

経済統計学講義ノート No.3

記述統計（2 変量）

蛭川雅之

2022 年 9 月 12 日

目 次

1. 散布図と相関
2. 共分散
3. 相関係数

1 散布図と相関

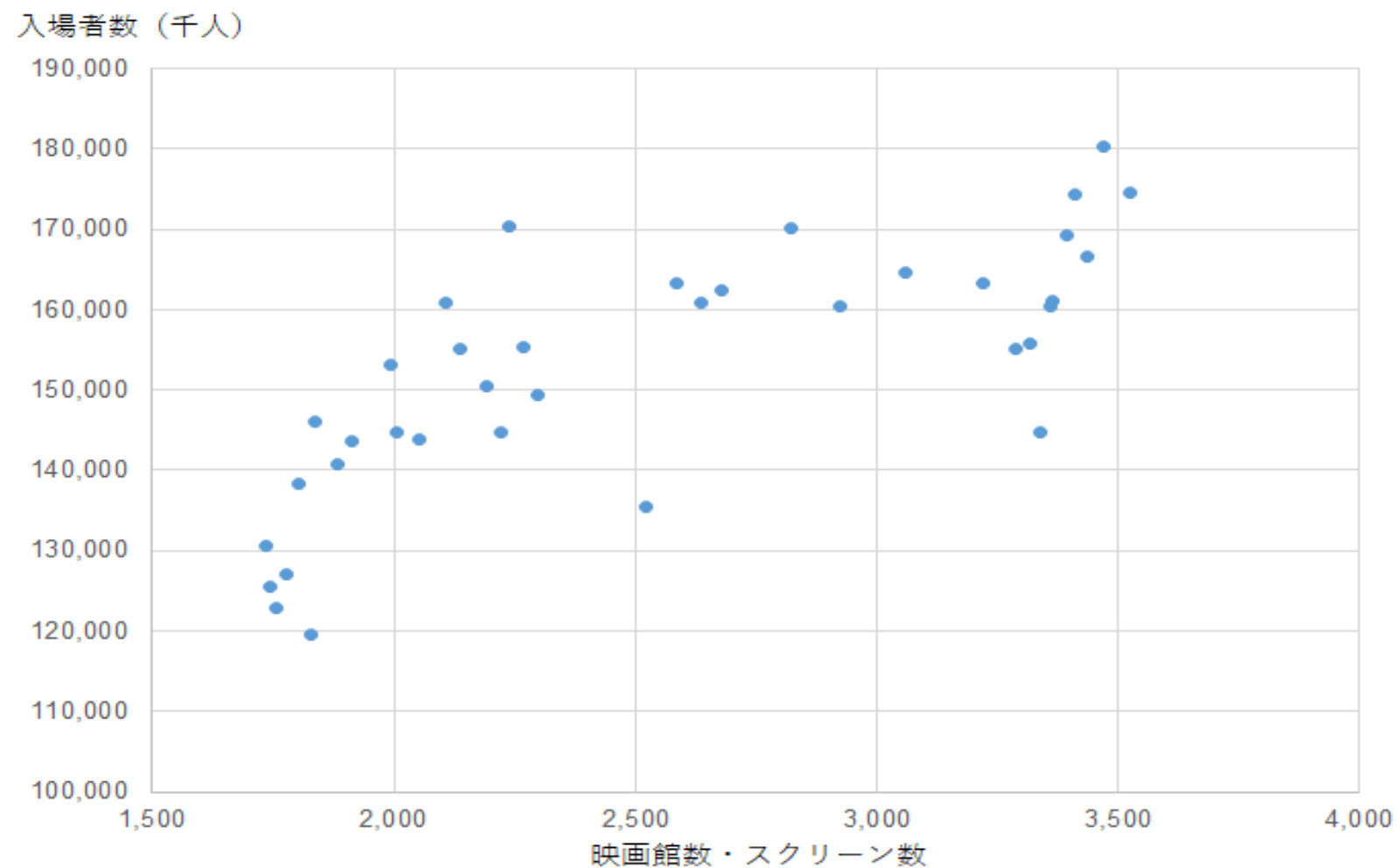
1.1 イントロダクション

- 前回までは 1 変数に話題を限定して、データの特徴を記述・描写する方法を学んできた。
- 実際には、複数の変数の関連（相互依存関係）に関心のある場合が多い。
 - － 身長と体重
 - － 複数の科目の試験の得点
 - － ある財の価格と取引量
 - － GDP と民間消費
 - － 利子率と設備投資
 - － 複数の銘柄の株価収益率...

1.2 映画に関する時系列データ

- Excel ファイル cinema2.xlsx は、1981 年から 2017 年までの 37 年分の全国の映画館数（1999 年以前）・スクリーン合計数（2000 年以降）と入場者数を一覧表にまとめたものである【出典：一般社団法人日本映画製作者連盟】。
- 表を一見しただけでこれらの関係を読み取るのは困難である。
 - － 2 つの変数の関係を把握しやすい図（**散布図**）が必要である。

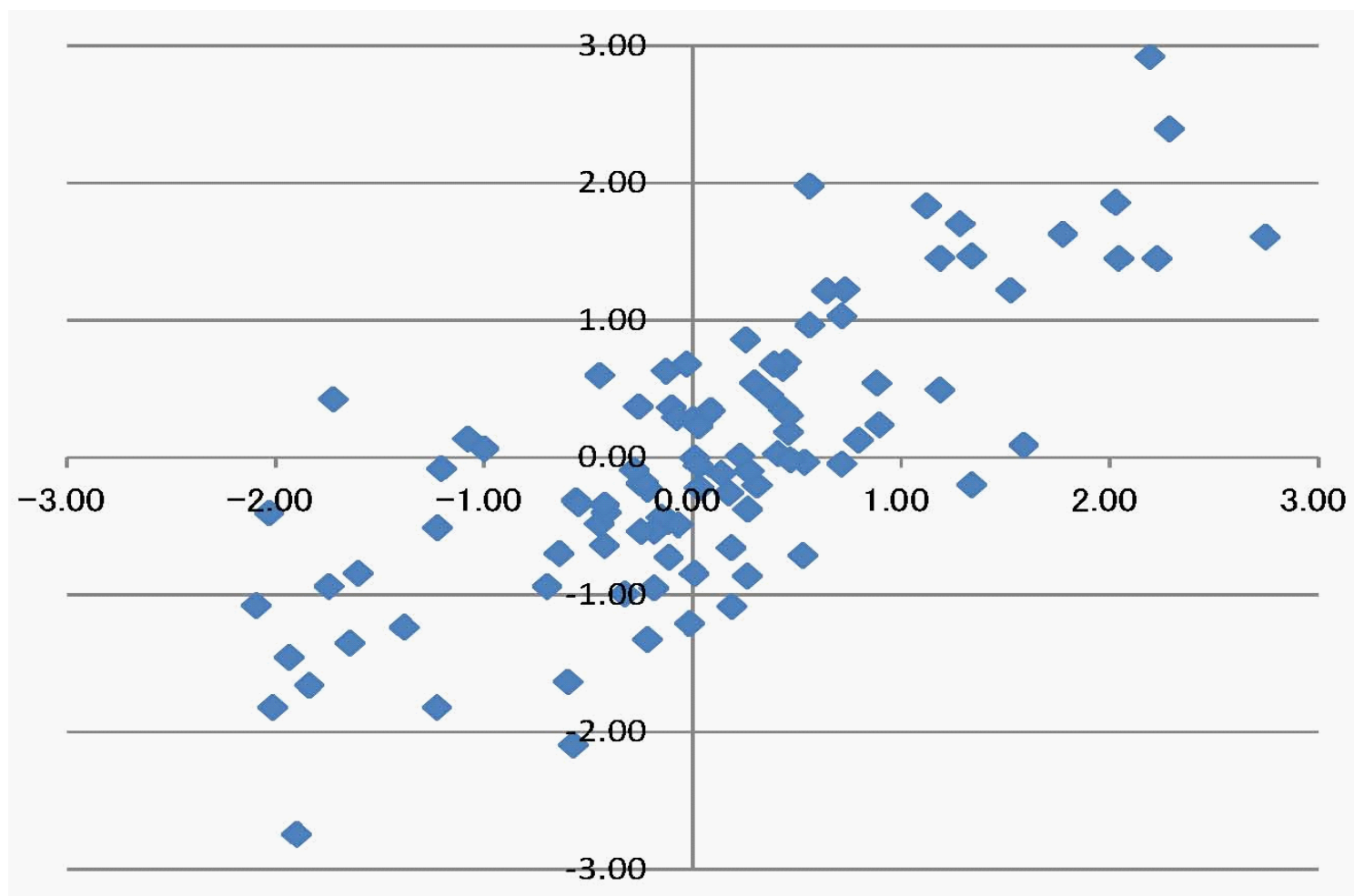
1.3 散布図（見本）



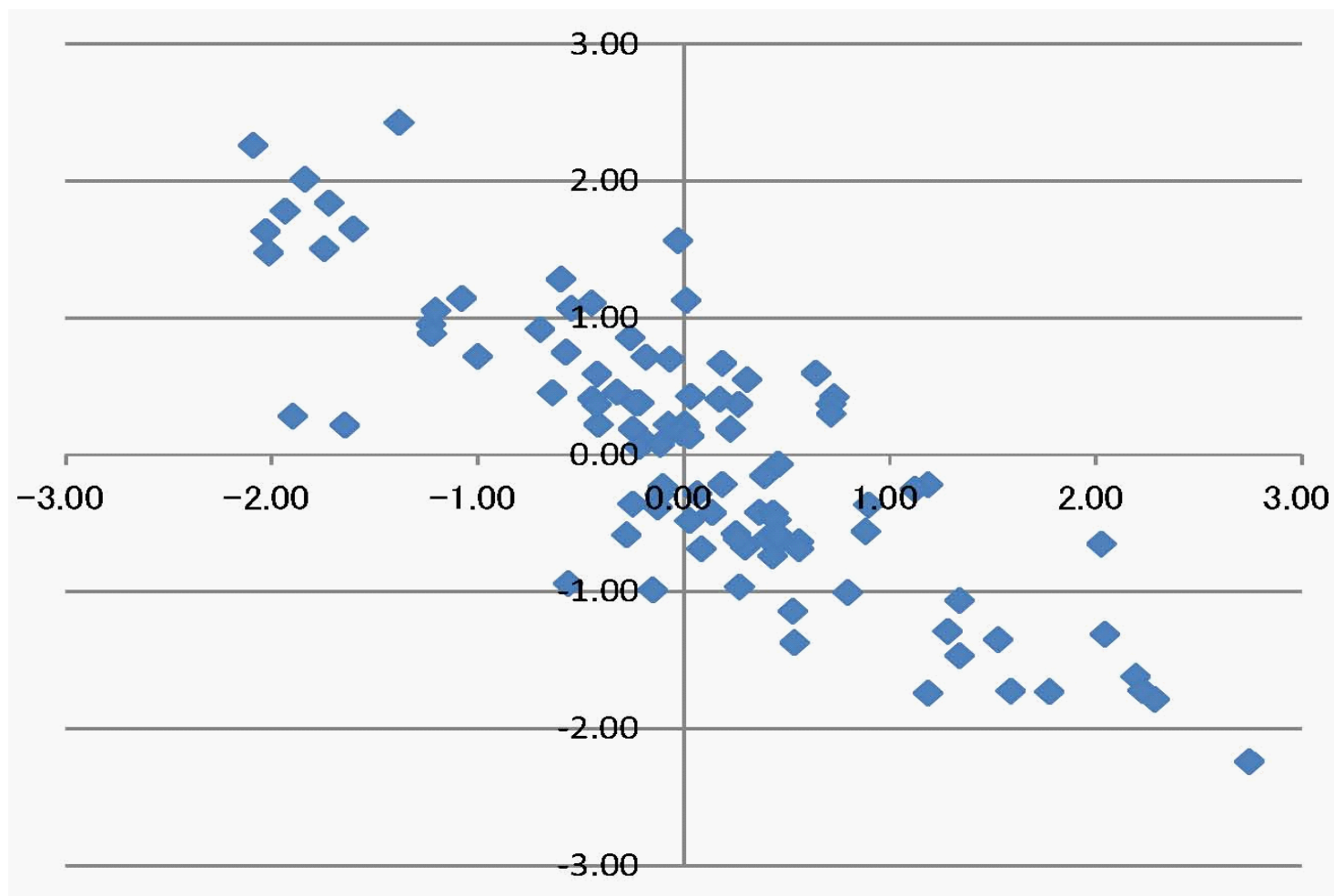
1.4 正の相関・負の相関

- 一方が増加（減少）すれば他方も増加（減少）する傾向が存在するとき、2つの変数の間に**正の相関**があるという。
 - － 例：身長と体重、価格と供給量、GDP と民間消費...
- 一方が増加（減少）すれば他方は減少（増加）する傾向が存在するとき、2つの変数の間に**負の相関**があるという。
 - － 例：価格と需要量、利子率と設備投資、物価上昇率と失業率（フィリップス曲線）...

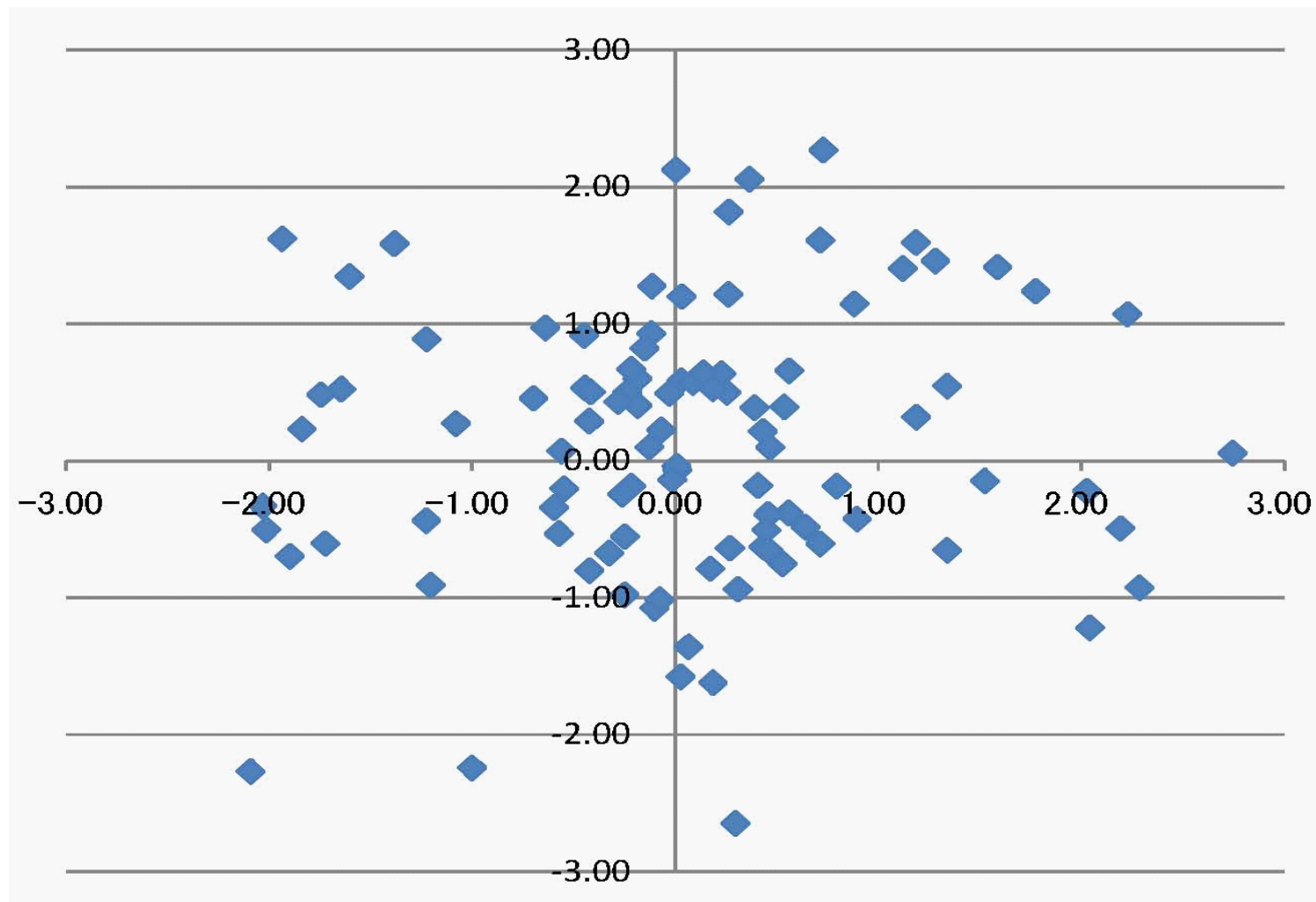
■正の相関（例）



■負の相関（例）



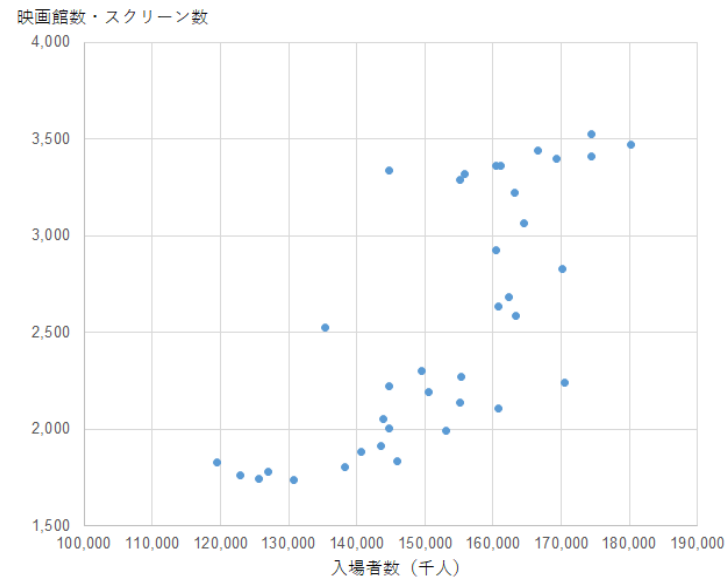
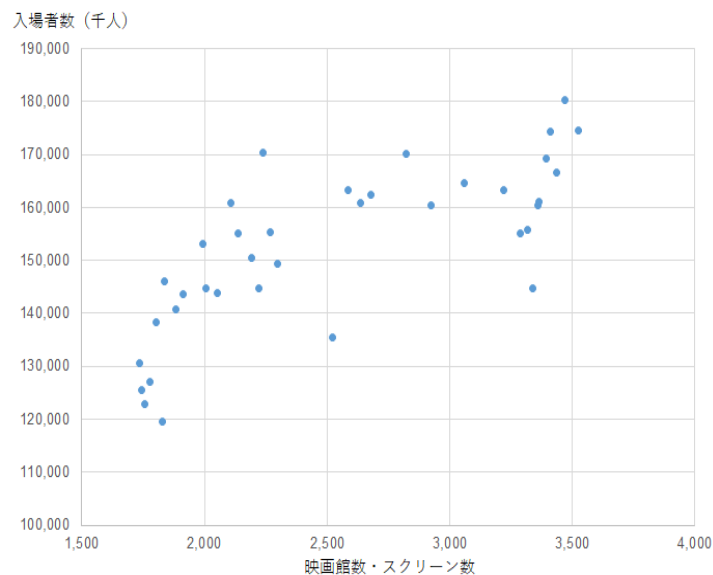
■無相関（例）



2 共分散

2.1 散布図の問題点

- 散布図を用いると、2 変数間の関係を視覚的に把握できる。
- その一方、作図の方法によっては情報を読み誤る危険がある。



2.2 共分散

- 2 変数間の関係の‘強さ’を数値で表すことを考える（＝代表値）。
- n 個のデータ $(X_1, Y_1), \dots, (X_n, Y_n)$ に対し、2 変数間の関係の尺度として**共分散**

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

を考える。

- － 標本平均回りの**偏差**の積の平均値である。
- － 関数 `covar` がこの数値を与える。

2.3 共分散の取りうる値

- 2 変数間に正の相関がある場合、共分散は**正**の値をとる。
- 2 変数間に負の相関がある場合、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

1. 2 変数間に非線形の関係がある。
 - 計算上無相関となる場合があるが、「関係が存在しない」とは言い切れない。
2. 2 変数が水平線（＝傾きゼロの直線）上に並ぶ。
 - 必ず無相関となる。

3 相関係数

3.1 共分散の問題点

- 共分散の正負は 2 変数間の正負の相関に対応する。
 - － 共分散の数値はそれぞれの変数の計測単位に依存する。
- 計測単位によらない尺度はないか？
 - － **ピアソンの積率相関係数**

$$r_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

は計測単位に依存しない。

- － 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。
- － 関数 `correl` がこの数値を与える。

3.2 ピアソンの積率相関係数を取りうる値

- ピアソンの積率相関係数は必ず-1 以上 1 以下の値をとる。
 - 2 変数間に正の相関がある場合、相関係数は正の値をとる。
 - 2 変数間に負の相関がある場合、相関係数は負の値をとる。
 - 2 変数が無相関の場合、相関係数はゼロとなる。
- 2 変数が正（負）の傾きを持つ直線上に並ぶ場合、ピアソンの積率相関係数は 1（-1）となる。
 - 2 変数が水平線（＝傾きゼロの直線）上に並ぶ場合、縦軸方向の変数の分散がゼロとなり、相関係数は定義されない。

3.3 スピアマンの順位相関係数

- **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。
- スピアマンの順位相関係数は、以下のような場合に利用される。
 - － 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...
 - － 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。
- スピアマンの順位相関係数は必ず-1 以上 1 以下の値をとる。

- スピアマンの順位相関係数は、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。
- n 個の順位データ $(X_1, Y_1), \dots, (X_n, Y_n)$ の中に同順位がないという仮定の下で、スピアマンの順位相関係数は

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

と書き直すことができる。

- 実際には、この結果を利用し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. 上式に基づいて順位相関係数を求める。