1　相関係数

前回，最小二乗誤差の考えをもとに，線形回帰分析により回帰直線を求めた．これにより，線形関係のある値間の大雑把な関係を求めることができた．

しかし，どの程度関係があるのかはわかっていない．求められた分析の結果を評価する場合に，相関係数を求めることが多い．相関係数とは二つのものの間にどの程度関連があるかを数値で客観的に表すものであり，統計学，機械学習などを扱う際よく使用される指標である．そのため，今回は相関係数について学んだ．

下記の図が使用するデータである．

顧客番号	年齢 (x)	年間支出額 (y)
1	32	40,000
2	23	38,000
3	39	46,000
4	45	47,000
5	47	48,000

表 1　顧客の年齢とその人が年間にかける化粧品代のデータ 1

顧客番号	年齢 (x)	年間支出額 (y)
1	32	39,500
2	23	39,000
3	39	47,000
4	45	47,000
5	47	47,600

表 2　顧客の年齢とその人が年間にかける化粧品代のデータ 2

2　共分散

共分散は一緒に変化する２つの変数の傾向を測る尺度であり，数列 X,Y が存在するとき，それら X,Y の平均からの距離である偏差は，$x_n - \bar{x}$ と $y_n - \bar{y}$ になる．

$\bar{x}$ は X の標本平均，$\bar{y}$ は Y の標本平均である．X と Y の変動が一緒であれば，これらの偏差は同じ符号になる．

２つの偏差を掛け合わせた場合，偏差が同じ符号を持つならば積は正になり，異なる符号を持っている場合，負の数になる．そのため，積の和が X と Y が一緒に変動するかどうかを表す尺度になる．共分散とはこれらの偏差の積の平均である．

以下が式となる．

$$(X, Y \text{ の共分散}) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

表１のデータから共分散を求めるプログラム

```
1  import numpy as np
2
3  X = np.array([32,      23,      39,      45,      47])
4  Y = np.array([40000, 38000, 46000, 47000, 48000])
5
6  # X,Y の平均を求める
7  meanX = np.mean(X)
8  meanY = np.mean(Y)
9
10 # X,Y の共分散を求める
11 covXY = np.dot(X−meanX, Y−meanY) / len(X)
12 print (covXY)
```

実行結果

```
1  34440.0
```

４，５行目のデータを表２のデータに書き換えた

```
1  X = np.array([32,      23,      39,      45,      47])
2  Y = np.array([39500, 39000, 47000, 47000, 47600])
```

実行結果

```
1  31696.0
```

共分散は，おおよその見当は付けられるものの，X と Y の単位の積であるため，今回は歳-円という単位になるが，この単位自体には意味がなく解釈が難しい．そのため次の相関係数といわれるものがよく統計の評価に使われる．

3　相関係数

共分散の単位の問題を解決するために，今回では X，Y の共分散から X の標準偏差と Y の標準偏差をかけたものを割る．標準偏差は各データの値を平均の差の２乗の合計をデータの個数で割った値の平方根したもので，以下の式で表される

$$\text{標準偏差} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

X の 標準偏差と Y の標準偏差をかけたものを割ることで単位が消され，数値として２変数間の関係が表すことができる．

$$\frac{(X, Y \text{ の共分散})(\cancel{\text{歳}} \cancel{\text{円}})}{(X \text{ の標準偏差})(\cancel{\text{歳}})(Y \text{ の標準偏差})(\cancel{\text{円}})}$$

よって X,Y の相関係数は以下のように表される．

$$(X, Y \text{ の相関係数}) = \frac{(X, Y \text{ の共分散})}{(X \text{ の標準偏差})(Y \text{ の標準偏差})}$$

表１のデータから相関係数を求めるプログラム

```
1  import numpy as np
2
3  X = np.array([32,      23,      39,      45,      47])
4  Y = np.array([40000, 38000, 46000, 47000, 48000])
5
6  # X,Y の標準偏差を求める
7  stdX = np.std(X)
8  stdY = np.std(Y)
9
10 # X,Y の共分散を求める
11 covXY = np.dot(X−np.mean(X), Y−np.mean(Y)) / len(X)
12
13 # X,Y の相関係数を求める
14 print ( covXY / (stdX ∗ stdY) )
```

実行結果

```
1  0.971548171177
```

４，５行目のデータを表２のデータに書き換えた

```
1  X = np.array([32,      23,      39,      45,      47])
2  Y = np.array([39500, 39000, 47000, 47000, 47600])
```

実行結果

```
1  0.920684967629
```

相関係数は，－ 1 以上 1 以下の実数を値にとり，相関係数が正のとき確率変数には正の相関が，負のとき確率変数には負の相関がある．

一般的には以下のような関係にあると言われている．だが変数間の因果関係を説明するものではないので注意が必要である．

相関係数	相関の強さ
0.0 ~ ±0.2	(ほとんど) 相関がない
0.2 ~ ±0.4	弱い相関がある
0.4 ~ ±0.7	相関がある
0.7 ~ ±0.9	強い相関がある
0.9 ~ ±1.0	(ほぼ) 完全な相関がある

表 3　相関係数と相関の関係を表したもの

しかし相関係数を使い，２つの変数の関係が詳しく調べると，一方から他方を予測することができる．今回であったら結果がそれぞれ 0.97, 0.92 だったので図 1 の左上のような散布図になることが予想することができる．

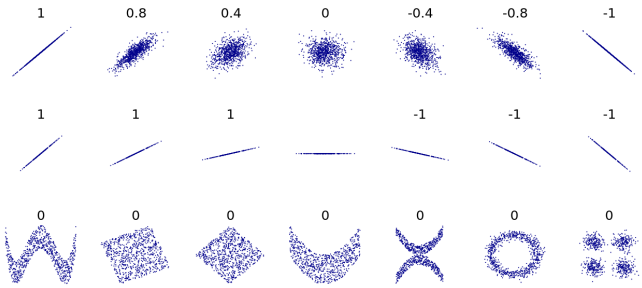

図 1　x，y の組と相関係数を表したもの [**?**]

参考文献

[1] NumPy Reference
https://docs.scipy.org/doc/numpy/reference/index.html

[2] Allen B. Downey (2015) Think Stats —— プログラマのための統計入門
(著，黒川 洋・訳，黒川 利明)　O'Reilly Japan

[3] 相関係数 - Wikipedia
https://ja.wikipedia.org/wiki/