



高知工科大学 経済・マネジメント学群

計量経済学

3. 二変数の関係

やない ゆう き
矢内 勇生



<https://yukiyanai.github.io>



yanai.yuki@kochi-tech.ac.jp



今日の目標

- 2つの変数の関係を調べる方法を理解する

1. 質的変数のとき

- クロス表（分割表）
- 独立性の検定

2. 量的変数のとき

- 散布図
- 相関係数

質的変数と量的変数

- 質的変数の例：性別、支持 vs. 不支持、大学の成績（S, A, B, C, F）、好きなスポーツ
- 量的変数の例：身長、体重、年齢、年収

クロス表（分割表, contingency table）

（例）性別と内閣支持の関係

	現在の内閣を		
	支持しない	支持する	計
男性	200	300	500
女性	250	250	500
計	450	550	1000

注目するのは行か列か（1）

- 問題ごとに行（row）と列（column）のどちらに注目するか考える
- 例の場合：
 - 行：性別ごとに内閣支持・不支持に差があるか
 - 列：内閣の支持・不支持によって女性の割合は異なるか

注目するのは行か列か (2)

行に注目

→行の合計を100%にする

	不支持	支持	計
男性	40%	60%	100%
女性	50%	50%	100%

列に注目

→列の合計を100%にする

	不支持	支持
男性	44%	55%
女性	56%	45%
計	100%	100%

性別によって内閣支持率は異なるか

- 標本：女性より男性のほうが内閣支持の割合が大きい

➡母集団でも男性の支持率のほうが高いといえる？

➡検定：独立性の検定

表：性別と内閣支持の関係

	不支持	支持	計
男性	200 (40%)	300 (60%)	500 (100%)
女性	250 (50%)	250 (50%)	500 (100%)
計	450 (45%)	550 (55%)	1000 (100%)

独立性の検定

- ・ クロス表で提示される2変数に関連があるかどうか調べるための検定
- ♣ 内閣支持率に男女間で差がない
- = 性別と内閣支持に関連がない
- = 性別と内閣支持は独立
- ➡ 「独立性の検定」
- ▶ χ^2 分布を利用するので、「 χ^2 [カイ二乗] 検定」とも呼ぶ

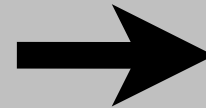
独立性の検定の帰無仮説と対立仮説

- 帰無仮説：2変数は独立である（関連がない）
 - 対立仮説：2変数は独立ではない（関連がある）
- (例)
- H_0 : 性別と内閣支持には関連がない
 - H_1 : 性別と内閣支持には関連がある

独立性の検定（ χ^2 検定）の考え方

帰無仮説

	不支持	支持	計
男性	45%	55%	100%
女性	45%	55%	100%
計	45%	55%	100%



実際に観測されたデータ

	不支持	支持	計
男性	200 (40%)	300 (60%)	500 (100%)
女性	250 (50%)	250 (50%)	500 (100%)
計	450 (45%)	550 (55%)	1000 (100%)

このようなサンプルはあり得ない？

帰無仮説が正しいとすれば

帰無仮説

	不支持	支持	計
男性	45%	55%	100%
女性	45%	55%	100%
計	45%	55%	100%



帰無仮説の下で 期待されるデータ

	不支持	支持	計
男性	225 (45%)	275 (55%)	500 (100%)
女性	225 (45%)	275 (55%)	500 (100%)
計	450 (45%)	550 (55%)	1000 (100%)

期待度数

帰無仮説が正しい場合の χ^2 値（検定統計量）を求める

$$\chi_0^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$$

▶ i は行を表す（ k は行の数）

▶ j は列を表す（ m は列の数）

– 観測度数 $_{ij}$ は i 行 j 列の観測度数

すべてのセルで $\frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}}$ を求めて、合計すればよい

例題の場合の検定統計量を求める

観測度数

	不支持	支持
男性	200	300
女性	250	250

期待度数

	不支持	支持
男性	225	275
女性	225	275

$$\begin{aligned}
 \chi_0^2 &= \sum_{i=1}^k \sum_{j=1}^m \frac{(\text{観測度数}_{ij} - \text{期待度数}_{ij})^2}{\text{期待度数}_{ij}} \\
 &= \frac{(200 - 225)^2}{225} + \frac{(300 - 275)^2}{275} + \frac{(250 - 225)^2}{225} + \frac{(250 - 275)^2}{275} \\
 &\approx 2.78 + 2.27 + 2.78 + 2.27 \\
 &= 10.1
 \end{aligned}$$

統計量を何と比較する？

- カイ二乗分布の臨界値と比較する
 - カイ二乗分布は自由度によって形が変わる
 - クロス表の場合：自由度 = (行数 - 1) × (列数 - 1)
 - 0からどれだけ離れた値を取るかを調べたいので、**棄却域を片側（右側）にとる**
- 「検定統計量 > 臨界値」なら帰無仮説を棄却する

χ^2 (カイ二乗) 分布 (chi-squared distribution)

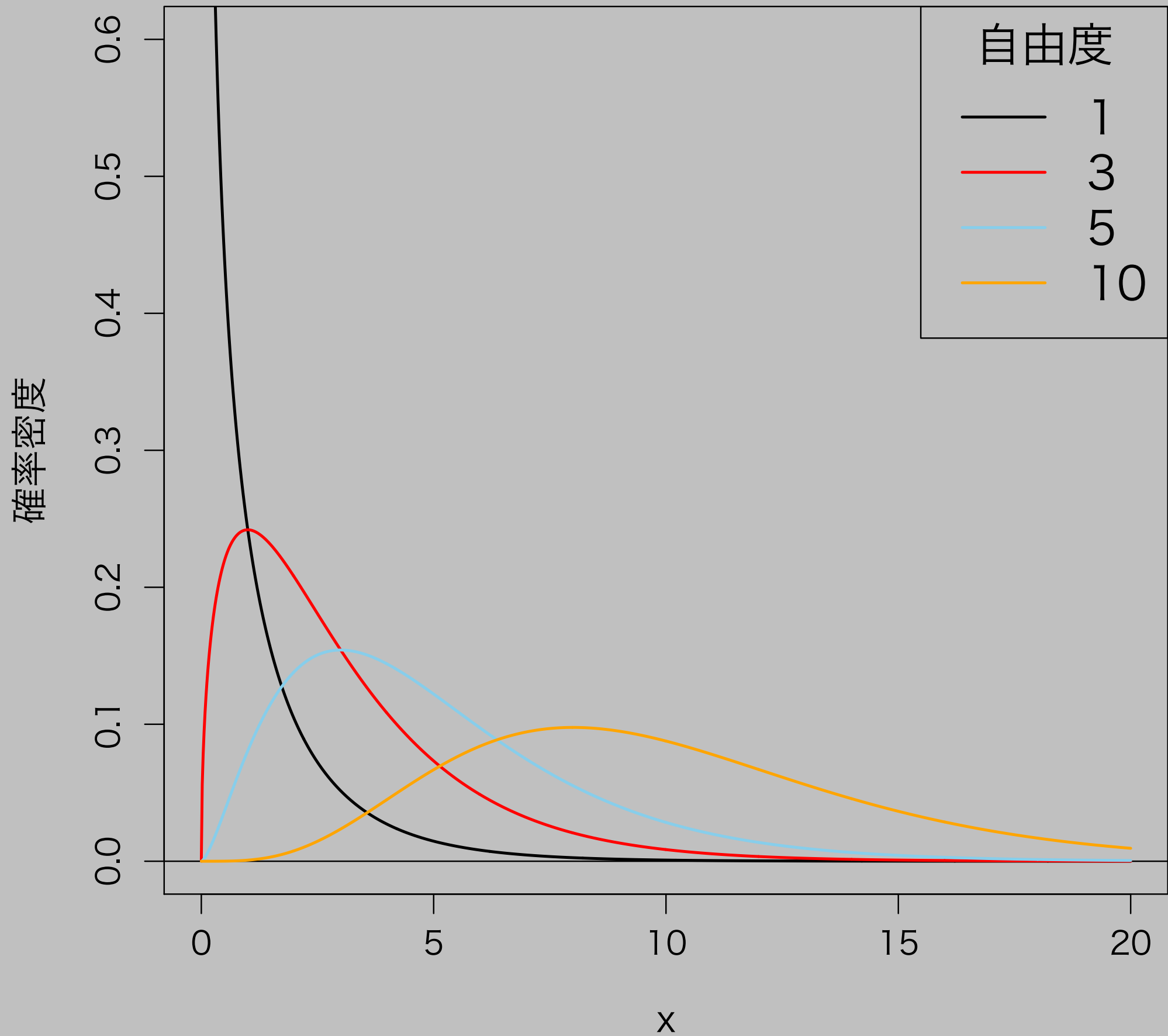
- 確率密度関数 $f(x)$ は、

$$f(x \mid k) = \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \quad (x > 0 \text{ のとき})$$

$$f(x \mid k) = 0 \quad (x \leq 0 \text{ のとき})$$

ただし、 k は x の自由度, $\Gamma(\cdot)$ はガンマ関数

χ^2 分布



自由度 (degree of freedom; df)

- 自由（独立）に動かせる値の数
- 各統計量に対して自由度が定められる
 - 例：サイズ N の標本の場合
 - ▶ 標本平均の自由度は N
 - ▶ 標本（不偏）分散の自由度は $N - 1$

例：有意水準5%で検定する

- 検定統計量：10.1
 - 2行2列の表 → 自由度 = $(2 - 1)(2 - 1) = 1$
- ➡ 有意水準5%の臨界値 = 3.84

```
qchisq(p = 0.05, df = 1, lower.tail = FALSE)
```

- ➡ 検定統計量 = $10.1 > 3.84$ = 臨界値
- ➡ 帰無仮説を棄却する
- ➡ 性別によって内閣支持率が異なる！

*フィッシャーの正確確率検定 (Fisher's exact test)

- 期待度数が5を下回るセルがあるとき
 - ➡ 検定統計量が大きめに出てしまうので、独立性の検定が使えない
 - ➡ フィッシャーの正確確率検定（直接確率法）を使う
(この授業では扱わない)

量的変数をクロス表にする

- ・情報が失われる

➡ 表にせずに関係を表す

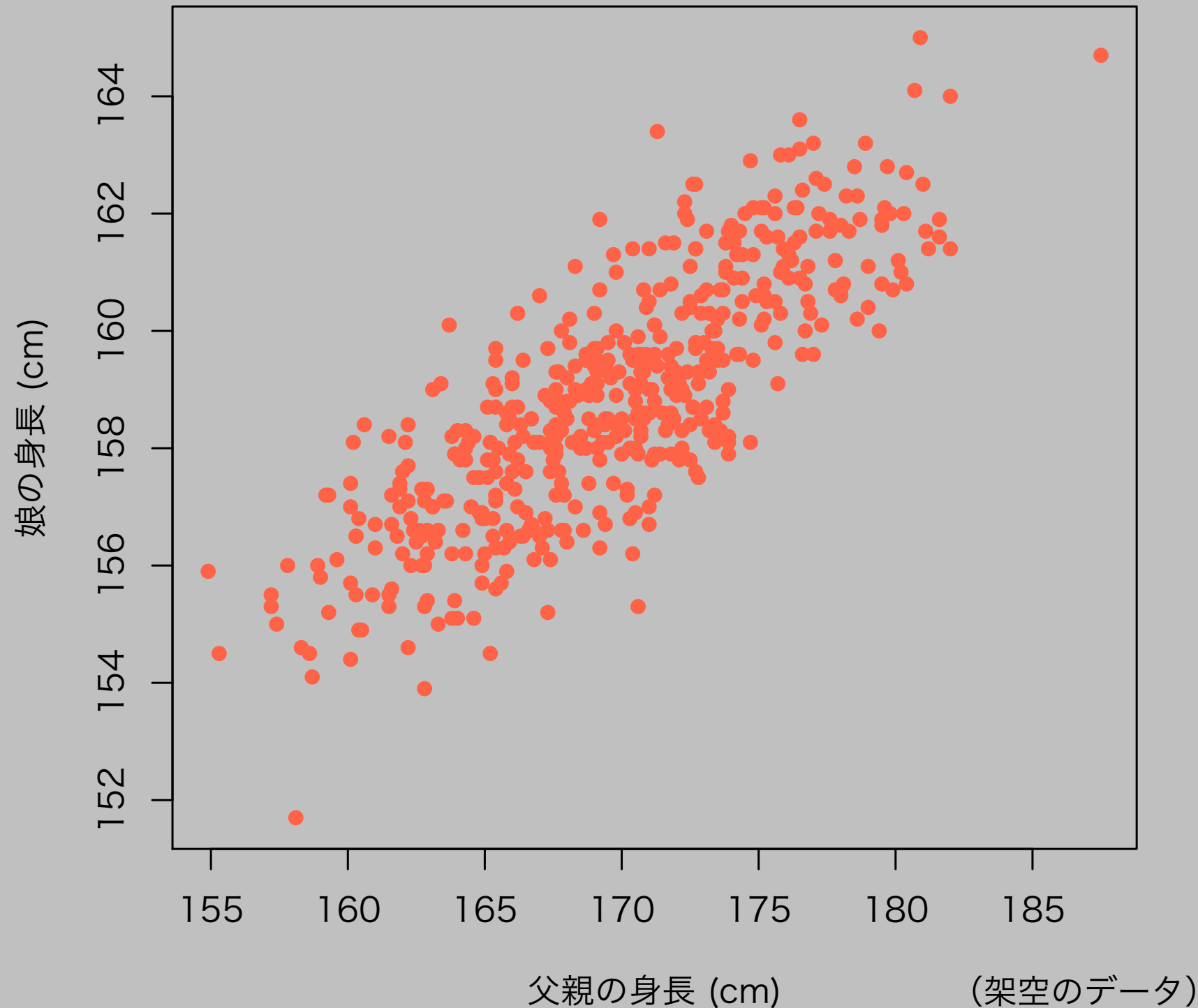
1.図示する：散布図

2.統計量を求める：相関係数

架空の例	年収		
	500未満	500～1000	1000以上
身長170cm未満	100	80	60
170cm以上	50	75	80

2変数の関係を図示する: 散布図 (scatter plot)

娘の身長と父親の身長の関係



相関関係

- 相関関係 (correlation) :
 - ▶ 2つの物事（変数）AとBの間の直線的な関係
 - ▶ Aの変化に合わせてBも変化する
 - ▶ 統計量：相関係数 r ($-1 \leq r \leq 1$)
 - ▶ Aが増える（減る）とき、Bも増える（減る）：正の相関 ($r > 0$)
 - ▶ Aが増える（減る）とき、Bが減る（増える）：負の相関 ($r < 0$)
 - ▶ $|r|$ が1に近いほど関係が強い

2変数の関係を表す統計量： 相関係数 (correlation coefficient)

- 変数 x と変数 y の相関係数 r

$$r = \frac{x \text{ と } y \text{ の共分散}}{\sqrt{x \text{ の不偏分散}} \sqrt{y \text{ の不偏分散}}}$$

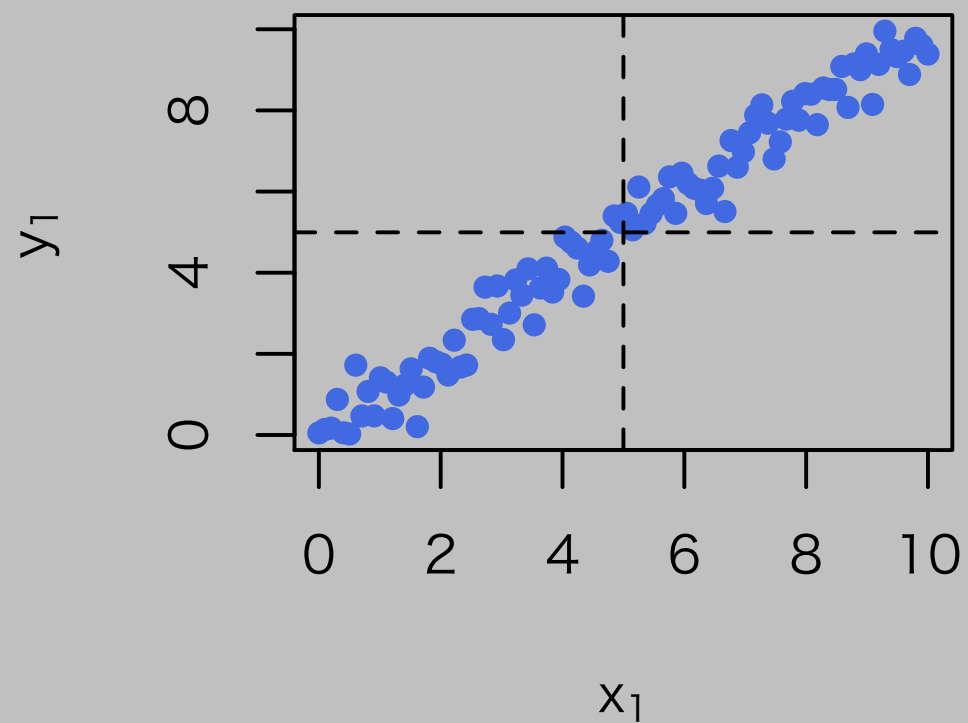
$$= \frac{\frac{\sum_{i=1}^N [(x_i - \bar{x})(y_i - \bar{y})]}{N-1}}{\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N-1}}}$$

$$= \frac{\sum_{i=1}^N [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

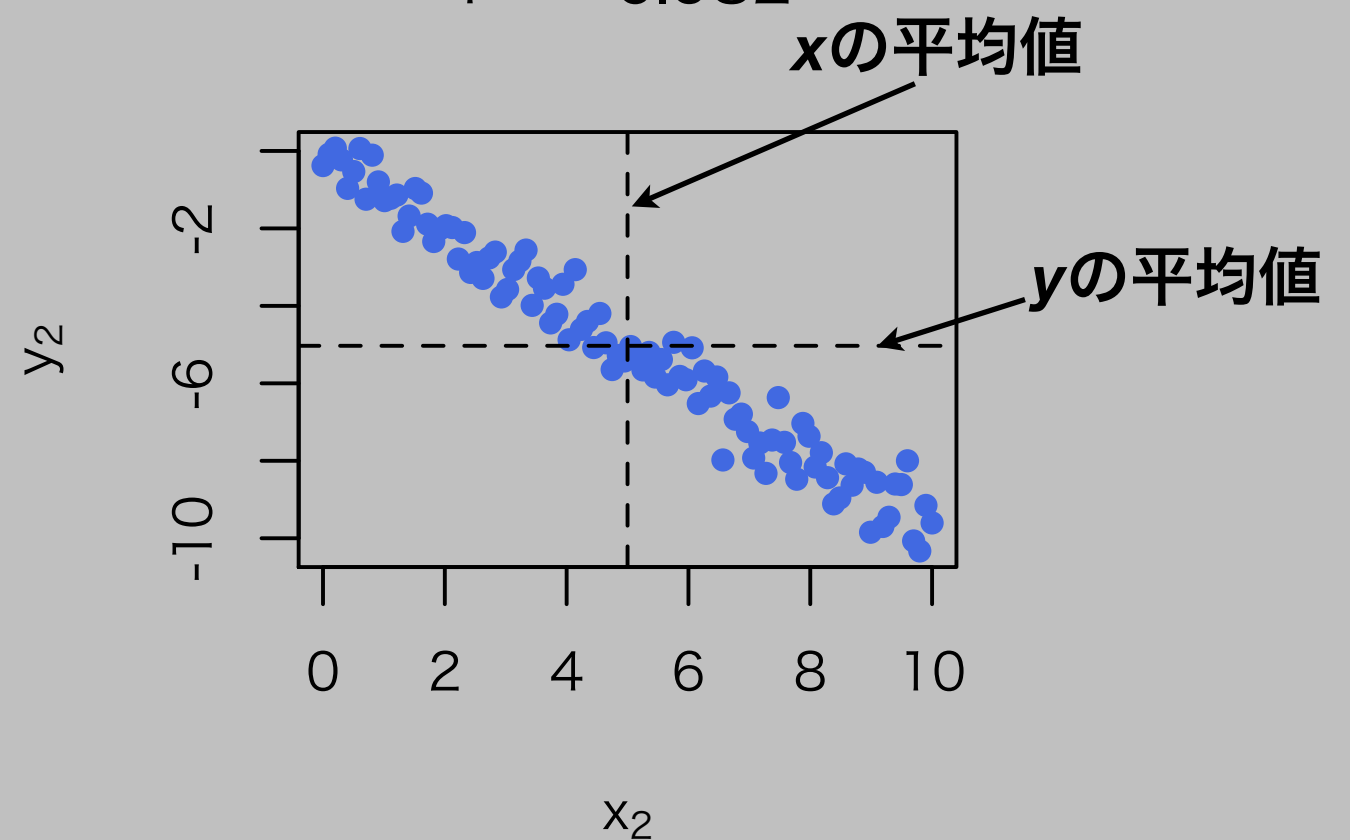
相関係数の特徴

- 2変数の**直線的な**関係の強さを表す
- 取り得る値の範囲は $[-1, 1]$
 - 1 : 正の直線的関係（一方が大きくなるとき、他方も大きくなる）が最も強い
 - -1 : 負の直線的関係（一方が大きくなるとき、他方が小さくなる）が最も強い
 - 0 : 直線的関係がない（曲線的関係は強いかもしれないことに注意）
- **因果関係はわからない**
- **因果関係を仮定するとして、原因が結果にどれだけ影響を与えるかはわからない**

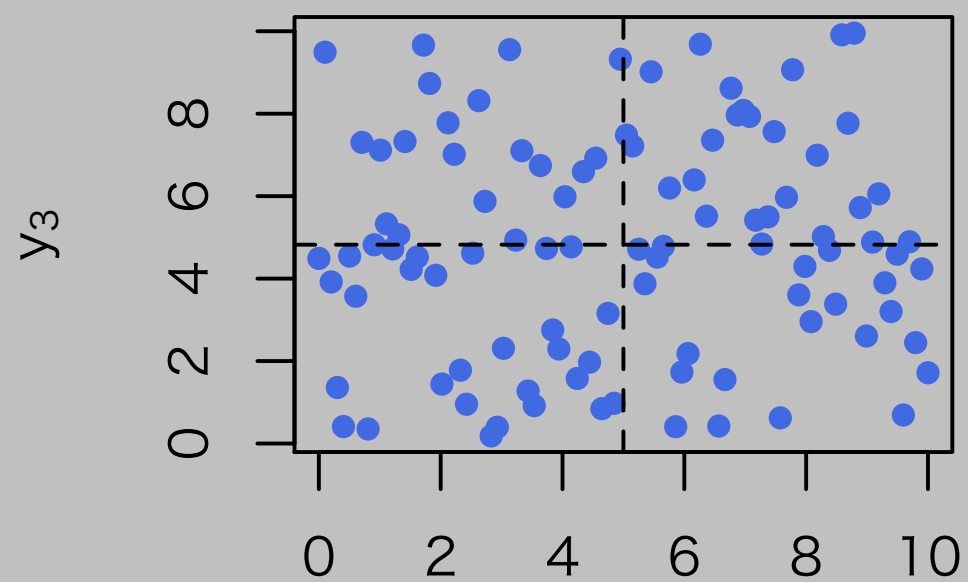
$$r = 0.986$$



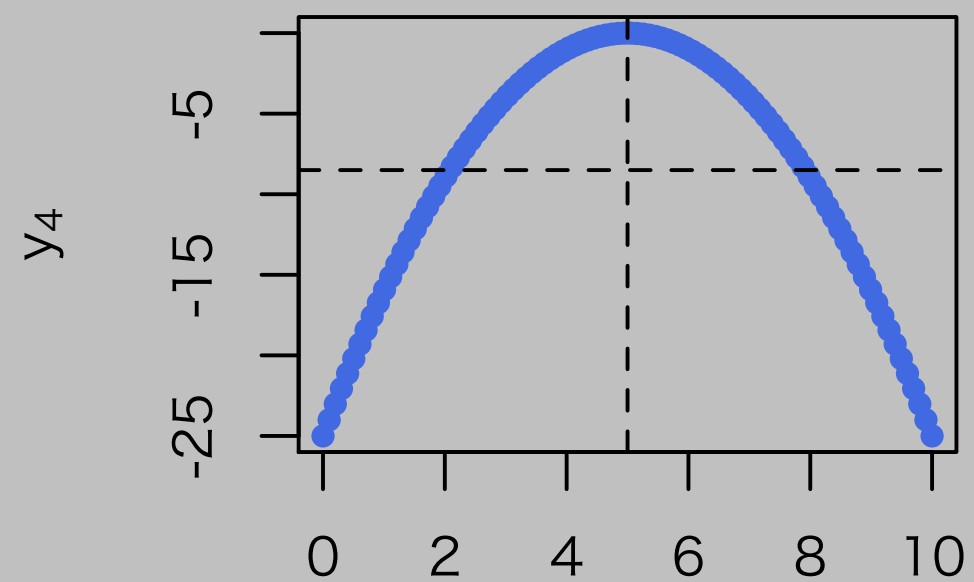
$$r = -0.982$$



$$r = 0.044$$



$$r = 0$$



今日のまとめ

- 2つの変数のまとめ方：変数の種類によって異なる
 - 質的変数：クロス集計表、独立性の検定（カイ二乗検定）
 - 量的変数：散布図、相関係数

▶ 散布図と相関係数は必ずセットで使う

次回

回帰分析の基礎