データの代表値

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　L02(2020-10-05 Mon)

最終更新: Time-stamp: "2020-10-03 Sat 13:34 JST hig"

今日の目標

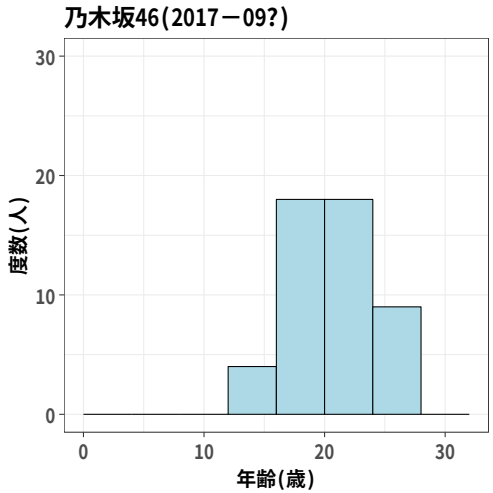
- 代表値:中央値, 四分位数, 平均値, 最頻値を手で求められる 岩薩林 確率・統計 §1.2 高校 数学 I
- 代表値を Excel で求められる.
- ヒストグラムを Excel で描ける

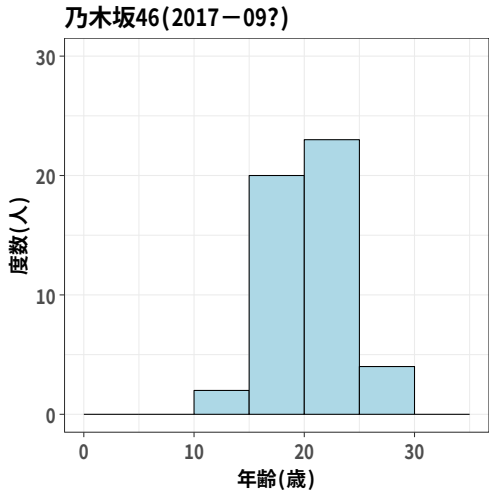
L01-Q1

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

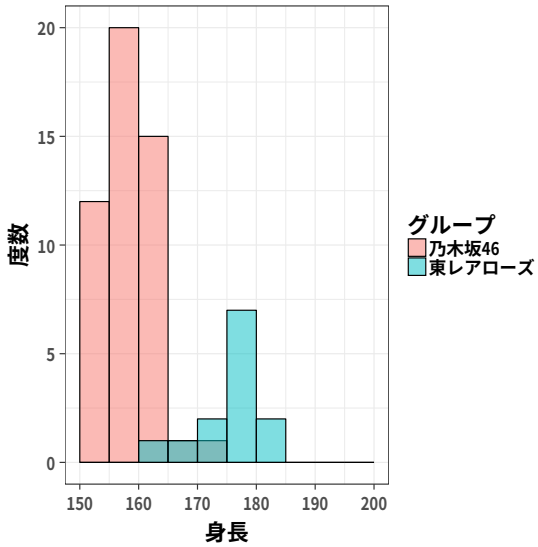
2. データの代表値
   - 中央値と四分位数
   - 平均値
   - Excel で統計

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)

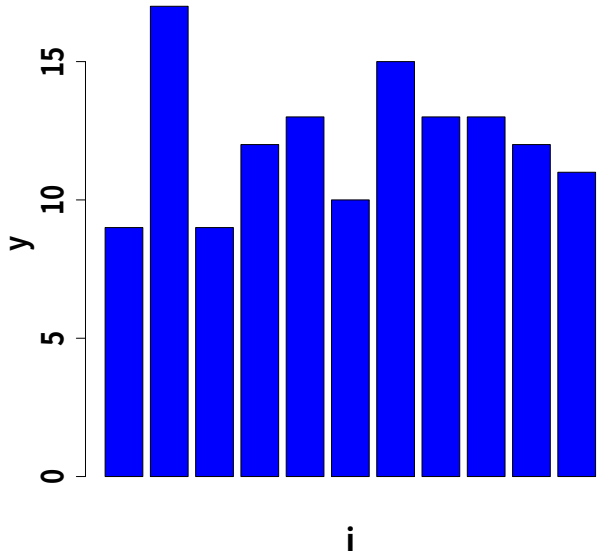
身長 y_i のデータ (n 個) を小さい順に並び替えたものを,
$x_0 \leq x_1 \leq \cdots \leq x_{n-1}$ とする.

例 $n = 11$

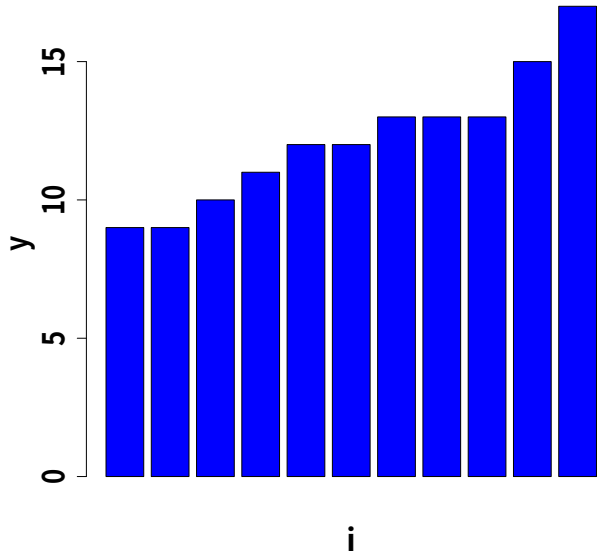
i	0	1	2	3	4	5	6	7	8	9	10
y_i	9	17	9	12	13	10	15	13	13	12	11

$\rightarrow$

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 $= x_{q\times(n-1)}$. $(0 \le q \le 1)$.

  全体の (1 に対する割合)q 番目 (下から) のデータの値

- 最小値 $Q_0 = x_0 = x_{\frac{0}{4}\times(n-1)} =$ (0.00-分位数).
- 第 1 四分位数 $Q_1 = x_{\frac{1}{4}\times(n-1)} =$ (0.25-分位数).
- 第 2 四分位数 $Q_2 = x_{\frac{2}{4}\times(n-1)}$=中央値=メディアン = (0.5-分位数).

  岩薩林 確率・統計練習問題 1-1

- 第 3 四分位数 $Q_3 = x_{\frac{3}{4}\times(n-1)} =$ (0.75-分位数).
- 最大値 $Q_4 = x_{\frac{4}{4}\times(n-1)} =$ (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.
-

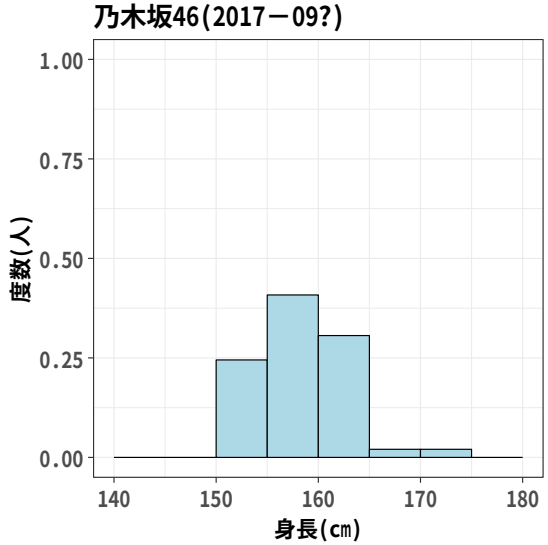
$$Q_2 = \begin{cases} x_{\frac{n-1}{2}} = \boxed{\text{真ん中の値}} & (n \text{ が奇}) \\ \frac{1}{2}\left(x_{\frac{n}{2}-1} + x_{\frac{n}{2}}\right) = \boxed{\text{真ん中 2 個の和/2}} & (n \text{ が偶}) \end{cases}$$

- 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



ヒストグラムの幅が 1 として, 全体の面積は 1

左側の面積が q になる x が q-分位数

ここまで来たよ

1. データの分布

2. データの代表値
   - 中央値と四分位数
   - 平均値
   - Excel で統計

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_{x} などという記号もある.

岩薩林 確率・統計例題 1.2(p.5)

平均値が中央値よりいい点

すべての値の影響がはいる, 計算しやすい

中央値が平均値よりいい点

例外的に大きい (小さい) 値に影響されにくい

度数分布表からの平均値の求め方

岩薩林 確率・統計 §1.2.3, (1.14)p.17

j 番目の階級の階級値 x_j, (相対) 度数 f_j. 近似的に,

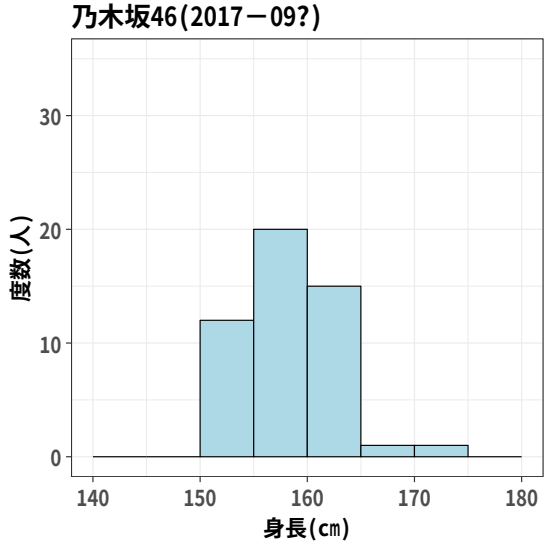
$$\overline{x} \approx \frac{1}{n}\sum_{j=1}^{k} x_j f_j = \frac{\sum_{j=1}^{k} x_j f_j}{\sum_{j=1}^{k} f_j}$$

例

階級 (cm)	階級値 $x_{(i)}$	度数 f_i	$x_{(i)} \times f_i$
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$=12122.5/77$

相対/度数の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 $m_j = f_j$

そこを支えると釣り合う点.

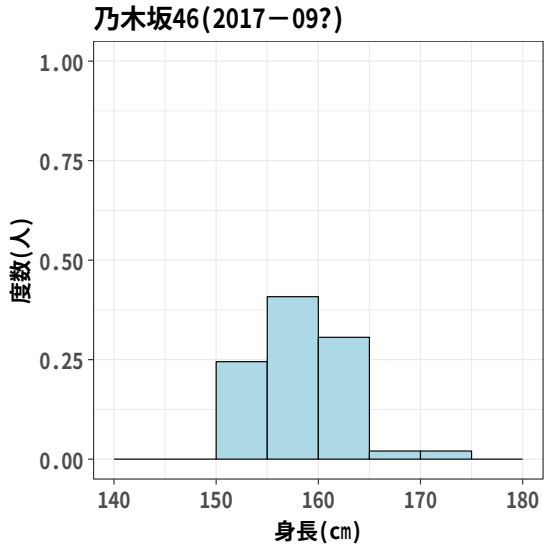
最頻値=mode 岩薩林 確率・統計練習問題 1-2

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 70

ヒストグラムの最頻値



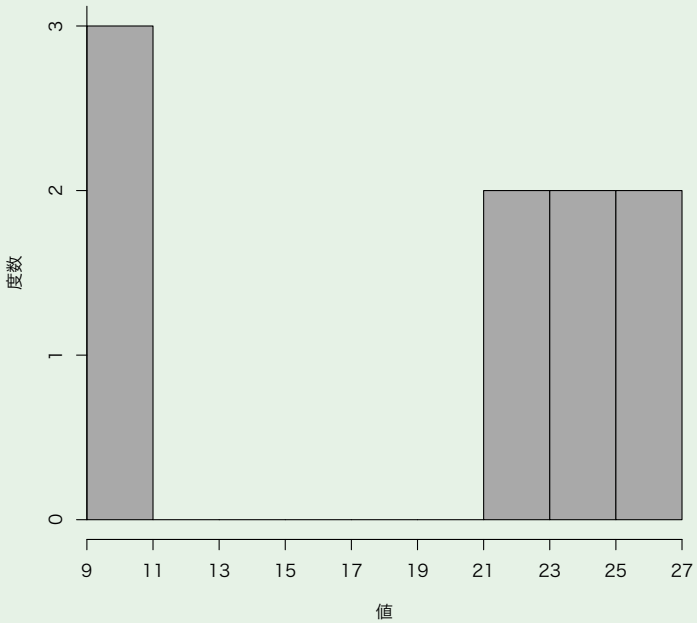
ヒストグラムの階級の取り方で変わる

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
   - 中央値と四分位数
   - 平均値
   - Excel で統計

Excel 使用の準備

統計ソフトウェア実習室にインストールされているのは

R 無料. オープンソース. 解説書が多い.
SPSS 伝統ある高級品.
Excel 表計算. 機能は限られ怪しいところもあるが, 普及率高い.
龍大では Office365 で無料. この科目ではこれ.

準備 (データ分析の有効化)
ファイル > オプション > アドイン > Excel のアドイン > 設定 > データ分析 に
チェックを入れて OK する.

Excel によるグラフ描画 挿入 > グラフ > (グラフの種類)

題名や軸の変数名の追加
挿入 > グラフ > グラフのデザイン > グラフ要素を追加

使用するデータの調整
挿入 > グラフ > グラフのデザイン > グラフデータの選択

表計算ソフトウェア (Excel) による分析 高校 数学 I

メニューからデータ範囲を指定, または関数の引数にデータ範囲を指定.

	メニューベース	関数ベース
平均値, 分散, 標準偏差	データ > 分析 > データ分析 > 基本統計量 > 統計情報	平均値 `average`, 分散 `var.p`, 標準偏差 `stdev.p`, 最頻値 `mode`
(四) 分位数	データ > 分析 > データ分析 > 順位と百分位数	中央値 `median`, 四分位数 `quartile`, 百分位数 `percentile.inc`
順位, 分位	データ > 分析 > データ分析 > 順位と百分位数	順位 `rank`, 百分位 `percentrank.inc`
ヒストグラム, 箱ひげ図	挿入 > グラフ > ヒストグラム, 箱ひげ図	グラフ

メニューベースでデータ分析をするときの注意

- 列=縦, データを n 個並べる.
  - 縦横を変えるときは, 形式を選択してペースト > 行列を入れ替える
- 「ラベル」は, 1 行目 (または 1 列目) に書かれている変数名 (身長) (データ (60 点) でなく). ラベルを範囲に含めるか含めないか, チェックボックスがあることが多い.