202221121 課題

1923078 室谷優

1)

平均値は、データと同一単位を持つ値であり、データの数値の合計をデータの個数で割った値である。すべてのデータに1/nのウェイトを与えているため、外れ値の影響を受けやすい。そのため直感と反する値を取ることがあるので代表値として用いる際には注意が必要である。

分散は、各データと平均値の差の2乗の和をデータの個数で割ることで求められる。散らばりを指標化するに当たり、中心を定め、各データが中心からどの程度離れているかを求めるという考えに基づいている。したがって、各データと平均値の差の2乗の平均値によってデータ分布全体の散らばりを計っている。

標準偏差は、分散の正の平方根である。分散が元データの2乗の単位を持つのに対し、標準偏差は、元データと同一の単位を持つ散らばりの指標である。分散と標準偏差は1：1の関係性にあるため、必要に応じ、適切な方を用いるべきである。

共分散行列は、2次元以上のデータにおいて、分散と共分散を同時に表したものである。対角頂にはそれぞれのデータの分散が入り、それ以外には共分散が入る。

相関行列は、共分散では値の尺度によって数値が変化するため、その値を-1~1にする。すなわち、相関係数に変化させたものである。

2)

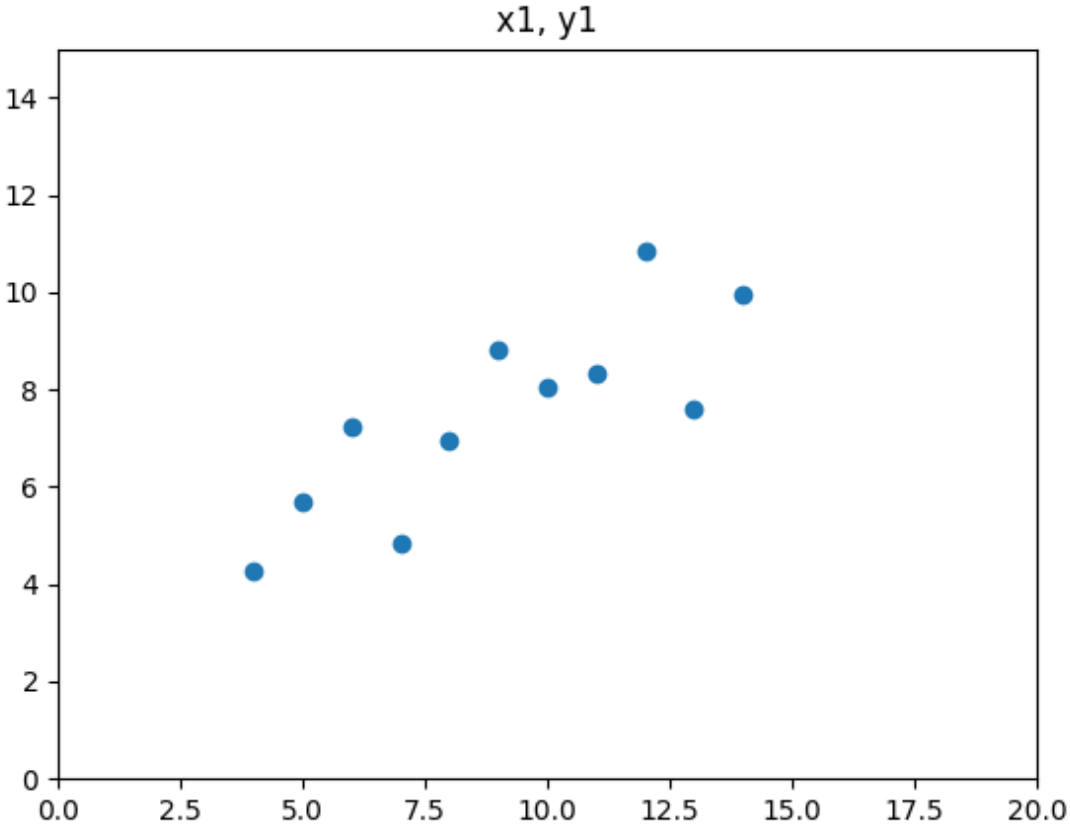
散布図の描画、及び値の算出には以下のプログラムを使用した。

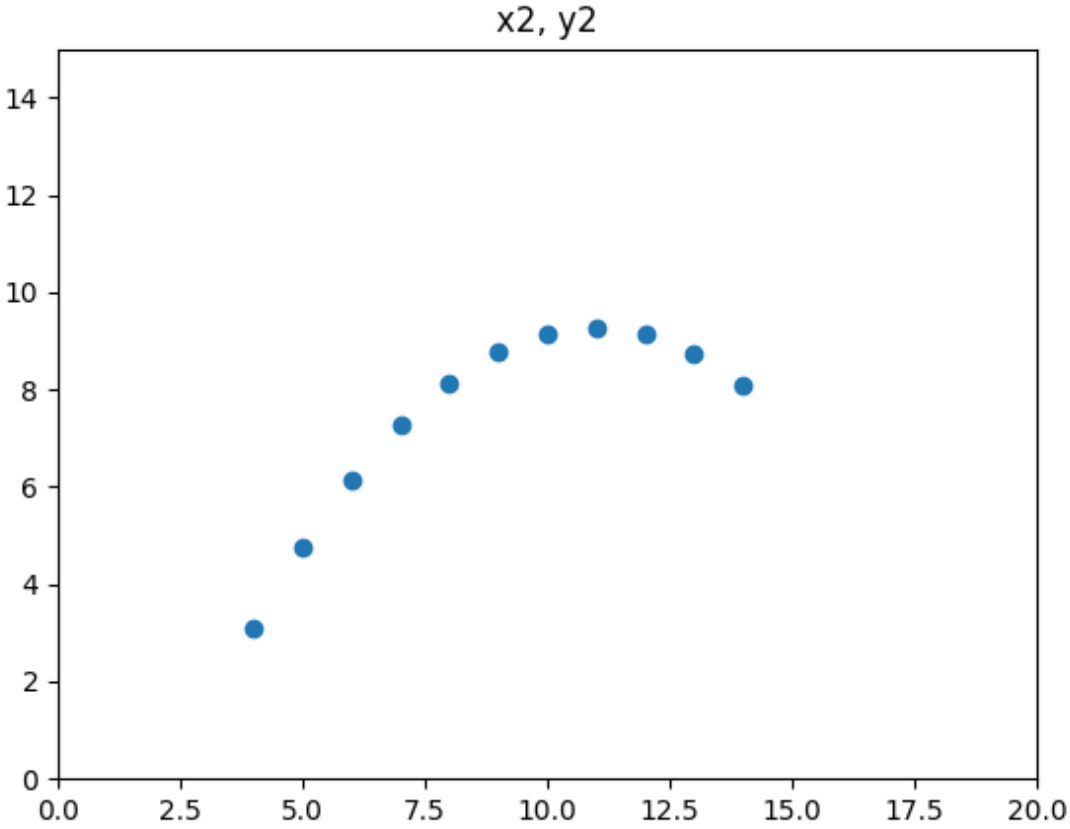
```
#coding: utf-8

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("~/docs/20221121_exam.csv")
for i in range(1, 5):
    x = df["x{}".format(i)].values
    y = df["y{}".format(i)].values
    print(" ----- data{} ----- ".format(i))
    print("平均値\tx:{:.15f}\ty:{:.15f}".format(x.mean(), y.mean()))
    print("分散\tx:{:.15f}\ty:{:.15f}".format(x.var(), y.var()))
    print("共分散行列\n{}".format(np.cov(x, y, ddof = 0)))
    print("相関行列\n{}".format(np.corrcoef(x, y, ddof = 0)))
    plt.figure()
    plt.title("x{0}, y{0}".format(i))
    plt.xlim(0, 20)
    plt.ylim(0, 15)
    plt.scatter(x, y)
    plt.savefig("./img/20221121/data{}.png".format(i))
```

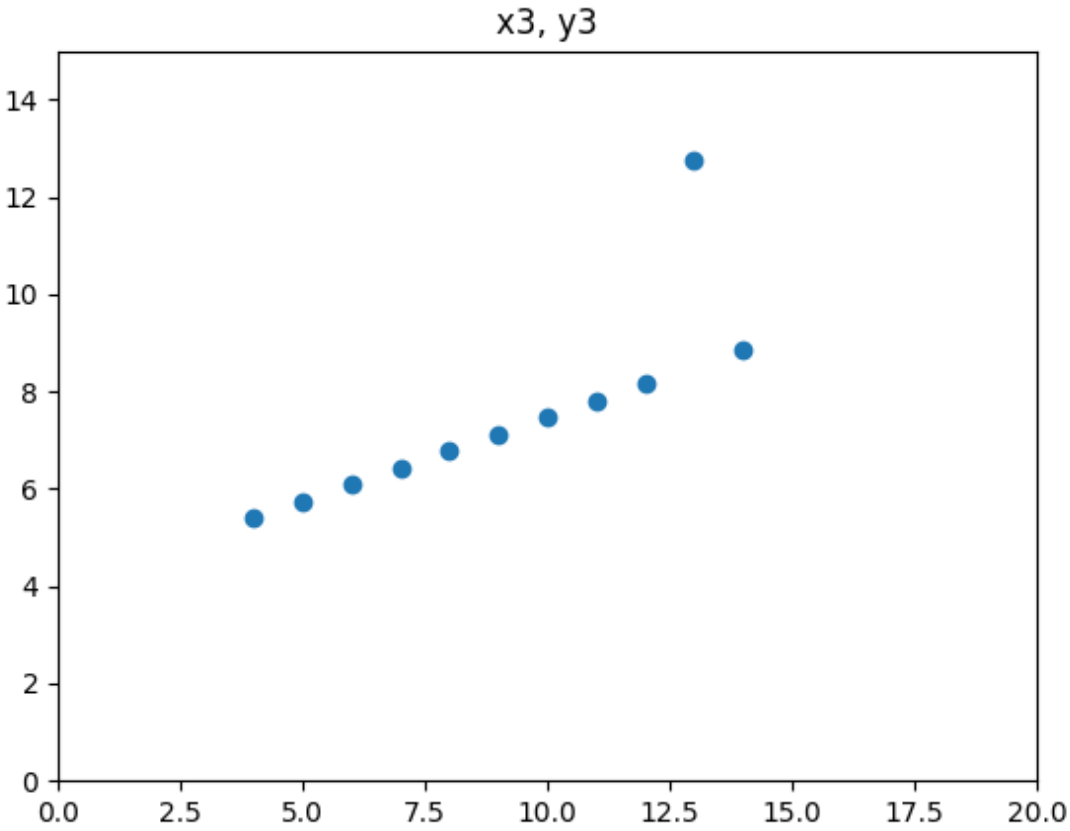
その結果、以下の数値が得られた。

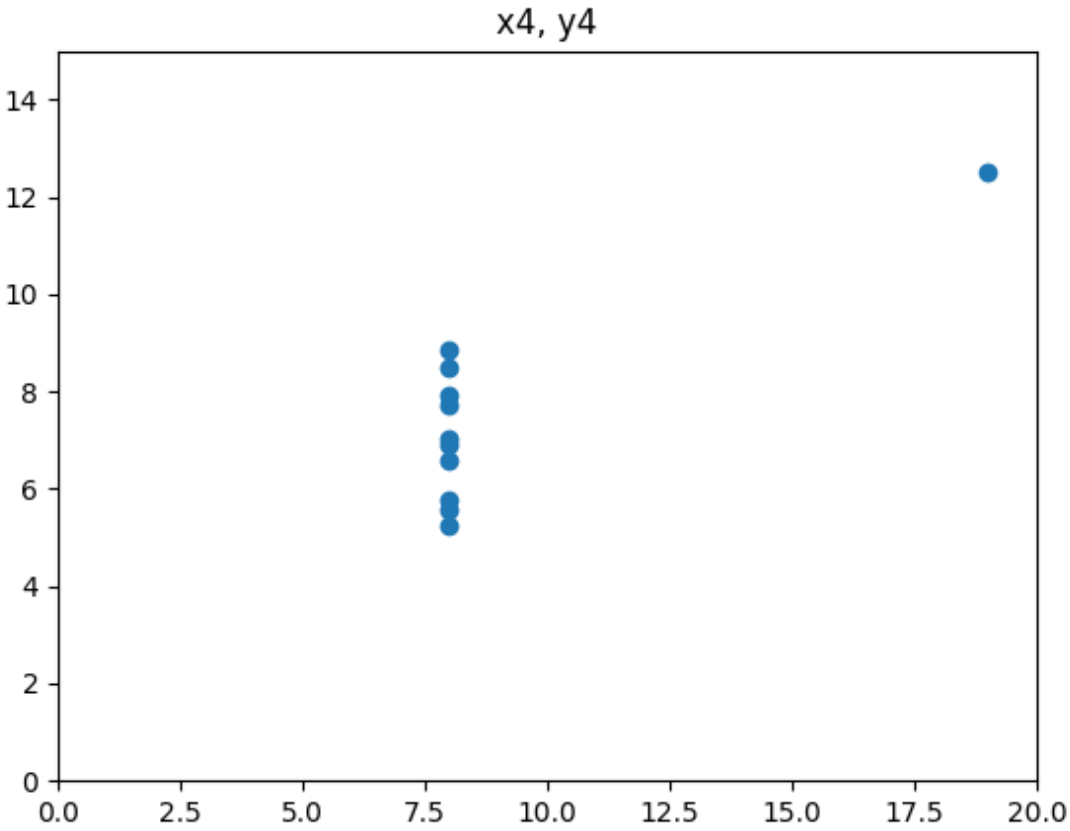
項目	x1, y1	x2, y2	x3, y3	x4, y4
平均値	x:9.000000000000000 y:7.500909090909093	x:9.000000000000000 y:7.500909090909090	x:9.000000000000000 y:7.500000000000000	x:9.000000000000000 y:7.500909090909091
分散	x:10.000000000000000 y:3.752062809917355	x:10.000000000000000 y:3.752390082644628	x:10.000000000000000 y:3.747836363636364	x:10.000000000000000 y:3.748408264462810
共分散行列	[[10. 5.00090909] [5.00090909 3.75206281]]	[[10. 5.] [5. 3.75239008]]	[[10. 4.99727273] [4.99727273 3.74783636]]	[[10. 4.99909091] [4.99909091 3.74840826]]
相関行列	[[1. 0.81642052] [0.81642052 1.]]	[[1. 0.81623651] [0.81623651 1.]]	[[1. 0.81628674] [0.81628674 1.]]	[[1. 0.81652144] [0.81652144 1.]]

また、描画した散布図は次のようになった。


x1, y1
0
2
4
6
8
10
12
14
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


x2, y2
0
2
4
6
8
10
12
14
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0





data1~4において、平均値、分散、共分散行列、相関行列はほとんど差異が見られなかった。しかし、散布図を見てみると4角データの組はそれぞれ異なる傾向を持っていると考えられる。したがって、データセットを

表現する場合には統計量だけでは不十分であると考えられる。

今回のデータセットはアンスコムの例(wikipedia)の一つである。