数を指定

✓ 例: ageの小さい人から10人を抽出する

```
df_smallest10 = df.nsmallest(10, 'age')
print(df_smallest10 )
```

課題7：部分抽出

- Rのsurvivalパッケージにあるpbcデータについて、
治療1を受けた人(`trt=1`)の中で年齢が若い人10人を抽出してみましょう
 - ✓ ヒント: `trt`は1か2を取りますが、文字列ではなく数値です

今回のまとめ

- ✓ 条件を指定するときの論理演算子は、対象集団を抽出したり
曝露要因やアウトカムを定義するときにも登場するので
使いこなせるようになりましょう
- ✓ エラーが出るときはまず、変数や値を引用符で括る必要があるか
を確認しましょう