データの代表値・ばらつきを表す値

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(2014-10-03 Fri)

今日の目標

- データから代表値:平均値, 四分位値, 中央値, 最頻値が手で求められる
- データからばらつきを表す量:分散, 標準偏差, 範囲, 四分位範囲が手で求められる
- データから箱ひげ図が手で描ける

http://hig3.net

ここまで来たよ

1. データの代表値・ばらつきを表す値
   - 代表値
   - データのばらつきを表す値
   - 箱ひげ図

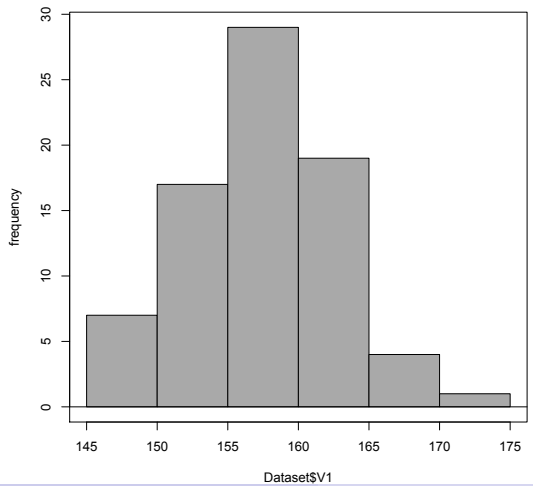
代表値:データを 1 個の値で代表させたい!

代表値某国民的アイドル集団の身長はだいたい 150cm? 170cm?
判断のもとになる情報は次のいずれかで与えられる.

- データ全体 $\rightarrow$ 定義
- 度数分布表
- ヒストグラム

データ全体の例 1(体重) 70 75 100 30 50 55 70 55 60 70 , $n = 10$.
度数分布表の例 2(身長) 某国民的アイドル集団

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	n =77



中央値 (median)

データ $x_1, x_2, \ldots, x_n$ 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 1

x: 70 75 100 30 50 55 70 55 60 70

y: 30 50 55 55 60 70 70 70 75 100

四分位数のアバウトな定義

- 最小値 $Q_0 = y_{\frac{0}{4}n}$
- 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 より小さいデータ (Q_2 は除く) の中央値
- Q_3 は, Q_2 より大きいデータ (Q_2 は除く) の中央値

例 1: 30 50 55 55 60 70 70 70 75 100

例 1': 30 50 55 55 60 70 70 70 75

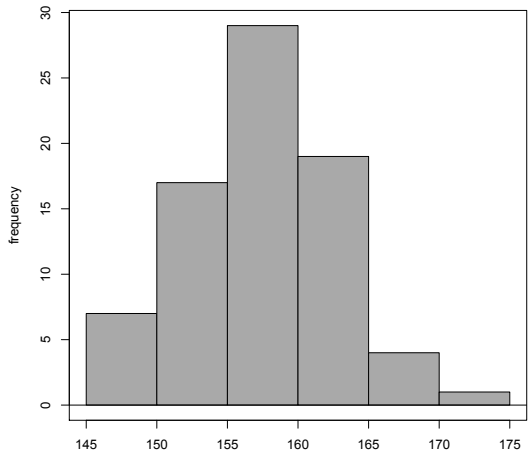
度数分布表からの中央値の (だいたいの) 求め方

端から度数を加えていって, データの総数の半分を超える階級の階級値

階級値=階級の (上限値+下限値)/2

j	階級	階級値 m_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
5	165 より大きく 170 以下		4
k =6	170 より大きく 175 以下		1
	合計 n	—	77

中央値のヒストグラム的意味



最頻値=mode

最頻値の定義

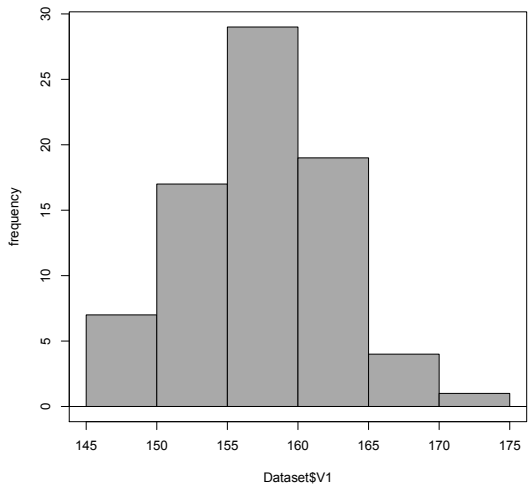
- ‘離散的な’ データのとき いちばん多く繰り返し現れる値
- ‘連続的な’ データのとき 度数分布表で, 度数最大の階級の階級値

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

度数分布表からの ‘最頻値’ の (だいたいの) 求め方

最頻値のヒストグラム的意味

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



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

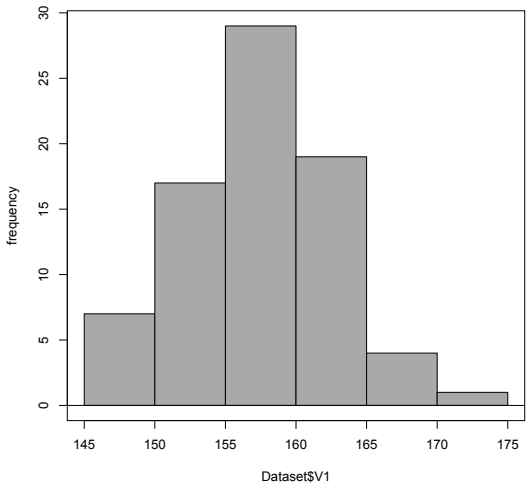
度数分布表からの平均値の (だいたいの) 求め方

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

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

平均値=12122.5/77

平均値のヒストグラム的意味



平均値のいい点

中央値のいい点

L02-Q1

Quiz(代表値)

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

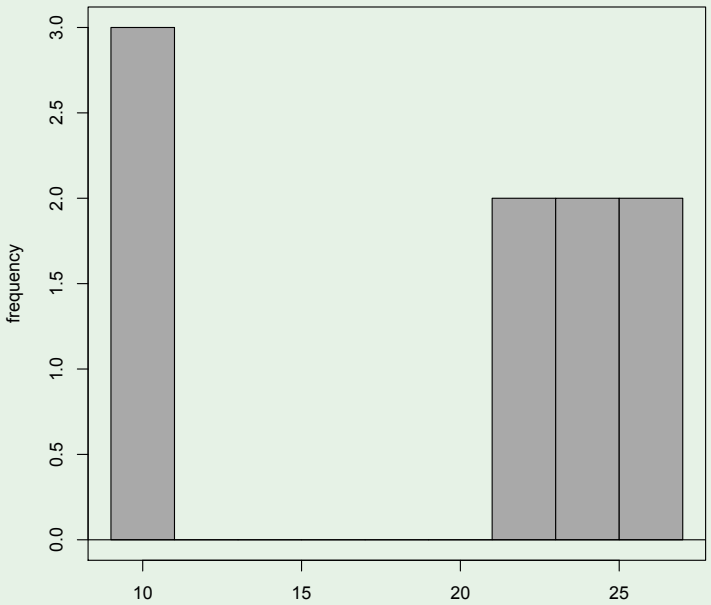
1. 四分位数 Q_1, Q_2, Q_3 を求めよう.
2. 最頻値を求めよう
3. 平均値を求めよう

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. 最頻値
3. 平均値

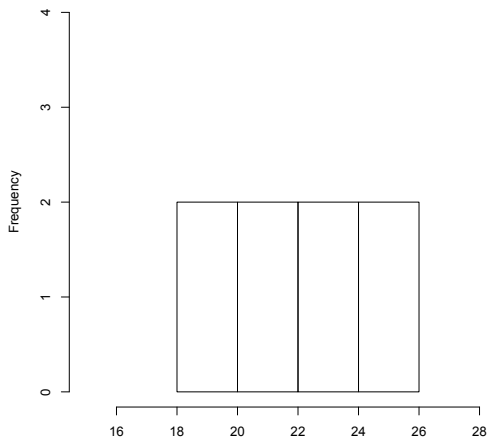


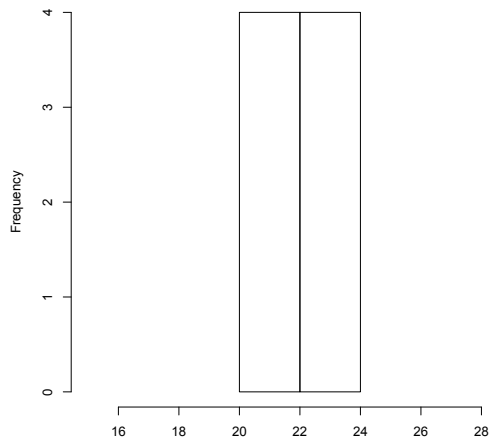
ここまで来たよ

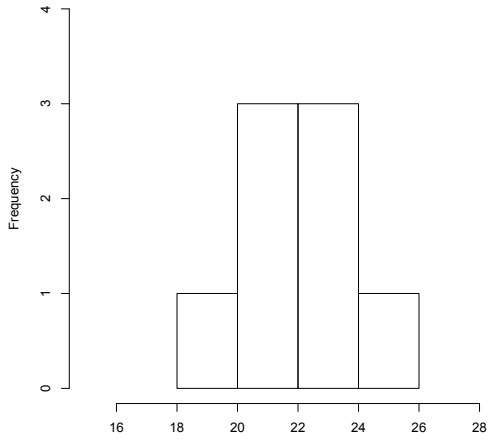
1. データの代表値・ばらつきを表す値
   - 代表値
   - データのばらつきを表す値
   - 箱ひげ図

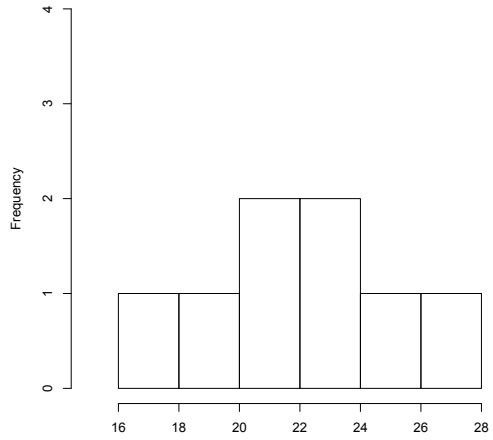
データの位置がすべてじゃない!

平均値が同じである分布









データのばらつきを表す値

範囲タイプの量の定義

- 範囲 (range) =
- 四分位範囲 interquartile range

  IQR= =

例 1: 30 50 55 55 60 70 70 70 75 100

L02-Q3

Quiz(範囲)

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

平均偏差と分散

平均値: $\overline{x}$ $(= m)$
準備: x_i の偏差 (deviation)$= x_i - \overline{x}$

偏差タイプの量の定義

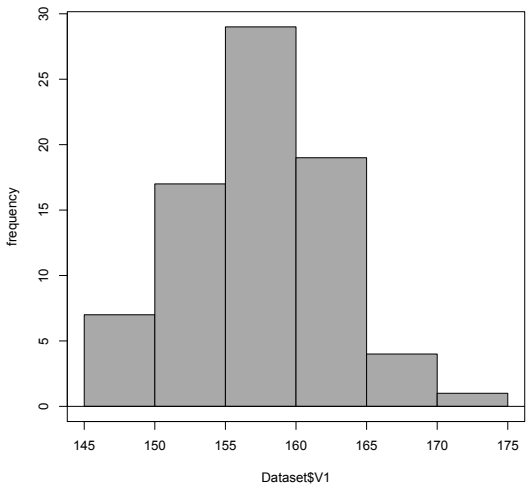
- データの平均偏差 (mean deviation): 偏差の絶対値の平均値

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

- データの分散 (variance): $(\text{偏差})^2$ 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)=



某国民的アイドル集団 (77 人) の身長

- 平均値 $m = \frac{148+148.5+\cdots+172.3}{77} = 158\text{cm}$
- 分散 $s^2 = \frac{(148-158)^2+(148.5-158)^2+\cdots+(172.3-158)^2}{77} = 26.0\ \text{cm}^2$
- 標準偏差 $s = \sqrt{26.0} = 5.1\ \text{cm}.$

$n-1=77-1$ 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．

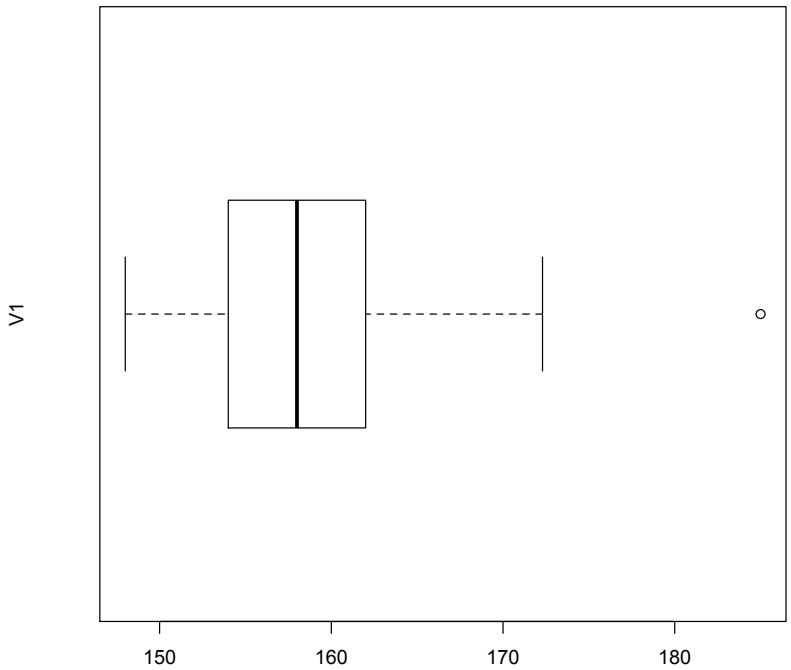
四分位範囲のいい点

標準偏差のいい点

ここまで来たよ

1. データの代表値・ばらつきを表す値
   - 代表値
   - データのばらつきを表す値
   - 箱ひげ図

箱ひげ図 (Box Plot)



横軸:身長 (cm), 縦軸:意味なし

四分位点 Q_1, Q_2, Q_3,　四分位範囲 IQR=$Q_3 - Q_1$

- Q_1 から下に, IQR の 1.5 倍より離れたデータ
- Q_3 から上に, IQR の 1.5 倍より離れたデータ

赤字部分を省略すると, 基本箱ひげ図. 高校の数学 I はそのレベル.

箱ひげ図を描く手順

- 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
- 平均値に $+$ を 1 個描く
- 外れ値を除いた最大値, 最小値までひげを描く
- 外れ値を $\circ$ で描く

L02-Q4

Quiz(ヒストグラムと箱ひげ図を描こう)

次のデータから作ろう.

1. 箱ひげ図
2. 度数分布表
3. ヒストグラム

14 14 15 16 18 18 18 25

連絡

- 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.
- Quiz の略解は授業終了後に http://hig3.net で配布しています.
- 次回からは, 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, 予習問題を RaMMoodle に金 15:30 までに公開. 翌週水 09:20 までにやってね. それまで何回でも「受験」できます. 最後の受験が点数になります.

来週の非相談非参照テスト

- 四分位値を求めよう (プチテストでは「すべての代表値」)
- ~~箱ひげ図を描こう~~
- (追加) 標準偏差を求めよう