データの代表値・散らばりの尺度

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(2017-09-27 Wed)

最終更新: Time-stamp: "2017-10-03 Tue 09:48 JST hig"

今日の目標

- データ, 度数分布表, ヒストグラムから
  - 中央値, 四分位数, 平均値, 最頻値を求められる 西川確率統計 5.1.2, 5.1.6 高校 数学 I
  - レンジ, 四分位範囲, 四分位偏差, 分散, 標準偏差を求められる 西川確率統計 5.1.3, 5.1.6 高校 数学 I

http://hig3.net

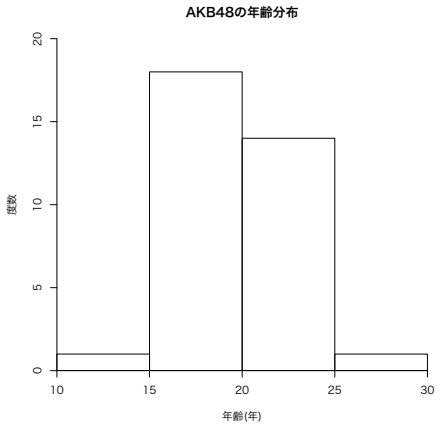
L01-Q1

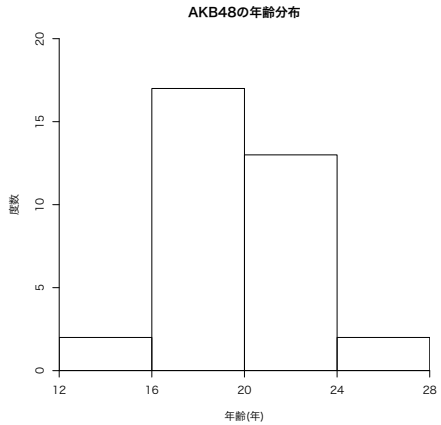
Quiz 解答:度数分布表とヒストグラムを作ろう

階級 (歳)	度数
10 より大きい 15 以下	1
15 より大きい 20 以下	18
20 より大きい 25 以下	14
25 より大きい 30 以下	1
計	34

階級 (歳)	度数
12 より大きい 16 以下	2
16 より大きい 20 以下	17
20 より大きい 24 以下	13
24 より大きい 28 以下	2
計	34

このデータの場合はたまたま, 以上未満でも同じ.





たまたま形が似たけど, 階級の取り方でヒストグラムの形は変化する.

ここまで来たよ

3. 略解:データの分布

4. データの代表値・散らばりの尺度
   - 中央値と四分位数
   - 最頻値
   - 平均値
   - レンジ (範囲,range)・四分位偏差
   - 分散・標準偏差・平均偏差

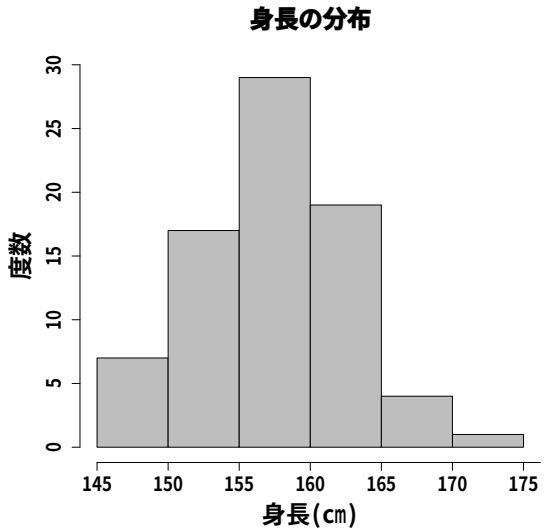
代表値:データを 1 個の値で代表させたい!

西川確率統計 5.1.2

縮約値=代表値某国民的アイドル集団の身長はだいたい 150cm? 170cm?

データ全体 148 152 . . . 170

階級 (cm)	度数 f_i
145 より大きく 150 以下	7
150 より大きく 155 以下	17
155 より大きく 160 以下	29
160 より大きく 165 以下	19
165 より大きく 170 以下	4
170 より大きく 175 以下	1
合計	77



今日やる様々な表現方法の間の変換

		箱ひげ図	ヒストグラム	度数分布表	(生)データ
代表値	中央値 (+四分位数) 平均値 最頻値 (ヒストグラム, データの)				
散らばりの尺度	レンジ, 四分位偏差, IQR 分散, 標準偏差, 平均偏差 —				

見やすい・直観的 $\leftrightarrow$ 詳しい・正確

代表値・散らばりの尺度 $\lesssim$ 箱ひげ図 $<$ ヒストグラム $\simeq$ 度数分布表 $<$ (生) データ

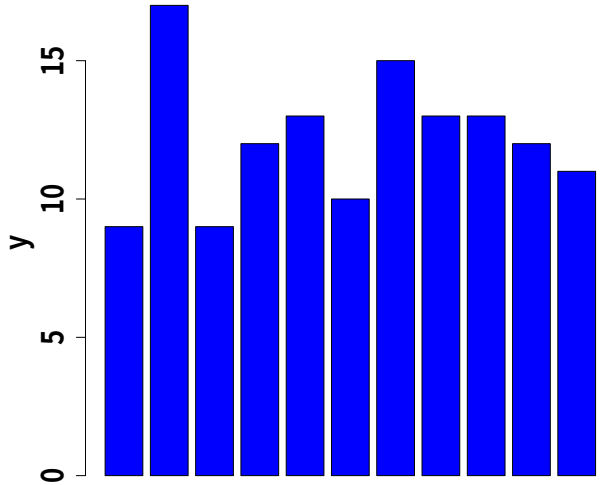
中央値 (median) と四分位数/値/点 (quartile)

身長 x のデータを小さい順に並び替えたものを,
$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$ とする.

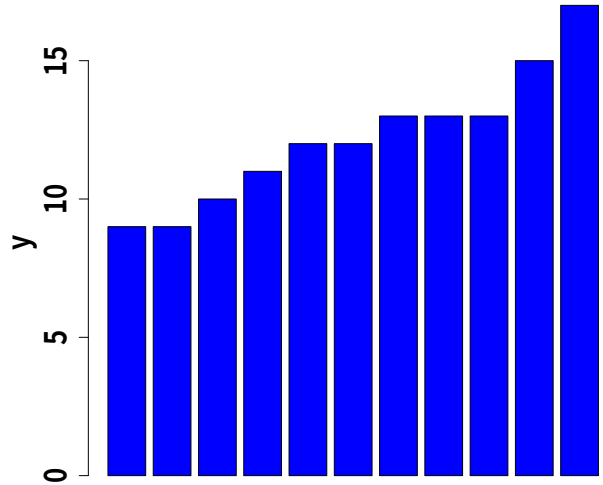
例 $n = 11$

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

i	1	2	3	4	5	6	7	8	9	10	11
$x_{(i)}$	9	9	10	11	12	12	13	13	13	15	17

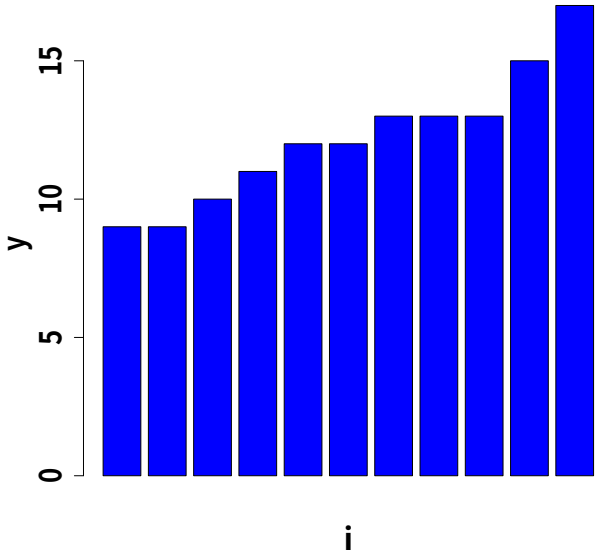


→ 順にならべる



四分位数のアバウトな定義 西川確率統計 p.96

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



四分位数の正確な定義 高校 数学 I 西川確率統計 p.96 注意 5

- Q_0, Q_4 さっきのまま.
-

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

- 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

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

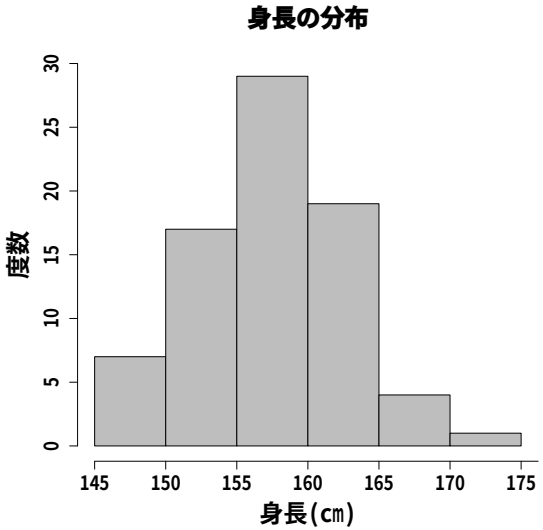
度数分布表からの中央値と四分位数の求め方

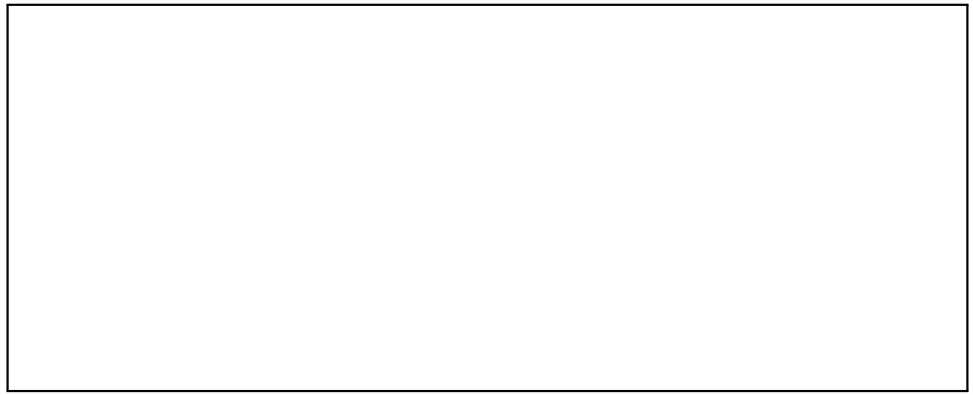
西川確率統計 5.6

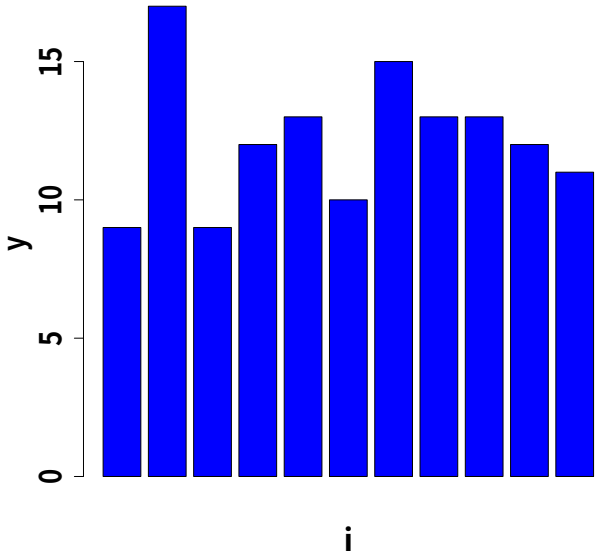
階級値=階級の (上限値+下限値)/2

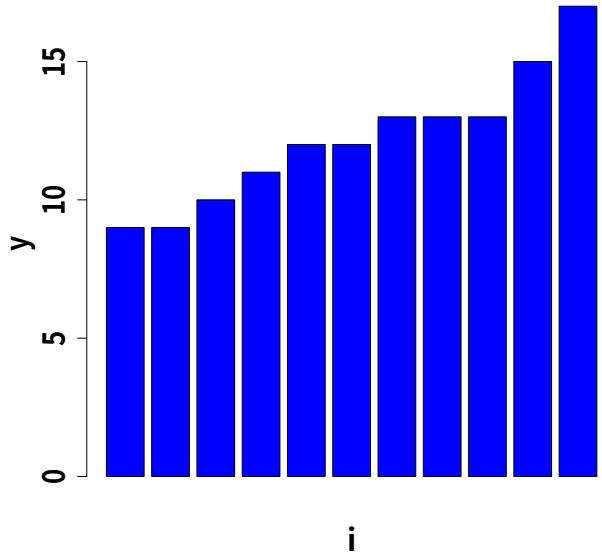
階級 (cm)	階級値 m_i	度数 f_i
145 より大きく 150 以下	147.5	7
150 より大きく 155 以下		17
155 より大きく 160 以下		29
160 より大きく 165 以下		19
165 より大きく 170 以下		4
合計 N	—	77

ヒストグラムからの中央値・四分位数の求め方









ここまで来たよ

3. 略解:データの分布

4. データの代表値・散らばりの尺度
   - 中央値と四分位数
   - 最頻値
   - 平均値
   - レンジ (範囲,range)・四分位偏差
   - 分散・標準偏差・平均偏差

最頻値=mode 西川確率統計なし

最頻値の定義

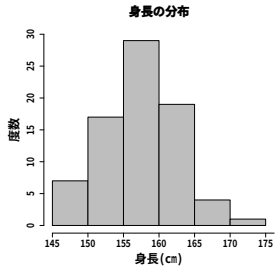
- 離散データの最頻値: ‘離散的な’ データのとき いちばん多く繰り返し現れる値
- ヒストグラムの最頻値: ‘連続的または離散的な’ データのとき 度数分布表/ヒストグラムで, 度数最大の階級の階級値

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の求め方

階級 (cm)	度数 f_i
145 より大きく 150 以下	7
150 より大きく 155 以下	17
155 より大きく 160 以下	29
160 より大きく 165 以下	19
165 より大きく 170 以下	4
170 より大きく 175 以下	1
合計	77

ヒストグラムの最頻値の意味



□

ここまで来たよ

3. 略解:データの分布

4. データの代表値・散らばりの尺度
   - 中央値と四分位数
   - 最頻値
   - 平均値
   - レンジ (範囲,range)・四分位偏差
   - 分散・標準偏差・平均偏差

平均値=mean

平均値の定義 西川確率統計 5.1.2

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-Q1

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

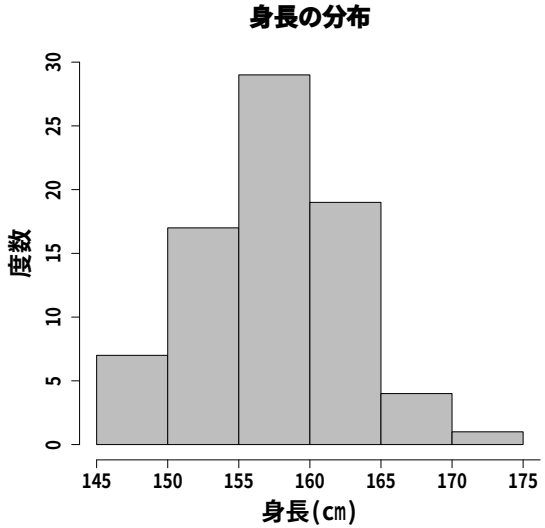
度数分布表からの平均値の求め方

西川確率統計 5.1.6

$$\overline{x} \approx \frac{1}{n}\sum_{i=1}^{k} m_i f_i = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$$

i 番目の階級の階級値 m_i, 度数 f_i.

ヒストグラムからの平均値の求め方



重心の座標 $x_{\mathrm{G}} = \dfrac{\sum_i x_i M_i}{\sum_i M_i}$ 力学

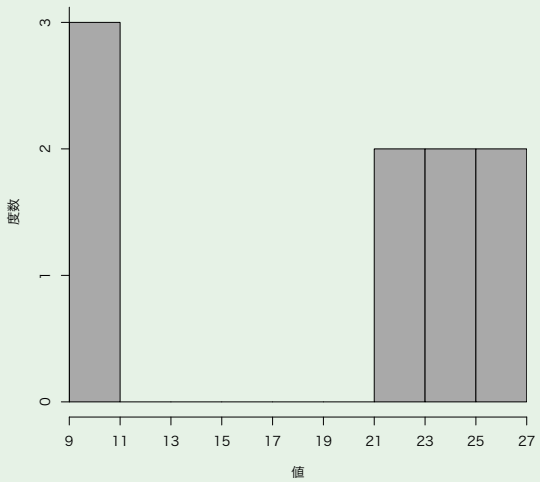
i 番目の質点の位置 x_i, 質量 M_i

L02-Q2

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

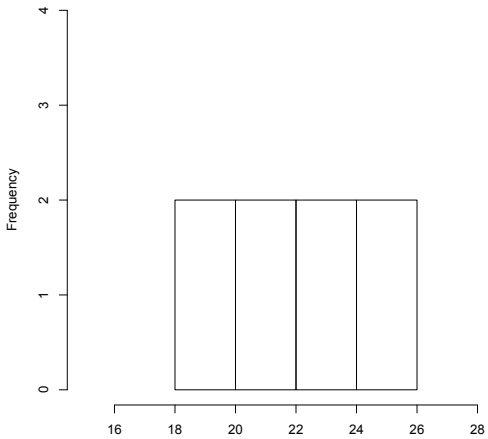


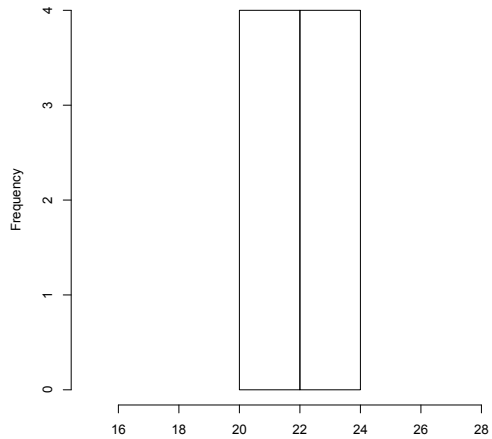
ここまで来たよ

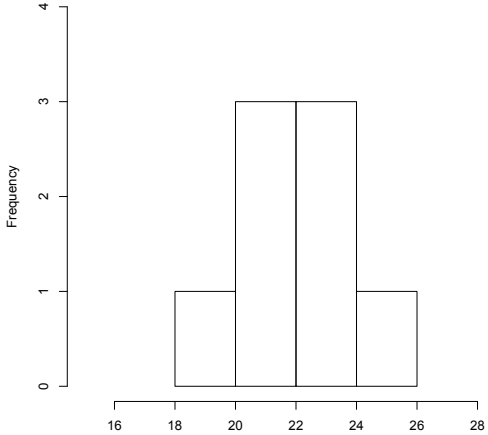
3. 略解:データの分布

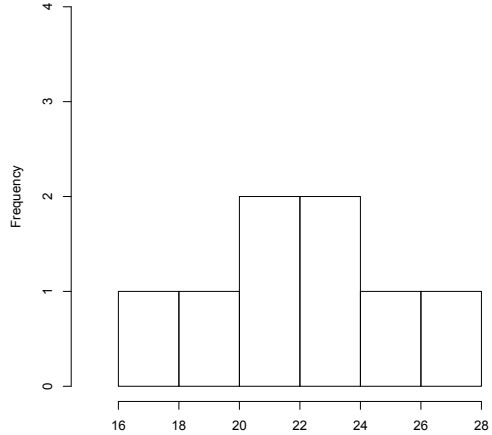
4. データの代表値・散らばりの尺度
   - 中央値と四分位数
   - 最頻値
   - 平均値
   - レンジ (範囲,range)・四分位偏差
   - 分散・標準偏差・平均偏差

平均値が同じでも分布はいろいろ









第 1,3 四分位数は?

散らばりの尺度が必要

レンジ・四分位偏差の定義 I

範囲タイプの量の定義 高校 数学 I 西川確率統計 p.97

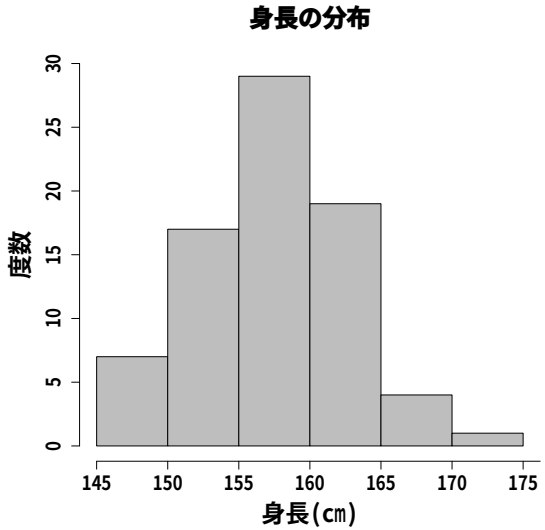
- 範囲 (レンジ) =
- 四分位範囲 (interquartile range) IQR=
- 四分位偏差 (quartile deviation) =

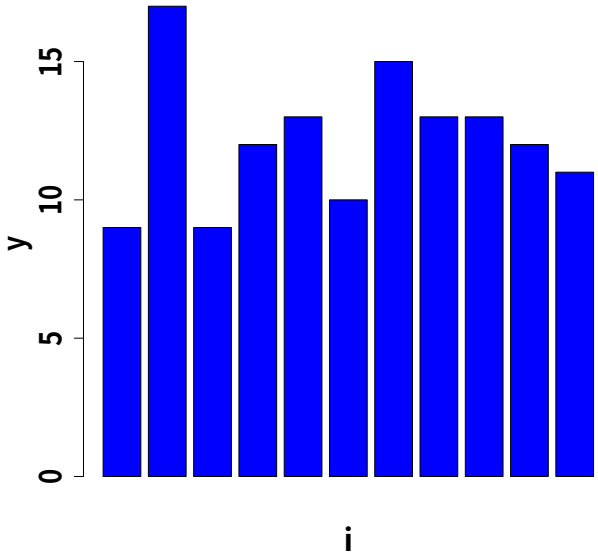
L02-Q3

Quiz(範囲)

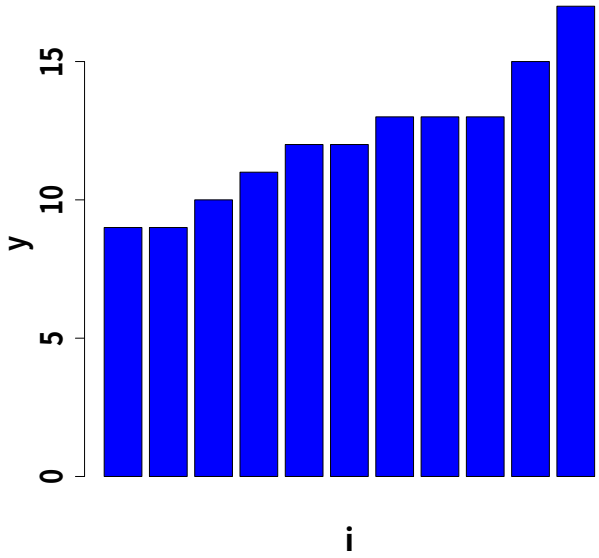
次のデータの, 範囲, 四分位範囲, 四分位偏差 を求めよう.
14 14 15 16 18 18 18 25

ヒストグラムからの範囲・四分位偏差の求め方





→ 並べかえ



ここまで来たよ

3. 略解:データの分布

4. データの代表値・散らばりの尺度
   - 中央値と四分位数
   - 最頻値
   - 平均値
   - レンジ (範囲,range)・四分位偏差
   - 分散・標準偏差・平均偏差

分散・標準偏差・平均偏差の定義 高校 数学 I 西川確率統計 p.98

データ: $x_1, x_2, \ldots, x_N$.

分散・標準偏差・平均偏差の定義

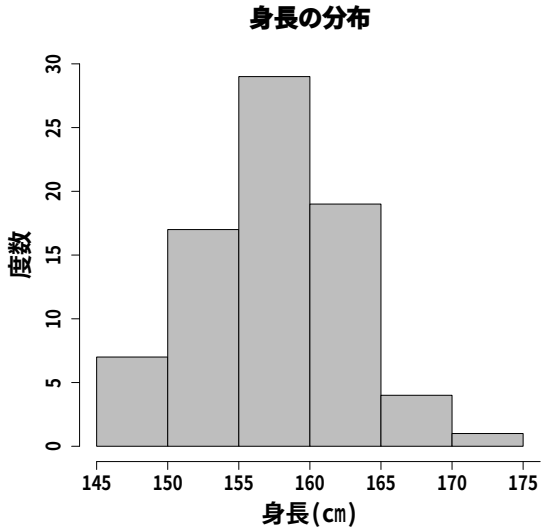
- データの分散 (variance): (偏差)^2 の平均

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

- データの標準偏差 (standard deviation)=
- データの平均偏差 (mean deviation):

$$d = \frac{1}{N}\sum_{i=1}^{N}|x_i - \overline{x}|$$

(例) 某国民的アイドル集団 (77 人) の身長 I



$n-1=77-1$ で割りたくなった人もいるかも．ここは 77 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 $\overline{x}=\frac{148+148.5+\cdots+172.3}{77}=158(\mathrm{cm})$
- 分散 $S^2=\frac{(148-158)^2+(148.5-158)^2+\cdots+(172.3-158)^2}{77}=26.0\ (\mathrm{cm}^2)$
- 標準偏差 $S=\sqrt{26.0}=5.1\ (\mathrm{cm})$

(例) 某国民的アイドル集団 (77 人) の身長 II

L02-Q4

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.

分散の便利な (こともある) 計算方法 高校 数学 I 西川確率統計定理 5.1(p.100)
度数分布表からの分散・標準偏差の求め方 高校 数学 I 西川確率統計 p.104
ヒストグラムからの標準偏差の求め方

連絡

- 配布資料は 1-503 向かいの引出や `http://hig3.net` で再配布.
- 加減乗除と平方根 (ルート) の使える電卓持ってきてね. 関数電卓でなくてもいいです. 携帯電話の機能・アプリでもかまいません.
- Learn Math Moodle の予習復習問題で来週の trial に備えてね.
- 樋口オフィスアワー月 3.5(1-539) 金 4(1-502), Math ラウンジ月-木昼 (1-614)
- 来週は教科書 西川確率統計 5.1.4, 5.1.5 読んできて

統計検定のディスカウント受験受付中 (– 2017-10-09 月) 樋口まで. 3 級合格者はプチテストの点数の一部として使用可.

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」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)