相関係数

関連性の尺度

- 問題
  - 安静時の最高血圧が高い人は排尿直後の最高血圧も高い，安静時の最高血圧が低い人は排尿直後の最高血圧も低いといった関連性がいえるだろうか？
  - 関連性があるとすれば，その強さをどのように表現するか？

- 20歳の女性の「安静時の最高血圧」と「排尿直後の最高血圧」を調べたデータ

被験者番号	1	2	3	4	5	6	7	8	9	10	11
安静時の最高血圧 x	100	118	110	114	106	106	116	94	98	102	124
排尿直後の最高血圧 y	110	150	144	130	140	108	124	114	122	120	146

変量間に関連性が見られるデータ

- 変量間に関連性が見られるデータ
  - 喫煙と肺がんの関係
  - 植物の栄養状態と子実生産量
  - 姉の身長と妹の身長
  - 入学時の成績と卒業時の成績
  - 耕地面積と農業所得

  等
- 関連性の強さの表現法について考える

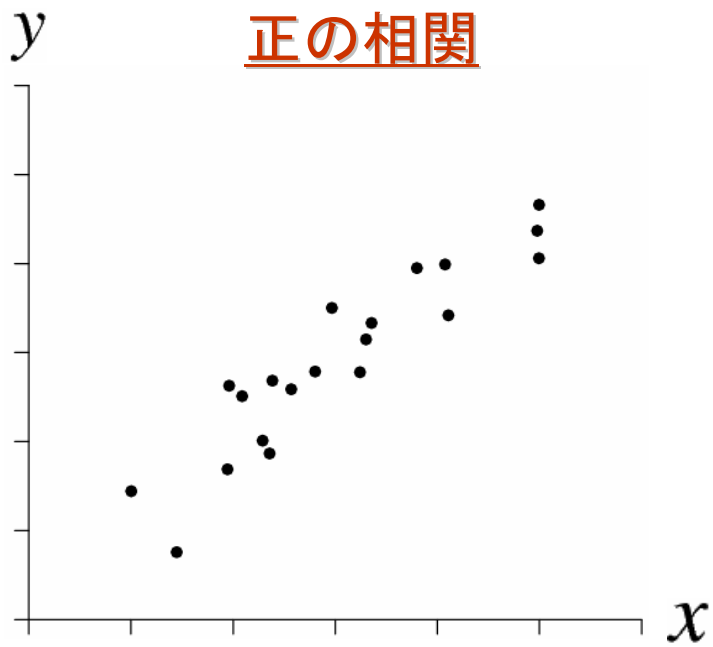
相関係数

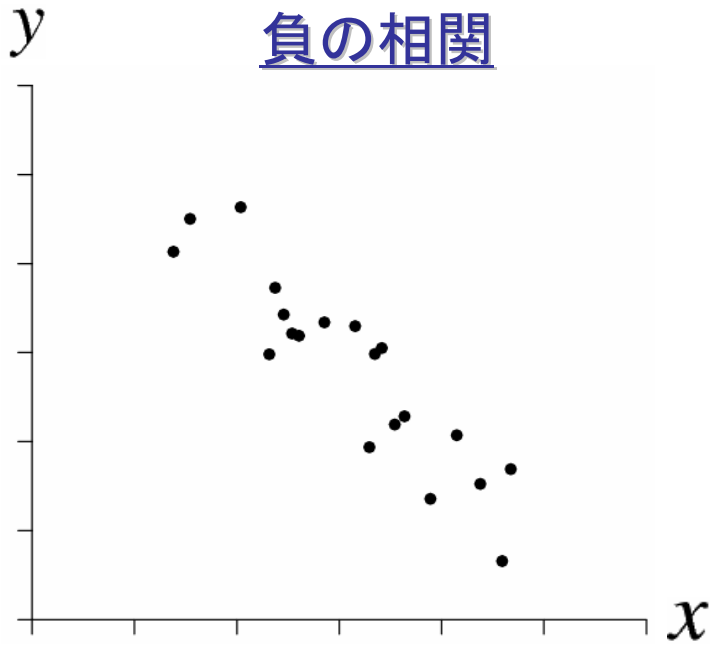
- 記号
  - 大きさ n の標本について，２つの変量 x と y とを調べた結果を $(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)$ とする
- 相関係数

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}\sqrt{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}}$$

  - 2つの変量の直線的な関連の強さを表す尺度
  - 正の相関：一方の値が増すとき，他方の値も増す関係
  - 負の相関：一方の値が増すとき，他方の値が減る関係

正の相関と負の相関



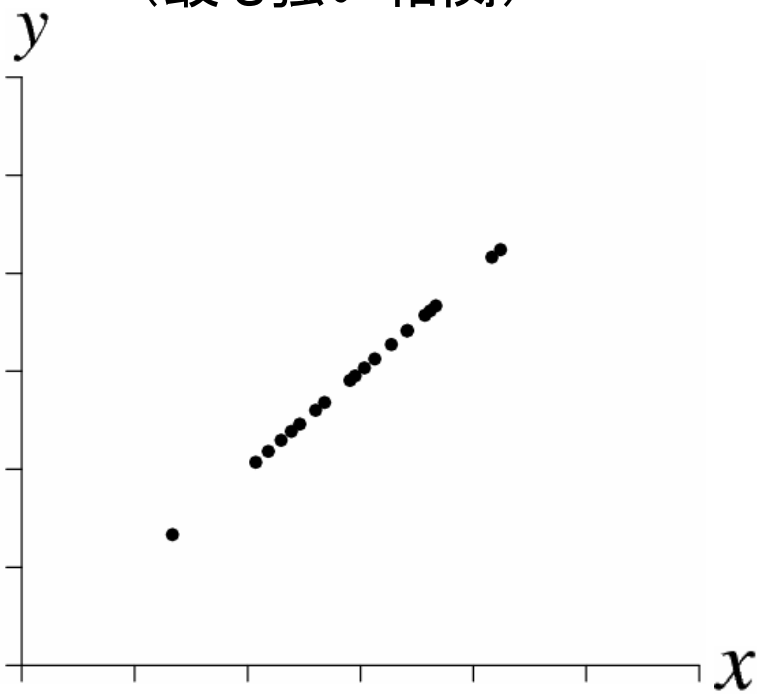


相関の強さ

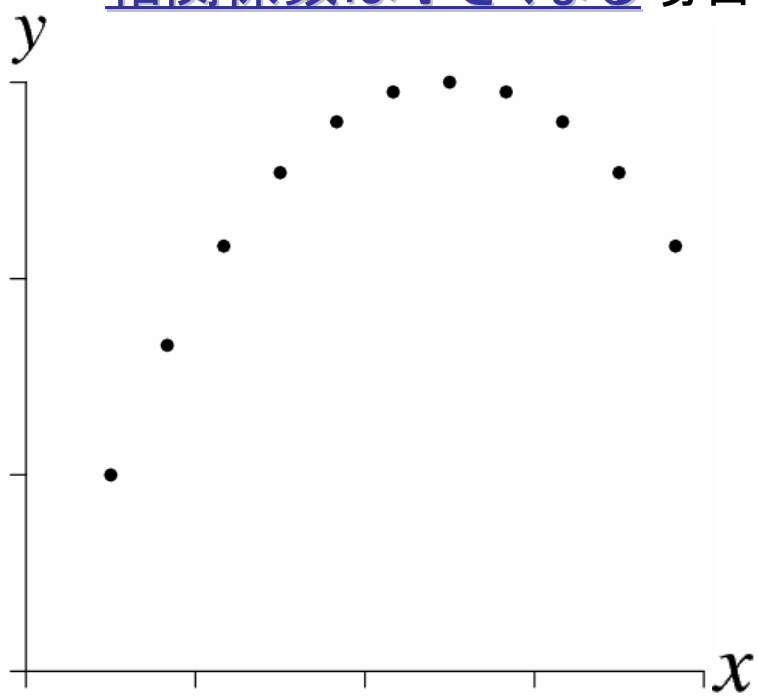
- 相関係数の取り得る値 : $-1 \leq r \leq 1$
- 相関の強さ
  - $|r|$ が大きいほど直線的関連が強い
  - $|r|$ が小さいほど直線的関連が弱い
  - $r = \pm 1$ のとき, すべての標本は一直線上に並ぶ
    一方の値が与えられれば, もう一方の値も定まる
    最も強い直線的関係
  - 相関係数は曲線的な関連性を表現することができない

相関の強さ

相関係数が1の場合
（最も強い相関）



曲線的に強い相関があるが
相関係数は小さくなる場合



例：相関係数の求め方

- 20歳の女性の「安静時の最高血圧」と「排尿直後の最高血圧」との相関係数を求める

被験者番号	1	2	3	4	5	6	7	8	9	10	11
安静時の最高血圧 x	100	118	110	114	106	106	116	94	98	102	124
排尿直後の最高血圧 y	110	150	144	130	140	108	124	114	122	120	146

- **相関係数の定義**

$$r = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}\sqrt{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}}$$

- 各変数の平均値 ： $\bar{x} = 108, \quad \bar{y} = 138$

相関係数の計算

- 相関係数の分母の計算
  - $$\sqrt{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2} = \sqrt{(100 - 108)^2 + \cdots + (124 - 108)^2}$$
    $$= \sqrt{(-8)^2 + \cdots + (16)^2}$$
    $$= \sqrt{864} = 29.4$$
  - $$\sqrt{(y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2} = \sqrt{(110 - 128)^2 + \cdots + (146 - 128)^2}$$
    $$= \sqrt{(-18)^2 + \cdots + (18)^2}$$
    $$= \sqrt{2248} = 47.4$$

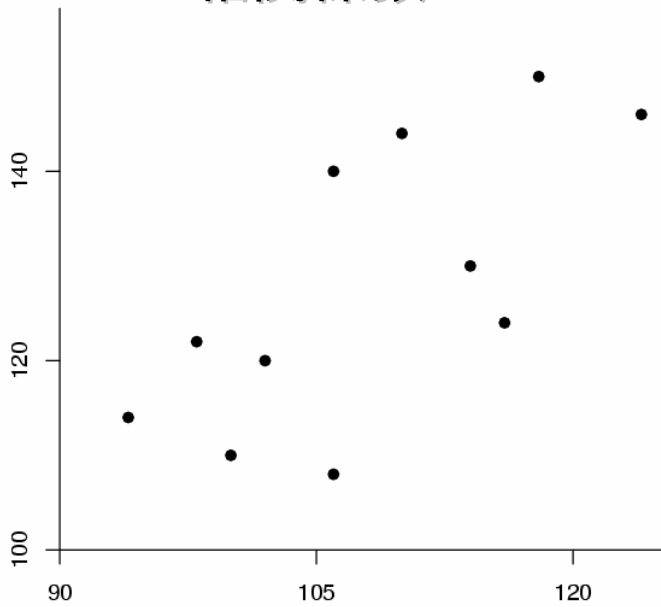
相関係数と散布図

- 相関係数の分子の計算
  - $$(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})$$
    $$= (-8) \times (-18) \times \cdots \times 16 \times 18 = 984$$

- 相関係数
  - $$r = \frac{984}{29.4 \times 47.4} = 0.71$$

右上がりの直線関係が見られる

相関係数　0.71



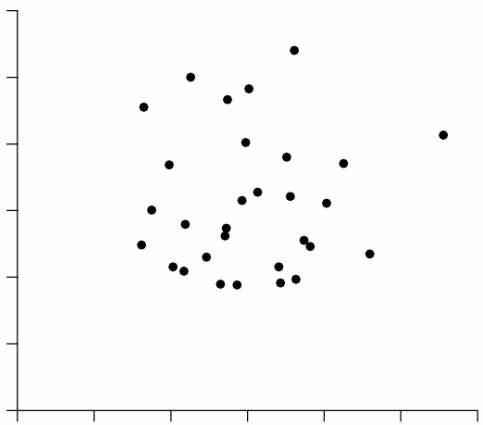
相関係数の大きさ

相関係数の大きさと関連の強さ

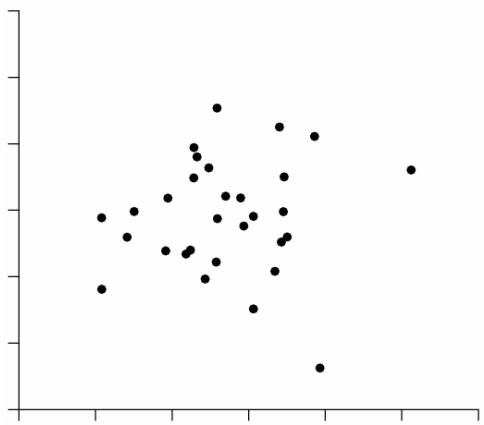
- 相関係数の大きさ
  - 相関係数の絶対値は0から1の間の値
  - 直線的な関連の強さは相関係数0．5が中間といえるか？
  - 実際に散布図を描くと，相関係数0．7程度が中間と考えられる
  - 相関係数がどの程度大きくなったときに，直線的な関連性を見て取れるか？

相関係数の大きさと散布図

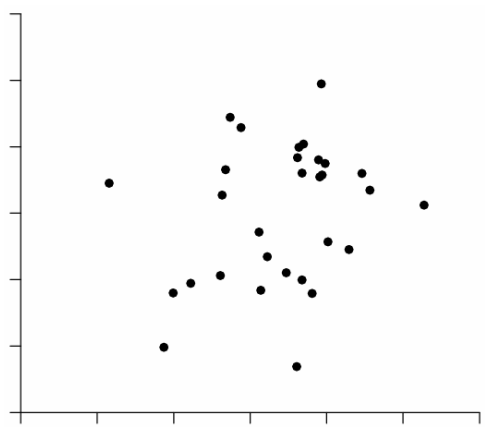
相関係数 0



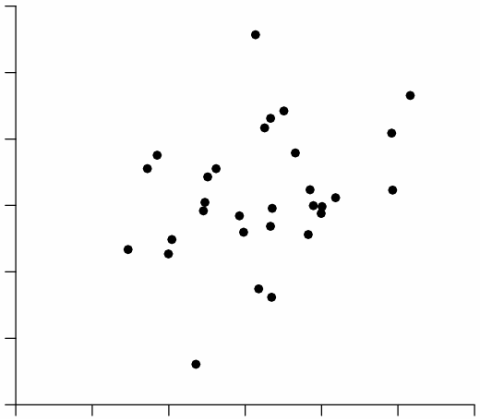
相関係数 0.1



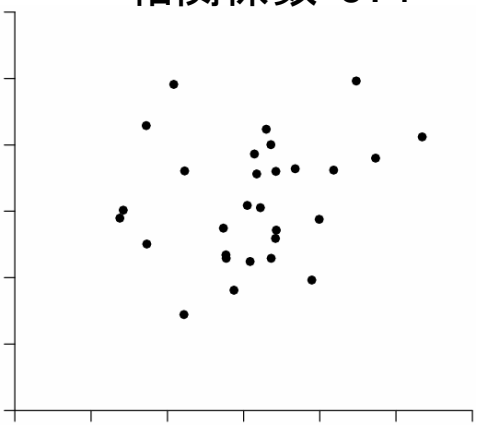
相関係数 0.2



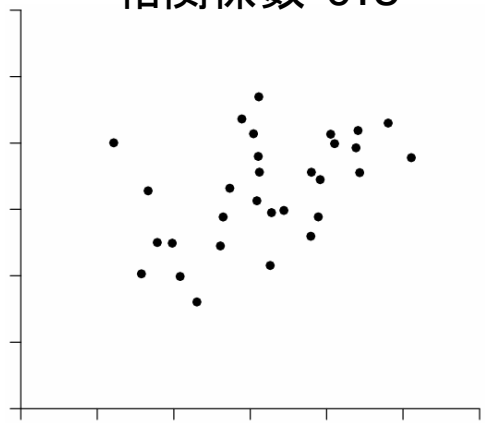
相関係数 0.3



相関係数 0.4

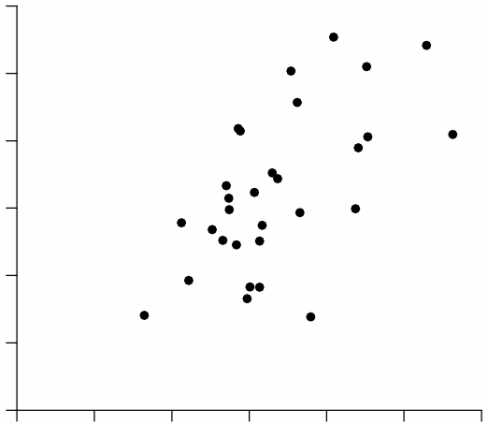


相関係数 0.5

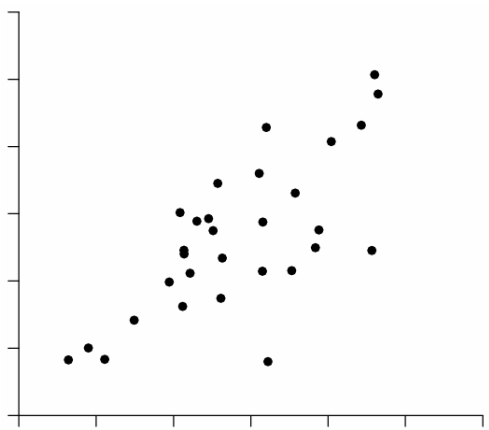


相関係数の大きさと散布図

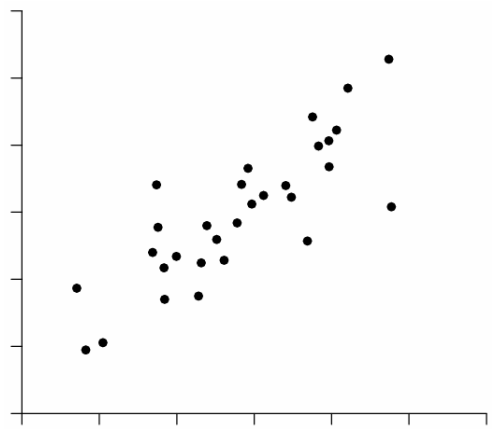
相関係数 0.6



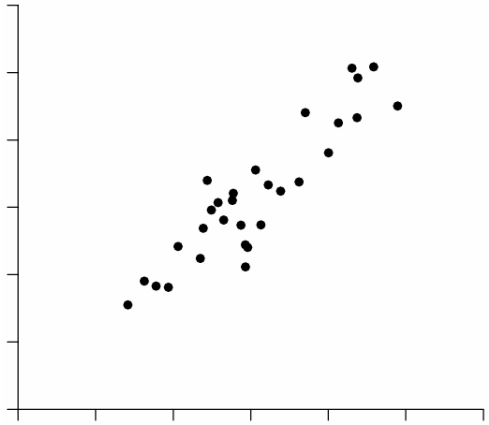
相関係数 0.7



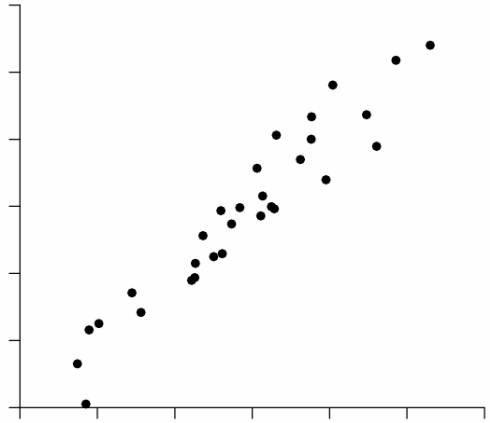
相関係数 0.8



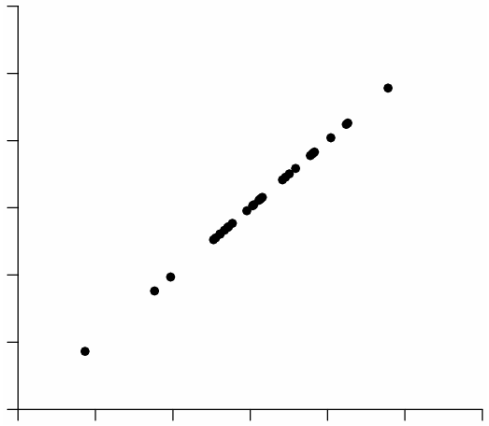
相関係数 0.9



相関係数 0.95



相関係数 1.0



相関係数の検定

相関係数の検定

目的：姉と妹の身長の間に相関関係があるか調べる

- データ
  - 姉か妹がいる年齢18歳から25歳の範囲の女性14組を無作為に抽出し，姉妹の身長を測定

姉妹の組	1	2	3	4	5	6	7
姉の身長 x	149	157	166	156	158	156	152
妹の身長 y	153	167	160	152	163	168	160
姉妹の組	8	9	10	11	12	13	14
姉の身長 x	148	155	160	165	152	163	163
妹の身長 y	162	154	157	162	152	157	155

帰無仮説と検定統計量

- 帰無仮説
  - 姉の身長と妹の身長の間には相関関係はない
  - $H_0 : \rho = 0, \quad \rho$: 相関係数
- 検定統計量とその分布
  - 検定統計量

$$\frac{\sqrt{n-2}r}{\sqrt{1-r^2}} \ (= t \text{とおく})$$

  - 帰無仮説のもとで, 検定統計量 t は自由度 $n-2$ の t 分布にしたがう

相関係数と検定統計量

- 相関係数
  - $r = \dfrac{54.56}{\sqrt{427.71}\,\sqrt{362.86}} = 0.138$
- 検定統計量
  - $t = \dfrac{\sqrt{n-2}\,r}{\sqrt{1-r^2}} = \dfrac{\sqrt{12} \times 0.138}{\sqrt{1-0.138^2}} = 0.483$
- t 分布の両側5%点

  (-2.179，2.179)

検定結果と考察

- 検定結果
  - 帰無仮説は有意水準5%で棄却できない
- 考察
  - 「妹と姉の身長には相関関係がある」と考えて「相関関係はない」という帰無仮説を立てたが，棄却できなかった
  - 今回の分析結果は，標本数が14と少ないことから，標本抽出の偶然性に左右された可能性がある