データの代表値

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(2015-09-25 Fri)

最終更新: Time-stamp: "2015-09-26 Sat 10:35 JST hig"

今日の目標

- データから 手で平均値, 離散データの最頻値, ヒストグラムの最頻値が求められる
- データから 手で中央値, 四分位数が求められる
- データから 手で高校レベルの箱ひげ図が描ける

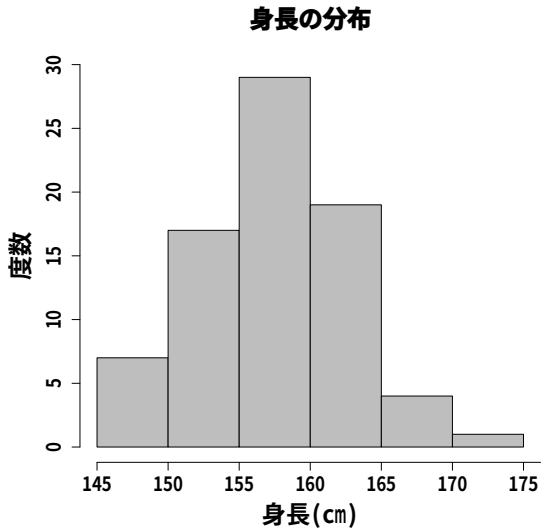
http://hig3.net

代表値:データを 1 個の値で代表させたい!

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	度数 f_j
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



ここまで来たよ

1. データの代表値
   - 中央値と四分位値
   - 最頻値と平均値
   - (高校レベル) 箱ひげ図

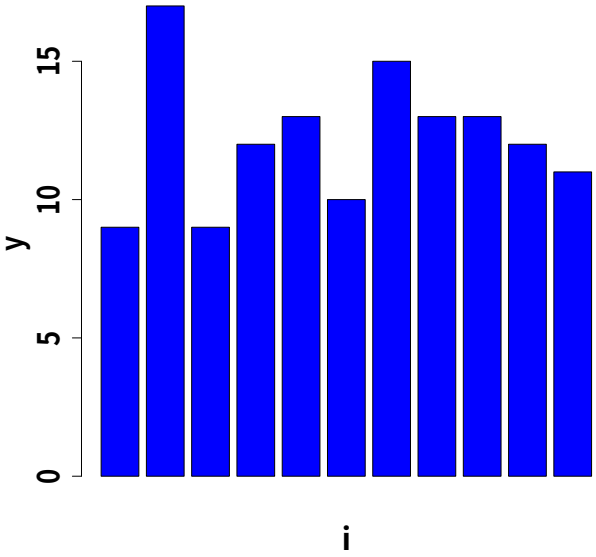
中央値 (median) と四分位数 (quantile)

データ $(1), (2), \ldots, (n)$ を小さい順に並び替えたものを,
$y_1 \leq y_2 \leq \cdots \leq y_n$ とする.

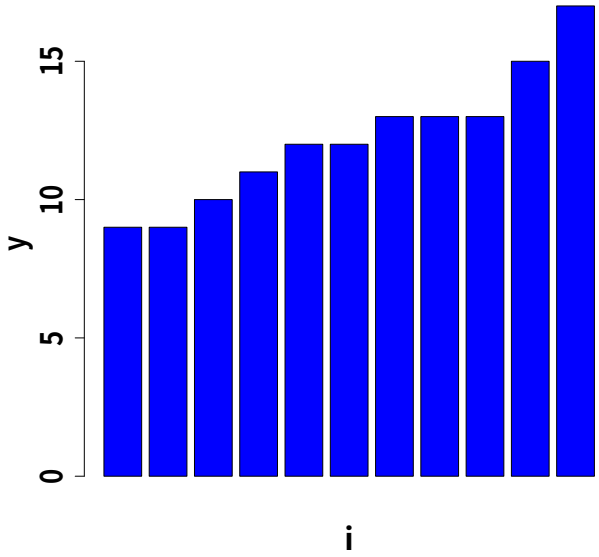
例

データ全体: 9 17 9 12 13 10 15 13 13 12 11

y: 9 9 10 11 12 12 13 13 13 15 17

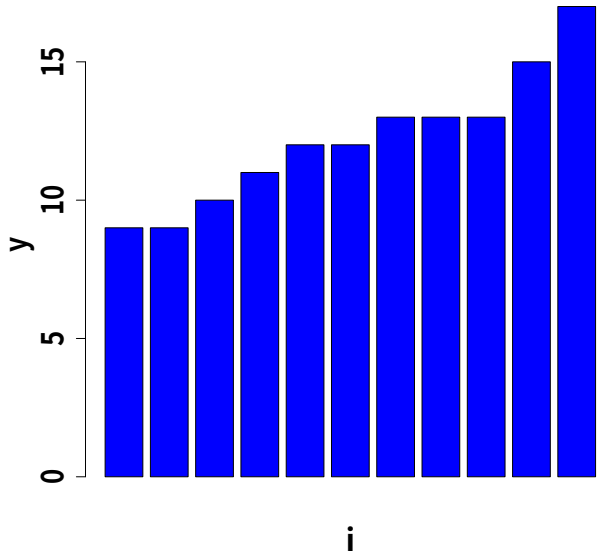


→ 順にならべる



四分位数のアバウトな定義

- 最小値 $Q_0 = y_1 \approx y_{\frac{0}{4}n}$
- 第 1 四分位数 $Q_1 = y_{\frac{1}{4}n}$
- 第 2 四分位数 $Q_2 = y_{\frac{2}{4}n}$=中央値
- 第 3 四分位数 $Q_3 = y_{\frac{3}{4}n}$
- 最大値 $Q_4 = y_{\frac{4}{4}n}$



四分位数の正確な定義

- Q_0, Q_4 さっきのまま.
-

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

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

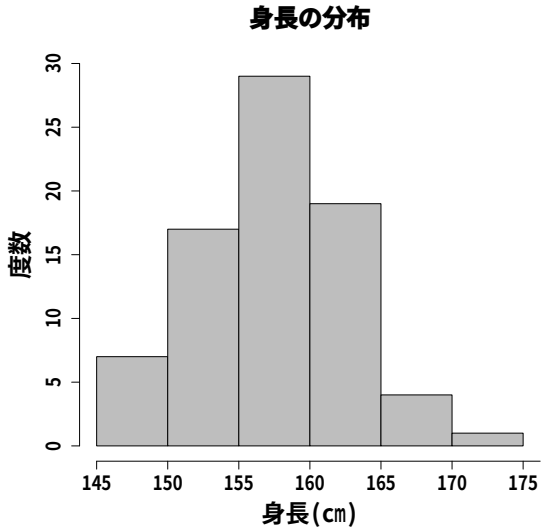
ちょっと変えた例: y 10 11 12 12 13 13 13 15 17

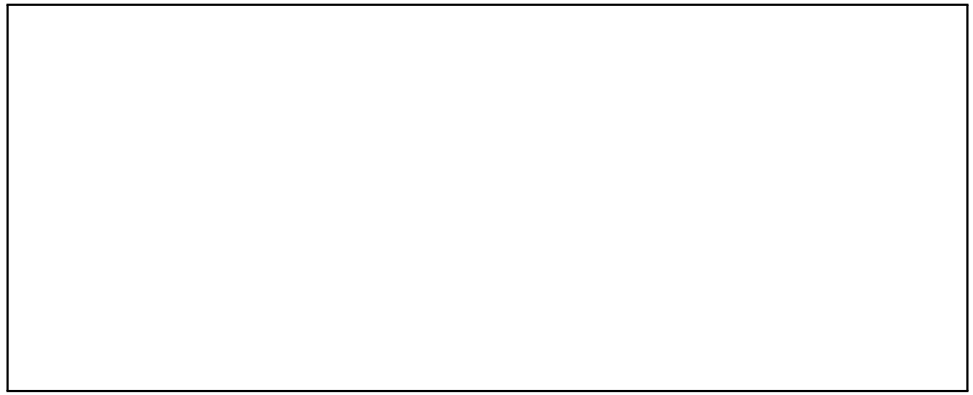
度数分布表からの中央値と四分位値の (だいたいの) 求め方

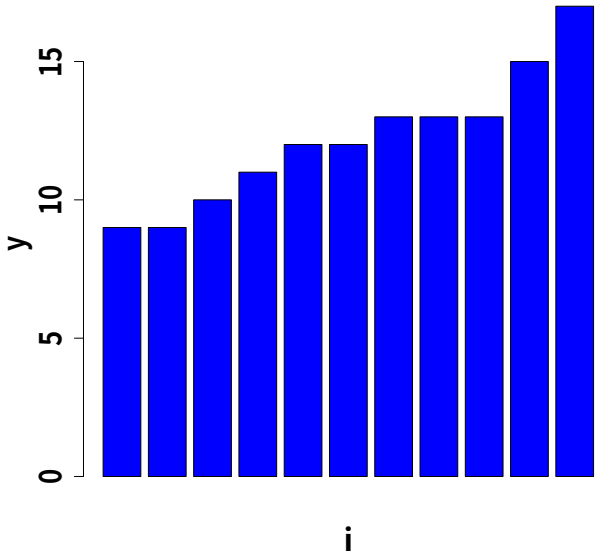
階級値=階級の (上限値+下限値)/2

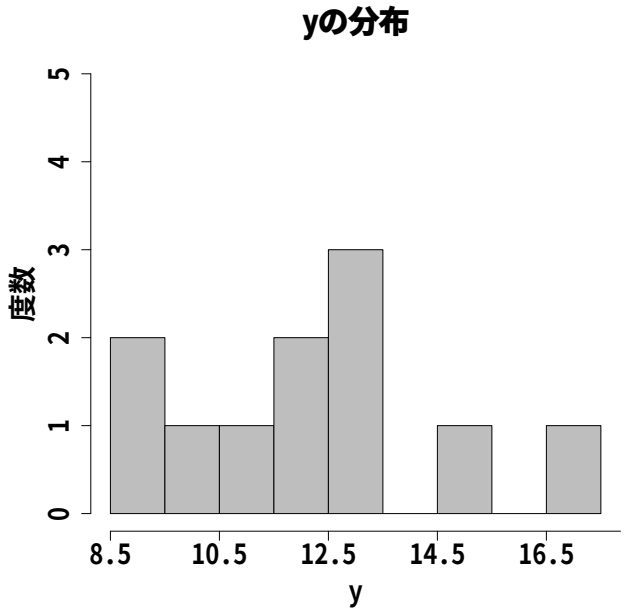
階級	階級値 m_j	度数 f_j
145 より大きく 150 以下	147.5	7
150 より大きく 155 以下		17
155 より大きく 160 以下		29
160 より大きく 165 以下		19
165 より大きく 170 以下		4
合計 n	—	77

中央値・四分位値のヒストグラム的意味









L02-Q1

Quiz(四分位値)

次のデータの四分位数 Q_1, Q_2, Q_3 を求めよう.
17 18 16 18 25 18 14 14 15

ここまで来たよ

1. データの代表値
   - 中央値と四分位値
   - 最頻値と平均値
   - (高校レベル) 箱ひげ図

最頻値=mode

最頻値の定義

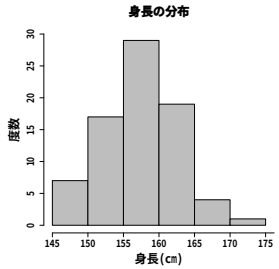
- 離散データの最頻値: ‘離散的な’ データのとき いちばん多く繰り返し現れる値
- ヒストグラムの最頻値: ‘連続的または離散的な’ データのとき 度数分布表/ヒストグラムで, 度数最大の階級の階級値

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	度数 f_j
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



平均値=mean

平均値の定義

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

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

例 1: 30 50 55 55 60 70 70 70 75 100 だと

平均値より中央値のいい点

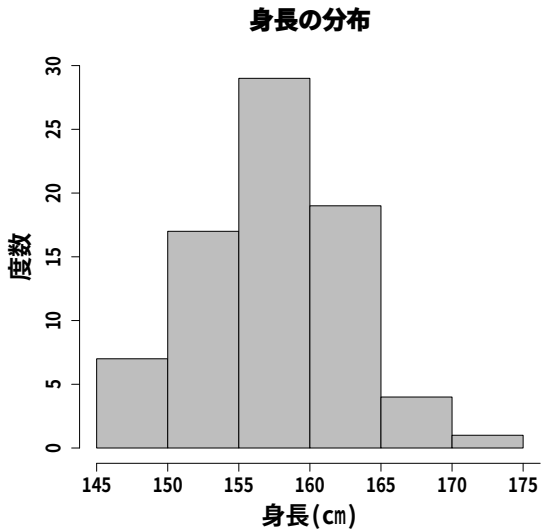
度数分布表からの平均値の (だいたいの) 求め方

$$m \approx \frac{1}{n}\sum_{j=1}^{k} m_j f_j$$

階級	階級値 m_j	度数 f_j	$m_j \times f_j$
145 より大きく 150 以下		7	1032.5
⋮			
170 より大きく 175 以下		1	172.5
合計		77	12122.5

平均値=12122.5/77

平均値のヒストグラム的意味



$x_{\mathrm{G}} = \frac{\sum_i m_i x_i}{\sum_i m_i}$ で, $m_i = 1$.

力学

L02-Q2

Quiz(代表値)

次のデータを考える.
14, 14, 15, 16, 18, 18, 18, 25

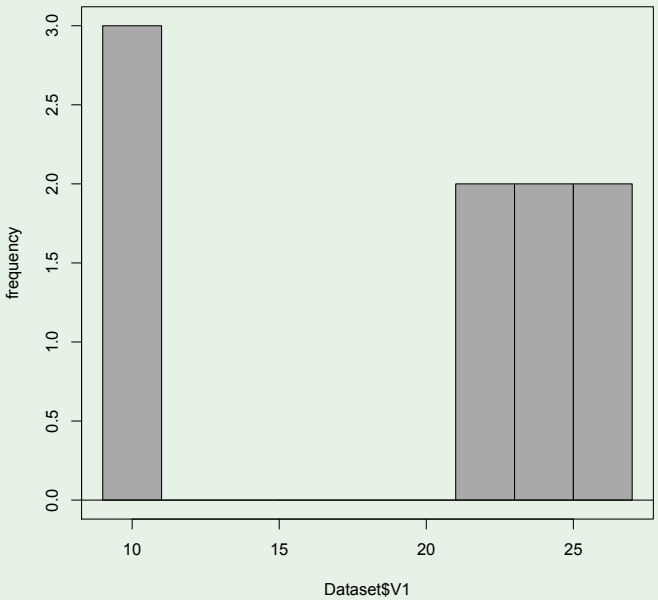
1. 四分位数 Q_1, Q_2, Q_3 を求めよう.
2. (離散データの) 最頻値を求めよう
3. 平均値を求めよう

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.

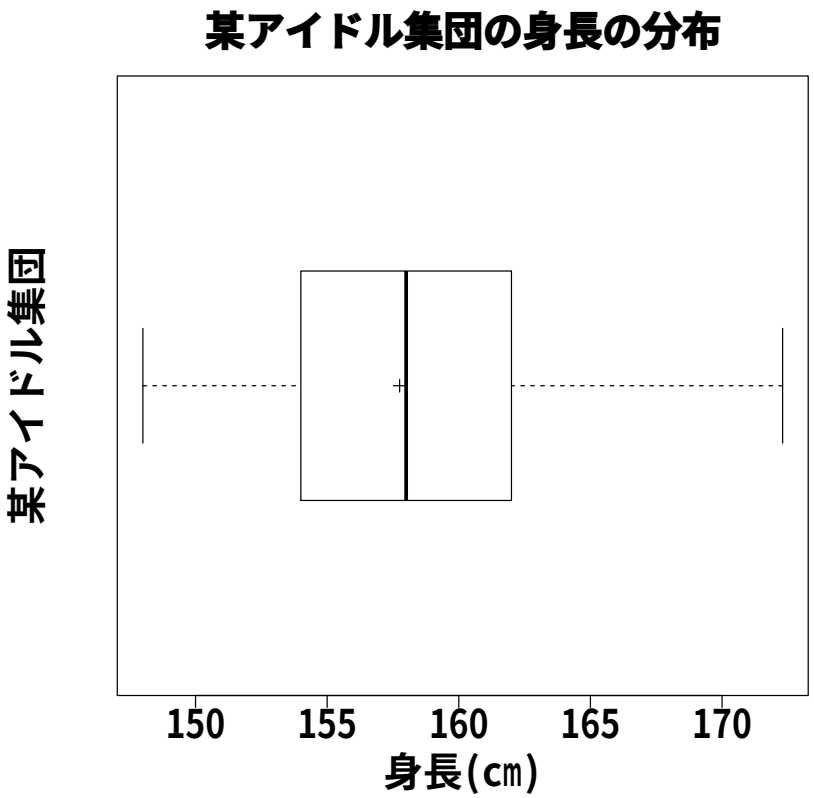
1. 中央値
2. (ヒストグラムの) 最頻値
3. 平均値



ここまで来たよ

1. データの代表値
   - 中央値と四分位値
   - 最頻値と平均値
   - (高校レベル) 箱ひげ図

箱ひげ図 (Box Plot)



最小最大値 Q_0, Q_4, 四分位点 Q_1, Q_2, Q_3

高校レベル箱ひげ図を描く手順

- Q_0, Q_4 Q_1,Q_2,Q_3 と平均値 m 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

いまの場合, 横軸:身長 (cm), 縦軸:意味なし

L02-Q4

Quiz(箱ひげ図)

下の 1 変量データについて, 3 つの四分位点を求め, 箱ひげ図を描こう.

2　8　10　11　12　12　12　14　15

連絡

- 次回は 7-002 講義室
- 配布資料は 1-503 向かいの引出, `http://hig3.net` で再配布しています.
- 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.
- 最初のころはいろいろ変更あるかも. メールに注意.
- 週のタイムラインで見たように, 非参照 Quiz 予習問題を RaMMoodle に金 17:00 ごろまでに公開. これで来週の Quiz に備えてね.
- 統計検定 申込締切 2015-10-16 金, 受験 2015-11-29 日. 3 級 or 2 級.
- オフィスアワー月 4 金 6(1-502)