

データサイエンスへの誘い

第4回: データの要約

瓜生真也（デザイン型AI教育研究センター・助教）

講義内容（予定）

講義に関する資料（スライド、補足資料等）を  GitHubに置いておきます

 <https://github.com/uribo/INNV1250>



ダウンロード可能



1. ガイダンス、データサイエンスとは何か
2. 現代社会におけるデータサイエンスの活用事例
3. データ処理の手法
4. データの要約
5. データの可視化
6. データと確率
7. データからの推論
8. 複数のデータを比較する

9. 統計のウソ
10. 統計的モデリング
11. 統計的学習
12. さまざまなデータサイエンスの手法
13. 機械学習とAI
14. コンピューターを用いた分析
15. ビッグデータの扱い
16. 期末試験（8月1日）

今日の目標

要約統計量の違いを理解し、

使い分けができるようになる

【課題】 データの特徴を表現する方法を理解する

提出期限: 来週の講義開始前まで

manabaのレポートとして提出してください

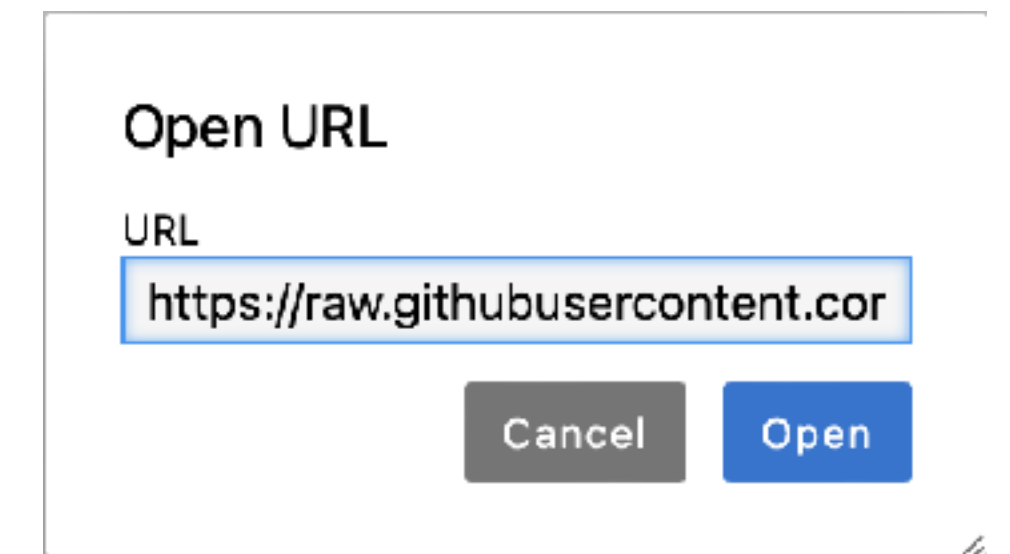
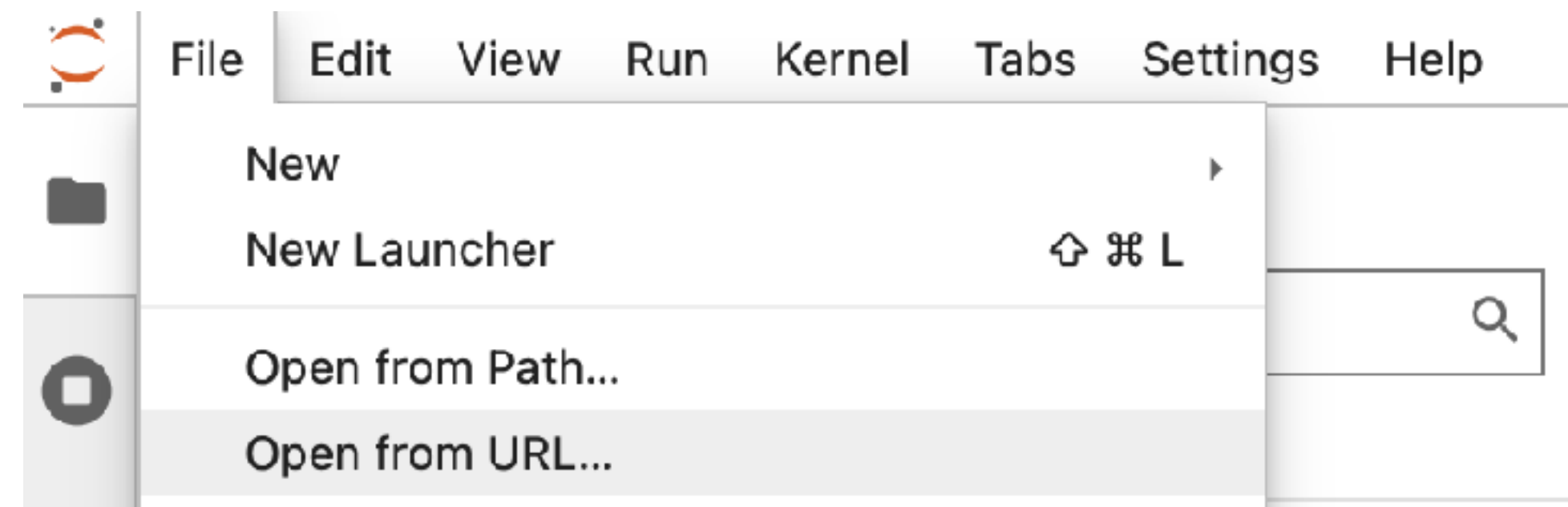
GitHubからyour_turn.ipynbをアップロードして記載

week04/your_turn.ipynb



Rawをクリックして表示先のURLをコピー

JupyterHubのサーバを起動、メニューのFileから “Open from URL…” を選択



注意: ファイル名は英数字のみにすること

日本語（漢字、片仮名、平仮名）、全角英数字、スペース、記号等は使わない

ファイルをダウンロードしても開けなくても問題ない（気にしない）

内容の確認、編集はJupyterHub上で行う

データの特徴を表現する

データの特徴を伝えるには？

📍 (参照) 第1回の講義

データ分析で扱うデータは一般的に膨大
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
→マジカルナンバー7
短期記憶内にとどめておける情報量の上限

📊 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

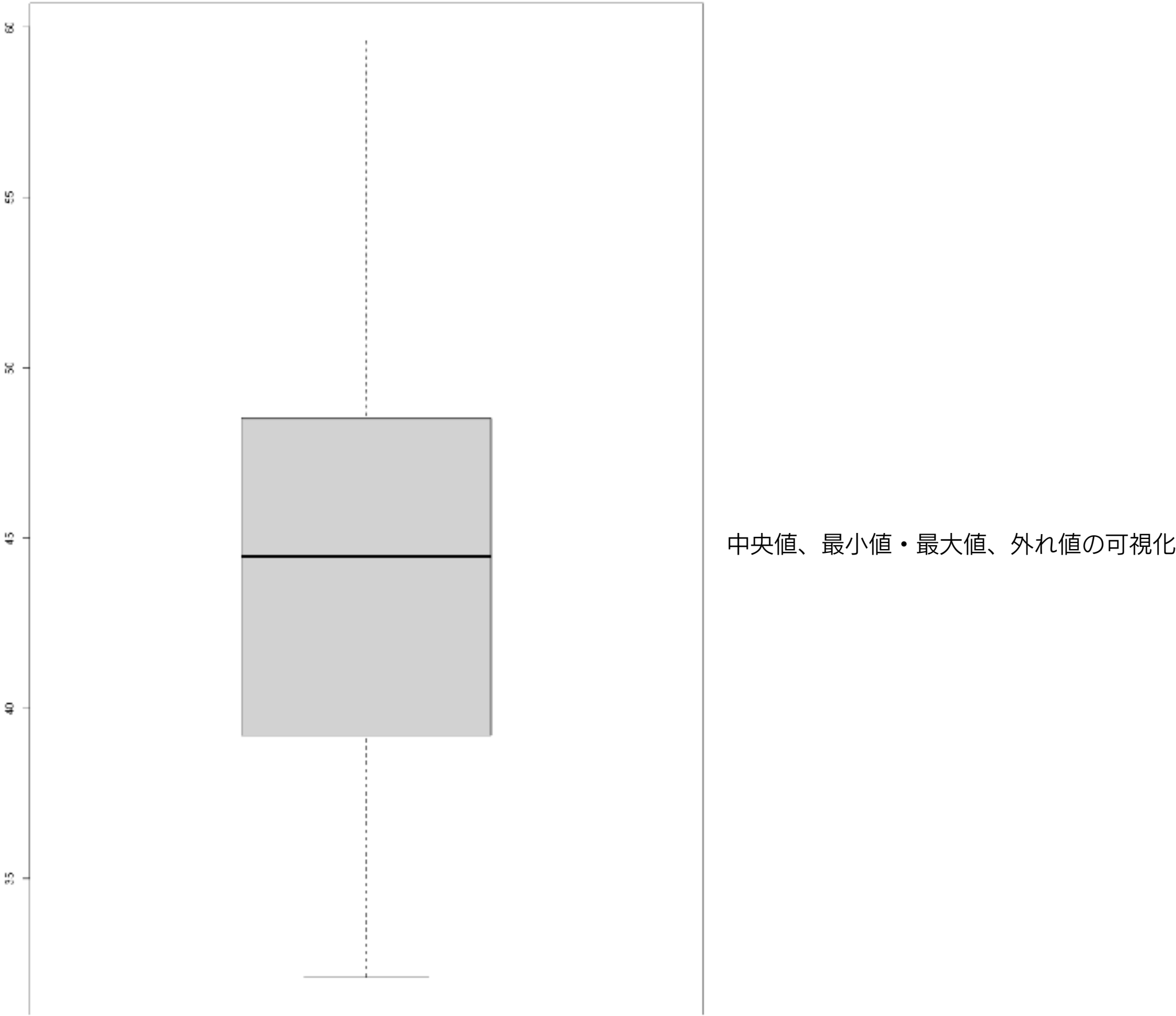
数値の羅列から特徴を読み取るのは困難・・・

39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, 42, 37.8, 37.8, 41.1, 38.6, 34.6, 36.6, 38.7, 42.5, 34.4, 46, 37.8, 37.7, 35.9, 38.2, 38.8, 35.3, 40.6, 40.5, 37.9, 40.5, 39.5, 37.2, 39.5, 40.9, 36.4, 39.2, 38.8, 42.2, 37.6, 39.8, 36.5, 40.8, 36, 44.1, 37, 39.6, 41.1, 37.5, 36, 42.3, 39.6, 40.1, 35, 42, 34.5, 41.4, 39, 40.6, 36.5, 37.6, 35.7, 41.3, 37.6, 41.1, 36.4, 41.6, 35.5, 41.1, 35.9, 41.8, 33.5, 39.7, 39.6, 45.8, 35.5, 42.8, 40.9, 37.2, 36.2, 42.1, 34.6, 42.9, 36.7, 35.1, 37.3, 41.3, 36.3, 36.9, 38.3, 38.9, 35.7, 41.1, 34, 39.6, 36.2, 40.8, 38.1, 40.3, 33.1, 43.2, 35, 41, 37.7, 37.8, 37.9, 39.7, 38.6, 38.2, 38.1, 43.2, 38.1, 45.6, 39.7, 42.2, 39.6, 42.7, 38.6, 37.3, 35.7, 41.1, 36.2, 37.7, 40.2, 41.4, 35.2, 40.6, 38.8, 41.5, 39, 44.1, 38.5, 43.1, 36.8, 37.5, 38.1, 41.1, 35.6, 40.2, 37, 39.7, 40.2, 40.6, 32.1, 40.7, 37.3, 39, 39.2, 36.6, 36, 37.8, 36, 41.5, 46.1, 50, 48.7, 50, 47.6, 46.5, 45.4, 46.7, 43.3, 46.8, 40.9, 49, 45.5, 48.4, 45.8, 49.3, 42, 49.2, 46.2, 48.7, 50.2, 45.1, 46.5, 46.3, 42.9, 46.1, 44.5, 47.8, 48.2, 50, 47.3, 42.8, 45.1, 59.6, 49.1, 48.4, 42.6, 44.4, 44, 48.7, 42.7, 49.6, 45.3, 49.6, 50.5, 43.6, 45.5, 50.5, 44.9, 45.2, 46.6, 48.5, 45.1, 50.1, 46.5, 45, 43.8, 45.5, 43.2, 50.4, 45.3, 46.2, 45.7, 54.3, 45.8, 49.8, 46.2, 49.5, 43.5, 50.7, 47.7, 46.4, 48.2, 46.5, 46.4, 48.6, 47.5, 51.1, 45.2, 45.2, 49.1, 52.5, 47.4, 50, 44.9, 50.8, 43.4, 51.3, 47.5, 52.1, 47.5, 52.2, 45.5, 49.5, 44.5, 50.8, 49.4, 46.9, 48.4, 51.1, 48.5, 55.9, 47.2, 49.1, 47.3, 46.8, 41.7, 53.4, 43.3, 48.1, 50.5, 49.8, 43.5, 51.5, 46.2, 55.1, 44.5, 48.8, 47.2, NA, 46.8, 50.4, 45.2, 49.9, 46.5, 50, 51.3, 45.4, 52.7, 45.2, 46.1, 51.3, 46, 51.3, 46.6, 51.7, 47, 52, 45.9, 50.5, 50.3, 58, 46.4, 49.2, 42.4, 48.5, 43.2, 50.6, 46.7, 52, 50.5, 49.5, 46.4, 52.8, 40.9, 54.2, 42.5, 51, 49.7, 47.5, 47.6, 52, 46.9, 53.5, 49, 46.2, 50.9, 45.5, 50.9, 50.8, 50.1, 49, 51.5, 49.8, 48.1, 51.4, 45.7, 50.7, 42.5, 52.2, 45.2, 49.3, 50.2, 45.6, 51.9, 46.8, 45.7, 55.8, 43.5, 49.6, 50.8, 50.2

平均値と標準偏差によってデータの特徴を把握する

```
# mean±sd
43.92193±5.459584
```

箱ひげ図を作成し、データの特徴を把握する



要約統計量(summary statistics)

数値を用いた統計的な指標

主に数値データの特徴を把握するのに用いられる

平均値 (mean)、中央値 (median)、最小値 (min)、最大値 (max) など

数値データのベクトルに対してsummary()関数を実行

```
summary(df_animal$body_length_cm)
```

```
#>      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
```

#>	1.20	63.62	82.50	102.87	133.00	250.00	4
----	------	-------	-------	--------	--------	--------	---



データフレームに対してsummary()関数を実行すると、各列についての要約統計量が表示される 

```
summary(df_animal)
```

#>	taxon	name	body_length_cm	weight_kg
----	-------	------	----------------	-----------

```
#>      Length:22      Length:22      Min.      : 1.20      Min.      : 0.90
```

```
#>      Class :character      Class :character      1st Qu.: 63.62      1st Qu.:  5.85
```

```
#> Mode :character Mode :character Median : 82.50 Median : 12.50
```

```
#>      Mean      :102.87      Mean      : 65.81
```

```
#> 3rd 0y.:133.00 3rd 0y.: 69.50
```

#> Max. :250.00 Max. :410.00

```
#>      NA's      :4      NA's      :2
```

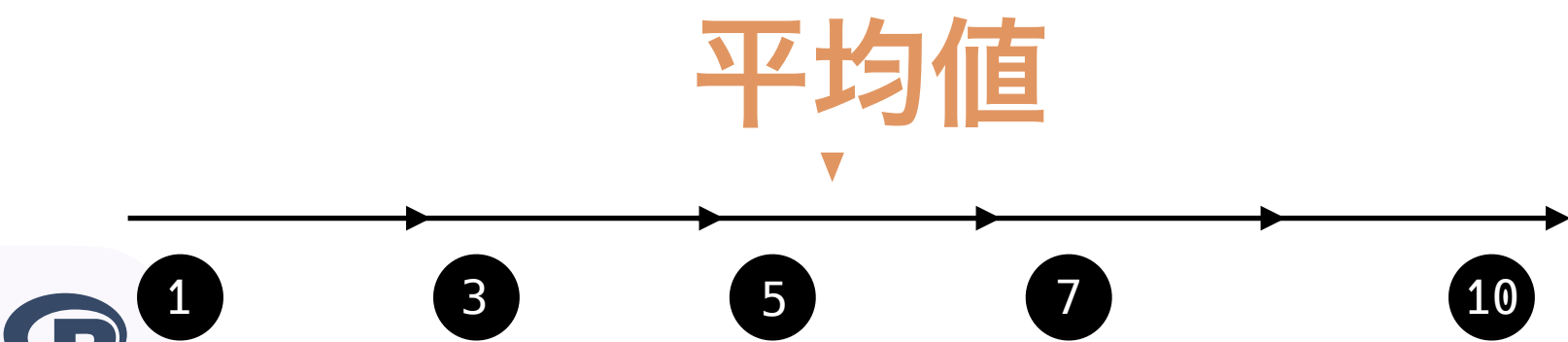

代表値の算出

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

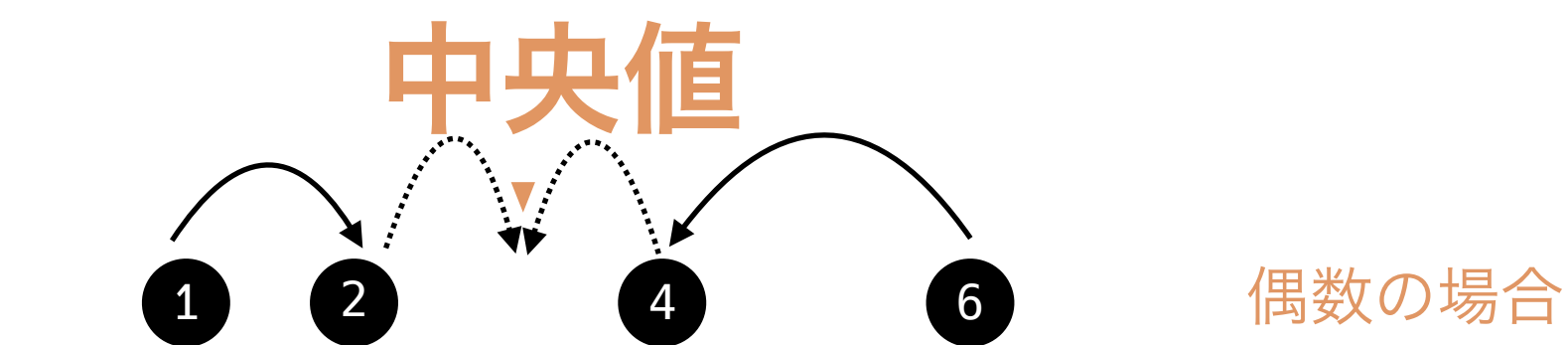
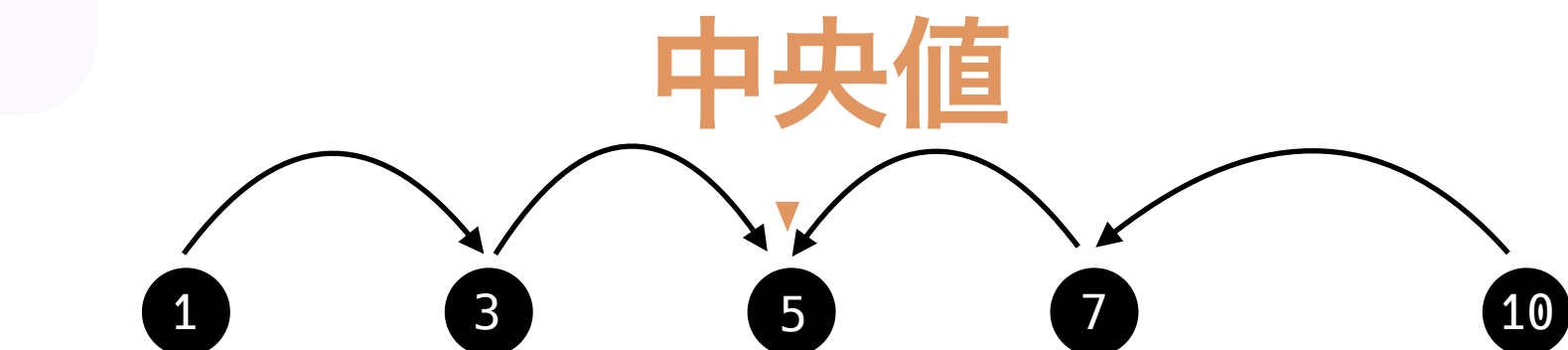
```
x <- c(1, 10, 5, 3, 7)
sum(x) / length(x)
#> [1] 5.2
# mean( )関数を用いて平均値を計算
mean(x)
#> [1] 5.2
```



中央値

データに含まれる数の真ん中となる値

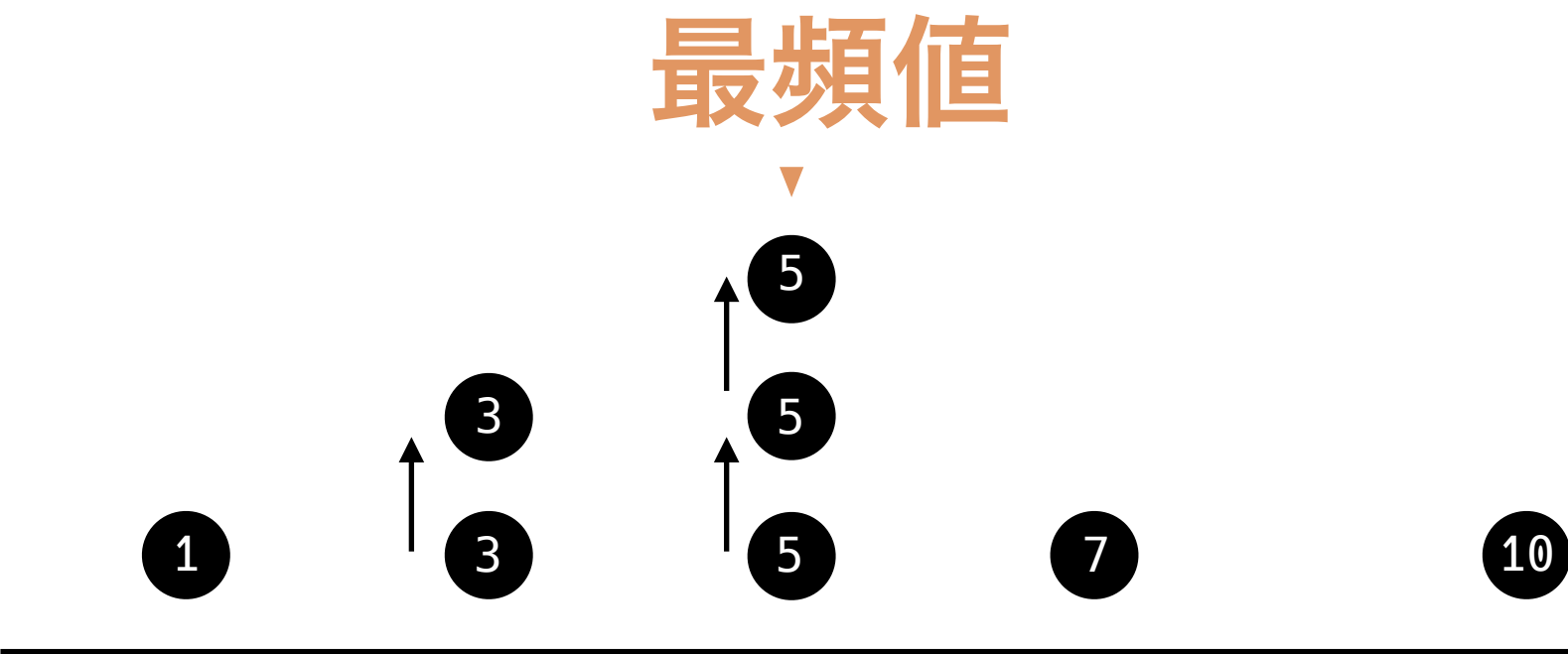
```
sort(x)[ceiling(length(x)/2)]
#> [1] 5
# median( )関数で数値ベクトルの中央値を計算
median(x)
#> [1] 5
```



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
```



平均値は外れ値の影響を受けやすい

種名	体重(kg)
ミーアキャット	0.9
リスザル	1.1
モルモット	1.5
コツメカワウソ	5.4
ホッキョクグマ	410

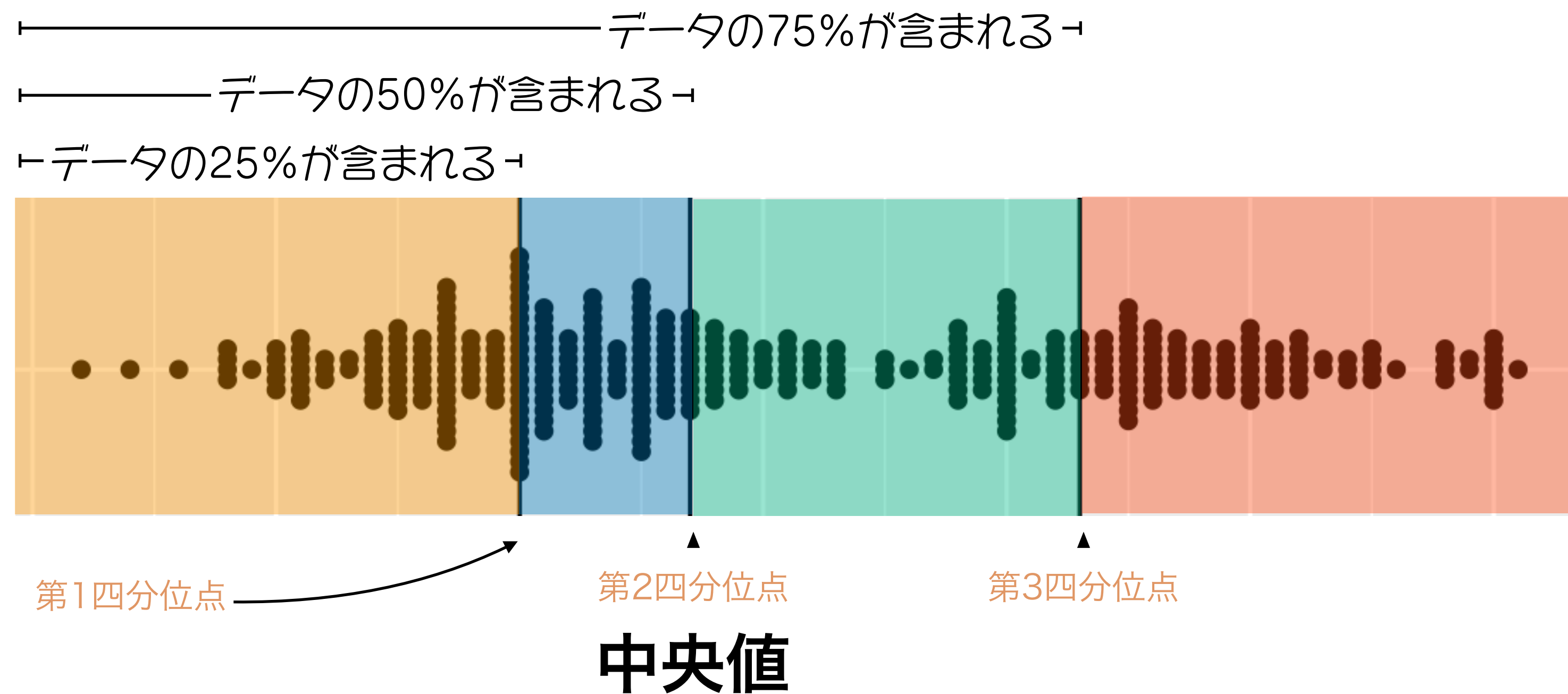



平均値より小さい動物は5種中4種

平均値が外れ値に引っ張られる

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
```


分散(variance)

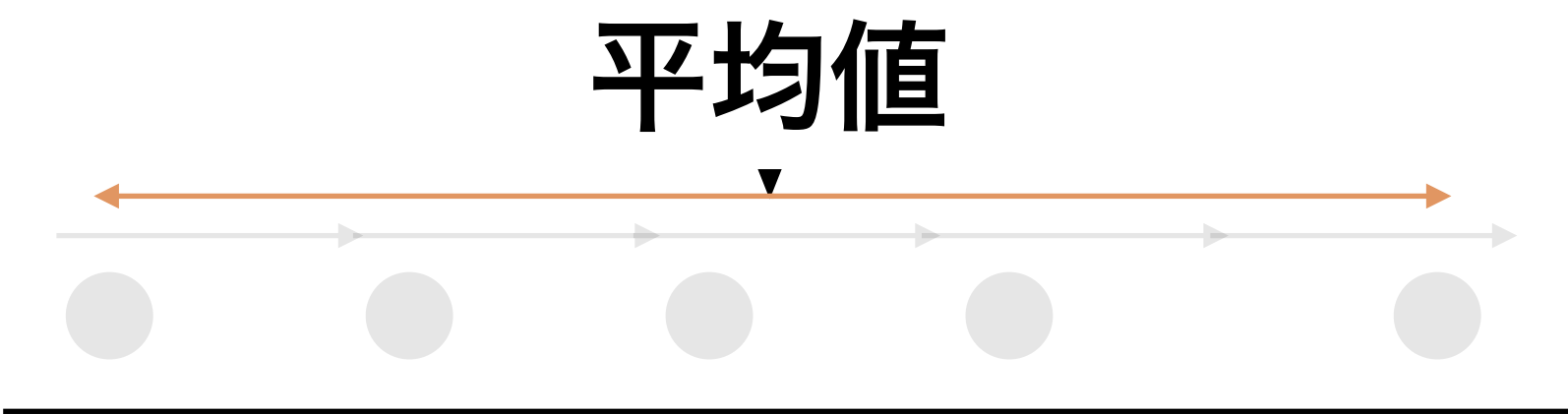
各値が平均値を中心としてどのように散らばっているかを示す

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？



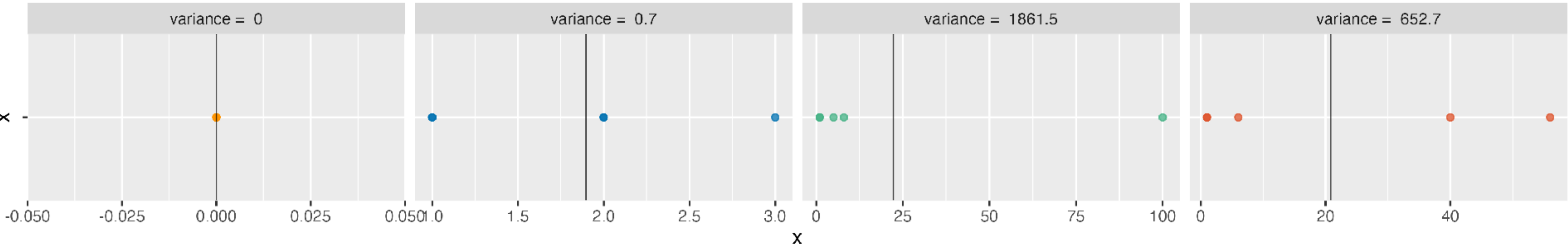
データの分布について具体的な説明ができるようになる

`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

分散を算出してみよう

ペンギンデータのうち、アデリーペンギンの5頭の体重(body_mass_g)について考える

```
R # library(dplyr)
df <-
  penguins |>
  filter(species == "Adelie") |>
  select(body_mass_g) |>
  filter(!is.na(body_mass_g)) |>
  slice_head(n = 5)

df
#> # A tibble: 5 × 1
#>   body_mass_g
#>   <int>
#> 1     3750
#> 2     3800
#> 3     3250
#> 4     3450
#> 5     3650
```



分散を算出してみよう

1. 変数の平均値を出す
2. 偏差を求める
3. 偏差を2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

```
R df <-  
  df |>  
  # 各値について偏差 deviation (平均よりもいくらか大きいか小さいか) を求める  
  
  mutate(deviation = body_mass_g - mean(df$body_mass_g, na.rm = TRUE))  
df  
#> # A tibble: 5 × 2  
#>   body_mass_g deviation  
#>   <int>      <dbl>  
#> 1      3750        170  
#> 2      3800        220  
#> 3      3250       -330  
#> 4      3450       -130  
#> 5      3650         70
```

偏差の特徴

正の値と負の値の両方が混ざる

⌘ 負の値でも2乗すると正の値になる

合計すると0になる



分散を算出してみよう

1. 変数の平均値を出す
2. 偏差を求める
3. 偏差を2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

Rの標準関数で分散を求める
※データの数 - 1で割る**不偏分散**

```
R df <-  
  df |>  
  mutate(deviation2 = deviation^2)  
  
df  
#> # A tibble: 5 × 3  
#>   body_mass_g deviation deviation2  
#>   <int>      <dbl>      <dbl>  
#> 1      3750        170      28900  
#> 2      3800        220      48400  
#> 3      3250       -330     108900  
#> 4      3450       -130      16900  
#> 5      3650         70       4900  
  
sum(df$deviation2) / nrow(df)  
#> [1] 41600
```

```
R var(df$body_mass_g)  
#> [1] 52000
```



標準偏差(standard deviation)

分散について平方根を求める



平均からの偏差 平均からの偏差の2乗

	body_mass_g	deviation	deviation^2
	3750	<small>3750-3580=</small> 170	<small>170x170=</small> 28900
	3800	<small>3800-3580=</small> 220	<small>220x220=</small> 48400
	3250	<small>3250-3580=</small> -330	<small>-330x-330=</small> 108900
	3450	<small>3450-3580=</small> -130	<small>-130x-130=</small> 16900
	3650	<small>3650-3580=</small> 70	<small>70x70=</small> 4900
total	17,900.00	0.00	208,000.00
mean	3,580.00	0.00	41,600.00

偏差の合計は0

平方根を利用する理由

分散を求めたときに2乗したものを元に戻すため

2乗すると単位が変わるものの影響を取り除く

$cm \rightarrow cm^2$

$\sqrt{cm^2} \rightarrow cm$

標準偏差
分散について平方根を求める

228

不偏分散

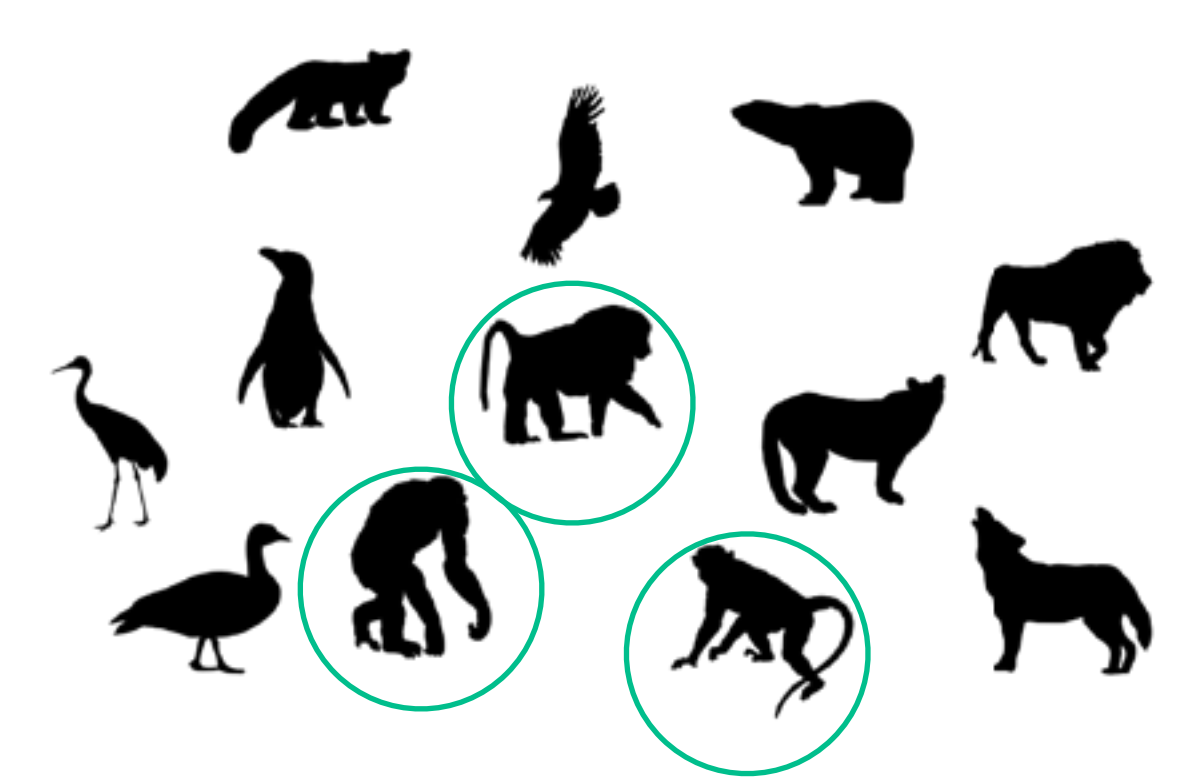
データの数 -1で割る
52000

分布の可視化

分布を視覚化する

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
度数の分布を表形式にまとめたもの… 度数分布表



この図では霊長類は3

```
R df_animal$taxon
#> [1] "食肉類" "鳥類" "食肉類" "鳥類" "霊長類" "霊長類"
#> [7] "霊長類" "食肉類" "齧歯類" "食肉類" "鳥類" "偶蹄類"
#> [13] "食肉類" "食肉類" "鳥類" "食肉類" "霊長類" "鳥類"
#> [19] "鯨偶蹄類" "奇蹄類" "齧歯類" "鯨偶蹄類"
```

taxon	frequency
食肉類	7
鳥類	5
霊長類	4
鯨偶蹄類	2
齧歯類	2
偶蹄類	1
奇蹄類	1

すべての分類群について度数をまとめる

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

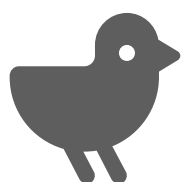


離散変数

サイコロの出目など

サイコロの出目	frequency
1	19
2	15
3	18
4	11
5	16
6	21

値を階級として直接用いる



連続変数

動物の体重など

class	frequency
(2000,3000]	11
(3000,4000]	159
(4000,5000]	111
(5000,6000]	59
(6000,7000]	2

適当な範囲を階級に用いる

```
R> weight_freq <-  
  table(cut(penguins$body_mass_g,  
            breaks = seq(2000,  
                          7000,  
                          by = 1000),  
            dig.lab = 4))  
tibble::tibble(  
  class = names(weight_freq),  
  frequency = weight_freq)
```


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

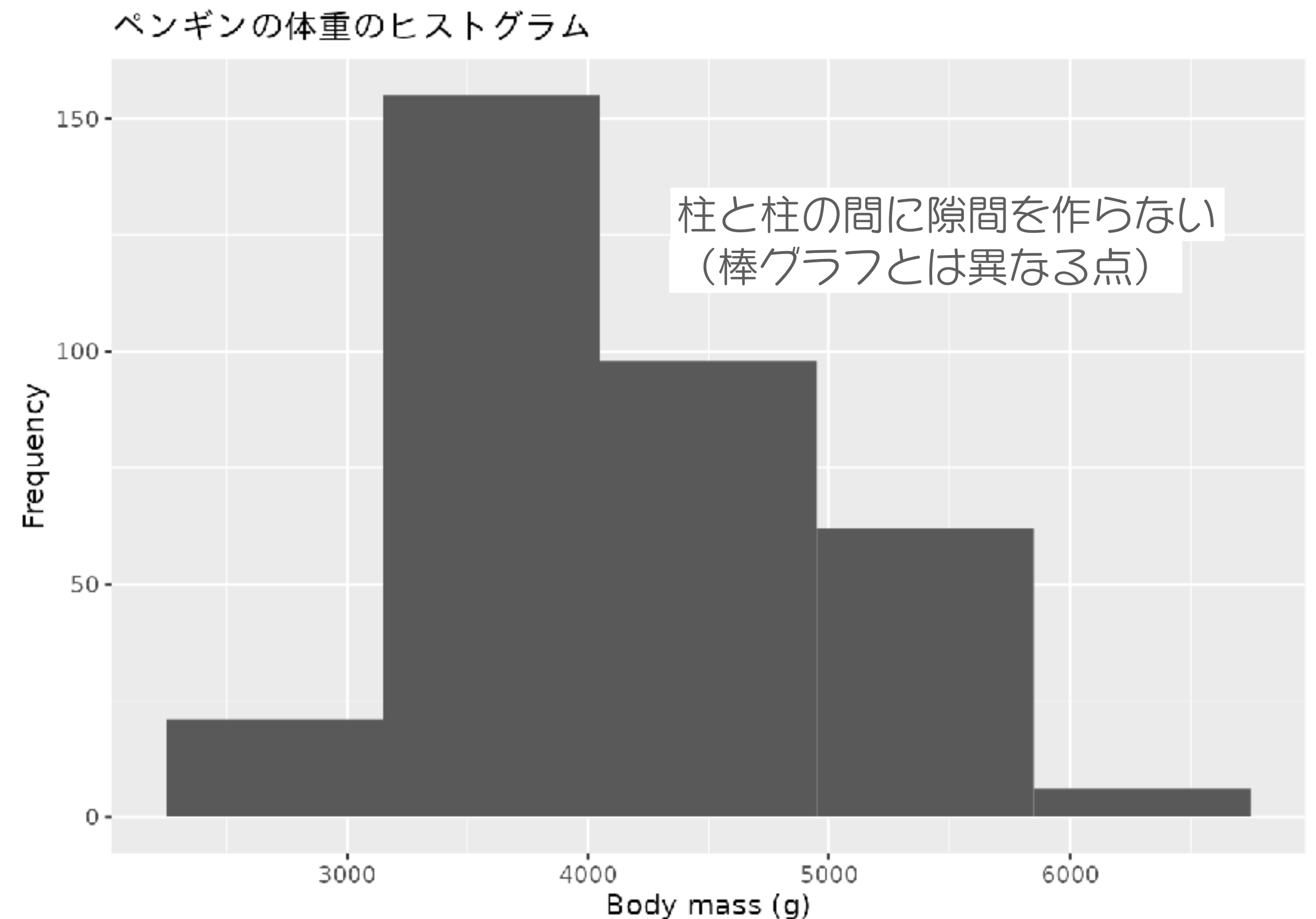
ヒストグラム(histogram)

度数分布表をもとにグラフを作成

階級ごとに柱を設け、柱の高さで度数を表現

class	frequency
(2000,3000]	11
(3000,4000]	159
(4000,5000]	111
(5000,6000]	59
(6000,7000]	2

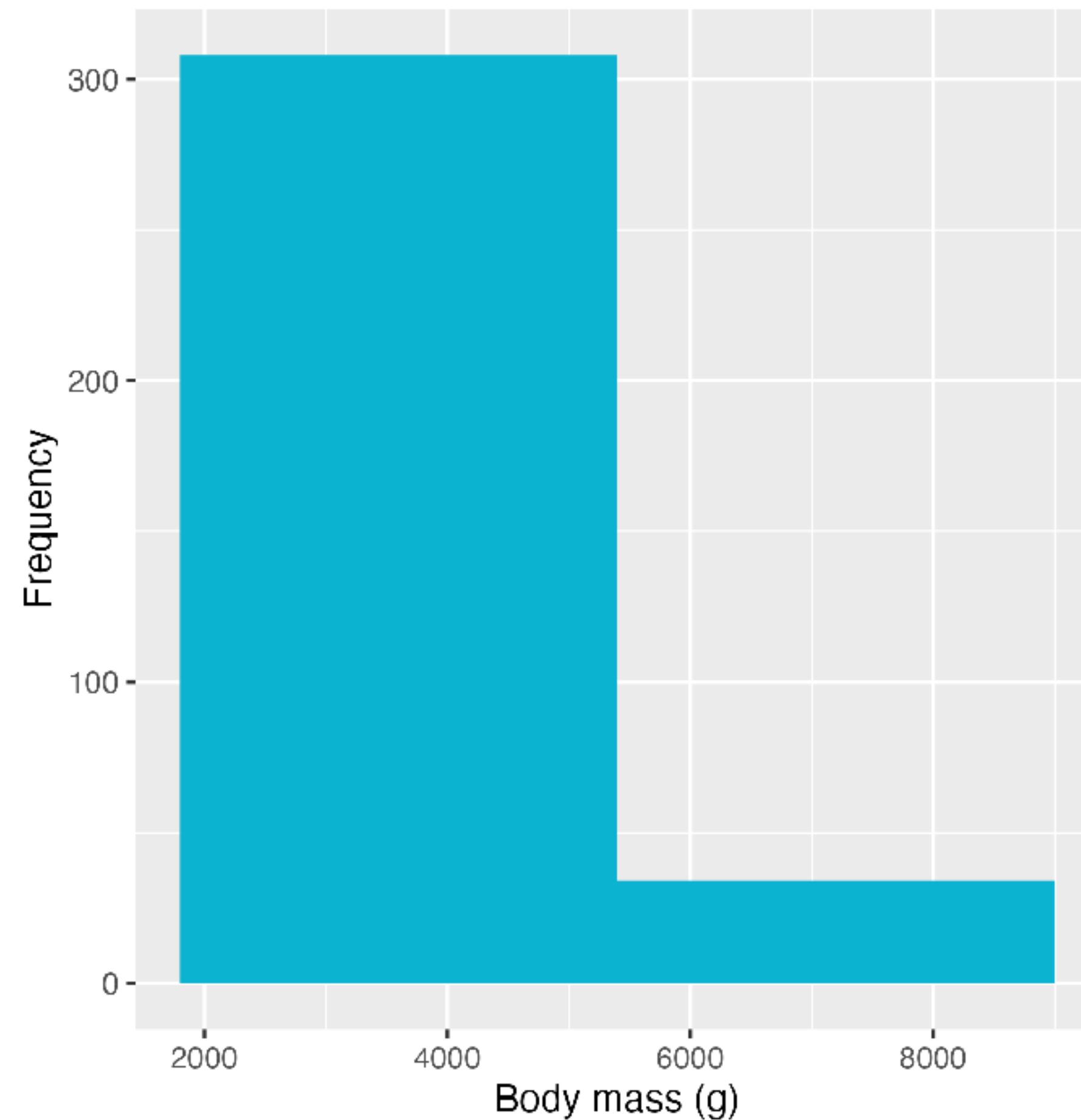
```
 penguins |>  
  ggplot(aes(body_mass_g)) +  
  # ヒストグラムでは柱の階級をビン bin と呼ぶ  
  
  geom_histogram(bins = 5) +  
  ylab("Frequency") +  
  xlab("Body mass (g)") +  
  labs(title = "ペンギンの体重のヒストグラム")
```



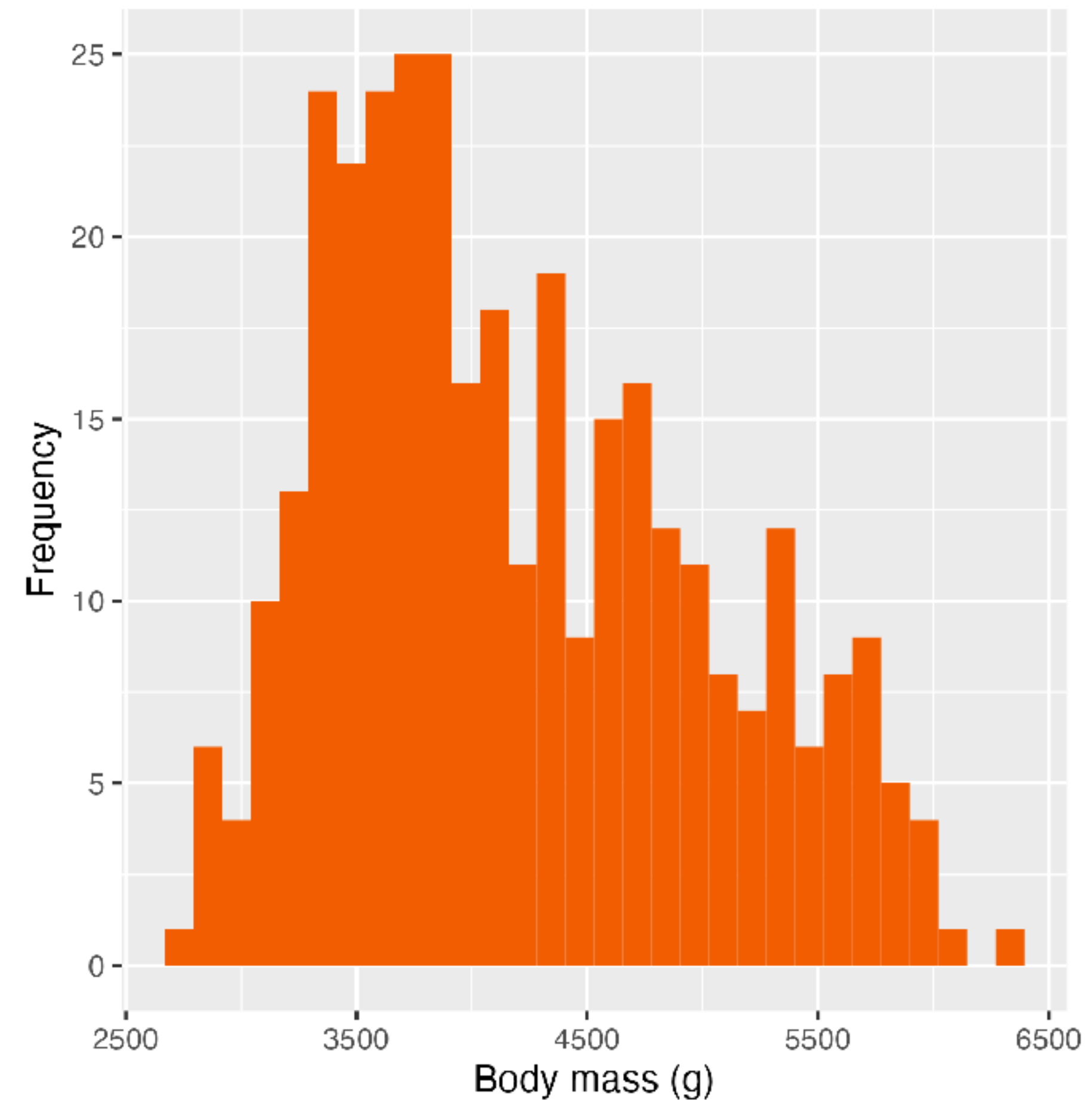
ヒストグラムの形いろいろ

ヒストグラムの階級数が異なると分布の形も変化することがある

ペンギンの体重のヒストグラム。階級数2



ペンギンの体重のヒストグラム。階級数30

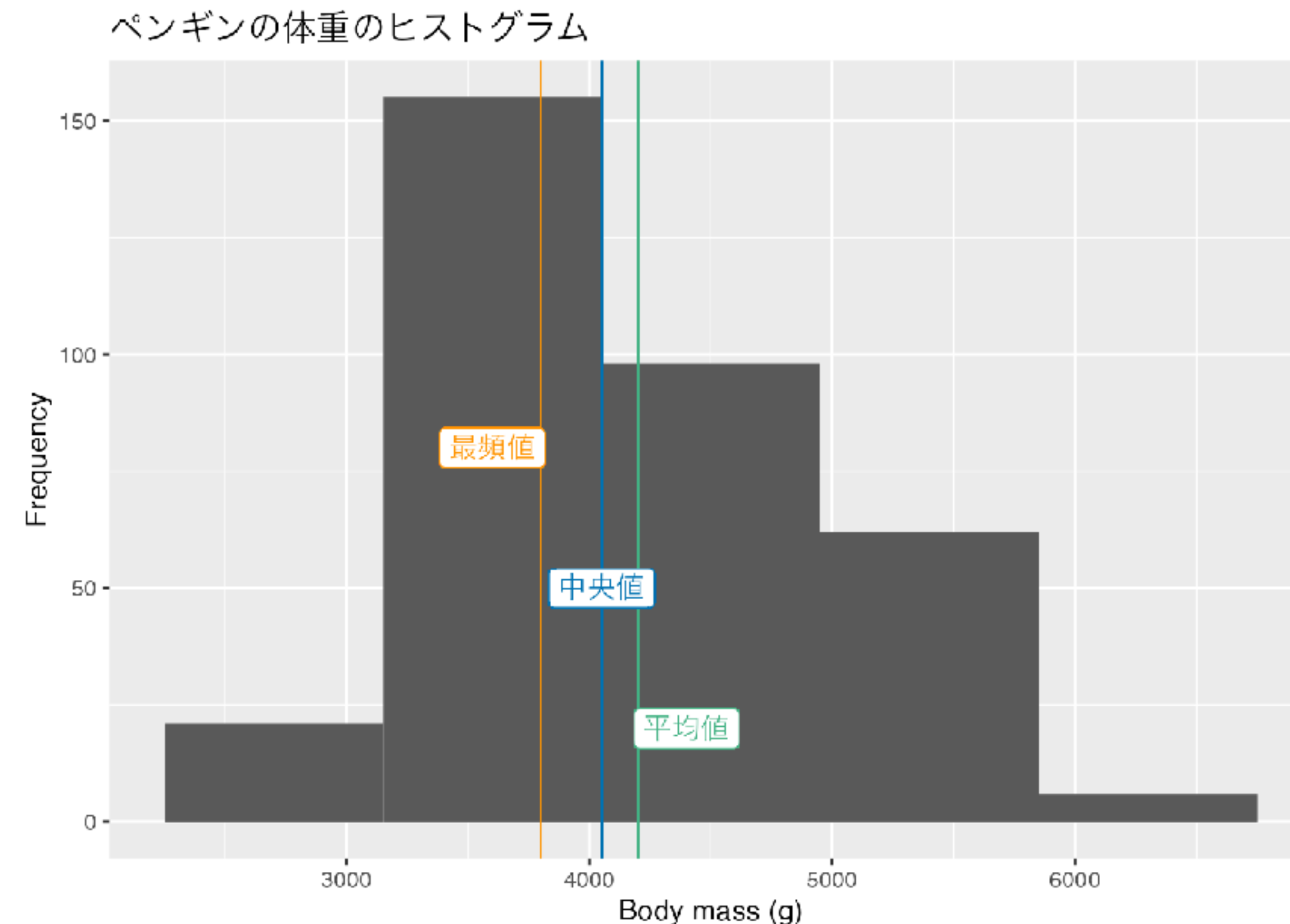


ヒストグラムの形いろいろ

データのばらつきに応じてデータの分布も異なる

右に裾（尻尾）が長い分布… ロングテール型

☞ 代表値が小さい方から最頻値、中央値、平均値の順に並ぶ



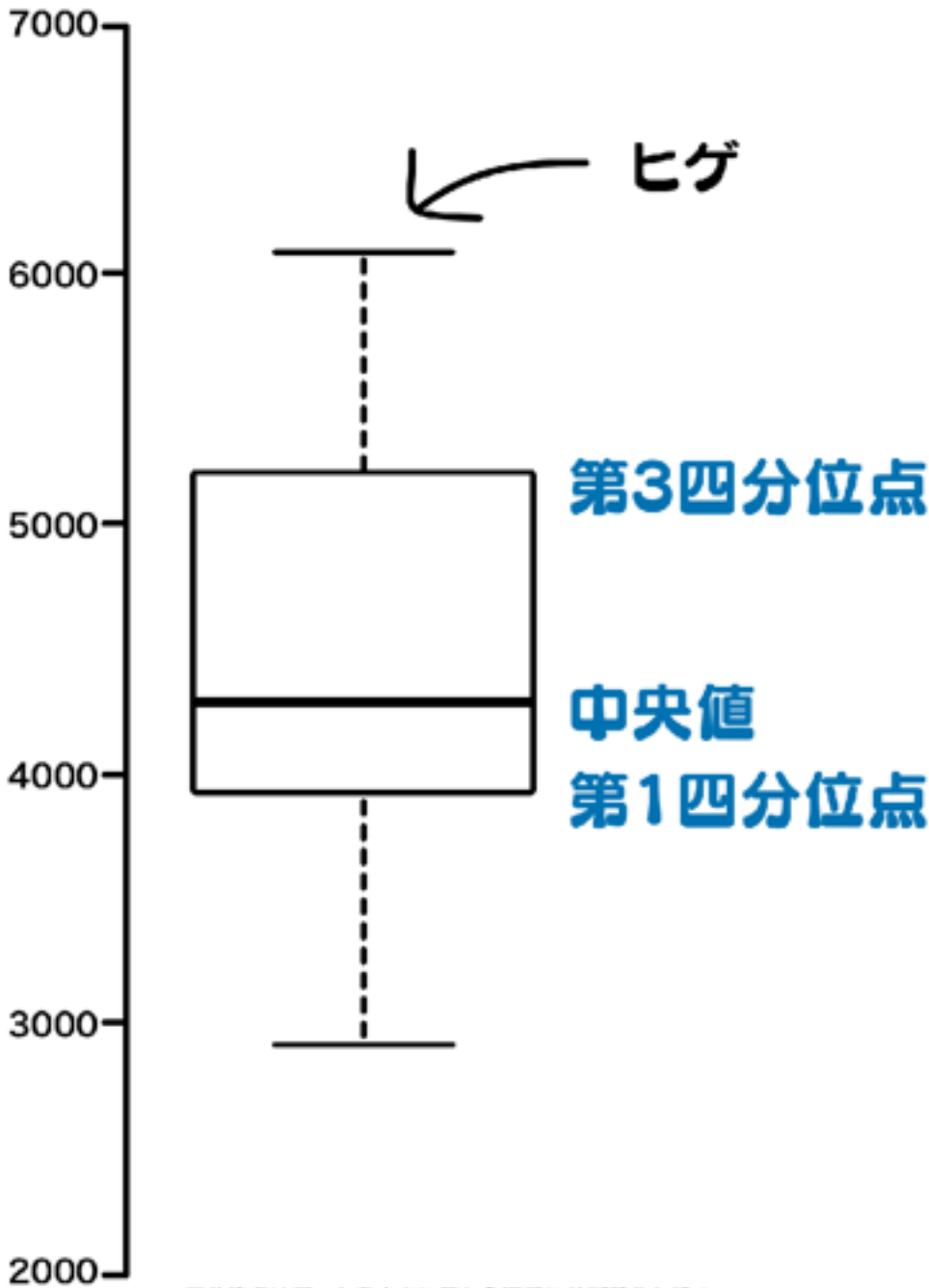
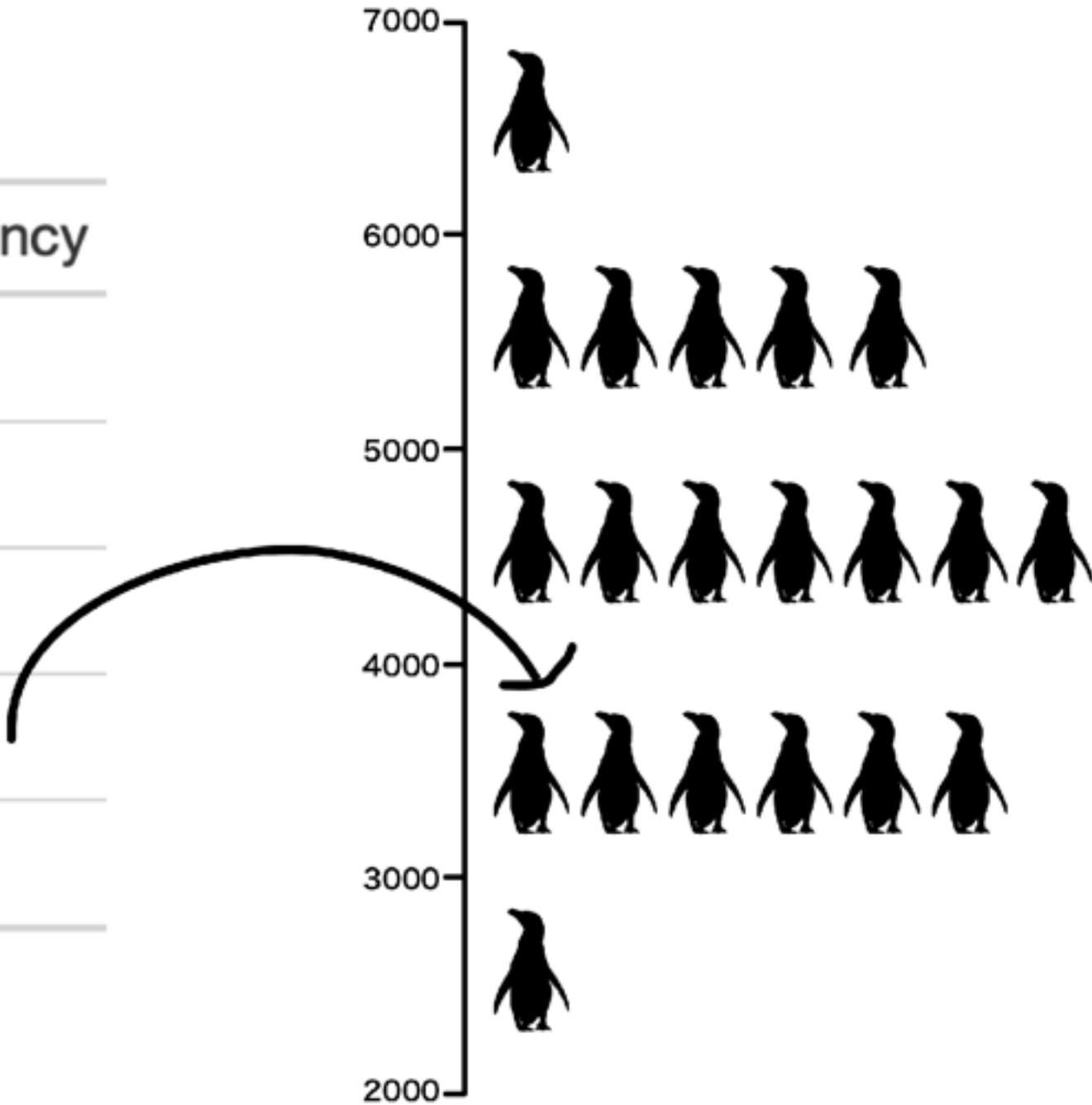
箱ヒゲ図

「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均等な数の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

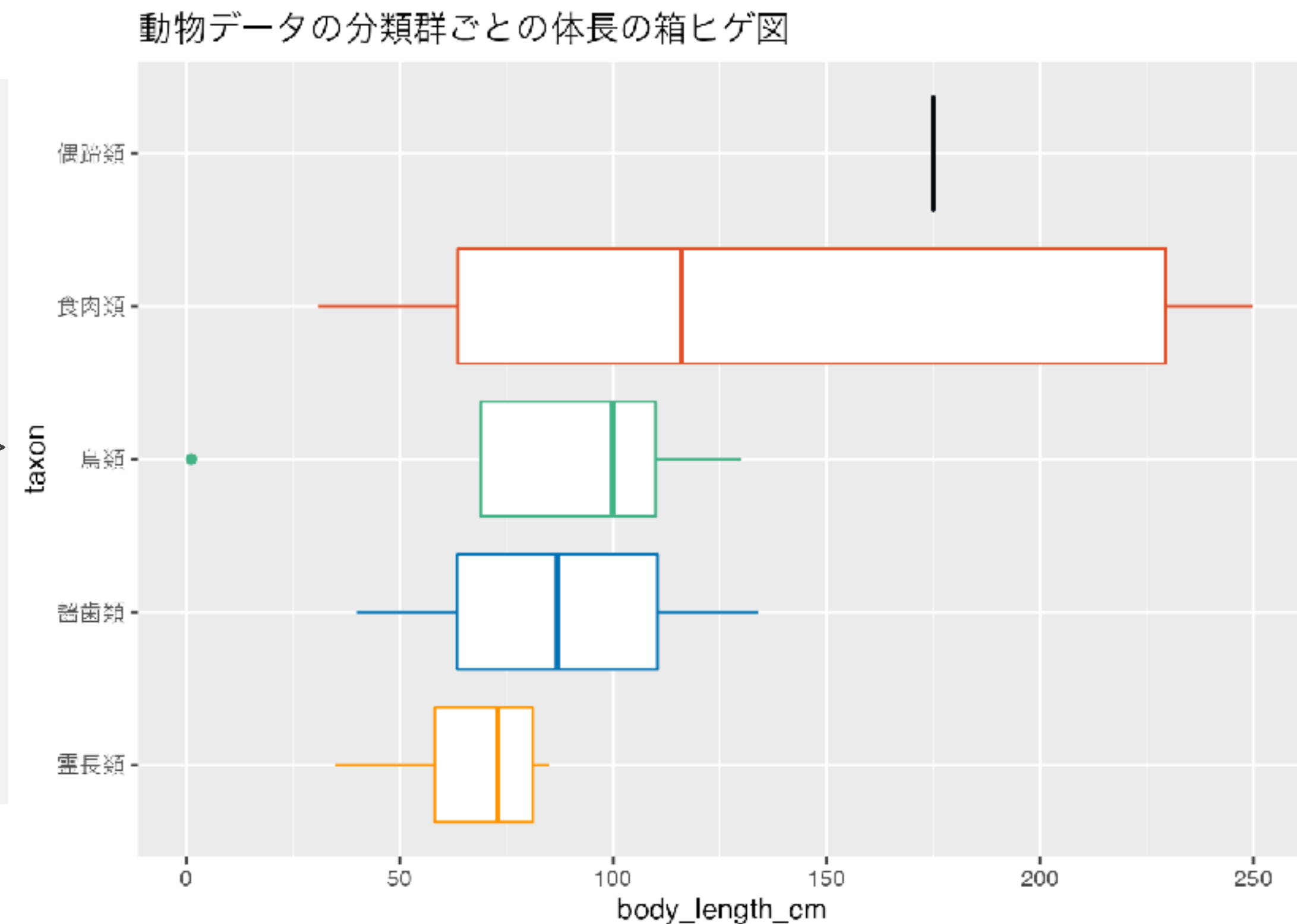
箱ヒゲ図

複数データのばらつきを比較する際にも有効

箱はデータの散らばりが小さい場合に小さく、散らばりが大きい時には大きくなる



```
df_animal |>
  filter(!is.na(body_length_cm)) |>
  group_by(taxon) |>
  mutate(body_length_median = median(body_length_cm)) |>
  ungroup() |>
  mutate(taxon = forcats::fct_reorder(taxon, body_length_median)) |>
  ggplot(aes(taxon, body_length_cm, color = taxon)) +
  geom_boxplot() +
  coord_flip() +
  scale_colour_tokupon() +
  guides(color = "none") +
  labs(title = "動物データの分類群ごとの体長の箱ヒゲ図")
```



参考資料・URL

- 📖 東京大学教養学部統計学教室（編）『基礎統計学I: 統計学入門』（1991）
東京大学出版会. ISBN: 4-13-042065-8
瓜生居室: あり、徳大図書館: あり、市立図書館: なし、県立図書館: あり
- 📖 Peter Bruce, Andrew Bruce, Peter Gedeck (著), 黒川利明 (訳), 大橋真也 (技術監修)
『データサイエンスのための統計学入門：予測、分類、統計モデリング、統計的機械学習とR/Pythonプログラミング』（2020）オライリー・ジャパン. ISBN: 978-4-87311-926-7
瓜生居室: あり（電子版第一版）、徳大図書館: あり（第一版）、市立図書館: なし、県立図書館: あり
- 📖 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（編）『データサイエンスの歩き方』（2022）学術図書出版社. ISBN: 978-4-7806-0936-3
瓜生居室: あり（電子版）、徳大図書館: あり、市立図書館: なし、県立図書館: なし

https://uribo.github.io/tokupon_ds/

https://github.com/uribo/cue2022aw_r104

動物のシルエットはPHYLOPIC <https://www.phylopic.org/> が
クリエイティブ・コモンズライセンスで提供するものです。

