

# データの分布

樋口さぶろお <https://hig3.net>

龍谷大学工学部数理情報学科

確率統計☆演習 I L01(2019-09-23 Mon)

最終更新: Time-stamp: "2019-09-23 Mon 13:58 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



## ここまで来たよ

- はじめに
  - この授業どんなのり?
  
- ① データの分布
  - データとは?
  - 度数分布表
  - ヒストグラム

## 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 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。

- 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, 人工知能 (AI), 機械学習が熱い! 精神は同じ
- 統計は科学技術の言葉 ⇨ 数理卒は当然期待されてる
- 統計検定 2,3 級 (瀬田学舎での 40%ディスカウント受験受付中)



## 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うていいの?
- 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は日向坂に櫛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?

## 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 Trial, 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
- 70 ピーナッツ: 定期試験. ただし, A パート, B パート からなり, A, B のうち低い方を採用します.
  - ▶ パート A は, 11 月にも実施します. A1, A2 のうち高いほうを採用します. 統計検定 3 級の合格により一部の問を正解とみなします.
  - ▶ パート B は, 1 月の定期試験前にも実施します. B1, B2 のうち高いほうを採用します. 統計検定 2 級の合格により全問正解とみなします.

定期試験 70 ピーナッツ分 =  $\min(\max(A1, A2), \max(B1, B2))$ .

**欠席届** 毎回出席を前提に進めます. やむを得ず欠席して, 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.ryukoku.ac.jp](mailto:hig-probstat@math.ryukoku.ac.jp)
- へや: 1-507
- 樋口オフィスアワー木 6(1-539) 金昼 (1-542), Math ラウンジ (1-536/538)
- Web ページ: <https://hig3.net> 演習の指示や, スケジュールもここから.



<https://hig3.net>

- → 確率統計☆演習 I (配布資料).
- → Note Math Moodle (全学認証だけど高橋先生とは別) → 確率統計



## 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)

イヤフォン実習室の回. PCにつながるもの. Bluetooth や Lightning はだめ.

電卓平方根 (ルート) のあるもの. スマホアプリでも可.

## 1 週間のタイムライン (学期前半)

- ① 火 09:00 ごろまでに 予習復習問題 (=Trial 予想問題) を Note Math Moodle で公開. 何回でもできるけど, 2 回目以降ちょっとずつ基礎点減少. Trial までの最高点を記録.
- ② 月 1 の最初 Trial(=小テスト) 紙 or PC 参照不可 相談不可 (水 1 までにやったことについて, 水 1 の最後に Trial するパターンも)
- ③ 月 1 チーム別エリア座席指定. 講義のような演習のような. チームで何かやったり.
- ④ 月 1 の最後 来週の Trial の予告
- ⑤ 樋口オフィスアワー木 6(1-539) 金昼 (1-542), Math ラウンジ (1-536/538)

## ここまで来たよ

- はじめに
  - この授業どんなのり?
- ① データの分布
  - データとは?
  - 度数分布表
  - ヒストグラム

# 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>

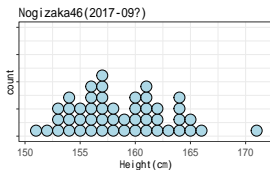
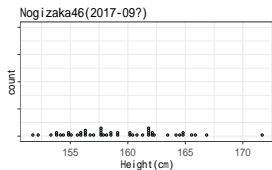
このコースの最後までいくと問えること (正確な表現ではありません)

- 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はグループごとに違うか?
- ⋮

## 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?



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、しばらく整数であるかのように思おう

## ここまで来たよ

- はじめに
  - この授業どんなのり?
- ① データの分布
  - データとは?
  - 度数分布表
  - ヒストグラム

## 度数分布表

高校 数学 I 岩薩林 確率・統計 §1.1 では縦横逆に書いている。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



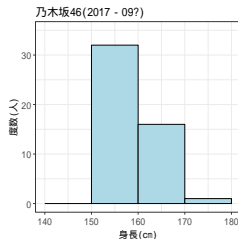
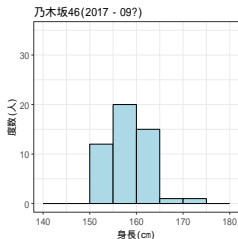
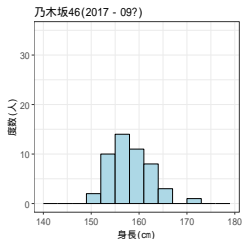
## ここまで来たよ

- はじめに
  - この授業どんなのり?
- ① データの分布
  - データとは?
  - 度数分布表
  - ヒストグラム

# 度数のヒストグラム

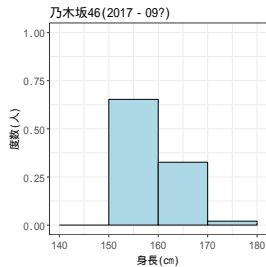
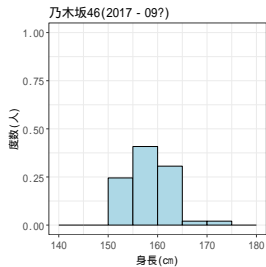
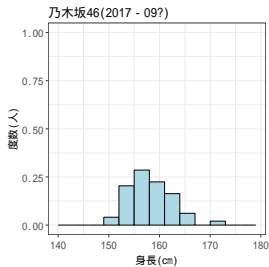
高校 数学 I

岩薩林 確率・統計 §1.1



- 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

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

## 連絡

- 次回はたぶん 3-202 講義室
- 樋口オフィスアワー木 6(1-539) 金昼 (1-542), Math ラウンジ (1-536/538)
- Trial 予告
- 来週は教科書 岩薩林 確率・統計 §1.2 読んできて。
- 統計検定. 2019-11-24 日 40%ディスカウント(団体受験受付中).

## 統計検定受験登録



## Moodle モバイルアプリ



で URL 指定 <https://note.math.ryukoku.ac.jp/moodle>  
または Safari/Chrome で.

