統計学ノートNo. 3
記述統計（2変量）

担当教員：蛭川雅之

研究室：紫英館3階345研究室

メール： hirukawa@econ.ryukoku.ac.jp

オフィスアワー：毎週金曜日3講時

1　散布図

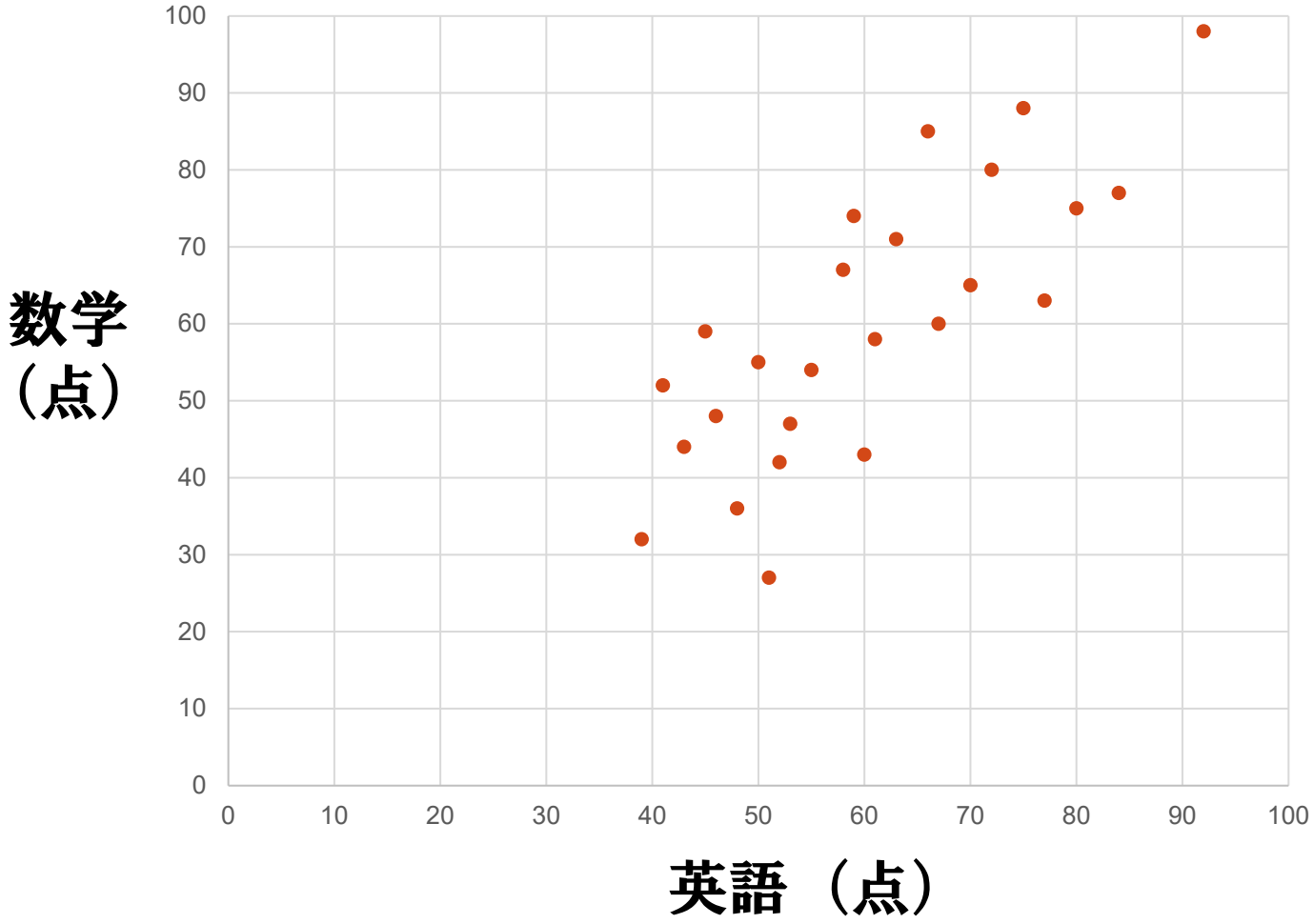
複数の変数の相互関係

- **これまで、1変数に焦点を絞って縮約を学んできた。**
- **現実には、2個以上の変数の関連に関心がある場合が多い。**
  - **身長と体重**
  - **ある科目の試験の得点と学習時間**
  - **複数の科目の試験の得点**
  - **複数の銘柄の株価収益率**

散布図

- データファイル“data2.xlsx”には生徒25人の英語と数学のテストの得点を示したものである。
- 表を一見しただけで英語・数学の得点の関係を読み取るのは難しい。
  - 2科目の得点の関係を把握しやすい図が必要である（⇒**散布図**）。

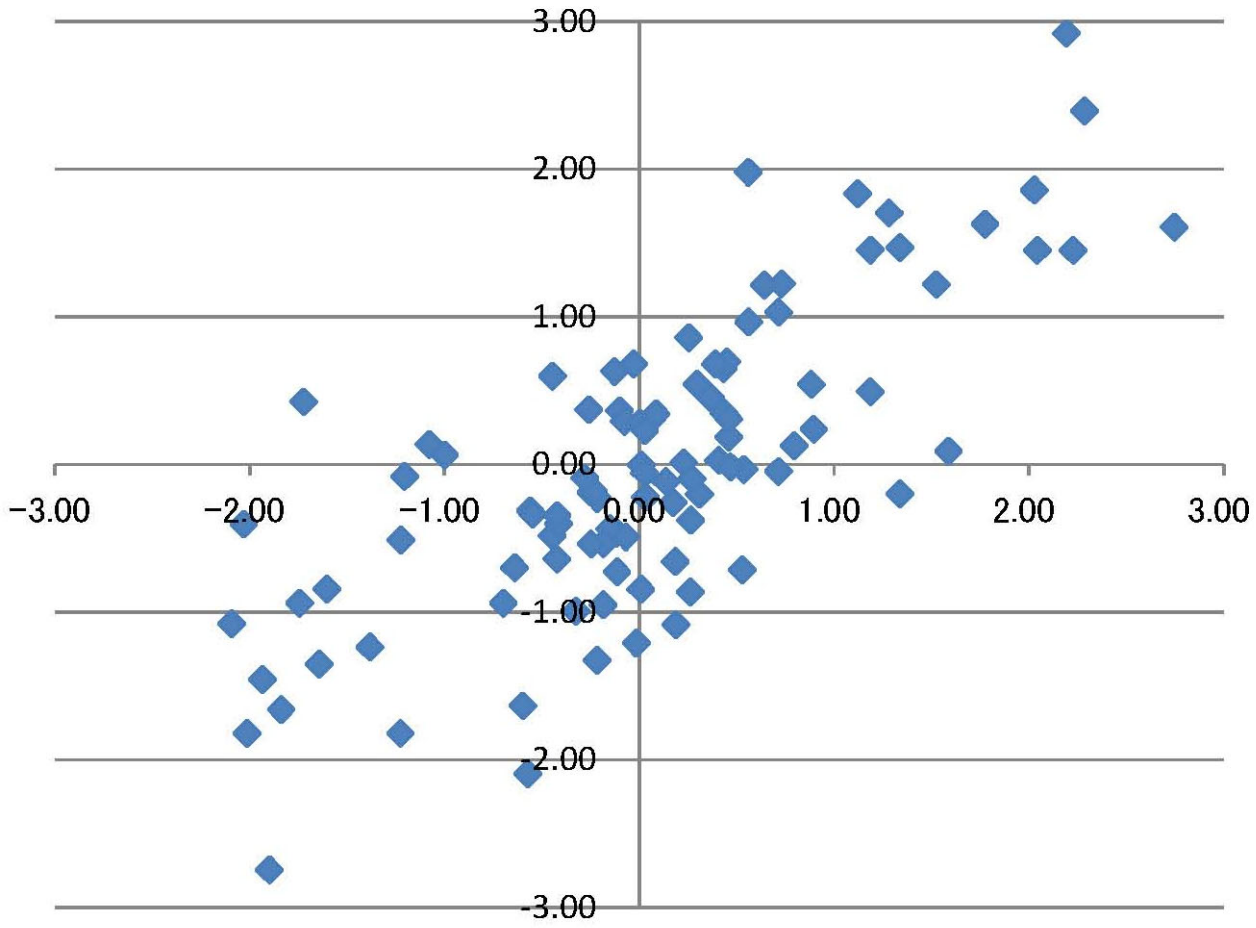
散布図（つづき）


数学（点）
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
英語（点）

正の相関

- **一方が増加（減少）すれば他方も増加（減少）する傾向があるとき、２変数間に正の相関があるという。**
- **例：**
  - **身長と体重**
  - **価格と供給量…**

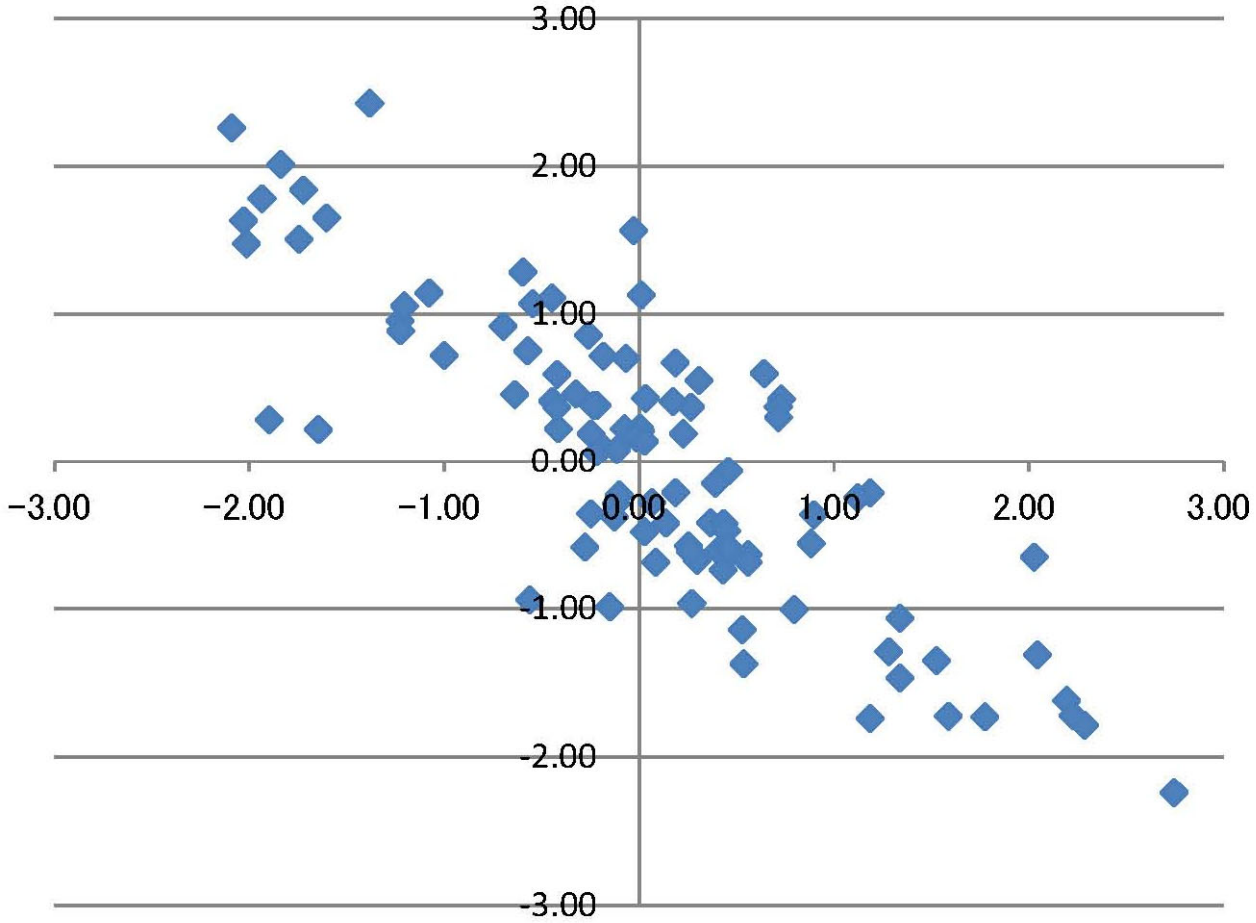
正の相関（例）



負の相関

- **一方が増加（減少）すれば他方が減少（増加）する傾向があるとき、２変数間に負の相関があるという。**
- **例：**
  - **価格と需要量**
  - **物価上昇率と失業率（フィリップス曲線）…**

負の相関（例）

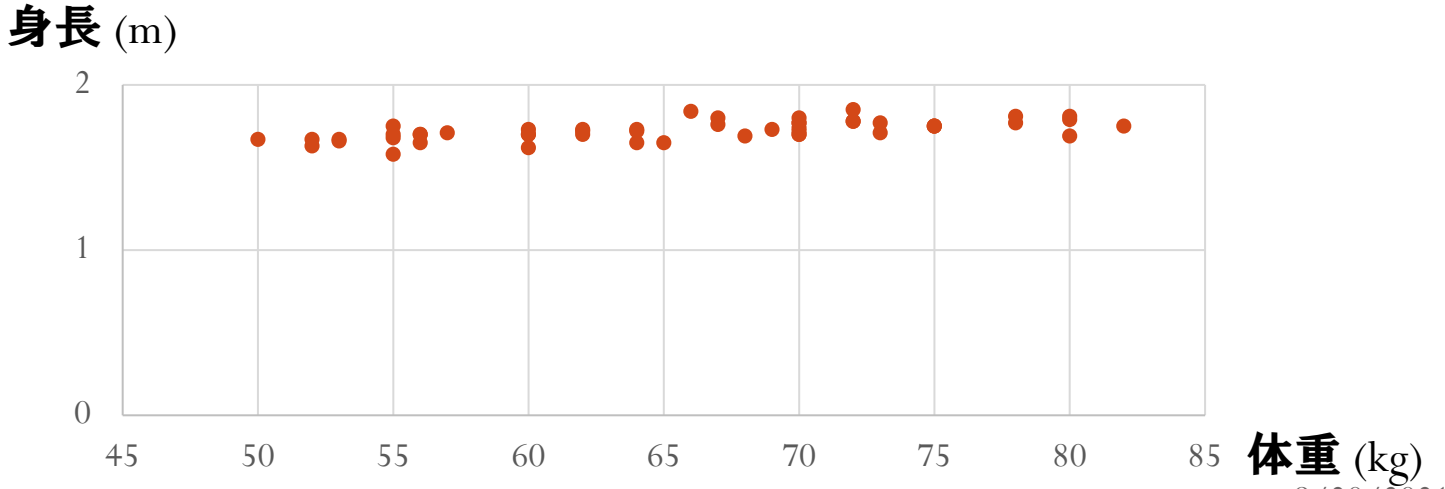


2　共分散と相関係数

散布図の問題点

- **散布図を用いると、２変数間の関連を視覚的に把握できる。**
- **散布図の軸のとり方や縦横の比によっては、情報を読み誤る危険性がある。**

■**悪い例：**



共分散

- 2変数の関係の強さを数値で表すことを考える。
- 2変数の関係の尺度として、**共分散**（covariance）を

$$(XY\text{の共分散}) = \frac{\text{“}(X\text{の偏差}) \times (Y\text{の偏差})\text{”の和}}{(\text{全データ数})}$$

と定義する。

共分散（つづき）

- **n組のデータ**

$$\{(X_i, Y_i)\}_{i=1}^{n} = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

に対し、共分散$\hat{\sigma}_{XY}$は

$$\hat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\left(\text{ただし}\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i\right)$$

と表現できる。

共分散（つづき）

- **分散の場合と同様、以下の関係を用いると共分散の計算が簡単になる場合が多い。**

$$\hat{\sigma}_{XY} = \frac{\textbf{(生データの積和)}}{\textbf{(全データ数)}} - \textbf{(平均値の積)}$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}$$

共分散：数値例

- **英語・数学の得点の例：**

$$\text{(共分散)} = \frac{\text{“(英語の偏差)} \times \text{(数学の偏差)” の和}}{\text{(全データ数)}}$$

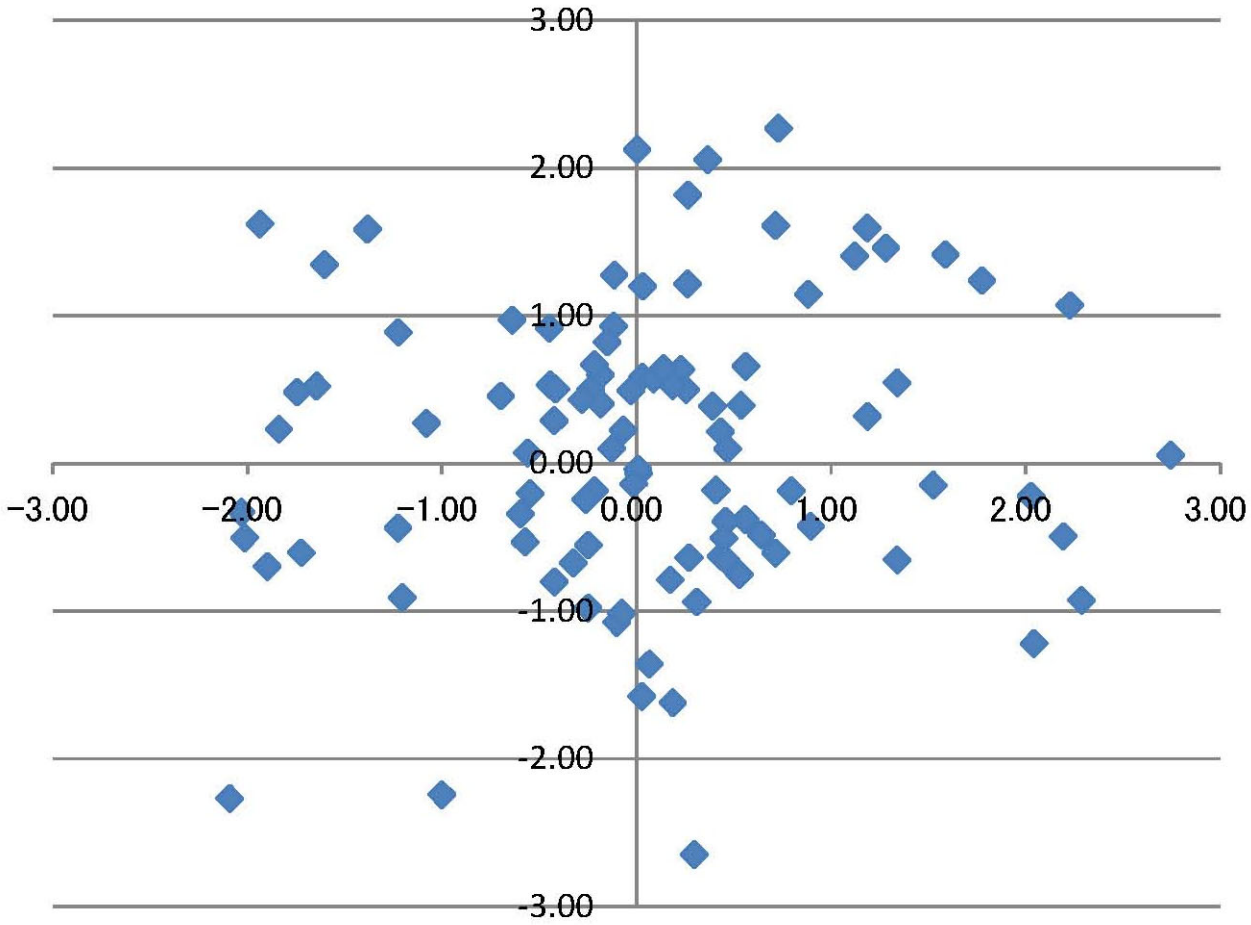
$$= \frac{4933}{25} = 197.32$$

- Excel**の関数**covar**からも同じ結果が得られる。**

共分散のとり得る値

- **2変数間に正の相関があるとき、共分散は正の値をとる。**
- **2変数間に負の相関があるとき、共分散は負の値をとる。**
- **2変数間の共分散がゼロとなる場合を無相関という。**
  - **注意：2変数が無相関であっても「関係がない」とは言い切れない場合がある。**

無相関（例）



相関係数

- **共分散の値は2変数の単位に依存する。**
- **変数の単位に依存しない尺度として、相関係数（correlation coefficient）を次のように定義する。**

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

相関係数（つづき）

- **相関係数r_{XY}は**

$$r_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

と表現できる。

相関係数：数値例

- **英語・数学の得点の例：**

$$\text{(相関係数)} = \frac{\text{(得点の共分散)}}{\text{(英語の標準偏差)(数学の標準偏差)}} \approx \frac{197.32}{13.901 \times 17.808} \approx 0.797$$

- Excelの関数`correl`からも同じ結果が得られる。

相関係数のとり得る値

- **相関係数は必ず-1以上1以下の値をとる。**
  - **２変数間に正の相関があるとき、相関係数は正の値をとる。**
  - **２変数間に負の相関があるとき、相関係数は負の値をとる。**

相関係数のとり得る値（つづき）

- **2変数が右上がり（右下がり）の直線上にある場合、相関係数は1（-1）となる。**
  - **<u>注意</u>：2変数が水平線（＝傾きゼロの直線）上にあるとき、相関係数は定義されない（縦軸方向の分散がゼロとなるため）。**

順位相関係数

- **スピアマンの順位相関係数とは「順位データに対する（ピアソンの積率）相関係数」と定義される。**
  - **通常、単に相関係数といえば「ピアソンの積率相関係数」を指す。**

順位相関係数（つづき）

- **順位相関係数は、以下のような場合に利用される。**
  1. **数量・計測値でなく順位のみが与えられている場合。**
     - **例：「好み」の順位、「レース」の着順…**
  2. **ピアソンの積率相関係数が「外れ値」の影響を受けていると思われる場合。**
- **順位相関係数は必ず-1以上1以下の値をとる。**

順位相関係数（つづき）

- **順位相関係数は、直接的には以下の手順で計算できる。**

1. **関数`rank`を用いて各変数を順位データに変換する。**
2. **2つの順位データに対し、関数`correl`を適用する。**

- **英語・数学の得点の順位相関係数は0.787となる。**

順位相関係数（つづき）

- **n組のデータ**

$$\{(X_i, Y_i)\}_{i=1}^{n} = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

の中に同順位がないと仮定する。

- **この場合、順位相関係数ρ_{XY}は**

$$\rho_{XY} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (X_i - Y_i)^2$$

と書き直すことができる。

順位相関係数（つづき）

- **実際には、前ページの結果を踏まえた以下の計算手順がよく用いられる。**
  1. **関数**`rank`**を用いて各変数を順位データに変換する。**
  2. **関数**`sumxmy2`**を用いて**$\sum_{i=1}^{n}(X_i - Y_i)^2$**を計算する。**
  3. 2**の結果を利用して、順位相関係数を求める。**
  - **英語・数学の得点の順位相関係数はやはり**0.787**となる。**

Computer Exercise 3

- データファイル“data2.xlsx”を講義ウェブページからダウンロードし、Excelを利用して英語・数学の得点に関する散布図を作成せよ。さらに、これらの

  ①共分散

  ②相関係数

  ③順位相関係数

  も計算せよ。

3　加工されたデータに対する平均値・標準偏差等の計算

各データに定数を加えた場合

- **各データに定数bを加える**と、

1. **新データの平均値は**

$$(\text{元データの平均値}) + b$$

2. **新データの分散は元データの分散と同一。**

3. **新データの標準偏差は元データの標準偏差と同一。**

- **偏差は変わらない点に注意！**

各データを定数倍した場合

- **各データを$a(>0)$倍すると、**

1. **新データの平均値は**

   $a \times$(元データの平均値)

2. **新データの分散は**

   $a^2 \times$(元データの分散)

3. **新データの標準偏差は**

   $a \times$(元データの標準偏差)

各データを一次変換した場合

- **各データを$a(>0)$倍しさらに定数bを加えると、**

1. **新データの平均値は**

$$a \times (\text{元データの平均値}) + b$$

2. **新データの分散は**

$$a^2 \times (\text{元データの分散})$$

3. **新データの標準偏差は**

$$a \times (\text{元データの標準偏差})$$

一次変換の例1：標準化

- **データを**

$$\frac{(\text{偏差})}{(\text{標準偏差})} = \frac{(\text{データ}) - (\text{平均値})}{(\text{標準偏差})}$$

と変形することを「標準化（基準化）する」という。

- **標準化して得られる新データの**

1. **平均値は0**
2. **分散・標準偏差はともに1**

一次変換の例2：2変量

- **次のような変形を考える。**
  - **データXを$a(>0)$倍しさらに定数bを加える。**
  - **データYを$c(>0)$倍しさらに定数dを加える。**
- **このとき、**
  1. **新データの共分散は**

     $$ac \times (\text{元データの共分散})$$

  2. **新データの相関係数（順位相関係数）は元データの相関係数（順位相関係数）と同一。**