広島大学

データサイエンス基礎

Fundamental Data Science

第4回　データの要約

広島大学　数理統計学グループ

私たちがナビゲートします！

今日の内容

- **データの要約（続き）**

前回の復習

- **ヒストグラム**

  分布を視覚的に表す

- **要約統計量**

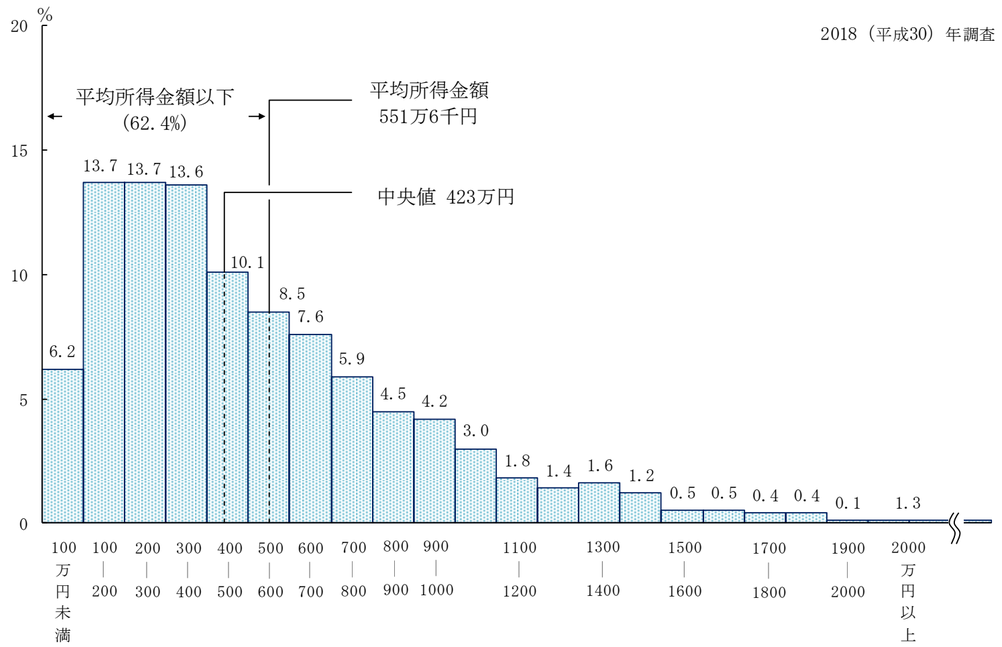
  分布の位置を数値で知る

  平均値, 中央値

普通の世帯の所得を知りたい ⇒ **中央値**

世帯全体の所得の中心を知りたい ⇒ **平均値**

図. 2018年所得金額階級別世帯数のヒストグラム

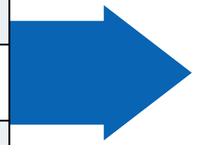


図引用: 厚生労働省 https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa18/index.html

練習問題 (1)

表. 10人の小テストA, B, C の成績 (10点満点)

生徒/小テスト	A	B	C
生徒1	0	0	3
生徒2	3	1	4
生徒3	3	2	4
生徒4	5	3	5
生徒5	5	5	5
生徒6	5	5	5
生徒7	5	7	5
生徒8	7	8	6
生徒9	7	9	6
生徒10	10	10	7



	A	B	C
平均値			
中央値			

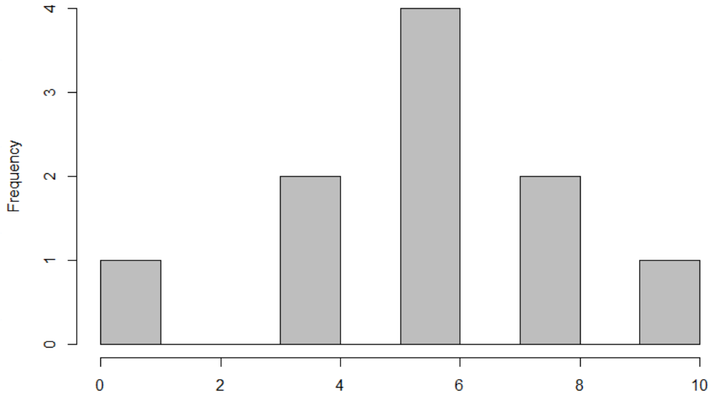
Q 上の表を埋めよ

Q どの小テストも
分布の特徴は同じ?

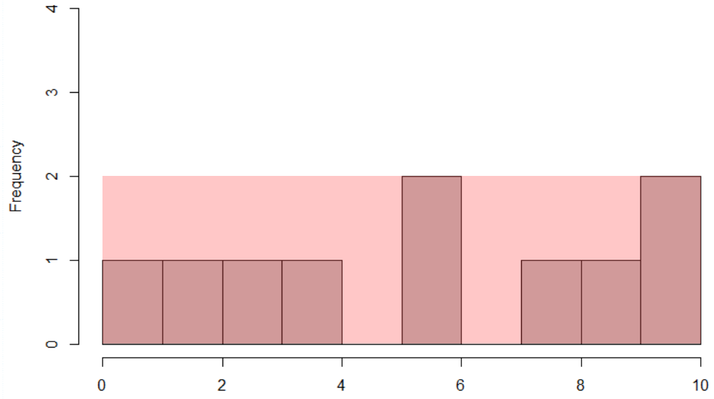
分布の散らばり

分布の位置は同じでも散らばり具合が違う

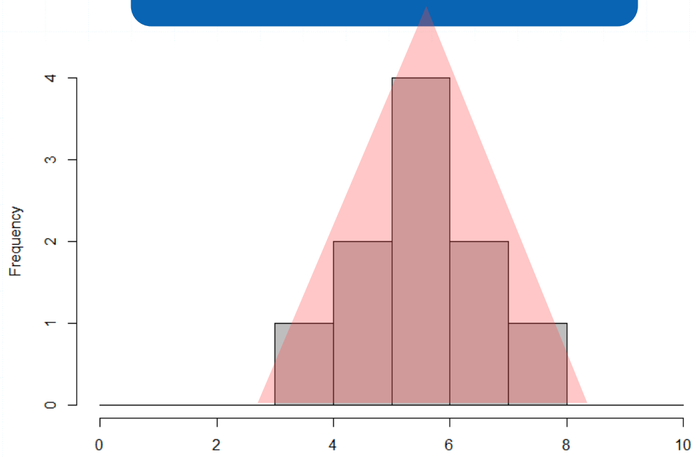
小テストA



小テストB



小テストC



Q 散らばりを表す要約統計量は?

分布の散らばり
分散，標準偏差

分散

分散 (variance)

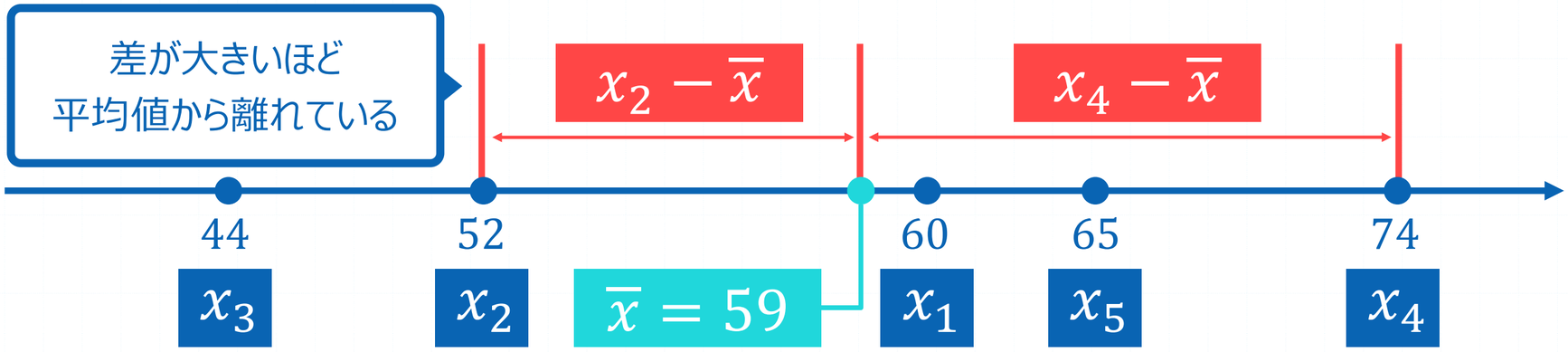
変数 x の n 個のデータ $x_1, \ldots, x_n$ に対する分散:

$$\text{分散 } s_x^2 = \frac{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}$$

$$\overline{x} = \frac{x_1+\cdots+x_n}{n}\text{: 平均値}$$

個体	変数 x
1	x_1
2	x_2
⋮	⋮
$n-1$	x_{n-1}
n	x_n

平均値周りの散らばり (バラつき) 具合を表す



人	体重 (kg)
1	60
2	52
3	44
4	74
5	65

分散

分散の解釈

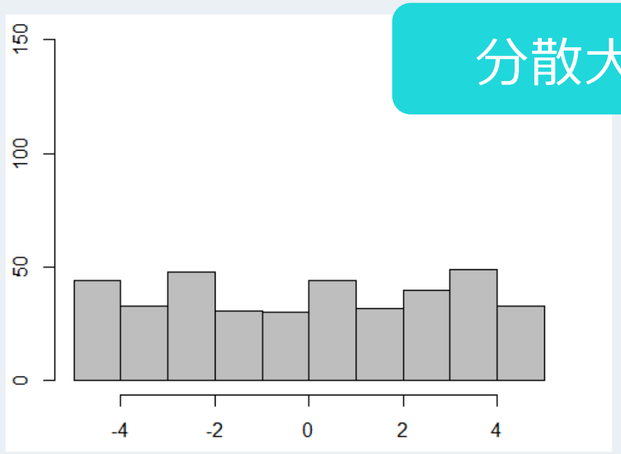
分散が大きい ⇔ データ全体が平均値から離れている傾向
⇔ **散らばりが大きい**

分散が小さい ⇔ データ全体が平均値に近い傾向
⇔ **散らばりが小さい**

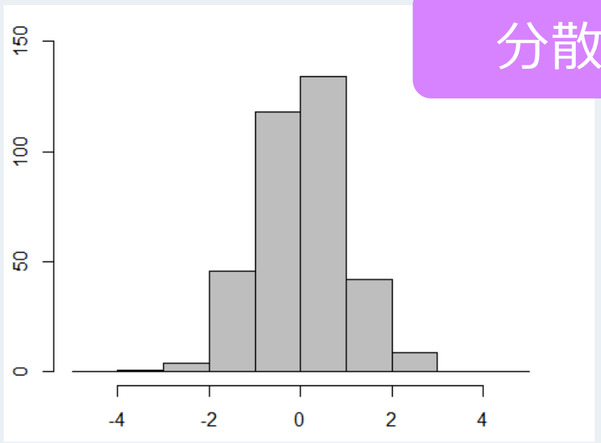
$$s_x^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$$

ヒストグラムによるイメージ図

分散大



分散小



分散

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

小テストA, B, C の分散:

小テストA: 分散 = 6.6

小テストB: 分散 = 10.8

小テストC: 分散 = 1.2

散らばり具合

C < A < B

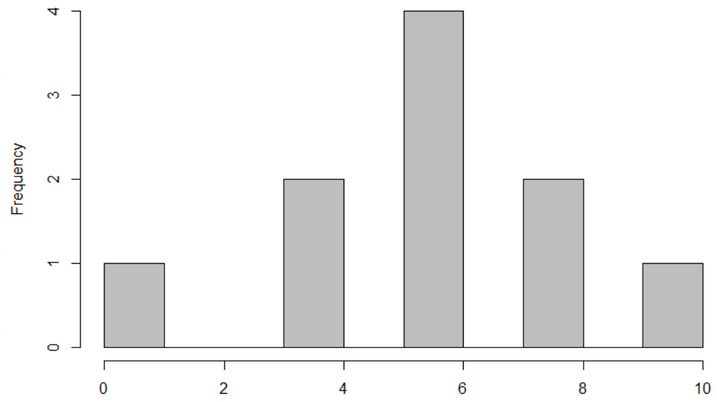
数値 (要約統計量) で
わかる!!

先ほどの
小テストの例で
考えてみましょう

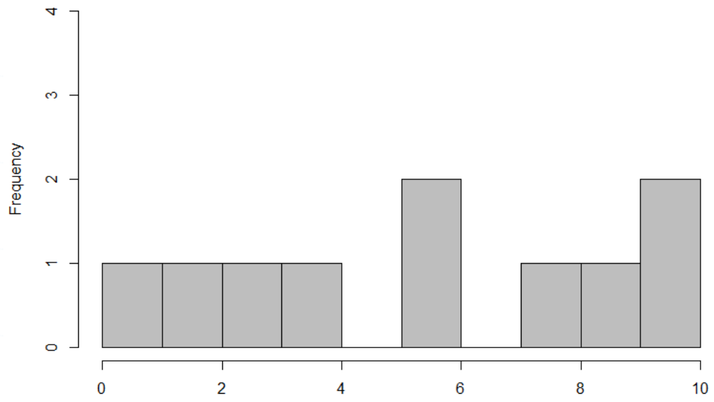
分散

小テストA, B, C の分散:

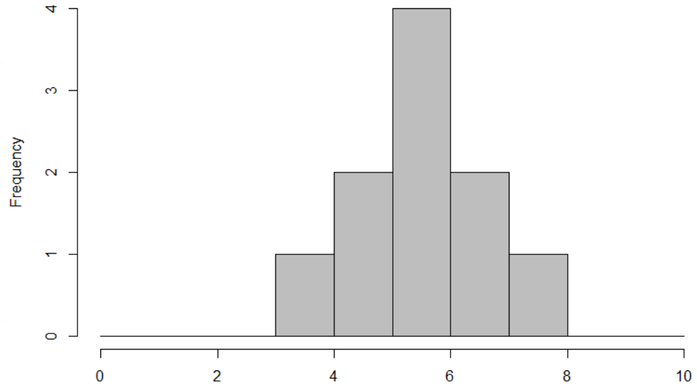
小テストA: 分散 = 6.6



小テストB: 分散 = 10.8



小テストC: 分散 = 1.2



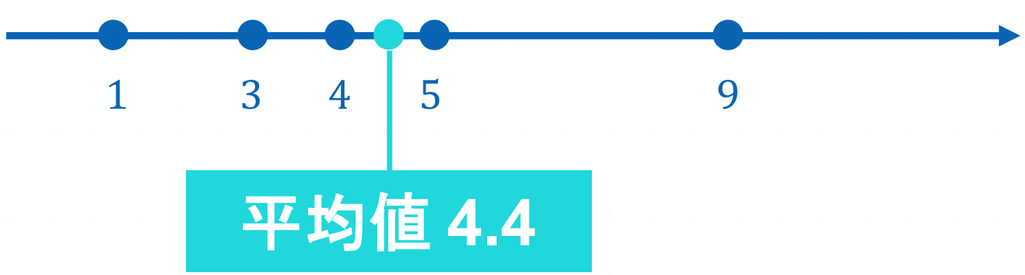
散らばり具合: **C < A < B**

分散のコメント1

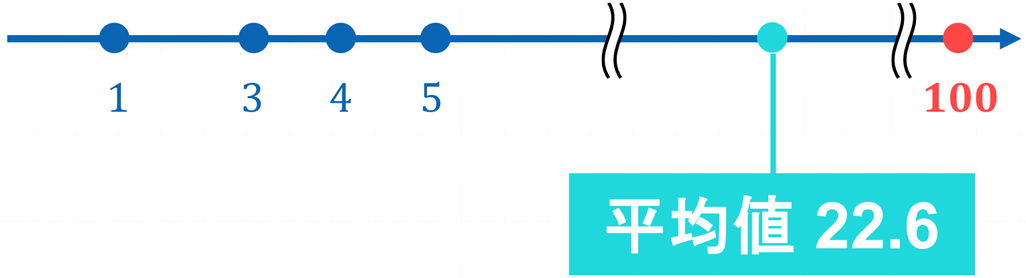
平均値と分散はセットで用いる

- データが全て同じ値なら, 分散 ＝ 0
- 分散は**外れ値**の影響を受けやすい

例 **1,3,4,5,9 の分散: 7.04**



1,3,4,5,100 の分散: 1499.44



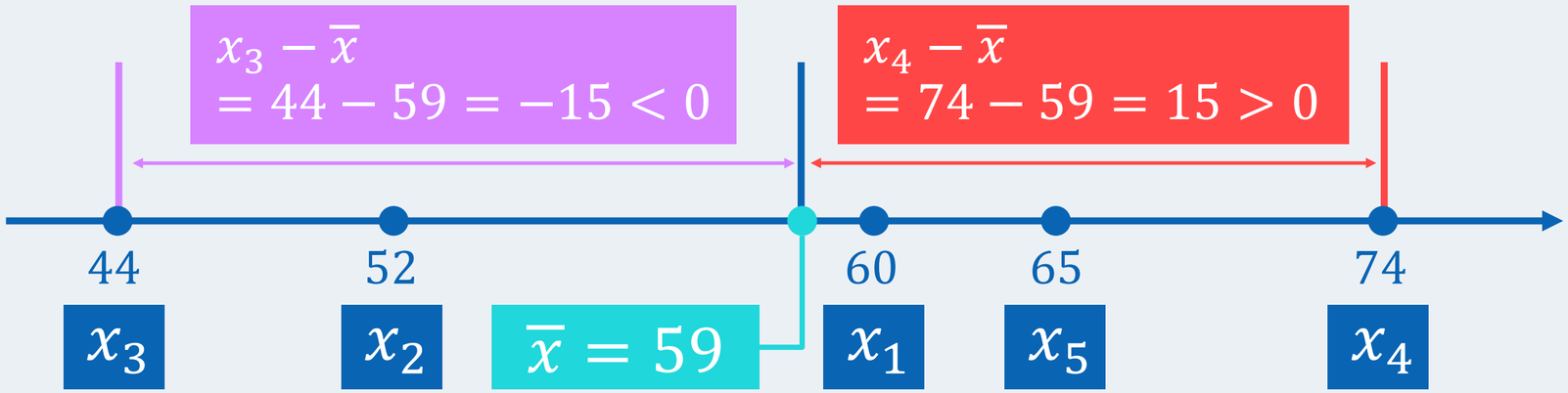
分散のコメント2

分散の式で2乗している理由

$$s_x^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$$

- $x_i - \overline{x}$ は負の値をとるかも?? (正と負の足し算で相殺される)

  → 相殺されないように**正の数**で表すため



人	体重 (kg)
1	60
2	52
3	44
4	74
5	65

- 絶対値で正にすると, 数学的に扱いにくい

標準偏差

標準偏差 (Standard Deviation; SD)

変数 x の n 個のデータ $x_1, \ldots, x_n$ に対する標準偏差:

標準偏差 $s_x = \sqrt{s_x^2}$

標準偏差 ＝ 分散の正の平方根

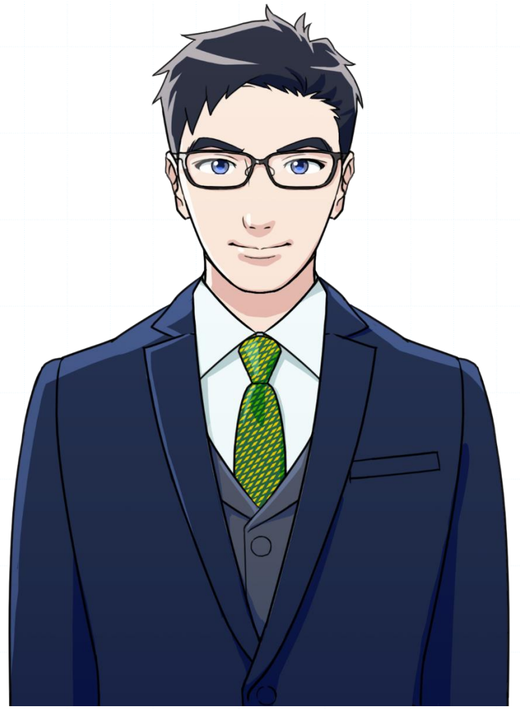
分散の単位を元の単位に戻したもの

例 **5cm, 10cm, 12cm, 13cm**

分散 **9.5 (cm²)**

標準偏差 **約 3.08 (cm)**

正の平方根とは
2乗してルートの中身のもの
になる正の値のことです



練習問題 (2)

Q 以下のデータの平均値, 分散, 標準偏差を求めよ

- 1,4,5,3,7

- 4,4,4,4,4

解答: 練習問題 (2)

- 1,4,5,3,7

$x_1 = 1,\ x_2 = 4,\ x_3 = 5,$
$x_4 = 3,\ x_5 = 7$ と思うと

平均値

分散

標準偏差

解答: 練習問題 (2)

- 4,4,4,4,4

$x_1 = 4,\ x_2 = 4,\ x_3 = 4,\ x_4 = 4,\ x_5 = 4$ と思うと

平均値

分散

標準偏差

実際に計算してみると
記憶に定着しやすいですね

分布の散らばり
範囲，四分位範囲

範囲

範囲 (range)

変数 x の n 個のデータ $x_1, \ldots, x_n$ の範囲:

範囲 R = (データの最大値) – (データの最小値)

範囲 = **全てのデータ100%**が含まれる区間の大きさ

範囲が大きい → 散らばり大

範囲が小さい → 散らばり小

外れ値の影響を受けやすい

例 **1,4,5,3,7 の範囲 R = 7 - 1 = 6**　　**1,4,5,3,100 の範囲 R = 100 – 1 = 99**

四分位点

四分位点 (quartile)

変数 x の n 個のデータ $x_1, \ldots, x_n$ の四分位点:

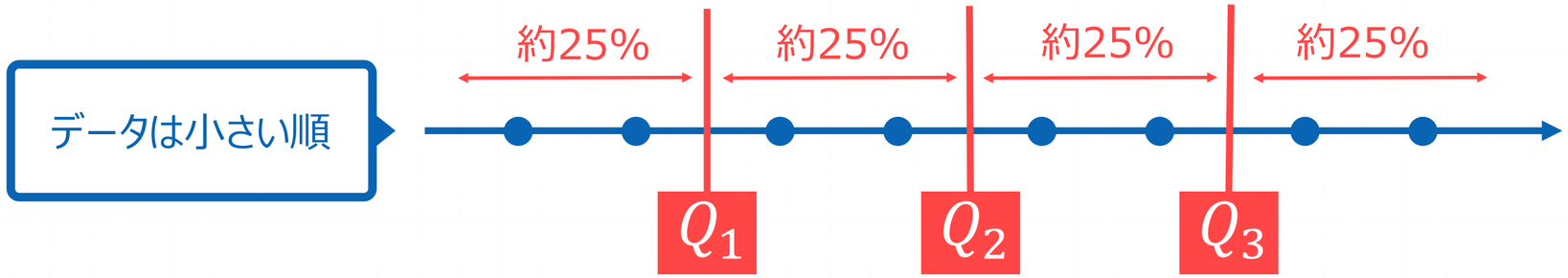
データを小さい順に並べたとき,

第1四分位点 Q_1: 真ん中より値が小さいデータの中央値

第2四分位点 Q_2: 中央値

第3四分位点 Q_3: 真ん中より値が大きいデータの中央値

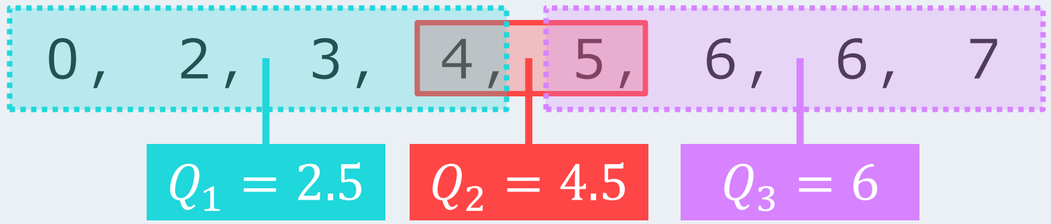
データを4分割したときの境界が四分位点 (だいたい)



さっそく, 計算例を見て
理解を深めていきましょう

四分位点

例 **データが偶数個のとき**

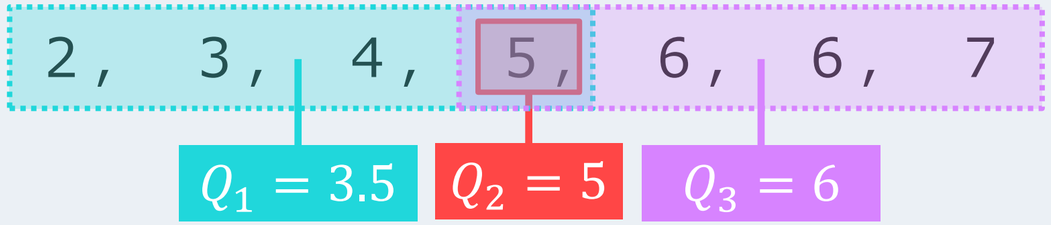


Q_1 は 0,2,3,4 の中央値

Q_3 は 5,6,6,7 の中央値

例 **データが奇数個のとき**

計算方法1
(**ヒンジ法**)

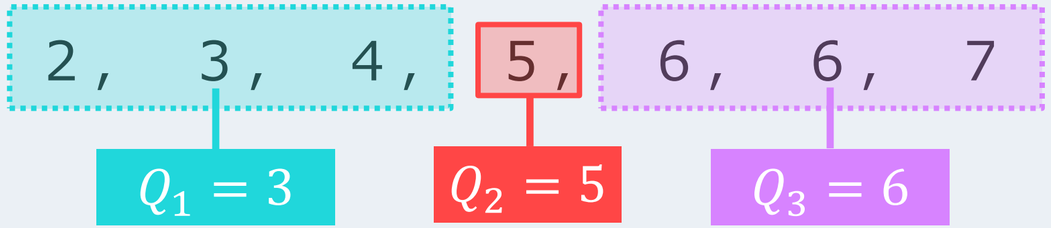


Q_1 は 2,3,4,5 の中央値

Q_3 は 5,6,6,7 の中央値

(**Q_2 を含める場合**)

計算方法2



Q_1 は 3,4,5 の中央値

Q_3 は 6,6,7 の中央値

(**Q_2 を含めない場合**)

四分位範囲

四分位範囲 (InterQuartile Range; IQR)

変数 x の n 個のデータ $x_1, \ldots, x_n$ の四分位範囲:

四分位範囲 ＝ (第3四分位点 Q_3) − (第1四分位点 Q_1)

四分位範囲 ＝ **真ん中のデータ (全データの約50%)** が含まれる区間の大きさ

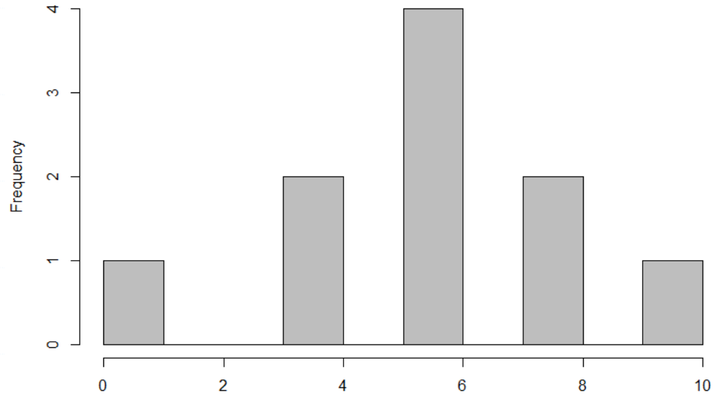
中央値周りの散らばり具合を表す

範囲よりも外れ値の影響を受けにくい

中央値と四分位範囲の計算はセットで

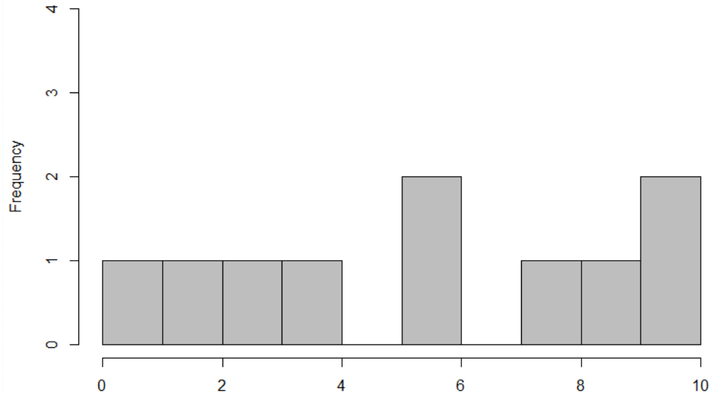
分布の散らばり

小テストA



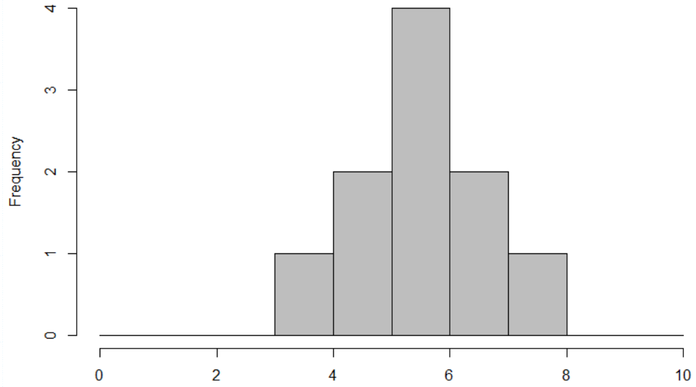
範囲: 10
IQR: 4

小テストB



範囲: 10
IQR: 6

小テストC



範囲: 4
IQR: 2

例題

表. サービス満足度のアンケート

(1:とても不満, 2:少し不満, 3:ふつう, 4:少し満足, 5:とても満足)

	サービスA	サービスB
顧客1	3	5
顧客2	4	2
顧客3	4	3
顧客4	5	4
顧客5	5	4
顧客6	1	5
顧客7	1	4
顧客8	5	3
顧客9	4	4
顧客10	4	3

Q どっちのサービスが安定して満足度が高い?

平均値… A: 3.6,　B: 3.7

分散…… A: 2.04, B: 0.81

~~だからサービスBのほう!!~~

! 順序変数だから平均値, 分散は適さない

練習問題 (3)

表. サービス満足度のアンケート

(1:とても不満, 2:少し不満, 3:ふつう, 4:少し満足, 5:とても満足)

	サービスA	サービスB
顧客1	3	5
顧客2	4	2
顧客3	4	3
顧客4	5	4
顧客5	5	4
顧客6	1	5
顧客7	1	4
顧客8	5	3
顧客9	4	4
顧客10	4	3

Q 中央値, IQR を計算してどっちのサービスが安定して満足度が高いか調べよ

この講義で紹介した方法で計算してみてください

解答: 練習問題 (3)

サービスA

1, 1, 3, 4, 4, 4, 4, 5, 5, 5

中央値

四分位点

IQR

サービスB

2, 3, 3, 3, 4, 4, 4, 4, 5, 5

中央値

四分位点

IQR

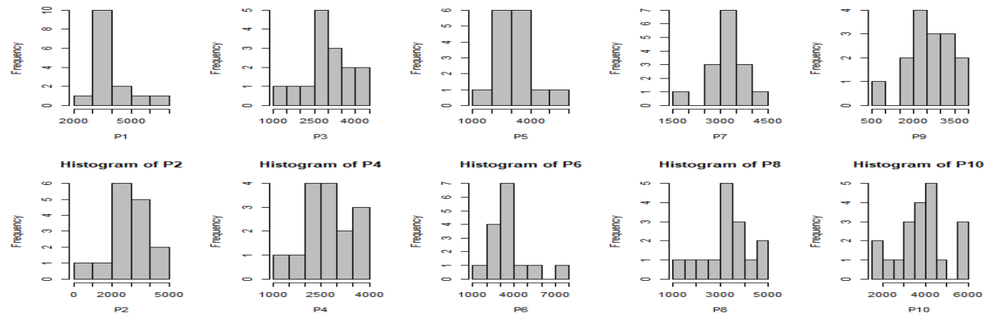
A

箱ひげ図

グラフ + 要約統計量

10地区の住宅価格の特徴を教えて

10地区のヒストグラムはこんなかんじで，
地区Aの平均値は~~~，分散は~~~，
中央値は~~~，四分位範囲は~~~，地区Bの…



部下

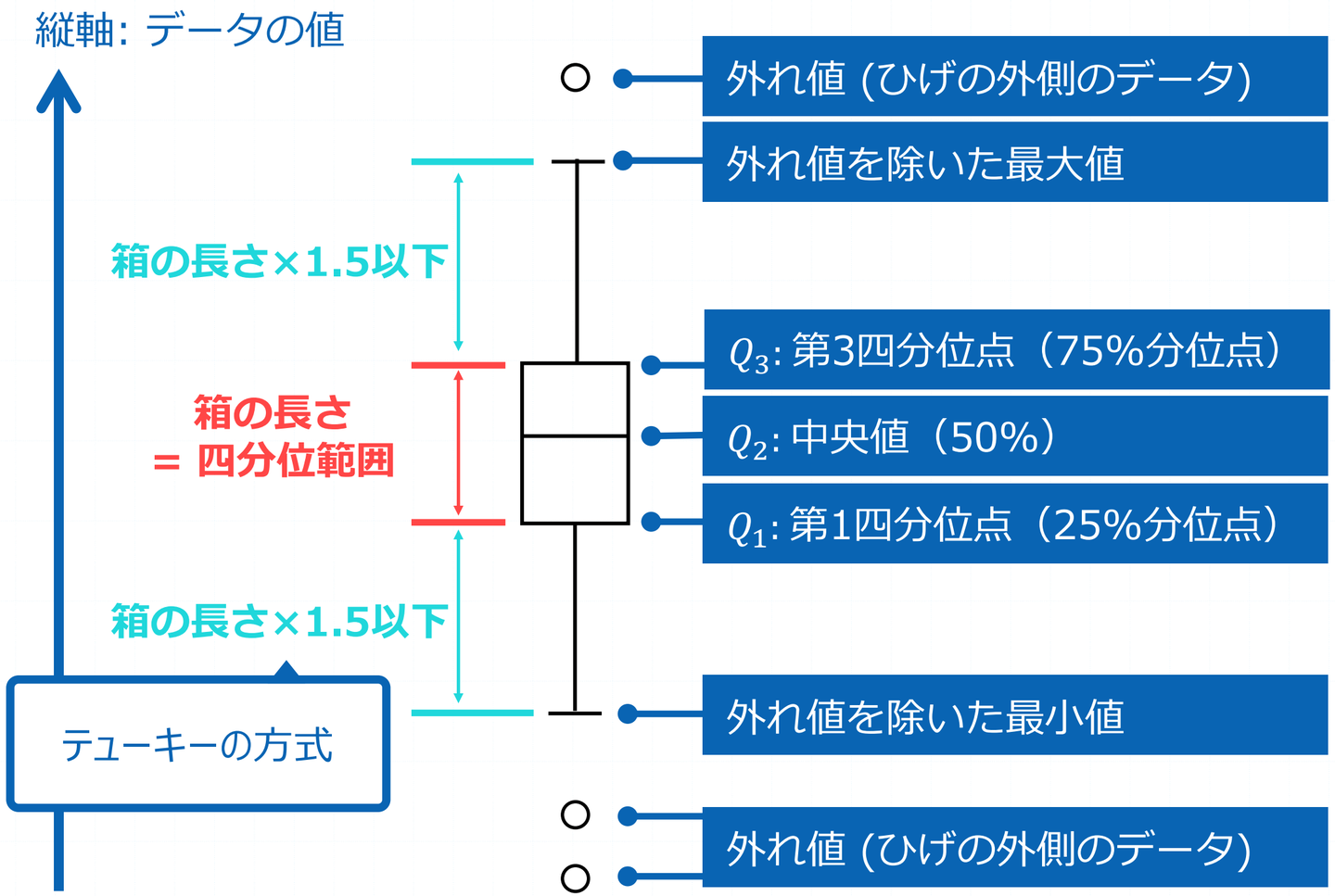
グラフが小さいしたくさんあって見にくい．
数値もいいけどもっとわかりやすく教えて

不動産会社の上司

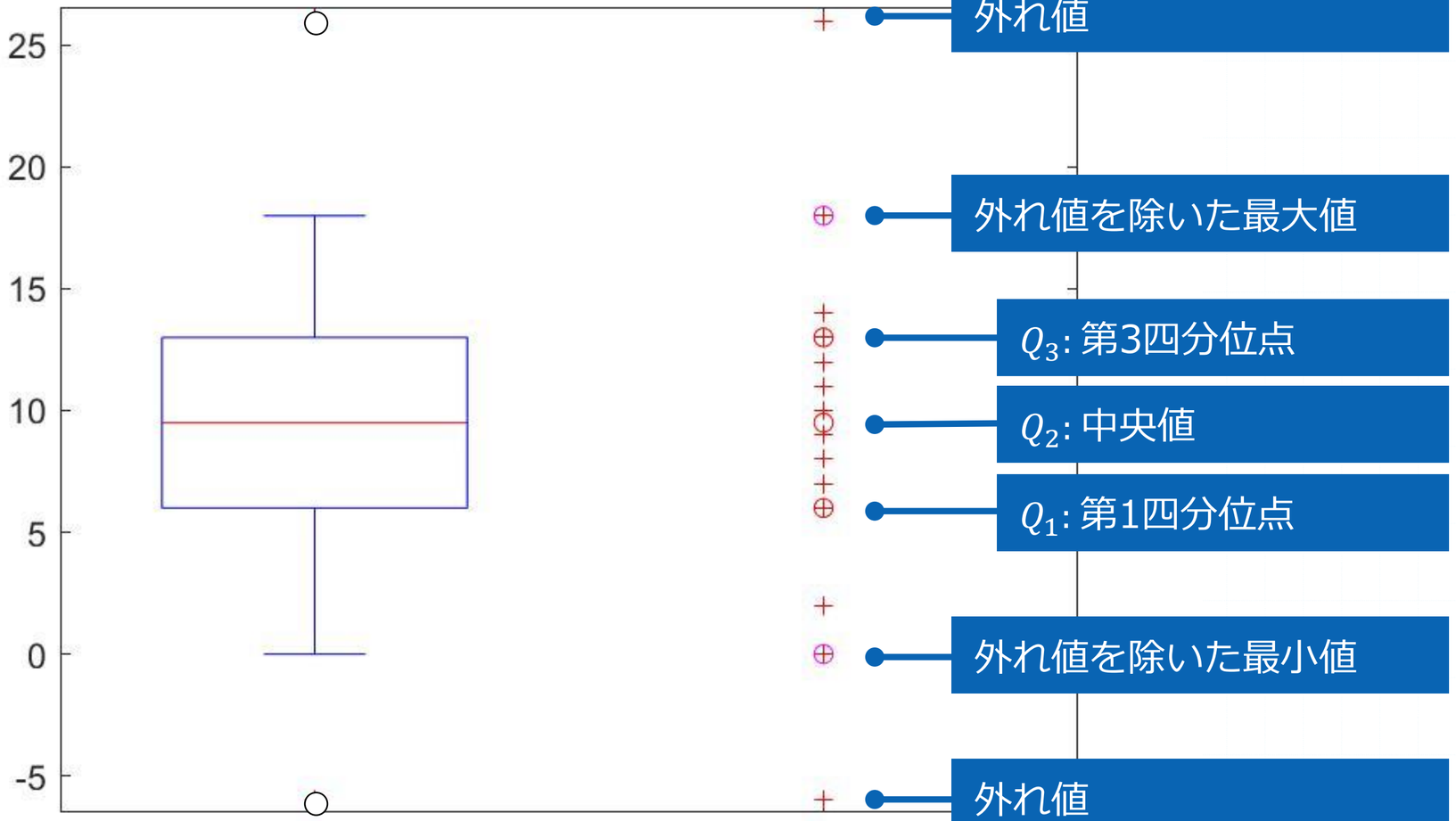
箱ひげ図

箱ひげ図
データの要点を
簡潔に表したグラフ

箱の長さ
＝ 散らばり具合



箱ひげ図



箱ひげ図

データ

外れ値

外れ値を除いた最大値

Q_3: 第3四分位点

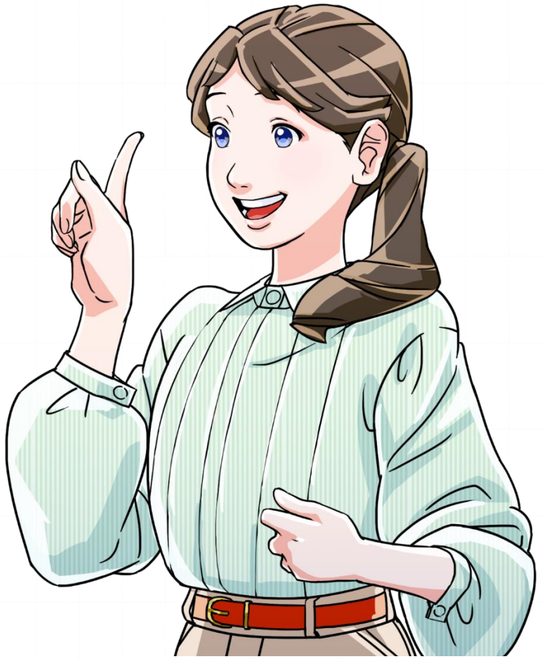
Q_2: 中央値

Q_1: 第1四分位点

外れ値を除いた最小値

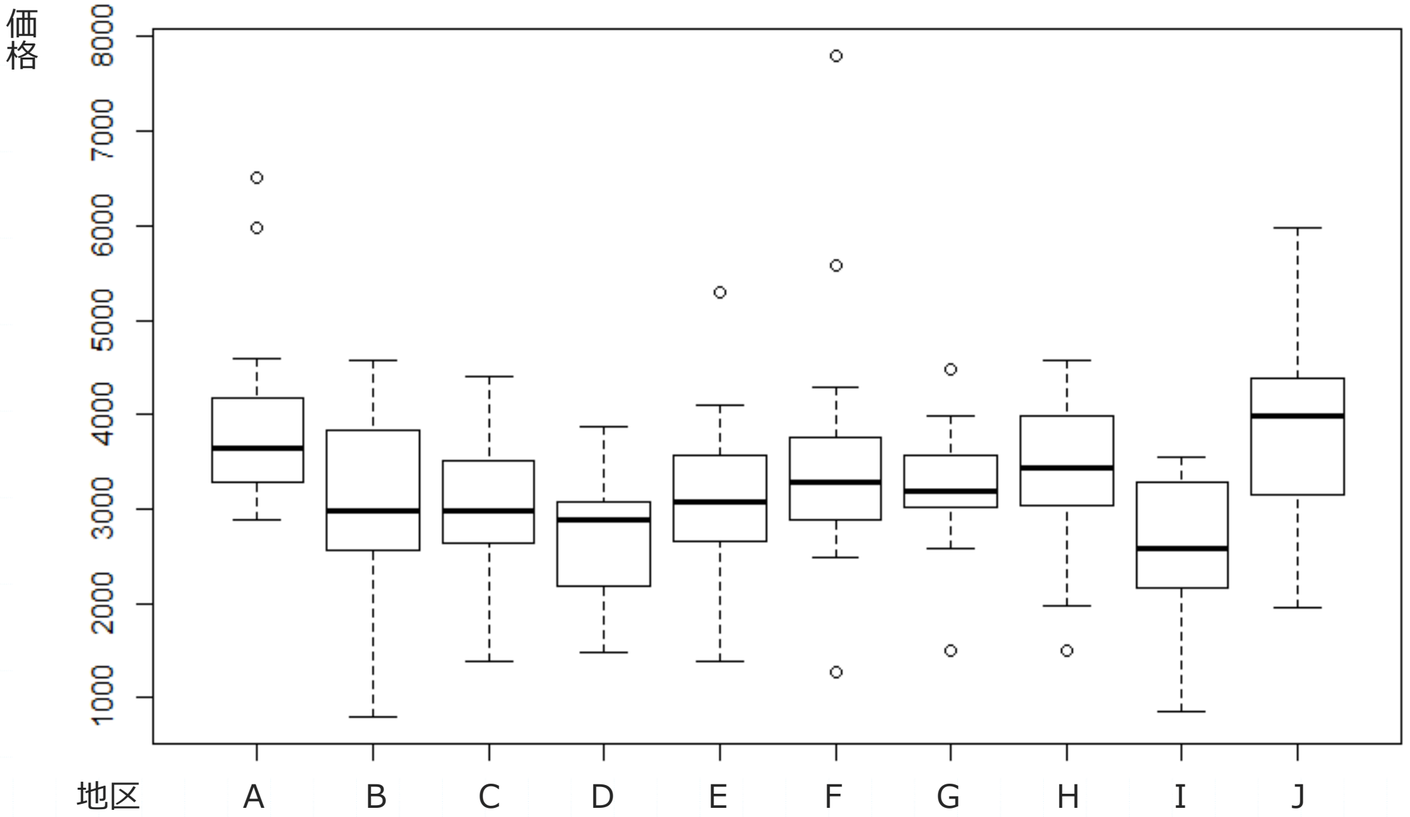
外れ値

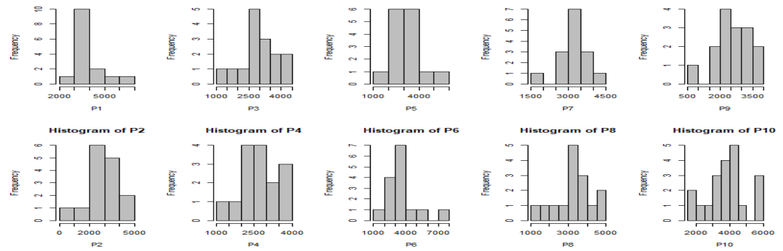
＋マークが
データを表していますね



箱ひげ図

複数グループを比較するとき箱ひげ図が便利





散布図

データの次元

1次元データ

1つの変数についてのデータ

例 学生の身長 (変数 x)

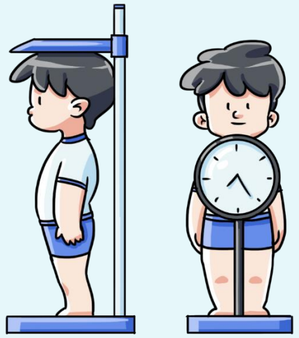
グラフ: ヒストグラム, 箱ひげ図など

個体	変数 x
1	x_1
$\vdots$	$\vdots$
n	x_n

2次元データ

2つの変数についてのデータ

例 学生の身長 (変数 x) と体重 (変数 y)



個体	変数 x	変数 y
1	x_1	y_1
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

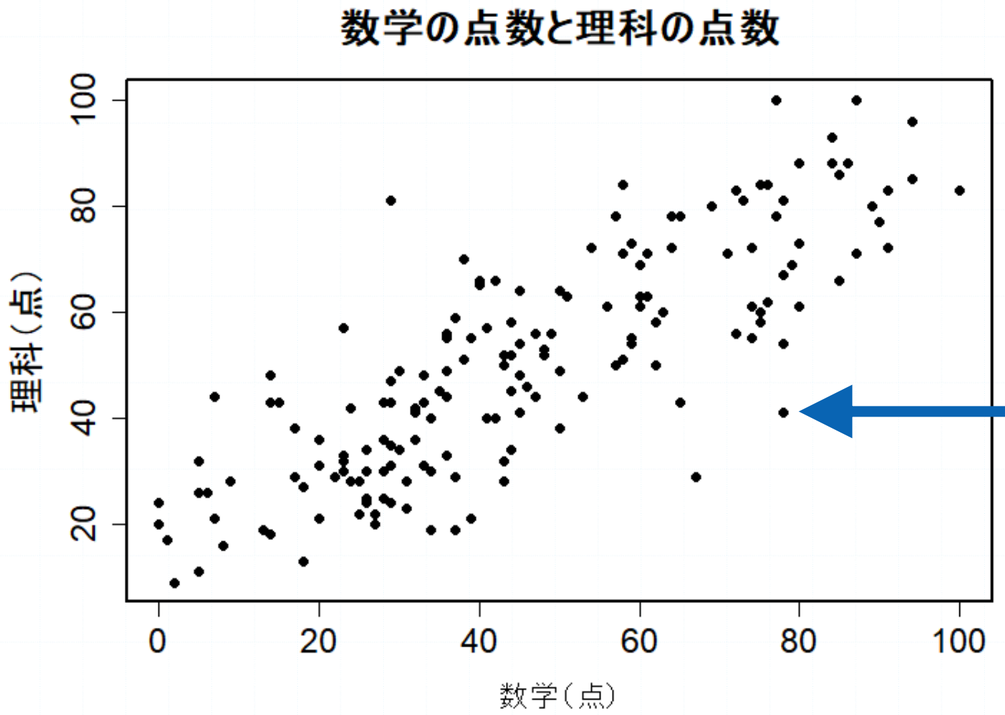
データの組: $(x_1, y_1), \ldots, (x_n, y_n)$

散布図

散布図

2つの変数の関係性を可視化したグラフ

例 数学の点数を横軸，理科の点数を縦軸にとってデータを平面上の点として表す



数学80点くらい，
理科40点くらいの人

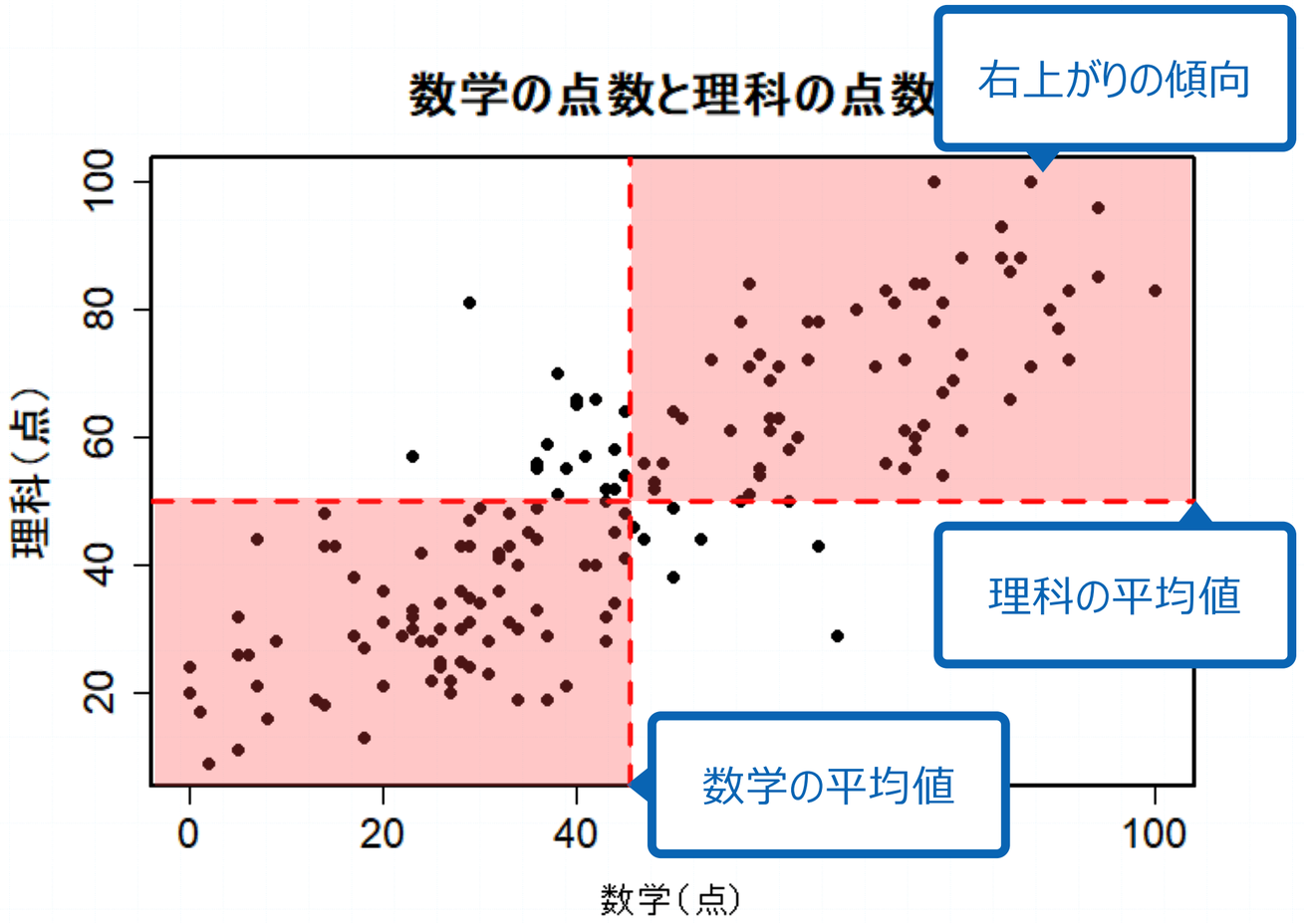
散布図の見方

直線的な関係性をみる方法

❶ 2つの変数の平均値の直線を
引いて4つの区画にわける

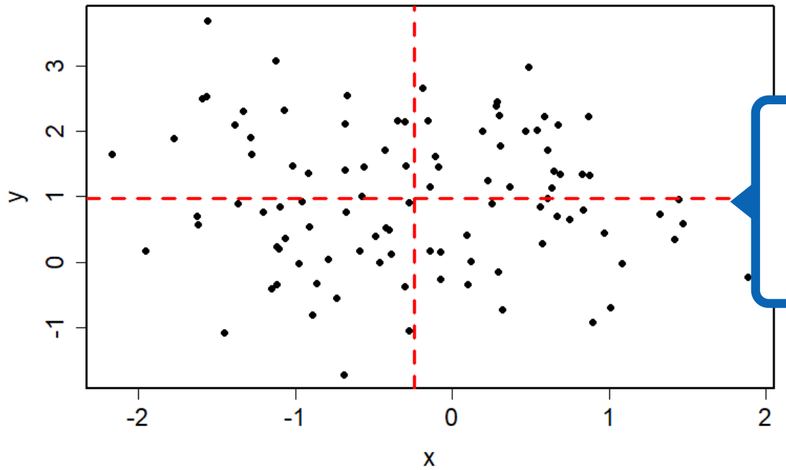
❷ データ点が**右上, 左下**に多い
→ **右上がりの (直線) 傾向**

右下, 左上に多い
→ **右下がりの (直線) 傾向**

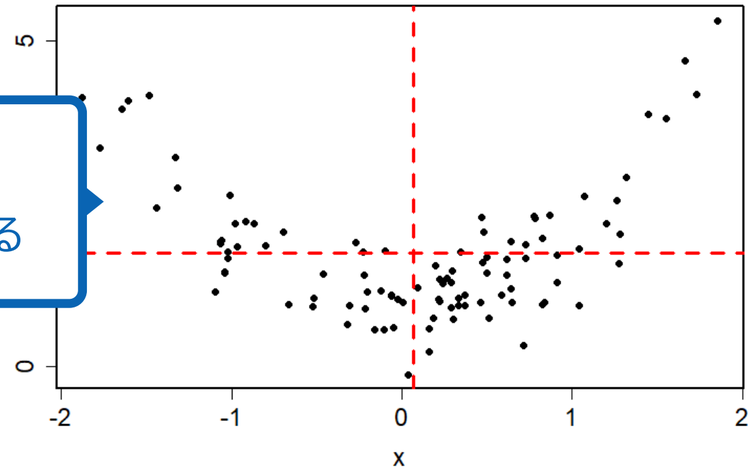


散布図の見方

直線的**でない**関係性も目でみて判断できる



数値を眺めても
わからないことを知れる



4つの区画に散らばっている
→ 右上がりでも右下がりでもない
→ **直線的な関係はなさそう**

4つの区画に散らばっている
→ 右上がりでも右下がりでもない
→ **そもそも曲線??**

2つの変数の関係性をみるにはまずは散布図を描こう!!

今日のまとめ

次回も一緒に
頑張りましょう!

- **要約統計量**

  分布の位置: 平均値, 中央値, 最頻値, 四分位点

  分布の散らばり: 分散, 標準偏差, 範囲, 四分位範囲

- **グラフ**

  ヒストグラム, **箱ひげ図** (1次元データ)

  散布図 (2次元データ)

要約統計量とグラフで分布の特徴を知る・伝える・比較する