

データの読み込み

```
In [1]: import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
# 以上3つのライブラリが「3種の神器」
```

```
In [2]: Data=pd.read_csv(
"D:2022_数理統計学/StatData/StatData02_2.csv",
skiprows=1, # 1行目を飛ばす
names=['mid (x)', 'final (y)']) # カラム名称 (英語が好ましい)
Data
```

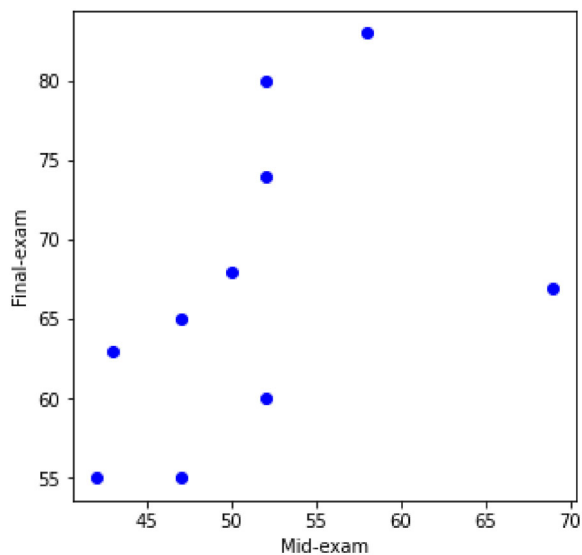
Out[2]:

	mid (x)	final (y)
0	50	68
1	58	83
2	52	74
3	52	80
4	43	63
5	47	55
6	52	60
7	69	67
8	47	65
9	42	55

散布図

```
In [3]: plt.figure(figsize=(5,5)) # 図の大きさ (6,4) がデフォルト
plt.scatter(Data['mid (x)'],Data['final (y)'],c='blue') # 散布図を表示
plt.xlabel('Mid-exam')
plt.ylabel('Final-exam')
```

Out[3]: Text(0, 0.5, 'Final-exam')



基本的な統計量の計算

```
In [4]: # 基本統計量 (1)
# あとで引用できるように統計量に名前をつけておく
size=len(Data['mid (x)']) # データのサイズ
Mid_mean=np.mean(Data['mid (x)']) # 変量 mid (x) の平均値
Mid_var=np.var(Data['mid (x)']) # 変量 mid (x) の分散
Mid_std=np.std(Data['mid (x)']) # 変量 mid (x) の標準偏差
print(size, Mid_mean, Mid_var, Mid_std)
```

10 51.2 55.36 7.4404300950953095

```
In [5]: # 基本統計量 (2)
# あとで引用できるように統計量に名前をつけておく
Final_mean=np.mean(Data['final (y)']) # 変量 mid (x) の平均値
Final_var=np.var(Data['final (y)']) # 変量 mid (x) の分散
Final_std=np.std(Data['final (y)']) # 変量 mid (x) の標準偏差
print(Final_mean, Final_var, Final_std)
```

67.0 83.2 9.121403400793104

```
In [6]: # 基本統計量 (3) 共分散と相関係数
MF_cov=Data['mid (x)'].cov(Data['final (y)'], ddof=0) # ddof=0 を要す。デフォルトは ddof=1 (不偏分)
MF_corr=Data['mid (x)'].corr(Data['final (y)'])
print(MF_cov, MF_corr)
```

32.0 0.4715093122829988

```
In [7]: # 基本統計量のまとめ
print(
    ['平均値', Mid_mean, Final_mean],
    ['分散', Mid_var, Final_var],
    ['標準偏差', Mid_std, Final_std],
    ['共分散', MF_cov],
    ['相関係数', MF_corr],
    ['サイズ', size]
)
```

['平均値', 51.2, 67.0] ['分散', 55.36, 83.2] ['標準偏差', 7.4404300950953095, 9.121403400793104]
['共分散', 32.0] ['相関係数', 0.4715093122829988] ['サイズ', 10]

```
In [8]: # 無駄に長い小数表示を整理すること
StatSummary=pd.DataFrame([
    ['平均値', np.round(Mid_mean,2), np.round(Final_mean,2)],
    ['分散', np.round(Mid_var,2), np.round(Final_var,2)],
    ['標準偏差', np.round(Mid_std,2), np.round(Final_std,2)],
    ['共分散', np.round(MF_cov,2)],
    ['相関係数', np.round(MF_corr,2)],
    ['データ数', size]
])
StatSummary=StatSummary.rename(columns={0:'統計量'})
StatSummary=StatSummary.rename(columns={1:'Mid-exam'})
StatSummary=StatSummary.rename(columns={2:'Final-exam'})
StatSummary
```

Out[8]:

	統計量	Mid-exam	Final-exam
0	平均値	51.20	67.00
1	分散	55.36	83.20
2	標準偏差	7.44	9.12
3	共分散	32.00	NaN
4	相関係数	0.47	NaN
5	データ数	10.00	NaN

標準化して散布図を描く

```
In [9]: # 標準化して散布図を描く
NX=(Data['mid (x)']-Mid_mean)/Mid_std # 変数 mid (x) の標準化
```

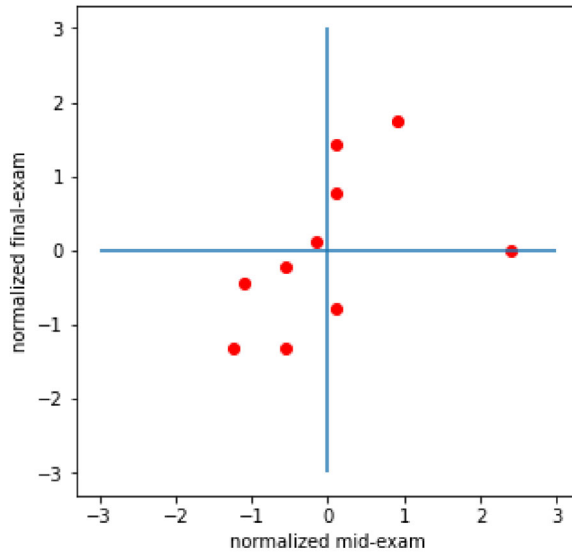
```
In [10]: NY=(Data['final (y)']-Final_mean)/Final_std # 変数 Final (y) の標準化
Data['x_normalized']=NX
Data['y_normalized']=NY
Data
```

Out[10]:

	mid (x)	final (y)	x_normalized	y_normalized
0	50	68	-0.161281	0.109632
1	58	83	0.913926	1.754116
2	52	74	0.107521	0.767426
3	52	80	0.107521	1.425219
4	43	63	-1.102087	-0.438529
5	47	55	-0.564483	-1.315587
6	52	60	0.107521	-0.767426
7	69	67	2.392335	0.000000
8	47	65	-0.564483	-0.219265
9	42	55	-1.236488	-1.315587

```
In [11]: plt.figure(figsize=(5, 5)) # 図の大きさ (6, 4) がデフォルト
plt.scatter(Data['x_normalized'], Data['y_normalized'], c='red') # 散布図を表示
plt.xlabel('normalized mid-exam')
plt.ylabel('normalized final-exam')
plt.hlines(0, -3, 3) # 直線 y=0 を -3<x<3 で図示
plt.vlines(0, -3, 3) # 直線 x=0 を -3<y<3 で図示
```

Out[11]: <matplotlib.collections.LineCollection at 0x2147d2de370>



In []: