データの代表値と散布度

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 I L02(2018-10-03 Wed)

最終更新: Time-stamp: "2018-10-03 Wed 07:37 JST hig"

今日の目標

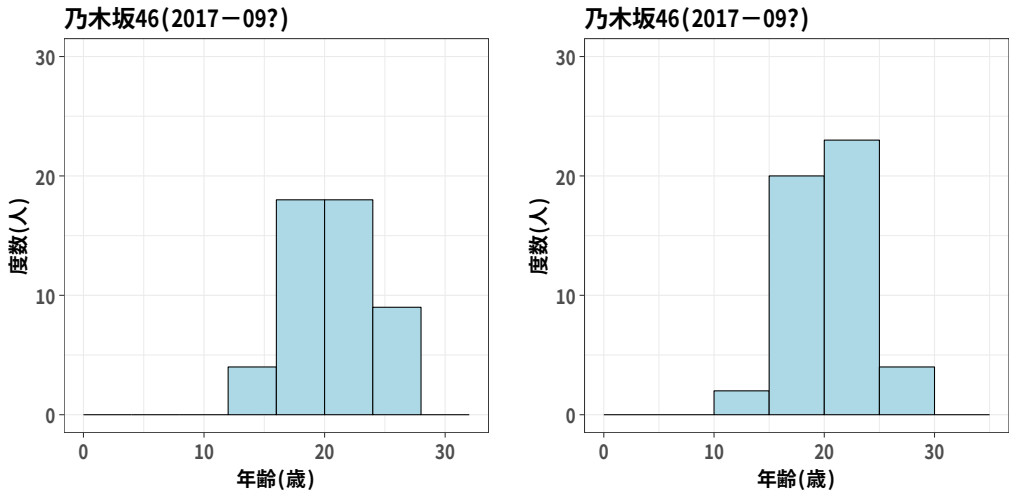
- 代表値:中央値, 四分位数, 平均値, 最頻値を求められる 前園確率統計 §4.1(p.66),§4.2(p.67) 高校 数学 I
- 散布度:レンジ, 四分位範囲, 分散, 標準偏差を求められる 前園確率統計 §4.1(p.66),§4.2(p.67) 高校 数学 I

L02-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

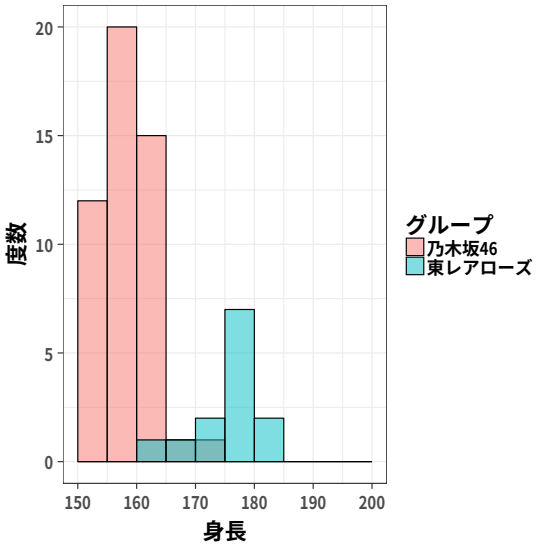
代表値:データを 1 個の値で代表させたい!

前園確率統計 §4.1(p.66)

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	平均タイプ	
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	レンジ，四分位範囲=IQR	分散, 標準偏差	

これらを度数分布表, 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 $\overset{>}{\lesssim}$ 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)

データ $y_0, y_1, \ldots, y_{N-1}$ (N データの個数)
小さい順に並び替えたもの
$\to x_0 \leq x_1 \leq \cdots \leq x_{N-1}$
例 (身長のデータ) $y_0 = 166, y_1 = 153, \ldots, y_{N-1} = 160$
$\to x_0 = 151 \leq x_1 = 152 \leq \cdots \leq 166 \leq x_{N-1} = 167$

分位数, 四分位数のアバウトな定義

前園確率統計 §4.2(p.67)

- q-分位数 $= x_{q\cdot(N-1)}$. $(0 \leq q \leq 1)$.
- 最小値 $Q_0 = x_0 = x_{\frac{0}{4}(N-1)}$
- 第 1 四分位数 $Q_1 = x_{\frac{1}{4}(N-1)}$
- 第 2 四分位数 $Q_2 = x_{\frac{2}{4}(N-1)}$=中央値
- 第 3 四分位数 $Q_3 = x_{\frac{3}{4}(N-1)}$
- 最大値 $Q_4 = x_{\frac{4}{4}(N-1)}$

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.
-

$$Q_2 = \begin{cases} x_{\frac{N-1}{2}} = \boxed{} & (N \text{ が奇}) \\ \frac{1}{2}(x_{\frac{N}{2}-1} + x_{\frac{N}{2}}) = \boxed{} & (N \text{ が偶}) \end{cases}$$

- 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

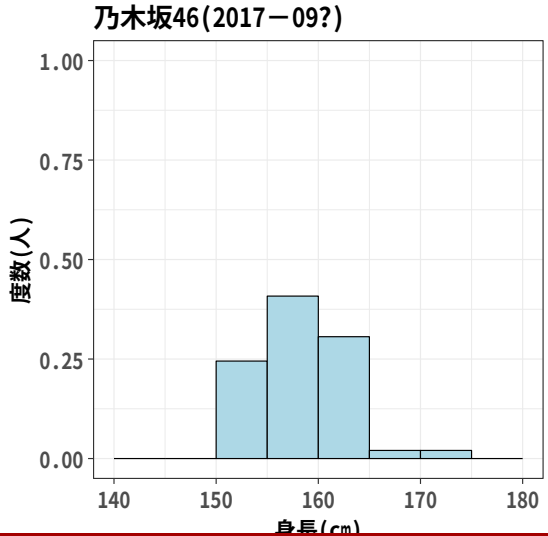
例: 9 9 10 11 12 12 13 13 13 15 17
ちょっと変えた例: 10 11 12 12 13 13 13 15 17

度数分布表からの q 分位数の求め方

階級値=階級の (上限値+下限値)/2

j	階級 (cm)	階級値 z_j	度数 f_j
1	145 より大きく 150 以下	147.5	7
2	150 より大きく 155 以下		17
3	155 より大きく 160 以下		29
4	160 より大きく 165 以下		19
k =5	165 より大きく 170 以下		4
	合計 N=	—	77

ヒストグラムからの q 分位数の求め方



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 前園確率統計 §4.1(p.66)

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_{x} などという記号もある.

例: 30 50 55 55 60 70 70 70 75 100 だと [　　]

中央値より平均値のいい点

[　　]

平均値より中央値のいい点

[　　]

L02-Q2

Quiz(代表値)

次のデータを考える.

14cm, 14cm, 15cm, 16cm, 18cm, 18cm, 18cm, 25cm

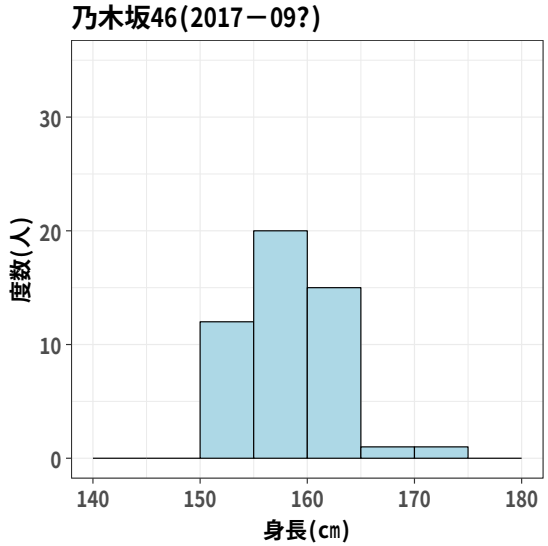
1. 四分位数 Q_1, Q_2, Q_3 を求めよう.
2. (離散データの) 最頻値を求めよう
3. 平均値を求めよう

度数分布表からの平均値の求め方 前園確率統計なし

$$\overline{x} \approx \frac{1}{N}\sum_{j=1}^{k} z_j f_j = \frac{\sum_{j=1}^{k} z_j f_j}{\sum_{j=1}^{k} f_j}$$

j 番目の階級の階級値 z_j, 度数 f_j.

ヒストグラムからの平均値の求め方



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 $x_j = z_j$, 質量 $m_j = f_j$

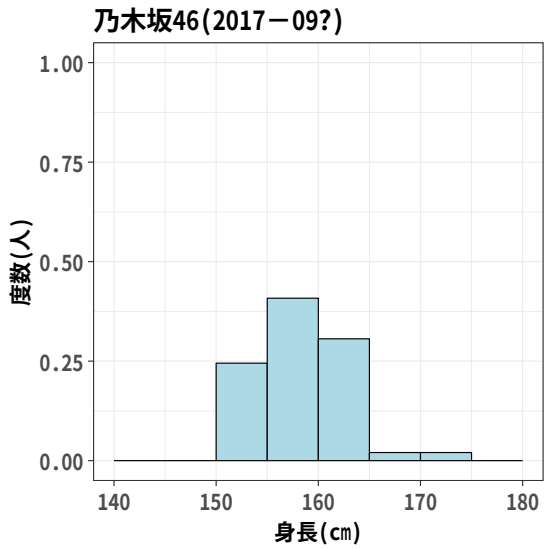
最頻値=mode 前園確率統計なし

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

離散的な例 1(30 50 55 55 60 70 70 70 75 100) だと□

ヒストグラムの最頻値

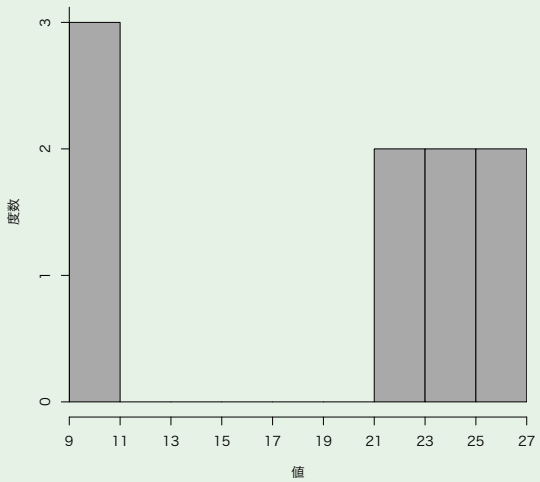


L02-Q3

Quiz(平均値中央値最頻値)

次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



2017 年 6 月統計検定 3 級問 5

2017 年 6 月統計検定 3 級問 5

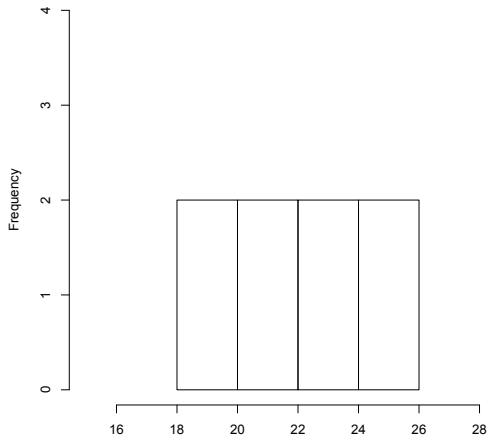
2017 年 6 月統計検定 3 級問 5

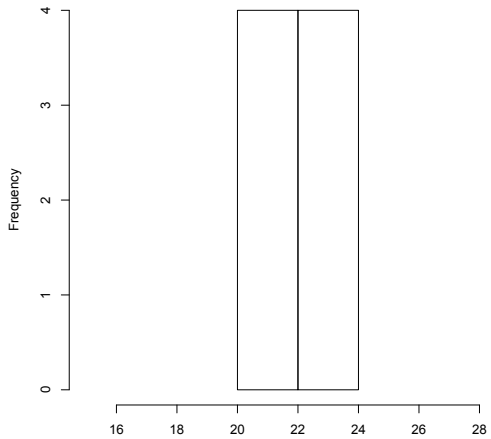
ここまで来たよ

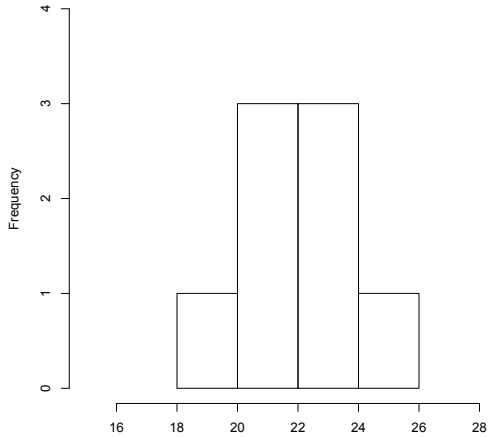
1. データの分布

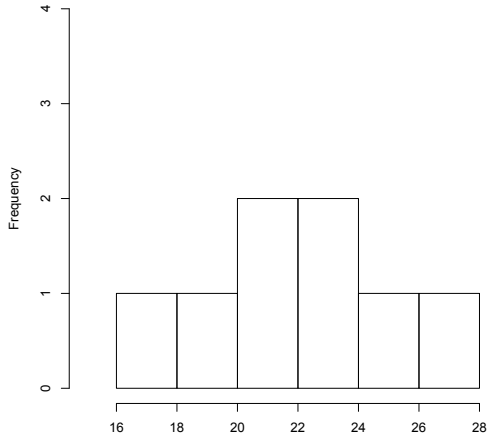
2. データの代表値と散布度
   - 中央値と四分位数
   - 平均値
   - レンジ (範囲,range)・四分位範囲 (IQR)
   - 箱ひげ図
   - 分散・標準偏差

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 前園確率統計なし

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

L02-Q4

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

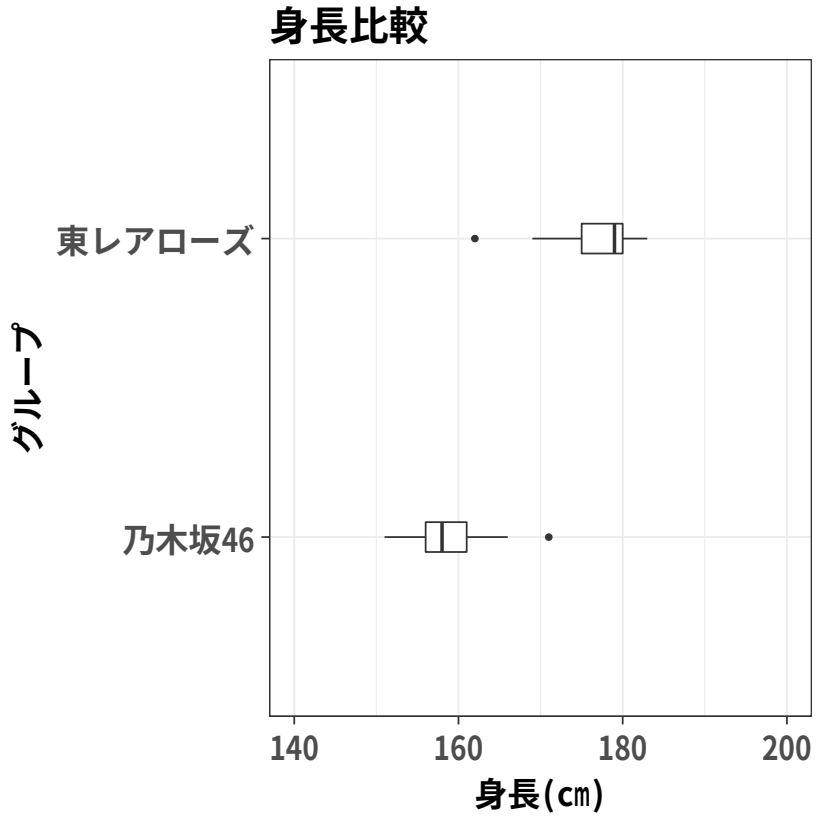
ここまで来たよ

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

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

前園確率統計 §4.2



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」)

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 | 前園確率統計 §4.1(p.66)

データ: $x_1, x_2, \ldots, x_N$.

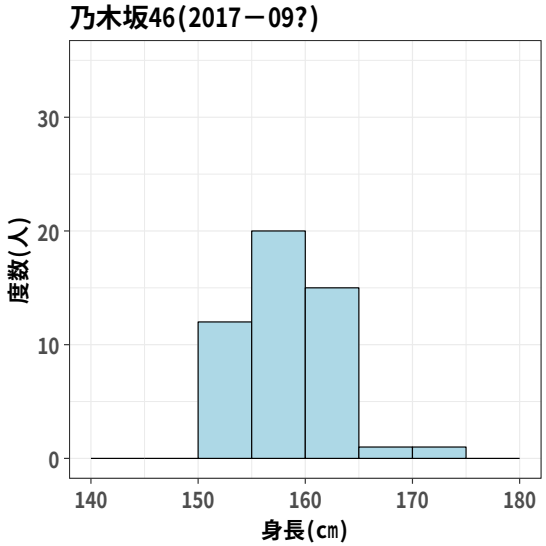
分散タイプの量の定義

- 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)=

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

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

ヒストグラムからの標準偏差の読み取り方

度数分布表からの分散・標準偏差の求め方 高校 数学 I 前園確率統計なし

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

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

i 番目の質点の位置 x_i, 質量 m_j

分散の便利な (こともある) 計算方法 高校 数学 I 前園確率統計なし

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

L02-Q5

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.

2017 年 6 月統計検定 3 級問 11

2017 年 6 月統計検定 3 級問 11

2017 年 6 月統計検定 3 級問 11

連絡

- 次回は 7-002 講義室
- 樋口オフィスアワー火昼 (1-539) 金 14:40-15:40(1-502), Math ラウンジ月-木昼 (1-614)
- Trial 予告
- Learn Math Moodle の予習復習問題で来週の trial に備えてね.
- 来週は教科書 前園確率統計 §4.3 読んできて
- 統計検定. 2018-11-25. 10%ディスカウント団体受験受付中, (– 2018-10-09 火)

過去の 2 年生の受験体験記より

僕は、数学教員を目指しており、数学を専門にするなら統計学の知識はある程度つけておきたいと思ったことと、いろいろと資格に挑戦しようと思い、3 級を受験しました。(略) また、僕は授業を受ける前に検定を受けたのですが、2 年の後期に「確率統計及び演習 I」という授業があり、この授業では 3 級や 2 級に出てくる公式や統計に関する知識を詳しく学ぶことができるので、この授業で検定の対策にするのも良いと思います。(以下略)