# データの代表値と散布度

樋口さぶろお　https://hig3.net

龍谷大学理工学部数理情報学科

確率統計☆演習 I　L02(2019-09-30 Mon)

最終更新: Time-stamp: "2019-10-01 Tue 08:33 JST hig"

## 今日の目標

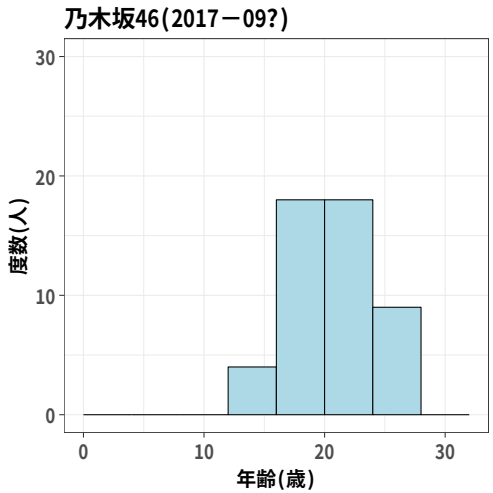
- 代表値:中央値, 四分位数, 平均値, 最頻値を求められる 岩薩林 確率・統計 §1.2 高校 数学 I
- 散布度:レンジ, 四分位範囲, 分散, 標準偏差を求められる 岩薩林 確率・統計 §1.2 高校 数学 I

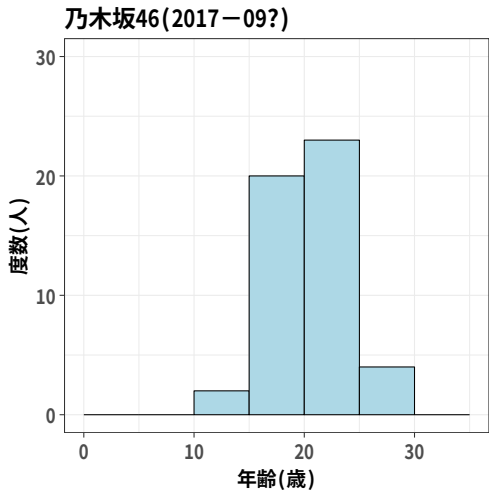
# L01-Q1

## Quiz 解答:度数分布表とヒストグラムを作ろう

度数分布表略.

例



# ここまで来たよ

1. データの分布

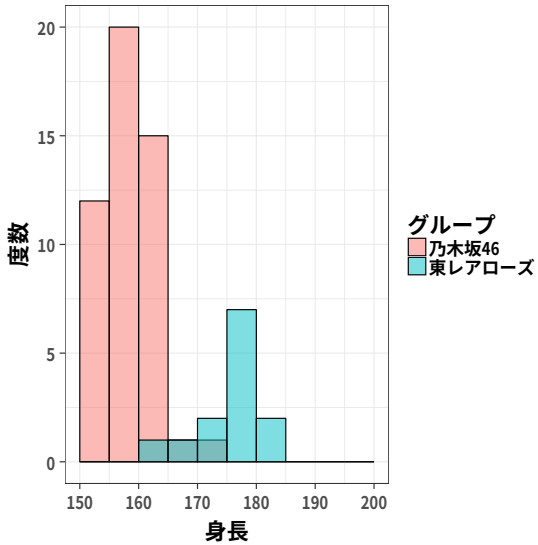
2. データの代表値と散布度
   - 中央値と四分位数
   - 平均値
   - レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
   - 分散・標準偏差

# 代表値:データを 1 個の値で代表させたい!

縮約値=代表値集団の身長はだいたい 150cm? 170cm?

| |
|---|
| 01 171cm |
| 02 166cm |
| 03 165cm |
| ⋮ |
| 49 151cm |

| |
|---|
| 01 179cm |
| 02 183cm |
| 03 182cm |
| ⋮ |
| 13 171cm |



# 今日やる様々な表現方法

| | 分位数タイプ 岩薩林 確率・統計 (少) | 平均タイプ 岩薩林 確率・統計 §1.2 | 岩薩林 確率・統計 (少) |
|---|---|---|---|
| 代表値 | 中央値, 四分位数 | 平均値 | 最頻値 (離散データの，ヒストグラムの) |
| 散布度 | 範囲=レンジ，四分位範囲=IQR | 分散, 標準偏差 | |

これらをデータ, 度数分布表, ヒストグラム (, 箱ひげ図) から読み取る

代表値・散布度 $\lesssim$ 箱ひげ図 $\gtrless$ ヒストグラム $\simeq$ 度数分布表 $<$ ストリップチャート $<$ (生) データ

情報が少ない, アバウト　$\leftrightarrow$　情報が多い, 詳しい

見やすい・直観的　$\leftrightarrow$　見にくい・直観に訴えない

# 中央値 (median) と四分位数/値/点 (quartile)

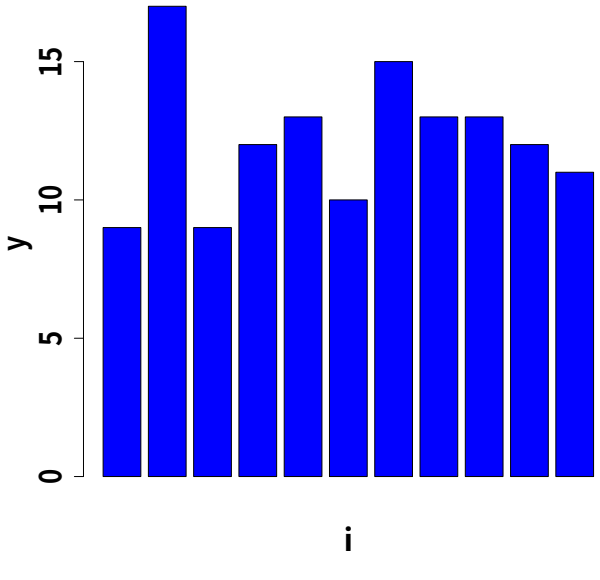
身長 $y_i$ のデータ ($n$ 個) を小さい順に並び替えたものを,
$x_0 \leq x_1 \leq \cdots \leq x_{n-1}$ とする.

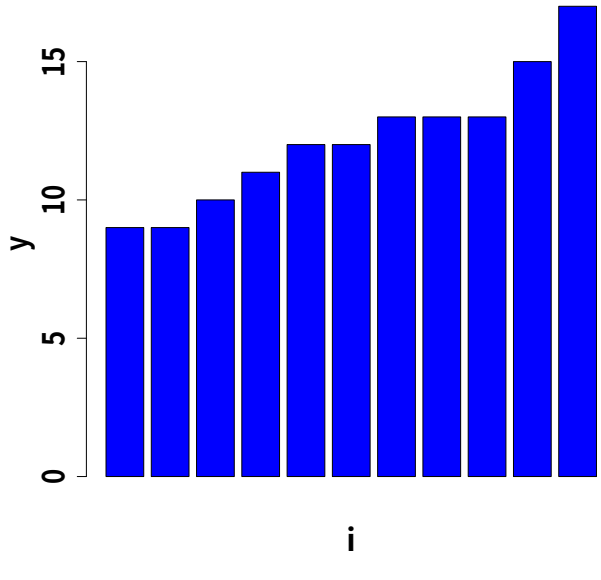
例 $n = 11$

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 9 | 17 | 9 | 12 | 13 | 10 | 15 | 13 | 13 | 12 | 11 |

$\rightarrow$

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 9 | 9 | 10 | 11 | 12 | 12 | 13 | 13 | 13 | 15 | 17 |




$\rightarrow$ 順にならべる

# 中央値 (median) と四分位数/値/点 (quartile)

## 分位数, 四分位数のアバウトな定義

- $q$-分位数 $= x_{q\cdot(n-1)}$. $(0 \leq q \leq 1)$.
- 最小値 $Q_0 = x_0 = x_{\frac{0}{4}\cdot(n-1)} =$ (0.0-分位数).
- 第 1 四分位数 $Q_1 = x_{\frac{1}{4}\cdot(n-1)} =$ (0.25-分位数).
- 第 2 四分位数 $Q_2 = x_{\frac{2}{4}\cdot(n-1)}$=中央値=メディアン $=$ (0.5-分位数).

  岩薙林 確率・統計練習問題 1-1
- 第 3 四分位数 $Q_3 = x_{\frac{3}{4}\cdot(n-1)} =$ (0.75-分位数).
- 最大値 $Q_4 = x_{\frac{4}{4}\cdot(n-1)} =$ (1.0-分位数).

## 高校数学における四分位数の定義 高校 数学 I

- $Q_0, Q_4$ さっきのまま.
-

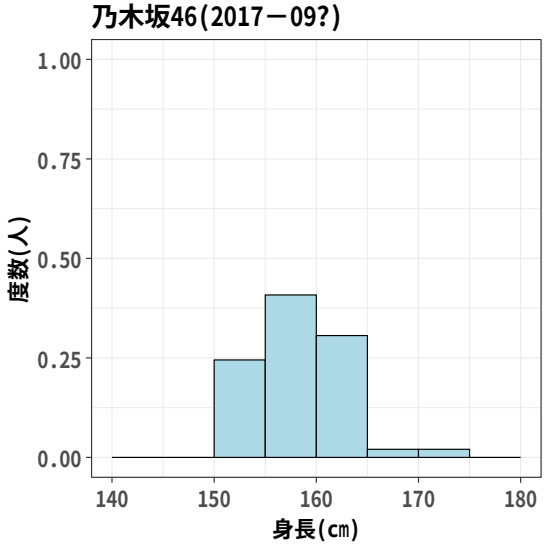
$$Q_2 = \begin{cases} x_{\frac{n-1}{2}} = \boxed{\phantom{00}} & (n \text{ が奇}) \\ \frac{1}{2}\left(x_{\frac{n}{2}-1} + x_{\frac{n}{2}}\right) = \boxed{\phantom{00}} & (n \text{ が偶}) \end{cases}$$

- $Q_1$ は, $Q_2$ の位置より前にあるデータ ($Q_2$ 自身は除く) の中央値
- $Q_3$ は, $Q_2$ の位置より後にあるデータ ($Q_2$ 自身は除く) の中央値

$Q_2$ と同じ値のデータが複数あるときも 1 個だけ除く

例: 9 9 10 11 12 12 13 13 13 15 17
ちょっと変えた例: 10 11 12 12 13 13 13 15 17

# 相対度数のヒストグラムからの $q$-分位数の求め方




全体の面積は 1

# ここまで来たよ

1. データの分布

2. データの代表値と散布度
   - 中央値と四分位数
   - 平均値
   - レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
   - 分散・標準偏差

# 平均 (値)=mean

## 平均 (値) の定義

岩薩林 確率・統計 (1.1)p.5

$n$ 個のデータ $x_1, x_2, \ldots, x_n$ に対して,

$$\text{平均値} \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\overline{x}$ のかわりに $m$, $m_{\mathrm{x}}$ などという記号もある.

岩薩林 確率・統計例題 1.2(p.5)

中央値より平均値のいい点

平均値より中央値のいい点

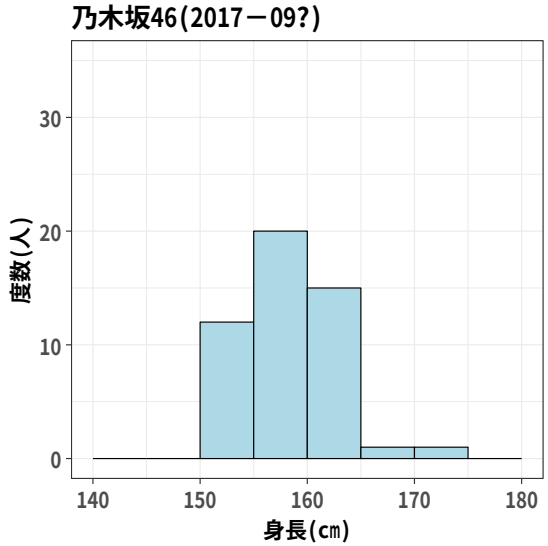
## 度数分布表からの平均値の求め方

岩薙林 確率・統計 §1.2.3, (1.14)p.17

$j$ 番目の階級の階級値 $x_j$, 度数 $f_j$. 近似的に,

$$\overline{x} \approx \frac{1}{n}\sum_{j=1}^{k} x_j f_j = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j}$$

## ヒストグラムからの平均値の求め方




### 力学のりで

$k$ 個の質点の重心の座標 $x_{\mathrm{G}} = \dfrac{\sum_{j=1}^{k} x_j m_j}{\sum_j m_j}$ 力学

$j$ 番目の質点の位置 $x_j$ = 階級値, 質量 $m_j$ = $f_j$

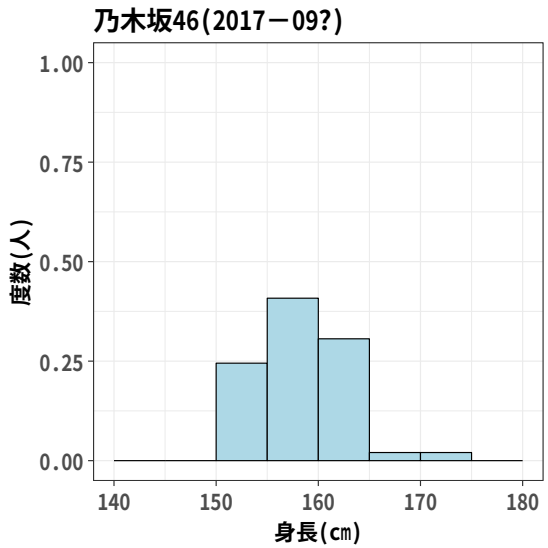
# 最頻値=mode 岩薙林 確率・統計練習問題 1-2

## 最頻値の定義

- 離散データの最頻値: ‘離散的な’ データのとき いちばん多く繰り返し現れる値
- ヒストグラムの最頻値: ‘連続的または離散的な’ データのとき 度数分布表/ヒストグラムで, 度数最大の階級の階級値

離散データの最頻値 (30 50 55 55 60 70 70 70 75 100) だと□

ヒストグラムの最頻値

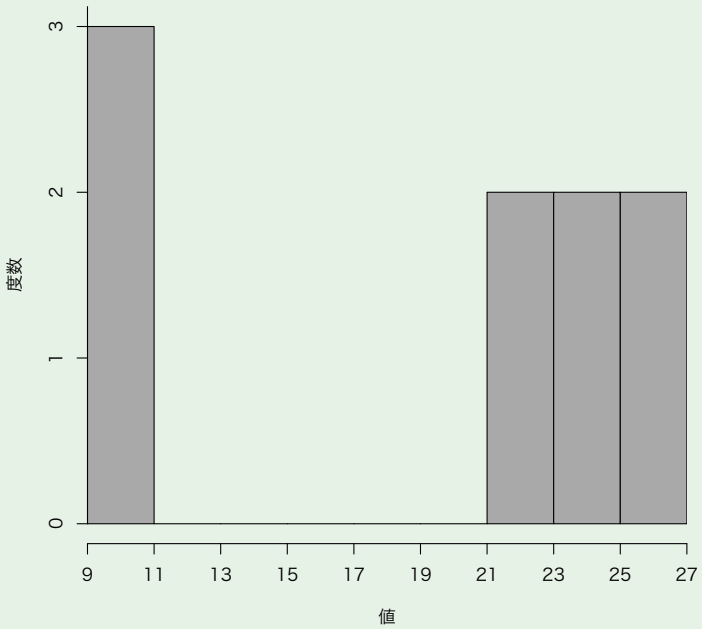


L02-Q1

## Quiz(平均値中央値最頻値)

次の代表値を, 下のヒストグラムから求めよう.

1. 中央値
2. (ヒストグラムの) 最頻値
3. 平均値

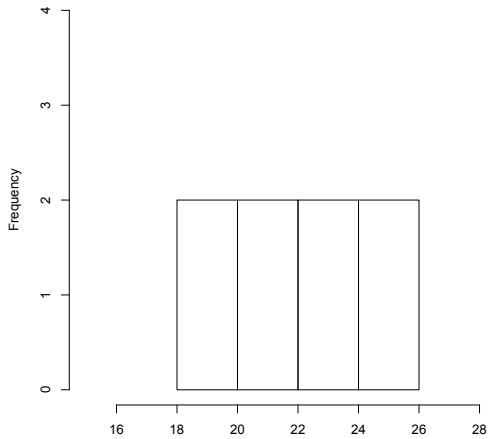


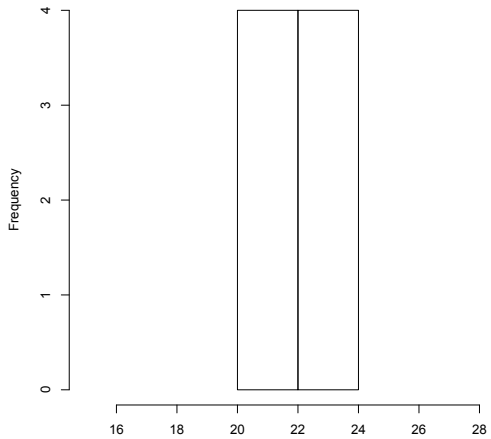
# ここまで来たよ

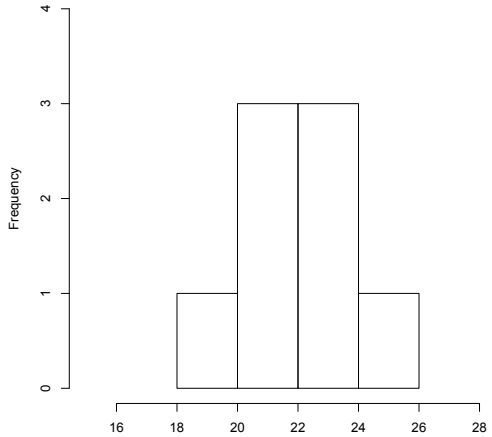
1. データの分布

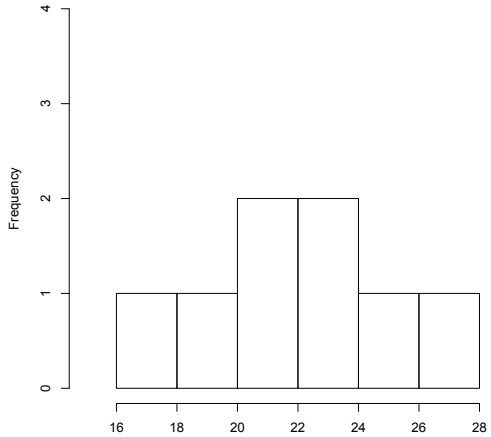
2. データの代表値と散布度
   - 中央値と四分位数
   - 平均値
   - レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
   - 分散・標準偏差

# 平均値が同じでも分布はいろいろ




第 1,3 四分位数は?

# 散布度:散らばりの尺度が必要

# レンジ・四分位範囲の定義 I

## 範囲タイプの量の定義 高校 数学 I 岩薩林 確率・統計なし

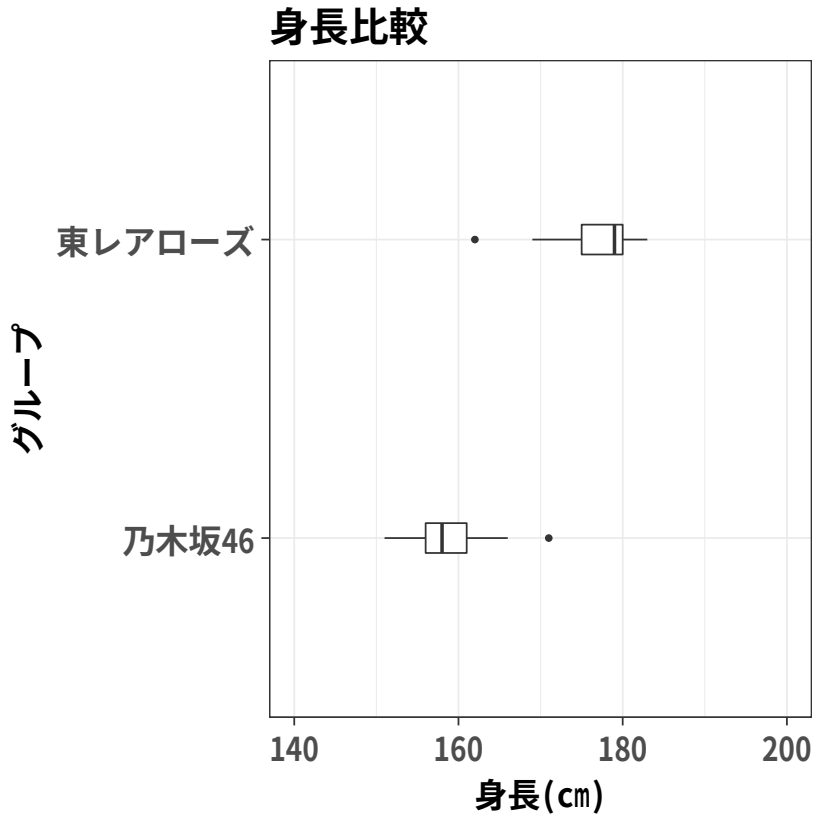
- 範囲 (レンジ) =
- 四分位範囲 (interquartile range) IQR=

L02-Q2

## Quiz(範囲)

次のデータの, 範囲, 四分位範囲, 四分位偏差 を求めよう.
14 14 15 16 18 18 18 25

# 箱ひげ図 (Box Plot, Box and Whisker diagram)




最小最大値 $Q_0, Q_4$, 四分位点 $Q_1, Q_2, Q_3$

**箱ひげ図を描く手順**高校 数学 I

- $Q_0, Q_4$ $Q_1$,$Q_2$,$Q_3$ と平均値 $\bar{x}$ を求める
- $Q_2$ に縦線をいれる
- $Q_1$,$Q_3$ を左右の端として箱を描く
- $Q_0$,$Q_4$ に短い縦線をいれ, 点線のひげで箱とつなぐ
- (平均値に $+$ を 1 個描く)
- (「外れ値を○で描く」)

外れ値=Q1,Q3 から IQR の $\alpha = 1.5$ 倍より離れているデータ

https://www.geogebra.org/m/dbfbkews

# ここまで来たよ

1. データの分布

2. データの代表値と散布度
   - 中央値と四分位数
   - 平均値
   - レンジ (範囲,range)・四分位範囲 (IQR)・箱ひげ図
   - 分散・標準偏差

# 分散・標準偏差の定義

高校 数学 I | 岩薙林 確率・統計 (1.4),(1.5)

データ: $x_1, x_2, \ldots, x_n$.

## 分散タイプの量の定義

- データの分散 (variance)

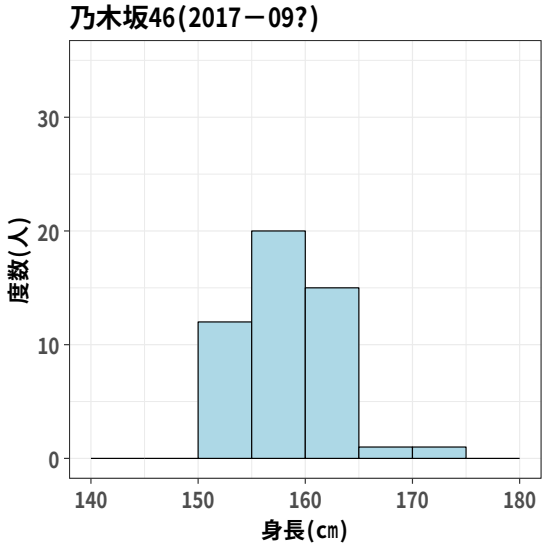
$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- データの標準偏差 (standard deviation)$= S = \sqrt{S^2} \geq 0$

岩薙林 確率・統計定理 1.1(p.6)

$$0 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^1$$

# (例) グループ (49 人) の身長 I




$n-1=49-1$ で割りたくなった人もいるかも．ここは 49 で OK

そのうちちゃんと区別を説明します．

データの単位 $\neq$ 分散の単位

- 平均値 $\overline{x}=\frac{171+166+165+\cdots+151}{49}=158.7(\text{cm})$
- 分散 $S^2=\frac{(171-158.7)^2+(166-158.7)^2+\cdots+(151-158.7)^2}{49}=17.7\ (\text{cm}^2)$
- 標準偏差 $S=\sqrt{17.7}=4.21\ (\text{cm})$

大注意: 平均値 158.7 cm を 159 や 160 に四捨五入すると，

□に加えて□の危険

数値計算法

## 度数分布表からの分散・標準偏差の求め方

高校 数学 I | 岩薩林 確率・統計 §1.2.3,(1.14)p.17

$$S^2 = \frac{1}{n}\sum_j (x_j - \overline{x})^2 f_j = \frac{\sum_j (x_j - \overline{x})^2 f_j}{\sum_j f_j}$$

.

## ヒストグラムからの標準偏差の読み取り方

## 力学のりでの分散の求め方

質点系の慣性モーメント $I = \frac{\sum_{j=1}^{k} (x_j - x_G)^2 m_j}{\sum_j m_j}$

力学

$j$ 番目の質点の位置 $x_j$, 質量 $m_j$

## 分散公式 (便利な (こともある) 計算方法)

高校 数学 I | 岩薩林 確率・統計定理 1.2(p.10)

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2$$

岩薩林 確率・統計例題 1.3(p.7), 岩薩林 確率・統計例題 1.4(p.11)

L02-Q3

## Quiz(平均値・分散・標準偏差)

データ 87kg, 93kg, 89kg, 91kg, 90kg の平均値・分散・標準偏差を求めよう.

# 連絡

- 次回はたぶん 3-202 講義室
- 樋口オフィスアワー木 6(1-539) 金昼 (1-542), Math ラウンジ (1-536/538)
- Trial 予告
- 来週は教科書 岩薙林 確率・統計 SS1.3,1.2.2 読んできて.
- 統計検定. 2019-11-24 日 40%ディスカウント団体受験受付中.

統計検定団体受験 申込は Moodle モバイルアプリ
来週月まで

で URL 指定

https://note.math.ryukoku.ac.jp/moodle
または Safari/Chrome で.