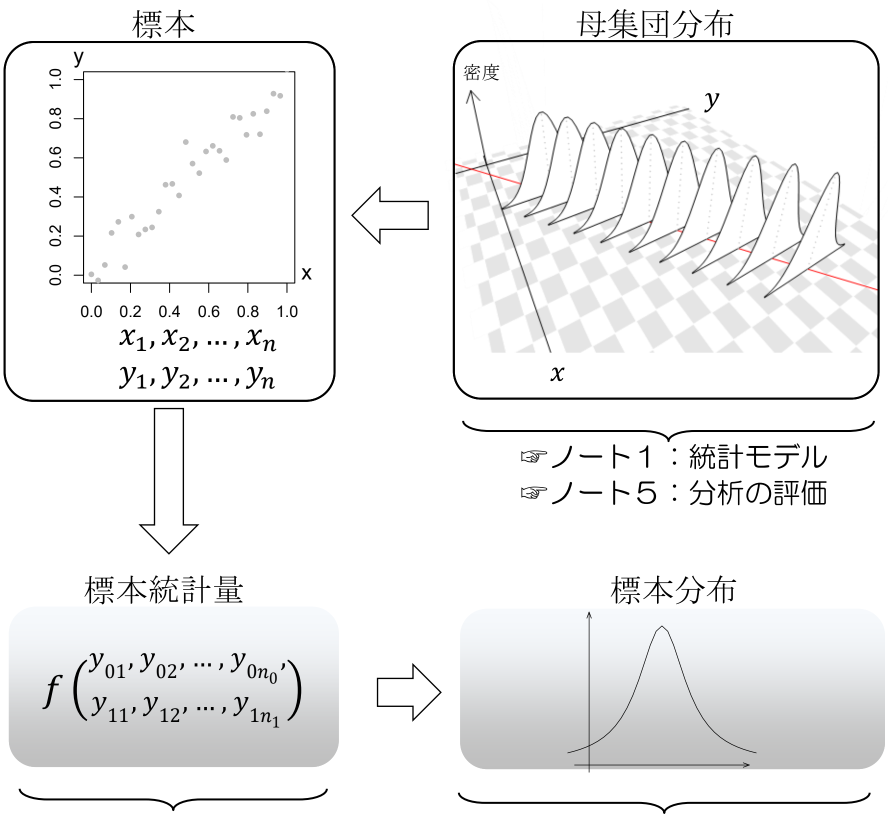
第4講　単回帰分析

学習の目標

- ☐ 研究者が母集団に対して仮定するモデルを**統計モデル**ということが分かる。
- ☐ 母集団において二つの変数 x、y の間に直線（一次関数）の関係があるという統計モデルを仮定し、これに関する統計的推測を行うことを**単回帰分析**と呼ぶことが分かる。
- ☐ 回帰分析の枠組みでは大きく二つのリサーチクエスチョンがあることが分かる。一つ目は、直線の切片と係数（回帰係数）の値が**どのくらいか**（how much）を求めることである。二つ目は、特に傾きが **0 か否か**（whether）である。
- ☐ 名義尺度変数を扱うために分類を 0, 1, …のように数値化したものを**ダミー変数**ということが分かる。
- ☐ 二群の差の検定で想定されている統計モデルを、ダミー変数を用いて数式で表すことができる。
- ☐ 単回帰分析で想定されている統計モデルを数式を用いて表すことができる。
- ☐ 回帰係数を求める方法として最小二乗法と呼ばれる点推定法が採用されることが分かり、なぜこの方法がよいのかが理解できる。
- ☐ 標本のデータをもとに構築した回帰直線と各データとの差を二乗基準で測ったものを残差と呼ぶことが分かる。
- ☐ 回帰係数と相関係数の関係が分かる。
- ☐ 不偏性という推定量（統計量）の望ましい性質が分かり、なぜ分散に、（標本）分散と不偏分散という二つの異なる指標が提案されているのかが理解できる。
- ☐ 作り上げた回帰係数を評価する軸として大きく三つの系統のものがあることが分かる。一つ目は、回帰係数の推定値の正確さの評価で、回帰係数の信頼区間の構成、予測値の信頼区間、データの予測区間を計算することができる。
- ☐ 二つ目は、作ったモデルがデータのばらつきのどのくらいを説明しているかという視点で、これが決定係数という指標で議論できることが分かる。
- ☐ 三つ目は、想定した統計モデルの適切さの評価で、これを吟味するために、残差プロット、Q-Q プロット、てこ値、クックの距離などが提案されていることが分かる。

見取り図



☞ノート１：統計モデル
☞ノート５：分析の評価

☞ノート２：統計量の基準
☞ノート３：回帰係数の点推定
☞ノート４：回帰係数の検定・区間推定

☞ノート４：t 分布

データの形式

ID	予測変数	応答変数
1	0.3	2.1
2	0.1	3.2
⋮	⋮	⋮
n	1.2	1.5

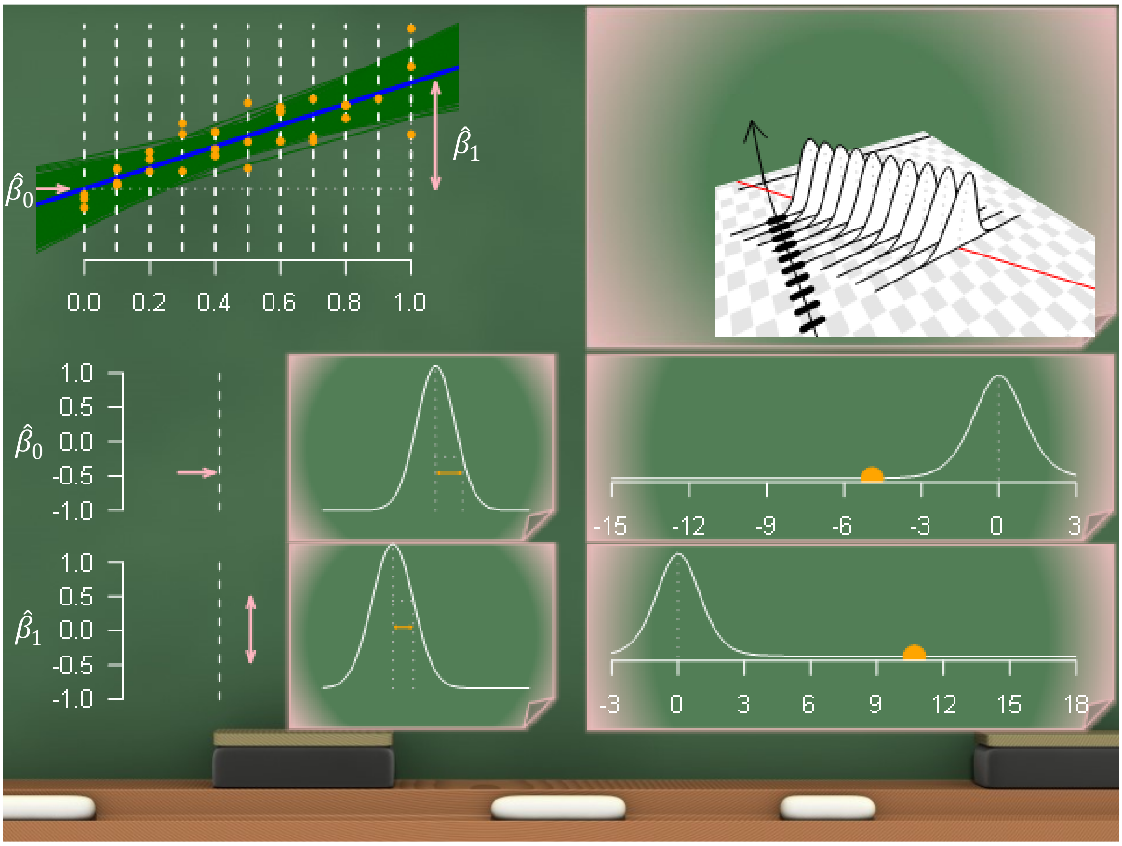
（１）目的（リサーチクエスチョン）

変数 x と変数 y が一次曲線の関係にあると仮定し、次の問題を考える。

（R1）How much 型：切片と傾きはどのくらいの大きさなのだろうか。

（R2）Whether 型：（切片と）傾きは 0 と見なしてよいか否か。

（２）考え方



母集団において二つの変数が直線で表されるような一次関数の関係にあると仮定してみる。そのときに、切片と傾きがどのようになり（How much；母集団効果量の推定）、とりわけ傾きが平らなのか（=0 なのか；回帰係数の t 検定）否か（whether）を考えるのがこの単回帰分析。

前講で扱った対応のない t 検定が扱った文脈は、x 軸の値が 0 と 1 しかとらなかったのに対して、この制約を取り払い、比率尺度データについても扱えるようになる（あるいは、t 検定が単回帰分析の特殊な場合だと見なすことができる）。

📖 ノート１　母集団に対する仮定：統計モデル

（１）統計モデル

これは、研究者が母集団に対して想定するモデル。

例１：t 検定

① ダミー変数

分類を表す質的変数（名義尺度変数）に便宜的に 0、1, …のように数値を与えコード化した変数。

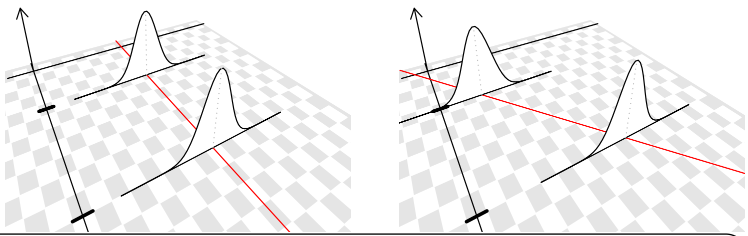
$$x_i = \begin{cases} 1 & \text{if } i\text{th person is \_\_\_\_} \\ 0 & \text{if } i\text{th person is not \_\_\_\_} \end{cases}$$

※1 か 0 のものを指示変数 Indicator variable と呼ぶ。

② 線形モデル

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i \\ \beta_0 \qquad\quad + \varepsilon_i \end{cases}$$



質問　ダミー変数はいつも０と１なんですか？

この他の組み合わせも可能です。ただ、係数の解釈が変わります。例えば、次の例ではβ_0は全体の平均、β_1は平均から上下に男女がどのくらい離れるかを表します。

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is \_\_\_\_} \\ -1 & \text{if } i\text{th person is not \_\_\_\_} \end{cases}$$

$$y_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i \\ \beta_0 - \beta_1 + \varepsilon_i \end{cases}$$

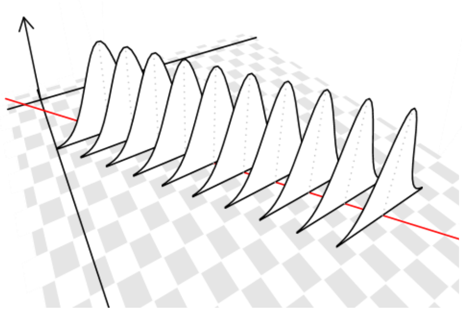
例２：単回帰分析

① 概要

これは「予測変数（x）と応答変数（y）が一次直線と基本とする関係にある」と想定する統計モデル。

② 線形モデル

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



（２）仮定

【標本の抽出の仕方に関する仮定】

① **独立性の仮定**：　各要素は互いに独立

② **同一分布の仮定**：　標本は同じ分布から抽出

③ **無作為性の仮定**：　標本はランダムに抽出

$$y_i \overset{i.i.d}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

【母集団の姿に関する仮定】

① 正規性の仮定：　母集団の分布は正規分布

② 分散の等質性の仮定：　二つの母集団の分散は同じ

📖 ノート２ 統計量：回帰係数の点推定（最小二乗法）

（１）最小二乗法 Least Squares Method

これは、 二乗基準 で測ったデータと直線の適合の悪さ（残差）を最小化することで点推定値を出す方法。

① 残差 Residuals

これは、予測の誤差のこと。

$$
\begin{aligned}
e_i &= y_i - \hat{y}_i \\
&= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)
\end{aligned}
$$

② 損失関数 Loss function

これは、適合の悪さを測る関数。一般的な単回帰では残差平方和Residual Sum of Squaresを損失関数とする。

$$
\begin{aligned}
L(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2
\end{aligned}
$$

（２）最小二乗推定量

これは、最小二乗法で求めた点推定量のこと。単回帰モデルの場合は、次のような結果になる。

① 傾き

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\
&= r_{xy} \frac{s_y}{s_x}
\end{aligned}
$$

② 切片

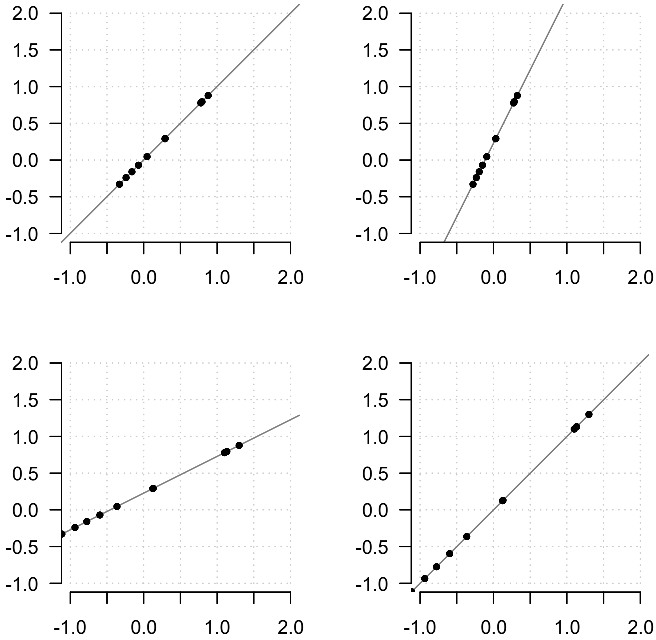
$$
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
$$

③ 誤差

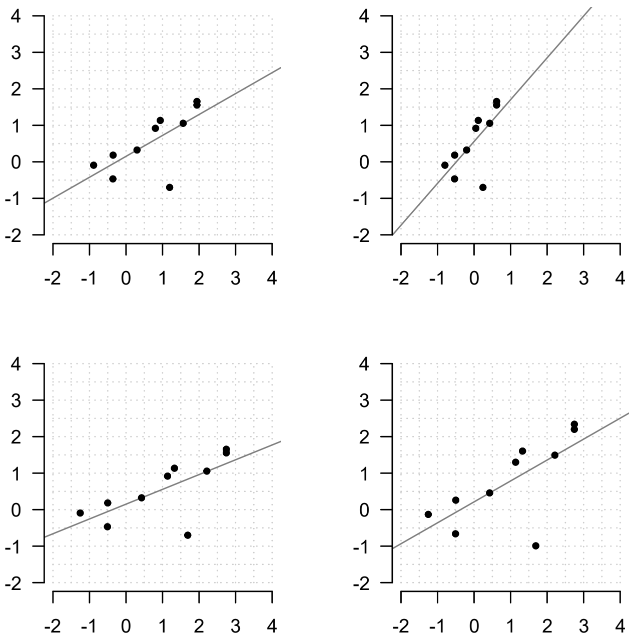
$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2} \times \sum_{i=1}^{n} e_i^2 \\
&= \frac{1}{n-2} \times \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2
\end{aligned}
$$

（３）回帰係数と相関係数の関係

① 相関係数 $r_{xy} = 1$



② 相関係数 $r_{xy} = .7$



📖 ノート３ 統計量：点推定と推定量の性質

（１）点推定

① 点推定

母集団に想定した統計モデルのパラメータの値がどのような値なのか点（統計量一つ）で推定する統計的推測。

② 推定量 Estimator

これは、点推定に採用される統計量のこと。

③ 推定量の作り方

これには、様々な方法が考案されている。

（例１）モーメント法

（例２）最小二乗法

（例３）最尤推定法

（２）推定量の選択基準Ⅰ：小標本特性

これは、標本を何度も何度も繰り返しとった場合に生じるメリットのこと（サンプルサイズは小さくてもよい）。

① 不偏性 unbiasedness

これは、その推定量の期待値が母パラメータに一致すること。不偏性を持つ推定量を不偏推定量という。

② 有効性 efficient estimator

これは、不偏性を持つ推定量の中で分散が最小のもののこと。有効性を持つ推定量を有効推定量という。

（３）推定量の選択基準Ⅱ：大標本特性

これは、サンプルサイズ n の大きい標本を採ったときに生まれるメリットのこと。

① 漸近的不偏性 asymptotic unbiasedness

これは、サンプルサイズが大きいとき、その推定量の期待値が母パラメータに一致すること。

② 一致性 consistency

これは、サンプルサイズが大きいとき、推定量と母パラメータに差が生じる確率は極めて小さくなること。

③ 漸近的有効性 asymptotic efficiency

これは、サンプルサイズが大きいときに、不偏推定量の中で最小分散となること。

（４）分散と不偏分散

① 分散

これは、漸近的不偏性を持つが不偏性は持たない。

② 不偏分散

これは、不偏性を持ち、当然漸近的不偏性も持つ。