データの分布と代表値

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　L01(2015-09-18 Fri)

最終更新: Time-stamp: "2015-09-26 Sat 10:37 JST hig"

今日の目標

- e ラーニングシステムで学習できる
- データから 手で度数分布表とヒストグラムが作れる
- データから 手で中央値, 四分位値が求められる

http://hig3.net

ここまで来たよ

学習目標

講義概要 → シラバス
現実世界の現象を理解し, 数理モデルとの関係を明らかにするためには, 観察・実験により取得した現象のデータを整理・解析することが必要です. データを表現する記述統計, 限られたデータから現象の性質を推測する推測統計を学びます. ただし, 量的 1 変数の場合を主に扱います. これに必要な範 囲で確率論を学びます. 数式を用いた解析, ソフトウェアによる解析の両方に習熟します.
到達目標 → シラバス
実験・観察により取得した (質的, 量的, 1 変量, 2 変量) データを統計的に整理して, 他者に対して表現できる. データから仮説を立てて検証し, 他者を説得できる.

確率統計☆演習 I を履修してはいけない理由

次のどれも響かない人は履修しないことを奨めます.

- 数学の教員免許に必要
- コア M
- (3 年前期) 確率統計☆演習 II, 計算科学☆実習 B の前提科目
- 中高の数学で統計はすでに強化されてる
- 教育の評価に統計は必要
- いま, 統計学が熱い!
- いま, ビッグデータ, 人工知能 (AI), 機械学習 (machine learning) が熱い!!
- 統計は科学技術の言葉 $\rightsquigarrow$ 数理卒は当然期待されてる
- 統計検定 2 級

こんなことに答えます

1. 高校の数学で, こういう教え方導入したら, ちょっとだけ平均点が上がった. これ効果あったって言っていいの?
2. YouTube から猫の動画を見つけるアルゴリズム, こう改良して, 100 個の入力画像で試したら, 判定精度がちょっとあがった. これで結論していいの? 10000 個でやり直すべき?
3. 秋元 P はチーム A にチーム K より身長高いメンバーをいれてる説を唱えたけどみんな信じてくれない…どうやって説得する?

確率統計☆演習 I ののり

成績計算難しくないけどとにかく注文の多い科目です…
科目の成績 100 ピーナッツは

- 30 ピーナッツ:毎回授業での非参照 quiz,e ラーニングの予習問題, 授業時間内の活動, それほどたいへんじゃないレポートなど
- 30 ピーナッツ:プチテスト (11 月)
- 40 ピーナッツ:ファイナルトライアル (定期試験期間)
- その他追加ピーナッツ. その時に説明.

その時点のピーナッツにかかわらず, ファイナルトライアルに参加しないと合格にはなりません. ファイナルトライアル時点で 20 ピーナッツ未満の人も, (平均点を上げるために) 参加をすすめますが, 追試験はなし.
欠席届ピーナッツ的に考慮されたい場合は, 専用用紙に事情を説明する書類を貼って, 授業前後各 5 分に提出 (事前事後とも可. ファイナルトライアルが締切). 欠席に事前連絡は原則不要. 何回欠席してもファイナルトライアル参加資格を失うことはありません.

担当者ののり

- なまえ: 樋口さぶろお `hig-probstat`@math.nyukoku.ac.jp
- へや: 1-502
- オフィスアワー: 月 4(1-502/1-539), 金 6(1-502). 1-502 に訪問歓迎な時間: 月火昼 (Math ラウンジに行ってることも). お弁当持参歓迎. お湯あげます.
- Web ページ: `http://hig3.net` (表紙に QR コード) 演習の指示や, スケジュールもここから.

1 週間のタイムライン

1. 金 17:00 ごろまでに Quiz 予習問題 (=非参照 Quiz 予想問題) を RaMMoodle で公開. (普通は)2 週間何度でも受験可能. 非参照 Quiz の満点の 1/3 まで得られます.
2. 金 2 の最初 非参照 Quiz(=テスト) 参照不可 相談不可
3. 金 2 部屋がかわったり座席指定があったりクラスで何かやったり…
4. 金 2 の最後 来週の Quiz の予告

RaMMoodle を使ってみよう

`http://hig3.net` → RaMMoodle (全学認証) → 確率統計☆演習 I

ここまで来たよ

1. はじめに
   - この授業どんなのり?
2. データの分布
   - データとは?
   - 度数分布表
   - ヒストグラム
   - 代表値:中央値と四分位値

1 変数の量的データ

某アイドル集団 (77 名)+某バレーボール選手 (1 名) の身長のデータ.

148cm
148.5cm
149cm
⋮
185cm

ps3id_raicho_1182 さん（最終更新日時:2012/3/20）投稿日：
2012/2/15 AKB48　身長　まとめ（研究生は 12.5 期まで）
http://note.chiebukuro.yahoo.co.jp/detail/n32745

このコースの最後までいくと問えること (正確な表現ではありません)

- オーディションにおいて, 身長は考慮されているか?
- チーム編成において, 身長は考慮されているか?
- ⋮

ここまで来たよ

1. はじめに
   - この授業どんなのり?
2. データの分布
   - データとは?
   - 度数分布表
   - ヒストグラム
   - 代表値:中央値と四分位値

階級	度数	相対度数
145 より大きく 150 以下	7	0.09
150 より大きく 155 以下	17	0.22
155 より大きく 160 以下	29	0.37
160 より大きく 165 以下	19	0.24
165 より大きく 170 以下	4	0.05
170 より大きく 175 以下	1	0.01
175 より大きく 180 以下	0	0.00
180 より大きく 185 以下	1	0.01
185 より大きく 190 以下	0	0.00
合計	78	1.00

- 階級幅は一定で
- 相対ナントカ (比率) の合計が 1 にならないとき. 度数分布表と限らず一般に, 無視して 1 と書くか, 相対誤差が小さい行で調整.

-
  - 以下, 以上, 未満 (=より小さい), より大きい

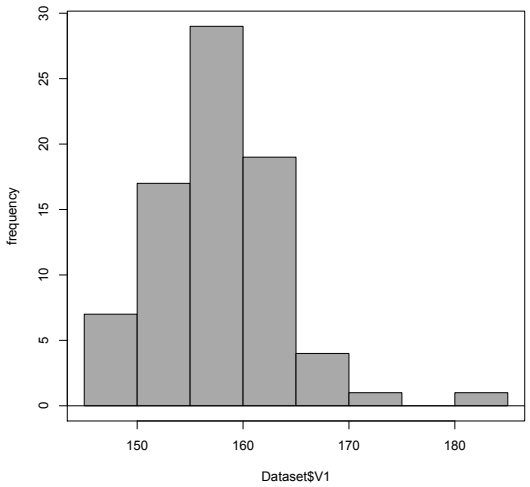
度数分布表の作り方

- 階級=一定間隔で区切った区間, 下品な?言葉 'bin' ビン. いくつに分けるか? 一概には言えないけど, 切りのいい値にしちゃっていい.
- 階級幅=区間の幅
- 階級値=その階級のまん中の値
- 度数=その範囲に入ってるデータの個数
- データ全体の個数=度数の合計 $= n$
- 相対度数=度数/データ全体の個数 (%で書くことも)

ここまで来たよ

1. はじめに
   - この授業どんなのり?
2. データの分布
   - データとは?
   - 度数分布表
   - ヒストグラム
   - 代表値:中央値と四分位値

ヒストグラム



- ‘度数分布表を棒グラフにしたもの’
- 必ず階級幅は一定
- 階級の個数や階級幅は指定がなければ, 見やすいように決めてよい.
  - 階級の幅=超大きい $\rightsquigarrow$ 長方形 1 個
  - 階級の幅=超小さい $\rightsquigarrow$??

手でやってみよう. 練習用データ. 小数点以下はでっち上げです.

名前	年齢
中野郁海	14.1
大和田南那	15.2
川本紗矢	16.4
大島涼花	16.2
加藤玲奈	17.9
宮脇咲良	17.1
小嶋真子	17.9

白間美瑠	17.6	入山杏奈	19.0	峯岸みなみ	22
高橋朱里	17.1	生駒里奈	19.0	指原莉乃	22.0
向井地美音	17.0	木崎ゆりあ	19.2	横山由依	22.1
森保まどか	17.6	川栄李奈	20.7	松井玲奈	23.7
松井珠理奈	18.1	武藤十夢	20.8	柏木由紀	23.2
渋谷凪咲	18.4	山本彩	21.4	須田亜香里	23.2
田野優花	18.4	島崎遥香	21.7	高橋みなみ	23.0
矢倉楓子	18.1	渡辺麻友	21.5	宮澤佐江	24.2
兒玉遥	18.4	渡辺美優紀	21.2	小嶋陽菜	26.9

Example (度数分布表とヒストグラムを作ろう)

度数分布表とヒストグラムを作ろう

- 学籍番号奇数の人は 5 刻みで. 10-15,15-20,. . . ,
- 学籍番号偶数の人は 4 刻みで. 12-16,16-20,. . . ,
- 以上, 以下, 未満, より大きい, は自分で正しく決めて.

ここまで来たよ

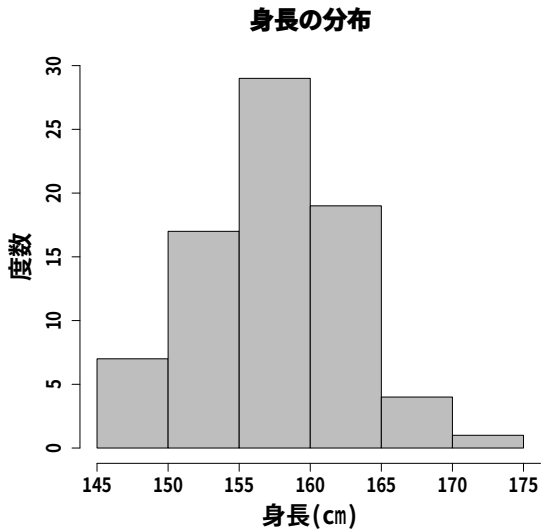
1. はじめに
   - この授業どんなのり?
2. データの分布
   - データとは?
   - 度数分布表
   - ヒストグラム
   - 代表値:中央値と四分位値

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



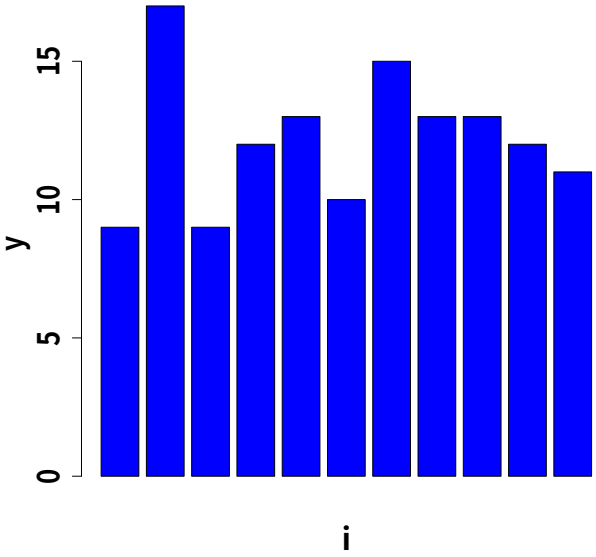
中央値 (median) と四分位数 (quartile)

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.

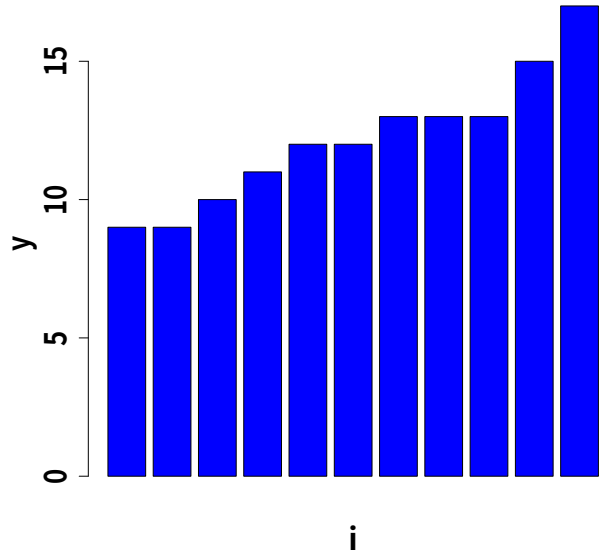
例

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

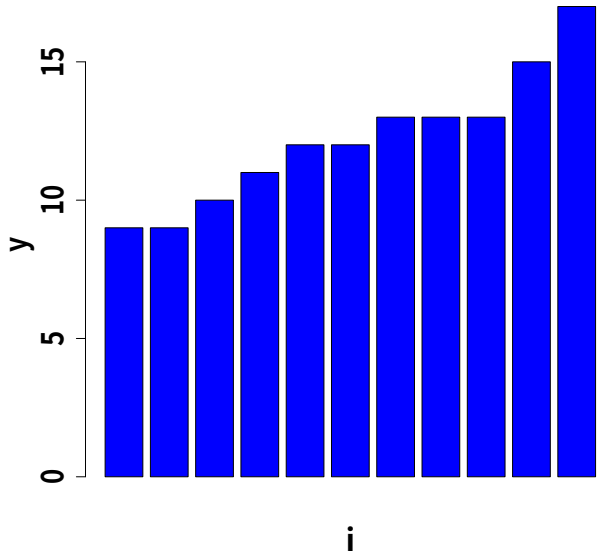


→ 順にならべる



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 Q_2
- Q_3 は, Q_2 より後ろにあるデータの (Q_2 自身は除く) の Q_2

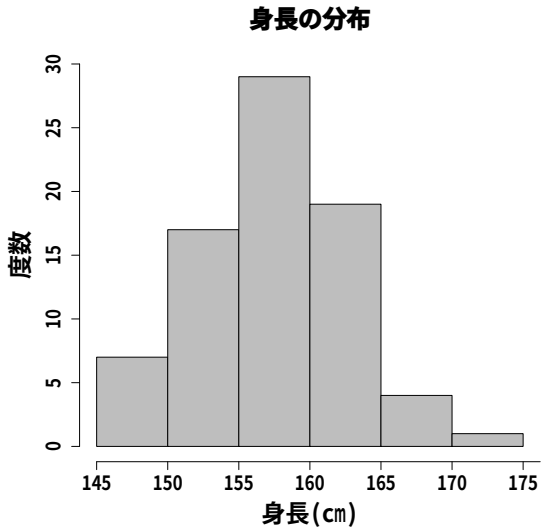
ちょっと変えた例: 10 11 12 12 13 13 13 15 17

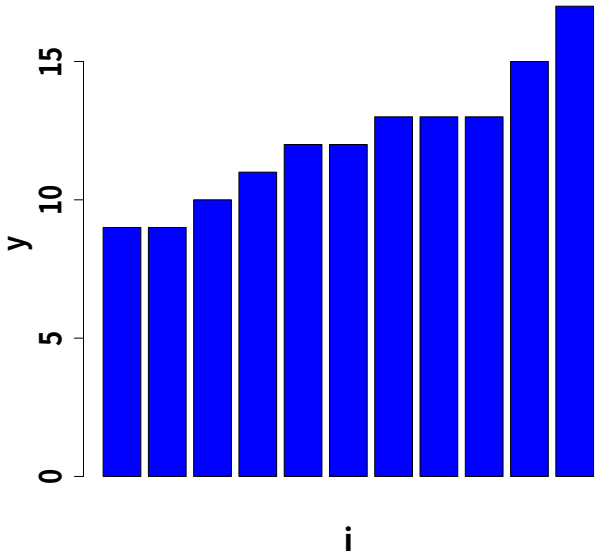
度数分布表からの中央値と四分位値の (だいたいの) 求め方

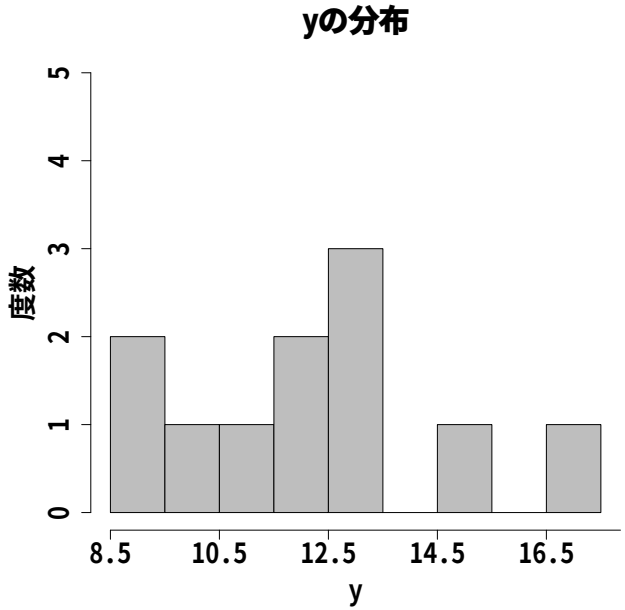
階級値=階級の (上限値+下限値)/2

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







L01-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

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.
- オフィスアワー月 4 金 6(1-502)
- 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 17:00 までに公開. これで来週の Quiz に備えてね.