

# 統計解析応用研究 記述統計

丸山 祐造 Yuzo Maruyama

神戸大学 大学院経営学研究科

# データ I

- ▶ 統計学で扱うデータ ⇒ 主に観測や実験で得られた数値データ
- ▶ 典型的には表形式に整理される
- ▶ 例：家計調査のある月の全調査世帯のデータ

世帯 \ 項目	食費	教育費	交際費	世帯人員	世帯主・性
1	75000	40000	66000	4	男
2	70000	80000	91000	1	女
⋮	⋮	⋮	⋮	⋮	⋮
n	30000	51000	65000	3	男

# データ II

## 用語

- ▶ 個体：各世帯．観測を行う個々の対象
- ▶ 変数 or 変量：調査項目
  - ↑ 典型的な統計データは「個体 × 変数」の形の表形式のデータ
- ▶ サンプルサイズ：個体の総数  
 $n$  が使われることが多い
- ▶ データの次元：変数の数  
 $p$  が使われることが多い

## データ III

- ▶ 一般に個々の観測値を  $x_{ij}$  とすれば，表形式のデータは  $n \times p$  の行列と見做せる

↑ ただし行列や線形代数の知識は不要

個体 \ 変数	1	2	...	$p$
1	$x_{11}$	$x_{12}$	...	$x_{1p}$
2	$x_{21}$	$x_{22}$	...	$x_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$n$	$x_{n1}$	$x_{n2}$	...	$x_{np}$

- ▶ データセット：表形式のデータで特にコンピュータのファイルとして記録されたもの

## データ IV

データセットの例：1990年代のT大学K学部の統計学の受講者のデータ

1. あなたの身長は何センチですか？
2. あなたの体重は何キロですか？
3. あなたの父親の身長は何センチですか？
4. あなたの母親の身長は何センチですか？
5. 通学時間は片道何分ですか？
6. アルバイトは週平均何時間ぐらいしていますか？
7. テレビを一日平均何分ぐらい見ますか？
8. 大学の授業は一般的に言って面白いですか？ 1～5のスケールで評価して下さい（1:大変面白い～5:大変つまらない）

## データ V

9. あなたの血液型は？ A,B,O,AB のいずれかで教えてください。
10. 好きな野球チームは？ 以下の数字で教えてください。(1:キョジン 2:ヨコハマ 3:チュウニチ 4:ハンシン 5:ヒロシマ 6:ヤクルト 7:キンテツ 8:セイブ 9:ダイエー 10:ニッポンハム 11:オリックス 12:ロッテ 13:特に無し)

以下1か0で教えてください。

11. 姉妹はいますか？ 1:はい, 0:いいえ
12. 煙草をすいますか？ 1:はい, 0:いいえ
13. 自宅ですか, 下宿ですか？ 1:自宅, 0:下宿
14. 自動車の免許を持っていますか？ 1:はい, 0:いいえ
15. 普段運転できる自動車が手近にありますか？ 1:はい, 0:いいえ

## データ VI

16. 恋人はいますか（両思いに限る）？ 1:はい, 0:いいえ

17. あなたの性別は？ 1:男性, 0:女性

データセットの最初の10行

---

172	70	165	163	30	0	60	3	a	1	0	0	0	0	0	0	1
176	69	150	155	50	4	60	3	a	2	0	0	0	0	0	0	1
170	70	170	158	60	30	60	5	a	10	1	1	0	1	0	1	1
174	70	165	154	50	2	120	3	a	1	0	1	1	1	1	1	1
170	62	163	158	40	15	60	4	o	1	1	1	0	1	0	1	1
167	50	165	158	45	0	120	4	ab	1	0	0	0	1	0	0	1
175	75	171	158	100	30	2	3	a	1	0	1	1	0	0	0	1
179	80	156	150	55	6	90	3	ab	3	0	0	0	1	0	0	1
162	60	160	160	60	2	120	5	b	13	1	0	0	0	0	0	1
169	80	165	162	45	0	60	4	o	6	1	0	0	1	1	0	1

---

# データ VII

## 量的変数と質的変数 より詳しい説明

- ▶ 量的変数：観測値自体が量として意味があるもの
  - ▶ 比例尺度：値の大小関係と値の差の大きさ・比に意味があり，値0が絶対的な意味をもつ．身長や体重，降水量など棒グラフは比例尺度の量を表すのに用い，棒の長さ（面積）が量に比例するように描くのが基本
  - ▶ 間隔尺度：値の大小関係と値の差の大きさに意味があり，値0は相対的な意味しかもたない．気温など折れ線グラフは，比例尺度 or 間隔尺度．0点から始める必要はない．特に時系列データについてよく使われる



## データ VIII

- ▶ 質的変数：性別，職業のように変数が分類を表すもの
  - ▶ 名義尺度：同じ値であるか否か以外に意味を持たない尺度。  
血液型「4 値」，性別「2 値」
  - ▶ 順序尺度：同じであるか否かに加えて，大小関係を持つ．満  
足度など．データ解析においては，取扱が難しい
- ▶ ダミー変数：2 値の名義尺度の質的変数に対して 0  
と 1 でコード化された変数  
↑特に回帰分析の文脈でこの用語が用いられる

## データ IX

- ▶ ダミー変数では，量的変数と同様の計算が可能
  - ▶ 自動車の免許を持っていますか？ 1：はい，0：いいえ
  - ↑ 算術平均は免許保有率である
- ▶ 満足度のような順序尺度で平均値に意味があるか？

$$1 \times \frac{15}{100} + 2 \times \frac{25}{100} + 3 \times \frac{30}{100} + 4 \times \frac{20}{100} + 1 \times \frac{10}{100}$$

# 統計計算のためのソフトウェアと R 言語 I

## 統計学の目的とコンピュータの利用

- ▶ データセットに対して様々な処理を施し、データの特徴を明らかにしようとするのが統計学
- ▶ 統計的な処理
  - ▶ データ行列に対する行列演算
  - ▶ 数列の和  $\sum$  などの四則演算の組み合わせ
- ▶ ただし個体数  $n$ , 変数の数  $p$  ともに現代においては大きい
- ▶ 手計算や電卓は非現実的

例:  $n = 1000$  における平均値  $(x_1 + \cdots + x_n)/n$

## 統計計算のためのソフトウェアと R 言語 II

- ▶ 伝統的な統計学の教科書
  - ▶  $n = 10$  程度で様々な説明
  - ▶ 理論重視で数値例は補助
- ▶ データセットに様々な処理を施し、データの特徴を明らかにしようとする統計学の目的からは不十分
- ▶ 統計学の理論と同時にその意味、目的を理解し、面白さを感じるにはコンピュータを用いて大きめのデータを統計処理することが望ましい

# 統計計算のためのソフトウェアと R 言語 III

excel に代表される表計算ソフト

- ▶ 縦横に罫線で区切られた「集計用紙」のイメージをコンピュータ上に再現
- ▶ 集計用紙での計算作業をコンピュータ上で実行
- ▶ データ自体とその計算処理が「集計用紙」の中に混在し，計算処理部分がプログラムの形に分離されていない
- ▶ 計算処理部分のアルゴリズムの明快な記述も不可能

# 統計計算のためのソフトウェアと R 言語 IV

## R 言語

- ▶ 教育現場でのライセンス制限なし
- ▶ プログラミングの要素を持つ
- ▶ 行列計算の簡単な記述
- ▶ データのグラフ表示が可能
- ▶ R studio cloud

# 推測統計とコンピュータ I

## ▶ 統計的な手法の分類

「記述統計的な手法」と「推測統計的な手法」

↑ 現段階では何を言っているかわからなくて良い

## 記述統計的な手法

- ▶ データを所与の値とみなしてその特徴を記述するための手法
- ▶ サンプルサイズ  $n$  が大きい場合には数字の羅列を眺めているだけでは全体の特徴を把握不可能
- ▶ データセットの効率的な記述方法が不可欠
- ▶ 記述統計的な手法の多くは行列計算の応用. コンピュータに馴染みやすい

# 推測統計とコンピュータ II

## 推測統計的な手法

- ▶ データの背後に確率を考えて、データを確率変数の実現値と見る
- ▶ サイコロを何度もふったときの目の出方  $2, 5, 3, 3, 4, 1, 2, 3, \dots$ ,  
これらの数字を記録してデータセットを作るとする  
↑ 個々の数字には興味がわかず、「それぞれの目が同様に出やすいか」「続いて同じ目が出やすいか」など目の出方の確率的な構造に興味を湧くはず
- ▶ 推測統計的な手法：このようにデータの背後に確率的な構造を考えて、確率的な構造を分析する
- ▶ 統計的モデル：データの背後に想定される確率的な構造



## 推測統計とコンピュータ III

- ▶ 必要とされる知識：確率論や微積分学
- ▶ ただし，コンピュータを用いる（モンテカルロ法あるいはシミュレーション）ことにより理解が容易となる
- ▶ シミュレーション：コンピュータ上で乱数を発生させて確率変数の振る舞いを実験的に観察する手法
- ▶ 中心極限定理のように数学的には抽象的な証明しか与えられない場合でも，視覚的にその意味を理解できる

学ぶ順番

記述統計 → 確率 → 推測統計

# 1 変量データの分布とヒストグラム I

- ▶ ある変数に関するサンプルサイズ  $n$  の観測値

$$x_1, \dots, x_n$$

- ▶ 分布： $n$  個の観測値がどの値を中心にしてどのように散らばっているか，その様子
- ▶ 度数分布，ヒストグラム：1 変量の観測値の分布を見る最も基本的な手法

# 1 変量データの分布とヒストグラム II

**度数分布**：実数軸を適当な区間に区切り，各区間に入る観測値の個数（度数）を数えたもの

- ▶ 区間たち  $I_0 = (-\infty, c_1)$ ,  $I_1 = [c_1, c_2)$ ,  $\dots$ ,  $I_k = [c_k, c_{k+1})$ ,  $I_{k+1} = [c_{k+1}, \infty)$
- ▶ 度数  $f_i$  :  $I_i = [c_i, c_{i+1})$  に落ちた観測値の数
- ▶ 多くの場合
  - ▶  $c_1, \dots, c_k$  は等間隔
  - ▶  $c_1 < \min x_i$ ,  $c_{k+1} > \max x_i$  として，区間  $[c_1, c_{k+1})$  に全ての観測値が入る（非有界な  $I_0, I_{k+1}$  は考えない）

# 1 変量データの分布とヒストグラム III

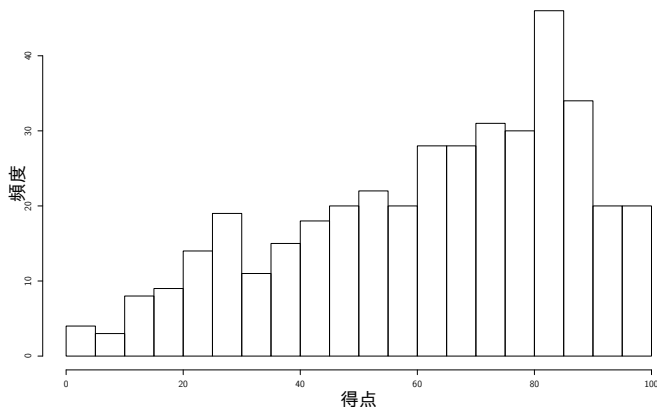
- ▶ 等間隔としてもその幅をどう決めるか？
- ▶ 理論的に正解はなく，その幅が変わればデータの特徴について異なった印象を与える

## ヒストグラム

- ▶ 各区間上に度数に比例する高さの棒グラフ
- ▶ 度数の区間幅の問題はヒストグラムによって可視化される

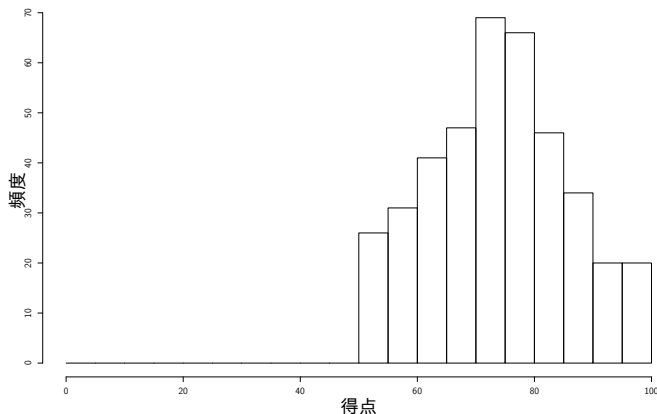
# 1 変量データの分布とヒストグラム IV

## 厳しい教員のテストのヒストグラム



# 1 変量データの分布とヒストグラム V

## 優しい教員のテストのヒストグラム



## 順序統計量，累積分布関数，分位点 I

- ▶ 度数分布において区間を非常に小さく取る
  - ▶ （同順がないとすると）全ての区間で度数 1
  - ▶ 観測値を小さい順に並び替えることと同じ
- ▶ 順序統計量：ソート（並び替え）して得られた値

最小値  $\Rightarrow x_{(1)} \leq x_{(2)} \cdots \leq x_{(n-1)} \leq x_{(n)} \Leftarrow$  最大値

- ▶ **中央値**：順序統計量の中で真ん中にある値

$$\text{med}(x) = \begin{cases} x_{(\{n+1\}/2)} & n \text{ が奇数} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & n \text{ が偶数} \end{cases}$$

## 順序統計量，累積分布関数，分位点 II

- ▶ これから四分位点・箱ひげ図及びその一般化を学ぶ  
中学生の学修内容！ ↑
- ▶ 累積分布関数（または経験分布関数）：ヒストグラムとともに分布の様子 of 把握に有用  
↑「特定の値」以下に落ちる観測値の割合
- ▶ 確率変数の分布関数も「累積分布関数」と呼ばれる。  
区別のための用語「経験分布関数」。経験 (empirical)  
はデータに基づいたという意味



# 順序統計量，累積分布関数，分位点 III

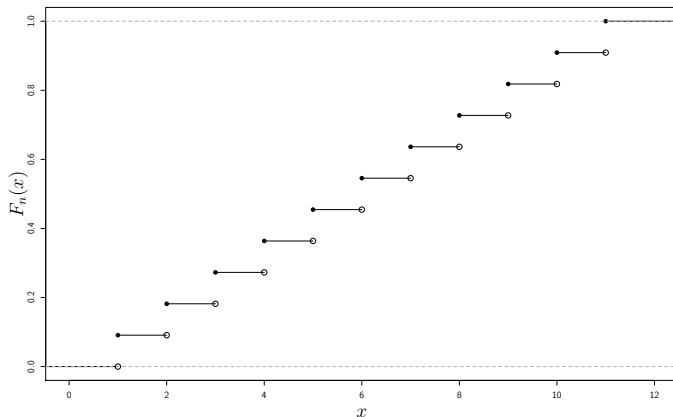
## ▶ データ $x_1, \dots, x_n$ に基づく累積分布関数

$$F_n(t) = \frac{1}{n}(\textcolor{red}{t}\text{以下の観測値の個数})$$
$$= \begin{cases} 0 & t < x_{(1)} \\ \frac{k}{n} & x_{(k)} \leq t < x_{(k+1)} \quad k = 1, \dots, n-1 \\ 1 & t \geq x_{(n)} \end{cases}$$

- ▶ 各観測値で  $1/n$  ジャンプする階段状の非減少関数
- ▶ グラフを描く場合，原点は  $(0, 0)$  ではなく， $y$  軸の  $x$  座標を  $x_{(1)}$  より小さく（例えば  $x_{(1)} - 1$ ）とる

# 順序統計量，累積分布関数，分位点 IV

人工データ 1, 2, 3..., 9, 10, 11 に対する経験分布関数



# 順序統計量，累積分布関数，分位点 $V$

中央値，分位点，経験分布関数

- ▶ 分位点 ( $100u\%$ 点)  $x_u$  :  $x$  以下に  $100u\%$  ( $0 < u < 1$ ) の観測値が落ちるような  $x$

下付き  $_u$  は小数なので混乱は生じないはず

- ▶ 経験分布関数  $F$  の逆関数として，分位点は定義できるように思われる

$$F(t) = \frac{1}{n}(t \text{ 以下の観測値の個数})$$

$\Rightarrow u$  が与えられたときに  $u = F_n(t)$  を満たす  $t$

# 順序統計量，累積分布関数，分位点 VI

- ▶ ただし  $F$  は階段関数
- ▶  $G: F$  を縦方向にもつないだ関数
- ▶  $x_u$  の決め方
  - ▶  $y$  軸上の座標  $(x_{(1)} - 1, u)$  から右に進み， $G$  とぶつかったら，その  $x$  座標を  $x_u$  とする
  - ▶ ただし，ぶつかった  $G$  が水平な部分 ( $x$  軸と平行) であった場合，その水平な部分の左端と右端の  $x$  座標の中点を  $x$  座標を  $x_u$  とする

# 順序統計量，累積分布関数，分位点 VII

## ▶ 順序統計量を用いた分位点の定義

$$x_u = \begin{cases} \frac{x_{(nu+1)} + x_{(nu)}}{2} & \text{if } u = \frac{1}{n}, \dots, \frac{n-1}{n} \\ x_{([nu]+1)} & \text{otherwise} \end{cases}$$

[ ] **ガウス記号** [ ] 内の値を越えない最大の整数値

- ▶ 中央値：中央値以下に50%の観測値が落ちるような点 50%点

# 順序統計量，累積分布関数，分位点 VIII

## 逆関数

- ▶ ある関数に対して「もとにもどす」関数

例： $y = 3x$  という関数は，1 を 3 に，2 を 6 にする  
ような関数．もとにもどす関数は，6 を 2 に，3 を 1  
にするような関数

グラフで表現すると， $y = 6$  に対応する  $x$  の値を探  
すために  $y$  軸上の点  $(0, 6)$  から左右どちらかに水平  
( $x$  軸と平行) に進み， $y = f(x)$  (今の場合  $y = 3x$ )  
にぶつかった点の  $x$  座標

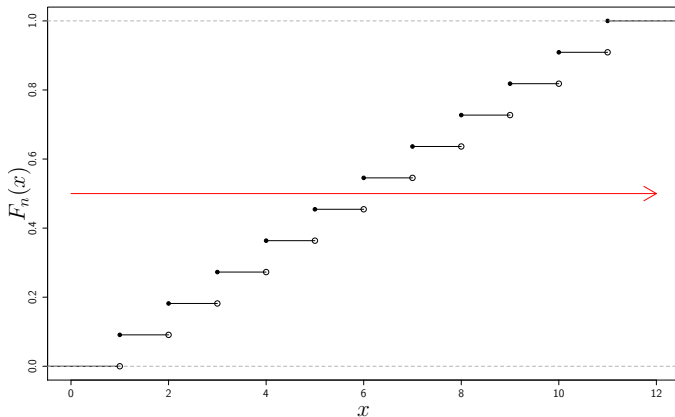
# 順序統計量，累積分布関数，分位点 IX

- ▶ 逆関数の計算： $y = f(x)$  を  $x$  について解き  $x = g(y)$  となったときの  $g$  が逆関数

例： $y = 3x$  の逆関数

- ▶  $y = 3x$  を  $x$  について解くと  $x = y/3$  なので，逆関数は  $y = x/3$
- ▶ 実際に 6 を 2 に，3 を 1 にするような関数になっている

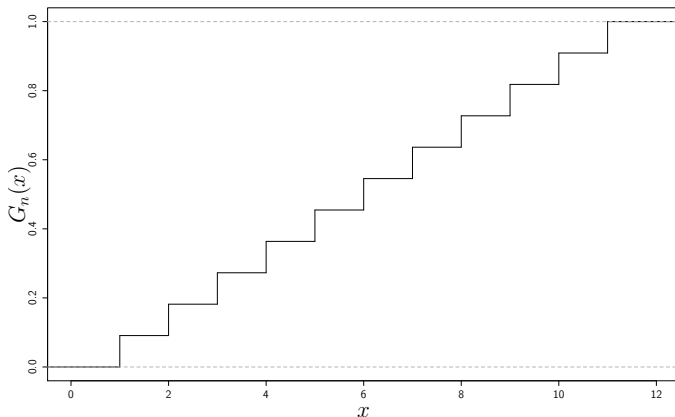
順序統計量，累積分布関数，分位点  $X$   
 $0.5 = F_n(x_u)$  を満たす  $x_u$  は？





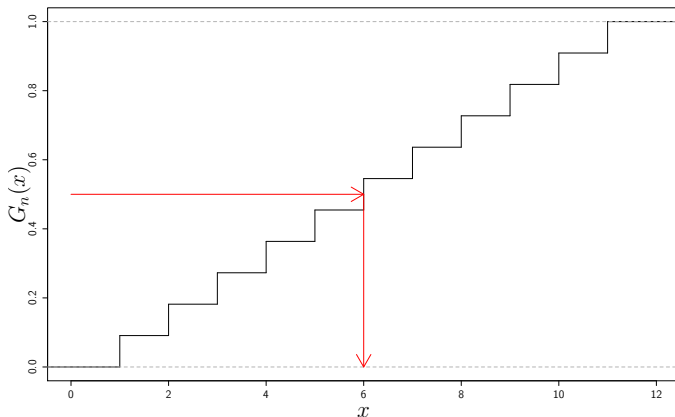
# 順序統計量，累積分布関数，分位点 XI

$G_n$  :  $F_n$  を縦方向にもつないだ関数



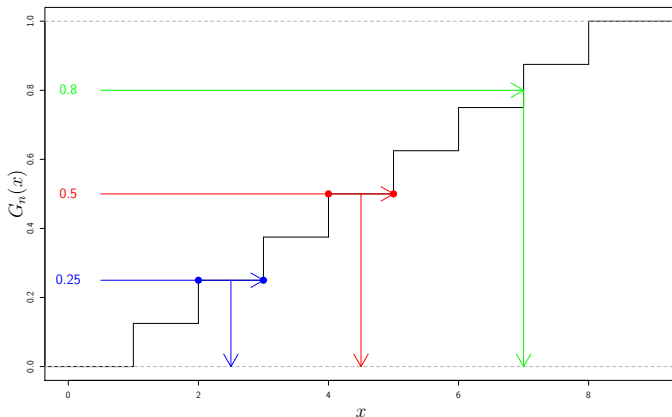
# 順序統計量，累積分布関数，分位点 XII

与えられた  $u$  に対して，対応する  $x_u$  を求める



# 順序統計量，累積分布関数，分位点 XIII

人工データ 1, 2, ..., 8 の場合



## 順序統計量, 累積分布関数, 分位点 XIV

- ▶  $x_u$  : 下側  $100u\%$  点,  $x_{1-u}$  : 上側  $100u\%$  点
- ▶ 下側四分位点 (下側 25%点) :  $x_{0.25}$
- ▶ 上側四分位点 (上側 25%点) :  $x_{0.75}$
- ▶ 最小値, 下側四分位点, 中央値, 上側四分位点, 最大値

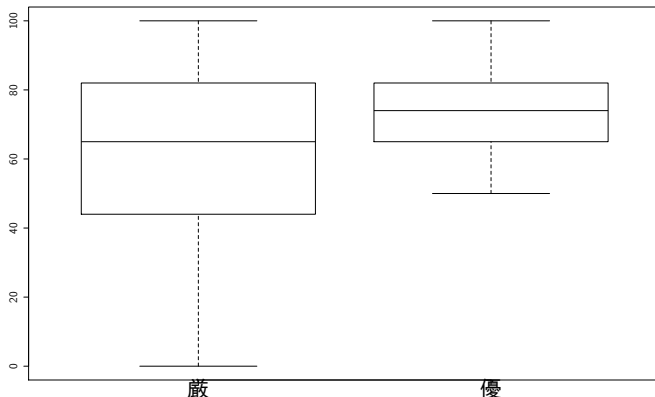
$$(x_{(1)}, x_{0.25}, x_{0.5}, x_{0.75}, x_{(n)})$$

は観測値を  $1/4$  ずつの割合に分ける点

- ▶ これらの値から分布の特徴をある程度掴める  
↑ 箱ひげ図 (ボックスプロット)

# 順序統計量，累積分布関数，分位点 XV

## 「厳しい」と「優しい」の箱ひげ図



# 平均と分散 I

- ▶ ヒストグラム, 分布関数, 箱ひげ図は分布を視覚的に把握
- ▶ 平均や分散は, より定量的に分布を把握
- ▶ 統計量: 分布の特徴を表すために計算される観測値  $x_1, \dots, x_n$  の関数

$$t(x_1, \dots, x_n)$$

# 平均と分散 II

## 平均

- ▶ 統計学の最も基本的な統計量
- ▶  $x_1, \dots, x_n$  に対する平均 math.pdf 1.1節

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- ▶ 平均は文字通り平均的な値,  $n$  個の値がどの辺りを中心として分布しているかを表す
- ▶ もちろん中央値も分布の位置を表す統計量

# 平均と分散 III

## 分散

- ▶ 観測値が分布の中心 (平均) からどのくらい離れる傾向にあるか, という分布のばらつきを表す代表的な統計量
- ▶  $x_1, \dots, x_n$  に対する分散  $s^2$

$$s^2 = s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- ▶ 分散:  $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2$  の平均
- ▶  $n - 1$  で割る流儀もあるが, ここでは説明省略



# 平均と分散 IV

## ▶ 知られた別表現 math.pdf 1.5節

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

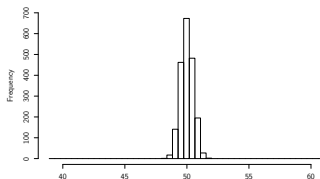
## ▶ 標準偏差 $s$ : 分散の平方根

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

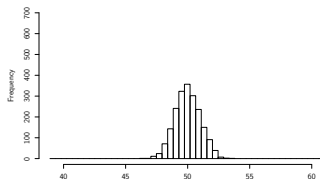
# 平均と分散 V

平均は共通，分散が違うデータのヒストグラム

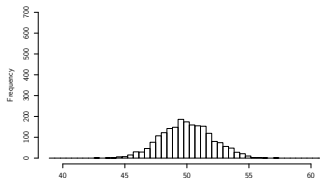
分散  $0.5^2$



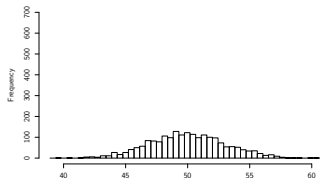
分散  $1^2$



分散  $2^2$



分散  $3^2$



# 線形変換，標準化，偏差値 I

## 線形変換

- ▶ 全ての観測値  $x \rightarrow a + bx$

$$x_1, x_2, \dots, x_n \rightarrow a + bx_1, a + bx_2, \dots, a + bx_n$$

- ▶ 位置の変換： $a$ を加えること
- ▶ 尺度の変換： $b$ をかけること

変換後の平均  $a + b\bar{x}$ ，分散  $b^2 s^2$

math.pdf 1.2節

## 線形変換，標準化，偏差値 II

▶ 標準化： $a = -\frac{\bar{x}}{s}$ ,  $b = \frac{1}{s}$  math.pdf 1.3節

$$z_i = a + bx_i = \frac{x_i - \bar{x}}{s}$$

$z_1, \dots, z_n$  の平均は 0, 分散は 1

▶ 偏差値： $a = 50 - 10\frac{\bar{x}}{s}$ ,  $b = \frac{10}{s}$  math.pdf 1.4節

$$h_i = a + bx_i = 50 + 10\frac{x_i - \bar{x}}{s} = 50 + 10z_i$$

$h_1, \dots, h_n$  の平均 50, 分散  $10^2$

# 線形変換，標準化，偏差値 III

$z_1, \dots, z_n$  に対するチェビシェフの不等式

- ▶  $k > 1$  とし， $|z_i| > k$  となる個体数  $n_k$  とする
- ▶ このとき  $\frac{n_k}{n} < \frac{1}{k^2}$  が成立

## 解釈

- ▶  $|z_1|^2, \dots, |z_n|^2$  の平均は 1  $\Leftarrow \frac{\sum_{i=1}^n |z_i|^2}{n} = 1$  より分かる
- ▶  $|z_i|$  が大きくなる個体の数はある程度コントロールされる

証明  $n = \sum_{i=1}^n z_i^2 = \sum_{i=1}^n |z_i|^2 > k^2 n_k$

$$\begin{cases} |z_i| \rightarrow k & \text{if } |z_i| > k \\ |z_i| \rightarrow 0 & \text{if } |z_i| \leq k \end{cases} \quad \uparrow$$

## 線形変換，標準化，偏差値 IV

- ▶ チェビシェフの不等式により， $|z_i|$  が大きくなる個体の数はある程度コントロールされるが，，，
- ▶  $|z_i| > 5$  は起こりうる  
or  $\max h_i > 100$  や  $\min h_i < 0$  は起こりうる

$$\text{最大化問題 } \max z_1 \quad \text{制約} \begin{cases} \sum_{i=1}^n z_i = 0 \\ \sum_{i=1}^n z_i^2 = n \end{cases}$$

$$\text{解 } z_1 = \sqrt{n-1}, \quad z_j = -\frac{1}{\sqrt{n-1}} \text{ for } i \neq 1$$

↑  $n$  人の得点  $100, 0, 0, 0, \dots, 0$

## 線形変換，標準化，偏差値 V

- ▶ 分布の位置を表す統計量として，平均と中央値があったように，ばらつきを表す統計量も複数考えられる

$$\text{四分位偏差} \quad \frac{x_{0.75} - x_{0.25}}{2}$$

$$\text{平均絶対偏差} \quad \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

- ▶ これらは標準偏差よりも理解しやすいが，実際にはそれほど用いられない  
⇒ 平均，分散が数学的に非常に扱いやすいため

# データの要約 I

- ▶ 標本平均・標本分散や箱ひげ図（最小値，下側四分位点，中央値，上側四分位点，最大値の組）は一次元データの「要約」
- ▶ 「要約」した情報から元データは復元できず，一意性がない．
- ▶ 例：9個の数値からなる一次元データセット
  - ▶ 値が互いに相異なる
  - ▶ 最小値，下側四分位点，中央値，上側四分位点，最大値がそれぞれ 1, 3, 5, 7, 9
  - ▶ 標本平均が 5



## データの要約 II

1	2	3	4	5	6	7	8	9
1	1.5	3	4.5	5	5.5	7	8.5	9
1		3		5		7		9

↑ 複数の（無限の）データセットが合致する

- ▶ 1次元だと要約のありがたみが分かりにくいが,,
- ▶ データを要約する意義は多次元になるにつれて飛躍的に高まる

## 2 変量データの散布図と相関係数 I

多変量データ（多次元データ）：単一の変数  $x$  だけでなく、 $p$  変数に関する  $n$  個のデータ

個体 \ 変数	1	2	...	$p$
1	$x_{11}$	$x_{12}$	...	$x_{1p}$
2	$x_{21}$	$x_{22}$	...	$x_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$n$	$x_{n1}$	$x_{n2}$	...	$x_{np}$

まず 2 変数に注目 ↑

## 2 変量データの散布図と相関係数 II

- ▶ 2 変量データ  $(x, y)$  :  $n$  個の観測値

$$(x_1, y_1), \dots, (x_n, y_n)$$

- ▶  $x$  と  $y$  それぞれに注目し, 1 変量データと見たときの要約統計量 (平均, 分散)

$$\bar{x} = \frac{\sum x_i}{n}, \quad s_x^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$
$$\bar{y} = \frac{\sum y_i}{n}, \quad s_y^2 = \frac{\sum (y_i - \bar{y})^2}{n}$$

の重要性は既に学んだ

- ▶ ここでは,  $x$  と  $y$  の関係に注目する

## 2 変量データの散布図と相関係数 III

### ▶ 散布図： $n$ 個の観測値

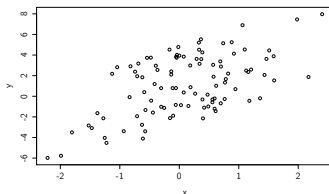
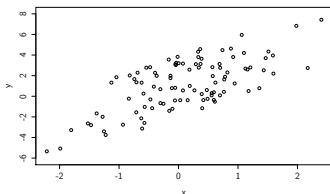
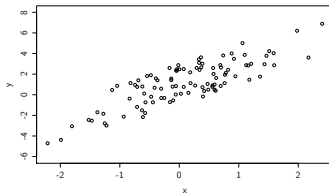
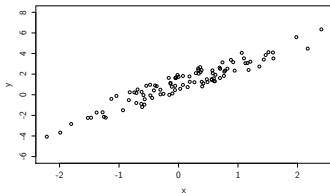
$$(x_1, y_1), \dots, (x_n, y_n)$$

を  $xy$  平面上の点とみて， $n$  個の点を平面上に打ったもの

- ▶ 二つの変数の大小に関連があることを，二つの変数の間に相関があるという
  - ▶ 正の相関：一方の変数の増加につれて他方の変数も増加する場合
  - ▶ 負の相関：逆に一方の変数の増加が他方の変数の減少に対応している場合

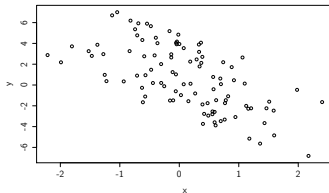
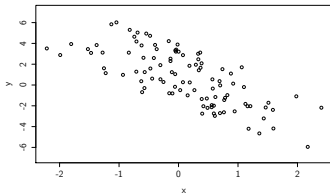
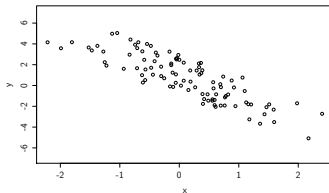
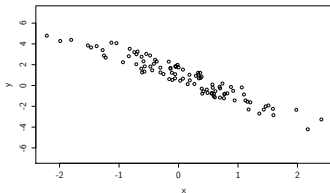
## 2変量データの散布図と相関係数 IV

### ► 散布図（正の相関）



# 2変量データの散布図と相関係数 $V$

## ▶ 散布図（負の相関）



## 2 変量データの散布図と相関係数 VI

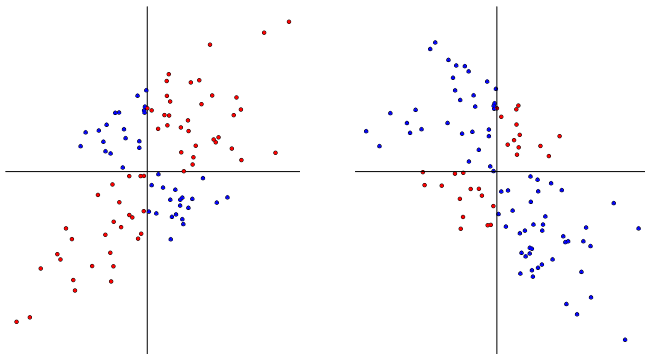
### ▶ 相関係数

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- ▶ (線形的な) 相関の強さを  $-1$  から  $1$  までの値で表現
- ▶ その正負は相関の正負に対応
- ▶ 絶対値  $|r_{xy}|$  が相関の強さを表現

# 相関係数の性質 I

## 正の相関, 負の相関 (原点 $(\bar{x}, \bar{y})$ )





# 相関係数の性質 II

## 相関係数の性質 (符号)

- ▶ 分母は常に正
- ▶ 分子  $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$  が  $r$  の符号を決める
- ▶  $(x_i - \bar{x})(y_i - \bar{y})$  の符号
  - ▶  $(x_i, y_i)$  が図2で原点を  $(\bar{x}, \bar{y})$  としたときの第1, 3象限の点であるとき正
  - ▶ 第2, 4象限のとき負
- ▶  $(\bar{x}, \bar{y})$  から見て、右上あるいは左下に多くの点が集まれば,  $r$  の分子は正
- ▶  $r > 0 \Leftrightarrow$  散布図の点が右上がりの傾向

# 相関係数の性質 III

## 相関係数の性質 (線形変換に関する不変性)

▶ 線形変換 
$$\begin{cases} z_i = a + bx_i & b > 0 \\ w_i = c + dy_i & d > 0 \end{cases}$$

このとき

$$z_i - \bar{z} = a + bx_i - \{a + b\bar{x}\} = b(x_i - \bar{x})$$

$$w_i - \bar{w} = c + dy_i - \{c + d\bar{y}\} = d(y_i - \bar{y})$$

▶  $r_{zw} = r_{xy}$

$$\begin{aligned} r_{zw} &= \frac{\sum_{i=1}^n (z_i - \bar{z})(w_i - \bar{w})}{\sqrt{\sum_{i=1}^n (z_i - \bar{z})^2 \sum_{i=1}^n (w_i - \bar{w})^2}} \\ &= \frac{\sum_{i=1}^n \{b(x_i - \bar{x})\} \{d(y_i - \bar{y})\}}{\sqrt{\{b^2 \sum_{i=1}^n (x_i - \bar{x})^2\} \{d^2 \sum_{i=1}^n (y_i - \bar{y})^2\}}} = r_{xy} \end{aligned}$$

## 相関係数の性質 IV

### 線形変換（単位変換を含む）の例

- ▶ 身長. 単位による絶対値の違い

$$1m = 100cm = 1000mm$$

- ▶ 気温. 摂氏と華氏  $C = \frac{5}{9}(F - 32)$

以下で相関係数は同じ！

- ▶ 身長と体重の相関係数 cm v.s. kg, m v.s. g
- ▶ 降水量と気温の相関係数 mm v.s. C, mm v.s. K

# 相関係数の性質 V

## 標準化と相関係数

▶ 標準化変換  $z_i = \frac{x_i - \bar{x}}{s_x}, w_i = \frac{y_i - \bar{y}}{s_y}$

$$\bar{z} = \frac{\sum z_i}{n} = 0, s_z = \frac{\sum (z_i - \bar{z})^2}{n} = \frac{\sum z_i^2}{n} = 1$$

$$\bar{w} = \frac{\sum w_i}{n} = 0, s_w = \frac{\sum (w_i - \bar{w})^2}{n} = \frac{\sum w_i^2}{n} = 1$$

▶  $(z_1, w_1), \dots, (z_n, w_n)$  の相関係数

$$r_{zw} = \frac{\sum_{i=1}^n (z_i - \bar{z})(w_i - \bar{w})}{\sqrt{\sum_{i=1}^n (z_i - \bar{z})^2 \sum_{i=1}^n (w_i - \bar{w})^2}} = \frac{1}{n} \sum_{i=1}^n z_i w_i$$

## 相関係数の性質 VI

範囲  $|r_{xy}| \leq 1$

▶ 以下の関係に注意

$$\begin{aligned} & \frac{1}{n} \sum (z_i \pm w_i)^2 \\ &= \frac{1}{n} \left( \sum z_i^2 \pm 2 \sum z_i w_i + \sum w_i^2 \right) \\ &= 2(1 \pm r_{zw}) = 2(1 \pm r_{xy}) \end{aligned}$$

↑ 左辺は常に非負（二乗の和）なので  $|r_{xy}| \leq 1$

# 相関係数の性質 VII

等号成立  $|r_{xy}| = 1$

▶  $r_{xy} = 1 \Leftrightarrow z_i = w_i \text{ for all } i = 1, \dots, n$

$$\Leftrightarrow y_i - \bar{y} = \frac{s_y}{s_x}(x_i - \bar{x}) \text{ for all } i = 1, \dots, n$$

$(x_1, y_1), \dots, (x_n, y_n)$  が全て右上がりの直線上

▶  $r_{xy} = -1 \Leftrightarrow z_i = -w_i \text{ for all } i = 1, \dots, n$

$$\Leftrightarrow y_i - \bar{y} = -\frac{s_y}{s_x}(x_i - \bar{x}) \text{ for all } i = 1, \dots, n$$

$(x_1, y_1), \dots, (x_n, y_n)$  が全て右下がりの直線上

# 共分散と相関係数 I

## ▶ $x$ と $y$ の間の共分散

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

相関係数の定義式の分子を  $n$  で割ったもの

## ▶ 記法 $s_{xy}$ . 定義より $s_{xy} = s_{yx}$


## 共分散と相関係数 II

▶ さらにこの記法によれば,

$$s_{x\textcolor{red}{x}} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(\textcolor{red}{x}_i - \bar{\textcolor{red}{x}}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad x \text{ の分散}$$

$$s_{\textcolor{red}{y}y} = \frac{1}{n} \sum_{i=1}^n (\textcolor{red}{y}_i - \bar{\textcolor{red}{y}})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad y \text{ の分散}$$

▶  $r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$  記法として  $s_x^2 = s_{xx}$  に注意

▶ 別表現  $s_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$   1.6節



# 相関と因果

## ▶ ニコラスケイジと溺死

↑ spurious correlation の一つの例

- ▶ 溺死者数とアイスクリーム消費量の相関（藪 66）
  - ▶ 両者に因果関係がある？
  - ▶ あるなら，溺死者数を減らすためにアイスクリームの消費の規制が有効
  - ▶ 気温↑⇒（アイスクリームの消費↑ & 海や川で泳ぐ人も増え溺死者数↑）
  - ▶ 規制を課してアイスクリーム消費量を減らしても，溺死者数は減らない
  - ▶ 気温：第3の変数や交絡因子

### 3 変量以上のデータの表示 I

- ▶ 3次元以上になると，変量間の関連を把握するのが難しくなるが， $p$ 個の変数から任意の2個を取り出して，その相関を調べることが基本

個体 \ 変数	1	2	...	$p$
1	$x_{11}$	$x_{12}$	...	$x_{1p}$
2	$x_{21}$	$x_{22}$	...	$x_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$n$	$x_{n1}$	$x_{n2}$	...	$x_{np}$

## 3 変量以上のデータの表示 II

### ▶ 平均ベクトル

$$\begin{pmatrix} \frac{1}{n} \sum_{t=1}^n x_{t1} \\ \vdots \\ \frac{1}{n} \sum_{t=1}^n x_{tp} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

- ▶ ベクトルにビビらないように
- ▶ 縦に  $p$  個並んだ箱に  $\bar{x}_1, \dots, \bar{x}_p$  を順に入れただけ

### 3 変量以上のデータの表示 III

#### ▶ 分散共分散行列

$$\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}, \quad s_{ij} = \frac{\sum_{t=1} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j)}{n}$$

- ▶ 右下方方向に昇順で  $(i, j)$  番地が割り振ってある  $p \times p$  の箱
- ▶  $(i, j)$  番地  $\leftarrow$  第  $i$  変数と第  $j$  変数の共分散  $s_{ij}$
- ▶ 対角線上の  $(i, i)$  番地は  $s_{ii}$ , つまり第  $i$  変数の分散

### 3 変量以上のデータの表示 IV

- ▶ 相関係数行列も同様に定義可能

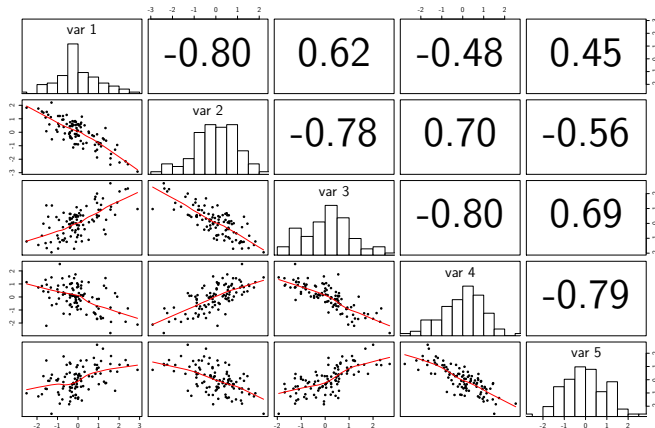
$$\begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}, \quad r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

- ▶ なぜ対角線上の番地は 1 なのか？

$$\frac{s_{i\textcolor{red}{i}}}{\sqrt{s_{ii}s_{\textcolor{red}{ii}}}} = \frac{s_{ii}}{s_{ii}} = 1$$

# 3 変量以上のデータの表示 V

## 散布図行列



# 分割表 I

- ▶ 散布図を描くのは主に連続変数の場合
- ▶ 質的変数や離散変数では多次元の度数分布が重要
- ▶ 分割表：多次元の質的変数の度数分布表

	恋人あり	恋人なし	行和
自宅	59	117	176
下宿	61	87	148
列和	120	204	324

## 分割表 II

### 用語

- ▶ セル：カテゴリのそれぞれの組み合わせ  
(例)「恋人あり × 自宅」のセルの度数 59
- ▶ 行和：各行の数字の和  
(例) 自宅生の総数は  $176 = 59 + 117$
- ▶ 列和：各列の数字の和  
(例) 恋人なしの総数は  $204 = 117 + 87$



## 分割表 III

行数  $r$ , 列数  $c$  とした  $r \times c$  分割表

第 1 変数 \ 第 2 変数	1	...	$c$	行和
1	$f_{11}$	...	$f_{1c}$	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
$r$	$f_{r1}$	...	$f_{rc}$	$f_{r\cdot}$
列和	$f_{\cdot 1}$	...	$f_{\cdot c}$	$n$

- ▶  $f_{ij}$  : 第 1 変数がカテゴリ  $i$ , 第 2 変数がカテゴリ  $j$  であるような組み合わせ,  $(i, j)$  セルの観測値の度数
- ▶  $f_{i\cdot} = \sum_{j=1}^c f_{ij}$ ,  $f_{\cdot j} = \sum_{i=1}^r f_{ij}$

## 分割表 IV

### 分割表の一つの見方

- ▶ 各行内，各列内での相対頻度を計算して比較
- ▶ 行方向
  - ▶ 自宅生の中で恋人ありの比率  $59/176 = 33.5\%$
  - ▶ 下宿生の中で恋人ありの比率  $61/148 = 41.2\%$
- ▶ 列方向
  - ▶ 恋人ありの中で自宅生の比率  $49.2\%$
  - ▶ 恋人なしの中で自宅生の比率  $57.4\%$

下宿生の方が恋人を持ちやすい傾向が見える

# 統計リテラシー，シンプソンのパラドクス

	受験者		男性		女性		
		受験者	合格者	合格率	受験者	合格者	合格率
A 専攻	100	80	40	50%	20	16	80%
B 専攻	50	20	2	10%	30	3	10%
	150	100	42	42%	50	19	38%

- ▶ 2 専攻合算での合格率男性 42%，女性 38%  
↑ 男性の方が合格率高い
- ▶ 専攻ごとの合格率  
↑ A 専攻では男:女 = 5 : 8, B 専攻では 1 : 1
- ▶ 専攻ごとと、全体とでは結果が逆  
↑ 矛盾のように見えるのでパラドクス

# シンプソンのパラドクス：ブログ記事より I

新薬 A				従来の治療		
	効果なし	効果あり	効あ割合	効果なし	効果あり	効あ割合
女性	3	37	92.5%	1	19	95%
男性	8	12	60%	12	28	70%
合計	11	49	82%	13	47	78%

## あるブログ

- ▶ 男性でも女性でも効かないが、人間（男女合計）には効果が高い新薬 A なるものが存在しうるのか？
- ▶ 男性でも女性でも効かないなら、集団全体で見ても効果がないと考えるのが自然な発想

# シンプソンのパラドクス：ブログ記事より II

## シンプソンのパラドックス

集団全体を見た時とその小集団（特定の質的変数の値で層別された小集団）に注目した時で一見矛盾した結論がデータから導かれてしまうこと

処理Ⅰ				処理Ⅱ		
	効果なし	効果あり	効あ割合	効果なし	効果あり	効あ割合
群 1	$A$	$B$	$\frac{B}{A+B}$	$C$	$D$	$\frac{D}{C+D}$
群 2	$a$	$b$	$\frac{b}{a+b}$	$c$	$d$	$\frac{d}{c+d}$
合計	$A+b$	$B+b$	$\frac{B+b}{A+a+B+b}$	$C+c$	$D+d$	$\frac{D+d}{C+c+D+d}$

# シンプソンのパラドクス：ブログ記事より III

## シンプソンのパラドクスの数学

$$\begin{cases} \frac{B}{A+B} < \frac{D}{C+D} \\ \frac{b}{a+b} < \frac{d}{c+d} \end{cases} \Rightarrow \frac{B+b}{A+a+B+b} > \frac{D+d}{C+c+D+d}$$

数学的には、左の2つの大小関係から右の大小関係が従うことに違和感を覚えるだけ

$$\text{同値} \begin{cases} \frac{B}{A} < \frac{D}{C} \\ \frac{b}{a} < \frac{d}{c} \end{cases} \Rightarrow \frac{B+b}{A+a} > \frac{D+d}{C+c}$$

# シンプソンのパラドクス：ブログ記事より IV

## Judea Pearl の説明

- ▶ 集団全体（例：男女合計）とその小集団内（例：男女別）で関連の方向性が逆転すること自体はパラドックスでない
- ▶ 数学

$$\frac{B}{A} < \frac{D}{C} \ \& \ \frac{b}{a} < \frac{d}{c} \text{ から } \frac{B+b}{A+a} < \frac{D+d}{C+c} \text{ は従わない}$$

シンプルに数学的な性質．関連が逆転したことをもってパラドックスと呼ぶのは正確ではない

# シンプソンのパラドクス：ブログ記事より V

- ▶ なぜ新薬 A の例ではデータが矛盾しているように感じたか？
  - ▶ データから得られる「結果」とその「解釈」
  - ▶ データから得られる結果
    - ▶ 男性：従来の治療のほうが効果があった人の割合が高い
    - ▶ 女性：従来の治療のほうが効果があった人の割合が高い
    - ▶ 集団全体：新薬 A のほうが効果があった人の割合が高い
- 「割合が高い」：完全に数学的表現。集団全体で関連の方向性が逆転したこと自体も数学的に起こりうる



# シンプソンのパラドクス：ブログ記事より VI

## ▶ データから得られる結果の「解釈」

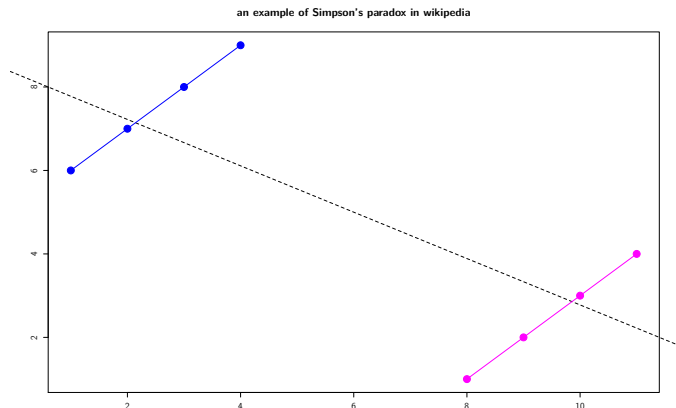
- ▶ 男性：従来の治療のほうが効き目がいい
- ▶ 女性：従来の治療のほうが効き目がいい
- ▶ 集団全体：新薬 A のほうが効き目がいい

「効き目がいい」という表現に，治療の種類と効果の有無の間の因果関係の想定．関連と因果は違う．

- ▶ データの結果に因果的な解釈を持ち込むことでデータが矛盾しているように感じるのが「パラドックス」が生じる原因

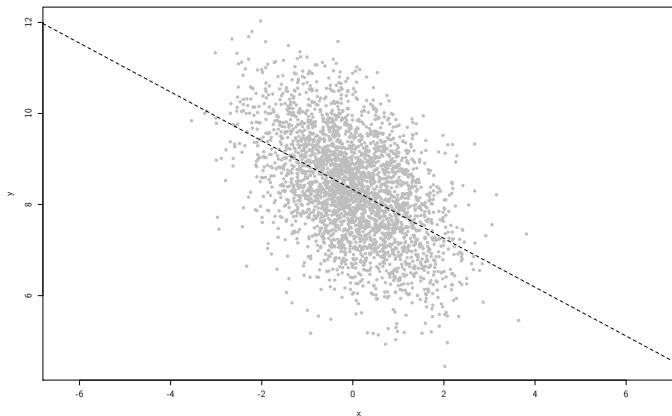
# シンプソンのパラドクス, 量的変数 I

## Wikipedia による散布図



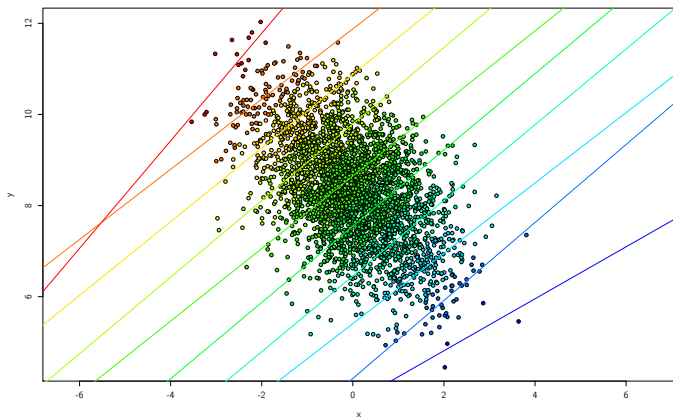
# シンプソンのパラドクス, 量的変数 II

Microsoft の datascientist の Bob Horton さんによる 散布図



# シンプソンのパラドクス, 量的変数 III

Microsoft の datascientist の Bob Horton さんによる **散布図**



# 回帰分析の概要

## 回帰分析

ある変数の値に基づいて、他の変数を説明したり  
予測したりするための手法

## 用語

- ▶ 説明変数 説明に用いる変数
- ▶ 目的変数 or 被説明変数 説明の対象となる変数
- ▶ 目的変数は一次元，説明変数は多次元の変数
- ▶ 単回帰分析 説明変数が単一の場合
- ▶ 重回帰分析 説明変数が複数の場合

# 単回帰分析 I

- ▶  $x$  の値を用いて  $y$  を予測  
↑  $x$  説明変数,  $y$  目的変数
- ▶ 予測式として最も単純なのは 1 次式であり,  $x$  と  $y$  の関係が近似的に

$$y \doteq a + bx \quad (1)$$

を満たすなら,  $a + bx$  により  $y$  の値を予測可能

- ▶ 記号  $\doteq$  は近似的に等しいことを表す
- ▶ 式 (1) のような直線を求めることを直線を当てはめるといふ
- ▶  $a$  定数項,  $b$  回帰係数

## 単回帰分析 II

- ▶ 既に得られている  $x$  と  $y$  についての  $n$  個のデータ

$$(x_1, y_1), \dots, (x_n, y_n)$$

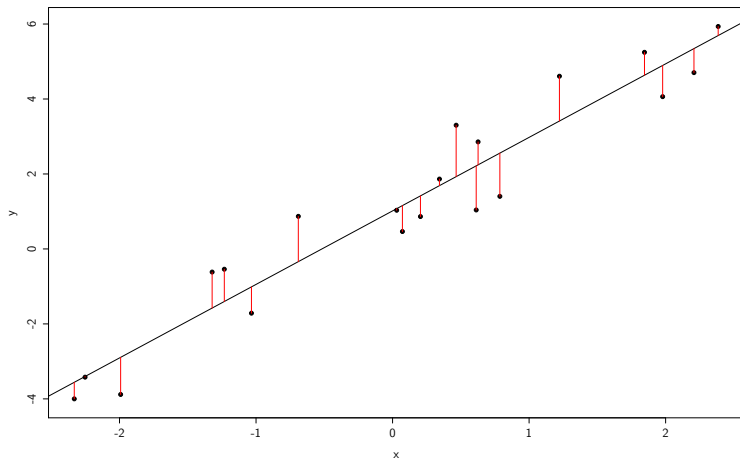
に基づいて, (1) のような直線を当てはめたい

- ▶  $x_i$  から予測される  $y$  の値  $a + bx_i$  と現実の値  $y_i$  の差

$$y_i - (a + bx_i), \quad i = 1, \dots, n$$

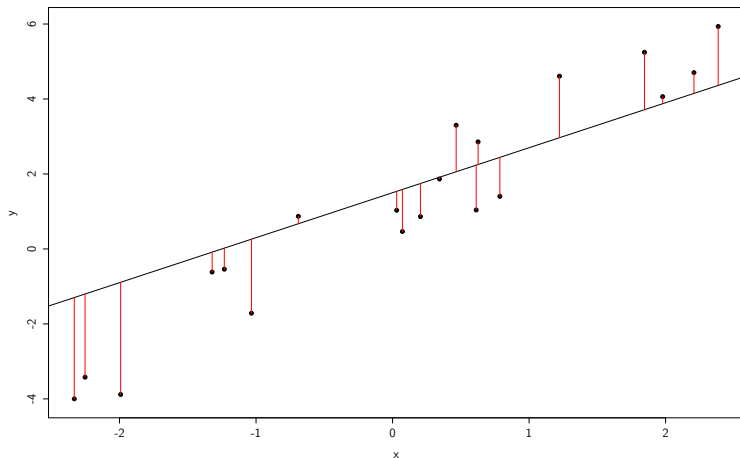
の絶対値が, (全体的な意味で) 小さくなるように直線を当てはめるのが自然

# 単回帰分析 III

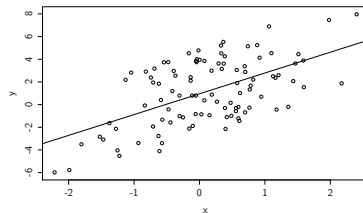
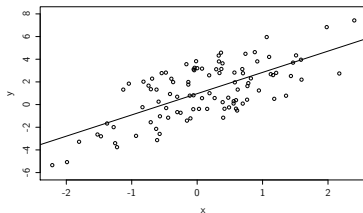
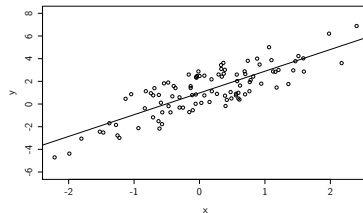
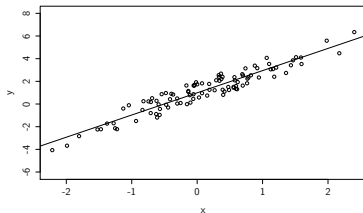




# 単回帰分析 IV



# 単回帰分析 V



# 単回帰分析 VI

## ▶ $n$ 個の絶対値

$$|y_1 - (a + bx_1)|, |y_2 - (a + bx_2)|, \dots, |y_n - (a + bx_n)|$$

## ▶ 数学的には，平方和

$$\begin{aligned} Q(a, b) &= \sum_{i=1}^n |y_i - (a + bx_i)|^2 \\ &= \sum_{i=1}^n \{y_i - (a + bx_i)\}^2 \end{aligned} \tag{2}$$

を最小にする  $a, b$  を求めることが扱いが容易

# 単回帰分析 VII

## $Q(a, b)$ の最小化問題

- ▶ 2変数関数  $Q(a, b)$  の最小化問題は、1変数関数のそれに比べてちょっと難しい
- ▶ でも、最小化問題において、微分が重要な役割を果たすことは知っているはず
- ▶ 今の場合は

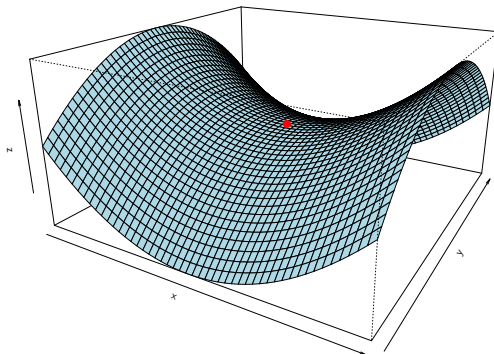
$$\frac{d}{da}Q(a, b) = 0, \quad \frac{d}{db}Q(a, b) = 0$$

の連立方程式の解が  $Q(a, b)$  を最小にする

↑ 最小化の証明 [math.pdf](#) 1.11節

# 単回帰分析 VIII

大域的な最小解ではないかもしれない



# 単回帰分析 IX

## 微分の性質 (和)

- ▶  $f(z)$  の微分を,  $f'(z)$  や  $\frac{d}{dz}f(z)$  と書く
- ▶ 関数の和  $f_1(z) + f_2(z)$  の微分は

$$f'_1(z) + f'_2(z)$$

- ▶  $\sum_{j=1}^m f_j(z) = f_1(z) + \cdots + f_m(z)$  の微分

$$\frac{d}{dz} \left( \sum_{j=1}^m f_j(z) \right) = f'_1(z) + \cdots + f'_m(z)$$

和の微分は微分の和

# 単回帰分析 X

## 微分の性質（合成関数）

▶  $f(g(z))$  の微分

$$\frac{d}{dv}f(v)|_{v=f(z)} \frac{d}{dz}g(z), \quad f'(g(z))g'(z)$$

▶ 例：  $f(cz)$  の場合

$$\frac{d}{dz}f(cz) = cf'(cz)$$

# 単回帰分析 XI

$$\frac{d}{da}Q(a, b)$$

- ▶  $Q$  の  $i$  番目の成分  $\{y_i - (a + bx_i)\}^2$
- ▶ その  $a$  に関する微分

$$2\{y_i - (a + bx_i)\} \times (-1) = 2(-y_i + a + bx_i)$$

- ▶ 和の微分は微分の和だから

$$\frac{d}{da}Q(a, b) = 2 \sum_{i=1}^n (-y_i + a + bx_i) = 2n(-\bar{y} + a + b\bar{x})$$

$$\uparrow \bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$



# 単回帰分析 XII

$$\frac{d}{db}Q(a, b)$$

- ▶  $Q$  の  $i$  番目の成分  $\{y_i - (a + bx_i)\}^2$
- ▶ その  $b$  に関する微分

$$2\{y_i - (a + bx_i)\} \times (-x_i) = 2x_i(-y_i + a + bx_i)$$

- ▶ 和の微分は微分の和だから

$$\begin{aligned}\frac{d}{db}Q(a, b) &= 2 \sum_{i=1}^n x_i(-y_i + a + bx_i) \\ &= 2 \left( - \sum_{i=1}^n x_i y_i + an\bar{x} + b \sum_{i=1}^n x_i^2 \right)\end{aligned}$$

## 単回帰分析 XIII

$$\text{連立方程式 } \frac{d}{da}Q(a, b) = 0, \frac{d}{db}Q(a, b) = 0$$

$$\begin{cases} a + b\bar{x} = \bar{y} \\ an\bar{x} + b\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$

$$\text{最小二乗解 } \check{a} = \bar{y} - \check{b}\bar{x}$$

$$\check{b} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

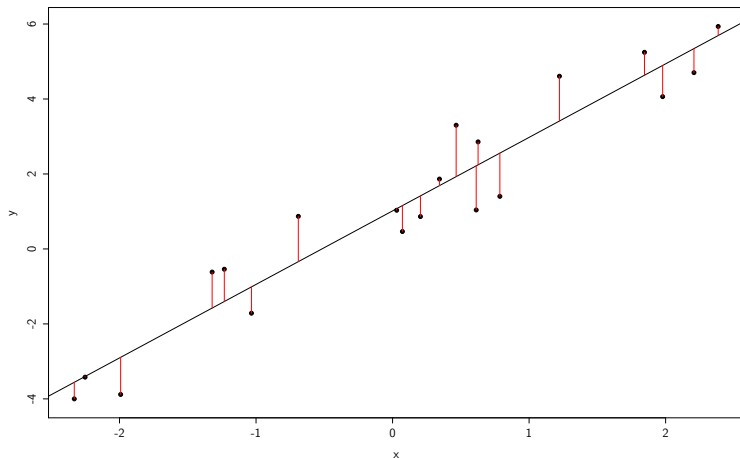
$\check{b}$  の表現における計算の確認 [math.pdf](#) 1.5, 1.6節

## 単回帰分析 XIV

当てはまりの程度

- ▶ 最小二乗法は  $Q(a, b) = \sum_{i=1}^n \{y_i - (a + bx_i)\}^2$  を最小にするという意味で最もデータに当てはまる直線を与える
- ▶ 回帰直線を求めた後には、その当てはまりの程度が問題
- ▶ 当てはまりの程度、関係の強さの尺度
  - ▶ よく当てはまっている場合には大
  - ▶ あまり当てはまっていない場合には小

# 単回帰分析 XV



# 単回帰分析 XVI

予測値, 残差

- ▶ 予測値  $\hat{y}_i$  : 回帰直線上の  $y$  の値

$$\hat{y}_i = \check{a} + \check{b}x_i, \quad i = 1, 2, \dots, n$$

最小二乗解  $\check{a}, \check{b}$  :  $Q(a, b)$  を最小化する  $a, b$

- ▶ 実測値  $y_i$  : 実際の観測値
- ▶ 残差 (予測式の外れ具合) : 実測値 - 予測値

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n$$

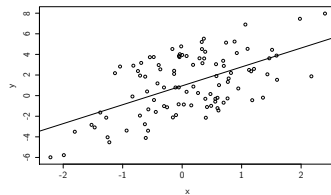
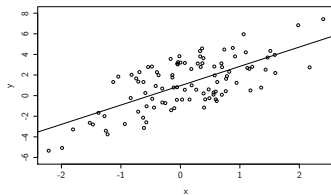
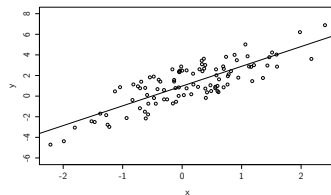
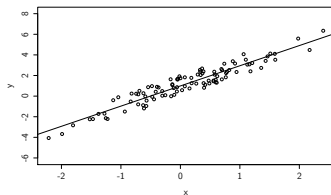
# 単回帰分析 XVII

## 残差平方和

$$\sum_{i=1}^n e_i^2 = \min_{a,b} Q(a, b)$$

- ▶ 残差平方和  $\sum_{i=1}^n e_i^2$  が小さいほど，回帰直線のあてはまりがよい気がする
- ▶ 残差平方和を「当てはまりの良さ」の指標として良いか？
- ▶ 左上 → 右上 → 左下 → 右下の順に残差平方和が大きくなる

# 単回帰分析 XVIII



# 単回帰分析 XIX

## 問題点

- ▶ 残差平方和は、 $y$  の単位の取り方（より一般に線形変換）に依存 [math.pdf](#) 1.7節 & 1.8節

$y_i \rightarrow cy_i + d$  のとき、残差平方和は  $c^2$  倍

### 単位や線形変換の例

- ▶  $y$  身長. 単位による絶対値の違い

$$1m = 100cm = 1000mm$$

- ▶  $y$  気温. 摂氏と華氏  $C = \frac{5}{9}(F - 32)$



## 単回帰分析 XX

$$\frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \Leftarrow y \text{ の線形変換に依存しない}$$

- ▶  $y_i \rightarrow cy_i + d$  に対し,

$$\sum_{i=1}^n (y_i - \bar{y})^2 \rightarrow \sum_{i=1}^n (cy_i - c\bar{y})^2 = c^2 \sum_{i=1}^n (y_i - \bar{y})^2$$

- ▶ 分母分子とも線形変換  $y \rightarrow cy + d$  により  $c^2$  倍

- ▶  $\frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$  は小さいほど嬉しい量

# 決定係数 I

## 決定係数

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

- ▶ 大きい方が嬉しい量
- ▶ 決定係数は、 $0 \leq R^2 \leq 1$  を満たし、1 に近い程、回帰式の当てはまりが良い

## 決定係数 II

決定係数  $R^2$  の性質  $0 \leq R^2 \leq 1$

- ▶ 1 以下  $\Leftarrow$  1 から非負の量を引いている
- ▶ 0 以上

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\min_{a,b} Q(a, b)}{Q(\bar{y}, 0)}$$

ただし,  $Q(a, b) = \sum (y_i - \{a + bx_i\})^2$

# 決定係数 III

決定係数の別表現 [math.pdf](#) 1.9, 1.10節

▶  $R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

▶  $R^2$  と相関係数

$$R^2 = \left( \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 = r_{xy}^2$$

- ▶ 頑張って  $R^2$  を考えたのは無駄？
- ▶ 次に習う重回帰分析においては、これに対応する関係はないので、無駄ではない

# 重回帰分析 I

- ▶  $p$  個の説明変数  $x_1, \dots, x_p$  に対して,

$$y \doteq b_0 + b_1x_1 + \dots + b_px_p$$

なる線形の近似式で予測

- ▶  $n$  組のデータに対し

$$y_i = (b_0 + b_1x_{i1} + \dots + b_px_{ip}), \quad i = 1, \dots, n$$

の絶対値が(全体的に)小さくなるように  $b_0, \dots, b_p$  を求めたい

## 重回帰分析 II

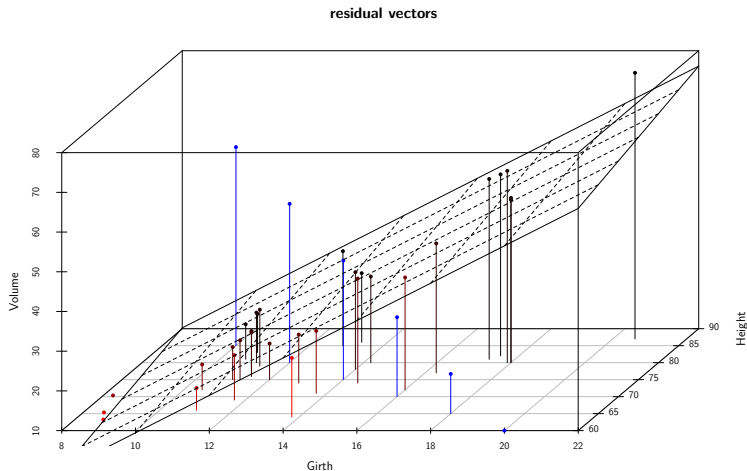
### ▶ 重回帰分析における $n$ 組のデータ

$$\begin{array}{ccccccccc} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{array}$$

### ▶ 最小二乗法 : $Q(b_0, b_1, \dots, b_p)$ の最小化

$$\begin{aligned} Q(b_0, b_1, \dots, b_p) \\ = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2 \end{aligned}$$

# 重回帰分析 III



# 重回帰分析 IV

最小二乗解      最小化の証明 [math.pdf](#) 1.11節

- ▶ 最小化の必要条件： $Q$  を  $b_0, \dots, b_p$  で（偏）微分した  $p + 1$  個を全て  $= 0$  とした連立方程式

$$\frac{d}{db_0} Q = 2 \sum_{i=1}^n (-1)(y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip}) = 0$$

$$\frac{d}{db_1} Q = 2 \sum_{i=1}^n (-x_{i1})(y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip}) = 0$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$\frac{d}{db_p} Q = 2 \sum_{i=1}^n (-x_{ip})(y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip}) = 0$$

連立方程式の解を  $\check{b}_0, \check{b}_1, \dots, \check{b}_p$  とする



# 重回帰分析 V

方程式の解  $\check{b}_0, \check{b}_1, \dots, \check{b}_p$  の具体的な表現

- ▶  $n$  次元ベクトル (たち) と行列

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}, \dots, \mathbf{x}_p = \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix}$$

$\mathbf{X} = (\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p) : n \times (p+1)$  行列

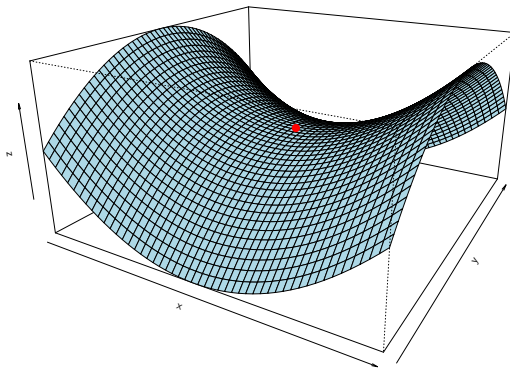
- ▶  $\mathbf{b} = (b_0, b_1, \dots, b_p)^T \in \mathbb{R}^{p+1}$  とするとき, 連立方程式は

$$\mathbf{X}^T \mathbf{X} \mathbf{b} - \mathbf{X}^T \mathbf{y} = \mathbf{0}$$

- ▶  $\mathbf{X}^T \mathbf{X}$  が正則 (逆行列をもつ) ならば, 最小二乗解は

$$\check{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

# 重回帰分析 VI



## 重回帰分析 VII

決定係数（単回帰の場合の繰り返し）

- ▶ 最小二乗法は  $Q(b_0, b_1, \dots, b_p)$  を最小にするという意味で最もデータに当てはまる平面を与える
- ▶ 回帰平面を求めた後には，その当てはまりの程度が問題
- ▶  $\check{b}_0, \dots, \check{b}_p$  : 最小二乗解
- ▶  $\hat{y}_i$  予測値 : 回帰平面上の  $y$  の値

$$\hat{y}_i = \check{b}_0 + \check{b}_1 x_{i1} + \dots + \check{b}_p x_{ip}, \quad i = 1, \dots, n$$

- ▶  $y_i$  : 実測値,  $e_i$  : 残差  $y_i - \hat{y}_i$

## 重回帰分析 VIII

▶ 決定係数  $R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

$$= 1 - \frac{\min_{b_0, b_1, \dots, b_p} Q(b_0, b_1, \dots, b_p)}{Q(\bar{y}, 0, \dots, 0)}$$

- ▶ 「 $0 \leq R^2 \leq 1$ 」また「1に近い程当てはまりが良い」という解釈は，単回帰の場合と同じ

▶ 決定係数の別表現  $R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$  math.pdf 1.13節

# 補遺 I

## 回帰係数の解釈

- ▶  $y \doteq b_0 + b_1x_1 + \cdots + b_px_p$  において  $i$  番目の説明変数  $x_i$  の係数  $b_i$  を考える
- ▶ 他の説明変数を固定して  $x_i$  だけを  $x_i \rightarrow x_i + 1$  と増加
- ▶ このとき  $y$  の予測値は  $b_i$  だけ増える
- ▶  $b_i$  :  $x_i$  を 1 単位増加させたときの  $y$  の平均的な増分

## 補遺 II

### $R^2$ についての補遺

- ▶ 説明変数の数を増やすと  $R^2$  は?
- ▶ 新たな説明変数  $z$ , その係数を  $d$  とすると,

$$\begin{aligned} & \min_{(c_0, \dots, c_p, d)} \sum_{t=1}^n \{y_t - c_0 - c_1 x_{t1} - \dots - c_p x_{tp} - dz_t\}^2 \\ & \leq \min_{(c_0, \dots, c_p)} \sum_{t=1}^n \{y_t - c_0 - c_1 x_{t1} - \dots - c_p x_{tp}\}^2 \end{aligned}$$

- ▶ 決定係数  $R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

## 補遺 III

- ▶ 決定係数  $R^2$  の定義から，説明変数を ( $y$  に関係あろうがなかろうが) 増やすと，決定係数  $R^2$  が大きくなることを示している
- ▶ 極端な話，説明変数を  $n - 1$  個用意して (各説明変数がベクトルとしてが一次独立となるように選んでやると)，残差は 0 (つまり決定係数は 1) になる
- ▶ 現段階では，このおかしさを修正する方法は説明できない
- ▶ 推測統計的な回帰分析の必要性

## 補遺 IV

### 発展

- ▶ そもそも  $y$  に関係ある説明変数が始めから分かっていたら簡単.
- ▶  $y$  を説明するのに関係あるかどうか分からない説明変数が大量にあって、その中から適切な（意味のある）説明変数の組を見つけて、 $y$  を言い当てるモデルを作りたい
- ▶ それ以外にも、そもそも  $x \rightarrow y$  の影響は線形なの？ など、いろいろ考えて拡張したくなりますよね？