データの分布

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　L01(2020-09-28 Mon)

最終更新: Time-stamp: "2020-09-28 Mon 06:50 JST hig"

今日の目標

- オンライン授業に参加できる
- データから度数分布表とヒストグラムが作れる 岩薩林 確率・統計 §1.1
- ヒストグラムから順位を読み取れる 岩薩林 確率・統計なし

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データを整理・解析することが必要です. データを整理して表現する記述統計と, 限られたデータから数理モデルのパラメタを推測する推測統計を説明します. 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説明します. 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てもらいます.
到達目標 → シラバス

- 実験・観察により取得した量的 1,2 変数データを統計的に整理して表現し, 他者に対して説明できる.
- データから数理モデルのパラメタを推測して, 根拠とともに他者に説明できる.
- 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.

- コア選択必修 M
- (3 年前期) 確率統計☆演習 II, 計算科学☆実習 B の前提科目
- 数学の教員免許の必修科目
- 高校の高校 数学 I(データの分析)=毎年センター試験に出題, 高校 数学 A(場合の数と確率), 高校 数学 B(確率分布と統計的推測)(選択)
- 教育の評価に統計は必要
- いま, データサイエンス, 統計が熱い!
- いま, 機械学習, 人工知能 (AI) が熱い! 生成モデルの芯
- 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 3 級 (今年は CBT のみ)

https://www.amazon.co.jp

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 3 個分あがった. これたまたま? 10000 個でやり直すべき?
3. 秋元 P は日向坂に櫻坂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?

一般的に言えば,

- 実験・観察により取得した量的 1,2 変数データを統計的に整理して表現し, 他者に対して説明できる.
- データから数理モデルのパラメタを推測して, 根拠とともに他者に説明できる.
- データから仮説を立てて検証し, 他者を説得できる.

オンライン授業ののり (週サイクル)

Hig3Moodle `https://moodle.hig3.net` のその週のセクションを上からやっていってください. Teams にログイン方法の動画.

- 2020-09-28 月 1(以前) 動画, 資料を Hig3Moodle で公開, メールと Teams **一般 ch** で連絡
- **次の週の** 2020-10-06 火 23:59 **練習問題**の受験期限.
  - 練習問題満点が Trial 受験のための条件
- **次の週の** 2020-10-11 日 23:59 **Trial** の受験期限, 3 回まで受験して最高点, 制限時間あり Trial でなく, 手書き答案提出, 相互採点, Excel の課題提出, などの週もある予定.
- 質問はいつでも Teams **質問と答 ch** または樋口 (a00010) への chat で. 樋口と TA で対応.
- 週に何講時か, TA または教員が音声とテキストで対応する確率統計☆演習**リアルタイムサポート (質問と答 ch)**. 初回は 28 月 4. 質問を書いておいてくれるのも歓迎.

練習問題, Trial は Moodle 上で実行する小テストのこと.

練習問題タイプ　期限まで何回でも受験, 制限時間なし.「チェック」で問ごとの正誤がわかりフィードバックがある. 誤答のあと正答だと減点.

Trial タイプ　期限まで 3 回だけ, 制限時間あり, 提出して初めて全体の合計の点数がわかる. 個々の問の正誤, 正答は表示しない.

オンライン授業ののり (成績計算)

科目の成績 100 ピーナツ =90 ピーナツ (15 回の Trial の平均得点率 ×90)
+ 10 ピーナツ (レポートや作文など)

まあ, Trial 1 回 6 ピーナツみたいなもの.
ただし, Trial 90 ピーナツでの得点率について大注意

- Trial の L01-L05, L06-L10 回, L11-15 の **3 つの平均得点率がすべて 0.5 を以上であることを合格の条件**とします. 記述統計-確率-推測統計の 3 分野すべて理解してないとだめ, という趣旨. 0.5 未満のものがあったら, 不合格で, **最低の**平均得点率を採用して 90 ピーナツを計算します.
- 上のルールと無関係に, 90 ピーナツ分の平均得点率を, 統計検定 3 級 (2020 年 4 月以降受験, 2021 年 1 月 31 日提出まで) のスコアレポートの得点率で置きかえられます. 受験料は各自負担.

オンライン授業ののり (教科書やその他の準備)

教科書 必須です. 岩薩林 確率・統計
岩佐-薩摩-林 理工系の数理 確率・統計, 裳華房 (2018)
L01 練習問題で教科書の図を参照しています.
L02 は教科書 岩薩林 確率・統計 §1.2 読んでおいて.

ノート

個人のやり方にお任せしますが, 板書を写す的なことは不要だろうと思います. 教科書や印刷した授業資料に書き込む, 問題を解くときの過程を 1 冊のノートにまとめて残す, などをお奨めします.

PC

Excel 使うのでスマホでは済まない回があります.
Excel は龍大の Office365 で無料でインストール. 中野先生の案内

ここまで来たよ

1 変数の量的データ

2017 年 9 月頃 (?) の某アイドルグループの身長

01 171.02cm
02 166.93cm
03 165.46cm
⋮
49 151.48cm

`https://nogizaka46.infonet.site/height.html`

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)

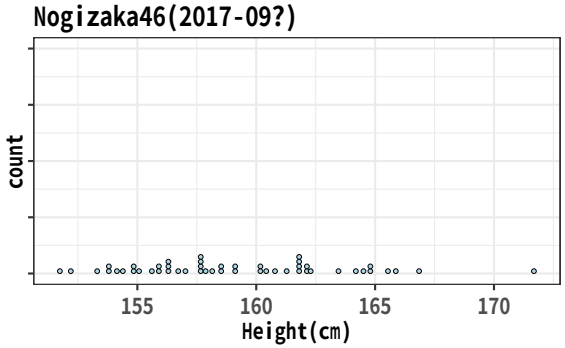
- オーディションにおいて, 身長は考慮されているか?
- オーディション基準の身長は期生ごとに違うか?
- ⋮

ストリップチャート

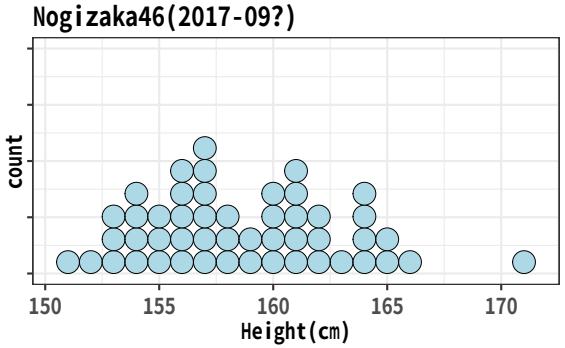
実軸上に, データに対応する点をマークする. 積み重ねる.

もっとデータが多くなったらどうするの?

小数点以下 2 桁まで



0 桁まで (整数)



本当は身長は無限桁ある実数だし, 身長計の測定誤差や有効数字もあるけど, しばらく厳密な有限小数であるかのように思おう

ここまで来たよ

度数分布表

高校 数学 I | 岩薙林 確率・統計 §1.1 では縦横逆に書いてる．どっちでも．

階級	階級値	度数	相対度数
145 より大きく 150 以下	147.5	0	0.00
150 より大きく 155 以下	152.5	12	0.24
155 より大きく 160 以下	157.5	20	0.41
160 より大きく 165 以下	162.5	17	0.35
165 より大きく 170 以下	167.5	1	0.02
170 より大きく 175 以下	172.5	1	0.02
175 より大きく 180 以下	177.5	0	0.00
合計		49	1.00?

- 元のデータより情報は減ってるけど見やすい
- 問: 身長が上から 5 位のメンバーは身長何 cm?
- 問: 身長の順位がちょうど真ん中のメンバーは身長何 cm?

度数分布表の作り方

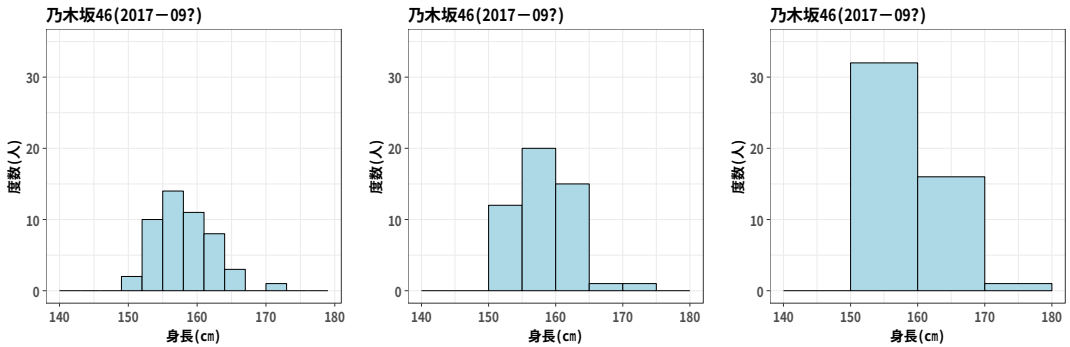
高校 数学 I 岩薩林 確率・統計 §1.1

- 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, 切りのいい値に自分で決めていい.
- 度数=階級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=度数/n. 端数で合計 1.00 にならないかも. 気にしてない.
- 階級値=その階級のまん中の値
- 以下, 以上, 未満 (=より小さい), より大きい

- 実数値なので 146 以上 150 以下, ではだめ

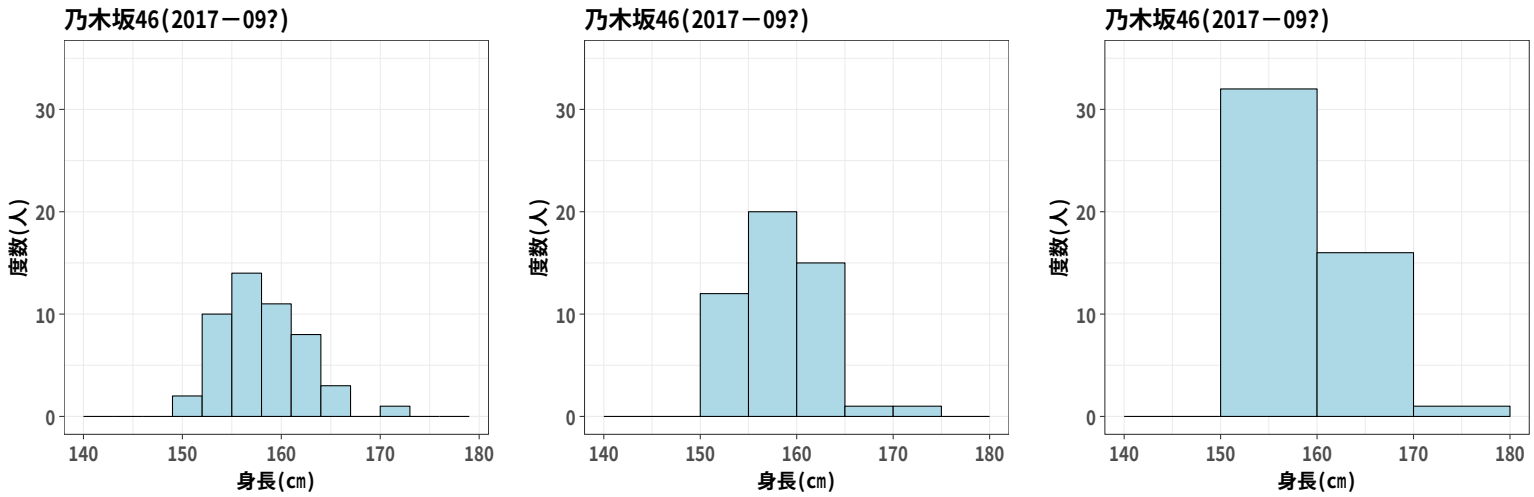
ここまで来たよ

度数のヒストグラム 高校 数学 I 岩薙林 確率・統計 §1.1



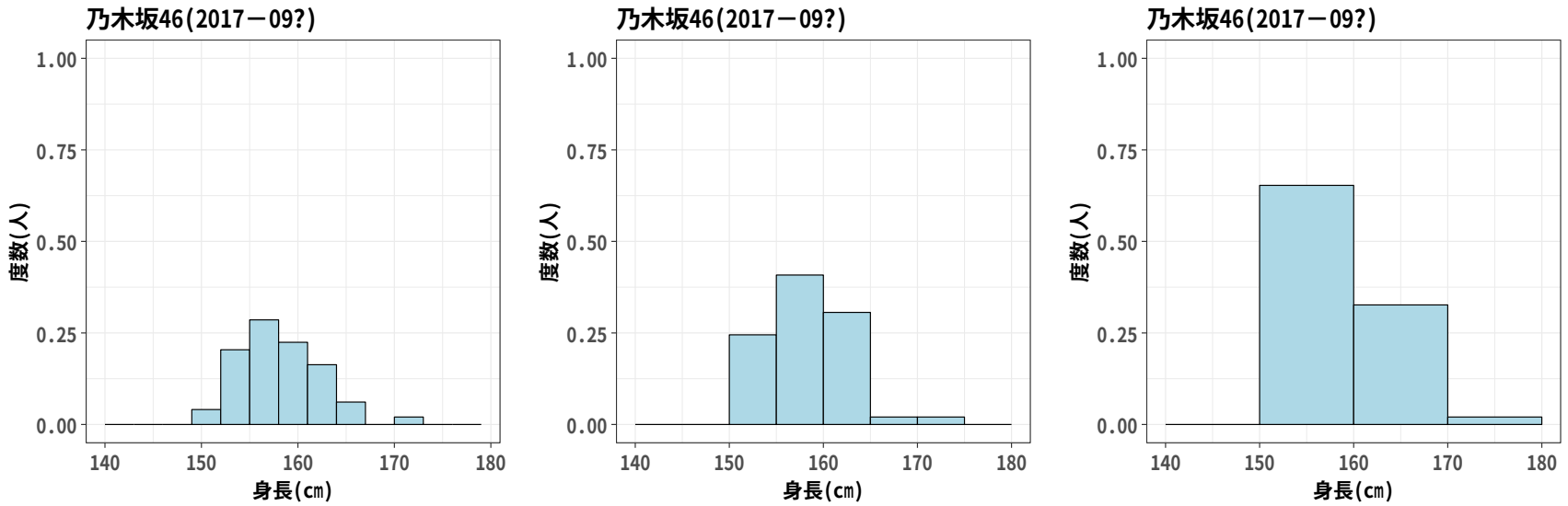
- 問: 身長が上から 5 位のメンバーは身長何 cm?
- 問: 身長の順位がちょうど真ん中のメンバーは身長何 cm?

- 度数分布表を '棒グラフ' にしたもの'.
- 元のデータより情報は減ってるけど見やすい
- 基本レベルでは階級幅は一定 $\rightsquigarrow$ 本当は高さでなく面積
- 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 $\rightsquigarrow$
  - `https://shiny.rstudio.com/gallery/faithful.html`
- ヒストグラムに限らず, グラフの縦軸横軸には量の名と単位を明示
- 棒の間にすきまなし. 値は棒と棒の境い目. 度数 zero の階級も消さない.



- 問: 身長が上から 5 位のメンバーは身長何 cm?
- 問: 身長の順位がちょうど真ん中のメンバーは身長何 cm?

相対度数のヒストグラム



高校 数学 I　岩薩林 確率・統計 §1.1

- 問: 身長の順位がちょうど真ん中のメンバーは身長何 cm?
- 問: 身長が下から 25%のメンバーは身長何 cm?

L01-Q1

Quiz(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手で作ろう. 2017-09 時点?

`http://nogizaka46.infonet.site/height.html`

名前	年齢
梅澤美波	18.73
斎藤ちはる	20.61
伊藤純奈	18.83
中村麗乃	16.00
相楽伊織	19.84
能條愛未	22.95
山崎怜奈	20.36
新内眞衣	25.69
橋本奈々未	24.61
衛藤美彩	24.74
深川麻衣	26.51
松村沙友理	25.09
白石麻衣	25.11
高山一実	23.64
吉田彩乃ク	22.07
佐々木琴子	19.09
阪口珠美	15.88
生田絵梨花	20.68
堀未央奈	20.96
佐藤楓	19.52
山下美月	18.18
西野七瀬	23.35
久保史緒里	16.21
川後陽菜	19.52
永島聖羅	23.37
中元日芽香	21.46
中田花奈	23.15
樋口日奈	19.66
若月佑美	23.26
和田まあや	19.44
北野日奈子	21.20
鈴木絢音	18.57
秋元真夏	24.11
川村真洋	22.19
斉藤優里	24.20
桜井玲香	23.38
大園桃子	18.04
伊藤万理華	21.61
寺田蘭世	19.02
岩本蓮加	13.65
伊藤かりん	24.35
井上小百合	22.79
齋藤飛鳥	19.14
伊藤理々杏	14.97
生駒里奈	21.75
向井葉月	18.10
与田祐希	17.40
星野みなみ	19.64
渡辺みり愛	17.91

- 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.

L01-Q2

岩薩林 確率・統計演習問題 1.1