

# ビジネスに役立つデータ分析 （入門編）

瓜生真也（徳島大学デザイン型AI教育研究センター）

# 諸注意

## 資料置き場: [https://github.com/uribo/cue2022aw\\_r104](https://github.com/uribo/cue2022aw_r104)

投影するプレゼンテーション、ソースコードを置いています  
(来週分は来週更新)

## Rコードと実行環境

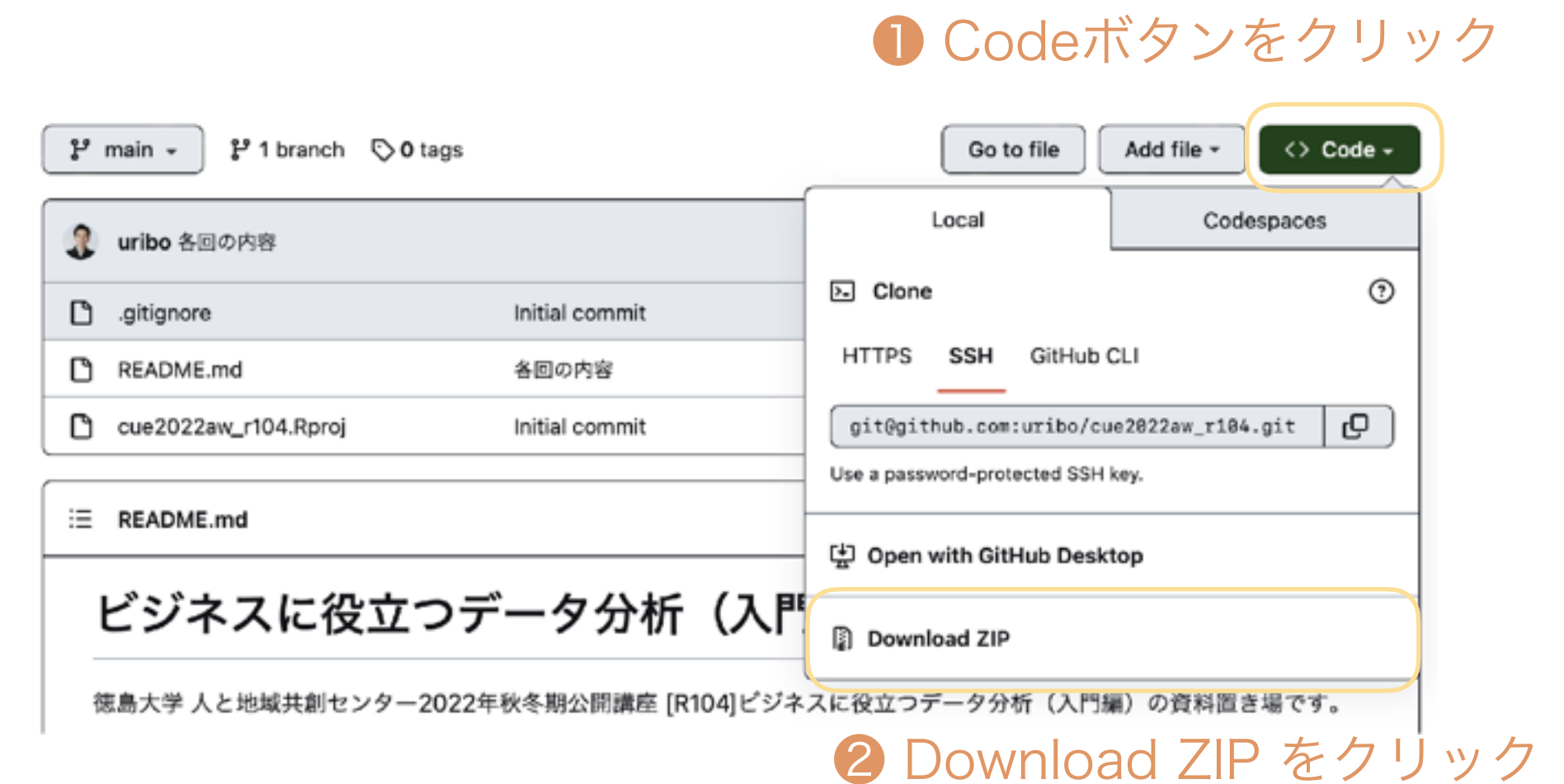
```
# 気温とアイスの相関係数を求める
cor(df_icecream_temperature$temperature_average_c,
    df_icecream_temperature$value)
#> [1] 0.9144466
```



Rでの実行コマンドであることを示します

[https://mybinder.org/v2/gh/uribo/cue2022aw\\_r104/main?urlpath=rstudio](https://mybinder.org/v2/gh/uribo/cue2022aw_r104/main?urlpath=rstudio)

から必要なパッケージ、データ、ソースコードを含んだRStudioが起動します



# 第二週: データを扱うためのリテラシー

瓜生真也（徳島大学デザイン型AI教育研究センター）

# 講座の内容

## 第一週

情報化社会におけるデータの利活用

## 第二週

データを扱うためのリテラシー

データの種類と表現方法

データの要約、 代表値

グラフの作り方、読み方

## 第三週

データサイエンス入門 (1)

## 第四週

データサイエンス入門 (2)

## 第五週

ツールの使い方

振り返り

# データの種類と表現方法

# データの種類: さまざまな変数

データを性質に応じていずれかの尺度水準に分類する

尺度水準に応じて、取り扱い方や用いる分析・表現手法が異なる

例) 名義尺度間での算術演算はできない、間隔尺度と比例尺度では統計量の利用ができる

変数の種類	尺度水準	判断の基準	例	
質的変数	名義尺度	対象が他とは異なるか同一か	性別、出身地	水準の高さ ↓
質的変数	順序尺度	対象が他より「大きい」、他より「良い」など	健康度、利便性	
量的変数	間隔尺度	対象は他よりもある単位によって～だけ多い（少ない）	温度、時刻、偏差値	
量的変数	比例尺度	対象は他よりある単位によって～倍だけ多い（少ない）	身長、絶対温度、年齢	

高い水準の尺度を、より低い水準の尺度に変換できる。  
例えば名義尺度である性別（「男」「女」と表現）を「男」= 0、「女」= 1のように

# 問題: 次のデータの各変数はどの尺度水準に分類されるか

## 四国4件の婚姻件数と婚姻率（2019年度）

年度	都道府県	人口順位	婚姻件数	婚姻率*
2019	徳島県	3	2,878	3.95
2019	香川県	2	4,237	4.43
2019	愛媛県	1	5,360	4.00
2019	高知県	4	2,630	3.77

婚姻率は人口千人当たりの婚姻件数と定義する

# データに潜む問題

データ分析で扱うデータにはさまざまな課題が含まれる

**欠損値** さまざまな理由により観測・測定されなかったデータを指す

問題: 欠損値を処理しないと統計的計算処理が不可能な場合がある… PCAなど

対処: 削除または補完による対処が求められる

**外れ値・異常値** 他の観測データに比して著しく乖離したデータ

問題: データ本来の性質とは異なる結果が導かれる可能性がある

対処: 外れ値を検出し、統計的アプローチなどを適用する

**誤差** データの観測・測定に伴う変動

同一条件下での測定においても高精度での繰り返しの記録には誤差が伴う

**個々の観測値 = 正確な値 + 誤差**

例) 同一個体の動物の体重を測定… 6.781kg, 6.789kg, 6.780kg 🐹



# データの特徴を捉える(1): 記述統計量と分布の可視化



# データの特徴を伝えるには？

データ分析で扱うデータは膨大（数百～数十万件）

これらのデータの内容を整理し、簡潔に伝えることが求められる



写真の魚の体長は？

人間が処理できる数値の数には限りがある

## データ分析の手法

### 代表値によるデータの集約

最小値・最大値

平均値

位置を伝える

### ばらつきの指標の計算による分布の推定

分散

標準偏差

範囲を伝える

### データ可視化

箱ヒゲ図

ヒストグラム

視覚的に伝える



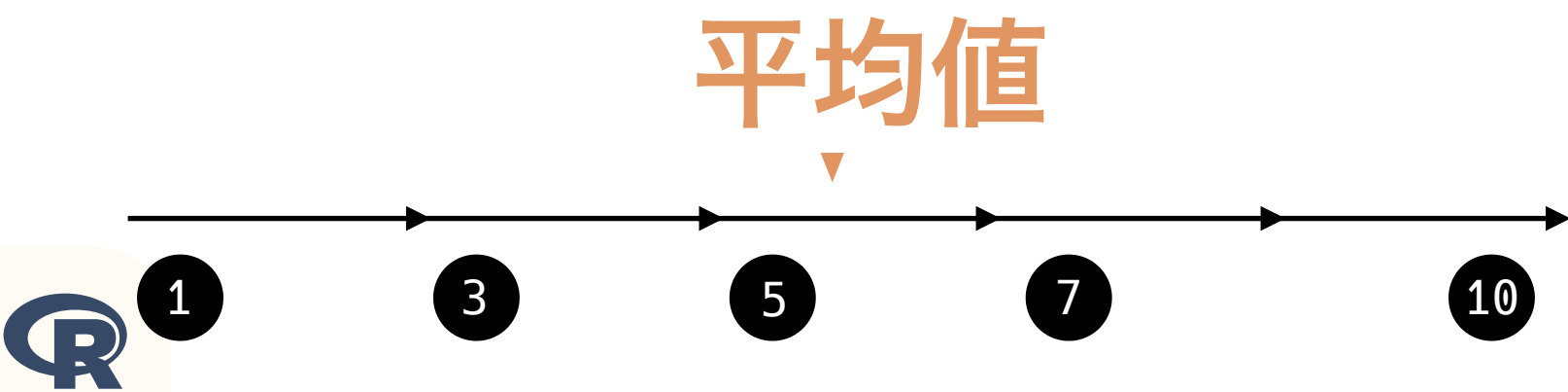
# 代表値の算出

## 平均値

データに含まれる値をすべて足し合わせて、データの数で割った値

平均値は必ずしもデータの真ん中を示す値ではない  
平均値は外れ値の影響を受けやすい

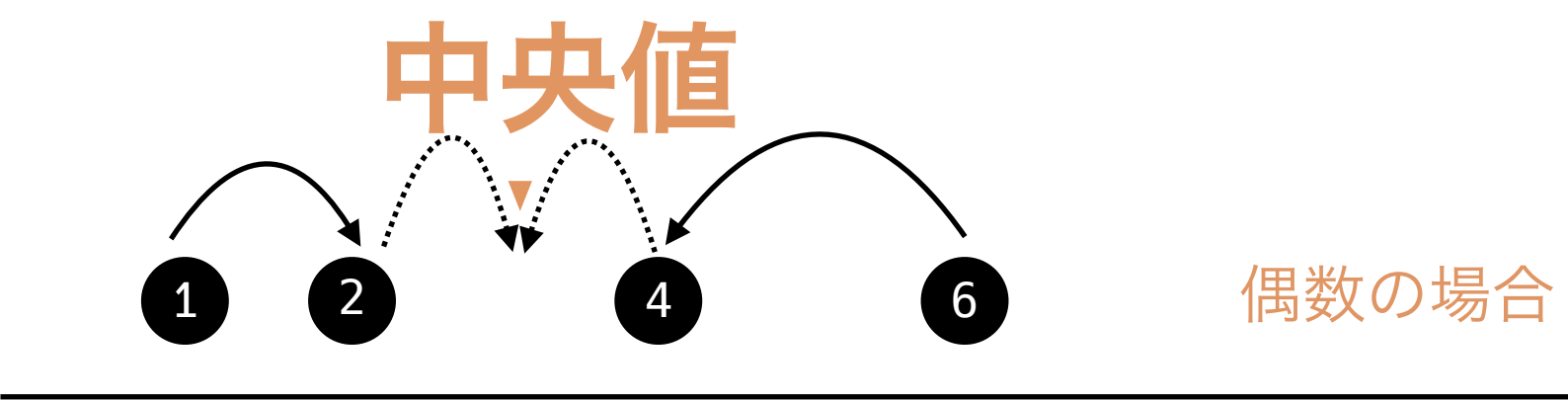
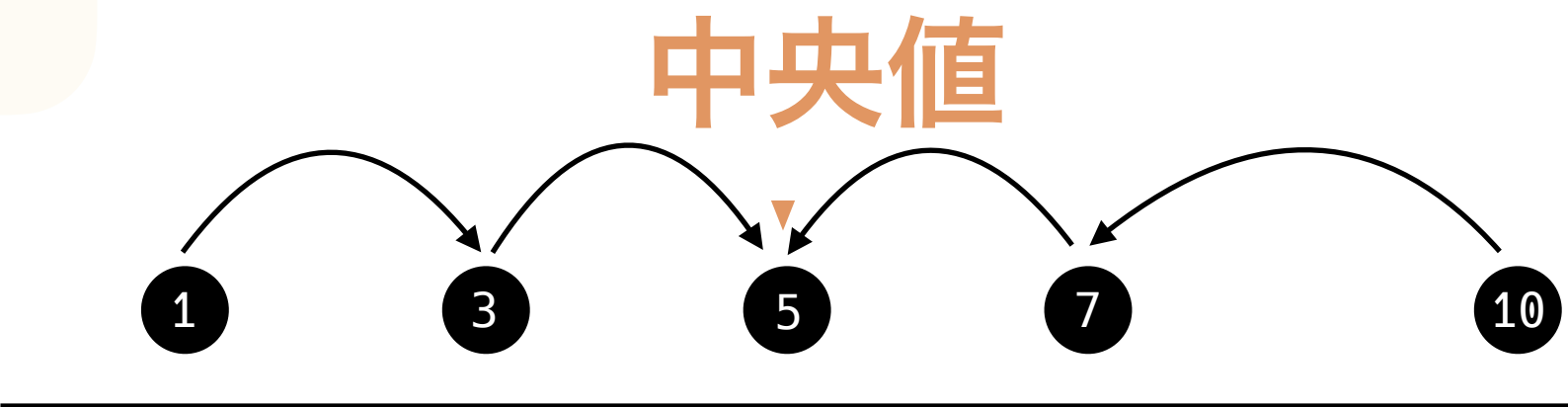
```
x <- c(1, 10, 5, 3, 7)
sum(x) / length(x)
#> [1] 5.2
# mean( )関数を用いて平均値を計算
mean(x)
#> [1] 5.2
```



## 中央値

データに含まれる数の真ん中となる値

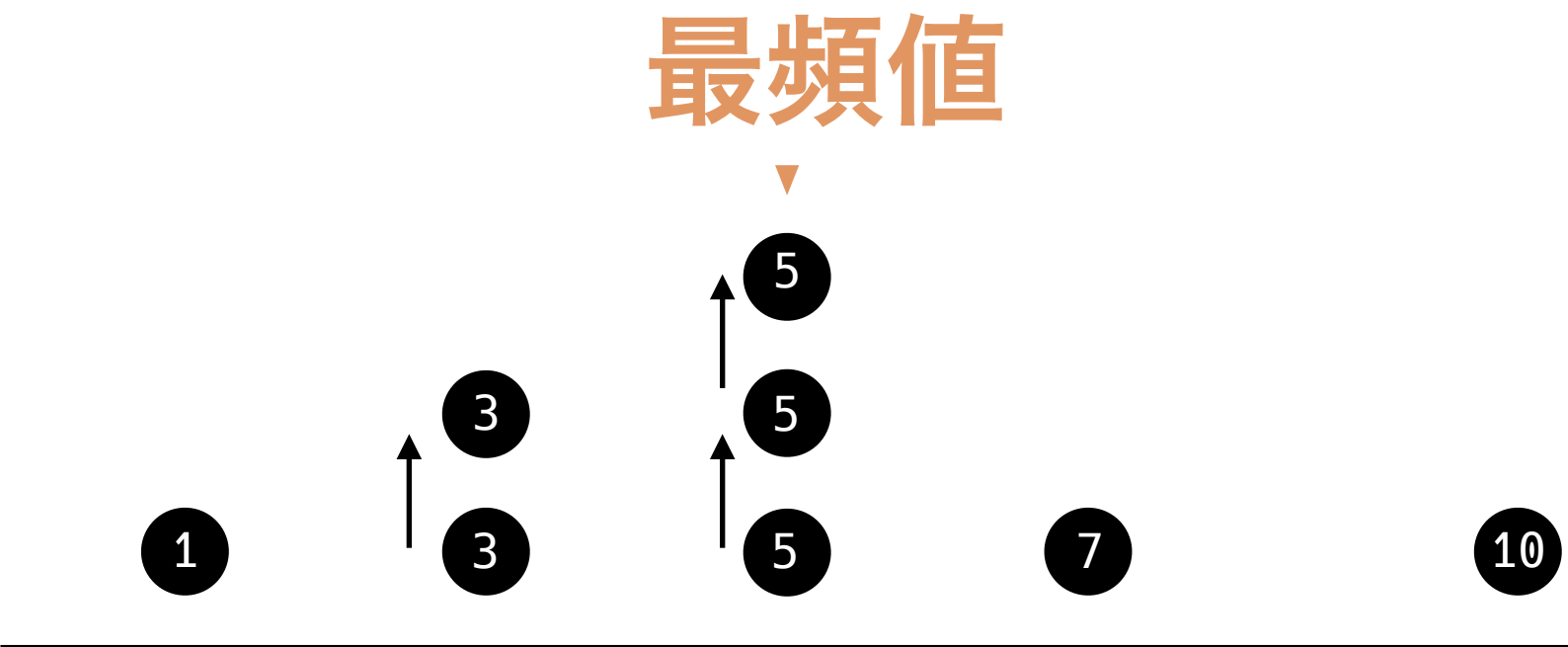
```
sort(x)[ceiling(length(x)/2)]
#> [1] 5
# median( )関数で数値ベクトルの中央値を計算
median(x)
#> [1] 5
```



## 最頻値

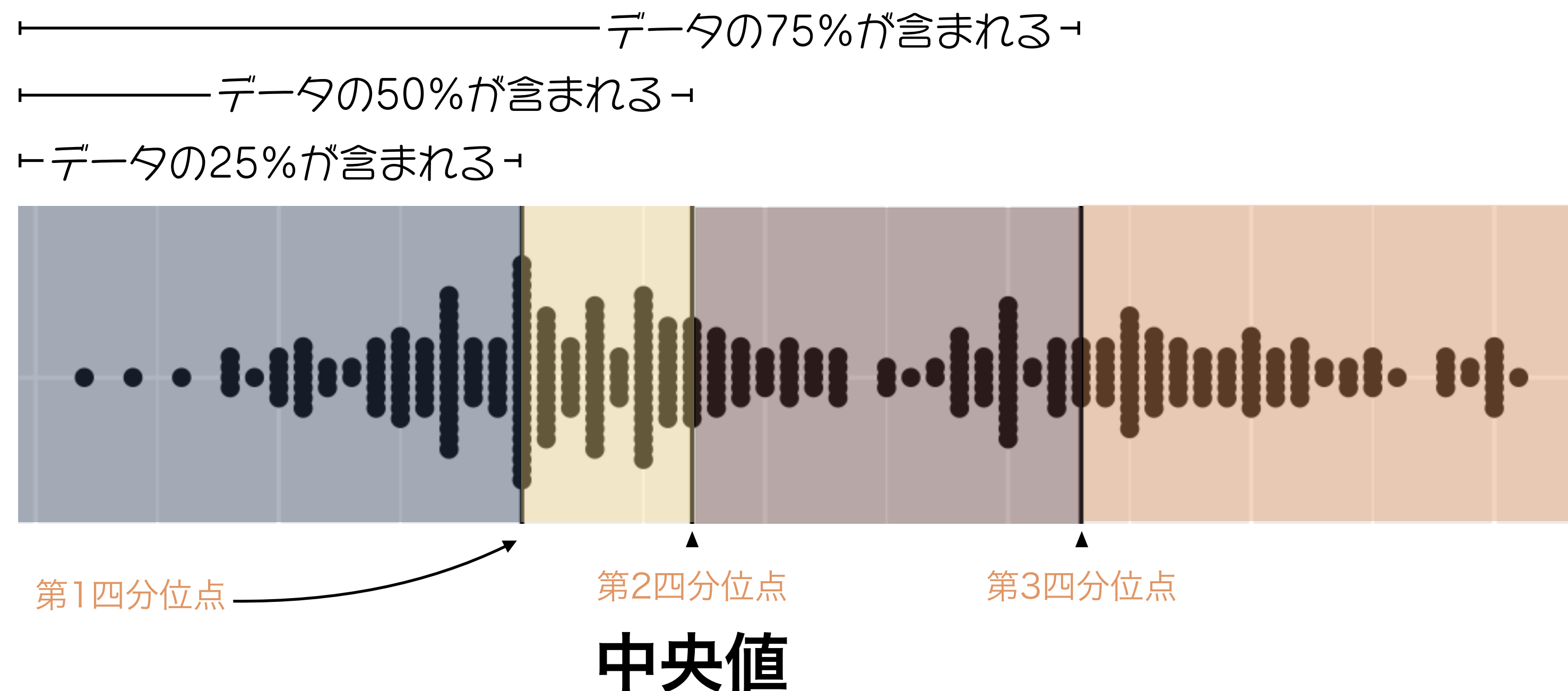
データに含まれる値の中で最も多い値


```
x <- c(1, 3, 3, 5, 5, 5, 7, 10)
as.numeric(names(which.max(table(x))))
#> [1] 5
```



# 中央値を拡張した考え方: 四分位点

データを値の小さい順に並び替えたとき、  
データ全体を均等な数からなる4つのグループに分ける  
このときのグループを分ける3つの点（値）を四分位点という



```
quantile(penguins$flipper_length_mm, na.rm = TRUE)   
#>      0%    25%    50%    75%   100%  
#>   172   190   197   213   231
```

# データのばらつき2: 分散 variance

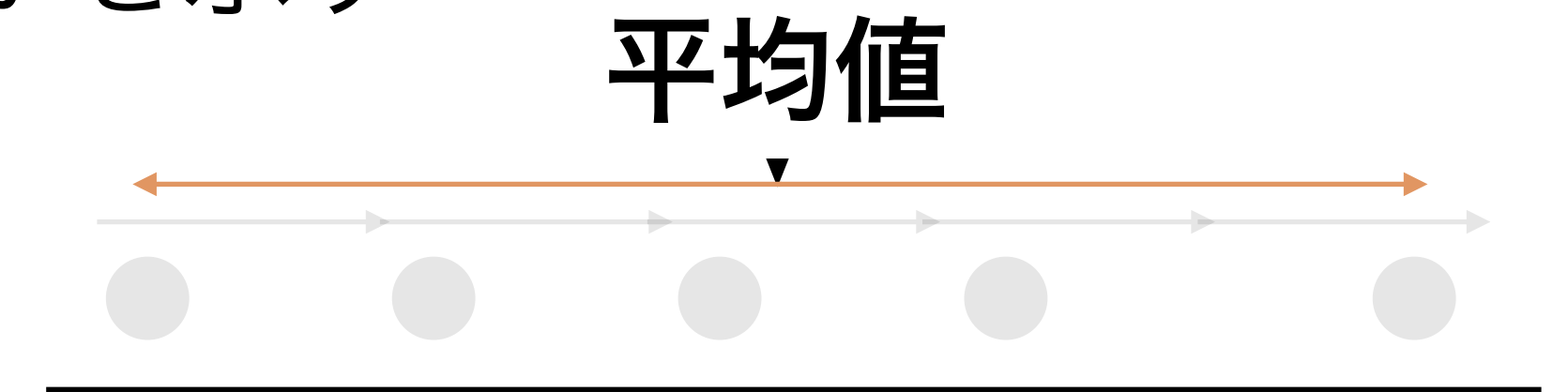
各値が平均値を中心としてどのように散らばっているかを示す

例) ペンギンの各個体の体長について

全般的に均一な値?

特定の個体が平均値よりも特段高い・低い?

体長が高い個体と低いバラバラ?



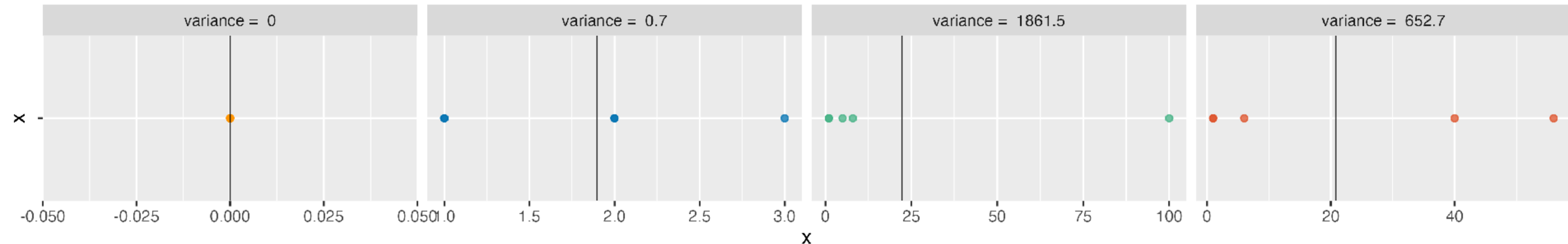
データの分布について具体的な説明ができるようになる

`c(0, 0, 0, 0, 0)`

`c(1, 2, 3, 2, 1)`

`c(1, 100, 5, 8, 1)`

`c(1, 6, 40, 56, 1)`



縦棒は平均値を示す

# 分散の求め方

$$\text{分散} = \frac{\text{変数の値と平均値の差の2乗の合計}}{\text{変数に含まれるデータ数}} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

1. 変数の平均値を出す
2. 変数の各値と平均値の差を求める（偏差）
3. 偏差を二乗する
4. **すべての値に対して1から3を繰り返し、合計する**
5. 合計した値をデータの数で割る

# 問題: 動物データから体長と体重の分散を求めよ

```
library(readr)
library(dplyr)

# 動物データを読み込む
animals <- read_csv("animals.csv")

# 体長と体重の平均と分散を計算する
summary(animals)

# 体長の平均と分散
mean_length <- mean(animals$body_length_cm)
var_length <- var(animals$body_length_cm)

# 体重の平均と分散
mean_weight <- mean(animals$body_weight_g)
var_weight <- var(animals$body_weight_g)
```

平均を求める関数は `mean()`

合計を求める関数は `sum()`

欠損値を除いた処理が必要（引数 `na.rm = TRUE` の指定）

階乗は `^` で指定（例. 2の二乗は `2^2` とする）

Rでの分散を求める関数は「不偏分散」の値を算出する

# ばらつきの程度を示す：標準偏差


分散の問題… 分散は比較ができるが、分散と平均を足したり、平均と比較することはできない

二乗しているため、単位が変わってしまう。単位が異なると解釈が難しい

例) 体長 (cm)の分散 …  $\text{cm}^2$

→分散に対して平方根を求める (**標準偏差**)

二乗された単位が元に戻り、比較が可能となる

```
sqrt(var(df_animal$body_length_cm, na.rm = TRUE))   
#> [1] 70.69994  
# Rの標準偏差を求める関数sd( )は不偏標準偏差として扱う  
sd(df_animal$body_length_cm, na.rm = TRUE)  
#> [1] 70.69994
```



# 度数分布表

ある値がデータに含まれる数… 度数または頻度 (frequency)

度数の分布を表形式にまとめたもの… **度数分布表**

動物データの分類群ごとの度数分布表

```
# 動物データの分類群ごとの頻度を数える
table(df_animal$taxon)
```

```
#>
#>   偶蹄類   奇蹄類   霊長類   食肉類  鯨偶蹄類   鳥類   齧歯類
#>      1      1      4      7      2      5      2
```

分類群	頻度
偶蹄類	1
奇蹄類	1
霊長類	4
食肉類	7
鯨偶蹄類	2
鳥類	5
齧歯類	2

# 度数分布表

量的変数に対して度数分布表を作成するときは

変数がとり得る値をいくつかの区間に分割した階級(class)を考える

```
body_length_freq <-  
  # 動物データの体長を40cm間隔の階級に分けて頻度を数える  
  table(cut(df_animal$body_length_cm,  
            breaks = seq(20,  
                          200,  
                          by = 40)))  
  
tibble(  
  class = names(body_length_freq),  
  frequency = body_length_freq)
```



動物データの体長の度数分布表（階級幅40）

階級	頻度
~60	3
60~100	7
100~140	3
140~180	2

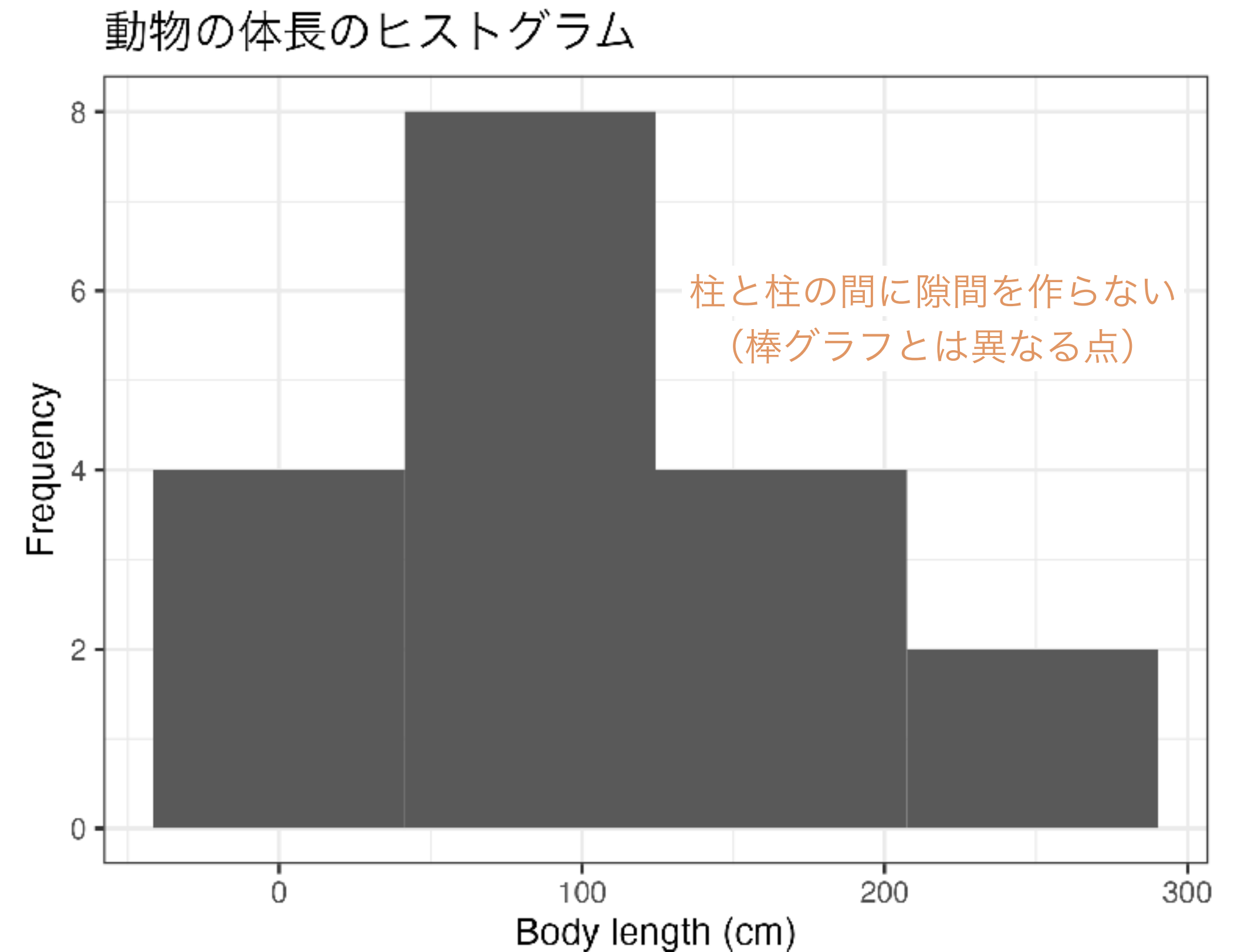
各度数に含まれる区間の幅を階級幅という

階級幅や階級数はデータの範囲を見て決める

# 度数分布表をもとにグラフを作成: ヒストグラム

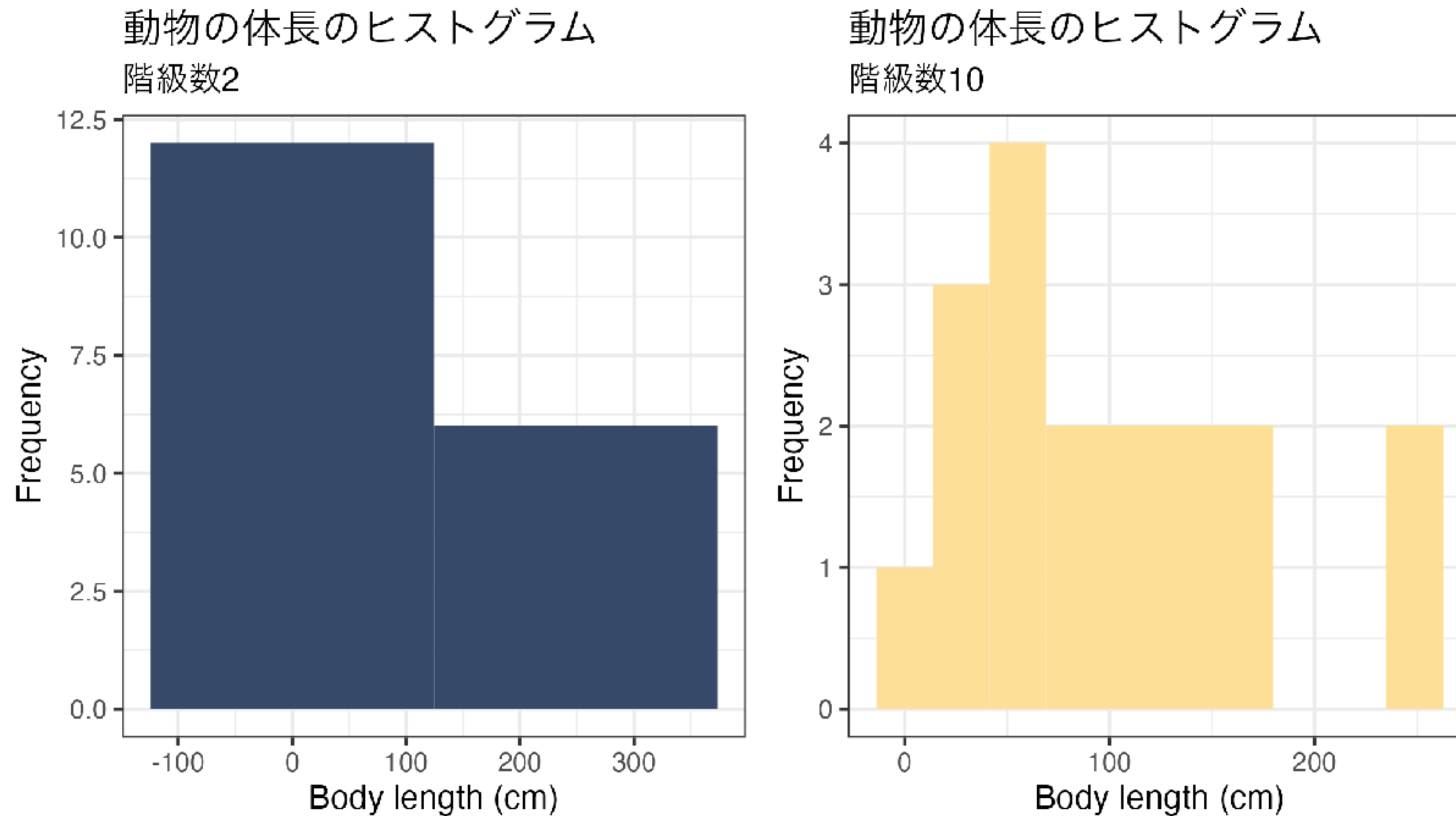
階級ごとに柱を設け、柱の高さで度数を表現

```
df_animal |>  
  ggplot(aes(body_length_cm)) +  
  # ヒストグラムでは柱の階級をビン bin と呼びます  
  geom_histogram(bins = 4) +  
  ylab("Frequency") +  
  xlab("Body length (cm)") +  
  labs(title = "動物の体長のヒストグラム")
```



# ヒストグラムの形からデータの分布を見る

ヒストグラムの階級数が異なると分布の形も変化する



データのばらつきに応じてヒストグラムの形も異なる… 分布が異なる

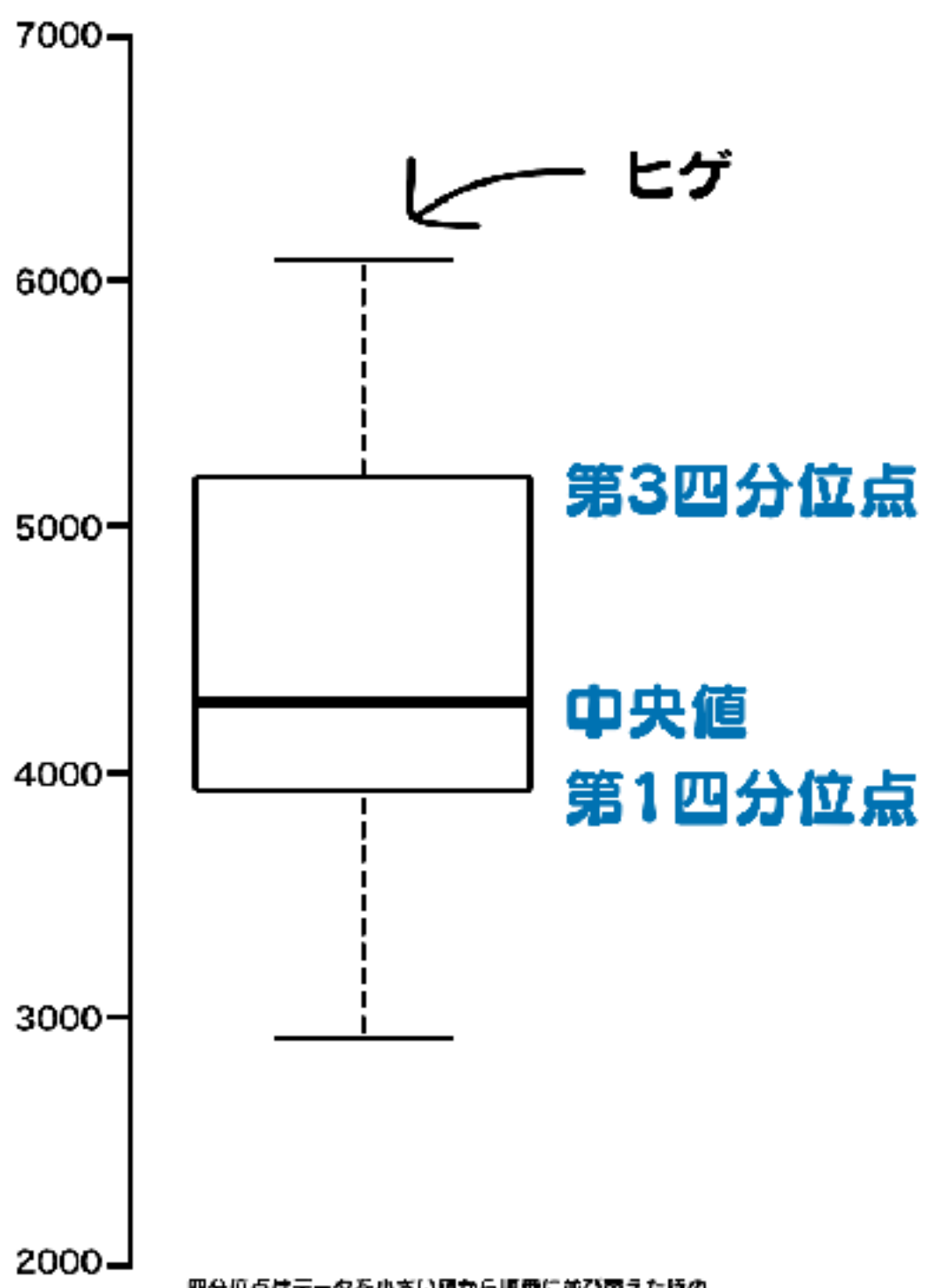
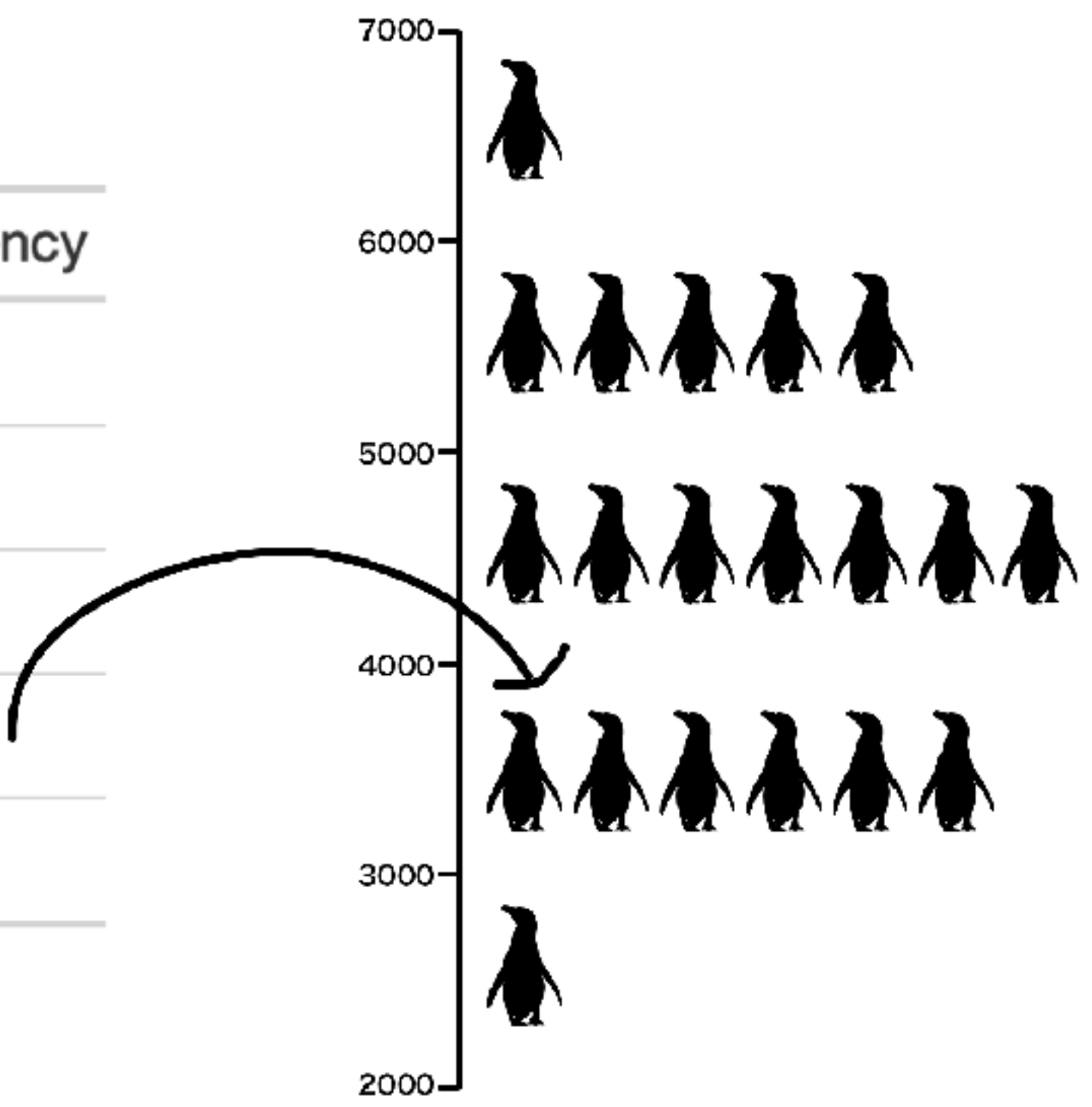
# 箱ヒゲ図

「箱」と「ヒゲ」を使ってデータの分布を表現するグラフ  
四分位点、外れ値の情報も可視化することができる

## 箱ヒゲ図の見方と作り方

- ①度数分布表からヒストグラムを作る
- ②四分位点を求める
- ③四分位範囲で箱を作り、箱の中に中央値の線を引く。  
最大値と最小値の位置までヒゲを描く

class	frequency
(6000,7000]	1
(5000,6000]	5
(4000,5000]	7
(3000,4000]	6
(2000,3000]	1



四分位点はデータを小さい順から順番に並び替えた時のデータ全体を同等数値の4つのグループに分ける3点（位置）。第1四分位点から第3四分位点の範囲を四分位範囲といい、データの50%が含まれる区間を示す

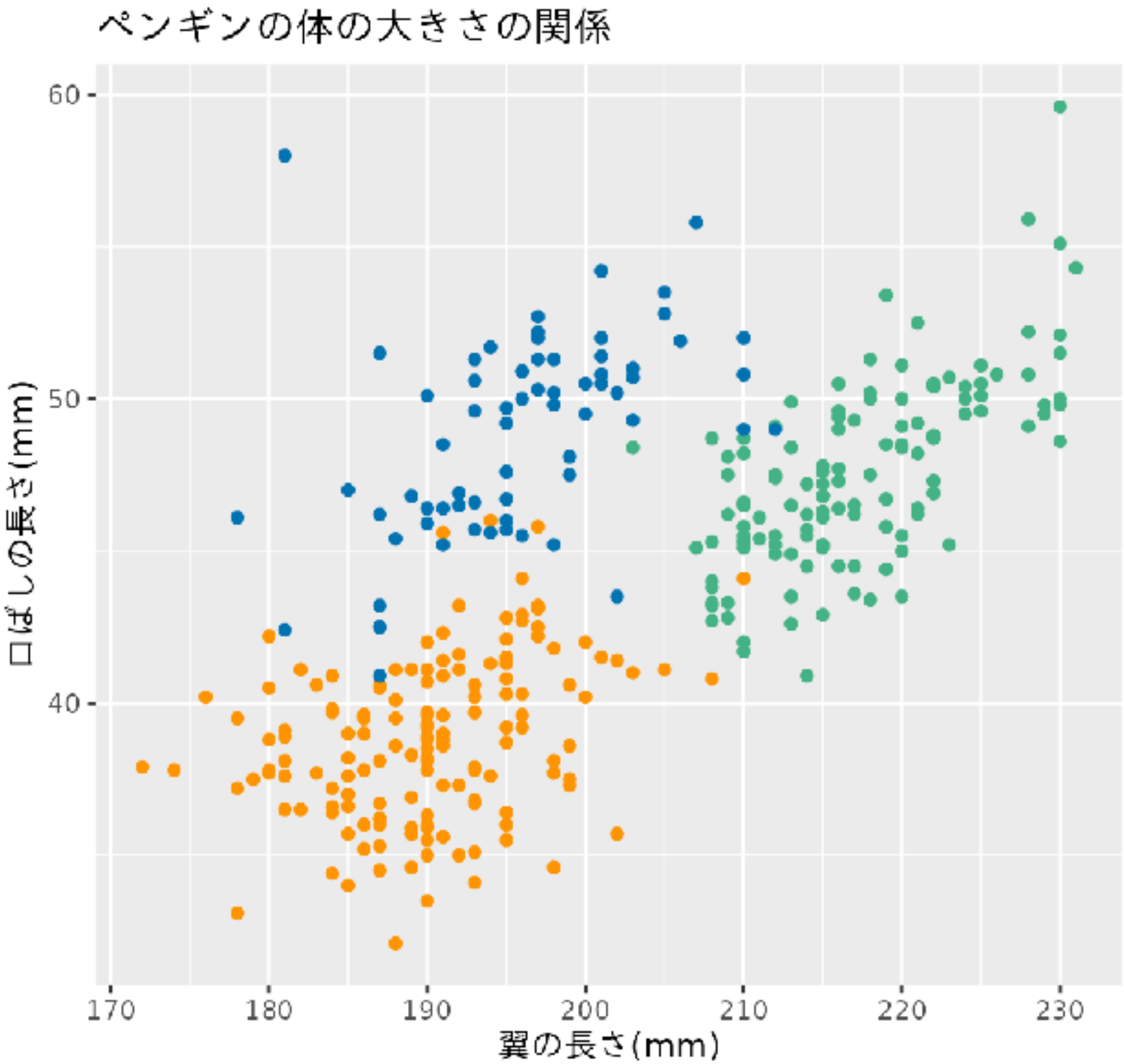
# データの背景を探る



# 複数の変数からデータを説明する

データがもつ意味、その背景を探る

複数のデータを比較し、その関係性を明らかにする



## 🔧 データ分析の手法

### 📊 関係の数値化

共分散                      相関係数

### 📈 グラフ、表による表現

散布図                      クロス集計表



# データ分析における2つの関係

複数の変数がともに変化する状態

データ分析では**相関関係**と**因果関係**の2つの関係を扱う（似て非なるもの）

## 相関関係

ある出来事や物事と別の出来事や物事の間に関係があるもの

ペンギン個体の翼の長さ  $\longleftrightarrow$  ペンギン個体のくちばしの長さ

## 因果関係

ある出来事や物事が**原因**となって、別の出来事や物事（**結果**）が起こるもの

ある水道会社の利用を止める  $\longrightarrow$  水道を利用していた地域のコレラ患者が減る

## 見せかけの相関

観測されていない第三の要因によって相関関係が因果関係のように見えるもの（疑似相関とも言う）

一人当たりのチョコレートの消費量が増える  $\cdots\cdots\longrightarrow$  ノーベル賞受賞者が増える

一人当たりのGDPが増える

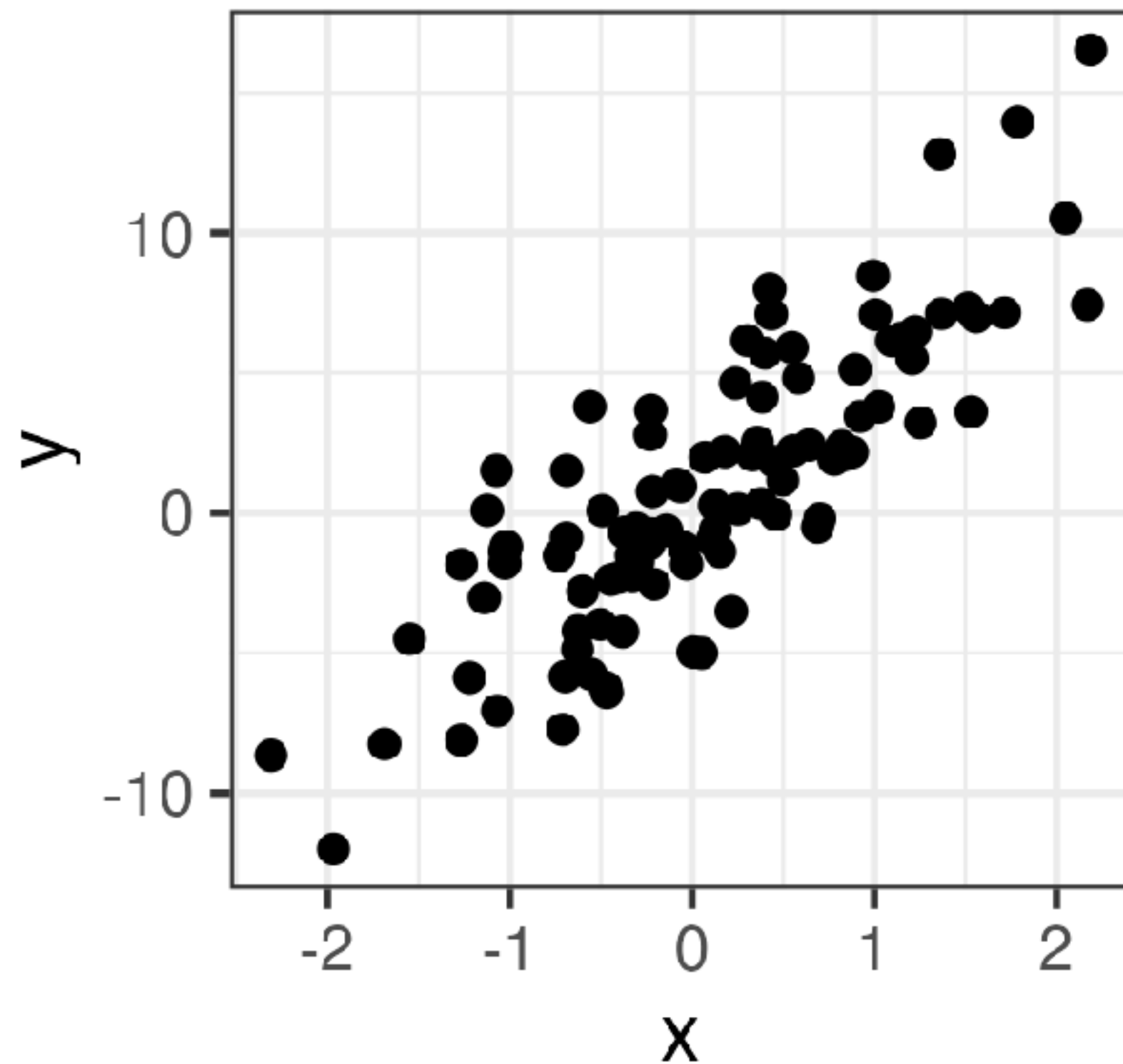


# 相関

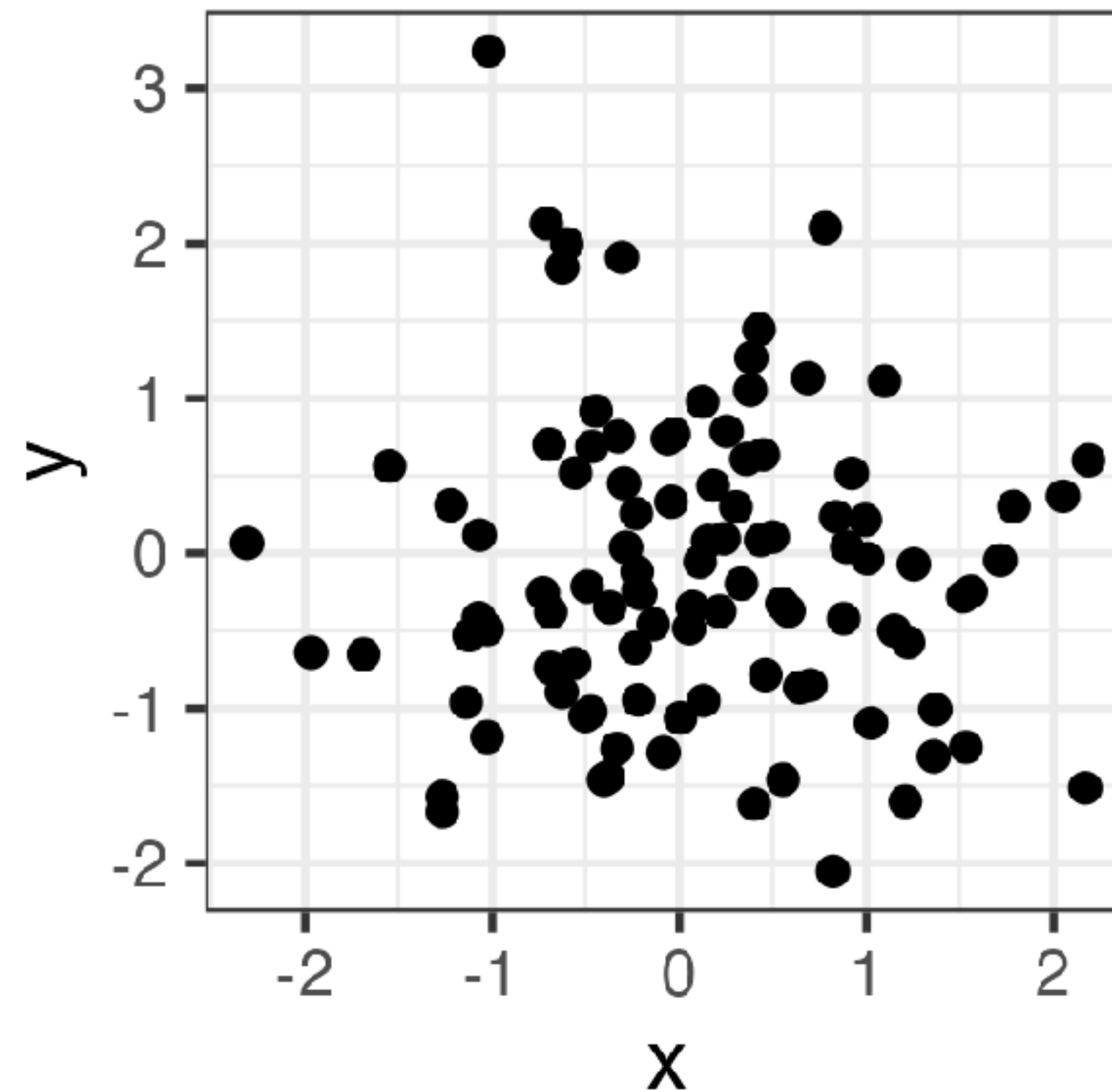
2つの変数間で起こる関係を表す

散布図（後述）としてグラフ上に可視化することで傾向を把握しやすくなる

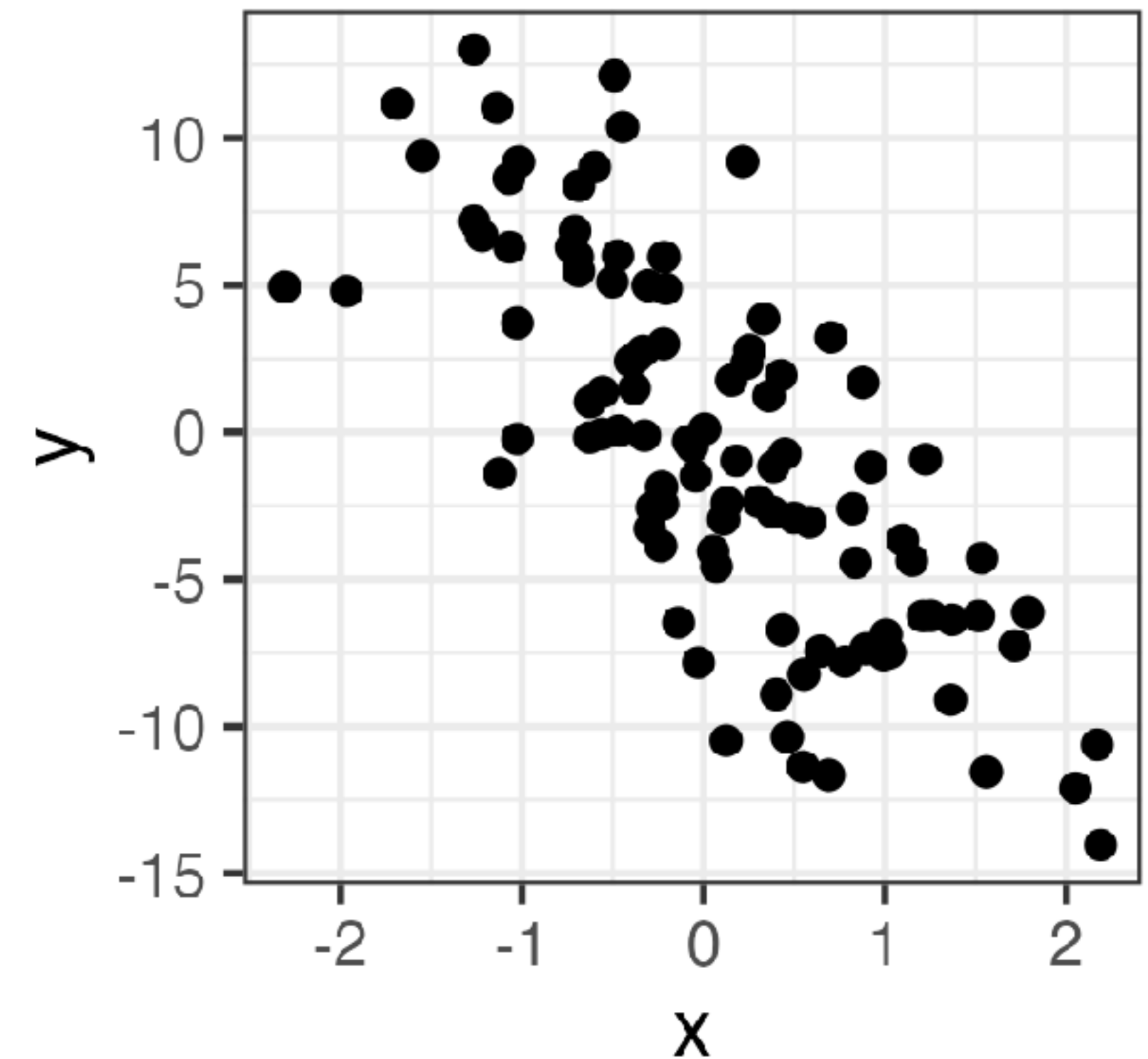
① 正の相関関係



② 無相関



③ 負の相関関係



2つの変数の関係を示す3つの状態

# 関係の数値化1: 共分散 covariance

2つの変数(xとy)についての共分散は次のように求められる

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

手順

①、②変数x(y)の値から変数x(y)の平均値を引く→偏差

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

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

$$Cov_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad ③ \text{ 偏差の積を求める}$$

# 関係の数値化1: 共分散 covariance

手順 (つづき) ④  $n$  (すべてのデータ) まで右の処理を行い、それを足し合わせる


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

データの  $i = 1$  番目から

⑤ 変数  $x$  と変数  $y$  の各値に対して偏差を求め、それを掛け合わせたものを足す

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

$n$ (データ数)で割る

```
# Rの標準関数で共分散を求めるとデータの数 - 1で割る不偏共分散になる   
cov(df_animal$body_length_cm,  
      df_animal$weight_kg,  
      use = "complete.obs")  
#> [1] 6619.572
```

# 共分散の特徴

値が大きいほど2変数の関係が強いことを示す

変数の単位に依存して値が変わる

```
cov(  
  # cm を m に  
  set_units(set_units(df_animal$body_length_cm, cm), m),  
  df_animal$weight_kg,  
  use = "complete.obs")  
#> [1] 66.19572  
cov(  
  # cm を m に  
  set_units(set_units(df_animal$body_length_cm, cm), m),  
  # kg を g に  
  set_units(set_units(df_animal$weight_kg, kg), g),  
  use = "complete.obs")  
#> [1] 66195.72
```



# 関係の数値化2: 相関係数

共分散の単位依存の問題を解消する指標

共分散を各変数の標準偏差の積で割ることで算出される

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

-1から1までの値をとる。変数の関係が強いほど絶対値が1に近づく

# 気温とアイスの相関係数を求める

```
cor(df_icecream_temperature$temperature_average_c,  
    df_icecream_temperature$value)
```

```
#> [1] 0.9144466
```



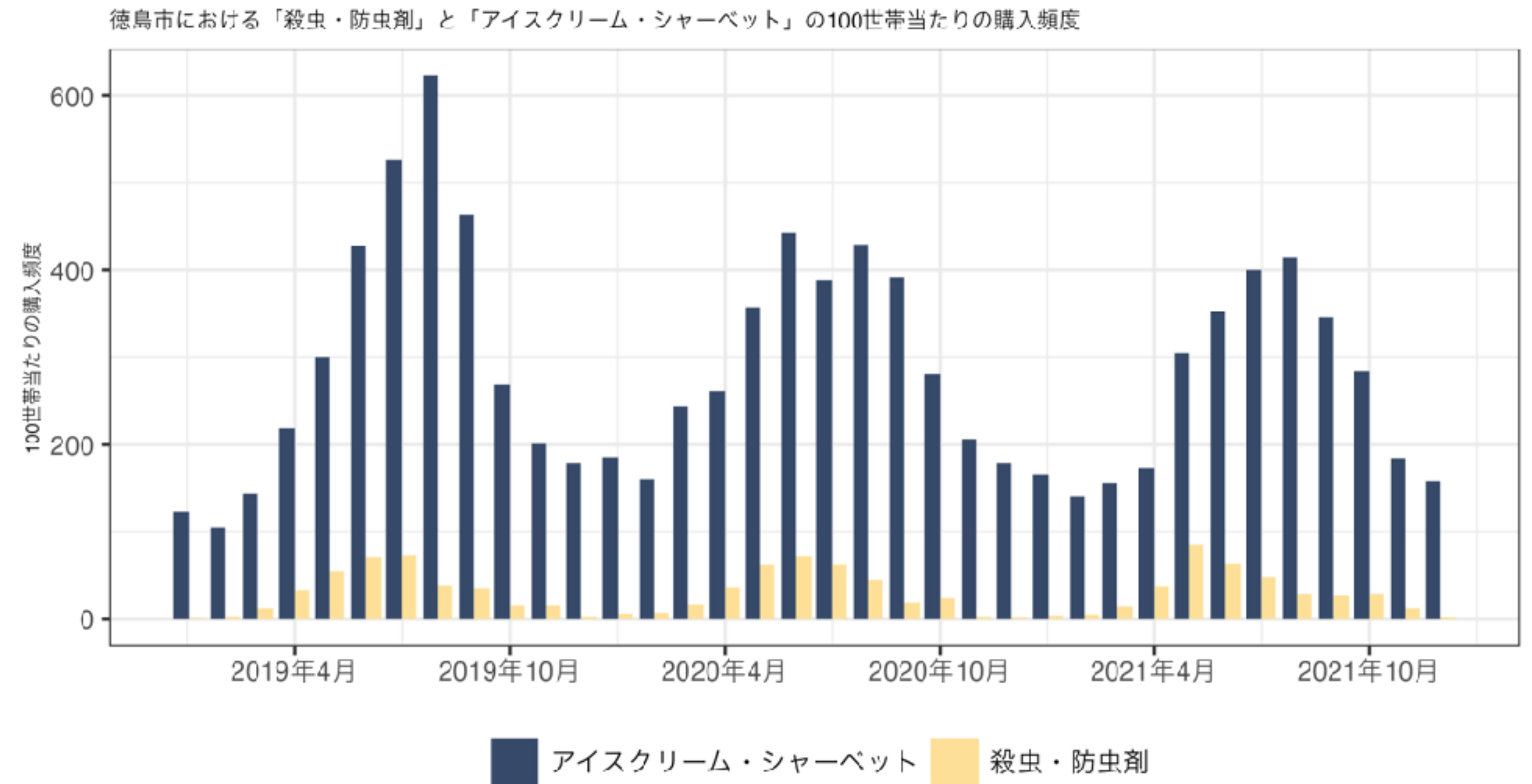
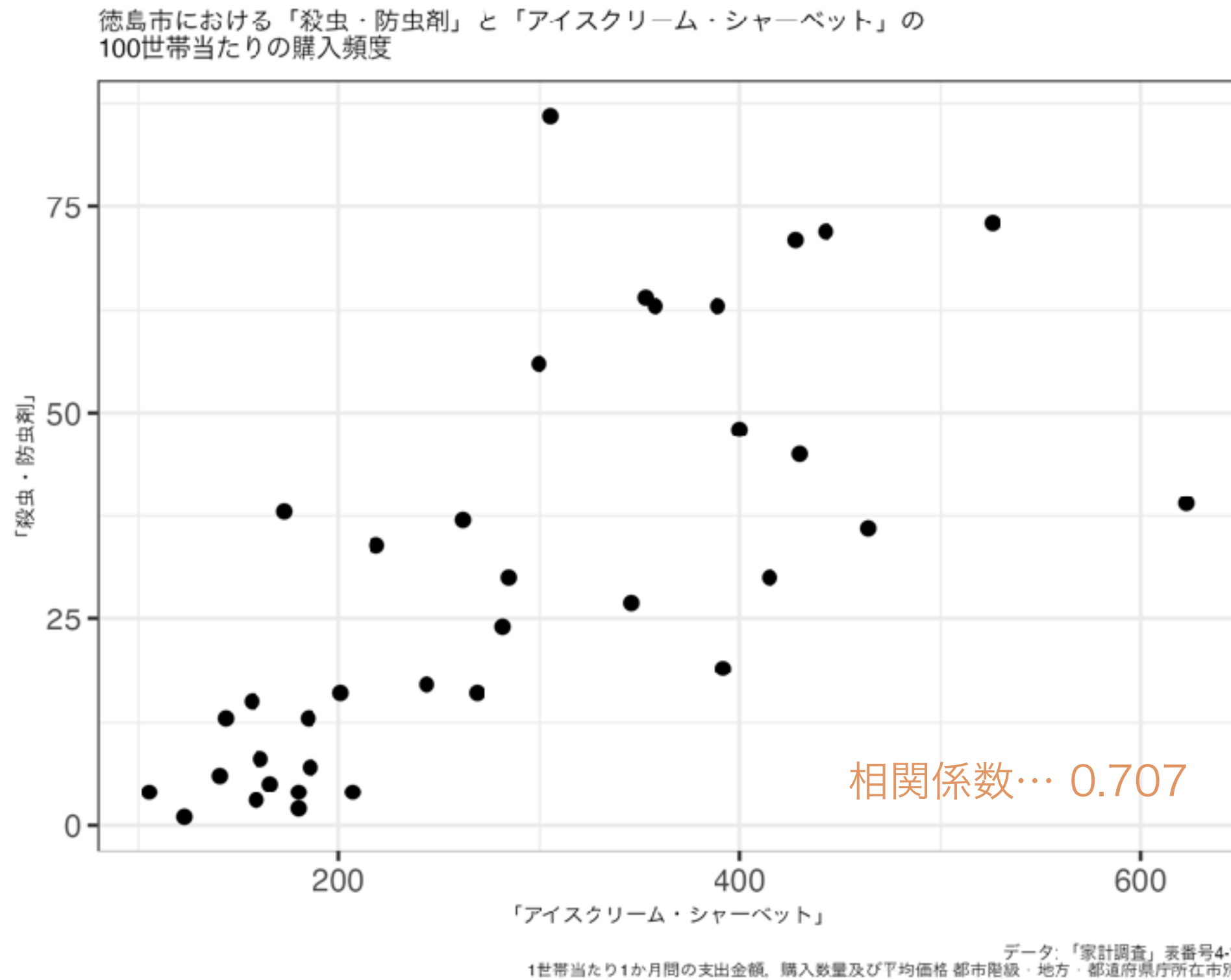
相関係数	相関の強さ
±0.7以上	とても強い
±0.4~0.7	やや強い
±0.2~0.4	弱い
±0.2以下	ほとんどなし



# 見せかけの相関

因果関係がありそうに見える二変数の関係が、  
観測されていない第三の変数（潜在変数）の効果によってもたらされるもの

例) 「殺虫・防虫剤」と「アイスクリーム・シャーベット」の購入頻度… **どちらも気温の影響を受ける**



データ:「家計調査」表番号4-1  
1世帯当たり1か月間の支出金額、購入数量及び平均価格 都市階級・地方・都道府県庁所在市別

→相関から因果関係を導き出すのは難しい

# 相関係数行列

変数のペアごとに計算した相関係数を行列形式で表現する


徳島市における月平均気温と  
「殺虫・防虫剤」（殺虫剤）と「アイスクリーム・シャーベット」（アイス）の  
購入頻度の相関係数行列

	殺虫剤	アイス	気温
殺虫剤	1.000	0.707	0.709
アイス	0.707	1.000	0.914
気温	0.709	0.914	1.000

気温は殺虫剤、アイスとも相関が強い… 見せかけの相関を示唆

# クロス集計

カテゴリ変数間の数値（個数や比率）を集計したもの  
集計結果をクロス集計表または分割表として表形式で表す

```
# 動物データから分類群と体重（200kg以上かどうか）ごとの件数を集計   
xtabs(~ taxon + heavy_mt_200kg,  
       data = df_animal)  
#>           heavy_mt_200kg  
#> taxon      FALSE TRUE  
#> 偶蹄類         0    1  
#> 霊長類         4    0  
#> 食肉類         5    2  
#> 鯨偶蹄類       1    0  
#> 鳥類           5    0  
#> 齧歯類         2    0
```

**目的に応じた軸（切り口）を設定することが重要**

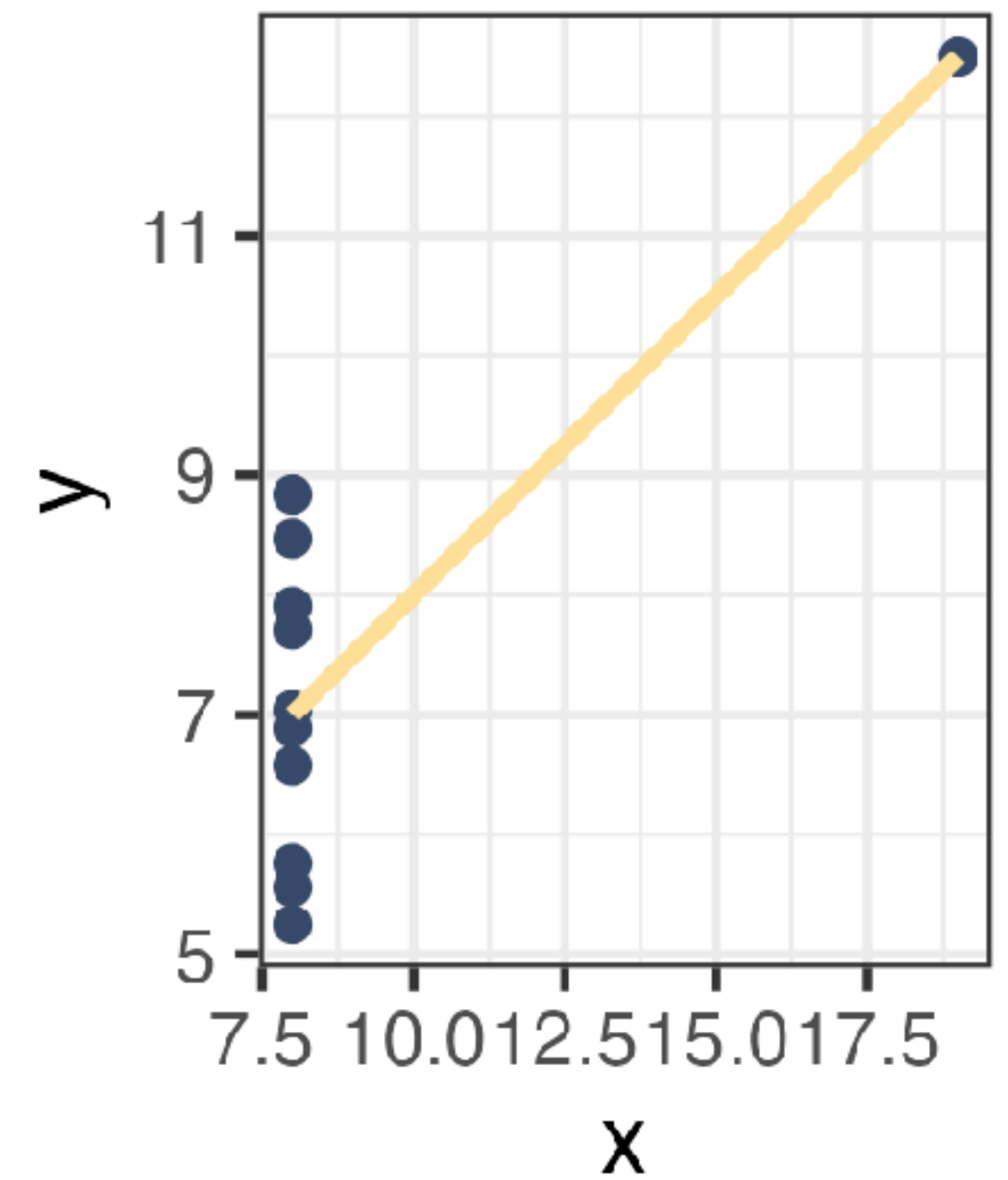
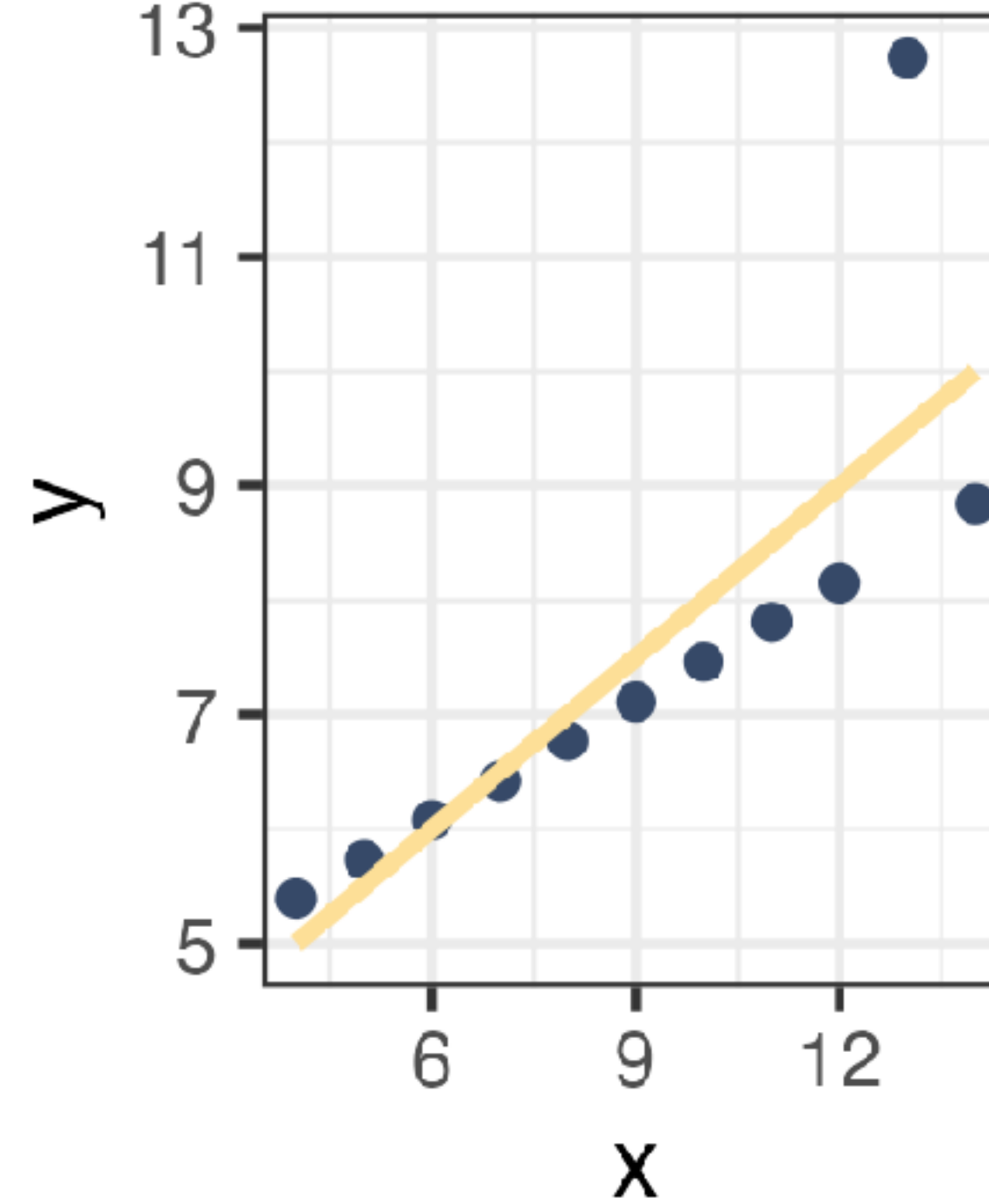
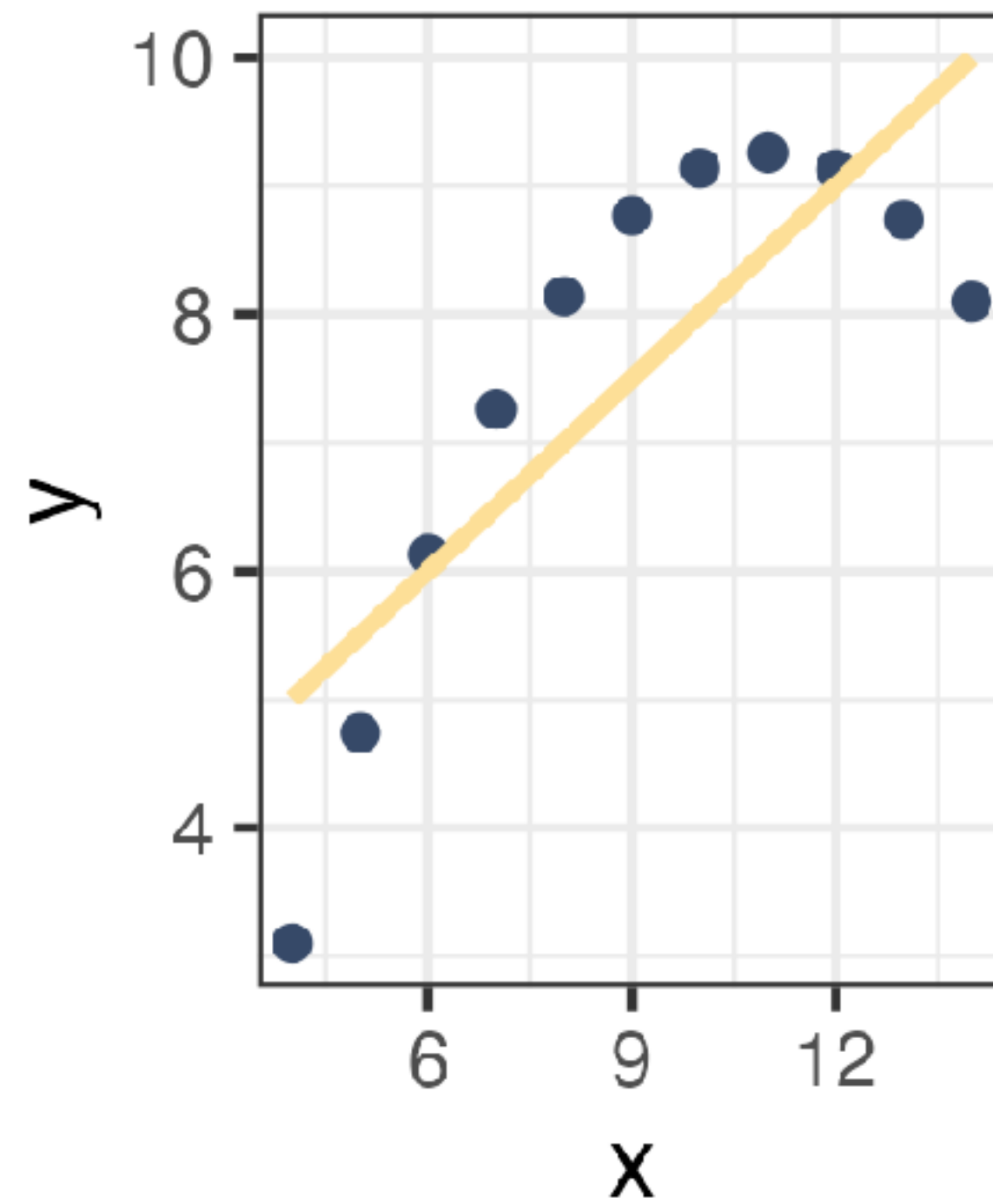
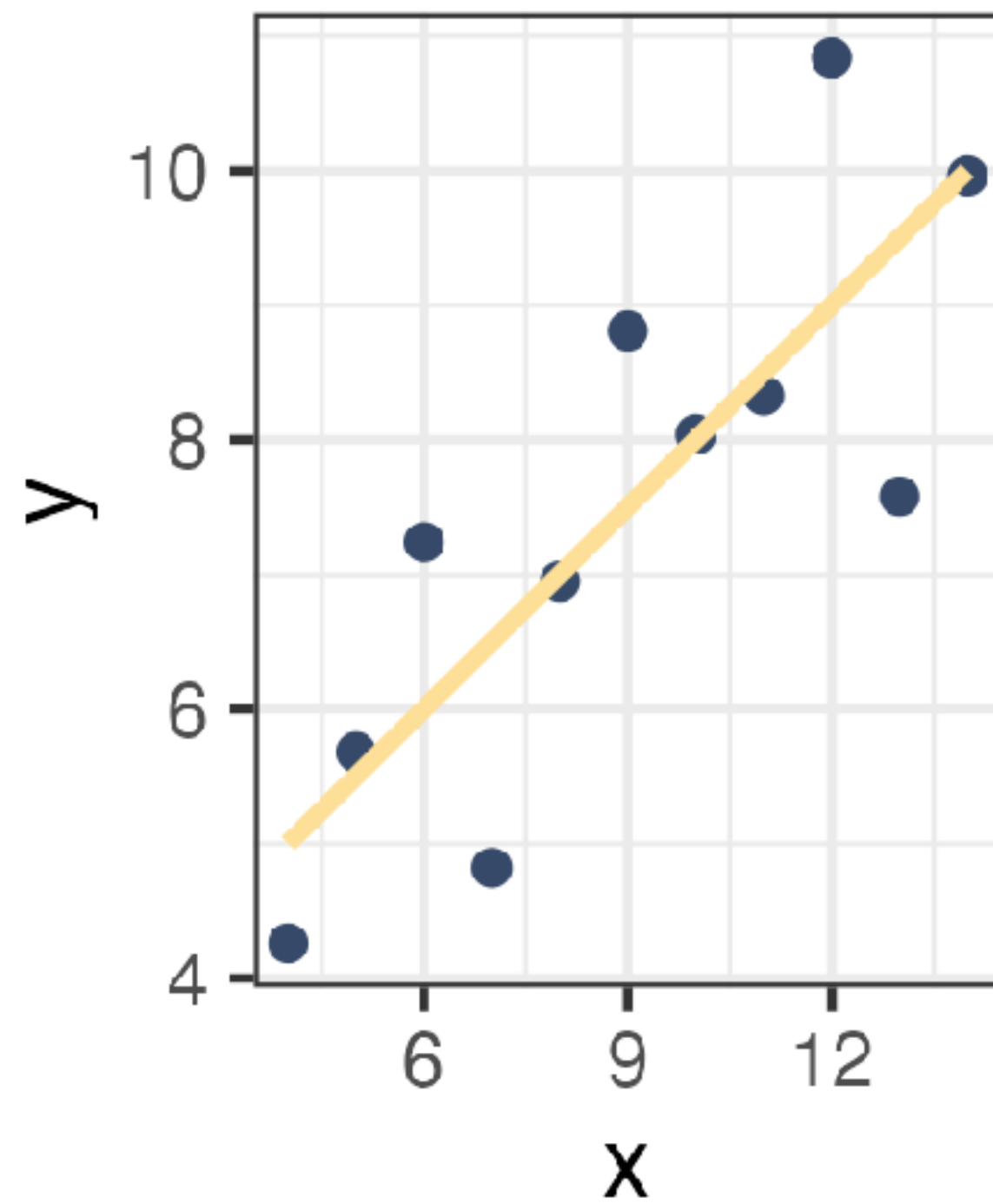
例）年齢を10代ごとにわけ、未成年、成人、高齢者（60歳以上）で分ける



# データの特徴を捉える(2): データ可視化

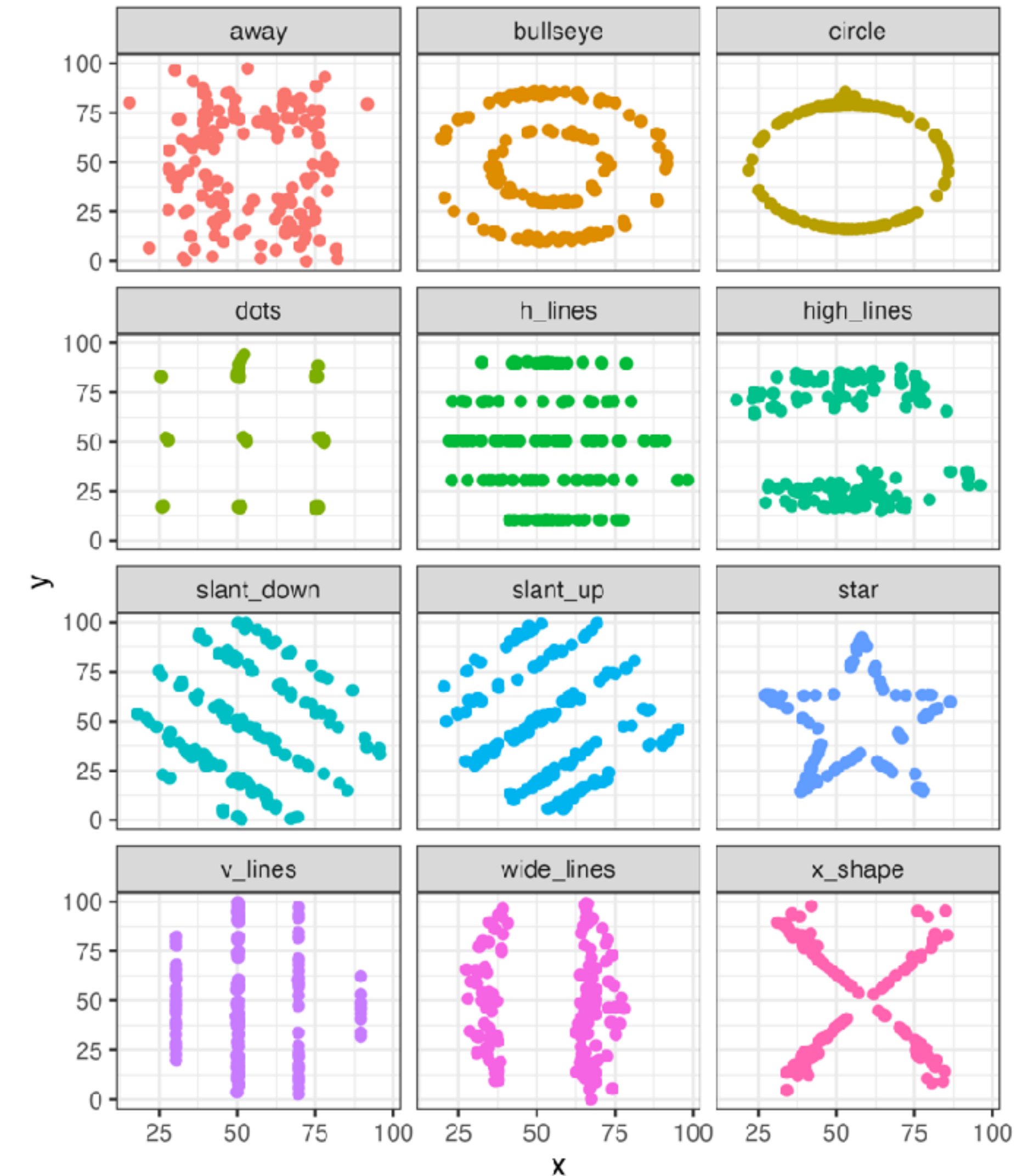
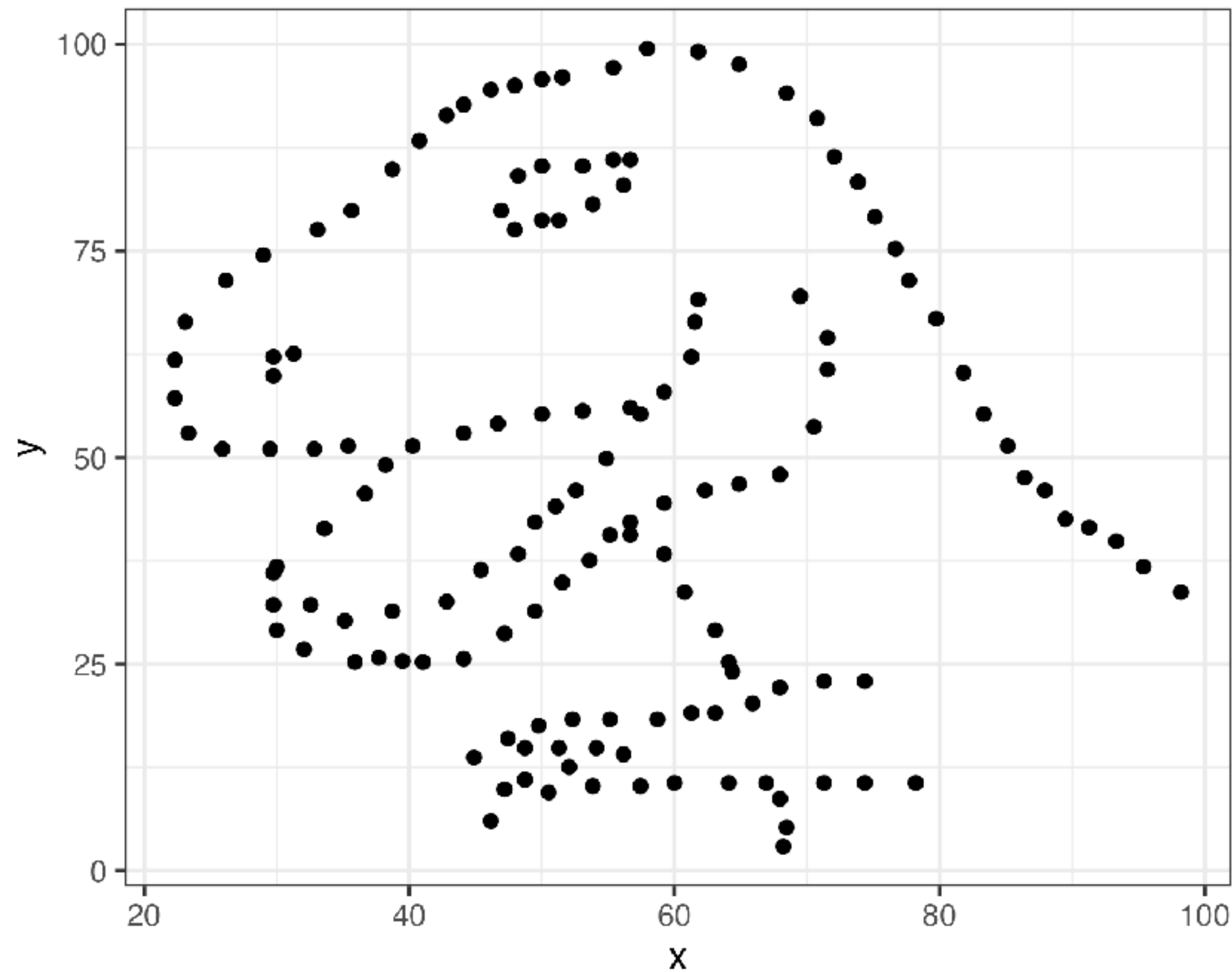
# データ可視化の重要性: アンスコムの例

記述統計量や相関係数がほぼ同じ値であっても、中身のデータが異なることを示す



# データ可視化の重要性: アンスコムサウルス

記述統計量がほぼ同じでありながら散布図にすると異なる図を描く  
データ生成のアルゴリズム



課題: 利用可能なデータから適当な作図をおこなう

# 参考文献・URL

滋賀大学データサイエンス学部・長崎大学情報データ科学部 共編 (2022). データサイエンスの歩き方 (学術図書出版社)  
ISBN: 978-4-7806-0936-3

石田基広、大藪進喜 監修・著 上田哲史、掛井秀一、金西計英、谷岡広樹、中山慎一、芳賀昭弘 著 (2021).  
情報科学入門: 統計・データサイエンス・AI (技術評論社)  
ISBN: 978-4-297-12040-5

北川源四郎、竹村彰通 編 内田誠一、川崎能典、孝忠大輔、佐久間淳、椎名洋、中川裕志、樋口知之、丸山宏 著 (2021).  
教養としてのデータサイエンス (講談社サイエンティフィク)  
ISBN: 978-4-06-523809-7

東京大学教養学部統計学教室 編 (1991). 基礎統計学 1 (統計学入門) (東京大学出版会) ISBN: 978-4-13-042065-5

田栗正隆、汪金芳 著 (2022). データサイエンスの基礎 (オーム社) ISBN: 978-4-274-22914-5

瓜生真也 「実践的データサイエンス」 <https://uribo.github.io/practical-ds/intro>

瓜生真也 「データ分析入門」 [https://uribo.github.io/tokupon\\_ds/](https://uribo.github.io/tokupon_ds/)

瓜生真也 「Rによるデータ可視化と地図表現」 [https://speakerdeck.com/s\\_uryu/rniyorudetake-shi-hua-todi-tu-biao-xian](https://speakerdeck.com/s_uryu/rniyorudetake-shi-hua-todi-tu-biao-xian)